

Introduction

This book takes a hands-on, example driven approach to the teaching of techniques required to analyse epidemiological data. No previous knowledge of statistics is required but we assume you are familiar with some basic epidemiological terms. The step-by-step instructions will enable you to reach a high level of statistical proficiency in a short period of time. Powerful and easy-to-use software is used to analyse real data and to illustrate statistical techniques. The aim is to illustrate those methods commonly employed by epidemiologists and medical statisticians, and to provide an opportunity to practise the application of these methods in a logical sequence. As such, this book is suitable as an introductory or refresher course for those studying or working in epidemiology, public health medicine, or environmental health.

You are advised to work through all sections of the book. We suggest that you either work through the material in a single sitting or in three separate sessions:

Session 1	Pages 1 - 37
------------------	---------------------

Identifying exposures or foods that are associated with illness in an outbreak of food-poisoning. The relative risk, odds ratio, confidence interval, chi-square test and test for linear trend are described and demonstrated using the course software.

Session 2	Page 38 - 51
------------------	---------------------

Measuring the simultaneous effect of two exposures on illness using stratified analysis. Confounding and interactions are defined and illustrated. The Mantel-Haenszel method of adjusting for confounding and testing for the significance of adjusted relative risks and odds ratios is described and illustrated using the course software.








Session 3	Page 52 - 88
------------------	---------------------

Advanced statistical methods. Building on the techniques introduced in sessions 1 and 2, the reader is now able to apply logistic regression to measure the simultaneous effect of several exposures on illness.

Learning to use epidemiological and statistical techniques should be an active process. Whilst reading this book you should have the course software running on your personal computer. Take time to become familiar with both the software and the techniques introduced.

Typographical conventions

Several typographical conventions are used throughout this book. The following table illustrates these conventions:

Typeface	Use	Example
Title	This typeface is used for page and section headings.	Indicator variables
VARIABLE	Variable names are shown in upper-case when referred to in the text.	toxin from the VANILLA ice-cream
MODULE	Modules of the course software are also shown in upper-case.	use ANALYSIS to obtain the table
emphasis	Bold letters are used for emphasis.	increased risk or a protective effect
<i>term</i>	The first occurrence of a statistical, epidemiological, or technical term is italicised.	an <i>interaction term</i> for EGGS and
output	This typeface is used to show output from one of the modules of the course software. Bold letters are used for emphasis.	Likelihood ratio = 6.7589
calculations	The same typeface is also used to show calculations, formulae, and worked examples.	$c / d = 3 / 18 = 0.17$
commands	This typeface is used to show commands which should be typed exactly as shown.	tables milk ill
 + 	This typeface is used to show keystrokes. Combinations are represented using a + sign.  +  means you should hold down the  key and press  .	Press  to leave ANALYSIS

The course software


The course software is supplied on a floppy disk and is made up of three separate modules. These are the ANALYSIS and STATCALC modules of Epi-Info, and a program, LOGISTIC, that performs logistic regression on data stored in Epi-Info .REC files.

To use the course software place the supplied disk in the **a:** drive of your computer and make this drive the default drive by typing:




a :

at the MSDOS prompt. Type:

episoft

at the MSDOS prompt and press  to start the course software. You will be presented with the main menu:

```
+----- EpiSoft -----+
|  Epi-Info ANALYSIS  |
|  Epi-Info STATCALC   |
|  LOGISTIC Regression  |
|      Quit             |
+-----+
```

To select an item from this menu you must first move the highlighted bar using the  and  keys. When the highlighted bar is over the name of the module you wish to use press the  key to select it. The selected module will then start. To stop using the course software select the menu option *Quit*. This will return you to the MSDOS prompt.

The course software is designed to run on an IBM compatible personal computer running MSDOS. You may also run the course software in a DOS window under Microsoft Windows, IBM OS/2 or using Soft-PC on a Macintosh, PowerPC, or UNIX machine.

The logistic regression module of the course software was written by Gerard Dallal and is part of his excellent StatTools™ suite of low-cost statistical and epidemiological software. The other programs in the StatTools™ suite are described in the file TOOLS.DOC on the course software disk. If you would like more information about StatTools™ contact:

Gerard E Dallal
54 High Plain Road
Andover, MA 01810
USA

You should acknowledge LOGISTIC (the logistic regression module of the course software) in any manuscript that makes use of its calculations. A suitable reference is:

Dallal GE (1988), 'LOGISTIC: A Logistic Regression Program for the IBM PC',
The American Statistician, 42, 272.

The file LOGISTIC.DOC on the course software disk contains the manual for the LOGISTIC module of StatTools™. You are free to use and copy the course software for non-commercial purposes only.

Sample datasets

Throughout this course you will work with data from three different outbreaks of food-poisoning and a study into the risk factors for HIV-2 infection. Each dataset, together with the name and structure of the sample data files, is described below. A sample dataset is included so that you may practise the techniques introduced in this course.

The OSWEGO outbreak

After a church supper in the small village of Lycoming, Oswego County, New York on April 18th, 1940, a majority of the seventy-five participants became ill with gastroenteritis. Data from this outbreak is stored in the file OSWEGO.REC. The variables in the file are:

ILL	Ill?
HAM	Baked ham
SPINACH	Spinach
POTATO	Potato salad
CABBAGE	Cabbage salad
JELLY	Jelly
ROLLS	Bread rolls
BREAD	Brown bread
MILK	Milk
COFFEE	Coffee
WATER	Water
CAKES	Cakes
VANILLA	Vanilla ice-cream (home-made)
CHOCOLATE	Chocolate ice-cream (home-made)
FRUIT	Fruit salad

Data is available for all seventy-five people who attended the church supper.

The BATEMAN outbreak

On Saturday, 21st April 1990, a luncheon was held in the home of Jean Bateman. There was a total of forty-five guests which included thirty-five members of the Department of Epidemiology and Population Sciences at the London School of Hygiene and Tropical Medicine. On Sunday morning, 22nd April 1990, Jean awoke with symptoms of gastrointestinal illness; her husband awoke with similar symptoms. The possibility of an outbreak related to the luncheon was strengthened when several of the guests telephoned Jean on Sunday and reported illness. On Monday, 23rd April 1990, there was an unusually large number of department members absent from work and reporting illness. Data from this outbreak is stored in the file BATEMAN.REC. The variables in the file are:

ILL	Ill?
CHEESE	Cheddar cheese
CRABDIP	Crab dip
CRISPS	Crisps
BREAD	French bread
CHICKEN	Chicken (roasted, served warm)
RICE	Rice (boiled, served warm)
CAESAR	Caesar salad
TOMATO	Tomato salad
ICECREAM	Vanilla ice-cream
CAKE	Chocolate cake
JUICE	Orange juice
WINE	White wine
COFFEE	Coffee

Data is available for all forty-five guests at the luncheon.

The SALEX outbreak

On Saturday 17th October 1992, eighty-two people attended a buffet meal at a sports club. Within fourteen to twenty-four hours fifty-one of the participants developed diarrhoea, with nausea, vomiting, abdominal pain and fever as other common symptoms.

Data from this outbreak is stored in the file SALEX.REC. The variables in the file are:

CASE	Case or control
HAM	Baked ham
BEEF	Roast beef
EGGS	Eggs
MUSHROOM	Mushroom flan
PEPPER	Pepper flan
PORKPIE	Pork pie
PASTA	Pasta salad
RICE	Rice salad
LETTUCE	Lettuce
TOMATO	Tomato salad
COLESLAW	Coleslaw
CRISPS	Crisps
PEACHCAKE	Peach cake
CHOCOLATE	Chocolate cake
FRUIT	Tropical fruit salad
TRIFLE	Trifle
ALMONDS	Almonds

Data is available for seventy-seven of the eighty-two people who attended the sports club buffet. This data was collected using a *case-control study* design.

The GUD / HIV study

Several studies have documented an association between genital ulcer disease (GUD) and HIV infection. A study of Gambian prostitutes documented an association between seropositivity for HIV-2 and antibodies against *Treponema pallidum* (a serological test for syphilis). Prostitutes are not the ideal population for such studies as they may have experienced multiple sexually transmitted infections and it is difficult to quantify the number of times they may have had sex with HIV-2 seropositive customers. A sample of males with sexually transmitted infections is easier to study as they have probably had fewer sexual partners than prostitutes and much less contact with sexually transmitted infection pathogens. In such a sample it is also easier to find and collect data.

The data stored in the file GUDHIV.REC has been adapted from a cross-sectional study of 435 male patients who presented with sexually transmitted infections at an outpatient clinic in The Gambia between August 1988 and June 1990. The variables in the file are:

MARRIED	Married
GAMBIAN	Gambian Citizen
GUD	History of genital ulcer disease (GUD) or syphilis
UTIGC	History of urethral discharge or gonorrhoea
CIR	Circumcised
TRAVOUT	Travelled outside of Gambia and Senegal
SEXPRO	Ever had sex with a prostitute
INJ12M	Injection in previous 12 months
PARTNERS	Number of sexual partners in previous 12 months
HIV	HIV-2 positive serology

Data is available for all 435 patients enrolled in the study.

The REDTIDE outbreak

On 30th July 1988, 120 cases of sudden neurological symptoms and twenty-six deaths occurred in a small fishing village in Guatemala. Paralytic shellfish poisoning caused by 'red tide' (toxic dinoflagellates) was suspected. The outbreak was referred to the Pan-American Health Organisation and investigated by epidemiologists from the Centers for Disease Control and Prevention (CDC).

Data from this outbreak is stored in the file REDTIDE.REC. The variables in the file are:

CASE	Case or Control
MILK	Milk
BEER	Beer (bottled)
COLDDRINK	Soft drink (bottled)
WATER	Water
BREAD	Bread
ARROZ	Rice
BEANS	Beans (re-fried)
CHEESE	Cheese
FISH	Fish (local caught)
SHRIMP	Shrimp (local caught)
LOBSTER	Lobster (local caught)
CLAMS	Clams (local caught)
CHICKEN	Chicken
BROTH	Broth
LOBSTRSOUP	Lobster soup (local caught)
CLAMSOUP	Clam soup (local caught)
VEGISOUP	Vegetable soup (freshly prepared)

This data was collected using a *case-control study* design and is included with the course software so that you may practise the techniques introduced in this course.

Measures of disease and exposure effects: 2-by-2 tables

Measures of disease and *exposure effects* are easily studied by means of *2-by-2 contingency tables*. Such a table is presented below:

		Outcome		
Exposure		Cases	Non-Cases	Totals
Present	a	b	a + b	
Absent	c	d	c + d	
Totals	a + c	b + d	a + b + c + d	

a = number exposed and ill
 b = number exposed and not ill
 c = number unexposed and ill
 d = number unexposed and not ill
 a + b = number exposed
 c + d = number unexposed
 a + c = number ill
 b + d = number not ill
 a + b + c + d = number in study

By organising data in this way it is possible to use a wide range of simple and powerful analytical techniques. The *2-by-2 table* format used above (*exposure* categories in rows and *outcome* categories in columns) and the method of labelling cells (a, b, c, and d) are used throughout this course.

The following table is from the Oswego Church Supper outbreak:

		ILL		
JELLY		+	-	Total
	+	16	7	23
	-	30	22	52
	Total	46	29	75

The exposure variable is JELLY and the outcome variable is ILL. The following values correspond to our notation:

a = **16** = number exposed and ill
 b = **7** = number exposed and not ill
 c = **30** = number unexposed and ill
 d = **22** = number unexposed and not ill
 a + b = **23** = number exposed
 c + d = **52** = number unexposed
 a + c = **46** = number ill
 b + d = **29** = number not ill
 a + b + c + d = **75** = number in study

Make sure that you can identify these figures in the table before proceeding.

Absolute risk

The *absolute risk* is the risk of an outcome to an individual belonging to a given *population*. This can be calculated from a 2-by-2 table as:

$$(a + c) / (a + b + c + d)$$

This is the number of people experiencing illness divided by the *total population at risk*. The following table is from the Oswego outbreak and shows the distribution of ill and not-ill outcomes among those who ate and did not eat VANILLA ice-cream:

VANILLA		ILL		Total
		-	+	
	+	11	43	54
	-	18	3	21
Total		29	46	75

The absolute risk can be calculated as:

$$(a + c) / (a + b + c + d) = (43 + 3) / (43 + 11 + 3 + 18) = 46 / 75 = 0.613$$

Each person attending the Oswego Church Supper ran a risk of 61% (0.61) of developing food-poisoning. This is the same as saying that 61% of people attending the Oswego Church Supper fell ill. The absolute risk makes no reference to a *risk factor* or exposure variable and remains the same for each tabulation of exposure by outcome. The following table shows the distribution of ill and not-ill outcomes among those who drank and did not drink MILK:

MILK		ILL		Total
		-	+	
	+	2	2	4
	-	27	44	71
Total		29	46	75

In this table the absolute risk can be calculated as:

$$(a + c) / (a + b + c + d) = (2 + 44) / (2 + 2 + 44 + 27) = 46 / 75 = 0.613$$

The absolute risk measures the *absolute magnitude* of a health problem in a population but provides **no** information regarding the association between risk factors and outcomes. It is simply the proportion of ill individuals in the *study population*. It is only valid to calculate an absolute risk with data from a *cross-sectional* or *cohort* study. It is **not** valid to calculate an absolute risk in this way with data from a *case-control* study (i.e. where members of the study population are selected according to their disease status). In a case-control study of the Oswego outbreak using 20 cases (a + c) and 20 controls (b + d) the absolute risk would appear to be:

$$(a + c) / (a + b + c + d) = 20 / 40 = 0.50$$

If 10 cases (a + c) and 20 controls (b + d) had been selected the absolute risk would appear to be:

$$(a + c) / (a + b + c + d) = 10 / 30 = 0.33$$

Both of these estimates differ from the *true* absolute risk among the 75 persons who attended the Oswego Church Supper. It is **not** valid to calculate an absolute risk with data from a case-control study.

Exposure-specific risk or attack rate

A more useful measure than the absolute risk is the *exposure-specific risk* or *attack rate*. This yields a measure that can be interpreted as the risk of an outcome given an exposure which can easily be calculated from a 2-by-2 table:

VANILLA		ILL		Total
		+	-	
	+	43	11	54
	-	3	18	21
Total		46	29	75

In this table the exposure-specific risk of becoming ill having eaten VANILLA ice-cream at the Oswego Church Supper can be calculated as:

$$a / (a + b) = 43 / (43 + 11) = 43 / 54 = 0.79$$

Each person attending the Oswego Church Supper ran a risk of 79% (0.79) of developing food-poisoning **if they consumed VANILLA ice-cream**. This is the same as saying that 79% of people who ate VANILLA ice-cream fell ill. This method makes specific reference to risk factors and will vary with each tabulation of exposure by outcome:

MILK		ILL		Total
		+	-	
	+	2	2	4
	-	44	27	71
Total		46	29	75

In this table the exposure-specific risk of becoming ill having consumed MILK can be calculated as:

$$a / (a + b) = 2 / (2 + 2) = 2 / 4 = 0.50$$

The exposure-specific risk *estimates* the risk of an outcome given exposure. It is a measure of *absolute risk* associated with an exposure and reveals little about the *excess risk* associated with exposure.

Relative risk or risk ratio

One way of estimating the excess risk due to exposure is to compare exposure-specific risks between *exposed* and *unexposed* groups. Given the following table:

VANILLA	ILL		Total
	+	-	
+	43	11	54
-	3	18	21
Total	46	29	75

it is possible to calculate the risk of becoming ill having consumed VANILLA ice-cream as:

$$a / (a + b) = 43 / (43 + 11) = 43 / 54 = 0.79$$

And the risk of becoming ill having **not** consumed VANILLA ice-cream as:

$$c / (c + d) = 3 / (3 + 18) = 3 / 21 = 0.14$$

The easiest way of comparing these two risks is as a *ratio*. This is the risk of becoming ill having been exposed **divided by** the risk of becoming ill having not been exposed and is calculated as:

$$(a / (a + b)) / (c / (c + d)) = 0.79 / 0.14 = 5.57$$

A person who consumed VANILLA ice-cream at the Oswego Church Supper was 5.57 times as likely to develop food-poisoning as a person who did not consume VANILLA ice-cream at the Oswego Church Supper.

This ratio of risks is described as the *relative risk* or *risk ratio*. This is always greater than zero. A relative risk of less than one implies a **decrease** in risk associated with a given exposure and that the exposure **may protect** against disease. A relative risk of greater than one implies an **increase** in risk associated with a given exposure. A relative risk of one implies **no** increase or decrease in risk associated with exposure. In summary:

Relative Risk	Interpretation
< 1	exposure may protect against disease
= 1	exposure not associated with disease
> 1	exposure may increase risk of disease

Relative risks of less than one do occur in food-borne outbreaks. In such cases it is unlikely that exposure protects against disease. It is far more likely to mean that a particular food and the contaminated food did **not** tend to be eaten by the same people.

Odds ratio

The examples given above assume that it is valid to calculate a relative risk of disease. It is only valid to calculate the relative risk of disease in a study which has **not** selected subjects according to their disease status (i.e. a study which has **not** selected *cases* or *controls*). The relative risk of disease is a valid measure of excess risk in a *cohort* study, where subjects have been selected according to their exposure status, or in a *cross-sectional* study.

In a case-control study an arbitrary number of cases and controls are selected for entry into the study. In this situation it is no longer reasonable to use the total numbers of ill and not-ill persons (cases and controls) to calculate exposure-specific risks and relative risks. Another measure of association must be used which uses only the internal (rather than total) values in the table. This measure is the *odds ratio*. Consider the following table from the Oswego Church Supper example:

VANILLA		ILL		Total
		+	-	
+		43	11	54
-		3	18	21
Total		46	29	75

In this example the *odds* of developing food-poisoning given consumption of VANILLA ice-cream can be calculated as:

$$a / b = 43 / 11 = 3.91$$

The odds of developing food-poisoning given no consumption of VANILLA ice-cream can be calculated as:

$$c / d = 3 / 18 = 0.17$$

The odds ratio can be calculated as:

$$(a / b) / (c / d) = 3.91 / 0.17 = 23.45$$

Eating VANILLA ice-cream at the Oswego Church Supper was associated with a 23.45 fold increase in odds of developing food-poisoning. This is **not** the same as a 23.45 fold increase in risk.

Odds ratio and relative risk

The odds ratio and relative risk measure excess risk in an exposed group compared to an unexposed group. In cohort and cross-sectional studies the odds ratio will *overestimate* the effect of exposure and the relative risk is preferred. In case-control studies the relative risk can **not** be estimated and the odds ratio must be used. In case-control studies of rare diseases the odds ratio provides a valid estimate of the relative risk.

The relative risk is a valid measure of excess risk to use with the Oswego Church Supper data because **all** seventy-five people who attended the supper were included in the study. It would not be valid to use the relative risk if a certain number of cases and controls had been selected for the study (e.g. 20 cases and 20 controls). In this case the odds ratio would be the valid measure of excess risk.

Epi-Info ANALYSIS

In this exercise you will use the Epi-Info ANALYSIS module to produce 2-by-2 tables using the OSWEGO dataset. You may need to use a calculator to determine exposure-specific risks (attack rates) and relative risks.

To complete this exercise you should be seated at your computer. Follow the step-by-step instructions to complete each task. You will also need a calculator and a pen or pencil.


Getting started

Start the course software. To use the course software place the course disk in the **a:** drive of your computer and make this drive the default drive by typing:


a:

at the MSDOS prompt. Type:

episoft

at the MSDOS prompt and press  to start the course software. You will be presented with the main menu:

```
+----- EpiSoft -----+
| Epi-Info ANALYSIS    |
| Epi-Info STATCALC    |
| LOGISTIC Regression   |
|           Quit        |
+-----+
```

Select **Epi-Info ANALYSIS**. To select an item from this menu you must first move the highlighted bar using the up and down arrow keys. When the highlighted bar is over the name of the module you wish to use press the  key to select it. The selected module will then start.

Epi-Info ANALYSIS

Once ANALYSIS has started the screen will be divided into distinct sections. On the top few lines of the screen will be the status information:

```
Dataset: <none>                               Free memory: 201K
Use READ to choose a dataset
```

The status information shows the name of the current data file and the amount of free memory in the system that ANALYSIS can use (this figure will depend upon how your computer has been configured). The bottom line of the screen shows the functions keys available from within Epi-Info ANALYSIS.

The rest of the screen is divided into two parts. The upper portion is labelled `Output`. This part of the screen displays the results (e.g. tables, statistics, and listings) of any commands entered. The label `Screen` next to the label `Output` tells us that the output of any commands we issue will be sent to the screen. You can send output to the screen, a printer, or a file on disk (using the **route** command).

The lower portion of the screen is labelled `Commands`. This is where you enter commands.

The `F1` key provides help. Press `F1` at any time for help on what you are doing. The help is context sensitive. If you type a command and then press `F1`, ANALYSIS will respond with help for that particular command. A menu of commands is available by pressing the `F2` key. A menu of variables in the current data file is available by pressing the `F3` key.

Entering commands in ANALYSIS

There are two methods of issuing commands to the ANALYSIS module. You may either type the command directly at the keyboard or choose commands and variables from menus.

Pressing the `F2` key brings up a menu of available commands. To select a command from the menu you must first move the highlighted bar using the `↑`, `→`, `↓`, and `←` keys. When the highlighted bar is over the command you wish to use press `ENTER` to select it. To execute the command press `ENTER` again. With most commands you will need to specify at least one variable.

Pressing the `F3` key brings up a menu of variables in the current dataset. Variables can be selected from the menu in the usual way. To select just one variable point to it using the `↑`, `→`, `↓`, and `←` keys and press `ENTER`. To select a group of variables point to each one in turn and press the `+` key. When you have selected all the variables press `ENTER`. If you select a variable by mistake you can deselect it using the `-` key.

Getting help on commands

There are two ways of getting help on commands in ANALYSIS. If you press `F1` before you have typed a command then ANALYSIS will present you with a menu with options for help on general topics and specific commands. Options are selected from the menu in the usual way. If you want help on a specific command then type the command (e.g. **freq**) and press `F1`. ANALYSIS will respond with help for the command you typed. If there is more than one screen of help on a particular topic or command then `PgDn` will be displayed in the bottom right hand corner of the help window. Pressing `PgDn` will display the next screen of information. To return to ANALYSIS press `ESC`.

Choosing a data file

Before performing any analysis we must first tell ANALYSIS which data file to work with (in this case OSWEGO.REC). The ANALYSIS command to retrieve a data file is **read** followed by the filename. Enter the command:

```
read oswego.rec
```

Once you have retrieved a dataset the status lines on the screen will change to show the name of the dataset and the number of records (observations) in the dataset:

```
Dataset:  A:\OSWEGO.REC (75 records)
Criteria: All records selected
```

The file OSWEGO.REC contains data collected from an outbreak of food-poisoning that followed a church supper in a small North American village.

For Epi-Info data files (created using the ENTER module of Epi-Info) you do not need to specify the .REC extension to the filename. ANALYSIS can also read files created using dBase. To retrieve a dBase file you must specify the .DBF extension to the filename rather than the .REC extension.

If you cannot remember the name of the file you wish to retrieve then issue the **read** command without a filename. ANALYSIS will respond with a menu listing all .REC files in the current directory. Issue the command **read *.dbf** to see a menu listing all .DBF (dBase) files in the current directory.

Examining data

Enter the command:

variables

This command displays a list of *variables* in the current data file together with information about their *type* (numeric, alphanumeric, date, yes/no) and *length* (how many numbers or letters a variable may hold). If ANALYSIS displays the message <more> at the bottom of the output screen press any key (apart from the **ESC** key) to see the next screen of the output.

In the OSWEGO dataset the ILL variable is the outcome of interest and the food variables (HAM, SPINACH, POTATO, CABBAGE, JELLY, ROLLS, BREAD, MILK, COFFEE, WATER, CAKES, VANILLA, CHOCOLATE, FRUIT) are the exposure variables. If you examine the output of the VARIABLES command you will notice that these variables are all of the *Yes/No* type. This means that each variable can hold one of two values (Y or + for YES, N or - for NO). *Yes/No* variables are sometimes called *binary variables*. You can examine the distribution of these variables using the **freq** command. Enter the command:

freq ill

to display a table listing *frequencies*, *percentages*, and *cumulative percentages* of the ILL variable:

ILL	Freq	Percent	Cum.
+	46	61.3%	61.3%
-	29	38.7%	100.0%
Total	75	100.0%	

Seventy-five (75) people attended the church supper. Forty-six (46) of these went on to develop food-poisoning. The proportion of ill people in the total population at risk is:

$$46 / 75 = 0.613 = 61.3\%$$

This is the absolute risk. Each person attending the Oswego Church Supper ran a risk of 61% of developing food-poisoning. This is the same as saying that 61% of people attending the Oswego Church Supper fell ill.

Examine the distribution of the exposure variables using the **freq** command with the command:

freq *

The asterisk (*) symbol can be used with the **freq** and **tables** commands to indicate all variables in the current dataset.

If the commands you issue produce output that will not fit onto a single screen then ANALYSIS will pause at the end of each screen of output and display the message <more> at the bottom of the output screen. If this happens press any key (apart from the **ESC** key) to see the next screen of the output. You can scroll through the output generated by previous commands using the **PG UP** and **PG DN** keys. These move through the output one screen at a time. You may also use **CTRL** + **PG UP** and **CTRL** + **PG DN** to move through the output one line at a time. You can scroll through previously entered commands using the **↑** and **↓** keys.

Using ANALYSIS to produce 2-by-2 tables

ANALYSIS provides the **tables** command to produce *cross-tabulations* or *contingency tables*. We can use the **tables** command to discover which foods are *associated* with illness for the Oswego Church Supper outbreak. Issue the command:

```
tables rolls ill
```

This command displays a table of the variable ROLLS by the variable ILL:

ROLLS	ILL		Total
	+	-	
+	21	16	37
-	25	13	38
Total	46	29	75

The interpretation of the output of the **tables** command depends on the orientation of the table. The correct format for the **tables** command is:

```
tables <exposure> <outcome>
```

The risk or exposure variable should **always** be the first variable you specify with the **tables** command.

If we examine this table closely we can extract the following information:

number exposed and ill	rolls = + and ill = +	a	21
number exposed and not ill	rolls = + and ill = -	b	16
number unexposed and ill	rolls = - and ill = +	c	25
number unexposed and not ill	rolls = - and ill = -	d	13
number exposed	rolls = +	a + b	37
number unexposed	rolls = -	c + d	38
number ill	rolls = +	a + c	46
number not ill	rolls = -	b + d	29
total number of records		a + b + c + d	75

Identify these figures in the displayed 2-by-2 table.

We can use this information to calculate various risk measures:

absolute risk	(a + c) / (a + b + c + d)	46 / 75	0.61
attack rate	a / (a + b)	21 / 37	0.57
relative risk	(a / a + b) / (c / c + d)	(21 / 37) / (25 / 38)	0.86
odds ratio	(a / b) / (c / d)	(21 / 16) / (25 / 13)	0.68

Check that you understand how to obtain these risk measures presented in this table before proceeding.

ANALYSIS and 2-by-2 tables

Use the **tables** command to investigate the association between the other exposure variables and reported illness. Calculate the absolute risk, attack rate (exposure-specific risk), and relative risk for each of the exposure variables and complete the table below:

	Absolute Risk (a + c) / (a + b + c + d)	Attack Rate a / (a + b)	Relative Risk (a / a + b) / (c / c + d)
HAM			
SPINACH			
POTATO			
CABBAGE			
JELLY			
ROLLS	0.61	0.57	0.86
BREAD			
MILK			
COFFEE			
WATER			
CAKES			
VANILLA			
CHOCOLATE			
FRUIT			

If the commands you issue produce output that will not fit onto a single screen then ANALYSIS will pause at the end of each screen of output and display the message <more> at the bottom of the output screen. If this happens press any key (apart from the **ESC** key) to see the next screen of the output.

You can scroll through the output generated by previous commands using the **PG UP** and **PG DN** keys. These move through the output one screen at a time. You may also use **CTRL** + **PG UP** and **CTRL** + **PG DN** to move through the output one line at a time.

You can scroll through previously entered commands using the **↑** and **↓** keys.

Leave ANALYSIS by issuing the command:

quit

or by pressing the **F10** key.

Confidence intervals

The exposure-specific risk associated with consuming VANILLA ice-cream was 80% (i.e. 80% of those who consumed VANILLA ice-cream subsequently fell ill). VANILLA ice-cream was found to be the only vehicle of food-poisoning at the Oswego Church Supper. But why is it that 20% of persons who consumed VANILLA ice-cream did not become ill? One possible explanation for this is *chance* or *random variation*. Individuals may have varied in their exposure (some individuals consuming more VANILLA ice-cream and more of the responsible organism or toxin) and in their *susceptibility* to disease (some individuals may be more susceptible to disease due to their age, their immune system, or other factors). The same exposure may have a different effect on different individuals. The exposure-specific risk associated with consuming VANILLA ice-cream will be affected by this random variation. If the whole of Oswego County had consumed the same VANILLA ice-cream it is likely that we would have found a slightly different exposure-specific risk due to this random variation.

Measures of disease and exposure effect are subject to random variation. The calculated relative risk will differ from the true population relative risk because of random variation. If the study were to be repeated using **different samples** from the **same population** the calculated relative risk would be slightly different for each sample. The sample relative risk is considered to be the best estimate of the true relative risk. It is called the *point estimate* of the relative risk. The effect of random variation can be accounted for statistically by calculating a range of values around the point estimate of the relative risk that has a specified *probability* of including the true value of the relative risk. The specified probability is called the *confidence level* and is usually 95%. The range of values that the true relative risk could take is called the *confidence interval*. The endpoints (maximum and minimum) of the confidence interval are called the *confidence limits*.

With a 95% confidence interval we are 95% sure that the true relative risk falls within the computed confidence interval. We do not know for certain that the true relative risk lies within the confidence interval. The chances are 95% that the true relative risk lies within the confidence interval. The chances are 5% that the true relative risk lies outside of the confidence interval.

Confidence intervals can be calculated for relative risks, odds ratios, absolute risks and exposure-specific risks. The calculation of the confidence interval is complicated and is best left to purpose-designed computer programs such as Epi-Info.

Interpretation of the confidence interval

The following relative risk and 95% confidence limits have been calculated from the OSWEGO data using the Epi-Info ANALYSIS package. For the association between VANILLA ice-cream and illness:

```
Relative risk of (ILL=+) for (VANILLA=+)          5.57
Greenland, Robins 95% conf. limits for RR  1.94 < RR < 16.03
```

the observed relative risk is 5.57. The true relative risk is likely to lie somewhere between 1.94 and 16.03. It is unlikely that the true relative risk is equal to one. It is reasonable to state that consumption of VANILLA ice-cream was associated with developing food-poisoning. For the association between JELLY and illness:

```
Relative risk of (ILL=+) for (JELLY=+)           1.21
Greenland, Robins 95% conf. limits for RR   0.84 < RR < 1.72
```

the observed relative risk is 1.21. The true relative risk is likely to lie somewhere between 0.84 and 1.72. It is **not** unlikely that the true relative risk is equal to one. It is **not** reasonable to state that consumption of JELLY was associated with food-poisoning.

The observed odds ratio is a point estimate of the true population odds ratio. Confidence intervals can be used in a similar way. The following odds ratio and 95% confidence limits for the association between eating VANILLA ice-cream and developing food-poisoning were calculated from the OSWEGO data using the Epi-Info ANALYSIS module:

```
Odds ratio                                23.45
Cornfield 95% confidence limits for OR   5.07 < OR < 125.19
```

The interpretation of an odds ratio and confidence interval is similar to that of the relative risk. The observed odds ratio is 23.45. The true odds ratio is likely to lie somewhere between 5.07 and 125.19. It is unlikely that the true odds ratio is equal to one. It is reasonable to state that consumption of VANILLA ice-cream was associated with developing food-poisoning.

Testing a hypothesis about association

Measures of disease and exposure effect are subject to random variation. We can allow for this random variation by calculating a confidence interval that is likely to include the true measure of disease or effect. The confidence interval can also be used to test whether this association is likely to be real or whether it is likely to have arisen by chance. **If the confidence interval for the relative risk (or odds ratio) does not include one then the association is likely to be real. If the confidence interval includes one then the association is unlikely to be real.**

Another way of assessing the association between exposure and outcome variables is to use *hypothesis testing* or *significance testing*. Statistical hypothesis testing relies on an assumption called the *null hypothesis*. This states that nothing interesting is happening in the data other than random variation (i.e. that there is no association between an exposure and an outcome). The null hypothesis is used to test an *alternative hypothesis* that states that something interesting or *systematic* is happening in the data. Statistical hypothesis testing involves comparing the observed data with what we would *expect* the data to look like if the null hypothesis were true.

Consider the following table from the Oswego Church Supper outbreak:

VANILLA			ILL		Total
		+	-		
	+	43	11		54
	-	3	18		21
Total		46	29		75

The **null hypothesis** is that consumption of VANILLA ice-cream is **not** associated with illness. The **alternative hypothesis** is that consumption of VANILLA ice-cream is associated with illness. The null hypothesis can be tested by comparing the numbers in each cell of the table with the numbers that we would expect to see if the null hypothesis were true.

From the table we can calculate the proportion of people falling ill (absolute risk):

$$46 / 75 = 0.61 = 61\%$$

If the null hypothesis were true, the *expected* number of people falling ill after eating VANILLA ice-cream would be 61% of 54, or:

$$54 * (46 / 75) = 33.12$$

Expected numbers in the other cells of the table can be calculated in the same way using the row and column totals of the cell in the *observed* table or by using the general formula:

$$\text{expected number} = (\text{row total} * \text{column total}) / \text{overall total}$$

giving the following *expected values*:

$$\text{Expected(a)} = (54 * 46) / 75 = 33.12$$

$$\text{Expected(b)} = (54 * 29) / 75 = 20.88$$

$$\text{Expected(c)} = (21 * 46) / 75 = 12.88$$

$$\text{Expected(d)} = (21 * 29) / 75 = 8.12$$

These expected values can then be compared with the *actual* or *observed* values from the data.

Comparing observed and expected values

Once the expected values have been calculated we can compare the table we observe from the data with the table we would expect to see if the null hypothesis were true:

OBSERVED				EXPECTED			
		ILL				ILL	
VANILLA		+	- Total	VANILLA		+	- Total
	+	43	11 54		+	33.12	20.88 54.00
	-	3	18 21		-	12.88	8.12 21.00
Total		46	29 75	Total		46.00	29.00 75.00

Subtracting the expected values from the values observed in the data gives:

		ILL			
VANILLA		+	- Total		
	+	9.88	-9.88 0.00		
	-	-9.88	9.88 0.00		
Total		0.00	0.00 0.00		

Positive and negative differences cancel each other out so we square the number in each cell to make them all positive numbers:

		ILL			
VANILLA		+	- Total		
	+	97.61	97.61 195.23		
	-	97.61	97.61 195.23		
Total		195.23	195.23 390.46		

Now divide each of these squared values by their expected values:

		ILL			
VANILLA		+	- Total		
	+	2.95	4.67 7.62		
	-	7.58	12.02 19.60		
Total		10.53	16.69 27.22		

The overall total of this table (27.22) is a measure of how much the observed data differs from the data expected under the null hypothesis. It is called the *chi-square statistic*:

$$X^2 = \sum [(Observed - Expected)^2 / Expected]$$

The Greek letter *sigma* (Σ) is used in statistical formulae to denote the sum of a series of numbers. Squaring a number (multiplying a number by itself) is a convenient way of turning negative numbers into positive numbers and is used in many statistical formulae for this purpose.

Under the null hypothesis there is a fixed probability of obtaining this particular chi-square value. If the probability of obtaining 27.22 under the null hypothesis is small then the null hypothesis is unlikely to be true and we would assume that there is an association between VANILLA and ILL. If the probability of obtaining 27.22 under the null hypothesis is large then the null hypothesis might be true and we would not assume that there is an association between VANILLA and ILL.

Chi-squares, degrees of freedom, and p-values

The probability of observing a particular chi-square value is determined by the *chi-square distribution*. There is a different chi-square distribution for each size of table. A large chi-square is more likely to arise from a large table (i.e. a table with many cells) than from a small table (i.e. a table with few cells such as a 2-by-2 table). The more rows and columns in a table the larger the chi-square value is likely to be. This is because there are more cells in which the observed values are *free to vary* from the expected values. The number of cells in which the observed values are free to vary from the expected values is called the *degrees of freedom*. The number of degrees of freedom for a table is calculated using the formula:

$$df = (\text{number of rows} - 1) * (\text{number of columns} - 1)$$

The degrees of freedom in a 2-by-2 table are:

$$\begin{aligned} df &= (\text{number of rows} - 1) * (\text{number of columns} - 1) \\ &= (2 - 1) * (2 - 1) \\ &= 1 \end{aligned}$$

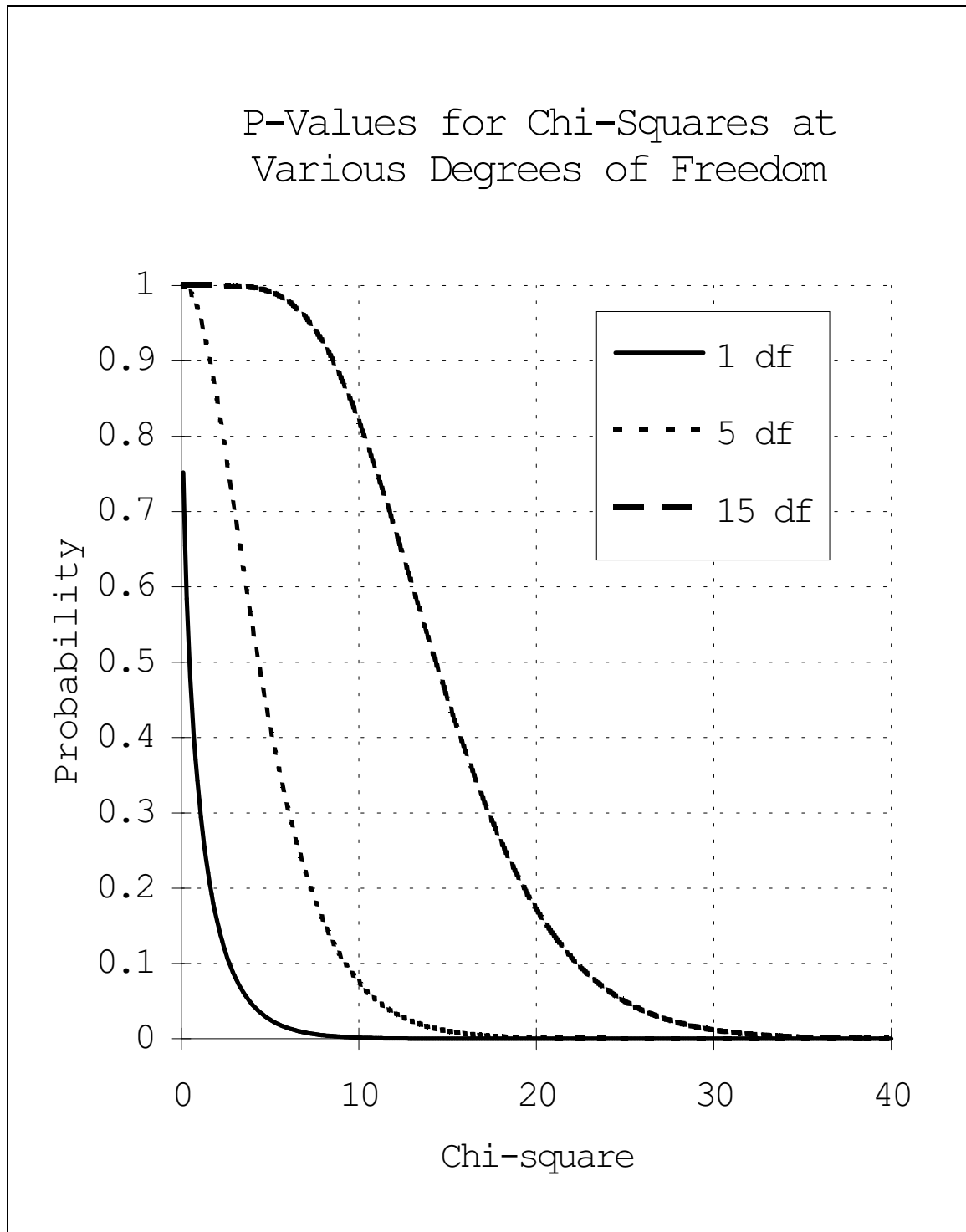
The chart on the next page shows the probability of observing chi-square values for different degrees of freedom. If you look at the chi-square distribution for one degree of freedom you can see that the probability of getting a chi-square value of 27.22 from a 2-by-2 table is very small. This probability is referred to as the *p-value* or *observed significance level*. It is the probability that the observed association arose by chance. If we refer to the chi-square distribution with one degree of freedom in a set of statistical tables we obtain a p-value of less than 0.001. The probability that the observed association arose by chance is less than 0.001. It is very unlikely that the observed association arose by chance. It is reasonable to reject the null hypothesis of no association between VANILLA and ILL. It is generally considered safe to reject the null hypothesis if the p-value is less than 0.05. When the p-value is less than 0.05, the association between the row and column variables is said to be *statistically significant*. Small p-values are **more** significant than large p-values: a p-value of 0.001 is **more** significant than a p-value of 0.01.

There are various chi-square tests. The one described above is called *Pearson's chi-square* and is used to test the association between two variables. Another test of association is *Yates' corrected chi-square* which is based on the same formula but with a *correction factor* to make the resulting chi-square slightly smaller. Yates' corrected chi-square is always slightly smaller than Pearson's chi-square. It always yields a slightly larger p-value than Pearson's chi-square. Most statisticians prefer to use Yates' corrected chi-square because it is more *conservative* and less likely to yield a significant p-value when the null hypothesis is true.

Another test of association is *Fisher's exact test*. This is based on a distribution called the *hypergeometric* distribution and is **not** a chi-square test. Chi-square tests can yield misleading results with small numbers. Fisher's exact test is more *robust* and can yield *reliable* results even with small numbers. Fisher's exact test should be used instead of a chi-square test when any of the cells in the 2-by-2 table has an expected value of less than five.

Chi-squares, degrees of freedom, and p-values

The chart below shows the probability of observing chi-square values for different degrees of freedom. Mark a chi-square value of 27.22 on the x-axis (chi-square axis) and estimate the probability of calculating this value from data in a 2-by-2 table:



Relative risk, confidence intervals, and tests with ANALYSIS

Start the course software and the Epi-Info ANALYSIS module as before. Issue the command:

```
read oswego.rec
```

to retrieve the OSWEGO dataset. Issue the command:

```
set percents = on
```

to instruct ANALYSIS to display row and column percentages in tables. Issue the command:

```
tables milk ill
```

This command now produces the following output:

MILK		ILL		Total
		+	-	
+		2	2	4
>		50.0%	50.0%	> 5.3%
		4.3%	6.9%	
-		44	27	71
>		62.0%	38.0%	> 94.7%
		95.7%	93.1%	
Total		46	29	75
		61.3%	38.7%	

The values in each cell are presented as *count*, *row percentage*, *column percentage*. The row percentages are the percentages in a row that add up to 100% (in this case 50.0% and 50.0%, 62.0% and 38.0%). The column percentages are the percentages in a column that add up to 100% (in this case 4.3% and 95.7%, 6.9% and 93.1%). Examine the table closely. The row percentage of the cell MILK = + and ILL = + is equivalent to the exposure-specific risk or attack rate:

```
exposure-specific risk = a / (a + b)
                        = 2 / 4
                        = 0.500
                        = 50.0%
```

The row percentage for the marginal total ILL = + is equivalent to the absolute risk:

```
absolute risk = (a + c) / (a + b + c + d)
               = (2 + 44) / (2 + 2 + 44 + 27)
               = 46 / 75
               = 0.613
               = 61.3%
```

If you find the tables with percentages difficult to read try issuing the command:

```
set lines = on
```

This command instructs ANALYSIS to print separating lines between the cells of tables.

Use the **tables** command with row and column percentages displayed to show the absolute risk and exposure-specific risk or attack rate for each of the exposure variables. Compare these with those you calculated earlier.

Producing statistics

Epi-Info ANALYSIS can calculate relative risks and odds ratios (with confidence intervals) and chi-square statistics (with p-values). First issue the commands:

```
set percents = off
set lines = off
```

to instruct ANALYSIS not to display row and column percentages and to stop printing separating lines between the cells of tables. Issue the command:

```
set statistics = on
```

This command instructs ANALYSIS to produce a full set of statistics in the output of each command. To see the effect of this command issue the command:

```
tables ham ill
```

This command now produces the following output:

HAM		ILL		
		+	-	Total
	+	29	17	46
	-	17	12	29
Total		46	29	75

Single Table Analysis

Odds ratio 1.20
Cornfield 95% confidence limits for OR 0.41 < OR < 3.50

Relative risk of (ILL=+) for (HAM=+) 1.08
Greenland, Robins 95% conf. limits for RR 0.74 < RR < 1.57
(Biometrics 1985;41:55-68)
Ignore relative risk if case control study.

	Chi-Squares	P-values
Uncorrected:	0.15	0.70170143
Mantel-Haenszel:	0.14	0.70360201
Yates corrected:	0.02	0.88899437

Examine the output carefully and identify the relative risk, the associated confidence interval, the chi-square statistic and its associated p-value. The **Uncorrected** chi-square is Pearson's chi-square. The **Yates corrected** chi-square is preferred for 2-by-2 tables. The **Mantel-Haenszel** chi-square test is **not** appropriate in this context and should be ignored. ANALYSIS will only calculate and display Fisher's Exact test if the 2-by-2 table yields any expected values that are less than five.

The interpretation of the relative risk produced by the **tables** command depends on the orientation of the table. The correct format for the **tables** command is:

```
tables <exposure> <outcome>
```

The risk or exposure variable should **always** be the first variable you specify with the **tables** command.

Relative Risk, confidence intervals, and significance tests

Use the **tables** command to investigate the association between the other exposure variables and reported illness and complete the table below:

	Lower 95% CL	Relative Risk	Upper 95% CL	Yates' X^2	p-value
HAM	0.74	1.08	1.57	0.02	0.889
SPINACH					
POTATO					
CABBAGE					
JELLY					
ROLLS					
BREAD					
MILK					
COFFEE					
WATER					
CAKES					
VANILLA					
CHOCOLATE					
FRUIT					

ANALYSIS will display two p-values for Fisher's exact test when any of the expected values on which the chi-square test is based is less than five. When an expected value is very small the standard chi-square tests are not valid. Fisher's exact test is a test of association for 2-by-2 tables that remains valid when an expected value for any of the cells of the table is less than five. There are two p-values associated with Fisher's exact test: the *two-tailed* p-value and the *one-tailed* p-value. A one-tailed p-value is only appropriate if you are testing for one type of association only (**either** an increased risk **or** a protective effect).

The one-tailed p-value is appropriate with data from food-borne outbreaks because we are interested in examining the effects of exposure variables in **one direction** (i.e. do they **increase** the risk of ILLness?).

In other types of study we are usually testing for any type of association and would normally use a two-tailed test. If in doubt use a two-tailed test. The two-tailed p-value is **not** appropriate here because we are interested in examining the effects of exposure variables in one direction only.

Apparent protective effects in food-borne outbreaks are usually due to people eating only one of each type of any one course. Consider, for example, the relative risks for VANILLA ice-cream (5.57) and CHOCOLATE ice-cream (0.71) in the Oswego Church Supper outbreak. The apparent protective effect of CHOCOLATE is due to the fact that those attending the Oswego Church Supper tended to eat either VANILLA ice-cream or CHOCOLATE ice-cream but not both (those that did eat both probably consumed a lower infective dose of the organism or toxin from the VANILLA ice-cream). You can verify this by entering the command:

```
tables vanilla chocolate
```

Leave ANALYSIS by issuing the command:

```
quit
```

or by pressing the **F10** key.

Confidence intervals and significance tests

Both confidence intervals and significance tests can be used to test for an association between exposure and outcome variables. They are related approaches to the same problem. If the 95% confidence interval for a relative risk does not include one then the chi-square will have a p-value of less than 0.05 (5%). The confidence interval (or *estimation*) approach is preferred as it provides an estimate of the *magnitude* and *direction* of the effect and the *inherent variability in the estimate*. The statistical significance testing approach only indicates the possible consistency of the data with the null hypothesis and provides no estimate of the magnitude or direction of the effect. Both approaches are influenced by *sample size*. Large numbers in cells will yield large chi-square values. Consider using a chi-square test to establish whether an association exists between eating CAKES and ILLness in the Oswego Church Supper outbreak:

CAKES	ILL		Total
	+	-	
+	27	13	40
-	19	16	35
Total	46	29	75

This table yields a chi-square statistic of 0.87 with a p-value of 0.35. If we assume that the church supper had been four times larger (by multiplying all cells in the table by four) the following table would result:

CAKES	ILL		Total
	+	-	
+	108	52	160
-	76	64	140
Total	184	116	300

which yields a chi-square statistic of 4.95 with a p-value of 0.03 (this is below the cut-off value of 0.05 which is *statistically significant*). The chi-square value has been influenced by the sample size. If we perform the same experiment but calculate the relative risk and confidence interval we get the following results for the original table:

Relative risk of (ILL=+) for (CAKES=+) 1.24
 Greenland, Robins 95% conf. limits for RR 0.86 < RR < 1.80

And the following results for the table with increased sample size:

Relative risk of (ILL=+) for (CAKES=+) 1.24
 Greenland, Robins 95% conf. limits for RR 1.03 < RR < 1.50

The point estimate is the same but the confidence interval is narrower. Increasing the sample size will give a smaller p-value and a narrower confidence interval.

You should never multiply the numbers in a 2-by-2 table to get a significant p-value or a narrower confidence interval. The **only** valid method of increasing the sample size is to collect more data!

Epi-Info STATCALC

The **tables** command in Epi-Info ANALYSIS produces contingency tables together with measures of effect (relative risks and odds ratios), confidence intervals, chi-square values, and p-values using data stored in an Epi-Info data (.REC) file. In some situations you may already have a 2-by-2 table but not the original data. It is not possible to use ANALYSIS with data that has already been tabulated. However, it is possible to use STATCALC to calculate measures of effect, confidence intervals, chi-square values and p-values with pre-tabulated data.

Start the course software and select the Epi-Info STATCALC module. Once STATCALC starts you will be presented with another menu:

```
+-----+
| Tables (2 x 2, 2 x n) |
| Sample size & power   |
| Chi square for trend  |
+-----+
```

STATCALC is an epidemiological calculator that produces statistics from summary data entered on the screen. Three types of calculations are available:

1. Statistics from 2-by-2 to 2-by-10 tables similar to those produced in ANALYSIS. 2-by-2 tables can be analysed to produce odds ratios and relative risks with confidence intervals and chi-square tests and associated p-values.
2. Sample size calculations for estimating a proportion in a population survey or for comparing proportions in case-control, cross-sectional, or cohort studies. This topic will not be covered in this course.
3. Chi-square test for linear trend. This tests for the presence of a *dose-response relationship* in epidemiological studies where a series of increasing or decreasing exposures is being studied.

STATCALC and 2-by-2 tables

In this exercise you will use the Epi-Info STATCALC module to examine the effect of sample size on the chi-square test and on the point estimate of the relative risk and its associated confidence interval. You will experiment with a table from the Oswego Church Supper outbreak.

From the STATCALC menu:

```
+-----+
| Tables (2 x 2, 2 x n) |
| Sample size & power   |
| Chi square for trend  |
+-----+
```

select the option **Tables (2 x 2, 2 x n)**. An empty 2-by-2 table appears on the screen. Enter the data from the WATER by ILLness table from the Oswego Church Supper outbreak:

WATER		ILL		Total
		+	-	
+		13	11	24
-		33	18	51
Total		46	29	75

by typing:

```
13 
11 
33 
18 
```

If you make a mistake when entering the data you will have to press the Quit key and start again.

When you have entered all of the data press the key again. STATCALC will display the odds ratio and relative risk with confidence intervals and chi-square tests with associated p-values:

```
Analysis of Single Table
Odds ratio = 0.64 (0.21 <OR< 1.94)
Cornfield 95% confidence limits for OR
Relative risk = 0.84 (0.55 <RR< 1.27)
Taylor Series 95% confidence limits for RR
Ignore relative risk if case control study.
```

	Chi-Squares	P-values
	-----	-----
Uncorrected :	0.76	0.3819645
Mantel-Haenszel:	0.75	0.3851566
Yates corrected:	0.38	0.5351719

Examine the output carefully and identify the relative risk, the associated confidence interval, the chi-square statistic (the Yates corrected statistic is preferred) and its associated p-value.

STATCALC and 2-by-2 tables

Press the **ENTER** key twice to get back to the empty 2-by-2 table and enter data for double the original sample size by typing:

26 **ENTER**
 22 **ENTER**
 66 **ENTER**
 36 **ENTER**

When you have entered all of the data press the **ENTER** key again. STATCALC will display the odds ratio and relative risk with confidence intervals and chi-square tests and associated p-values:

```

      Analysis of Single Table
      Odds ratio = 0.64 (0.30 <OR< 1.37)
      Cornfield 95% confidence limits for OR
      Relative risk = 0.84 (0.62 <RR< 1.13)
      Taylor Series 95% confidence limits for RR
      Ignore relative risk if case control study.
  
```

	Chi-Squares	P-values
	-----	-----
Uncorrected :	1.53	0.2163018
Mantel-Haenszel:	1.52	0.2178394
Yates corrected:	1.12	0.2906442

Examine the output carefully and identify the relative risk, the associated confidence interval, the chi-square statistic (the Yates corrected statistic is preferred) and its associated p-value. Examine the effect that doubling the sample size has on these measures. Try doubling the sample size three more times:

	+	-		+	-		+	-	
+	52	44		104	88		208	176	
-	132	72		264	144		528	288	
	300			600			1200		

Observe the effect this has on the calculated relative risk, confidence interval, chi-square test, and p-value by completing the table below:

Sample Size	Lower 95% CL	Relative Risk	Upper 95% CL	χ^2	p-value
75	0.55	0.84	1.27	0.38	0.535
150	0.62	0.84	1.13	1.12	0.291
300					
600					
1200					

Increasing the sample size gives a narrower confidence interval and a smaller p-value. **You should never multiply the numbers in a 2-by-2 table to get a significant p-value or a narrower confidence interval.** The **only** valid method of increasing the sample size is to collect more data!

Press **F10** twice to leave STATCALC and return to the EpiSoft main menu.

Analysing data with more than one level of exposure

The previous examples assume that the exposure variable is binary (exposed / not exposed). This will not always be the case. There may be several values for several different levels of exposure based on extent or duration of exposure. If exposure is a *continuous* variable, a series of levels can be formed by grouping the variable. As an example consider exposure to tap-water (in terms of the number of glasses drank per day) in the following table:

WATER	ILL		Total
	+	-	
NONE	37	20	57
1, 2, 3	70	23	93
4, 5, 6	34	7	41
MORE	17	3	20
Total	158	53	211

The choice of groups is somewhat arbitrary. You should use groups that are natural or *plausible*. It might be appropriate to use common groups (e.g. ages are often grouped into five or ten year age-bands).

Do not group together categories which are very different. If the exposure-specific risk (or odds) for those who drank two glasses of tap-water was very different to the exposure-specific risk (or odds) for those who drank three glasses of tap-water then it would **not** be appropriate to put them into the same group. The easiest way of checking this is to plot the exposure-specific risks for a series of very small groups.

Do not allow the size of any group to be too small. Check that no cells have an expected value below about five. In the table above the cell containing the value three (3) has the smallest expected value but this is greater than five:

$$\begin{aligned}
 \text{expected value} &= (\text{row total} * \text{column total}) / \text{overall total} \\
 &= 20 * 53 / 211 \\
 &= 5.02
 \end{aligned}$$

In the absence of an obvious or natural choice of groups you could use percentiles or quartiles to divide the exposure variable into a number of equal sized groups.

There are two tests that you can use with this sort of data. These are the *chi-square test of association* and the *chi-square test for linear trend in the odds ratios (or proportions)*.

The chi-square test of association

Pearson's chi-square works with tables with any number of rows and columns. It can be used to test for an association between an exposure variable with more than two levels and an outcome variable. It tests the null hypothesis that the ratio of cases to controls (i.e. ill to not-ill persons) is the same for each exposure level (i.e. the ratio of cases (ill) to controls (not-ill) does not depend on the level of exposure). It does this by comparing the observed and expected numbers in each cell using the standard formula:

$$X^2 = \sum [(\text{Observed} - \text{Expected})^2 / \text{Expected}]$$

Applying this formula to the data gives a chi-square of 5.52 (degrees of freedom = (4 - 1) * (2 - 1) = 3) and a p-value of 0.14. There is, therefore, *insufficient evidence* to reject the null hypothesis of no association. There is no evidence of an association between drinking tap-water and illness. It is possible to use STATCALC to perform this test.

Using STATCALC to perform the chi-square test of association

Start the course software and select the Epi-Info STATCALC module. Once STATCALC starts you will be presented with another menu:

```
+-----+
| Tables (2 x 2, 2 x n) |
| Sample size & power   |
| Chi square for trend  |
+-----+
```

Select the option **Tables (2 x 2, 2 x n)** and enter the example table:

WATER	ILL		Total
	+	-	
NONE	37	20	57
1,2,3	70	23	93
4,5,6	34	7	41
MORE	17	3	20
Total	158	53	211

by typing:

```
37 
20 
70 
23 
34 
7  
17 
3  
```

If you make a mistake when entering the data you will have to press the Quit key and start again.

When you have entered all of the data press the key again. STATCALC will display the chi-square statistic and its associated p-value:

```
Chi square =      5.52
3 degrees of freedom.
p value = 0.13751692
```

This is the chi-square test of association. Epi-Info ANALYSIS also produces this statistic if you issue the command **set statistics = on** before any **tables** commands.

Press the key to return to the STATCALC menu. Press again to leave STATCALC and return to the EpiSoft main menu.

The chi-square test for linear trend

The chi-square test of association will test for the existence of **any type** of association. With the Oswego Church Supper outbreak data it was used to show that VANILLA ice-cream was associated with food-poisoning. It cannot be used to determine the **nature** (in terms of *direction* and *magnitude*) of this association.

The chi-square test of association is not always specific enough. A test which could detect a particular association, such as a *dose-response relationship*, might be more appropriate. The chi-square test of association is *insensitive* when it comes to detecting a dose-response relationship. A more powerful test of a dose-response relationship is provided by the *chi-square test for linear trend* which makes better use of the data.

The chi-square test for linear trend is only appropriate if you suspect that there might be a dose-response relationship between exposure and outcome. Use the table below to examine the proportion of persons ILL for each level of WATER and determine whether a dose-response test is appropriate.

WATER	ILL			Total	
	+	-			
NONE	37	20		57	Proportion ILL = (37 / 57) * 100 = 65%
1, 2, 3	70	23		93	Proportion ILL = (70 / 93) * 100 = 75%
4, 5, 6	34	7		41	Proportion ILL = (34 / 41) * 100 = 83%
MORE	17	3		20	Proportion ILL = (17 / 20) * 100 = 85%
Total	158	53		211	

We can see that the proportion ILL increases as the level of exposure increases. The proportion ILL increases from 65% of those who drank no tap-water, to 75% of those who drank between one and three glasses, to 83% of those who drank between four and six glasses, and to 85% of those who drank more than six glasses. There appears to be a dose-response relationship.

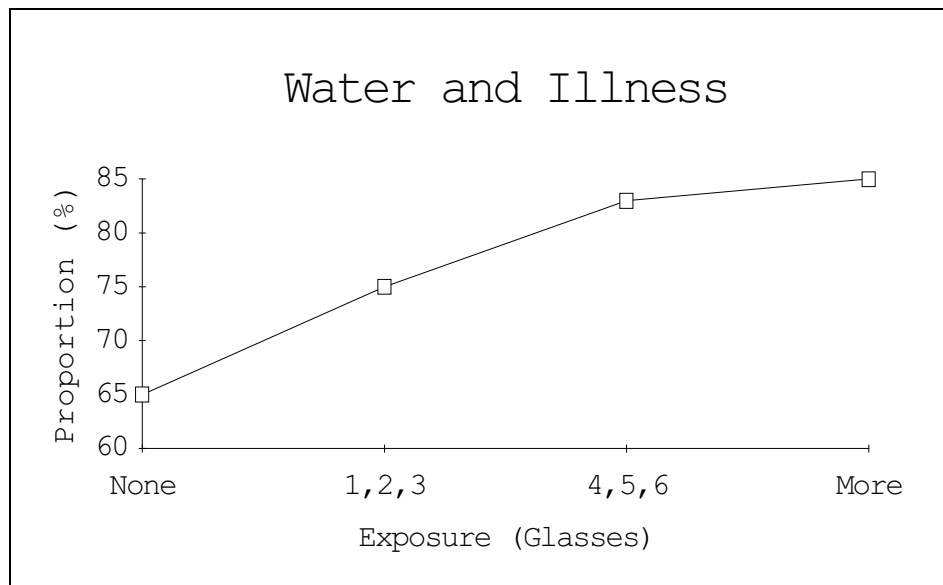
Before calculating the chi-square test for trend it is important to examine the data to see if it is reasonable to use a test for *linear* trend. The proportions are clearly increasing, but does the increase look linear? Once you have calculated the proportion of cases at each exposure level it is a good idea to graph the results to check that the data does not exhibit any obvious *non-linear* characteristics.

These proportions are subject to random variation. The proportion ILL observed at each level of WATER might not be exactly equal to the true proportion ILL. The graph is unlikely to exhibit exact linearity even if there is a true linear trend in these proportions. In this context *linear* means a **straight** line. In practice this means a **reasonably** straight line through the observed proportions.

If the graph exhibits marked *non-linearity* you should **not** use the chi-square test for linear trend. The test for trend is appropriate **only** if the proportions increase (or decrease) in a straight line.

The chi-square test for linear trend

If we graph the proportions from our example table we can see that it is reasonable to use the chi-square test for linear trend with this data:



When calculating the chi-square test for linear trend, scores are assigned to each level of exposure. The usual choice is 1, 2, 3, 4 etc. This represents a trend that increases (or decreases) with each row in the table. These scores are then regarded as observations and the mean scores for cases and controls are compared. Calculation of the chi-square test for linear trend is complicated and is best left to purpose-designed computer programs such as STATCALC. Entering the example table:

WATER	ILL			Total	
	+	-			
NONE	37	20		57	Proportion ILL = (37 / 57) * 100 = 65%
1, 2, 3	70	23		93	Proportion ILL = (70 / 93) * 100 = 75%
4, 5, 6	34	7		41	Proportion ILL = (34 / 41) * 100 = 83%
MORE	17	3		20	Proportion ILL = (17 / 20) * 100 = 85%
Total	158	53		211	

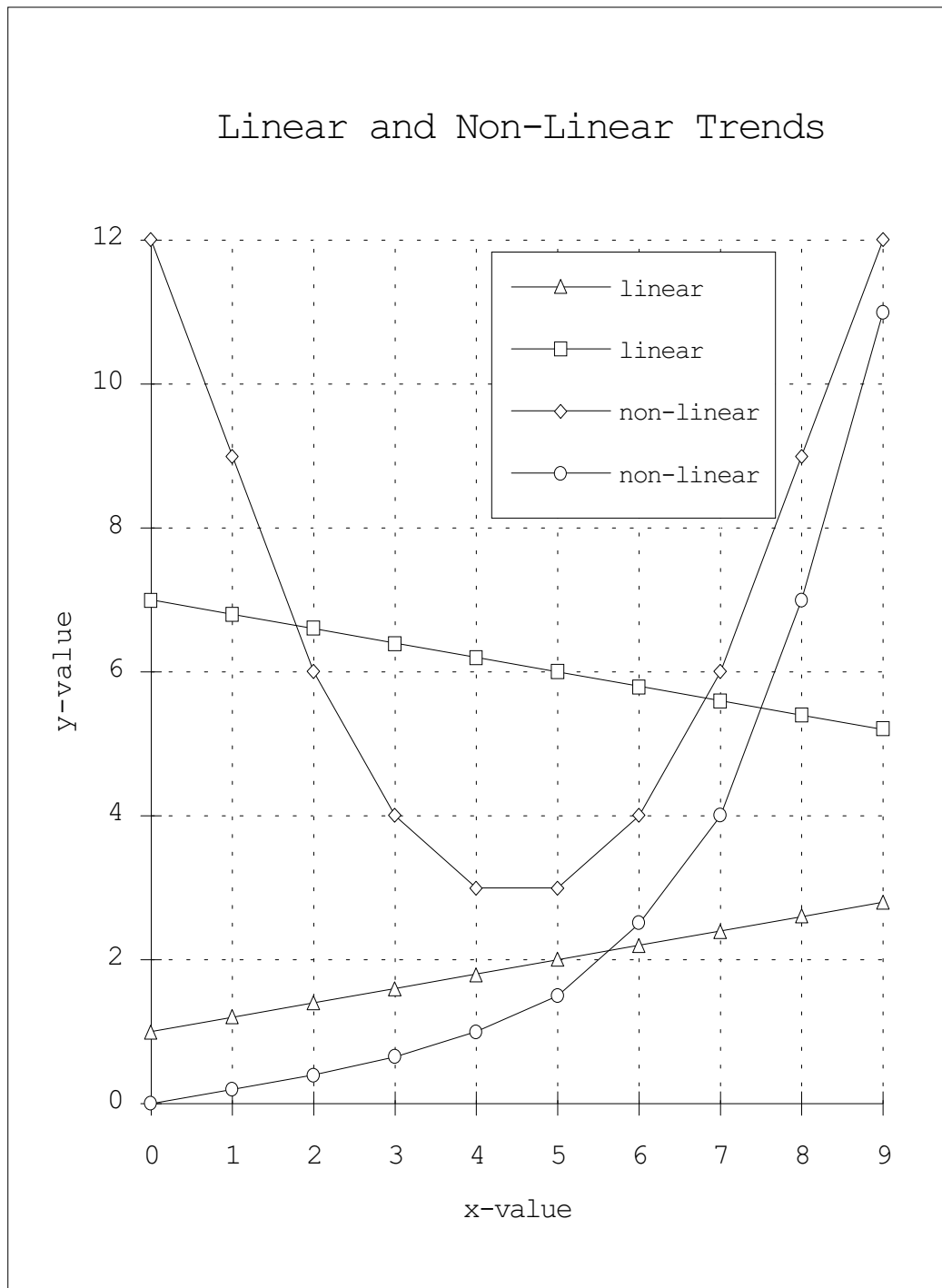
into Epi-Info STATCALC, a chi-square (for linear trend) of 5.124 (degrees of freedom = 1) with a p-value of 0.02 was obtained. This indicates a *linear trend in the proportion of cases over increasing levels of exposure* to tap-water. By looking at these proportions, it appears that this is an increasing trend. The proportion ILL increases as the number of glasses of tap-water drank per day increases. This provides evidence of a dose-response effect. Note the chi-square test for trend always has one degree of freedom irrespective of the number of rows in the table.

Epi-Info ANALYSIS does **not** calculate the chi-square test for linear trend. The chi-square test results that are displayed with the **tables** command in Epi-Info ANALYSIS are for tests of association. If you want to perform the chi-square test for linear trend with your own untabulated data, you should use ANALYSIS to obtain the table and then enter the cell counts from this table into STATCALC.

The chi-square test for trend is only valid when investigating a dose-response relationship. It is **not** valid to use the chi-square test for trend with *non-ordered* data such as ethnicity, marital status, or eye colour. In these cases the chi-square test of association is the appropriate test.

Linear and non-linear trends

The graph below shows some possible trends in proportions. Two of the lines show a linear trend. It would be reasonable to use the chi-square test for linear trend in these circumstances. The other two lines are non-linear. It would **not** be reasonable to use the test for linear trend in these circumstances.



It is only valid to use the chi-square test for trend when the trend in the proportion of cases appears reasonably linear. The trend in proportions does not need to be perfectly linear in order to use this test.

Non-linear data

If your data exhibits marked non-linearity it may be appropriate to collapse the data into a 2-by-2 table (e.g. exposure levels of 0 - 3 glasses, and 4 or more glasses) for further analysis. The methods described earlier should be used to choose the *threshold* values for the two categories. The following analysis is appropriate if it appears that the exposure-specific risk increases greatly for exposure to four or more glasses of tap-water compared to three or less glasses of tap-water. This table was created using the example data:

WATER	ILL		Total
	+	-	
4 or MORE	51	10	61
0 - 3	107	43	150
Total	158	53	211

Odds ratio 2.05
Cornfield 95% confidence limits for OR 0.90 < OR < 4.75

The 4 or More group is placed in the first row because this is the *exposed* or *most-exposed* category. There is a two-fold increase in odds associated with drinking 4 or more glasses of tap-water.

Non-ordered data

Here is a table from a hypothetical study of the diets of children in lone-parent households. The outcome variable is IRON (where + is used to indicate that a diet was considered to be deficient in iron). The exposure variable is ETHNIC which indicates the self-defined ethnicity of the head of household:

ETHNIC	IRON		Total
	+	-	
Black	37	200	237
Chinese	70	230	300
Indian	31	70	101
White	153	270	423
Total	291	770	1061

Proportion IRON = (37 / 237) * 100 = **16%**
Proportion IRON = (70 / 300) * 100 = **23%**
Proportion IRON = (31 / 101) * 100 = **31%**
Proportion IRON = (153 / 423) * 100 = **36%**

There is **no** natural order to the explanatory (or exposure) variable. The apparent trend in proportions is an *artefact* of the way the categories of the explanatory variable have been sorted (alphabetically). If the categories were sorted in a different way the apparent trend would disappear. It does **not** make sense to speak of a trend in proportions in this context.

The chi-square test for trend is only valid when investigating a dose-response relationship. It is **not** valid to use the chi-square test for trend with non-ordered data such as ethnicity. In this case the chi-square test of association is the appropriate test.

Using STATCALC to calculate a chi-square test for trend

From the STATCALC main menu select the option Chi square for trend and enter the data from the tap-water example table replacing the exposure categories with numeric trend scores (1, 2, 3, and 4):

WATER	ILL		Total
	+	-	
1	37	20	57
2	70	23	93
3	34	7	41
4	17	3	20
Total	158	53	211

by typing:

```

1  [ENTER]
37 [ENTER]
20 [ENTER]
2  [ENTER]
70 [ENTER]
23 [ENTER]
3  [ENTER]
34 [ENTER]
7  [ENTER]
4  [ENTER]
17 [ENTER]
3  [ENTER]

```

If you make a mistake when entering the data you can move between the cells of the table using **[←]** and **[SHIFT] + [→]**. When you have entered all of the data press the **[F4]** Calc key. STATCALC will display the chi-square statistic and its associated p-value together with the odds ratio for each exposure level:

```

Analysis For Linear Trend In Proportions

Chi Square for
linear trend : 5.124

p value      : 0.02359

Exposure
Score      Odds Ratio
1.00      1.00
2.00      1.65
3.00      2.63
4.00      3.06

```

Examine the output carefully and identify the chi-square test for linear trend and its associated p-value. The output shows odds ratios because they are a valid measure of effect for all types of study.

The Epi-Info ANALYSIS module does not calculate the chi-square test for trend. If you need to analyse data in this way first produce the table from your data using ANALYSIS and then use STATCALC to calculate the chi-square test for linear trend. The chi-square statistic produced by ANALYSIS is for the chi-square test of association.

Press the **[F10]** key to return to the STATCALC menu. Press **[F10]** again to leave STATCALC and return to the course software main menu.

Analysing two exposure variables at the same time

Using the methods covered so far, you can find the effect of an exposure on the disease (using the odds ratio or relative risk) and test the statistical significance of this effect (using confidence intervals or statistical hypothesis testing). These methods are appropriate for analysing the effect of exposure variables *independently* of one another. For example, you looked at the effect of becoming ill at the Oswego Church Supper after eating each particular food without considering the effects of having eaten other foods at the supper. After analysing the effect of each food at the Oswego Church Supper it became apparent that only one food (VANILLA ice-cream) was associated with becoming ill. This will not always be the case. Typically you may find that two or more exposures are associated with your outcome variable.

The Bateman outbreak

Start the course software and select Epi-Info ANALYSIS from the main menu. Once ANALYSIS has started issue the command:

```
read bateman.rec
```

to retrieve the data for the Bateman outbreak. Enter the command:

```
variables
```

to display a list of variables in the Bateman outbreak data file together with information about their type and length. In the BATEMAN dataset the ILL variable is the outcome of interest and the food variables (CHEESE, CRABDIP, CRISPS, BREAD, CHICKEN, RICE, CAESAR, TOMATO, ICECREAM, CAKE, JUICE, WINE, COFFEE) are the exposure variables.

If you examine the output of the variables command you will notice that these variables are all of the *Yes/No* type. This means that each variable can hold only one of two values (Y or + for YES, N or – for NO). You can examine the distribution of these variables using the **freq** command.

Enter the command:

```
freq ill
```

to display a frequency table of the ILL variable. The table lists frequencies, percentages, and cumulative percentages:

ILL	Freq	Percent	Cum.
+	29	64.4%	64.4%
-	16	35.6%	100.0%
Total	45	100.0%	

From this table we can see that we have data on forty-five (45) people who attended the luncheon party. Twenty-nine (29) of these went on to develop food-poisoning. The proportion of ill people in the total population at risk is $29 / 45 = 0.644 = 64.4\%$. This is the absolute risk and is interpreted as: each person attending the Bateman luncheon ran a risk of 64.4% of developing food-poisoning. Identify this figure in the displayed frequency table. It is valid to calculate the absolute risk because we have data for **all** persons who attended the luncheon.

Producing statistics with 2-by-2 tables

Issue the command:

```
set statistics = on
```

to instruct ANALYSIS to produce a full set of statistics in the output of each command. Use the **tables** command to investigate the association between CHEESE and ILL by issuing the command:

```
tables cheese ill
```

which produces the following output:

CHEESE		ILL		Total
		+	-	
+		15	7	22
-		14	9	23
Total		29	16	45

Single Table Analysis

Odds ratio 1.38
Cornfield 95% confidence limits for OR 0.34 < OR < 5.69

Relative risk of (ILL=+) for (CHEESE=+) 1.12
Greenland, Robins 95% conf. limits for RR 0.73 < RR < 1.73
(Biometrics 1985;41:55-68)
Ignore relative risk if case control study.

	Chi-Squares	P-values
Uncorrected:	0.26	0.60848285
Mantel-Haenszel:	0.26	0.61249393
Yates corrected:	0.04	0.84089968

Examine the output carefully and identify the relative risk, the associated confidence interval, the chi-square statistic and its associated p-value. The **Uncorrected** chi-square is Pearson's chi-square. The **Yates corrected** chi-square is preferred for 2-by-2 tables.

Relative risk, confidence intervals, and significance tests

Use the **tables** command to investigate the association between each of the other exposure variables and reported ILLness and complete the table below.

Recall that the Yates' corrected chi-square and its corresponding p-value should be used. For some of the 2-by-2 tables ANALYSIS will display two p-values for Fisher's exact test when any of the expected values on which the chi-square test is based is less than five. When an expected value is small the standard chi-square tests are not valid. Fisher's exact test remains valid when an expected value for any of the cells in the table is less than five. There are two p-values associated with Fisher's exact test: the two-tailed p-value and the one-tailed p-value. The one-tailed p-value is appropriate here because we are interested in examining the effects of the exposure variables in **one direction only** (i.e. do they **increase** the risk of ILLness?).

	Lower 95% CL	Relative Risk	Upper 95% CL	Yates' X^2	p-value
CHEESE	0.73	1.12	1.73	0.04	0.841
CRABDIP					
CRISPS					
BREAD					
CHICKEN					
RICE					
CAESAR					
TOMATO					
ICECREAM					
CAKE					
JUICE					
WINE					
COFFEE					

If the commands you issue produce output that will not fit onto a single screen then ANALYSIS will pause at the end of each screen of output and display the message <more> at the bottom of the output screen. If this happens press any key (apart from the **ESC** key) to see the next screen of the output.

You can scroll through the output generated by previous commands using the **PG UP** and **PG DN** keys. These move through the output one screen at a time. You may also use **CTRL** + **PG UP** and **CTRL** + **PG DN** to move through the output one line at a time.

You can scroll through previously entered commands using the **↑** and **↓** keys.

When you have filled in the table above, leave ANALYSIS by issuing the command:

quit

or by pressing the **F10** key.

Confounded associations

The following table is from the Bateman outbreak:

CAESAR		ILL		Total	
		+	-		
	+	26	5	31	Relative Risk = 3.91 Lower 95% confidence limit = 1.42 Upper 95% confidence limit = 10.80
	-	3	11	14	
Total		29	16	45	

The relative risk of becoming ill after eating CAESAR salad is 3.91. The 95% confidence interval for the relative risk is 1.42 to 10.80 which suggests that CAESAR salad is significantly associated with becoming ill. The chi-square statistic for this table is 13.80 which gives a p-value of less than 0.001 at one degree of freedom. There is a cell (cell (d) with the observed value of 11) which has an expected value equal to:

$$(\text{row total} * \text{column total}) / \text{overall total} = 14 * 16 / 45 = 4.98$$

So it might be wise to use Fisher's Exact test:

```
Fisher exact: 1-tailed P-value: 0.0001002 <---
                2-tailed P-value: 0.0001002 <---
```

The (one-tailed) p-value for Fisher's Exact test is less than 0.001 which is still highly significant.

The following table is also from the Bateman outbreak:

TOMATO		ILL		Total	
		+	-		
	+	24	6	30	Relative Risk = 2.40 Lower 95% confidence limit = 1.15 Upper 95% confidence limit = 2.40
	-	5	10	15	
Total		29	16	45	

The relative risk of becoming ill after eating TOMATO salad is 2.40 (95% CI: 1.15 - 5.02), which suggests that eating TOMATO salad was also significantly associated with becoming ill. The chi-square statistic for this table is 7.58 which gives a p-value of 0.006 at one degree of freedom.

Assuming that these two associations are real (i.e. they did not arise by chance), there are three possible explanations for these results:

1. **Both CAESAR salad and TOMATO salad caused food-poisoning.**
2. **CAESAR salad caused food-poisoning** and TOMATO salad did not.
3. **TOMATO salad caused food-poisoning** and CAESAR salad did not.

Another explanation is that neither CAESAR salad nor TOMATO salad alone were sufficient to cause ILLness and that consumption of both of these foods was necessary to cause ILLness. This situation is rare in food-borne outbreaks but can occur with cases of chemical or environmental poisoning or with *multifactorial* diseases. The independent operation of two or more factors to produce, prevent, enhance, or limit an effect is known as *interaction* or *effect-modification*.

Confounded associations

The two situations:

1. **CAESAR salad caused food-poisoning** and TOMATO salad did not,
2. **TOMATO salad caused food-poisoning** and CAESAR salad did not

could arise if eating CAESAR salad was associated with eating TOMATO salad (and vice versa). This would cause TOMATO salad to *confound* the relationship between CAESAR salad and food-poisoning. It would also cause CAESAR salad to *confound* the relationship between TOMATO salad and food-poisoning. *Confounding* between variables may cause *spurious associations* to arise between exposures and the outcome. It may be that CAESAR salad caused the food-poisoning and, because many of the people who ate CAESAR salad also ate TOMATO salad, it appeared that TOMATO salad was also associated with food-poisoning. *Negative confounding* between variables may cause true associations between risk factors and the disease to be concealed. A *confounding variable* is associated with **both** the outcome variable **and** the exposure variables. TOMATO salad is a *potential confounder* if it is associated with food-poisoning and eating CAESAR salad. CAESAR salad is a potential confounder if it is associated with food-poisoning and eating TOMATO salad.

You can use the Epi-Info ANALYSIS module to investigate whether an association exists between two exposure variables using the **tables** command. The command **tables caesar tomato** produces the following output:

CAESAR		TOMATO		Total
		+	-	
+	27	4	31	
-	3	11	14	
Total	30	15	45	

Single Table Analysis

Odds ratio 24.75
Cornfield 95% confidence limits for OR 3.78* < OR < 200.29*
*May be inaccurate

Relative risk of (TOMATO=+) for (CAESAR=+) 4.06
Greenland, Robins 95% conf. limits for RR 1.48 < RR < 11.18
(Biometrics 1985;41:55-68)
Ignore relative risk if case control study.

	Chi-Squares	P-values
Uncorrected:	18.72	0.00001517 <---
Mantel-Haenszel:	18.30	0.00001888 <---
Yates corrected:	15.88	0.00006759 <---

Fisher exact: 1-tailed P-value: 0.0000344 <---
2-tailed P-value: 0.0000344 <---

Eating CAESAR salad is strongly associated with eating TOMATO salad (and vice versa). The relative risk (or odds ratio) is not used in this context as neither variable can be properly classified as the *exposure* or *outcome* variable.

If ANALYSIS does not produce the statistics after the tables then you should issue the command **set statistics = on** and issue the **tables** command again.

The Mantel-Haenszel relative risk and significance test

CAESAR salad and TOMATO salad are both associated with food-poisoning. The relative risk of food-poisoning for CAESAR salad is 3.91. To determine whether TOMATO salad is a confounder of this relationship, you should look at the relationship between CAESAR salad and ILL according to whether individuals ate TOMATO salad or not. This is called a *stratified analysis* because it is an analysis of one risk factor by outcome done separately for the different *strata* (i.e. different values) of another risk factor. The 2-by-2 table below shows the association between CAESAR salad and food-poisoning for those who ate TOMATO salad:

TOMATO = + ALL individuals ate TOMATO salad

CAESAR		ILL		Total	
		+	-		
+		23	4		27
-		1	2		3
Total		24	6		30

Relative Risk = 2.56
Lower 95% confidence limit = 0.51
Upper 95% confidence limit = 12.76

There were 30 people who ate TOMATO salad of whom 24 became ill. The relative risk for CAESAR salad in this group is 2.56 (95% CI: 0.51 - 12.76).

The following 2-by-2 table shows the association between CAESAR salad and food-poisoning for those who did not eat TOMATO salad:

TOMATO = - NO individuals ate TOMATO salad

CAESAR		ILL		Total	
		+	-		
+		3	1		4
-		2	9		11
Total		5	10		15

Relative Risk = 4.13
Lower 95% confidence limit = 1.04
Upper 95% confidence limit = 16.32

There were 15 people who did not eat TOMATO salad of whom 5 became ill. The relative risk for CAESAR salad in this group is 4.13 (95% CI: 1.04 - 16.32). The relative risk is based on small numbers and should be interpreted cautiously.

It is **not** appropriate to combine these two relative risks if the relative risks for those who ate TOMATO salad and those who did not are very different. It is valid to combine these two relative risks to get a *pooled relative risk* only if the relative risks for those who ate TOMATO salad and those who did not are similar.

In this case the relative risks for those who ate TOMATO salad and for those who did not eat TOMATO salad are sufficiently similar (both relative risks are substantially greater than one and their confidence intervals overlap substantially). The pooled relative risk is an *average* of the two relative risks. In preference to a simple average, we calculate a *weighted average* that gives more *weight* to the strata with more data. A simple weighted average would be:

$$RR_{\text{pooled}} = ((RR_1 * N_1) + (RR_2 * N_2)) / (N_1 + N_2)$$

where RR_n is the relative risk in the n^{th} stratum and N_n is the number of cases in the n^{th} stratum. In this example:

$$RR_{\text{pooled}} = ((2.56 * 30) + (4.13 * 15)) / (30 + 15) = 3.08$$

The Mantel-Haenszel relative risk and significance test

The simple weighted average approach runs into problems when strata contain few subjects and reliable estimates of stratum-specific relative risks cannot be calculated. A better weighting scheme is proposed by Mantel and Haenszel:

$$RR_{MH} = \Sigma [(a * (c + d)) / n] / \Sigma [(c * (a + b)) / n]$$

where summation is over the different strata. In this example:

$$RR_{MH} = ((23 * 3) / 30 + (3 * 11) / 15) / ((1 * 27) / 30 + (2 * 4) / 15) = 3.14$$

The *Mantel-Haenszel relative risk* is also called the *adjusted relative risk* or the *summary relative risk*. It provides a means of summarising relative risks from different strata of a factor and in so doing, it *adjusts* for the confounding effect of this factor. An equivalent technique can be used to calculate an *adjusted odds ratio*. The crude (unadjusted) and adjusted relative risks should be compared to determine whether a variable is a confounder. If the adjusted relative risk is much smaller than the crude relative risk (at least 15-20% smaller) then there is evidence of confounding. If the adjusted relative risk is a little smaller than the crude relative risk (less than 15% smaller) then there is little or no evidence of confounding. If the adjusted relative risk is much larger (at least 15-20% larger) than the crude relative risk then there is evidence of negative confounding.

A significance test (the *Mantel-Haenszel summary chi-square*) can be used to test whether the Mantel-Haenszel relative risk or the Mantel-Haenszel odds ratio are significantly different from one (i.e. no association between exposure and outcome). **The Mantel-Haenszel summary chi-square test is not a significance test for confounding. A large difference between the crude and adjusted relative risk is evidence of confounding.** A significant Mantel-Haenszel summary chi-square suggests that the adjusted relative risk is significantly different from one. This usually happens when there is little confounding but it is possible for the Mantel-Haenszel chi-square to be significant when there is substantial confounding.

When the Mantel-Haenszel relative risk was calculated for the above example with Epi-Info ANALYSIS, the following results were obtained:

```

SUMMARY RELATIVE RISKS
(Ignore if Case-Control Study)

Crude RR                                3.91
Summary RR of (ILL=+) for (CAESAR=+)    3.14
95% confidence limits for summary RR    1.06 < RR < 9.28
(Greenland, Robins Biometrics 1985;41:55-68)

M-H Summary Chi Square                    5.75
P value                                0.01647023 <---
```

The relative risk calculated from the table for CAESAR by ILL (without adjusting for the effect of TOMATO) is called the *crude* or *unadjusted* relative risk. The crude relative risk = 3.91 which suggests that those persons who ate CAESAR salad were more likely to become ill than those who did not. The Mantel-Haenszel relative risk = 3.14 which suggests that after *adjusting* for the effect of TOMATO salad, CAESAR salad still increased the risk of becoming ill. The fact that the relative risk decreased by about 20% from 3.91 to 3.14 suggests that there was substantial confounding. What appeared to be the effect of eating CAESAR salad on food-poisoning may have actually been due to eating TOMATO salad. The adjusted relative risk is still much larger than one and still statistically significant. After adjusting for TOMATO salad, those persons who ate CAESAR salad still had three times the risk of acquiring food-poisoning, than those who did not eat CAESAR salad.

Testing for an interaction: Woolf's test

The Mantel-Haenszel relative risk provides a means of summarising the relative risks from different strata of a factor and in so doing, it adjusts for the effect of this factor. It is only appropriate to summarise the relative risks from different strata if they are **not** very different from one another.

If the effect of factor A on the disease is different for the different strata of factor B (i.e. the relative risks in the strata of factor B are very different), then there may be an interaction between factor A and factor B. If the effect of eating CAESAR salad on food-poisoning is different for those who ate TOMATO salad and those who did not eat TOMATO salad, then there may be an interaction between TOMATO salad and CAESAR salad.

To test for an interaction between TOMATO salad and CAESAR salad, you must compare the relative risk for CAESAR salad for those who ate TOMATO salad with the relative risk for CAESAR salad for those who did **not** eat TOMATO salad. Even if the *true* relative risks for those who ate TOMATO salad and those who did not eat TOMATO salad are equal they will differ due to *random variation*. In this example there is a difference between the CAESAR salad relative risk among those who ate TOMATO salad (relative risk = 2.56, 95% CI: 0.51 - 12.76) and among those who did not eat TOMATO salad (relative risk = 4.13, 95% CI: 1.04 - 16.32). These relative risks are based on small numbers and have wide and overlapping confidence intervals. This suggests that the difference in relative risks is probably due to random variation (since both are based on small numbers) rather than an interaction

There is a way of testing more formally whether these two relative risks are equal. This is *Woolf's test for heterogeneity of odds ratios or relative risks*. Woolf's test is another chi-square test and has degrees of freedom equal to the number of relative risks (or odds ratios) being compared minus one. Since we are comparing two relative risks Woolf's test has one degree of freedom. A Woolf's chi-square value with a p-value of less than 0.05 suggests that there is a significant difference between the two relative risks (i.e. there is an interaction). Some statisticians regard a p-value of less than 0.10 (rather than 0.05) as evidence of an interaction. This is because the test is based on stratum-specific relative risks or odds ratios. These are based on small numbers and are therefore less likely to yield small p-values.

When Woolf's test was carried out on the relative risks for CAESAR salad for those who ate TOMATO salad and those who did not eat TOMATO salad, the following results were obtained:

WOOLF'S TEST FOR HETEROGENEITY OF ODDS RATIOS

Woolf's Chi Square	0.01
P value	0.93390382
Test does not suggest multiplicative interaction.	

Woolf's chi-square was equal to 0.01 which gave a p-value of 0.933 suggesting that the observed difference between the two relative risks probably arose by chance and that there is unlikely to be an interaction between CAESAR salad and TOMATO salad.

If Woolf's chi-square had yielded a significant p-value, then an interaction between CAESAR salad and TOMATO salad could not be ruled out. If an interaction does exist or is suspected, then it is not appropriate to combine the two relative risks. If the effect of CAESAR salad is different depending on whether TOMATO salad had been eaten, then this should be shown in the results by presenting the relative risks separately. If Woolf's test yields a significant p-value, then the relative risks (or odds ratios) should be presented separately and the Mantel-Haenszel summary measures should not be quoted.

Testing for an interaction: Woolf's test

It is likely that CAESAR salad was one of the vehicles of food-poisoning. Those persons who ate CAESAR salad were nearly four times as likely to get food-poisoning than those who had not eaten CAESAR salad (relative risk = 3.91). It is possible that this relative risk might vary according to whether or not WINE was also consumed at the Bateman luncheon. To investigate this further we should perform an analysis of CAESAR by ILL stratified by WINE. The following 2-by-2 table shows the association between CAESAR salad and food-poisoning for those who drank WINE:

WINE = + ALL persons drank WINE

CAESAR		ILL		Total	
		+	-		
+		21	3		24
-		1	9		10
Total		22	12		34

Relative Risk = 8.75
 Lower 95% confidence limit = 1.35
 Upper 95% confidence limit = 56.52

The relative risk for CAESAR salad in this group is 8.75. The following 2-by-2 table shows the association between CAESAR salad and food-poisoning for those who did not drink WINE:

WINE = - NO persons drank WINE

CAESAR		ILL		Total	
		+	-		
+		5	2		7
-		2	2		4
Total		7	4		11

Relative Risk = 1.43
 Lower 95% confidence limit = 0.48
 Upper 95% confidence limit = 4.23

The relative risk for CAESAR salad in this group is 1.43. This relative risk is based on small numbers and should be interpreted with caution. When the Mantel-Haenszel summary relative risk was calculated for the above example using Epi-Info ANALYSIS, the following results were obtained:

```

SUMMARY RELATIVE RISKS
(Ignore if Case-Control Study)

Crude RR                                3.91
Summary RR of (ILL=+) for (CAESAR=+)    4.04
95% confidence limits for summary RR    1.39 < RR < 11.71
(Greenland, Robins Biometrics 1985;41:55-68)

M-H Summary Chi Square                   13.18
P value                                0.00028319 <---

```

When Woolf's test was carried out on the relative risks for CAESAR salad for those who drank WINE and those who did not drink WINE, the following results were obtained:

```

WOOLF'S TEST FOR HETEROGENEITY OF ODDS RATIOS

Woolf's Chi Square                       3.26
P value                                0.07090245

```

The effect of CAESAR salad appears much stronger for those who drank WINE compared to those who did not drink WINE. A p-value of 0.07 from Woolf's test is evidence of an interaction between CAESAR salad and WINE. It is **not** appropriate to present the adjusted relative risk. The relative risks for CAESAR salad should be presented for the separate levels of drinking WINE. Interactions are often called *effect modifiers*: drinking WINE *modifies the effect* of eating CAESAR salad.

Stratified analysis with ANALYSIS

Start the course software and select Epi-Info ANALYSIS from the main menu. Once ANALYSIS has started issue the command:

```
read bateman.rec
```

to retrieve the data for the Bateman outbreak. Enter the command:

```
set statistics = on
```

to instruct ANALYSIS to produce a full set of statistics in the output of each command. Use the **tables** command to investigate the association between CAESAR and ILL controlling for TOMATO by issuing the command:

```
tables caesar ill tomato
```

which, along with tables of CAESAR by ILLness for each level of TOMATO, produces the following output:

```

              SUMMARY ODDS RATIOS
Crude OR for sets with discordant results          19.07
Mantel-Haenszel Weighted Odds Ratio                12.50
95% confidence limits for M-H OR          1.88 < OR < 83.16
      (Robins, Greenland, Breslow  AJE 1986;124:719-23)

              SUMMARY RELATIVE RISKS
              (Ignore if Case-Control Study)
Crude RR                                           3.91
Summary RR of (ILL=+) for (CAESAR=+)              3.14
95% confidence limits for summary RR          1.06 < RR < 9.28
      (Greenland, Robins  Biometrics 1985;41:55-68)

M-H Summary Chi Square                           5.75
P value                                           0.01647023 <---

              WOOLF'S TEST FOR HETEROGENEITY OF ODDS RATIOS

Woolf's Chi Square                               0.01
P value                                           0.93390382
      Test does not suggest multiplicative interaction.
      (Schlesselman, JJ, Case-Control Studies. NY,Oxford U. Press,1982;p.194)
```

Examine the output of this command carefully and identify the crude relative risk, the Mantel-Haenszel summary relative risk and its confidence limits, the Mantel-Haenszel summary chi-square and its associated p-value, and Woolf's test and its associated p-value.

The interpretation of the relative risk produced by the **tables** command depends on the orientation of the table. The correct format for the **tables** command is:

```
tables <exposure> <outcome> <confounder1> .. <confounder3>
```

The risk or exposure variable should **always** be the first variable. The outcome variable must **always** be the second variable. The stratifying or confounding variables may follow. You may specify up to **three** stratifying or confounding variables with the **tables** command in Epi-Info ANALYSIS.

Stratified analysis with ANALYSIS

Investigate the association between TOMATO and ILL controlling for CAESAR by issuing the command:

```
tables tomato ill caesar
```

Examine the output of this command carefully and identify the crude relative risk, the Mantel-Haenszel summary relative risk and its confidence limits, the Mantel-Haenszel summary chi-square and its associated p-value, Woolf's chi-square and its associated p-value. Complete the table below:

Exposure:	CAESAR
Outcome:	ILL
Controlling for :	TOMATO
Crude RR:	3.91
M-H summary RR:	3.14
95% CI:	1.06 - 9.28
M-H chi-square:	5.75
p-value:	0.016
Woolf's chi-square:	0.01
p-value:	0.934

Exposure:	TOMATO
Outcome:	ILL
Controlling for :	CAESAR
Crude RR:	
M-H summary RR:	
95% CI:	
M-H chi-square:	
p-value:	
Woolf's chi-square:	
p-value:	

These results suggest that TOMATO salad was a confounder between CAESAR salad and food-poisoning (20% difference between crude and adjusted relative risk). After adjusting for TOMATO salad, CAESAR salad was still significantly associated with food-poisoning.

CAESAR salad was a confounder between TOMATO salad and food-poisoning (49% difference between crude and adjusted relative risk). After adjusting for CAESAR salad, TOMATO was not significantly associated with food-poisoning.

It is likely that CAESAR salad was a vehicle of food-poisoning, and that TOMATO salad was not a vehicle of food-poisoning. Many of those at the luncheon ate CAESAR salad **and** TOMATO salad. CAESAR salad confounded the relationship between TOMATO salad and ILLness. This resulted in a spurious association between TOMATO salad and ILLness.

Stratified analysis with STATCALC

In this exercise you will use the Epi-Info STATCALC module to perform a stratified analysis and calculate the Mantel-Haenszel chi-square and Mantel-Haenszel summary risk measures using the example data from the BATEMAN outbreak.

Start the course software and select the Epi-Info STATCALC module. Once STATCALC starts you will be presented with another menu:

```
+-----+
| Tables (2 x 2, 2 x n) |
| Sample size & power   |
| Chi square for trend  |
+-----+
```

Select the option Tables (2 x 2, 2 x n) and enter the example tables:

TOMATO = +				TOMATO = -			
		ILL				ILL	
CAESAR		+	- Total	CAESAR		+	- Total
	+	23	4 27		+	3	1 4
	-	1	2 3		-	2	9 11
Total		24	6 30	Total		5	10 15

Enter the data for the first table by typing:

```
23 
4  
1  
2  
```

Press the key again. STATCALC displays statistics for the first table:


```
Analysis of Single Table
Odds ratio = 11.50 (0.58 <OR< 429.25*)
Cornfield 95% confidence limits for OR
*Cornfield not accurate. Exact limits preferred.
Relative risk = 2.56 (0.51 <RR< 12.76)
Taylor Series 95% confidence limits for RR
Ignore relative risk if case control study.





Chi-Squares      P-values
-----
Uncorrected      :      4.54      0.0331690 <---
Mantel-Haenszel:      4.39      0.0362394 <---
Yates corrected:      1.88      0.1709035
Fisher exact: 1-tailed P-value: 0.0935961
                2-tailed P-value: 0.0935961

An expected cell value is less than 5.
Fisher exact results recommended.

F2 More Strata; <Enter> No More Strata; F10 Quit
```

Stratified analysis with STATCALC

Press the  Stratum key to enter another stratum or table. Enter the sample data for the second table by typing:

3 
 1 
 2 
 9 


Press the  key again. STATCALC displays statistics for the second table:

Odds ratio = 13.50 (0.59 <OR< 673.24*)
 Cornfield 95% confidence limits for OR
 *Cornfield not accurate. Exact limits preferred.
Relative risk = 4.13 (1.04 <RR< 16.32)
 Taylor Series 95% confidence limits for RR
 Ignore relative risk if case control study.

	Chi-Squares	P-values
	-----	-----
Uncorrected :	4.26	0.0389886 <---
Mantel-Haenszel:	3.98	0.0461182 <---
Yates corrected:	2.09	0.1484537
Fisher exact: 1-tailed P-value:	0.0769231	
2-tailed P-value:	0.0769231	

An expected cell value is less than 5.
 Fisher exact results recommended.

F2 More Strata; <Enter> No More Strata; F10 Quit

Press the  key to instruct STATCALC to calculate and display the results of a stratified analysis:

***** Stratified Analysis *****
 Summary of 2 Tables


 Crude odds ratio for all strata = 19.07
 Mantel-Haenszel Weighted Odds Ratio = 12.50
 Cornfield 95% Confidence Limits
 1.46 < 12.50 < 133.21
Mantel-Haenszel Summary Chi Square = 5.75
P value = 0.01646683 <---

Crude RR for all strata = 3.91
Mantel-Haenszel Weighted Relative Risk
of Disease, given Exposure = 3.14
Greenland/Robins Confidence Limits =
1.06 < MHRR < 9.28


<Enter> for more; F10 to quit.

Examine the output carefully and identify the crude relative risk, the Mantel-Haenszel summary relative risk and its confidence limits, and the Mantel-Haenszel summary chi-square and its associated p-value.

Stratified analysis with STATCALC

Once you have examined the output of the stratified analysis carefully and identified the crude relative risk, the Mantel-Haenszel summary relative risk and its confidence limits, and the Mantel-Haenszel summary chi-square and its associated p-value, press  and STATCALC displays the message:

Press "E" for Exact Confidence Limits or <Enter>

In situations where you are using the odds ratio rather than the relative risk it is possible to obtain a more accurate confidence interval for the odds ratio in STATCALC than the one calculated by ANALYSIS. If you press , STATCALC calculates and displays an *exact* confidence interval for the odds ratio:

```
***Exact Confidence Limits***  
  
Mehta CR, Patel NR, Gray R,  
J. Am. Stat. Assoc., 1985, 78, 969-973.  
Pascal program by ELF Franco & N Campos-Filho  
Ludwig Cancer Institute, Sao Paulo, Brazil
```

```
Exact Lower 95% Confidence Limit = 1.35  
Mantel-Haenszel Weighted Odds Ratio = 12.50  
Exact Upper 95% Confidence Limit = 130.05
```

<Enter> to continue.....

These *exact* confidence limits are different from the confidence limits that STATCALC calculated and displayed earlier. They are also different from the confidence limits that ANALYSIS calculates with the same data. This is because they are calculated using the *exact* or *actual distribution* of the data rather than using an *approximation* such as the *normal* or *chi-square* distribution. The results of exact calculations are always preferred to approximate results, especially when some cells in the table have small expected values. Epi-Info STATCALC will calculate exact confidence limits for the Mantel-Haenszel weighted odds ratio only.

STATCALC does not calculate Woolf's test for heterogeneity of odds ratios.

Analysing several exposure variables at the same time

Using the methods covered so far, you can analyse the effect of an exposure on the disease (using the odds ratio or relative risk), examine its effect having adjusted for other exposures (using the Mantel-Haenszel odds ratio or relative risk) and test whether there is an interaction between these exposures (using Woolf's test).

These stratified analysis techniques can be extended to study the effect of several exposures simultaneously. For example, for the Bateman outbreak, it would be possible to calculate the relative risk for CAESAR salad on food-poisoning after adjusting for the effect of TOMATO salad **and** orange JUICE. Four tables would be produced, one for every combination of having eaten TOMATO salad or having drank orange JUICE.

With several potential confounders stratified analysis methods result in the analysis of many tables which are difficult to interpret and can yield unreliable results. Four potential confounders each with two levels would produce sixteen tables. Some of these tables are likely to contain small numbers which may produce unstable or unreliable results. A better approach would be to use *logistic regression*. This is illustrated in the following pages using data from the SALEX outbreak.

The SALEX outbreak

On Saturday 17th October 1992, eighty-two people attended a buffet meal at a sports club. Within fourteen to twenty-four hours fifty-one of the participants developed diarrhoea, with nausea, vomiting, abdominal pain and fever as other common symptoms.

Data from this outbreak is stored in the file SALEX.REC. The variables in the file are:

CASE	Case or control
HAM	Baked ham
BEEF	Roast beef
EGGS	Eggs
MUSHROOM	Mushroom flan
PEPPER	Pepper flan
PORKPIE	Pork pie
PASTA	Pasta salad
RICE	Rice salad
LETTUCE	Lettuce
TOMATO	Tomato salad
COLESLAW	Coleslaw
CRISPS	Crisps
PEACHCAKE	Peach cake
CHOCOLATE	Chocolate cake
FRUIT	Tropical fruit salad
TRIFLE	Trifle
ALMONDS	Almonds

Data is available for seventy-seven of the eighty-two people who attended the sports club buffet. This data was collected using a case-control study design (51 cases, 26 controls).

Using ANALYSIS to produce 2-by-2 tables

Start the course software and select Epi-Info ANALYSIS from the main menu. Once ANALYSIS has started issue the command:

```
read salex.rec
```

to retrieve the data for the SALEX outbreak. Enter the command:

```
set statistics = on
```

to instruct ANALYSIS to produce a full set of statistics in the output of each command. Issue the command:

```
tables * case
```

to examine the association between each of the exposure variables (* = all variables) and the outcome variable (CASE). Examine the output of this command carefully and complete the following table:

	Lower 95% CL	Odds Ratio	Upper 95% CL	Yates' X^2	p-value
HAM					
BEEF					
EGGS					
MUSHROOM					
PEPPER					
PORKPIE					
PASTA					
RICE					
LETTUCE					
TOMATO					
COLESLAW					
CRISPS					
PEACHCAKE					
CHOCOLATE					
FRUIT					
TRIFLE					
ALMONDS					

We are interested in the odds ratio because this is a **case-control study**. The relative risk would **not** be a valid measure of effect. This simple *univariate* analysis shows associations between several exposure variables (HAM, EGGS, PEPPER, PASTA, RICE, LETTUCE, COLESLAW) and the outcome variable (CASE). All of these associations may be real but it is more likely that some of these associations arose due to confounding and that only one or two of these foods caused the food-poisoning. This is because persons who ate these foods also ate some of the others. It is possible to perform a stratified analysis to obtain the odds ratio of eating each of these seven foods after adjusting for the effect of eating the other six foods. This procedure would involve breaking up the data into many 2-by-2 tables. This is cumbersome to do, difficult to interpret, and may produce unreliable results. An alternative method for simultaneously analysing the effect of several exposures is to use *logistic regression*.

Note that Epi-Info ANALYSIS states that the odds ratios for some exposures may have inaccurate confidence intervals. This is because such odds ratios are based on small numbers. Better estimates of these confidence intervals would be the exact confidence intervals which can be obtained in STATCALC.

The logistic model for a single exposure variable

Logistic regression is a statistical technique used for simultaneously estimating the effects of **several** exposure variables on an outcome variable. We will first use this technique to estimate the effect of a **single** exposure (EGGS) on the outcome variable (CASE). This is the 2-by-2 table for EGGS by CASE:

EGGS	CASE		Total
	+	-	
+	40	6	46
-	10	20	30
Total	50	26	76

Various measures of disease (exposure-specific risk and absolute risk) and measures of effect (relative risk and odds ratio) can be calculated from this table. We want to relate a measure of disease or effect to the exposure of having eaten EGGS. Does eating EGGS increase the risk of being a case? The relationship between the risk of being a case and eating EGGS can be represented by a *model*.

As an example of a model, consider the relationship between *body mass index* (BMI) and weight and height. BMI can be expressed in terms of weight and height using the following relationship or model:

$$\text{BMI} = \text{weight} / \text{height}^2$$

It is possible to work out the BMI for every possible pair of values of weight and height. The model provides a useful way of summarising all possible combinations of BMI, weight, and height in one simple relationship.

We might represent the relationship between the risk of being a case (y) and eating EGGS ($x = 1$ for someone who ate EGGS, $x = 0$ for someone who did not eat EGGS) using the following model:

$$y = \alpha + \beta x$$

An important difference between these two models is that the former is *deterministic*. For any given combination of weight and height, the model will *determine* a single value of BMI. The latter model is *probabilistic*. For a given value of EGGS the model can only give a *probability* for the risk of being a case. The relationships between exposures and diseases are usually represented by probabilistic models. The models covered in this book are all probabilistic models.

Consider the model:

$$y = \alpha + \beta x$$

for the relationship between the risk of being a case and eating EGGS. There are techniques available for estimating suitable values of α and β . If we plotted this model for particular values of x and y , the graph would be *linear*. Such a *linear* model runs into problems when values of y (in this case, the risk of disease) have to lie within a certain range (such as between 0 and 1). The *linear* model assumes that y can take any value. There are no measures of disease or effect which meet this constraint. However, there are *functions* of these measures which do meet this constraint. One such function is the *natural logarithm* (denoted here by \log) of the *odds of disease*.

The logistic model for a single exposure variable

Using the standard notation for a 2-by-2 table:

Exposure	Outcome		Totals
	Cases	Non-Cases	
Present	a	b	a + b
Absent	c	d	c + d
Totals	a + c	b + d	a + b + c + d

a = number exposed and ill
 b = number exposed and not ill
 c = number unexposed and ill
 d = number unexposed and not ill
 a + b = number exposed
 c + d = number unexposed
 a + c = number ill
 b + d = number not ill
 a + b + c + d = number in study

the odds of disease (or the odds of being a case) for those exposed is a / b . The odds of disease for those not exposed is c / d . The overall *odds of disease* is the odds of disease irrespective of exposure and this is used to model the relationship between being a CASE and eating EGGS. The overall *odds of disease* is estimated as:

$$(a + c) / (b + d)$$

The *log odds of disease* is estimated as:

$$\log((a + c) / (b + d))$$

In the example table:

EGGS	CASE		Total
	+	-	
+	40	6	46
-	10	20	30
Total	50	26	76

the odds of disease is equal to:

$$50 / 26 = 1.92$$

So the log odds of disease is equal to:

$$\log(1.92) = 0.65$$

Logistic regression is a technique which allows us to relate the log odds of disease to an explanatory variable, x , using the *linear* model:

$$\log \text{ odds of disease} = \alpha + \beta x$$

where α and β are *coefficients* to be determined.

Estimating the coefficients in the model

Consider the situation with one explanatory variable (having eaten EGGS):

EGGS	CASE		Total
	+	-	
+	40	6	46
-	10	20	30
Total	50	26	76

We will use the data shown in this table to estimate suitable values for the coefficients α and β in this model:

$$\log \text{ odds of disease} = \alpha + \beta x$$

There are only two possible values of x for this model: these are $x = 0$ for someone who did not eat EGGS and $x = 1$ for someone who ate EGGS. We can calculate suitable values of α and β by calculating the log odds of disease at $x = 0$ and $x = 1$ for our data. The estimated value of the log odds of disease at $x = 0$ is used to find a suitable value of α . The log odds of disease for someone who did not eat EGGS ($x = 0$) is:

$$\begin{aligned} \log \text{ odds of disease} &= \alpha + \beta x \\ &= \alpha + \beta * 0 \\ &= \alpha \end{aligned}$$

$$\begin{aligned} \log \text{ odds of disease} &= \log(c / d) = \log(10 / 20) = -0.69 \\ \alpha &= -0.69 \end{aligned}$$

A suitable value for α is - 0.69. We can find a suitable value for β by calculating the log odds of disease at $x = 1$ and using $\alpha = - 0.69$. The log odds of disease for someone who ate EGGS (for $x = 1$) is:

$$\begin{aligned} \log \text{ odds of disease} &= \alpha + \beta x \\ &= \alpha + \beta * 1 \\ &= \alpha + \beta \end{aligned}$$

$$\begin{aligned} \log \text{ odds of disease} &= \log(a / b) = \log(40 / 6) = 1.90 \\ \alpha + \beta &= 1.90 \\ \beta &= 1.90 - \alpha = 1.90 - (-0.69) = 1.90 + 0.69 \\ \beta &= 2.59 \end{aligned}$$

A suitable value for β , which is a measure of the excess risk associated with eating EGGS, is 2.59. The *logistic model* for this table can be written as:

$$\log \text{ odds of disease} = -0.69 + 2.59x$$

This means that for someone who ate EGGS ($x = 1$), the log odds of disease is equal to $-0.69 + 2.59 = 1.90$ and for someone who did not eat EGGS, the log odds of disease is equal to -0.69 . We have used the log odds of disease to estimate suitable values for α and β . Conversely, given values for α and β we can estimate the log odds of disease and the odds ratio:

$$\begin{aligned} \alpha &= \text{the log odds of disease in the unexposed group} \\ \beta &= \text{the log odds ratio associated with being exposed} \end{aligned}$$

The calculation of the coefficients of the model (α and β) is complicated when there is more than one exposure variable and is best left to purpose-designed computer programs such as LOGISTIC.

Fitting the model using LOGISTIC

Here are the results from the logistic regression of EGGS carried out using the course software:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-.6931	.3873	-1.7897	.0735
EGGS	2.5903	.5845	4.4315	.0000
95.0-% confidence limits				
	Coefficient		Odds ratio	
	lower limit	upper limit	lower limit	upper limit
EGGS	1.4447	2.5903	3.7359	41.9249

The column labelled *coefficients* in the above output lists the *estimates* of α and β : α is equal to the coefficient for the CONSTANT (-0.69) and β is equal to the coefficient for eating EGGS (2.59). The model which relates eating EGGS to the log odds of disease is therefore:

$$\log \text{ odds of disease} = -0.69 + 2.59x$$

This is the same as the logistic model for the 2-by-2 table shown previously. The coefficient for β is equal to the log(odds ratio) which means that the odds ratio for eating eggs is:

$$\exp(2.5903) = e^{2.5903} = 13.3333$$

where \exp denotes the *exponential* function (or *anti-logarithm*). We can check this figure against the one calculated directly from the 2-by-2 table:

EGGS	CASE		Total
	+	-	
+	40	6	46
-	10	20	30
Total	50	26	76

$$\text{Odds ratio} = (40 / 6) / (10 / 20) = 13.3333$$

The logistic regression module of the course software also calculates a 95% confidence interval for the odds ratio:

Odds ratio		
lower limit		upper limit
4.2404	13.3333	41.9249

Significance tests

Once the model which best *describes* the data has been found, it is important to test the significance of those variables included in the model. The logistic regression module gives the results of two tests of significance of exposure variables.

The first test of significance is the *likelihood ratio* test. This measures the significance of a variable by comparing the model which includes this variable to the model which does not include this variable. For example, to assess the significance of EGGS in the model, you would compare the model with EGGS to the model without EGGS.

The model with EGGS is:

$$\log \text{ odds of disease} = -0.69 + 2.59 * \text{EGGS}$$

The model without EGGS is of the form:

$$y = \alpha$$

which is the appropriate model when the probability (risk) of disease is constant and not dependent upon having eaten EGGS. If you fit this model to the data, you would get:

$$\log \text{ odds of disease} = \log ((a + c) / (b + d)) = \log (50 / 26) = 0.654$$

The significance of EGGS is assessed by comparing the probability that:

$$\alpha = -0.69, \beta = 2.59$$

with the probability that:

$$\alpha = 0.65, \beta = 0$$

These two probabilities are called *likelihoods* and a good measure of their difference is given by r , the ratio of one to the other. This is the likelihood of the model without EGGS divided by the likelihood of the model with EGGS. For example, if the likelihood of the model without EGGS was 0.1 and the likelihood of the model with EGGS was 0.5, the *ratio of likelihoods* would be $0.1 / 0.5 = 0.2$ which indicates that the model without EGGS was only 20% as likely as the model with EGGS. A formal measure of the difference in likelihoods is given by the *likelihood ratio statistic* which is equal to:

$$-2 * \log(r)$$

and has a chi-square distribution with one degree of freedom. Large values of the likelihood ratio statistic indicate that the model with EGGS is more probable than the model without EGGS for the given data (i.e. that EGGS is associated with being a case). The likelihood ratio test assesses the significance of EGGS by comparing the likelihood ratio statistic with a chi-square distribution with one degree of freedom. Here is the result of the likelihood ratio test for EGGS:

H0:	coeff = 0	lr statistic (1 df)	P-value
	EGGS	23.8338	.0000

The likelihood ratio statistic for EGGS is 23.83. This indicates that there is a large difference between the likelihoods of the two models. A chi-square statistic of 23.83 with one degree of freedom is highly significant ($p < 0.0001$). This result suggests that EGGS is strongly associated with being a case.

Significance tests

The second test of significance is *Wald's test*. This test measures the significance of a variable by testing whether its coefficient in the model is significantly different from zero. Since the coefficient is equal to the log of the odds ratio and $\log(1) = 0$ this is equivalent to testing whether an odds ratio is significantly different from one.

To assess the significance of EGGS in the model, you would test whether the coefficient of EGGS ($\beta = 2.59$) is significantly different from zero. A good measure of the difference between a coefficient and zero is given by R , the ratio of the coefficient to its *standard error* (a measure of how precisely the coefficient can be estimated from the sample). Large values of R indicate that the coefficient is not equal to zero. Wald's test assesses the significance of EGGS by comparing the square of R to a chi-square distribution with one degree of freedom. Here is the result of Wald's test for EGGS:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-.6931	.3873	-1.7897	.0735
EGGS	2.5903	.5845	4.4315	.0000

Wald's test $(\text{Coef/SE})^2$ gives 4.4315^2 which equals 19.6382 which, when compared to a chi-square distribution with one degree of freedom, is also highly significant ($p < 0.0001$). It is unlikely that the observed association between eating EGGS and being a case arose by chance alone. Note that comparing $(\text{Coef/SE})^2$ to a chi-square distribution with one degree of freedom is equivalent to comparing (Coef/SE) to a Normal distribution. Some textbooks and computer programs describe the latter procedure but the p-values from the two are identical.

Most statisticians recommend the use of the likelihood ratio test rather than Wald's test. This is because Wald's test is based on an *approximation* and the results may not be as accurate as those from the likelihood ratio test. In some situations, the p-value from Wald's test is larger than it should be. This may cause a significant variable to appear to be non-significant. This can occur when the absolute value of the coefficient is very large. In most situations, the p-value from Wald's test is similar to that from the likelihood ratio test.

Starting LOGISTIC

Start the course software and select the LOGISTIC Regression module from the main menu. Once LOGISTIC starts you will be presented with its start-up screen:

```
-----
##          #####          #####          #####          #####          #####          #####
##          ##          ##          ##          ##          ##          ##          ##          ##
##          ##          ##          ##          ##          ##          ##          ##          ##
##          ##          ##          ##          ##          ##          ##          ##          ##
##          ##          ##          ##          ##          ##          ##          ##          ##
#####          #####          #####          #####          #####          #####          #####
-----

                          Version  3.11Ef
                    (c) 1993  by Gerard E. Dallal

                "One of many STATOOLS(tm)..."

Please acknowledge LOGISTIC in any manuscript that makes
use of its calculations.  A suitable reference is Dallal
GE (1988), "LOGISTIC: A Logistic Regression Program for
the IBM PC," The American Statistician, 42, 272.
```

At the bottom of the screen is the command prompt:

```
LOGIT>
```

This is where you enter commands.

LOGISTIC can read Epi-Info data files and automatically converts Epi-Info *Yes/No* type variables into *binary* (1 = YES, 0 = NO) variables for use in logistic models.

Using LOGISTIC to fit a logistic regression model

The first command you should enter is:

```
read salex
```

to retrieve the data file SALEX.REC. You should **not** specify the .REC extension when retrieving a file in LOGISTIC as doing so will confuse the program. LOGISTIC responds with a list of the variables in the data file:

CASE	HAM	BEEF	EGGS	MUSHROOM
PEPPER	PORKPIE	PASTA	RICE	LETTUCE
TOMATO	COLESLAW	CRISPS	PEACHCAKE	CHOCOLATE
FRUIT	TRIFLE	ALMONDS		

and displays the command prompt:

```
LOGIT>
```

ready to receive new commands. After we have retrieved the data file we must define our model. The most simple model is the model with no exposure variables:

```
log odds of disease =  $\alpha$ 
```

This is defined by entering the command:

```
model case = constant
```

where CASE is the variable that defines whether a particular person belongs to the case or control group. Having defined the model, instruct LOGISTIC to fit the model by issuing the command:

```
estimate
```

LOGISTIC responds with information about the fitted model:

	Coefficient	Standard Error	Coef/SE	"P value"		
CONSTANT	.6737	.2410	2.7958	.0052		
95.0-% confidence limits						
	Coefficient		Odds ratio			
	lower limit	upper limit	lower limit		upper limit	
CONSTANT	.2014	.6737	1.1460	1.2231	1.9615	3.1457

In this simple model the odds ratio is the ratio of cases to controls ($51 / 26 = 1.96$) in the dataset. This model is very simple and tells us nothing about the factors that contributed to the development of food-poisoning. It suggests that the odds of food-poisoning for any person at the supper is 1.96 irrespective of what was eaten. This interpretation of the odds is only valid for a cohort study. This interpretation is **not** valid for a case-control study (where the investigator has control over the ratio of cases to controls and, hence, the odds of being a case).

Adding variables to the model

This simple model (log odds of disease = 0.67, odds of disease = $\exp(0.67) = e^{0.67} = 1.96$) is unlikely to be the most appropriate model for the data. It is likely that particular foods are associated with food-poisoning. These foods should be incorporated as *terms* or exposure variables in the model. We will investigate whether each particular food is associated with food-poisoning.

We will now use LOGISTIC to investigate the association between eating HAM and developing food-poisoning. To do this we must instruct LOGISTIC to add HAM as a term in the model using the command:

add ham

and then instruct LOGISTIC to estimate the model again by typing the command:

estimate

LOGISTIC responds with information about the fitted model:

```
Log likelihood    =    -45.8599
Likelihood ratio =      6.7589      1 df  (P = .0093)

Dependent Variable =      CASE

      Coefficient      Standard      Coef/SE      "P value"
      Error
CONSTANT      -.5878      .5578      -1.0538      .2920
HAM          1.5832      .6258      2.5298      .0114

      95.0-% confidence limits
      Coefficient      upper      lower      Odds ratio      upper
      lower      limit      limit      limit      limit
HAM      .3566      1.5832      2.8098      1.4285      4.8706      16.6071
```

Examine this output carefully. The odds ratio for HAM is 4.87 (95% CI = 1.43 - 16.61). Wald's test $(\text{Coef/SE})^2$ gives 6.3998 (2.5298^2) which is significant ($p < 0.05$). It is unlikely that the observed association between eating HAM and being a case arose by chance alone. You can also examine the results of the likelihood ratio test by issuing the command:

lr

LOGISTIC responds with a likelihood ratio test for HAM:

```
H0:  coeff = 0      lr statistic (1 df)      P-value
      HAM          6.7589          .0093
```

The likelihood ratio for HAM is 6.76 which, when compared to a chi-square with one degree of freedom, is also significant ($p < 0.01$). The model with HAM is more probable than the model without HAM suggesting that HAM is associated with being a case.

Leave LOGISTIC and return to the course software main menu by typing the command:

quit

The logistic model for two exposure variables

Consider the situation with two explanatory variables, say, eating EGGS and eating PASTA salad. You want to determine whether or not eating EGGS and / or PASTA salad increases the probability of being a case. Logistic regression can be used to relate the log odds of disease to the two explanatory variables x_1 ($x_1 = 1$ for someone who ate EGGS and $x_1 = 0$ otherwise) and x_2 ($x_2 = 1$ for someone who ate PASTA salad and $x_2 = 0$ otherwise) using the linear model:

$$\log \text{ odds of disease} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

where α , β_1 and β_2 are coefficients to be determined. There are many values which α , β_1 and β_2 could take. The probability of obtaining a particular set of values for α , β_1 and β_2 is called the likelihood. Logistic regression finds the set of values for α , β_1 and β_2 which has the maximum likelihood. This is called the *method of maximum likelihood*. The maximum likelihood estimates of β_1 and β_2 are:

$$\beta_1 = \log (\text{odds ratio for EGGS adjusted for PASTA})$$

$$\beta_2 = \log (\text{odds ratio for PASTA adjusted for EGGS})$$

Here are the results from this logistic regression carried out with LOGISTIC:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-1.3768	.5051	-2.7255	.0064
EGGS	2.7079	.6514	4.1569	.0000
PASTA	2.2042	.7824	2.8171	.0048

95.0-% confidence limits						
	Coefficient			Odds ratio		
	lower limit		upper limit	lower limit		upper limit
EGGS	1.4311	2.7079	3.9846	4.1834	14.9974	53.7648
PASTA	.6707	2.2042	3.7378	1.9555	9.0632	42.0055

The coefficients in the above output are the maximum likelihood estimates of α , β_1 and β_2 : α is equal to the coefficient for the constant (-1.38), β_1 is equal to the coefficient for eating EGGS (2.71), and β_2 is equal to the coefficient for eating PASTA salad (2.20). Since EGGS and PASTA salad have been included in the model, the resulting coefficients are said to have been *adjusted* for one another. The coefficients for β_1 and β_2 are the log(adjusted odds ratio) for eating EGGS and PASTA salad respectively. The adjusted odds ratios and their 95% confidence intervals are also shown in the LOGISTIC output: for EGGS, the adjusted odds ratio = 14.99 (95% CI = 4.18 - 53.76) and for PASTA salad, the adjusted odds ratio = 9.06 (95% CI = 1.96 - 42.01). The adjusted odds ratio will be different from the Mantel-Haenszel summary odds ratios:

	Crude OR	Adjusted OR	M-H OR
EGGS	13.33	14.99	14.63
PASTA	7.67	9.06	7.83

because different methods (maximum likelihood and weighted averages) were used to obtain them.

Significance tests for the model with two exposure variables

The likelihood ratio statistic for the model with **both** EGGS and PASTA salad appears at the top of the output from the **estimate** command:

```
Log likelihood      =      -31.8752
Likelihood ratio =      33.8978      2 df  (P = .0000)
```

This likelihood ratio statistic is comparing the likelihood of the model with two variables (EGGS and PASTA) to the likelihood of the model with no exposure variables. There are two variables so the chi-square has two degrees of freedom. The likelihood ratio statistic for the model with both EGGS and PASTA salad is 33.89 which, when compared to a chi-square distribution with two degrees of freedom, is highly significant ($p < 0.0001$). This result suggests that EGGS or PASTA salad or both are strongly associated with being a case.

It is possible to determine the effect of each variable separately using the likelihood ratio test of the models with and without each variable. To assess the effect of EGGS after adjusting for the effect of PASTA salad, the model with EGGS and PASTA salad is compared to the model with PASTA salad alone:

```
H0:  coeff = 0      lr statistic (1 df)      P-value
      EGGS          21.7764          .0000
      PASTA          10.0640          .0015
```

The likelihood ratio statistic for the model with EGGS and PASTA salad compared to the model with PASTA salad alone is 21.78 which, when compared to a chi-square with one degree of freedom, is highly significant ($p < 0.0001$). This result suggests that after adjusting for the effect of PASTA salad, EGGS is still significantly associated with being a case.

The likelihood ratio statistic for the model with EGGS and PASTA salad compared to the model with EGGS alone is 10.06 which, when compared to a chi-square with one degree of freedom, is highly significant ($p = 0.0015$). This result suggests that after adjusting for the effect of EGGS, PASTA salad is still significantly associated with being a case.

Wald's test gives 17.23 (4.1569^2) for EGGS ($p < 0.0001$) and 7.94 (2.8171^2) for PASTA salad ($p = 0.0048$), which support the results from the likelihood ratio tests:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-1.3768	.5051	-2.7255	.0064
EGGS	2.7079	.6514	4.1569	.0000
PASTA	2.2042	.7824	2.8171	.0048

Interactions

It is only appropriate to calculate odds ratios adjusted for the effect of another variable if there is no *interaction* between the explanatory variables. It is possible to test for an interaction using the logistic model. To test for an interaction between EGGS and PASTA salad, an *interaction term* for EGGS and PASTA salad is added to the model. The model with the interaction term is of the form:

$$\log \text{ odds of disease} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

where:

$$\begin{aligned} \beta_1 &= \text{coefficient for EGGS} \\ \beta_2 &= \text{coefficient for PASTA} \end{aligned}$$

and:

$$\beta_{12} = \text{coefficient for the interaction between EGGS and PASTA}$$

The interaction term ($x_1 x_2$) is the *product* of EGGS and PASTA salad:

	PASTA = 1	PASTA = 0
EGGS = 1	$x_1 x_2 = 1 * 1 = 1$	$x_1 x_2 = 1 * 0 = 0$
EGGS = 0	$x_1 x_2 = 0 * 1 = 0$	$x_1 x_2 = 0 * 0 = 0$

This model is appropriate when the combination of EGGS and PASTA salad has a greater or lesser effect on being a case than either EGGS or PASTA salad alone. In this situation, it is appropriate to adjust the main effects of EGGS (β_1) and PASTA salad (β_2) so that the effect caused by the interaction of EGGS and PASTA salad (β_{12}) is taken into account.

LOGISTIC does not calculate the likelihood ratio test for the interaction term so Wald's test should be used to assess the significance of the interaction term:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-1.5041	.5528	-2.7210	.0065
EGGS	2.9392	.7438	3.9518	.0001
PASTA	2.6027	.9860	2.6396	.0083
EGGS * PASTA	-1.0933	1.5075	-.7253	.4683

Wald's test of the interaction gives -0.7253^2 ($= 0.5261$) for the interaction term EGGS * PASTA ($p = 0.4683$) which suggests that the coefficient of the interaction term is not significantly different from zero and that there is no interaction between EGGS and PASTA salad. The model without the interaction term is more appropriate.

If the interaction is significant then it is appropriate to fit separate models of the effect of EGGS for those persons who ate PASTA salad and those who did not (or to fit separate models of the effect of PASTA salad for those persons who ate EGGS and those who did not).

Using LOGISTIC to fit a logistic model with two exposure variables

Start LOGISTIC and at the command prompt type:

```
read salex
```

to retrieve the SALEX data file. Specify the logistic model by typing the command:

```
model case = constant + pepper + pasta
```

and instruct LOGISTIC to fit the model by typing the command:

```
estimate
```

LOGISTIC responds with information about the model:

```
Log likelihood      =      -38.0704
Likelihood ratio =      16.8414      2 df  (P = .0002)

Dependent Variable =      CASE

      Coefficient      Standard      Coef/SE      "P value"
      Error

CONSTANT      -.2578      .3333      -.7735      .4392
PEPPER      1.7206      .7047      2.4416      .0146
PASTA      1.4908      .7138      2.0885      .0368

      95.0-% confidence limits
      Coefficient      Odds ratio
      lower      upper      lower      upper
      limit      limit      limit      limit

PEPPER      .3394      1.7206      3.1018      1.4041      5.5878      22.2370
PASTA      .0917      1.4908      2.8900      1.0961      4.4409      17.9927
```

Examine the output carefully. Identify the likelihood ratio test for this model. Identify the odd ratios and Wald's test statistic for each of the exposure variables and complete the table below:

	Lower 95% CL	Odds Ratio	Upper 95% CL	Wald's Test	p-value
PEPPER	1.4041	5.5878	22.2370	2.4416 ²	0.0146
PASTA					

Using LOGISTIC to fit a logistic model with two exposure variables

Instruct LOGISTIC to display the likelihood ratio tests for each possible model by issuing the command:

```
lr
```

LOGISTIC responds by displaying the likelihood ratio tests for each variable in the model:

H0:	coeff = 0	lr statistic (1 df)	P-value
	PEPPER	7.1696	.0074
	PASTA	5.0571	.0245

The likelihood ratio statistic for PEPPER assesses the effect of PEPPER after adjusting for the effect of PASTA salad. The likelihood ratio statistic for the model with PEPPER and PASTA compared to the model with PASTA alone is 7.17 which, when compared to a chi-square with one degree of freedom is highly significant ($p = 0.0074$). This result suggests that after adjusting for the effect of PASTA salad, PEPPER is still significantly associated with being a case.

The likelihood ratio statistic for PASTA is 5.06 which, when compared to a chi-square with one degree of freedom is also highly significant ($p = 0.0245$). This suggests that after adjusting for the effect of PEPPER, PASTA salad is still significantly associated with being a case.

Fitting a model with an interaction term

Before trying to fit a model with an interaction term you should note that LOGISTIC sometimes has problems adding or removing terms in the model *after* an interaction term has been specified. If this happens you should quit and restart the program. LOGISTIC does not allow you to use the commands **add** or **delete** to add or remove interaction terms. Interaction terms must be specified on the **model** command line.

To test for an interaction between PEPPER flan and PASTA salad we need to redefine our model by issuing the command:

```
model case = constant + pepper + pasta + pepper*pasta
```

The term PEPPER*PASTA defines the interaction term in the model. You cannot use the **add** command to add an interaction term to the model. You must add an interaction term to the model by redefining the model using the **model** command. Fit the new model with the command:

```
estimate
```

LOGISTIC responds with information about the fitted model:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-.2877	.3416	-.8422	.3996
PEPPER	1.8971	.8466	2.2410	.0250
PASTA	1.6740	.8612	1.9438	.0519
PEPPER * PASTA	-.6444	1.5534	-.4148	.6783

95.0-% confidence limits						
	Coefficient		Odds ratio			
	lower limit	upper limit	lower limit	upper limit	lower limit	upper limit
PEPPER	.2379	1.8971	3.5564	1.2686	6.6667	35.0351
PASTA	-.0139	1.6740	3.3619	.9862	5.3333	28.8439
PEPPER * PASTA	-3.6890	-.6444	2.4003	.0250	.5250	11.0261

Identify Wald's test for the interaction term and its associated p-value. The values suggest that there is no interaction between PEPPER flan and PASTA salad ($p = 0.6783$). In this case, the model without the interaction term is more appropriate.

Leave LOGISTIC and return to the course software main menu by typing the command:

```
quit
```

The logistic model for several exposure variables

Consider the situation with several explanatory variables: HAM, EGGS, PEPPER flan, PASTA salad, RICE, LETTUCE, and COLESLAW. Each of these variables appear to be significantly associated with being a case although some of these associations may be due to confounding.

You want to determine whether or not eating any combination of HAM, EGGS, PEPPER flan, PASTA salad, RICE, LETTUCE and COLESLAW increases the probability of being a case. Logistic regression can be used to relate the log odds of disease to these seven explanatory variables x_1, \dots, x_7 using the linear model:

$$\log \text{ odds of disease} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

where $\alpha, \beta_1, \beta_2, \beta_3, \dots, \beta_7$ are coefficients to be determined. This model allows you to assess the effect of each variable after adjusting for the effect of the other six variables. There are many possible values which $\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$, and β_7 could take. Logistic regression uses the method of maximum likelihood to find the most probable values of $\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$, and β_7 for our data. The maximum likelihood estimates of β_1 to β_7 are:

$\beta_1 = \log$ (odds ratio for HAM adjusted for the other 6 variables)
 $\beta_2 = \log$ (odds ratio for EGGS adjusted for the other 6 variables)
 $\beta_3 = \log$ (odds ratio for PEPPER adjusted for the other 6 variables)
 $\beta_4 = \log$ (odds ratio for PASTA adjusted for the other 6 variables)
 $\beta_5 = \log$ (odds ratio for RICE adjusted for the other 6 variables)
 $\beta_6 = \log$ (odds ratio for LETTUCE adjusted for the other 6 variables)
 $\beta_7 = \log$ (odds ratio for COLESLAW adjusted for the other 6 variables)

Here are the results from this logistic regression carried out with LOGISTIC:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-2.1228	.9166	-2.3160	.0206
HAM	.3579	1.1184	.3200	.7490
EGGS	2.4839	1.0291	2.4136	.0158
PEPPER	.2213	.9212	.2402	.8101
PASTA	1.3960	.9958	1.4019	.1609
RICE	-.3054	.9403	-.3248	.7454
LETTUCE	3.1396	1.1781	2.6649	.0077
COLESLAW	.3113	.9545	.3262	.7443

95.0-% confidence limits						
	Coefficient		Odds ratio			
	lower limit	upper limit	lower limit	upper limit		
HAM	-1.8341	.3579	2.5498	.1598	1.4303	12.8049
EGGS	.4669	2.4839	4.5009	1.5950	11.9875	90.0948
PEPPER	-1.5842	.2213	2.0269	.2051	1.2477	7.5903
PASTA	-.5557	1.3960	3.3478	.5737	4.0391	28.4388
RICE	-2.1484	-.3054	1.5377	.1167	.7368	4.6537
LETTUCE	.8305	3.1396	5.4487	2.2946	23.0954	232.4614
COLESLAW	-1.5593	.3113	2.1820	.2103	1.3653	8.8643

Interpreting the coefficients in the model

The coefficients in the output:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-2.1228	.9166	-2.3160	.0206
HAM	.3579	1.1184	.3200	.7490
EGGS	2.4839	1.0291	2.4136	.0158
PEPPER	.2213	.9212	.2402	.8101
PASTA	1.3960	.9958	1.4019	.1609
RICE	-.3054	.9403	-.3248	.7454
LETTUCE	3.1396	1.1781	2.6649	.0077
COLESLAW	.3113	.9545	.3262	.7443

are the maximum likelihood estimates of α , and β_1 to β_7 :

α	is equal to the coefficient for the constant	(-2.1228)
β_1	is equal to the coefficient for eating HAM	(.3579)
β_2	is equal to the coefficient for eating EGGS	(2.4839)
β_3	is equal to the coefficient for eating PEPPER flan	(.2213)
β_4	is equal to the coefficient for eating PASTA salad	(1.3960)
β_5	is equal to the coefficient for eating RICE	(-.3054)
β_6	is equal to the coefficient for eating LETTUCE	(3.1396)
β_7	is equal to the coefficient for eating COLESLAW	(.3113)

Since all seven food items have been included in the model, the resulting coefficients are said to have been adjusted for one another. The coefficients for β_1 to β_7 are the log(adjusted odds ratio) for eating HAM, EGGS, PEPPER flan, PASTA salad, RICE, LETTUCE, and COLESLAW respectively. The adjusted odds ratios and their 95% confidence intervals are also shown in the LOGISTIC output:

	95.0-% confidence limits		Odds ratio	
	lower limit	upper limit	lower limit	upper limit
HAM	-1.8341	.3579	.1598	1.4303
EGGS	.4669	2.4839	1.5950	11.9875
PEPPER	-1.5842	.2213	.2051	1.2477
PASTA	-.5557	1.3960	.5737	4.0391
RICE	-2.1484	-.3054	.1167	.7368
LETTUCE	.8305	3.1396	2.2946	23.0954
COLESLAW	-1.5593	.3113	.2103	1.3653

Significance tests

The likelihood ratio statistic for the model appears at the top of the output from the **estimate** command:

```
Log likelihood   =      -23.7165
Likelihood ratio =      44.6887      7 df  (P = .0000)
```

The likelihood ratio statistic for this model is 44.6887 which, when compared to a chi-square with seven degrees of freedom, is highly significant ($p < 0.0001$). The likelihood ratio test is testing seven variables so the chi-square has seven degrees of freedom. This result suggests that a particular combination of these seven foods is strongly associated with being a case.

It is possible to determine the effect of each variable separately using the likelihood ratio test of the models with and without that variable respectively. For example, to assess the effect of EGGS, the model with EGGS and the other six variables is compared to the model with the other six variables alone.

The likelihood ratio statistic for the model with EGGS and the other six variables compared to the model with the other six variables (i.e. without EGGS) is 6.8292 which, when compared to a chi-square with one degree of freedom, is significant ($p < 0.01$). The likelihood ratio statistics associated with the other six variables are:

```
H0:  coeff = 0      lr statistic (1 df)      P-value
      HAM           .1032                   .7480
      EGGS          6.8292                   .0090
      PEPPER        .0577                   .8102
      PASTA          2.1203                   .1454
      RICE           .1063                   .7444
      LETTUCE       11.4560                  .0007
      COLESLAW       .1065                   .7441
```

These results suggest that EGGS and LETTUCE are significantly associated with being a case after adjusting for the effects of the other foods in the model. This is reflected in the odds ratios and confidence limits reported by LOGISTIC:

		95.0-% confidence limits			Odds ratio	
		Coefficient				
	lower		upper	lower		upper
	limit		limit	limit		limit
HAM	-1.8341	.3579	2.5498	.1598	1.4303	12.8049
EGGS	.4669	2.4839	4.5009	1.5950	11.9875	90.0948
PEPPER	-1.5842	.2213	2.0269	.2051	1.2477	7.5903
PASTA	-.5557	1.3960	3.3478	.5737	4.0391	28.4388
RICE	-2.1484	-.3054	1.5377	.1167	.7368	4.6537
LETTUCE	.8305	3.1396	5.4487	2.2946	23.0954	232.4614
COLESLAW	-1.5593	.3113	2.1820	.2103	1.3653	8.8643

Variables which are not significant and do not add anything to the model's description of the data should be removed from the model. It is best to remove one variable at a time from the model so that any changes and effects can be attributed to that one variable being removed.

Using LOGISTIC to fit a model with several exposure variables

Start LOGISTIC and at the LOGIT> prompt type:

```
read salex
```

to retrieve the SALEX data file. Specify the logistic model and instruct LOGISTIC to fit the model by typing the commands:

```
model case = constant + ham + eggs + pepper + pasta + rice  
add lettuce  
add coleslaw  
estimate
```

There is nothing special about using the **add** command in this way. It is used here to extend the model defined by the **model** command. The **constant** term must always come first on the **model** command line but the order in which the variables are listed is not important. LOGISTIC responds with information about the model:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-2.1228	.9166	-2.3160	.0206
HAM	.3579	1.1184	.3200	.7490
EGGS	2.4839	1.0291	2.4136	.0158
PEPPER	.2213	.9212	.2402	.8101
PASTA	1.3960	.9958	1.4019	.1609
RICE	-.3054	.9403	-.3248	.7454
LETTUCE	3.1396	1.1781	2.6649	.0077
COLESLAW	.3113	.9545	.3262	.7443

95.0-% confidence limits						
	Coefficient		Odds ratio			
	lower limit	upper limit	lower limit	upper limit	lower limit	upper limit
HAM	-1.8341	.3579	2.5498	.1598	1.4303	12.8049
EGGS	.4669	2.4839	4.5009	1.5950	11.9875	90.0948
PEPPER	-1.5842	.2213	2.0269	.2051	1.2477	7.5903
PASTA	-.5557	1.3960	3.3478	.5737	4.0391	28.4388
RICE	-2.1484	-.3054	1.5377	.1167	.7368	4.6537
LETTUCE	.8305	3.1396	5.4487	2.2946	23.0954	232.4614
COLESLAW	-1.5593	.3113	2.1820	.2103	1.3653	8.8643

Instruct LOGISTIC to display the likelihood ratio tests for each possible model by issuing the command:

```
lr
```

LOGISTIC responds by displaying the likelihood ratio tests for each variable in the model:

H0: coeff = 0	lr statistic (1 df)	P-value
HAM	.1032	.7480
EGGS	6.8292	.0090
PEPPER	.0577	.8102
PASTA	2.1203	.1454
RICE	.1063	.7444
LETTUCE	11.4560	.0007
COLESLAW	.1065	.7441

Removing non-significant associations

The EGGS and LETTUCE variables are significant. PEPPER fln is the least significant variable. This variable adds nothing to the model's description of the data and should be removed from the model. Remove the PEPPER fln variable from the model and re-fit the model by issuing the following commands:

```
delete pepper  
estimate
```

LOGISTIC responds with information about the model:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-2.1938	.9270	-2.3667	.0179
HAM	.3864	1.1093	.3483	.7276
EGGS	2.3299	.9421	2.4731	.0134
PASTA	1.4873	.9595	1.5501	.1211
RICE	-.1108	.9089	-.1219	.9029
LETTUCE	3.1548	1.1652	2.7074	.0068
COLESLAW	.4764	.8927	.5337	.5936

95.0-% confidence limits						
	Coefficient		Odds ratio			
	lower limit	upper limit	lower limit	upper limit	lower limit	upper limit
HAM	-1.7879	.3864	2.5606	.1673	1.4716	12.9437
EGGS	.4834	2.3299	4.1763	1.6216	10.2765	65.1258
PASTA	-.3933	1.4873	3.3679	.6748	4.4252	29.0173
RICE	-1.8922	-.1108	1.6705	.1507	.8951	5.3149
LETTUCE	.8710	3.1548	5.4386	2.3892	23.4483	230.1300
COLESLAW	-1.2732	.4764	2.2260	.2799	1.6103	9.2626

Instruct LOGISTIC to display the likelihood ratio tests for each possible model by issuing the command:

```
lr
```

LOGISTIC responds by displaying the likelihood ratio tests for each variable in the model:

H0: coeff = 0	lr statistic (1 df)	P-value
HAM	.1224	.7264
EGGS	7.2523	.0071
PASTA	2.5732	.1087
RICE	.0149	.9028
LETTUCE	12.1079	.0005
COLESLAW	.2871	.5921

The EGGS and LETTUCE variables remain significant. RICE is the least significant variable. This variable adds nothing to the model's description of the data and should be removed from the model.

Removing non-significant associations

Remove the RICE variable from the model and re-fit the model by issuing the following commands:

```
delete rice
estimate
```

LOGISTIC responds with information about the model:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-2.2027	.9256	-2.3797	.0173
HAM	.3967	1.1049	.3590	.7196
EGGS	2.2956	.8958	2.5626	.0104
PASTA	1.4415	.8821	1.6341	.1022
LETTUCE	3.1415	1.1607	2.7066	.0068
COLESLAW	.4761	.8920	.5337	.5935

95.0-% confidence limits						
	Coefficient		Odds ratio			
	lower limit	upper limit	lower limit	upper limit	lower limit	upper limit
HAM	-1.7690	.3967	2.5623	.1705	1.4869	12.9656
EGGS	.5398	2.2956	4.0513	1.7157	9.9301	57.4719
PASTA	-.2874	1.4415	3.1703	.7502	4.2269	23.8156
LETTUCE	.8666	3.1415	5.4164	2.3788	23.1391	225.0742
COLESLAW	-1.2722	.4761	2.2244	.2802	1.6098	9.2479

Instruct LOGISTIC to display the likelihood ratio tests for each possible model by issuing the command:

```
lr
```

LOGISTIC responds by displaying the likelihood ratio tests for each variable in the model:

H0:	coeff = 0	lr statistic (1 df)	P-value
	HAM	.1302	.7182
	EGGS	7.7055	.0055
	PASTA	2.9120	.0879
	LETTUCE	12.2488	.0005
	COLESLAW	.2871	.5921

The EGGS and LETTUCE variables remain significant. HAM is the least significant variable. This variable adds nothing to the model's description of the data and should be removed from the model.

Removing non-significant associations

Remove the HAM variable from the model and re-fit the model by issuing the following commands:

```
delete ham  
estimate
```

LOGISTIC responds with information about the model:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-1.9662	.6141	-3.2019	.0014
EGGS	2.4582	.7903	3.1105	.0019
PASTA	1.4839	.8788	1.6885	.0913
LETTUCE	3.1429	1.1582	2.7136	.0067
COLESLAW	.4404	.8835	.4985	.6181

95.0-% confidence limits						
	Coefficient		Odds ratio			
	lower limit	upper limit	lower limit	upper limit	lower limit	upper limit
EGGS	.9092	2.4582	4.0071	2.4824	11.6832	54.9857
PASTA	-.2386	1.4839	3.2064	.7878	4.4101	24.6895
LETTUCE	.8728	3.1429	5.4130	2.3937	23.1717	224.3107
COLESLAW	-1.2912	.4404	2.1720	.2750	1.5533	8.7755

Instruct LOGISTIC to display the likelihood ratio tests for each possible model by issuing the command:

```
lr
```

LOGISTIC responds by displaying the likelihood ratio tests for each variable in the model:

H0: coeff = 0	lr statistic (1 df)	P-value
EGGS	11.1267	.0009
PASTA	3.1085	.0779
LETTUCE	12.2833	.0005
COLESLAW	.2500	.6171

The EGGS and LETTUCE variables remain significant. Note that PASTA is approaching statistical significance. COLESLAW is the least significant variable. This variable adds nothing to the model's description of the data and should be removed from the model.

Removing non-significant associations

Remove the COLESLAW variable from the model and re-fit the model by issuing the following commands:

```
delete coleslaw
estimate
```

LOGISTIC responds with information about the model:

	Coefficient	Standard Error	Coef/SE	"P value"		
CONSTANT	-1.9710	.6146	-3.2071	.0013		
EGGS	2.6391	.7334	3.5985	.0003		
PASTA	1.6646	.8376	1.9873	.0469		
LETTUCE	3.1956	1.1516	2.7749	.0055		
95.0-% confidence limits						
	Coefficient			Odds ratio		
	lower limit		upper limit	lower limit		upper limit
EGGS	1.2017	2.6391	4.0765	3.3258	14.0008	58.9407
PASTA	.0229	1.6646	3.3062	1.0232	5.2835	27.2824
LETTUCE	.9385	3.1956	5.4527	2.5561	24.4247	233.3914

Instruct LOGISTIC to display the likelihood ratio tests for each possible model by issuing the command:

```
lr
```

LOGISTIC responds by displaying the likelihood ratio tests for each variable in the model:

H0: coeff = 0	lr statistic (1 df)	P-value
EGGS	15.6575	.0001
PASTA	4.4146	.0356
LETTUCE	13.2211	.0003

The EGGS and LETTUCE variables remain significant. PASTA salad is also significantly associated with being a case. Since all variables in the model are significant and contribute to the model's description of the data it would **not** be appropriate to remove any more variables from the model.

Here are the likelihood ratio tests from the original logistic regression model with all seven variables:

H0: coeff = 0	lr statistic (1 df)	P-value
HAM	.1032	.7480
EGGS	6.8292	.0090
PEPPER	.0577	.8102
PASTA	2.1203	.1454
RICE	.1063	.7444
LETTUCE	11.4560	.0007
COLESLAW	.1065	.7441

If we had simultaneously removed all of the variables which appeared to be non-significant, PASTA would have been removed. After the other non-significant variables had been removed from the model, PASTA became significant. This shows that the significance of a variable can be reduced when it is adjusted for non-significant variables. It is best to remove non-significant variables from the model **one at a time** in case any of them later become significant.

Enter, backward elimination, and forward selection

The simplest way of using multiple logistic regression is sometimes called the *enter* method. All variables are entered into the model simultaneously. This technique is liable to make some significant associations appear non-significant. Missing values in any variable in a record (observation) will lead to that record being excluded from the model. The model may then be based on a small set of observations and produce unreliable results.

A better technique, *backward elimination*, has been presented. The process starts with all variables in the model (as in the enter method). Candidate variables for removal from the model are evaluated at each step. Wald's test statistic, the likelihood ratio test, or the odds ratio (with confidence limits) may be used to select variables for removal. A single variable is removed from the model at each step until all variables remaining in the model are significant. This method starts with the same model as the enter method and the same problem with missing values can occur with models early in the elimination process. To overcome this problem it is best to set the threshold significance level for the likelihood ratio test or Wald's test high (e.g. $p=0.10$ or $p=0.15$) for the first few models

A similar technique, *forward selection*, may also be used. With Forward Selection you start off with a model that contains the CONSTANT term and the most significant exposure variable. Each of the other exposure variables are then tested to see if they should be added to the model by fitting a set of test models with each of the remaining exposure variables added and removed in turn. To begin with, the model has the form:

$$\log \text{ odds of disease} = \alpha + \beta_1 x_1$$

and the test models have the form:

$$\log \text{ odds of disease} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

where x_2 is the variable under test. The most significant x_2 variable is then added to the model and a new set of models is fitted with each of the remaining exposure variables added and removed in turn. The model now has the form:

$$\log \text{ odds of disease} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

and the test models now have the form:

$$\log \text{ odds of disease} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where x_3 is the variable under test. The most significant x_3 variable is then added to the model. This process continues until there are no significant variables to be added to the model. The criteria for adding a variable is usually the significance level of the Wald's test statistic or the likelihood ratio statistic. Non-significant variables may also be removed at each step. Forward selection requires the fitting and evaluation of a great many logistic models and offers no significant advantages over the backward elimination technique.

An epidemiological approach

A more epidemiological strategy would be to remove variables on the basis of changes in odds ratios rather than the significance of p-values. Those variables which, when included in the model, change the odds ratios for other foods should be retained in the model. The table below shows the odds ratios obtained using logistic regression to adjust for all other foods in a particular column:

	OR	OR	OR	OR	OR
PEPPER	1.25	REMOVED			
RICE	0.74	0.90	REMOVED		
HAM	1.43	1.47	1.49	REMOVED	
COLESLAW	1.37	1.61	1.61	1.55	REMOVED
PASTA	4.04	4.43	4.23	4.41	5.28
EGGS	11.99	10.28	9.93	11.68	14.00
LETTUCE	23.10	23.45	23.14	23.17	24.42

We could argue that COLESLAW should be retained in the model because it is a *substantial confounder*. The changes in the odds ratios between the model with COLESLAW and the model without COLESLAW are greater than 15% for two of the foods:

	With COLESLAW	Without COLESLAW	Percentage Change
PASTA	4.41	5.28	$(5.28 - 4.41) / 5.28 = 0.16 = 16\%$
EGGS	11.68	14.00	$(14.00 - 11.68) / 14.00 = 0.16 = 16\%$
LETTUCE	23.17	24.42	$(24.42 - 23.17) / 24.42 = 0.05 = 5\%$

This approach gives better estimates of the magnitude and direction of an effect than the p-value approach. It is best suited to the analysis of data from epidemiological studies. For outbreak investigation it is often sufficient to identify the exposure variables which are associated with illness and the p-value approach does this simply and effectively.

Analysing data with more than one level of exposure

The previous examples assume that the exposure variable is binary (exposed / not exposed). This will not always be the case. There may be several values for several different levels of exposure based on extent or duration of exposure.

The GUD / HIV study

Several studies have documented an association between genital ulcer disease (GUD) and HIV infection. A study of Gambian prostitutes documented an association between seropositivity for HIV-2 and antibodies against *Treponema pallidum* (a serological test for syphilis). Prostitutes are not the ideal population for such studies as they may have experienced multiple sexually transmitted infections and it is difficult to quantify the number of times they may have had sex with HIV-2 seropositive customers. A sample of males with sexually transmitted infections is easier to study as they have probably had fewer sexual partners than prostitutes and much less contact with sexually transmitted infection pathogens. Such a sample is also easy to find and collect data from.

The data stored in the file GUDHIV.REC has been adapted from a cross-sectional study of 435 male patients who presented with sexually transmitted infections at an outpatient clinic in The Gambia between August 1988 and June 1990. The variables in the file are:

MARRIED	Married
GAMBIAN	Gambian Citizen
GUD	History of genital ulcer disease (GUD) or syphilis
UTIGC	History of urethral discharge or gonorrhoea
CIR	Circumcised
TRAVOUT	Travelled outside of Gambia and Senegal
SEXPRO	Ever had sex with a prostitute
INJ12M	Injection in previous 12 months
PARTNERS	Number of sexual partners in previous 12 months
HIV	HIV-2 positive serology

Data is available for all 435 patients enrolled in the study.

Using ANALYSIS to produce tables

Start the course software and select Epi-Info ANALYSIS from the main menu. Once ANALYSIS has started issue the command:

```
read gudhiv.rec
```

to retrieve the data for the GUD / HIV Study. Enter the command:

```
set statistics = on
```

to instruct ANALYSIS to produce a full set of statistics for each command. Issue the command:

```
tables * hiv
```

to examine the association between each of the exposure variables and the outcome variable (HIV). Examine the output of this command carefully and complete the following table:

	Lower 95% CL	Relative Risk	Upper 95% CL	X^2	p-value
MARRIED					
GAMBIAN					
GUD					
UTIGC					
CIR					
TRAVOUT					
SEXPRO					
INJ12M					
PARTNERS					

This is a **cross-sectional** study. We are interested in the relative risk. This simple univariate analysis shows associations between several exposure variables (GUD, CIR, and TRAVOUT). A protective effect is associated with CIR. We are interested in this effect because it is *plausible* that circumcision (CIR) may provide some protection against sexually transmitted infections. The situation with the PARTNERS variable is more complicated:

PARTNERS	HIV		Total
	+	-	
1	1	60	61
2	1	128	129
3	2	131	133
4	3	68	71
5	4	21	25
6	3	3	6
7	4	2	6
8	1	1	2
9	2	0	2
Total	21	414	435

because several expected values are less than five and the chi-square statistic is not reliable:

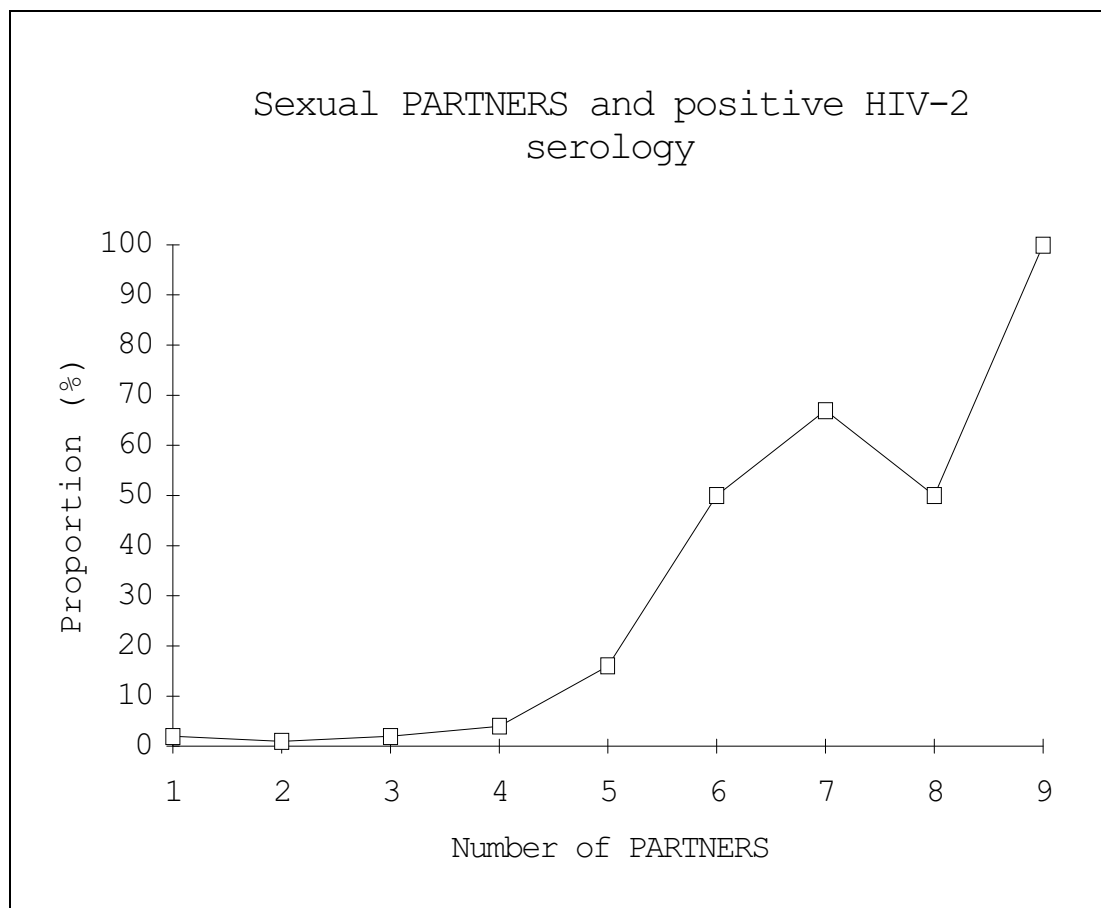
```
An expected value is < 5. Chi square not valid.
Chi square =      140.90
Degrees of freedom =      8
p value = 0.00000000 <---
```

Analysing data with more than one level of exposure

There appears to be an association between the number of sexual PARTNERS in the previous twelve months and positive HIV serology:

PARTNERS	HIV		Total	
	+	-		
1	1	60	61	Proportion HIV = (1 / 61) * 100 = 2%
2	1	128	129	Proportion HIV = (1 / 129) * 100 = 1%
3	2	131	133	Proportion HIV = (2 / 133) * 100 = 2%
4	3	68	71	Proportion HIV = (3 / 71) * 100 = 4%
5	4	21	25	Proportion HIV = (4 / 25) * 100 = 16%
6	3	3	6	Proportion HIV = (3 / 6) * 100 = 50%
7	4	2	6	Proportion HIV = (4 / 6) * 100 = 67%
8	1	1	2	Proportion HIV = (1 / 2) * 100 = 50%
9	2	0	2	Proportion HIV = (2 / 2) * 100 = 100%
Total	21	414	435	

The proportion with positive HIV-2 serology increases as the number of sexual partners increases. It is also difficult to use a chi-square test for trend with this data. This is because the numbers in some of the exposure levels are small and it is difficult to tell whether or not the proportions increase in a linear fashion:



The small numbers in the least exposed group also make it difficult to use this group as the *baseline* or *least-exposed* group.

Logistic regression and data with more than one level of exposure

It is possible to use logistic regression to analyse continuous or ordered data with more than one level of exposure. The form of the logistic model does not change from the basic form:

$$\log \text{ odds of disease} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

where x_1, x_2, x_3 , and x_4 are the variables GUD, CIR, TRAVOUT, and PARTNERS, and $\alpha, \beta_1, \beta_2, \beta_3, \beta_4$ are coefficients to be determined. This model allows us to assess the effect of each variable after adjusting for the effect of the other three variables. The maximum likelihood estimates of β_1 to β_4 are:

$$\begin{aligned}\beta_1 &= \log (\text{odds ratio for GUD adjusted for the other 3 variables}) \\ \beta_2 &= \log (\text{odds ratio for CIR adjusted for the other 3 variables}) \\ \beta_3 &= \log (\text{odds ratio for TRAVOUT adjusted for the other 3 variables}) \\ \beta_4 &= \log (\text{odds ratio for PARTNERS adjusted for the other 3 variables})\end{aligned}$$

Here are the results from this logistic regression carried out with LOGISTIC:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-9.0891	1.9291	-4.7116	.0000
GUD	1.1858	.6439	1.8417	.0655
CIR	-.6789	.9929	-.6837	.4941
TRAVOUT	2.1665	1.0045	2.1567	.0310
PARTNERS	1.2024	.2233	5.3846	.0000

95.0-% confidence limits						
	Coefficient		Odds ratio			
	lower limit	upper limit	lower limit	upper limit	lower limit	upper limit
GUD	-.0761	1.1858	2.4478	.9267	3.2734	11.5628
CIR	-2.6248	-.6789	1.2671	.0725	.5072	3.5506
TRAVOUT	.1977	2.1665	4.1354	1.2186	8.7279	62.5141
PARTNERS	.7648	1.2024	1.6401	2.1485	3.3282	5.1558

PARTNERS is incorporated into the logistic model as a continuous variable. The odds ratio reported for PARTNERS is the odds ratio associated with a *unit increase* in the number of sexual PARTNERS: a man reporting five sexual partners, for example, was over three times as likely (odds ratio = 3.3282) to have a positive HIV-2 serology than a man reporting four sexual partners.

Using logistic regression in this way is similar to testing for a linear trend in proportions and assumes that the proportion of cases at each exposure level increases (or decreases) in a linear fashion. If the data exhibits marked non-linearity you should **not** use logistic regression in this way.

Analysing continuous data as binary data

An alternative approach to dealing with data that is non-linear (or where small numbers make it difficult to tell if a trend is linear) is to collapse the data into a 2-by-2 table for further analysis. Examine the example table again:

PARTNERS	HIV		Total	
	+	-		
1	1	60	61	Proportion HIV = (1 / 61) * 100 = 2%
2	1	128	129	Proportion HIV = (1 / 129) * 100 = 1%
3	2	131	133	Proportion HIV = (2 / 133) * 100 = 2%
4	3	68	71	Proportion HIV = (3 / 71) * 100 = 4%
5	4	21	25	Proportion HIV = (4 / 25) * 100 = 16%
6	3	3	6	Proportion HIV = (3 / 6) * 100 = 50%
7	4	2	6	Proportion HIV = (4 / 6) * 100 = 67%
8	1	1	2	Proportion HIV = (1 / 2) * 100 = 50%
9	2	0	2	Proportion HIV = (2 / 2) * 100 = 100%
Total	21	414	435	

The proportion of patients with positive HIV-2 serology appears to increase quite sharply for those reporting five or more sexual PARTNERS. Collapsing the table around this threshold value results in the following 2-by-2 table:

PARTNERS	HIV		Total
	+	-	
5 or MORE	14	27	41
1 - 4	7	387	394
Total	21	414	435

Entering this table into STATCALC yields the following results:

```

Analysis of Single Table
Odds ratio = 28.67 (9.78 <OR< 86.74)
Cornfield 95% confidence limits for OR
Relative risk = 19.22 (8.23 <RR< 44.89)
Taylor Series 95% confidence limits for RR
Ignore relative risk if case control study.

Chi-Squares      P-values
-----
Uncorrected      :      84.69      0.0000000 <---
Mantel-Haenszel:      84.49      0.0000000 <---
Yates corrected:      77.79      0.0000000 <---
Fisher exact: 1-tailed P-value: 0.0000000 <--
                  2-tailed P-value: 0.0000000 <--

An expected cell value is less than 5.
Fisher exact results recommended.

```

suggesting that patients reporting five or more sexual partners in the previous twelve months were at a significantly higher risk of testing positive for HIV-2 than patients reporting less than five sexual partners in the previous 12 months. Note that the choice of the threshold value (5) was somewhat arbitrary.

Indicator variables

Variables which have more than two levels but are **not** ordered can be analysed using *indicator variables*. Here is a table from a hypothetical study of the diets of children in lone-parent households. The outcome variable is IRON (where + is used to indicate that a diet was considered to be deficient in iron). The exposure variable is ETHNIC which indicates the self-defined ethnicity of the head of household:

ETHNIC	IRON		Total	
	+	-		
Black	37	200	237	Proportion IRON = (37 / 237) * 100 = 16%
Chinese	70	230	300	Proportion IRON = (70 / 300) * 100 = 23%
Indian	31	70	101	Proportion IRON = (31 / 101) * 100 = 31%
White	153	270	423	Proportion IRON = (153 / 423) * 100 = 36%
Total	291	770	1061	

To analyse this sort of data with logistic regression you must first create a set of *indicator variables*. An indicator variable indicates whether a subject belongs to one group (category of a variable) or another. The number of indicator variables required to represent a categorical variable is one less than the number of categories:

ETHNIC	Indicator Variables		
	ETHNIC01	ETHNIC02	ETHNIC03
Chinese	1	0	0
Indian	0	1	0
White	0	0	1

The category **Black** is represented by a zero in each of the indicator variables:

ETHNIC	Indicator Variables		
	ETHNIC01	ETHNIC02	ETHNIC03
Black	0	0	0
Chinese	1	0	0
Indian	0	1	0
White	0	0	1

The category **Black** is the *baseline* category. The logistic model for this data would be:

$$\log \text{ odds of iron deficient diet} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where x_1 , x_2 , and x_3 are the indicator variables (ETHNIC01, ETHNIC02, and ETHNIC03) and α , β_1 , β_2 , and β_3 are coefficients to be determined. The maximum likelihood estimates of β_1 , β_2 , and β_3 are:

$$\begin{aligned} \beta_1 &= \log (\text{odds ratio for ETHNIC01}) <- \text{Chinese (relative to Black)} \\ \beta_2 &= \log (\text{odds ratio for ETHNIC02}) <- \text{Indian (relative to Black)} \\ \beta_3 &= \log (\text{odds ratio for ETHNIC03}) <- \text{White (relative to Black)} \end{aligned}$$

The odds ratios reported for each of the indicator variables (ETHNIC01, ETHNIC02, and ETHNIC03) will be the odds ratio of each group relative to the **Black** group.

Indicator variables

Indicator variables allow you to analyse non-ordered categorical data using logistic regression. They can also be used to analyse non-linear ordered or continuous data. The following exercise illustrates how to create and analyse indicator variable sets using data from the GUD/HIV study. Start Epi-Info ANALYSIS and issue the following commands to create a set of indicator variables from the PARTNERS variable in the GUDHIV dataset:

```
read gudhiv.rec
part3 = 0
part6 = 0
if partners >= 3 and partners <= 5 then part3 = 1
if partners >= 6 then part6 = 1
```

These commands set up two indicator variables for different levels of the PARTNERS variable:

PARTNERS	PART3	PART6
1	0	0
2	0	0
3	1	0
4	1	0
5	1	0
6	0	1
7	0	1
8	0	1
9	0	1

The *baseline* category (one or two partners) is represented by a zero in each of the indicator variables. You can check this and examine the values of the indicator variables with the following commands:

```
tables partners part3
tables partners part6
```

The indicator variables should be saved in a **new** data file. Issue the commands:

```
erase gudnew.rec
route gudnew.rec
write recfile
```

to do this and then quit ANALYSIS.

Indicator variables

Start LOGISTIC and issue the following commands to read in the new dataset, define the logistic model, and fit the logistic model:

```
read gudnew
model hiv = constant + part3 + part6 + gud + cir + travout
estimate
```

Examine the output of these commands carefully:

	Coefficient	Standard Error	Coef/SE	"P value"
CONSTANT	-6.5469	1.6855	-3.8842	.0001
PART3	2.0492	1.0688	1.9172	.0552
PART6	6.0631	1.2218	4.9623	.0000
GUD	1.3260	.6164	2.1513	.0315
CIR	-.9278	1.0901	-.8511	.3947
TRAVOUT	2.2109	.9730	2.2723	.0231

95.0-% confidence limits						
	Coefficient			Odds ratio		
	lower limit		upper limit	lower limit		upper limit
PART3	-.0457	2.0492	4.1440	.9554	7.7615	63.0563
PART6	3.6684	6.0631	8.4579	39.1876	429.7115	4712.0029
GUD	.1179	1.3260	2.5341	1.1252	3.7660	12.6049
CIR	-3.0644	-.9278	1.2088	.0467	.3954	3.3495
TRAVOUT	.3039	2.2109	4.1178	1.3551	9.1236	61.4265

The odds ratios reported for variables PART3 and PART6 are for each group relative to the *baseline* group of one or two partners. The confidence intervals for the two groups are wide because there are very few cases with positive HIV-2 serology in the dataset. You might try experimenting with different groups derived from the PARTNERS variable.

The variables CIR and PART3 are candidates for removal from the model using the backward elimination technique. Use the backward elimination technique to remove non-significant variables from the model.

A strategy of analysis

Assessing the independent effect of many variables on the occurrence of a disease involves fitting many combinations of variables and interaction terms. For a large number of explanatory variables this would be very time consuming. Here is a strategy of analysis using the course software which you may find useful:

Examine the data using 2-by-2 tables

This should always be done before the multivariate analysis. It is important that the relationship between all potential risk factors and the explanatory variables is studied using 2-by-2 tables. Study the crude odds ratio to measure the magnitude of any effect and its 95% confidence interval or p-value to assess its significance.

Select the variables to be included in the multivariate analysis

Only a selection of the many potential explanatory variables would normally be included in the multivariate analysis. Variables which are not statistically significant (or are not potential confounders) are first eliminated from further analysis. The statistical significance of each variable can be assessed using the likelihood ratio test from fitting a logistic model with that variable alone or using the chi-square test from the 2-by-2 table. It is best to use a p-value of 0.10 as the cut off for statistical significance to ensure that potentially important variables are not excluded from further analysis.

Perform the multivariate analysis

Many statisticians use the backward elimination procedure. A logistic model containing all the variables to be included in the multivariate analysis is fitted using the course software and variables which are not statistically significant are removed from the model one at a time. The likelihood ratio test is used to assess the significance of each variable after adjusting for the effects of the other variables in the model. A more epidemiological strategy would be to remove variables on the basis of changes in odds ratios rather than the significance of p-values.

Check the final model

Including interaction terms in logistic models rapidly increases the number of coefficients to be estimated and makes interpretation of the model complicated. Interaction terms are usually only fitted for interactions which are known or suspected to exist. It is also worth checking if there are any interactions between any pairs of variables in the final model by fitting the *pairwise* interaction terms one at a time. If any interaction terms are significant, then separate models should be fitted for different strata of one of the interaction variables.

Stratified analysis

Stratified analysis using Mantel-Haenszel methods is a useful alternative to logistic regression if your study has a small number of explanatory variables. Stratified analysis can also be used in conjunction with logistic regression to examine interactions or associations in detail. It may prove useful to use stratified analysis as part of the selection process so that associations which are no longer statistically significant after adjusting for one particular variable are not included in the multivariate analysis.

Presentation of results

Once you have completed the data analysis, you will have accumulated a lot of statistical results. As well as the number of cases and controls who ate each food, you will have obtained crude odds ratios for each food and adjusted odds ratios for those foods which were independently associated with food-poisoning. There are also confidence intervals associated with each crude and adjusted odds ratio and various significance tests for the crude odds ratios (chi-square test) and adjusted odds ratios (likelihood ratio test and Wald's test).

The main results of the SALEX study, for example, are best presented in tabular form. You might only show the results for the three foods which were independently associated with food-poisoning:

Table 1: Number (%) of cases and controls who had eaten each food associated with food-poisoning and the odds of being a case for each of these foods.

Food	Cases (%)	Controls (%)	Crude Odds Ratio	Adjusted [†] Odds Ratio (95% CI)
EGGS				
No	10 (20)	20 (77)	1	1
Yes	40 (80)	6 (23)	13.3	14.0 (3.3 - 58.9) p < 0.001 [‡]
PASTA				
No	26 (53)	23 (88)	1	1
Yes	25 (49)	3 (12)	7.4	5.3 (1.0 - 27.3) p = 0.036 [‡]
LETTUCE				
No	23 (45)	25 (96)	1	1
Yes	28 (55)	1 (4)	30.4	24.4 (2.5 - 233.4) p < 0.001 [‡]

[†] odds ratio adjusted for eggs, pasta salad, and lettuce

[‡] p-values refer to the likelihood ratio test

It is often useful to present non-significant results as well as significant results. Readers might be interested to see which foods were not associated with food-poisoning and whether any associations observed originally were due to confounding. In this case, it would be appropriate to present a table showing results for all the foods investigated.

For a cohort or cross-sectional study, it would not be appropriate to look at the distribution of each food for the cases and controls. Instead, you would present the number (and percentage) who were ill for those who did and did not eat each food.