Analysing Data

A practical primer using EpiInfo

Mark Myatt & Sue Ritter

First published in 1996 by Brixton Books, Station Building, Llanidloes, Powys SY18 6EB

This edition published 1999 in partnership with Truman State University

A CIP catalogue record for this book is available from the British Library

ISBN 1-873937-46-6

‘There is one outstandingly important fact regarding Spaceship Earth, and that is that no instruction book came with it.’

R. Buckminster Fuller, Operating Manual for Spaceship Earth

Contents

Introduction 5
Categorical and continuous variables 9
Files, cases, and variables 10
Retrieving and examining a dataset in ANALYSYS 11
Working with subsets of data 13
Producing charts 16
Chart types and their uses 20
Summarising distributions
- Frequency distributions 20
- Categorical and continuous variables 22
- The seven figure summary 23
- How the standard deviation measures variability 25
- Calculating the standard deviation - worked example 28
- Standard deviation and degrees of freedom 27
- Problems with the mean and standard deviation 28
- Calculating summary measures in ANALYSIS 29
- Summarising distributions with charts 30

Creating and recoding variables 32
Two-by-two tables
- Naming of parts 35
- Absolute risk 36
- Exposure specific risk 37
- Relative risk 38
- Odds ratio 39
- Measures of risk (summary) 40
- Confidence intervals 42
- Testing a hypothesis about association 43
- Chi-square, degrees of freedom, and p-values 45
- Confidence intervals and significance tests 47
- An example of ANALYSIS output 48

Contingency tables 49
Tables with subsets of data 54
Differences in means
- Categorical and continuous data 54
- Analysis of variance 55
- A worked example of ANOVA calculations 56
- An example of ANALYSIS output 57
- What ANOVA assumes about your data 59
- Testing the ANOVA assumptions 60
- When ANOVA is unsuitable 62
- Transforming the data 63
- Checking data for skew and unequal variance 64
- Applying a logarithmic transformation 65
- The antilogarithm of the mean of the logarithms (geometric mean) 66
- Difficult distributions 67
- Another problem with ANOVA / Non-parametric statistics 68

Two continuous variables
- Correlation and regression analysis 72
- Using scatter plots to examine associations 73
- Assessing the strength of an association 74
- The correlation coefficient and non-linear associations 75
- Transformation and non-linear associations 76
- Scatter plots and correlation coefficients 78
- Interpreting the correlation coefficient 79
- The regression line 80
- Linear regression in ANALYSIS 82
- What the regression line does not explain 83

Solution to practical exercises 86
Statistical tables 97

Introduction

How to use this book

This is a self-contained course in the basic techniques of statistics and data analysis. Whilst much theoretical material is introduced, extensive use is made of computer based practical exercises. In this way you will learn the basic techniques of statistics and data analysis whilst actually doing statistics and data analysis yourself. All techniques are introduced in the context of analysing a real-world study into stress levels amongst student groups. This highly practical bias means that you should be seated at your personal computer as you work through this book.

Typographical Conventions

Several typefaces are used throughout this book. The following table illustrates their uses:

Typeface	Use	Example
body text	Main text.	examine the residuals
Title	Page and section headings.	**Distributions**
VARIABLE	Variables referred to in the text.	outcome is CHQCASE
COMMAND	ANALYSIS commands referred to in the text.	the REGRESS command
MODULE	Modules of the course software.	the ANALYSIS module
emphasis	**Bold** letters are used for emphasis.	**not** normally distributed
term	Important technical terms are *italicised*.	*positively skewed*
`output`	Output from the course software.	`p-value = 0.758942`
`calculations`	Calculations and worked examples.	`2 - 1 = 1`
`commands`	Commands - type exactly as shown.	**`scatter mee resid`**
KEYS	Keystrokes	press ENTER to start
Formula	Statistical formulae	$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$

The course software

The course software

The course software and example dataset are supplied on the floppy disk. The course software is the ANALYSIS module of a public domain database and statistics package called EpiInfo. The full version of EpiInfo including data-entry, data-checking, data-management, statistics, and sample size modules is available from Brixton Books.

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:

```
analysis
```

at the MSDOS prompt and press ENTER to start the course software (if your are using Windows or OS/2 then you will need to start an 'MSDOS Prompt' session in order to use the course software). You may also copy the course software and example dataset (all files on the floppy disk) to a directory on your hard disk and run the software from there.

The course software is designed to run on any 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 a PC emulator such as SoftPC or SoftWindows on a Macintosh, PowerPC, or UNIX machine.

ANALYSIS can read both EpiInfo and dBase format files.

Please acknowledge EpiInfo in any manuscript that makes use of its calculations. A suitable reference is:

> Dean AG, Dean JA, Coulombier D, Brendel KA, Smith DC, Burton AH, Dicker RC, Sullivan K, Fagan RF, Arner TG (1994), *EpiInfo: A word processing, database and statistics program for epidemiology on microcomputers*, Centres for Disease Control and Prevention, Atlanta, Georgia, USA.

The example dataset

Introduction to the study

Several studies have investigated the issue of stress in the caring professions. Much of this work has been focused on general nurses and junior hospital doctors. Few quantitative studies have focused on social workers. The data presented here was collected as part of a study of social work students.

Three groups of students were surveyed:

1. Fifty (50) social work students studying for a diploma in social work.

2. Sixty-nine (69) undergraduate psychology students studying for a university degree course.

3. Thirty-three (33) graduate trainees following a teacher training (PGCE) course.

All students were in their first year of study. The PGCE students were all graduates. Study subjects were approached via their course tutors and asked to complete four separate questionnaires:

1. **Demographic Questionnaire:** A short five item form asking about sex, age, marital status, living situation, and ethnic origin.

2. **General Health Questionnaire (GHQ-28):** A twenty-eight item self-administered screening questionnaire developed by Goldberg (Goldberg D & Williams P, *A User's Guide to the General Health Questionnaire*, NFER-Nelson, Windsor, 1988). It is aimed at detecting psychological ill-health in people in non-psychiatric settings, including community, primary care, and general hospital units.

3. **Rosenberg Self-Esteem Scale:** A widely used questionnaire used to assess levels of self-esteem that was first developed for use with adolescents (Rosenberg M, *Society and the Adolescent Self-Concept*, Princeton University Press, Princeton NJ, 1965). The ten item version used in this study was developed by Wycherley (Wycherley B, *The Living Skills Pack*, South East Thames Regional Health Authority, Bexhill-on-Sea, 1987).

4. **Maslach Burnout Inventory:** This is the name popularly given to Maslach and Jackson's twenty-two item questionnaire used for the *Human Services Survey* (Maslach C & Jackson S, *Maslach Burnout Inventory*, Consulting Psychologists Press, Palo Alto CA, 1986). Its purpose is to measure the feelings of workers in the caring professions about their jobs, colleagues, and users of their services. There are three subscales: *Emotional Exhaustion*, *Depersonalisation*, and *Personal Accomplishment*. This questionnaire was not given to the psychology students because they were not in contact with clients or pupils on a regular basis.

Objectives of the study

The objectives of the study were:

1. To determine whether students undertaking professional training exhibited higher stress scores than undergraduate students.

2. To determine whether students undertaking professional training in social work exhibited higher stress levels than those undertaking postgraduate teacher training.

Further details of the study can be found in:

> Tobin PJ & Carson J, *Stress and the Student Social Worker*, Social Work & Social Sciences Review, Vol 5(3), pp 246-255, 1994.

Variable names, types, and ranges

The structure of the SSSW.REC dataset

Data from this study is stored in an EpiInfo data file called SSSW.REC. The following table lists the *names*, *types*, and *ranges* of variables in the SSSW.REC dataset. The variable types and the EpiInfo *pictures* that would be used to define variables in .QES files or with the DEFINE command in ANALYSIS are also listed:

Variable	Contents	Type	Picture	Values
SUBJECT	Subject ID number	Number	<idnum>	1 thru 152
COHORT	Study group	Character	<AAAAAAAAAAA>	ACADEMIC PGCE SOCIAL WORK
SEX	Sex	Number	#	1 = Male 2 = Female
AGE	Age	Number	#	1 = 25 and under 2 = Over 25
MARITAL	Marital status	Number	#	1 = Married 2 = Single 3 = Divorced 4 = Separated 5 = Cohabiting 6 = Widowed
LIVING	Living with ...	Number	#	1 = Alone 2 = Parents or siblings 3 = Partner 4 = Partner and children 5 = Children 6 = Friends or colleagues
ETHNIC	Ethnic origin	Number	##	1 = African 2 = West-Indian 3 = Indian 4 = Pakistani 5 = Bangladeshi 6 = East African Asian 7 = Chinese 8 = Cypriot 9 = Black European 10 = White European 11 = Other
GHQ	GHQ-28	Number	##	Score (0-99)
ROSENBERG	Self-esteem	Number	##	Score (0-99)
MEE	Emotional-exhaustion	Number	##	Score (0-99)
MDP	Depersonalisation	Number	##	Score (0-99)
MPA	Accomplishment	Number	##	Score (0-99)

This table is reproduced on the inside back cover of this book.

Categorical and continuous variable

The two broad groups

Variables can be split into two broad groups. These are:

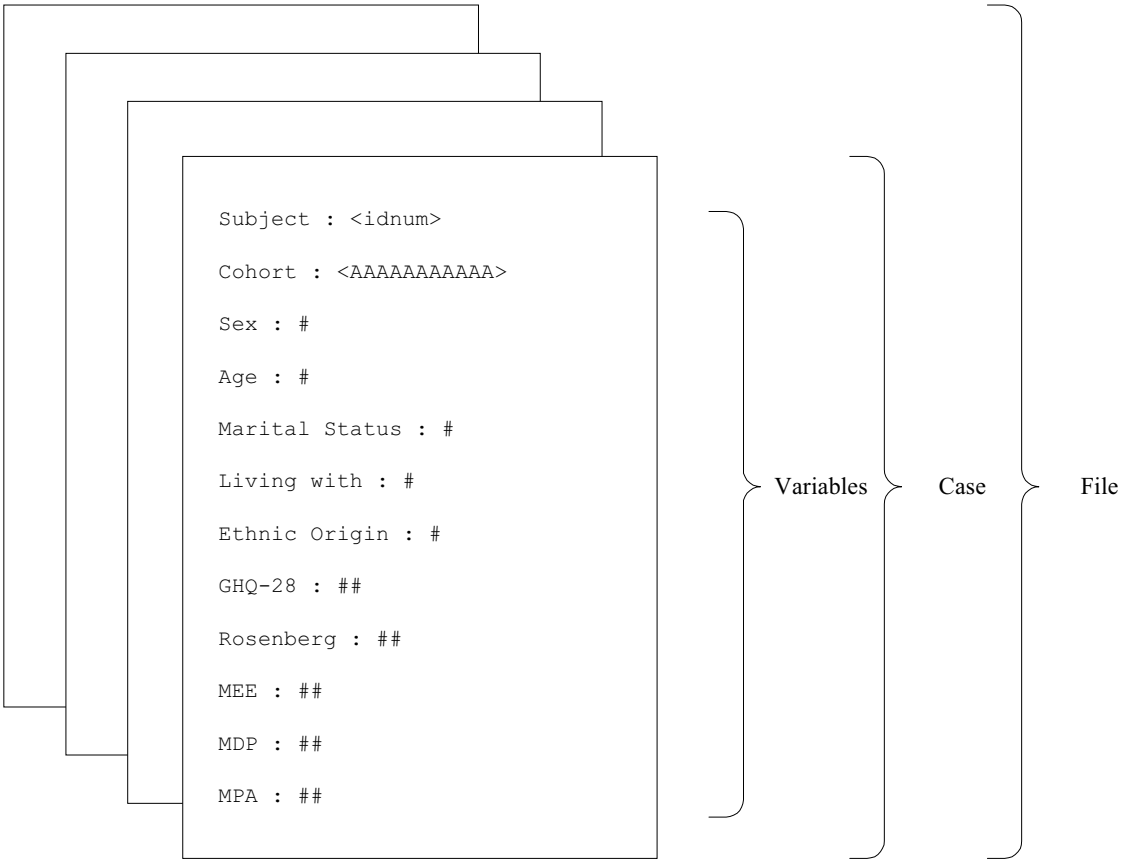
> Variables that hold codes that indicate membership of a defined category or group. The variables COHORT, SEX, AGE, MARITAL, LIVING, and ETHNIC are examples of *categorical variables* in the **Stress and the Student Social Worker** dataset. A particularly common type of categorical variable is a *binary categorical variable* where cases can belong to one of two alternative groups. The variables SEX and AGE are examples of binary categorical variables in the **Stress and the Student Social Worker** dataset. Categorical variables can be *ordered* (as with AGE) or *non-ordered* (as with COHORT, SEX, MARITAL, LIVING, and ETHNIC).
>
> Variables hold data that is measured and stored on an ordered scale. *Discrete variables* are limited to whole numbers. This includes data such as scores and counts The variables GHQ, ROSENBERG, MEE, MDP, and MPA are examples of discrete variables in the **Stress and the Student Social Worker dataset** *Continuous variables* hold data that is measured on a fully continuous scale that is not constrained to whole numbers. Data such as height and weight would be continuous variables.

It is important that you appreciate the distinction between these two broad groups of variable as different techniques are used to describe and analyse them.

Files, cases, and variables

Files, cases, records, variables, and fields

A set of data (*dataset*) is stored in the computer as a *file*. A file consists of a collection of *cases* (or *records*), one for each individual. Each case contains the data on that individual in a series of *variables* (or *fields*):



In our example, the cases would be individuals interviewed and the variables would be the answers to the questions asked (or scores calculated from those answers).

Data is usually represented in a table in which each row represents an individual case or record and each column represents a variable or field. Here are the first three cases in the Stress and the Student Social Worker study:

Subject	Cohort	Sex	Age	Marital	Living	Ethnic	GHQ	Rosenberg	MEE	MDP	MPA
1	SOCIAL WORK	2	2	2	5	2	2	15	14	0	2
2	SOCIAL WORK	2	2	1	4	2	17	13	12	6	3
3	SOCIAL WORK	2	2	1	1	11	21	26	29	2	3

Retrieving a dataset in ANALYSIS

Retrieving a dataset

ANALYSIS is the data analysis module of EpiInfo. With ANALYSIS you can use simple commands to produce lists, frequencies, statistics and graphs. Type

```
analysis
```

at the MSDOS prompt and press ENTER to start ANALYSIS (see page 8).

Once ANALYSIS starts the screen is divided into several distinct sections. The upper section is headed OUTPUT and the lower section is headed COMMANDS. At the top of the screen are two lines giving the status information:

```
Dataset:  <None>                                          Free memory: 250K
Use READ to choose a dataset
```

The status information shows the name of the current data file (dataset) and the amount of free memory in the system (this will depend on how your computer has been set up and is likely to differ from system to system). The bottom line of the screen shows the function keys available from within ANALYSIS:

Key	Function
F1	Help
F2	Menu of commands
F3	Menu of variables in the current dataset
F4	Browse the current dataset in 'spreadsheet' format
F5	Printer on / Printer off
F9	Temporary DOS session (type EXIT to get back to ANALYSIS)
F10	Quit ANALYSIS

We need to tell ANALYSIS the name of the dataset we want to work with. The status information reminds you to do this:

```
Dataset:  <None>                                          Free memory: 250K
Use READ to choose a dataset
```

The data for the **Stress and the Student Social Worker** study is stored in an EpiInfo file called SSSW.REC. Type:

```
read sssw.rec
```

and press ENTER. The name of the file and the number of records will appear at the top of the screen, indicating that the file has been read:

```
Dataset:  A:\SSSW.REC (152 records)                       Free memory: 250K
Criteria: All records selected
```

Note that you need to press the ENTER key to instruct ANALYSIS to act on any of the commands you type. You can edit the command line before you press ENTER. The command is executed only after you press the ENTER key. You can clear (i.e. delete) the command line by pressing ESC. ANALYSIS maintains a list of previously entered commands which you can scroll through using the ↑ and ↓ keys.

Examining a dataset

The BROWSE command

Look at the menu bar at the bottom of the screen showing the function keys available within ANALYSIS. Press the F4 key to browse through the data. The first twenty-one records appear on the screen, one record per line, with variables running across the line. Move around the file using the ↑, →, ↓, ←, PG UP, PG DN, END and HOME keys. Their uses are:

Keys	Action
PG UP, PG DN	move up & down through the records (rows)
HOME, END	move to the top & bottom records (rows)
↑, →, ↓, ←	move between variables (columns) or records (rows)
CTRL + ←, CTRL + →	move to the first & last variables (columns)

Press F4 again to select full screen mode. You now have a single record displayed above the menu bar. Look at the menu bar and work out how to move to the next record and back to the previous one. Try this a few times. Press F4 again to return to 'spreadsheet' mode. When you have finished looking at the data press F10 to return to the main ANALYSIS screen. Note that the word BROWSE appeared in the COMMANDS window, as a result of pressing F4. We could have typed the command BROWSE instead of pressing the F4 key.

The LIST command

If we want to see the variables in a dataset and examine them variable-by-variable and record-by-record we would use the BROWSE command. If we want to see data as a list we would use the LIST command. Type:

```
list subject cohort sex age marital living ethnic
```

The SUBJECT, COHORT, SEX, AGE, MARITAL, LIVING, and ETHNIC variables (together with the internal record number) for all cases in the dataset are listed on the output screen. Since it is impossible to view all 152 cases on the screen at the same time, they are listed a page at a time (each page showing fourteen cases). The word <more> at the bottom of the output screen means there are more pages of output to come. Press any key to see the next page of output. If you do not want to see any more output, press ESC. When <more> no longer appears at the bottom of the output screen you are at the bottom of the listing. You can scroll back through the listing using the PG UP and PG DN keys. You can also use CTRL with the PG UP and PG DN keys to scroll up and down the output screen one line at a time.

BROWSE and LIST commands

You can use the BROWSE and LIST commands on their own to view the entire dataset (i.e. all variables for all cases) or you can specify a list of variables to view. The command:

```
list mee mdp mpa
```

and the command:

```
browse mee mdp mpa
```

both allow you to examine the contents of the MEE, MDP, and MPA variables. The BROWSE command is more flexible ('spreadsheet' and full-screen modes, easy scrolling between variables and cases, output not limited to the width of the screen) but sends output to the screen **only**. If you want to list variables and cases to a printer or file then you should use the LIST command.

Subsets of data

The SELECT command

If we wanted to work with data for a *subset* of cases we would use the SELECT command to select only those cases we wanted to work with. Type:

```
select sex = 2
```

Any subsequent commands we type will be performed **only** with the data for females (SEX = 2). Note that the status information at the top of the screen changes to reflect the SELECTion *criteria*:

```
Dataset:  A:\SSSW.REC (152 records)                          Free memory: 250K
Criteria: sex = 2
```

Type the command:

```
list subject age sex marital
```

If we page through the output screen, we can see that only females are listed. Now suppose we want to work with the male subjects (SEX = 1). If we type:

```
select sex = 1
```

No records will be SELECTed (check this using the LIST or BROWSE commands). This is because there can be **no** cases where SEX = 2 **and** SEX = 1 **at the same time:**

```
Dataset:  A:\SSSW.REC (152 records)                          Free memory: 250K
Criteria: (sex = 2) AND (sex = 1)
```

What we must do is cancel the previous SELECT statement, so that all records are SELECTed again. Type:

```
select
```

The SELECT command on its own instructs ANALYSIS to work with all the cases in a dataset. Note that the SELECTion criteria at the top of the screen now shows that all records are selected:

```
Dataset:  A:\ SSSW.REC (152 records)                         Free memory: 250K
Criteria: All records selected
```

Issue another SELECT command to SELECT the males:

```
select sex = 1
```

Examine the SELECTion criteria at the top of the screen. Use the BROWSE or LIST command and scroll through the listing to check that this new SELECT command has worked as expected.

Issue the SELECT command on its own:

```
select
```

to clear the current SELECTion criteria and continue working with all records in the dataset.

Subsets of data

More complex selections

The expression (e.g. 'SEX = 1') used with the SELECT command can be a complex *expression* using a combination of *inequality operators*:

Operator	Meaning	Example
=	equal to	`sex = 1`
<>	not equal to	`sex <> 1`
<	less than	`ghq < 5`
>	greater than	`ghq > 5`
<=	less than or equal to	`ghq <= 5`
>=	greater than or equal to	`ghq >= 5`

and *Boolean* operators:

Operator	Meaning	Example
`and`	both expressions must be true	`sex = 1 and ghq < 5`
`or`	either of the expressions may be true	`sex = 1 or ghq < 5`
`not`	the expression must not be true	`sex = 1 and not ghq < 5`

Boolean operators are symbols used to express logic and to test the *validity* (*truth*) of a *logical expression* in an algebraic way. Boolean operators always separate expressions that can only be *true* or *false* (e.g. 'Is the GHQ-28 score less than five?') and always return true or false values. Here are the *truth tables* for the AND and OR Boolean operators:

AND		
First Expression	**Second Expression**	**Resulting Expression**
`false`	`false`	`false`
`true`	`false`	`false`
`false`	`true`	`false`
`true`	`true`	`true`

OR		
First Expression	**Second Expression**	**Resulting Expression**
`false`	`false`	`false`
`true`	`false`	`true`
`false`	`true`	`true`
`true`	`true`	`true`

The NOT operator reverses the resulting expression (i.e. true becomes false and false becomes true).

Parentheses can be used to create complex expressions or to group expressions. Examples are:

Example	Selected Cases
`sex = 1`	Males
`sex = 1 and cohort = "PGCE"`	Male PGCE students
`sex = 1 and cohort = "PGCE" and ethnic = 1`	Male African PGCE students
`sex = 1 or cohort = "PGCE"`	Males and PGCE students
`sex = 1 or (cohort = "PGCE" and ethnic = 1)`	Males and African PGCE students
`sex = 2`	Females
`sex = 2 and ethnic = 3`	Indian females
`sex = 2 and ethnic = 3 and age = 2`	Indian females over 25
`sex = 2 or ethnic = 3`	Females and Indians
`sex = 2 or (ethnic = 3 and age = 2)`	Females and Indians over 25

The *Boolean* operators may appear to work against common sense but it is not difficult to build expressions using them as long as you remember that AND **excludes** cases (**all** expressions must be true) whereas OR **includes** them (only **one** expression need be true).

Subsets of data

When to use quotes with SELECT

You may have noticed that some of the SELECTion expressions in the examples had selection values enclosed in double-quote characters (") whilst others did not. The use of double-quote characters depends on the variable type. You should use double-quotes with character variables but **not** with number variables. You can list variable names, types and lengths using the command:

```
variables
```

Number variables are marked `Integer` or `Real`. Character variables are marked `ALPHA` or `string`. If you use double-quote characters with a number variable the SELECT command will function correctly **only** with the *equality* (=) operator. Note that single inverted commas (e.g. COHORT = 'PGCE') will **not** work with either type of variable.

Entering commands in ANALYSIS

There are two methods of issuing commands to the ANALYSIS program. 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 arrow 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 arrow 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 you need 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. LIST) 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 PG DN will display the next screen of information.

Press ESC to return to ANALYSIS.

Leaving ANALYSIS

To leave ANALYSIS press F10 or type the command:

```
quit
```

and press ENTER. This will return you to the MSDOS prompt.

Producing charts

The shape of data

Charts are an ideal way of looking at how data in a variable is *distributed*. The *distribution* of a variable can be thought of as the shape of the data in that variable. Using charts we can see the shape of the data and answer the following types of questions about a variable:

Are the data values similar to each other?

Are the data values different from each other?

How different are the data values from each other?

Is there a single group of data values?

Are there several groups or clusters of data values?

Are a few data values very different from the majority of data values?

Questions like these cannot easily be answered by looking at a list of the data values in a variable. Charts allow us to see important features in the data. A chart is a picture of the data that allows us to see the overall structure of data rather than the confusing detail of individual data values.

Producing charts

ANALYSIS chart types

You can produce five different types of chart with ANALYSIS:

ANALYSIS Command	Chart Type
pie *variable*	Pie chart of the specified variable
bar *variable*	Bar chart of the specified variable
histogram *variable*	Histogram of the specified variable
line *variable-1 [variable-2]*	Line plot of variable-1 [by levels of variable-2]
scatter *x-variable y-variable [/r]*	scatter plot of two variables [with regression line]

Start ANALYSIS and retrieve the **Stress and the Student Social Worker** dataset by typing:

```
read sssw.rec
```

Suppose we want to examine the distribution of the ETHNIC (ethnic origin) variable. We could produce a BAR chart by typing:

```
bar ethnic
```

After examining the chart, we can remove it from the screen by pressing ESC (the graph does not appear in the output screen). You can produce a HISTOGRAM of the MDP (depersonalisation) variable using the HISTOGRAM command:

```
histogram mdp
```

You can produce a PIE chart of the MARITAL (marital status) variable using the PIE command:

```
pie marital
```

or a LINE plot (*frequency polygon*) of the MDP (depersonalisation) variable using the LINE command:

```
line mdp
```

You can also use the LINE command to examine the difference between groups. The command:

```
line mdp sex
```

shows separate lines for males and females. The labels for the lines (1 and 2) are taken from the values of the SEX variable (i.e. 1 = MALES, 2 = FEMALES).

To produce a SCATTER plot of two variables, showing how one varies with the other, use the SCATTER command which takes the general form:

```
scatter x-variable y-variable
```

The first variable listed (*x-variable*) will be plotted along the horizontal (*x*) axis and the second variable listed (*y-variable*) will be plotted along the vertical (*y*) axis as in:

```
scatter ghq mee
```

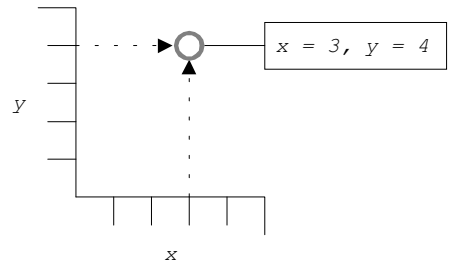
If you add '/r' to the end of a SCATTER command it will display a *regression* line showing the *linear* (straight line) relationship between the two variables:

```
scatter ghq mee /r
```

Chart types and their uses

Analysis chart types and their uses

Each type of chart has a different use. The table below summarises their use:

Chart	Examples	Description
PIE	**pie cohort** **pie living**	Used to display the *proportion* (percentage) of cases for each value of a categorical variable. Segments of the pie represent the proportion of cases with different values of the variable. Pie charts should be used for non-ordered data (e.g. sex). They are not appropriate for ordered data (e.g. age groups) and can be difficult to read when used to chart variables with more than five or six values (categories). Bar charts are preferred to pie charts when used to chart ordered data or non-ordered data with more than five or six categories.
BAR	**bar age** **bar living** **bar ethnic**	Used to display the count of cases in each category of a categorical variable. Vertical bars represent counts of cases with different values of the variable. Bar charts differ from histograms in having bars separated by spaces and in omitting counts for values (categories) with a count of zero.
HISTOGRAM	**histogram ghq** **histogram mee**	Used to display the distribution of values in a continuous variable. Vertical bars represent counts of records with different values of the variable. Histograms differ from bar graphs in having the bars contiguous (not separated by spaces) and in displaying counts for values with a count of zero.
LINE	**line ghq** **line ghq sex**	Used to display the distribution of values in a continuous variable. Points on the line represent counts of cases with different values of the variable. More than one line may be charted by giving the name of an additional (categorical) *stratifying* variable. Charts with multiple lines can be difficult to read when the stratifying variable has more than three or four categories (i.e. when the chart displays more than three or four lines). Line charts are often used to show trends or movement in the value of a variable over time (*time-series*).
SCATTER	**scatter mee ghq**	Used to investigate and display the relationship between two continuous variables. Points on the scatter plot represent pairs of values (i.e. the values for each of the variables within the same case). The first value (*x*) defines the horizontal position of the point. The second value (*y*) defines the vertical position of the point. For example, the value pair (*3,4*) would be represented as:  A non-random appearance to the plot indicates that there is an *association* between the two variables.

Practical exercises

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

Exercise 1 - Examining a dataset

How many cases are there in the dataset stored in SSSW.REC?

SELECT only those persons who are LIVING alone. Use BROWSE or LIST to see how many records are selected? How many records are selected in each of the ACADEMIC, PGCE, and SOCIAL WORK groups?

Clear the SELECTion so that you can continue to work with all of the cases.

SELECT only those persons who are **not** White Europeans. Use BROWSE or LIST to see how many records are selected? How many records are selected in each of the ACADEMIC, PGCE, and SOCIAL WORK groups?

Clear the SELECTion so that you can continue to work with all of the cases.

SELECT only those persons who are both West-Indian and female. Use BROWSE or LIST to see how many records are selected? How many records are selected in each of the ACADEMIC, PGCE, and SOCIAL WORK groups?

Clear the SELECTion so that you can continue to work with all of the cases.

Exercise 2 - Producing charts

Produce a BAR chart of MARITAL status. Which is the most common marital status? Are there groups with less than 10 persons - if so, which ones?

Produce a PIE chart of AGE. What percentage of the study population was over 25 years old?

Produce a HISTOGRAM, PIE chart, and LINE plot of the GHQ-28 (GHQ) score. Which chart do you think is the most appropriate and why? What was the highest and lowest GHQ-28 score recorded for any person?

Produce a SCATTER plot of the GHQ-28 (GHQ) variable and the emotional exhaustion (MEE) variable. Produce the same chart again but with a regression line.

Summarising distributions

Frequency distributions

The *frequency distribution*, *percentage frequency distribution* and *cumulative percentage frequency distribution* of a variable can be produced using the FREQ command:

```
freq marital
```

This can be done for several variables in the same command:

```
freq marital age ethnic
```

The FREQ command produces a table listing counts, percentages, and cumulative percentages for each value of the specified variable. If the variable is numeric, the FREQ command will also display some *summary statistics* which may (or may not be) meaningful:

```
MARITAL ¦  Freq  Percent   Cum.
--------+----------------------
1       ¦    32   21.1%    21.1%
2       ¦    90   59.2%    80.3%
3       ¦     5    3.3%    83.6%
4       ¦     4    2.6%    86.2%
5       ¦    21   13.8%   100.0%
--------+----------------------
  Total ¦   152  100.0%

Sum                =    348.00
Mean               =      2.29
Standard deviation =      1.23
```

In this case the summary statistics (*sum*, *mean*, *standard deviation*) are meaningless because MARITAL (marital status) is a *categorical* variable (i.e. it contains codes that indicate membership of mutually exclusive categories). The quoted value for the mean (2.29) does **not** point somewhere between being single and divorced. The mean and standard deviation are **not** of use when dealing with categorical data. Many ANALYSIS commands will provide summary statistics and the results of statistical tests whether they are appropriate or not. It is up to you to decide if the output of any command is relevant.

Use the F1 help key to find out more about the FREQ command. Practice using the FREQ command with other variables.

Examine the output of the FREQ command carefully, making sure that you can interpret it.

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

Exercise 3 - Summarising distributions

Use the FREQ command to examine the distribution of the ROSENBERG score. How many people had a ROSENBERG score of less than or equal to fifteen? What was the most common ROSENBERG score? How many people had a ROSENBERG score greater than thirty?

Use the FREQ command to examine the distribution of the ETHNIC variable. What percentage of subjects were Indian? What percentage were Bangladeshi? What percentage were **not** White European?

The table below was produced using the command FREQ ETHNIC. Examine the table carefully and mark on the table the positions of the values of the variable, the counts for each value of the variable, the percentage for each value of the variable, the cumulative percentages, and the total number of cases:

```
ETHNIC ¦  Freq  Percent   Cum.
-------+----------------------
 1     ¦     1    0.7%     0.7%
 2     ¦     6    3.9%     4.6%
 3     ¦     9    5.9%    10.5%
 4     ¦     2    1.3%    11.8%
 5     ¦     3    2.0%    13.8%
 6     ¦     1    0.7%    14.5%
 7     ¦     2    1.3%    15.8%
 9     ¦     1    0.7%    16.4%
10     ¦   110   72.4%    88.8%
11     ¦    17   11.2%   100.0%
-------+----------------------
 Total ¦   152  100.0%

Sum                 =   1379.00
Mean                =      9.07
Standard deviation  =      2.54
```

What problems might there be in using the sum, mean and standard deviation to summarise this distribution?

Use the BAR or PIE command to chart the distribution of the ETHNIC variable.

Summarising distributions

Categorical variables

Variables such as ETHNIC, SEX, MARITAL, and the AGE variable (which is grouped into '25 and under' and 'over 25') are stored as a set of mutually exclusive codes. Such variables are known as *categorical* variables. The MARITAL variable, for example, is coded:

Code	Meaning
1	Married
2	Single
3	Divorced
4	Separated
5	Cohabiting
6	Widowed

With categorical variables it makes sense to use the FREQ command to produce a frequency distribution as this allows us to investigate the count (i.e. number) and proportion of cases that belong to each group and to identify groups with many or few members:

```
MARITAL ¦  Freq  Percent   Cum.
--------+----------------------
1       ¦    32   21.1%    21.1%
2       ¦    90   59.2%    80.3%   <- Largest (modal) group
3       ¦     5    3.3%    83.6%
4       ¦     4    2.6%    86.2%   <- Smallest group
5       ¦    21   13.8%   100.0%
--------+----------------------
  Total ¦   152  100.0%

Sum                =    348.00     <- Meaningless for categorical variables
Mean               =      2.29     <- Meaningless for categorical variables
Standard deviation =      1.23     <- Meaningless for categorical variables
```

Beware of the summary statistics presented below this frequency table as they are meaningless for categorical variables such as ETHNIC, SEX, and MARITAL.

Continuous variables

Some variables, such as GHQ, ROSENBERG, MEE, MDP, and MPA, hold *continuous* data. A frequency distribution is of different use with such variables as each value will account for only a small proportion of cases and does little to help us understand how the variable is distributed. As an example, produce a frequency distribution of the MPA variable by issuing the command:

```
freq mpa
```

If you study the resulting table closely enough you will be able to extract some useful information from it.

Frequency distributions of continuous data

Summarising continuous data with a frequency table

This is the frequency distribution of the MPA variable. Important information has been highlighted:

```
MPA   |  Freq  Percent   Cum.
------+----------------------
 5    |     1    1.2%     1.2%   <- minimum value
13    |     1    1.2%     2.4%
14    |     2    2.4%     4.8%
20    |     1    1.2%     6.0%
22    |     1    1.2%     7.2%
23    |     2    2.4%     9.6%
25    |     3    3.6%    13.3%
26    |     2    2.4%    15.7%
27    |     2    2.4%    18.1%
28    |     1    1.2%    19.3%
29    |     3    3.6%    22.9%
30    |     4    4.8%    27.7%
31    |     4    4.8%    32.5%
32    |     5    6.0%    38.6%
33    |     2    2.4%    41.0%
34    |    10   12.0%    53.0%   <- most common (modal) & middle (median) values
35    |     6    7.2%    60.2%
36    |     4    4.8%    65.1%
37    |     6    7.2%    72.3%
38    |     4    4.8%    77.1%
39    |     5    6.0%    83.1%
40    |     5    6.0%    89.2%
42    |     3    3.6%    92.8%
43    |     2    2.4%    95.2%
44    |     2    2.4%    97.6%
45    |     1    1.2%    98.8%
46    |     1    1.2%   100.0%   <- maximum value
------+----------------------
Total |    83  100.0%            <- number of observations (cases)

Sum                 =   2754.00
Mean                =     33.18   <- mean
Standard deviation  =      7.47   <- standard deviation
```

The highlighted information is known as the *seven figure summary* and can be used to summarise the distribution of continuous data. The seven summary measures are:

summary measure	formula	description
minimum	NONE	lowest value in the dataset
median	NONE	middle item in the dataset
mode	NONE	most common value in the dataset
maximum	NONE	highest value in the dataset
number of observations (n)	NONE	number of observations (cases)
mean ($\bar{x}$)	$\frac{\sum x}{n}$	sum of the observations divided by the number of observations
standard deviation	$\sqrt{\frac{\sum (x-\bar{x})^2}{(n-1)}}$	a measure of how much the distribution varies around the mean

Some texts refer to this as the *six figure summary*. This is because they replace the *minimum* and *maximum* with the *range* which is the difference between the maximum value and the minimum value.

Note that in this example the *modal* (most frequent) and *median* (middle) values are the same. This will **not** always be the case.

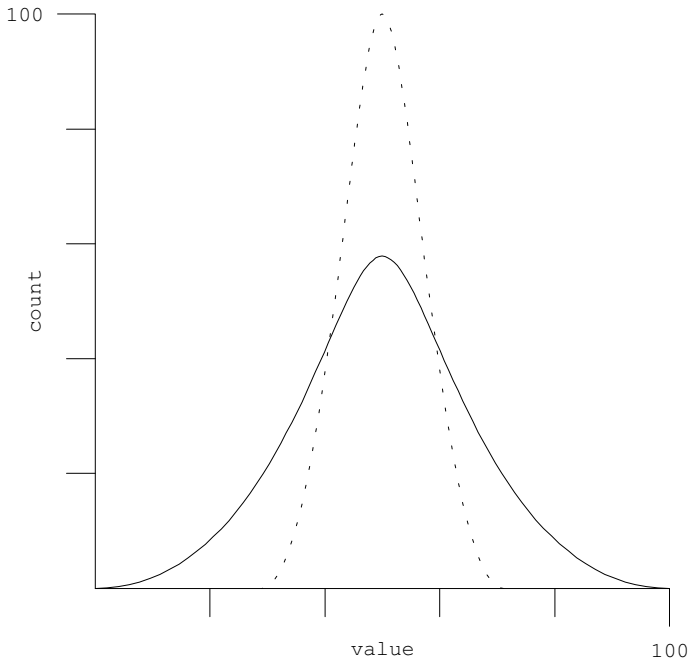
Summarising distributions of continuous data

The seven figure summary

With continuous data we are not interested in the counts of cases for each particular height, weight, or score. We are more interested in the complete distribution. This can be summarised with two separate measures: A measure of the *central tendency* and a measure of the *variability* of the data. The seven figure summary yields both measures:

Measuring	Measure
central tendency	mode median mean
variability	range (maximum - minimum) standard deviation

To understand why you need to examine both *central tendency* and *variability* consider the following two distributions:



The two distributions shown on the graph have similar *mean*, *median*, and *modal* values. If we just used these measures of central tendency to describe the distributions we would not be describing them properly. We would conclude that they were the same as each other. As well as reporting the central tendency we also need to report the variability of the two distributions using the *range* and the *standard deviation*:

measure	wide	narrow
mean	50	50
median	50	50
mode	50	50
range	100	50
standard deviation	15	8

This allows us to see that one distribution has a wider range of values (i.e. it is more variable) than the other.

Summarising distributions of continuous data

How the standard deviation measures variability

The formula for the standard deviation appears complicated:

$$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$$

but it can be easily understood once broken down into its component steps. The formula relies on subtracting the mean from each data value:

$$x-\bar{x}$$

and *summing* (adding up) the *deviations* (differences):

$$\sum(x-\bar{x})$$

The positive and negative deviations will cancel each other out (giving zero) so each of the deviations is squared (multiplied by itself) to give positive values before summation:

$$\sum(x-\bar{x})^2$$

The sum of the squared deviations is divided by the number of cases contributing to the mean minus one:

$$\frac{\sum(x-\bar{x})^2}{(n-1)}$$

This gives a measure of *average* deviation from the mean which is known as the *variance*. The variance is expressed in *squared* units which are difficult to interpret. To get over this problem, the square root of the variance is used:

$$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$$

This measure is known as the *standard deviation* and is a measure of average deviation from the mean expressed in the original unit of measure. The more variable the data, the larger the standard deviation will be.

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.

Standard deviation - worked example

Calculating the standard deviation

The table below shows four data points, their sum, and their mean:

	Data
	3
	7
	6
	4
Sum	20
Mean	5

The first step in calculating the standard deviation is to subtract the mean from each data point to give the deviation of each data point from the mean:

	Data	point - mean
	3	−2
	7	+2
	6	+1
	4	−1
Sum	20	0
Mean	5	

Because the deviations add-up to zero, they are squared to make them all positive numbers:

	Data	point - mean	(point - mean)^2
	3	−2	4
	7	+2	4
	6	+1	1
	4	−1	1
Sum	20	0	10
Mean	5		

Dividing the sum of the squared deviations by the number of data points minus one gives us the variance:

$$10 / (4 - 1) = 3.33$$

Taking the square root of the variance gives us the standard deviation:

$$\sqrt{3.33} = 1.82$$

Standard deviation and degrees of freedom

Why use *n - 1* instead of *n* to calculate the standard deviation?

The standard deviation is calculated by taking the square root of the sum of the squared deviations from the mean divided by the number of cases minus one:

$$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$$

Note that *n - 1* is used to calculate the average squared deviation. This is because subtracting the mean from each value leads to the 'loss' of one item of data. The deviations may not be any group of *n* numbers. The sum of the deviations is always zero. When *n - 1* deviations are specified, the last one is *determined* by the condition that they sum to zero:

$$\sum(x-\bar{x}) = 0$$

The final deviation is the sum of the other deviations with a change of sign (i.e. a positive number will become negative or a negative number will become positive). This is known as the loss of a *degree of freedom*. The deviations have *n - 1* degrees of freedom because the final deviation has to have a particular value once *n - 1* deviations are specified. You can see this 'loss' in this example with four data points:

	Data	*n* deviations	*n - 1* deviations
	3	−2	−2
	7	+2	+2
	6	+1	+1
	4	−1	**??**
Sum	20	0	+1
Mean	5		

The sum of *n - 1* deviations determines the value of the missing deviation which is marked '??' in the table. The sum of *n - 1* deviations is +1 so the missing deviation must be -1 because the sum of all deviations must be zero. As the missing deviation can be determined from *n - 1* deviations and the constraint that the sum of deviations must be zero, there are really only *n - 1* pieces of information (degrees of freedom) for the deviations.

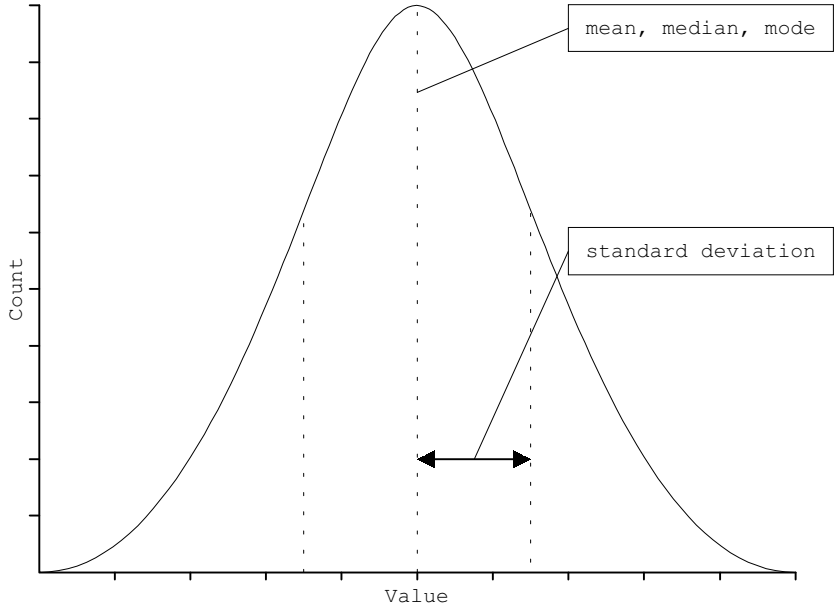
The range and the standard deviation

Both the range and the standard deviation measure variability. The range is simple to calculate but has an undesirable property. It is calculated using the two extreme values only (the minimum value and the maximum value). Any *outliers* (extremely high or low values) will be included in these extremes making the range extremely sensitive to (potentially erroneous or unrepresentative) outliers in the data. The standard deviation is often preferred to the range because all data points (not just two) are used and it is less effected (but may still be distorted) by outliers in the data.

Problems with the mean and standard deviation

The normal distribution

The mean and standard deviations describe a particular type of distribution very well. This distribution is *symmetrical* and bell shaped with mean median and mode similar to each other:



and is called the *normal distribution*. If your data has a very different type of distribution then the mean and standard deviation may not describe the distribution well. With non-normal distributions it may be best to use alternative measures of central tendency (e.g. the median) and variability (e.g. the range or *inter-quartile* distance) or apply a *transformation* to your data to bring it closer to the normal distribution (types of distribution and the range of transformations that can be applied to them are covered on pages 64 - 69).

Fortunately, the normal distribution is very common in nature. Many of the variables you collect will be normally distributed and you will be able to use the mean and standard deviation to describe them accurately.

It is very unlikely that you will see a perfectly normal distribution each time you chart your data. *Random variation* in the data will tend to make the distributions look ragged and may also make the distribution slightly non-symmetrical. There will usually be a few *extreme* or *outlier* values in the left and right *tails* of the distribution. If the data values tend to cluster in a symmetrical manner around a central value and there are progressively fewer and fewer data values as you move outward from the middle value and there are no data values that are very far way from the bulk of the other data values then you may safely assume that the data come from a variable that is normally distributed and that it is reasonable to use the mean and standard deviation to summarise the distribution.

Summarising distributions of continuous data

Calculating summary measures in analysis

ANALYSIS provides the MEANS command for dealing with continuous data. The MEANS command produces summary measures by different levels of a grouping variable (e.g. a mean MPA score for each SEX). If we want to look at the summary measures for a variable we must first create a new variable that contains the same value for each case (this is a limitation of ANALYSIS). Issue the command:

```
define all <AAAAAAAAA>
all = "ALL CASES"
```

These two commands create a variable called ALL and set it to hold the value "ALL CASES" for every case in the dataset. Issue the command:

```
means mpa all
```

to instruct ANALYSIS to calculate and display summary measures for the MPA variable. The command produces a frequency table and a table showing summary measures:

```
ALL              Obs      Total       Mean   Variance    Std Dev
ALL CASES         83       2754     33.181     55.784      7.469

ALL          Minimum      25%ile     Median      75%ile    Maximum       Mode
ALL CASES      5.000      30.000     34.000      38.000     46.000     34.000
```

It is always a good idea to look at the frequency table as ANALYSIS will only report the **first** modal value if there is more than one *mode*.

You can instruct ANALYSIS not to produce the frequency table with the MEANS command by adding '/n' to the end of the command:

```
means mpa all /n
```

Examine the output carefully and make sure that you can identify all of the measures of the seven figure summary. Note that there are only eighty-three cases. This is because MPA was not collected for undergraduate psychology students.

In addition to the seven figure summary, ANALYSIS also reports the cut-point values for the *quartiles*:

```
ALL              Obs      Total       Mean   Variance    Std Dev
ALL CASES         83       2754     33.181     55.784      7.469

ALL          Minimum      25%ile     Median      75%ile    Maximum       Mode
ALL CASES      5.000      30.000     34.000      38.000     46.000     34.000
```

These values divide the distribution into four equal sized groups and can also be a useful way of summarising a distribution. The 25th percentile (lower quartile) and the 75th percentile (upper quartile) are the boundary values for the middle 50% of the data. The *interquartile range* or *interquartile distance* (i.e. 75th percentile - 25th percentile) is sometimes used as an alternative measure of variability. The advantage of using the interquartile range (instead of the range or the standard deviation) is that it is **not** effected by outliers.

Summarising distributions with charts

Categorical data

You can summarise categorical data using BAR and PIE charts. This is because these charts plot absolute counts (BAR charts) and counts as a proportion (percentage) of the sample population (PIE charts). Issue the following commands:

```
pie marital
bar marital
```

Both charts show the same thing. They are graphical equivalents of the FREQ command. The BAR chart shows counts and the PIE chart shows proportions (percentages).

Continuous data

Continuous data can be summarised using either histograms or line charts. Issue the following commands:

```
histogram mdp
line mdp
```

This sort of line chart is sometimes called a frequency polygon. You can also use the LINE command to show the distribution of a continuous variable grouped by *levels* (values or categories) of a categorical variable. Issue the command:

```
line mdp sex
```

The labels for the lines (1 and 2) are taken from the values of the SEX variable.

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

Exercise 4 - Summarising distributions

Use the MEANS command to produce the seven figure summaries for the GHQ, ROSENBERG, MEE, MDP, and MPA variables and complete the following table:

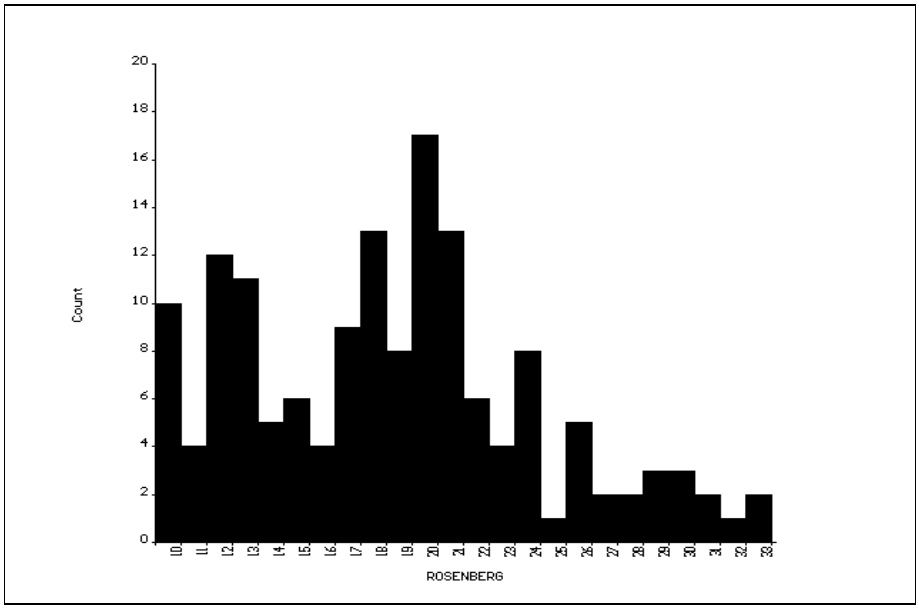
	GHQ	ROSENBERG	MEE	MDP	MPA
number of cases					
mode					
median					
mean					
minimum					
maximum					
standard deviation					

Use the HISTOGRAM command to graph the distributions of the GHQ, ROSENBERG, MEE, MDP, and MPA variables.

The command:

```
histogram rosenberg
```

produces the following graph:



Mark the mode, median, and mean, minimum, and maximum (from the previous part of this exercise) on this graph.

Creating and recoding variables

The DEFINE command

To create a new variable (e.g. GHQCASE) you first need to tell ANALYSIS what kind of variable GHQCASE will be using the DEFINE command. Variable types that could be used for the new variable are:

Name	Typical Values	Variable Type	DEFINE 'Code'
GHQCASE	1, 2	Number	#
GHQCASE	Y, N	Logical (Yes/No)	<Y>
GHQCASE	+, -	Character	<A>
GHQCASE	"CASE", "NOT CASE"	Character	<AAAAAAAA>

Issue the command:

```
define ghqcase <AAAAAAAA>
```

to create a new character variable called GHQCASE. The number of characters between the angle brackets ('<' and '>') defines the length of new variable. Use the ANALYSIS help system to discover how other codes can be used with the DEFINE command.

Press F4 BROWSE. You should see a blank column at the right hand side of the data file which has the heading GHQCASE (you may need to press → or CTRL+→ to see this new column). ANALYSIS has assigned a new character column for this variable. At present all the values are blank (or 'missing').

Recoding continuous variables into categorical variables

Issue the command:

```
recode ghq to ghqcase 0-4="NOT CASE" 5-hi="CASE"
```

to RECODE the number variable GHQ into the character variable GHQCASE. The RECODE command must be issued **after** defining the new variable. Your data file now has the original variable GHQ and a new variable GHQCASE. The choice of *cut-points* for group membership is often *arbitrary* and can make a difference to the results of analysis. With variables such as GHQ it is valid to choose a cut-point of five because the GHQ-28 score was developed to ensure that this cut-point was not arbitrary. Sometimes an agreed definition of cut-points for (e.g.) high and low groups for a particular continuous measure will not be available. With these variables, you might choose cut-points such as the mean, median, mode, or quartiles.

After using RECODE you should always check to see if the command has had the expected result using the LIST or BROWSE commands:

```
list ghq ghqcase
browse ghq ghqcase
```

An alternative way of recoding data is to use the IF ... THEN command. Try this by typing:

```
define meegroup <AAAAAAAAAAAA>
if mee <= 16 and mee <> . then meegroup = "LOW"
if mee >= 17 and mee <= 33 then meegroup = "INTERMEDIATE"
if mee >= 34 then meegroup = "HIGH"
```

In this example the lower and upper quartiles (found using the MEANS command) have been chosen as cut-points. Check the results of the IF ... THEN commands using the F4 BROWSE command (you will need to press → or CTRL+→ to view the new column). Examine the distribution of MEEGROUP using the FREQ command.

Creating and recoding variables

Another use for RECODE

A problem with the LINE and HISTOGRAM commands in ANALYSIS is that they do not *group* data automatically. This can often make it difficult to see the underlying distribution. Enter the command:

```
histogram mee
```

It is difficult to see the underlying distribution. Issue the commands:

```
define meegrp ##
recode mee to meegrp 0-9=0 10-19=10 20-29=20 30-39=30 40-49=40 50-59=50
```

to group the values of the MEE variable into six groups (or *intervals*). Use the BROWSE or LIST command to check that the commands have had the desired effect. Issue the command:

```
histogram meegrp
```

The underlying distribution is now easier to see.

There are four rules that you should follow when recoding continuous data into intervals:

1. The groups you create must be mutually exclusive. The same value must **not** be assigned to more than one group. For example, a RECODE command with the form:

   ```
   recode mee to meesame 0-10=0 10-20=10 20-30=30 ...
   ```

   is *ambiguous* for cases where MEE = 10 and MEE = 20.

2. The intervals should be of the same width. Different interval widths can make the histogram difficult to interpret. The commands:

   ```
   define meediff ##
   recode mee to meediff 0-9=0 10-19=10 20-24=20 25-29=25 30-60=30
   histogram meediff
   ```

   give a wrong picture of the underlying distribution of the MEE variable.

3. You should choose enough groups (i.e. the groups should be narrow enough) to show some of the variation in the data. The commands:

   ```
   define meefew ##
   recode mee to meefew 0-32=0 30-60=30
   histogram meefew
   ```

   also give a wrong picture of the underlying distribution of the MEE variable. Between five and ten groups is usually sufficient.

4. The groups must cover all values of the variable being grouped.

These rules only apply to investigating the underlying distribution of a variable and need not apply to grouping a variable in another context. ANALYSIS provide a quick way of recoding data into equal width intervals using the BY keyword. The commands:

```
define ghqquick <AAAAAAAA>
recode ghq to ghqquick by 5
```

Will recode the GHQ variable into groups of width five.

Creating and recoding variables

Recoding categorical variables

So far we have used the RECODE and IF ... THEN commands to RECODE continuous variables into categorical variables. You can also use these commands to RECODE categorical variables.

One of the objectives of the **Stress and the Student Social Worker** study was 'To determine whether students undertaking professional training exhibited higher stress scores than undergraduate students'. The COHORT variable records which group each individual case belongs to:

```
COHORT      ¦  Freq  Percent   Cum.
------------+----------------------
ACADEMIC    ¦    69   45.4%    45.4%
PGCE        ¦    33   21.7%    67.1%
SOCIAL WORK ¦    50   32.9%   100.0%
------------+----------------------
      Total ¦   152  100.0%
```

These three groups need to be collapsed into two groups in order to explore this hypothesis. To do this we must first define a new variable:

```
define profession <A>
```

and then use the IF ... THEN command to create the two groups:

```
if cohort = "ACADEMIC" then profession = "-"
if cohort = "PGCE" then profession = "+"
if cohort = "SOCIAL WORK" then profession = "+"
```

and finally use the BROWSE command to check the effect of these commands:

```
browse cohort profession
```

Now that we have two variables GHQCASE and PROFESSION which are coded as *binary categorical variables* we can use the TABLES command to produce a *two-by-two table* to investigate the *hypothesis* that students undertaking professional training exhibit higher stress scores than undergraduate students:

```
tables profession ghqcase
```

Two-by-two tables

Two-by-two tables : Naming of parts

Measures of association between two variables are easily studied by means of *two-by-two 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

With the **Stress and the Student Social Worker** study data, the command:

```
tables profession ghqcase
```

produces the following table:

	GHQCASE	
PROFESSION | CASE	NOT CASE |	Total
+ |	62	21 | 83
- |	29	40 | 69
Total |	91	61 | 152

as well as some summary statistics which we will cover later in the book. In this table our *exposure* variable is PROFESSION and our *outcome* variable is CHQCASE. Using the terms introduced above we have:

a = **62** = number exposed and ill
b = **21** = number exposed and not ill
c = **29** = number unexposed and ill
d = **40** = number unexposed and not ill
a + b = **83** = number exposed
c + d = **69** = number unexposed
a + c = **91** = number ill
b + d = **61** = number not ill
a + b + c + d = **152** = number in study

In this example *exposed* means ‘undergoing professional training’ and *ill* means ‘having a GHQ-28 score of five or more’. Make sure that you can identify these figures in the table before proceeding.

Two-by-two tables

The absolute risk

The *absolute risk* is the risk of an outcome happening 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 number of people experiencing illness divided by the total population at risk. From the table:

PROFESSION	CASE	GHQCASE NOT CASE	Total
	+	62	21	83
	-	29	40	69
Total		**91**	61	**152**

The absolute risk can be calculated as:

```
(a + c) / (a + b + c + d) = 91 / 152 = 0.599 = 59.9%
```

Each person in the **Stress and the Student Social Worker** study ran a risk of 59.9% of having a GHQ-28 score of five (5) or more. This is the same as saying that 59.9% of people in the study can be considered as 'cases' as defined by their GHQ-28 score.

The absolute risk measures the *absolute magnitude* of (e.g.) a health problem in a population but provides **no** information regarding the association between *risk factors* (exposure variables) and outcomes. It is simply the proportion of individuals with a particular outcome in the *study population*.

It is only valid to calculate an absolute risk with data from a *cross-sectional* or *cohort* study (i.e. where members of the study population are selected according to their exposure status). 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 outcome status). In a case-control study of the stress among social work trainees 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 152 persons in the study. It is **not** valid to calculate an absolute risk with data from a case-control study.

Two-by-two tables

Exposure specific risks

A more useful approach is to calculate *exposure specific risks* or *attack rates*. This yields a measure that can be interpreted as the risk of an outcome given an exposure which can easily be calculated from a two-by-two table:

	GHQCASE		
PROFESSION	CASE	NOT CASE	Total
+	**62**	21	**83**
-	29	40	69
Total	91	61	152

In this table the exposure specific risk of being a 'case' can be calculated as:

```
a / (a + b) = 62 / (62 + 21) = 62 / 83 = 0.747 = 74.7%
```

Each person in the study undergoing professional training (i.e. "PGCE" or "SOCIAL WORK") ran a risk of 74.7% of being defined as a ‘case’ based on their GHQ-28 score. This is the same as saying that 74.7% of people who are undergoing professional training are ‘cases’. This method makes specific reference to risk factors and will vary with each tabulation of exposure by outcome.

The exposure specific risk will be different for each exposure whereas the absolute risk does not vary with exposure. The exposure specific risk estimates the risk of an outcome given an exposure but reveals little about the *excess risk* associated with exposure.

It is **not** valid to calculate an exposure specific risk with data from a case-control study.

Two-by-two tables

Relative risk

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:

	GHQCASE		
PROFESSION	CASE	NOT CASE	Total
+	**62**	21	**83**
-	**29**	40	**69**
Total	91	61	152

It is possible to calculate the risk of being a 'case' for those undergoing professional training as:

```
a / (a + b) = 62 / (62 + 21) = 62 / 83 = 0.746 = 74.7%
```

and the risk of being a 'case' for those **not** undergoing professional training as:

```
c / (c + d) = 29 / (29 + 40) = 29 / 69 = 0.420 = 42.0%
```

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

```
(a / (a + b)) / (c / (c + d)) = (62 / 83) / (29 / 69) = 0.747 / 0.420 = 1.78
```

A person undergoing professional training was 1.78 times as likely to be a 'case' as an academic undergraduate.

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 an outcome. 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 outcome
= 1	exposure **not** associated with outcome
> 1	exposure **may** increase risk of outcome

Two-by-two tables

Odds ratio

The examples given above assume that it is valid to calculate a relative risk of outcome. It is only valid to calculate the relative risk of outcome in a study which has **not** selected subjects according to their outcome status (i.e. a study which has **not** selected *cases* or *controls*). The relative risk of an outcome 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:

		GHQCASE	
PROFESSION	CASE	NOT CASE	Total
+	**62**	**21**	83
-	**29**	**40**	69
Total	91	61	152

In this example the *odds* of being a 'case' given professional training can be calculated as:

```
a / b = 62 / 21 = 2.95
```

The odds of being a 'case' in the absence of professional training can be calculated as:

```
c / d = 29 / 40 = 0.73
```

The odds ratio can be calculated as:

```
(a / b) / (c / d) = (62 / 21) / (29 / 40) = 2.95 / 0.73 = 4.07
```

Undergoing professional training was associated with a 4.07 fold increase in the odds of being a 'case'. The odds ratio is interpreted a similar way to the relative risk:

Odds Ratio	Interpretation
< 1	exposure **may** protect against outcome
= 1	exposure **not** associated with outcome
> 1	exposure **may** increase risk of outcome

Odds ratios and relative risks

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 **cannot** be estimated and the odds ratio must be used. In case-control studies of rare outcomes 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 **Stress and the Student Social Worker** data because subjects were selected by their exposure status ("ACADEMIC", "PGCE", or "SOCIAL WORK") rather than their outcome status (e.g. their GHQ-28 score). 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.

Two-by-two tables

Measures of risk

If we examine the two-by-two table:

PROFESSION	GHQCASE CASE	NOT CASE	Total
+	62	21	83
-	29	40	69
Total	91	61	152

of PROFESSION by GHQCASE closely we can extract the following information:

exposed and ill	profession = + **and** ghqcase = CASE	a	**62**
exposed and not ill	profession = + **and** ghqcase = NOT CASE	b	**21**
unexposed and ill	profession = - **and** ghqcase = CASE	c	**29**
unexposed and not ill	profession = - **and** ghqcase = NOT CASE	d	**40**
exposed	profession = +	a + b	**83**
unexposed	profession = -	c + d	**69**
ill	ghqcase = CASE	a + c	**91**
not ill	ghqcase = NOT CASE	b + d	**61**
total records	sum of all the cells in the table	a + b + c + d	**152**

Identify these figures in the displayed table.

We can use this information to calculate various risk measures:

absolute risk	(a + c) / (a + b + c + d)	91 / 152	**0.60**
exposure specific risk	a / (a + b)	62 / 83	**0.75**
relative risk	(a / a + b) / (c / c + d)	(62 / 83) / (29 / 69)	**1.78**
odds ratio	(a / b) / (c / d)	(62 / 21) / (29 / 40)	**4.07**

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

Two-by-two tables

Measures of risk

ANALYSIS calculates relative risks and odds ratios for two-by-two tables together with *confidence limits*:

```
Odds ratio                                                    4.07
Cornfield 95% confidence limits for OR      1.93 < OR < 8.68

Relative risk of (GHQCASE=CASE) for (PROFESSION=+)      1.78
Greenland, Robins 95% conf. limits for RR   1.31 < RR < 2.41
                (Biometrics 1985;41:55-68)
```

and will calculate the absolute risk and exposure specific risk if you issue the command:

```
set percents = on
```

before the TABLES command. This command instructs ANALYSIS to display *row* and *column percentages* in tables. To see the effect of this issue the command:

```
tables profession ghqcase
```

This command now produces the following output:

```
                        GHQCASE
PROFESSION |CASE    NOT CASE | Total
-----------+-----------------+------
         + |      62      21 |    83   <- Counts
           >   74.7%   25.3% > 54.6%   <- Row percentages
           |   68.1%   34.4% |         <- Column percentages
         - |      29      40 |    69
           >   42.0%   58.0% > 45.4%
           |   31.9%   65.6% |
-----------+-----------------+------
     Total |      91      61 |   152
           |   59.9%   40.1% |
```

The values in each cell of the table are presented as *count*, *row percentage*, *column percentage*. Examine the table. The row percentage of the cell PROFESSION = "+" and GHQCASE = "CASE" is equivalent to the exposure specific risk or attack rate (74.7%) and the row percentage of the column total for CHQCASE = "CASE" is the absolute risk (59.9%). 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. To instruct ANALYSIS to stop displaying cell percentages and separating lines issue the commands:

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

Use the ANALYSIS help system to find out what else you can change using the SET command.

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 risk-variable outcome-variable
```

The risk or exposure variable (e.g. PROFESSION) should **always** be the first variable you specify with the TABLES command. The outcome variable (e.g. GHQCASE) should always be the second variable you specify with the TABLES command.

Two-by-two tables

Confidence intervals

Measures of exposure effect are subject to *random variation*. The calculated relative risk will differ from the *true* or *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% or, sometimes, 99%. 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. There is a 2½% chance of the true relative risk being below the lower confidence limit and a 2½% chance of the true relative risk being above the upper confidence limit.

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 ANALYSIS.

Interpretation of the confidence interval

The following relative risk and 95% confidence limits have been calculated from the **Stress and the Student Social Worker** data (using the command TABLES PROFESSION GHQCASE) for the association between undergoing professional training and having a GHQ Score of five (5) or higher:

```
Relative risk of (GHQCASE=CASE) for (PROFESSION=+)      1.78
Greenland, Robins 95% conf. limits for RR   1.31 < RR < 2.41
```

The observed relative risk is 1.78. The true relative risk is likely to lie somewhere between 1.31 and 2.41. The true relative risk is unlikely to be equal to one. It is reasonable to state that undergoing professional training was associated with being a ‘case’.

The following relative risk and 95% confidence limits have been calculated from the **Stress and the Student Social Worker** data for the association between gender and having a GHQ Score of five (5) or higher:

```
Relative risk of (GHQCASE=CASE) for (SEX=1)             0.99
Greenland, Robins 95% conf. limits for RR   0.74 < RR < 1.33
```

The observed of the relative risk is 0.99. The true relative risk is likely to lie somewhere between 0.74 and 1.33. It is **not** unlikely that the true relative risk is equal to one. It is **not** reasonable to state that gender was associated with being a ‘case’.

The odds ratio is also a point estimate and the confidence interval is used in the same way.

Two-by-two tables

Testing a hypothesis about association

Measures of 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 exposure 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 hypothesis 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:

		GHQCASE	
PROFESSION	CASE	NOT CASE	Total
+	62	21	83
-	29	40	69
Total	91	61	152

The *null hypothesis* is that undergoing professional training is **not** associated with being a 'case'. The *alternative hypothesis* is that undergoing professional training **is** associated with being a 'case'. 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 the proportion of people defined as 'cases' (*absolute risk*) is:

```
(a + c) / (a + b + c + d) = 91 / 152 = 0.598 = 59.8%
```

If the null hypothesis were true, the *expected* number of people being defined as 'cases' whilst undergoing professional training would be 59.8% of 83 (i.e. the risk of being a case would **not** vary with exposure), or:

```
Expected[a] = 83 * (91 / 152) = 49.69
```

Expected numbers for each of the cells in the table can be calculated in a similar way:

```
Expected[a] = 83 * (91 / 152) = 49.69

Expected[b] = 83 * (61 / 152) = 33.31

Expected[c] = 69 * (91 / 152) = 41.31

Expected[d] = 69 * (61 / 152) = 27.69
```

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

Two-by-two tables

Testing a hypothesis about association

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

PROFESSION	CASE	NOT CASE	Total
+	62	21	83
-	29	40	69
Total	91	61	152

EXPECTED

PROFESSION	CASE	NOT CASE	Total
+	49.69	33.31	83.00
-	41.31	27.69	69.00
Total	91.00	61.00	152.00

by subtracting the expected values from the values observed in the data:

PROFESSION	CASE	NOT CASE	Total
+	12.31	-12.31	00.00
-	-12.31	12.31	00.00
Total	00.00	00.00	00.00

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

PROFESSION	CASE	NOT CASE	Total
+	151.54	151.54	303.08
-	151.54	151.54	303.08
Total	303.08	303.08	606.16

before dividing them by the *expected* values:

PROFESSION	CASE	NOT CASE	Total
+	3.05	4.55	7.60
-	3.67	5.47	9.14
Total	6.72	10.02	16.74

The overall total of this table (16.74) 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*:

$$X^2 = \sum\left[\frac{(O-E)^2}{E}\right]$$

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 16.74 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 undergoing professional training and being a ‘case’. If the probability of obtaining 16.74 under the null hypothesis is large then the null hypothesis might be true and we would not assume that there is an association between professional training and being a ‘case’.

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

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*. In this table:

		GHQCASE	
PROFESSION	CASE	NOT CASE	Total
+	62	**??**	83
-	**??**	**??**	69
Total	91	61	152

knowing the value of one internal cell allows us to determine the values of the other three internal cells. This table has one degree of freedom. The degrees of freedom in a table is calculated using the formula:

```
df = (number of rows - 1) * (number of columns - 1)
```

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

```
df = (number of rows - 1) * (number of columns - 1) = (2 - 1) * (2 - 1) = 1
```

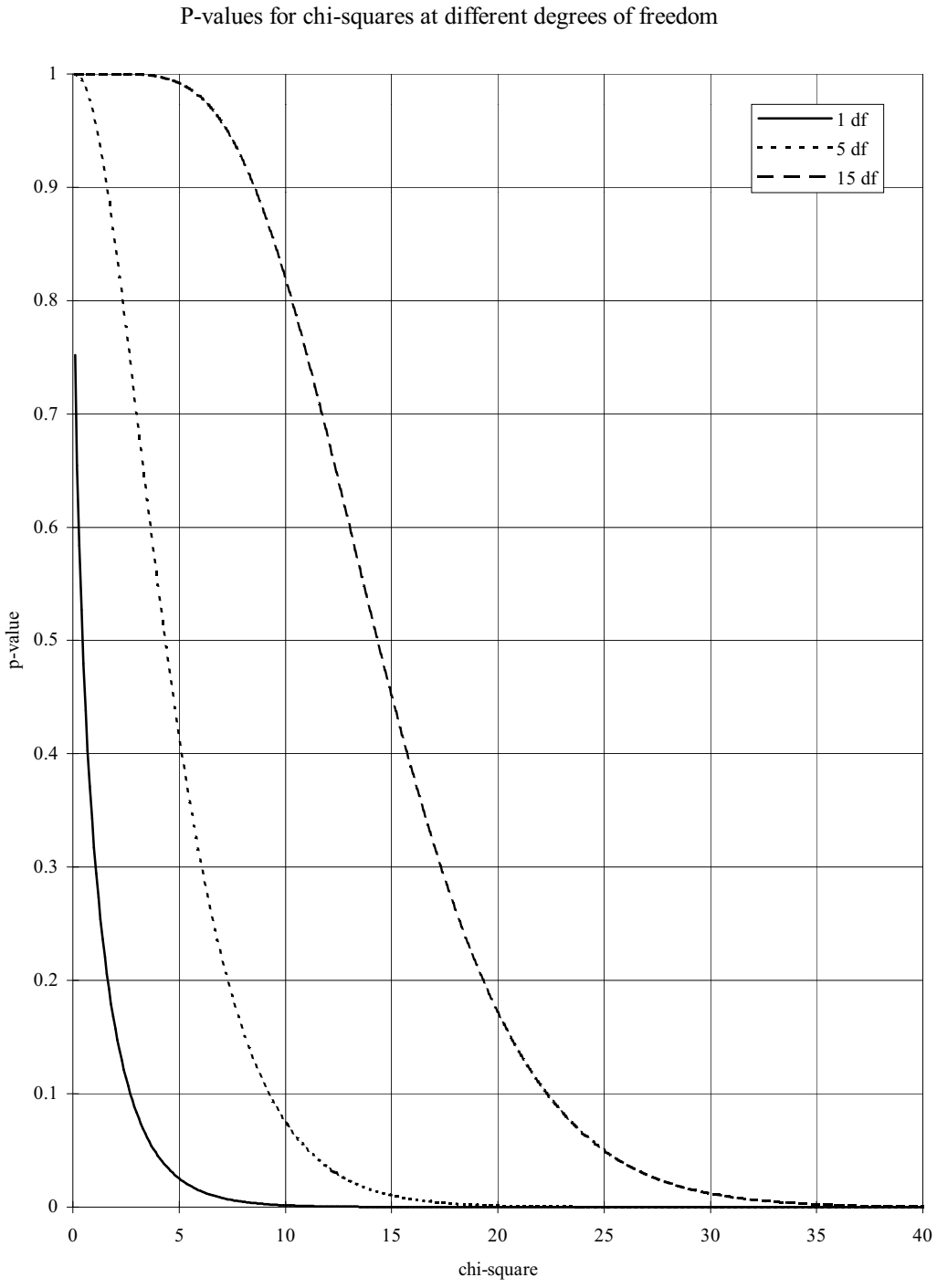
The chart on the next page shows the probability of observing chi-square values at different degrees of freedom. Look at the chi-square distribution for one degree of freedom. The probability of getting a chi-square value of 16.74 from a 2-by-2 table is very small. This probability is referred to as the *p-value* and 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 (i.e. less than one in a thousand). It is unlikely that the observed association arose by chance. It is reasonable to reject the null hypothesis of no association. It is generally considered safe to reject the null hypothesis with a p-value less than 0.05 (i.e. less than one in twenty). 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 several chi-square tests. The one described here is called *Pearson's chi-square*. Another test of association is *Yates' corrected chi-square* which uses the same formula but includes a *correction factor* and is used for two-by two tables. Yates' corrected chi-square is always slightly smaller than Pearson's chi-square and will always yields a slightly larger p-value. Most statisticians prefer 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 when expected numbers are small. Fisher's exact test is more *robust* and can yield *reliable* results with small expected 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. ANALYSIS will report Pearson's chi-square for all tables and Yate's corrected chi-square for two-by-two tables. If an expected value in a two-by-two table is less than five then ANALYSIS will also report Fisher's exact test. There are two p-values associated with Fisher's exact test: the *one-tailed* p-value and the *two-tailed* p-value. The one-tailed p-value is appropriate when we are interested in examining the effects of exposure variables in **one direction** only (e.g. do they **increase** the risk of ILLness?). In other types of study we are usually testing for any association and would use a two-tailed p-value. If in doubt you should use the two-tailed p-value. The one-tailed p-value would be appropriate here because we are interested in examining the effects of an exposure variables in one direction only (i.e. does undergoing professional training increase the probability of scoring five or more on GHQ-28).

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

The chi-square distribution

The chart below shows the probability of observing chi-square values at three different degrees of freedom:



Mark a chi-square value of 16.74 on the x-axis (chi-square on *x-axis*) and estimate the probability of calculating this value from data in a 2-by-2 table (p-value on *y-axis*) if the null hypothesis were true.

Two-by-two tables

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 *variability* of the estimate. The statistical significance testing approach only indicates the 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 AGE and being a 'case':

```
                          GHQCASE
AGE         ¦CASE    NOT CASE ¦ Total
------------+-----------------+------
          1 ¦      39      36 ¦    75
          2 ¦      52      25 ¦    77
------------+-----------------+------
      Total ¦      91      61 ¦   152
```

This table yields a chi-square statistic of 3.82 with a p-value of 0.05. If we assume that the sample size of the study had been three times larger (by multiplying all cells in the table by three) the following table would result:

```
                          GHQCASE
AGE         ¦CASE    NOT CASE ¦ Total
------------+-----------------+------
          1 ¦     117     108 ¦   225
          2 ¦     156      75 ¦   231
------------+-----------------+------
      Total ¦     273     183 ¦   456
```

which yields a chi-square statistic of 11.45 (3 * 3.82) with a P-value of 0.0007. 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:

```
Relative risk of (GHQCASE=CASE) for (AGE=1)             0.77
Greenland, Robins 95% conf. limits for RR   0.59 < RR < 1.01
```

for the original table and:

```
Relative risk of (GHQCASE=CASE) for (AGE=1)             0.77
Greenland, Robins 95% conf. limits for RR   0.66 < RR < 0.90
```

for the table with increased sample size. 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!

Two-by-two tables

An example of ANALYSIS output

The following table was produced using the command TABLES PROFESSION GHQCASE. Important information has been highlighted:

```
                        GHQCASE
PROFESSION |CASE    NOT CASE | Total
-----------+-----------------+------
         + |      62      21 |    83
         - |      29      40 |    69
-----------+-----------------+------
     Total |      91      61 |   152

                  Single Table Analysis
Odds ratio                                              4.07 <- Odds Ratio
Cornfield 95% confidence limits for OR      1.93 < OR < 8.68 <- Confidence Limits

Relative risk of (GHQCASE=CASE) for (PROFESSION=+)      1.78 <- Relative Risk
Greenland, Robins 95% conf. limits for RR   1.31 < RR < 2.41 <- Confidence Limits
                (Biometrics 1985;41:55-68)
        Ignore relative risk if case control study.

                        Chi-squares   P-values
                        -----------   --------
        Uncorrected:        16.74     0.00004292 <---        <- Chi-square/p-value
        Mantel-Haenszel:    16.63     0.00004549 <---
        Yates corrected:    15.41     0.00008674 <---        <- Corrected values
```

Notice that ANALYSIS produces three chi-square measures and p-values. With two-by-two tables you may quote either the *uncorrected* or *corrected* values. The corrected value is preferred as it is more *conservative* and less prone to error.

Summary of measures introduced

The table below summarises the measures of effect and association introduced so far:

Measure	Value	Interpretation	Notes
Relative risk	< 1	decrease in risk	**not** valid for case-control studies
	= 1	no effect	
	> 1	increase in risk	
Odds ratio	< 1	decrease in risk	valid for case-control studies
	= 1	no effect	
	> 1	increase in risk	
Confidence interval	includes 1	observed association could be due to random variation	applies to relative risk / odds ratio
	excludes 1	observed association unlikely to be due to random variation	
P-value	> 0.050	observed association could be due to random variation	applies to test statistic
	< 0.050	observed association unlikely to be due to random variation	
	< 0.010	observed association very unlikely to be due to random variation	
	< 0.001	observed association extremely unlikely to be due to random variation	

Contingency tables

Tables larger than two-by-two

In tables larger than two-by-two it is difficult to calculate risk measures and so we must rely on significance testing. Issue the command:

```
tables marital ghqcase
```

to investigate the association between marital status and being a 'case'. The following table is produced:

```
                        GHQCASE
MARITAL     |CASE     NOT CASE | Total
------------+------------------+------
          1 |       23       9 |    32
          2 |       48      42 |    90
          3 |        3       2 |     5
          4 |        4       0 |     4
          5 |       13       8 |    21
------------+------------------+------
      Total |       91      61 |   152


          Chi square =       6.24        <- Chi-square
Degrees of freedom =          4
            p value = 0.18210692        <- p-value
```

If you examine the statistics that follow the table you will see that MARITAL status does not appear to be associated with being a 'case'. The p-value that ANALYSIS reports for the chi-square statistic is 0.18210692 (i.e. $p > 0.05$) indicating that any differences between the values we observe in the table and the values we would expect if the null hypothesis of no association were true could be due to random variation.

You can get a better understanding of what is going on in this table if you issue the command:

```
set percents = on
```

before issuing the TABLES command and examining the row percentages:

```
                        GHQCASE
MARITAL     |CASE     NOT CASE | Total
------------+------------------+------
          1 |       23       9 |    32   <- Counts
            >   71.9%   28.1% > 21.1%   <- Row percentages
            |   25.3%   14.8% |         <- Column percentages
          2 |       48      42 |    90
            >   53.3%   46.7% > 59.2%
            |   52.7%   68.9% |
          3 |        3       2 |     5
            >   60.0%   40.0% >  3.3%
            |    3.3%    3.3% |
          4 |        4       0 |     4
            >  100.0%    0.0% >  2.6%
            |    4.4%    0.0% |
          5 |       13       8 |    21
            >   61.9%   38.1% > 13.8%
            |   14.3%   13.1% |
------------+------------------+------
      Total |       91      61 |   152
            |   59.9%   40.1% |
```

If you examine the row percentages, you will see that the proportions of cases are similar for **all** values of the MARITAL status variable (in this example they are always larger than the proportion of non-cases). In this case GHQCASE is said to be *independent* of MARITAL status (i.e. GHQCASE and marital status are **not** associated with each other). This observation can be quantified by the chi-square statistic and p-value associated with the table.

Writing-up

What you should quote in reports

If you are using two-by-two tables to analyse your data you should quote the relative risk (or odds ratio) and the associated confidence interval. Remember to specify the confidence level as this can vary. A typical quote might be:

> *'We found professional training to be associated with a score of five or higher on the GHQ-28 scale (RR = 1.78, 95% C.I. 1.31 - 2.41).'*

Some journals may require you to quote a chi-square and p-value. A typical quote might be:

> *'We found professional training to be associated with a score of five or higher on the GHQ-28 scale (Yates chi-square = 15.41, p < 0.0001).'*

Some may want both:

> *'We found professional training to be associated with a score of five or higher on the GHQ-28 scale (RR = 1.78, 95% C.I. 1.31 - 2.41, Yates chi-square = 15.41, p < 0.0001).'*

If you are using contingency tables larger than two-by-two you should quote the chi-square, degrees of freedom, and the associated p-value. A typical quote might be:

> *'We found no association between marital status and a score of five or higher on the GHQ-28 scale (chi-square = 6.24, df = 4, p = 0.18).'*

When quoting p-values it is conventional to quote the calculated value if it is above $p = 0.05$ or the threshold value ($P < 0.05$, $p < 0.01$, $p < 0.001$, $p < 0.0001$) if it is below the threshold value (i.e. for *significant* results).

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

Exercise 5 - Recoding variables and two-by-two tables

Use the RECODE or IF ... THEN commands to recode the ROSENBERG variable into two groups ('HIGH' and 'LOW-NORMAL'). Use to lower three quartiles of the distribution of the ROSENBERG variable for the 'LOW-NORMAL' group and the upper quartile for the 'HIGH' group. You can find the quartiles of the distribution of the ROSENBERG variable using the MEANS or FREQ commands. Make sure that the cut-points you specify with the RECODE or IF … THEN commands are not ambiguous (i.e. ensure the two groups do not overlap).

Use the LIST or BROWSE command to check that the RECODE or IF ... THEN commands you used have worked correctly.

Use the TABLES command to investigate the association between PROFESSION and your new variable.

Use the TABLES command to investigate the association between AGE, SEX, MARITAL, and ETHNIC and your new variable.

Tables with subsets of data

Selecting cases for ANALYSIS

The second question of interest to the researchers in the **Stress and the Student Social Worker** study was 'To determine whether students undertaking professional training in social work exhibited higher stress levels than those undertaking postgraduate teacher training'. The COHORT variable records which group each individual case belongs to:

```
COHORT      ¦  Freq  Percent   Cum.
------------+----------------------
ACADEMIC    ¦    69   45.4%    45.4%
PGCE        ¦    33   21.7%    67.1%
SOCIAL WORK ¦    50   32.9%   100.0%
------------+----------------------
      Total ¦   152  100.0%
```

Use the SELECT command to instruct analysis to ignore the data from the ACADEMIC students:

```
select cohort <> "ACADEMIC"
```

and use the TABLES command to investigate the association between COHORT and GHQCASE:

```
tables cohort ghqcase
```

which produces the following two-by-two table:

```
                          GHQCASE
COHORT      |CASE     NOT CASE  | Total
------------+-------------------+------
PGCE        |       30        3 |    33
SOCIAL WORK |       32       18 |    50
------------+-------------------+------
      Total |       62       21 |    83

                    Single Table Analysis
Odds ratio                                                  5.63
Cornfield 95% confidence limits for OR   1.34* < OR < 27.19*
                     *May be inaccurate

Relative risk of (GHQCASE=CASE) for (COHORT=PGCE)       1.42
Greenland, Robins 95% conf. limits for RR   1.12 < RR < 1.80
                (Biometrics 1985;41:55-68)
        Ignore relative risk if case control study.

                         Chi-squares   P-values
                         -----------   --------
        Uncorrected:          7.62     0.00578402 <---
        Mantel-Haenszel:      7.52     0.00608612 <---
        Yates corrected:      6.26     0.01235539 <---
```

Note that the SELECT command has excluded those cases where COHORT = "ACADEMIC" from the analysis (the '< >' operator means 'not equal to ...').

Examine the relative risk and its associated confidence limits. The relative risk is greater than one indicating a higher risk of being a 'case' for PGCE students than for SOCIAL WORK trainees. The confidence interval does not include one indicating that the observed association (or increase in risk) is unlikely to be due to random variation. The p-value for Yates chi-square is less than 0.05 which also indicates that the observed association (or increase in risk) is unlikely to be due to random variation.

The data do **not** support the researchers' hypothesis. PGCE students are more likely to be 'cases' than SOCIAL WORK trainees (RR = 1.42, 95% C.I. 1.12 - 1.80, Yates chi-square = 6.26, $p < 0.05$).

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

Exercise 6 - Recoding, subsets, and two-by-two tables

Use the RECODE or IF ... THEN commands to recode the MEE and MDP variables into two groups ('CASE' and 'NOT CASE'). Use to lower three quartiles of the distributions of these variables for the 'NOT CASE' group and the upper quartile for the 'CASE' group. You can find the quartiles of the distribution of a variable using the MEANS or FREQ commands. Make sure that the cut-points you specify with the RECODE or IF … THEN commands are not ambiguous (i.e. ensure the two groups do not overlap).

Use the LIST or BROWSE command to check that the RECODE or IF ... THEN commands you used have worked correctly.

Use the TABLES command to investigate the association between COHORT and your new variables. Does the evidence support the researchers' hypothesis?

Categorical and continuous data

Recoding and information loss

So far we have been comparing groups and testing hypotheses using categorical data that we derived from continuous data. The problem with this approach is that when we collapse a continuous variable into groups we are losing information. Also, it is not always possible to set sensible *threshold* values that divide the data into groups. Recoding the GHQ score variable was not problematic as it is conventional to define a 'case' as a person scoring five or higher using the GHQ-28 scale. With other data it will not always be so easy to define groups. In such cases you will need to use analytical techniques that deal with continuous data. One technique that can be applied to such data is *analysis of variance* (also called *ANOVA*). This technique is used to test the null hypothesis that the several population means are equal.

Have a look at the mean GHQ score for each category of the COHORT variable by issuing the command:

```
means ghq cohort /n
```

Ignore the lengthy statistical output and concentrate on the summary statistics for each group:

COHORT	Obs	Total	Mean	Variance	Std Dev
PGCE	33	384	11.636	38.614	6.214
SOCIAL WORK	50	422	8.440	42.660	6.531
Difference			3.196		

COHORT	Minimum	25%ile	Median	75%ile	Maximum	Mode
PGCE	0.000	8.000	10.000	17.000	24.000	8.000
SOCIAL WORK	0.000	2.000	8.000	14.000	22.000	0.000

Note that only two groups are shown. This is because of the SELECT statement:

```
select cohort <> "ACADEMIC"
```

we issued earlier. Examine the summary statistics carefully and make sure that you can identify the number of observations, minimum, maximum, mean, standard deviation, median, and mode for each of the two groups. Another way of looking at the same data is to display it as line chart. Issue the command:

```
line ghq cohort
```

and examine the resulting graph. You might like to group the GHQ variable to make the graph easier to read.

The '/n' at the end of the means command:

```
means ghq cohort /n
```

instructs ANALYSIS **not** to display a table of GHQ by COHORT before displaying the summary statistics for each group. This table is often of little use.

Differences in means

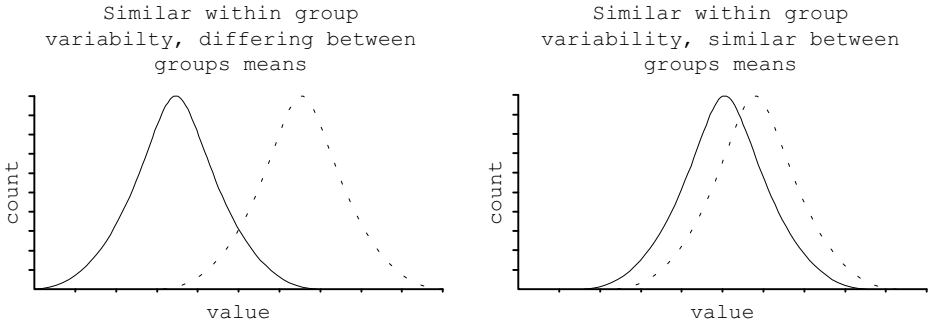
Analysis of variance

With analysis of variance the variability of the data is split into two components:

1. Variability of the data **<u>within</u>** each group around the group mean

2. Variability of the means **<u>between</u>** groups

If the GHQ-28 score doesn't vary much between individuals within the same professional group (e.g. PGCE students yield similar scores to each other and SOCIAL WORK trainees yield similar scores to each other) but the group means differ a great deal, there is evidence that the population means are not equal:



The variability within each group is called the *within-groups sum of squares* and is calculated as:

$$SSW = \sum_{i=1}^{k} (N_i - 1) S_i$$

where S_i is the *variance* of group *i* about its mean, N_i is the number of cases in group *i*, and *k* is the number of groups . For us:

```
SSW = (33 - 1) * 38.614 + (50 - 1) * 42.660 = 3325.988
```

The variability of the group means is measured by the *between-groups sum of squares* and is calculated as:

$$SSB = \sum_{i=1}^{k} N_i (\overline{X_i} - \overline{X})^2$$

where $\overline{X_i}$ is the mean of group *i* and $\overline{X}$ is the mean of the entire sample. For our data:

```
SSB = 33 * (11.636 - 9.711)^2 + 50 * (8.440 - 9.711)^2 = 203.057
```

A test statistic is calculated as the ratio of the *mean squares* of these two numbers. The *means squares* are calculated by dividing each sum of squares figure by its *degrees of freedom.* The *between-groups degrees of freedom* is *k-1* where *k* is the number of groups. The within-groups degrees of freedom is *N - k* where *N* is the number of cases in the entire sample. The ratio of these *mean squares* is called the *F* statistic:

$$F = \frac{between \ \text{groups means square}}{within \ \text{groups means square}} = \frac{203.057}{41.062} = 4.945$$

The *p-value* can be obtained by comparing the calculated *F* statistic against the *F* distribution (at *k - 1* and *N - k* degrees of freedom) in a set of statistical tables. Fortunately ANALYSIS calculates the *F* statistic and the associated p-value for us.

ANOVA - worked example

A worked example of ANOVA calculations

The table below shows twenty values of the GHQ for the PGCE cohort and eighteen values of GHQ for the SOCIAL WORK cohort:

PGCE	5	6	8	8	9	10	10	11	12	13
	14	15	17	18	18	20	20	21	22	24
SOCIAL WORK	2	2	2	4	4	4	5	6	7	7
	8	9	10	13	13	14	14	20		

The means and variances for this data are:

Group	Mean	Variance
BOTH	11.184	37.938
PGCE	14.050	32.366
SOCIAL WORK	8.000	26.000

The within groups sum of squares is:

```
SSW = (20 - 1) * 32.366 + (18 - 1) * 26.000 = 1056.954
```

The within group degrees of freedom is:

```
20 + 18 - 2 = 36
```

The within group means square is:

```
1056.954 / 36 = 29.360
```

The between groups sum of squares is:

$$SSB = 20 * (14.050 - 11.184)^2 + 18 * (8.000 - 11.184)^2 = 347.761$$

The between group degrees of freedom is:

```
2 - 1 = 1
```

The between groups mean square is:

```
347.761 / 1 = 347.761
```

The ratio of the between groups mean square and the within groups mean square is the *F* statistic:

```
F = 347.761 / 29.360 = 11.811
```

The *observed significance value* or *p-value* can be obtained by comparing the calculated *F* statistic against an *F* distribution (at $k - 1$ and $N - k$ degrees of freedom) in a set of statistical tables - in this example $p < 0.005$.

ANOVA in ANALYSIS

An example of ANALYSIS output

The command:

```
means ghq cohort /n
```

produces the following summary statistics and *ANOVA table*:

```
COHORT            Obs      Total       Mean   Variance    Std Dev
PGCE               33        384     11.636     38.614      6.214
SOCIAL WORK        50        422      8.440     42.660      6.531
Difference                             3.196

COHORT         Minimum     25%ile     Median     75%ile    Maximum       Mode
PGCE             0.000      8.000     10.000     17.000     24.000      8.000
SOCIAL WORK      0.000      2.000      8.000     14.000     22.000      0.000

                                ANOVA
                    (For normally distributed data only)
          The p value is equivalent to that for the Student's T Test,
                       since there are only 2 samples.

Variation         SS  df       MS  F statistic    p-value
Between      203.104   1  203.104        4.946   0.027183  <- F-Statistic / p-value
Within      3325.956  81   41.061
Total       3529.060  82
```

In this case the p-value is less that 0.05 which is evidence that the observed difference in the means of the two groups is unlikely to have arisen by chance. This is the same as saying that the PGCE students yielded significantly higher GHQ-28 scores than the SOCIAL WORK Students.

Issue the command:

```
means ghq cohort /n
```

and examine the output carefully making sure you can identify the *F*-statistic and its associated p-value.

What you should quote in reports

If you are using ANOVA techniques to analyse your data you should quote the means for each group together with the *F*-statistic, degrees of freedom, and p-value. A typical quote might be:

> *'The PGCE group had a higher mean GHQ-28 score than the social work trainee group (PGCE mean = 11.636, Social Work mean = 8.440, F = 4.946, df = 1, p < 0.05)'*

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

Exercise 7 - Analysis of variance (ANOVA)

Use the MEANS command to test the null hypothesis that the means of the GHQ, ROSENBERG, MEE, MDP, and MPA variables for PGCE and SOCIAL WORK Students are the same.

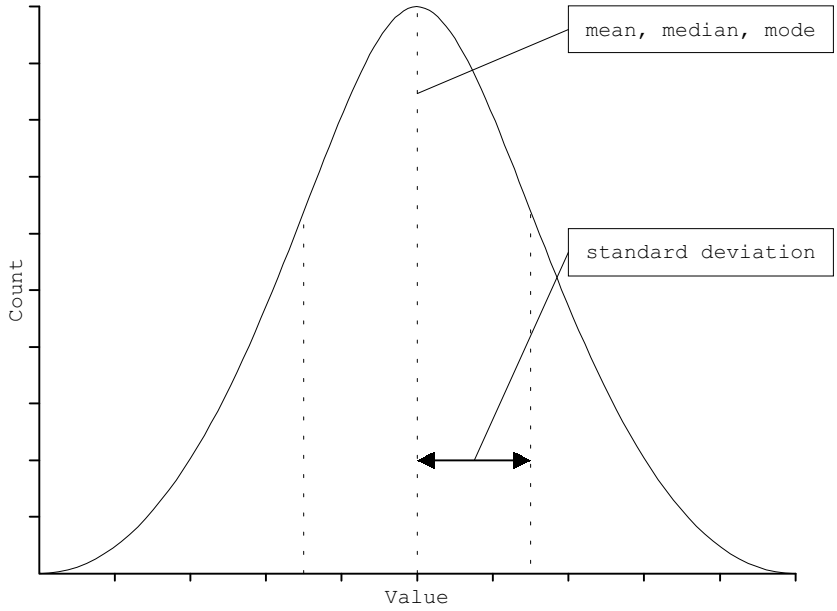
Problems with ANOVA

What ANOVA assumes about your data

ANOVA is a fairly *robust* statistical test. This means that the test can be applied to a wide range of data without it reporting misleading values. ANOVA does, however, make several *assumptions* about data. These are:

1. The groups are *independent* of each other. This means that the groups should be mutually exclusive. In our data this means that we should **not** include the same individual in the PGCE group and the SOCIAL WORK group. You should **not** use ANOVA techniques with repeated measurements from the same subjects.

2. The variable being studied should be (more-or-less) *normally* distributed. A *normal distribution* is a symmetrical bell-shaped distribution where the mean, median, and mode are similar to each other:



3. The variable being studied should have similar variances (or standard deviations) for each group.

If **any** of these assumptions are violated then the ANOVA technique may produce *unreliable* results. You should always test these assumptions before placing any trust in the results of any ANOVA based analysis.

Testing the ANOVA assumptions

Examining the output

The following output was produced using the command MEANS GHQ COHORT:

MEANS of GHQ for each category of COHORT

COHORT	Obs	Total	**Mean**	**Variance**	Std Dev
PGCE	33	384	**11.636**	**38.614**	6.214
SOCIAL WORK	50	422	**8.440**	**42.660**	6.531
Difference			**3.196**		

COHORT	Minimum	25%ile	**Median**	75%ile	Maximum	**Mode**
PGCE	0.000	8.000	**10.000**	17.000	24.000	**8.000**
SOCIAL WORK	0.000	2.000	**8.000**	14.000	22.000	**0.000**

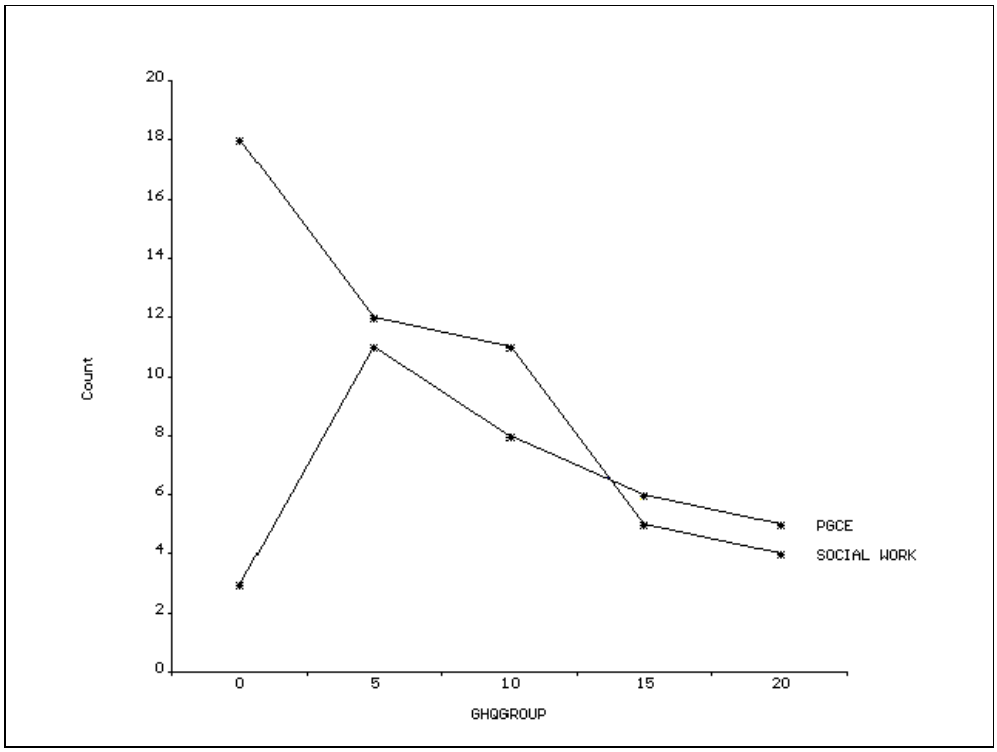
We can examine the assumption of *normality* by comparing the mean, mode and median within each group. If the data is *normally* distributed we would expect the mean, median, and mode to take similar values. In our data this is **not true** for the SOCIAL WORK group. The assumption that the variances are equal can be tested by comparing the values in the variance column of the two groups. In our data the variances are quite similar. There is a *formal* statistical test that tests the hypothesis that the variances are equal. This is *Bartlett's Test* and is displayed below the ANOVA table by ANALYSIS:

Bartlett's test for homogeneity of variance
Bartlett's chi square = 0.094 deg freedom = 1 p-value = 0.758942

The variances are homogeneous with 95% confidence.
If samples are also normally distributed, ANOVA results can be used.

Bartlett's test confirms that the variances are similar enough for us to use the ANOVA technique.

You can also test the assumption of normality by examining a HISTOGRAM or LINE chart of the distribution of the variable under study. Here the GHQ variable has been grouped to better show the underlying distribution:



This also shows that the GHQ data in the both groups is **not** normally distributed. This means that it may not be valid to use the ANOVA technique to analyse our data.

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

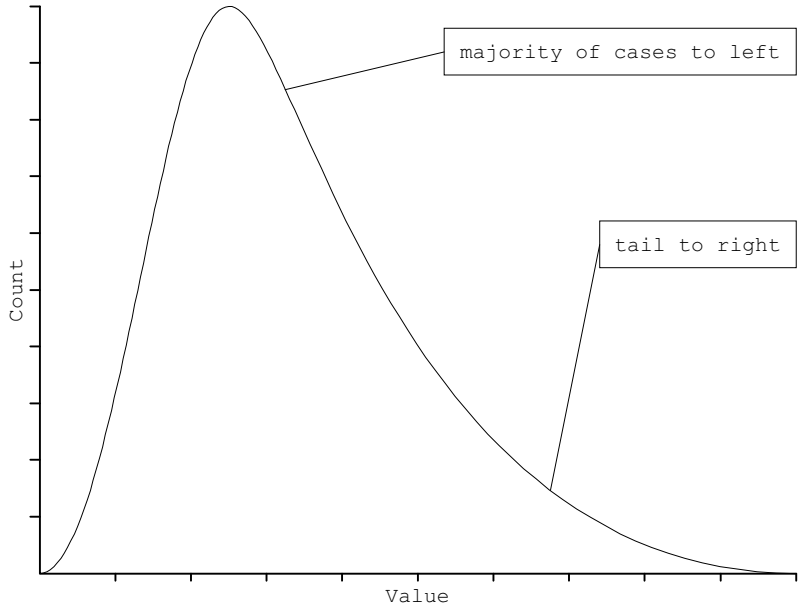
Exercise 8 - Testing the ANOVA assumptions

Use the MEANS command to test the ANOVA assumptions for ROSENBERG, MEE, MDP, and MPA grouped for PGCE and SOCIAL WORK Students. Test the ANOVA assumption using the comparison of means, medians, modes, and standard deviations and with LINE charts or HISTOGRAMs. You may need to group these variables before producing LINE charts and HISTOGRAMs.

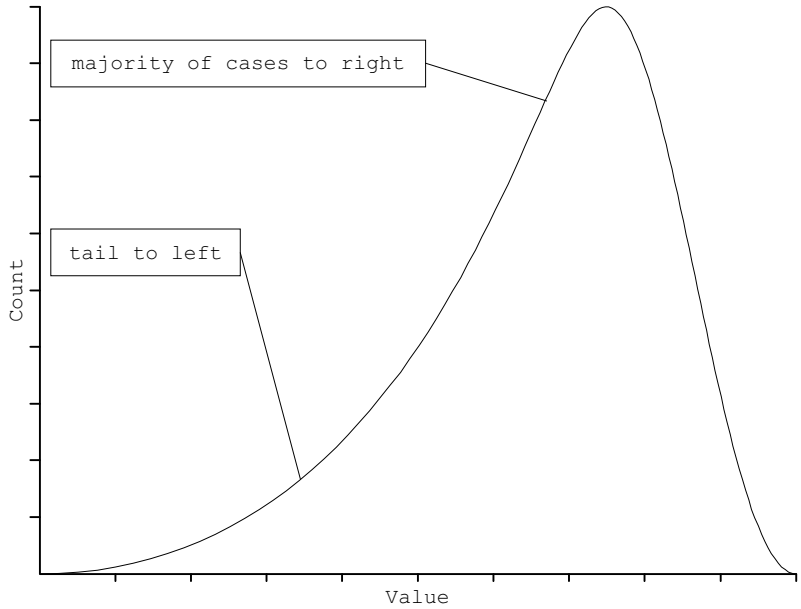
When ANOVA is unsuitable

It's a common problem

The assumption of normality is frequently *violated*. This is often a problem with data which are based on scores or counts. This data tends to be distributed so that most individuals have low scores, counts, or values and a few individuals have high scores, counts, or values. This type of distribution is often part of the design of an instrument such as GHQ-28. Such distributions are called *skewed distributions*. A distribution with many individuals with low values is called a *positively skewed* or *skewed to the right* distribution:



and is typical of data collected using instruments such as GHQ-28 and biological measures such as triceps skin fold. A distribution with many individuals with high values is called a *negatively skewed* or *skewed to the left* distribution:



and is typical of biological measures such as gestation period. Both of these distribution types are common. With these types of distribution, the mean and standard deviation do **not** describe the population well and statistical techniques that rely on them (such as ANOVA) are **not** reliable.

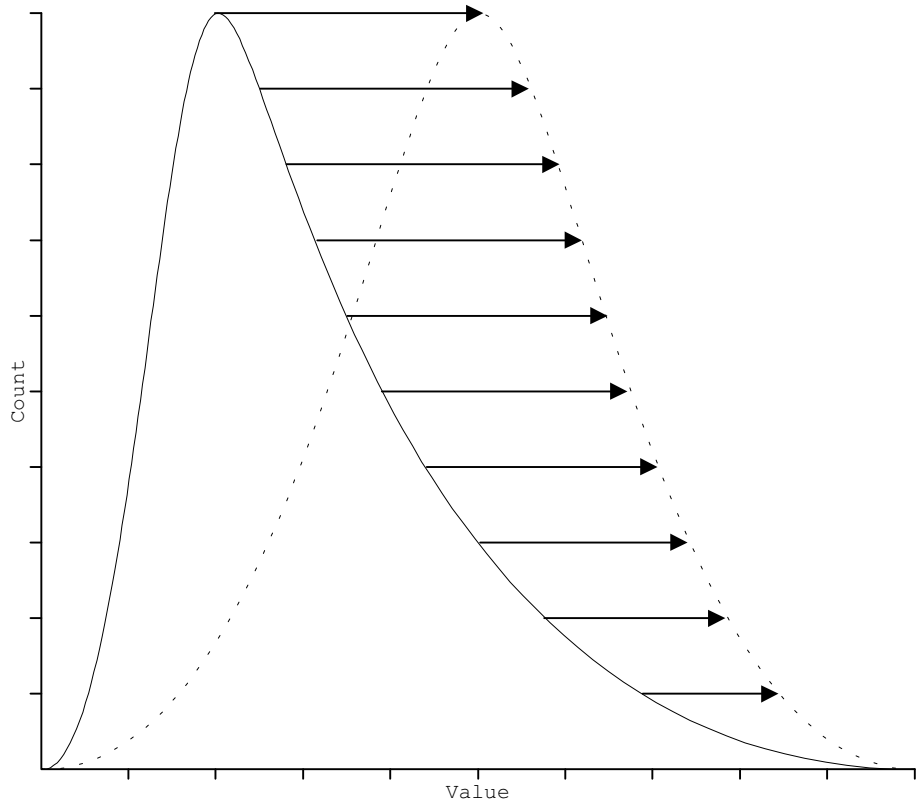
What can be done?

Transforming the data

If the data do not fit the ANOVA assumptions we could attempt to change the data to make it more suitable. Changing the data in this way is called *transforming* the data and it is achieved by applying a simple mathematical formula to each data point. The table below shows the most useful transformations that can be applied to data that do not fit the ANOVA assumptions:

Problem	Severity / Nature	Transformation	Formula
Positively skewed	severe	reciprocal	$-1/x$
	moderate	logarithmic	$log(x)$
	slight	square root	$\sqrt{x}$
Negatively skewed	severe	cubic	x^3
	moderate	square	x^2
Unequal variances	proportional to mean	logarithmic	$log(x)$
	proportional to the square of the mean	reciprocal	$-1/x$
	proportional to the square root of the mean	square root	$\sqrt{x}$

For example, with a moderate positive (right) skew:



a transformation of the variable (such as a logarithmic transformation) will move the higher data values less than the lower data values bringing the distribution closer to the normal distribution.

The most commonly used transformation is the *logarithmic* transformation. Always try this transformation first as it will solve the majority of problems associated with non-normal data.

When applying transformations you should note that you cannot take logarithms of negative numbers or zero, you cannot take a square root of a negative number, and you cannot take the reciprocal of zero. If your data contains negative or zero values then you may need to add a constant to each observation before transforming the data.

Checking the data

Checking the data for skew and unequal variance

Issue the command:

```
means mdp cohort /n
```

and examine the output carefully. The mean, median, and mode for each group are quite different from each other:

COHORT	Obs	Total	**Mean**	Variance	**Std Dev**
PGCE	33	155	**4.697**	11.468	**3.386**
SOCIAL WORK	50	349	**6.980**	32.224	**5.677**
Difference			-2.283		

COHORT	Minimum	25%ile	**Median**	75%ile	Maximum	**Mode**
PGCE	0.000	2.000	**4.000**	7.000	13.000	**2.000**
SOCIAL WORK	0.000	3.000	**5.000**	10.000	28.000	**4.000**

The standard deviations in the two groups are also very different from each other. This is confirmed by Bartlett's test:

```
            Bartlett's test for homogeneity of variance
Bartlett's chi square =   9.155  deg freedom =  1   p-value = 0.002480

    Bartlett's Test shows the variances in the samples to differ.
         Use non-parametric results below rather than ANOVA.
```

Produce a HISTOGRAM of the distribution of the MDP variable:

```
histogram mdp
```

You can see that this variable has a *positively skewed* distribution.

Produce a LINE plot of the MDP variable for each value of the COHORT variable:

```
line mdp cohort
```

You can see that the SOCIAL WORK group has a wider range of values than the PGCE group and that both distributions are positively skewed.

Despite the problems of skew and unequal variance with this data, it may still be possible to use ANOVA techniques after applying a logarithmic transformation.

Transforming the data

Applying a logarithmic transformation

Create a new variable called LOGMDP to hold the logarithm of the MDP variable:

```
define logmdp #.####
```

and assign a value to this variable:

```
logmdp = log(mdp + 1)
```

The variable LOGMDP now contains the logarithm of the MDP variable. We need to add 1 to MDP so that we do not try to take the logarithm of zero (a non-existent number). We would need to do the same if we were to use a *reciprocal* transformation as one divided by zero is an infinitely large number. Examine the distribution of the LOGMDP variable:

```
histogram logmdp
line logmdp cohort
```

The distribution is now more normal. Issue the command:

```
means logmdp cohort /n
```

and examine the output carefully. The mean, median, and mode for each group are now more similar:

```
COHORT            Obs      Total       Mean   Variance    Std Dev
PGCE               33         22      0.669      0.088      0.297
SOCIAL WORK        50         40      0.794      0.106      0.326
Difference                           -0.125

COHORT         Minimum     25%ile     Median     75%ile    Maximum       Mode
PGCE             0.000      0.477      0.699      0.903      1.146      0.477
SOCIAL WORK      0.000      0.602      0.778      1.041      1.462      0.699
```

The standard deviations in the two groups are also similar. This is confirmed by Bartlett's test:

```
            Bartlett's test for homogeneity of variance
Bartlett's chi square =   0.321  deg freedom =  1   p-value = 0.571246

         The variances are homogeneous with 95% confidence.
 If samples are also normally distributed, ANOVA results can be used.
```

Despite the original problems of skew and unequal variance with this data, ANOVA techniques may still be used after applying a logarithmic transformation.

ANOVA and transformed data

What is the antilogarithm of the mean of the logarithms?

Now that we have transformed the data to meet the assumptions of the ANOVA technique we can use this new transformed data to test our hypothesis. Issue the command:

```
means logmdp cohort /n
```

which produces the following output:

```
                MEANS of LOGMDP for each category of COHORT

COHORT            Obs      Total       Mean   Variance    Std Dev
PGCE               33         22      0.669      0.088      0.297
SOCIAL WORK        50         40      0.794      0.106      0.326
Difference                           -0.125

COHORT        Minimum     25%ile     Median     75%ile    Maximum       Mode
PGCE            0.000      0.477      0.699      0.903      1.146      0.477
SOCIAL WORK     0.000      0.602      0.778      1.041      1.462      0.699

                              ANOVA
                    (For normally distributed data only)
          The p value is equivalent to that for the Student's T Test,
                      since there are only 2 samples.

Variation          SS   df          MS  F statistic    p-value
Between         0.309    1       0.309        3.119   0.077453
Within          8.012   81       0.099
Total           8.320   82
```

We are working with the *logarithmically* transformed data. The summary measures are for this data. To get back to sensible values we need to use a calculator (or a set of 'log' tables) to calculate the antilogarithms of the quoted values. For the PGCE group the 'average' MDP score is $10^{0.669} = 4.667$ and for the SOCIAL WORK group it is $10^{0.794} = 6.223$. These values are different from the arithmetic mean of the untransformed data. They are known as *geometric means*. If you transform your data in this way make sure that you explicitly state that you are using *geometric means* in reports and papers.

What you should quote in reports

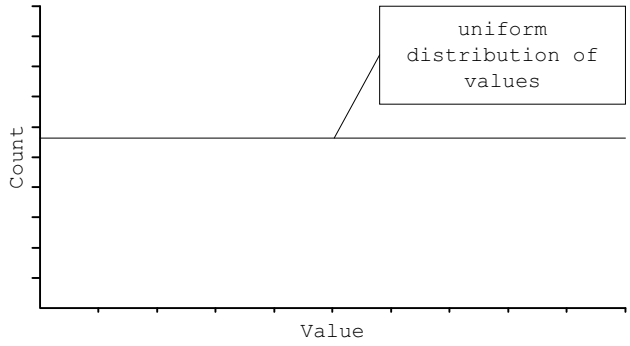
If you are using ANOVA techniques to analyse your data and have applied a logarithmic transformation to your data you should quote the geometric means for each group together with the *F*-statistic, degrees of freedom, and p-value. A typical quote might be:

> *'The social work trainees and PGCE students had similar geometric mean MDP scores (Social Work = 6.223, PGCE = 4.667, F = 3.119, df = 1, p = 0.077453)'*

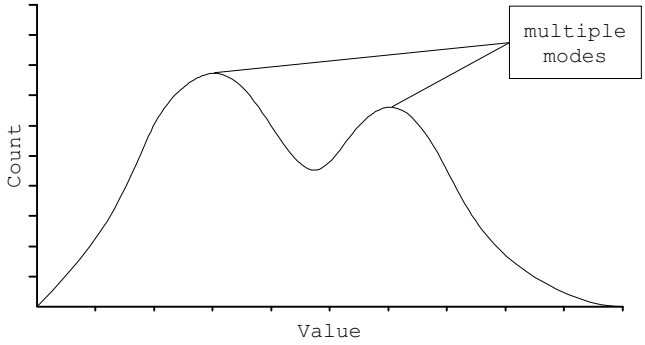
Can't transform! Won't transform!

Difficult distributions

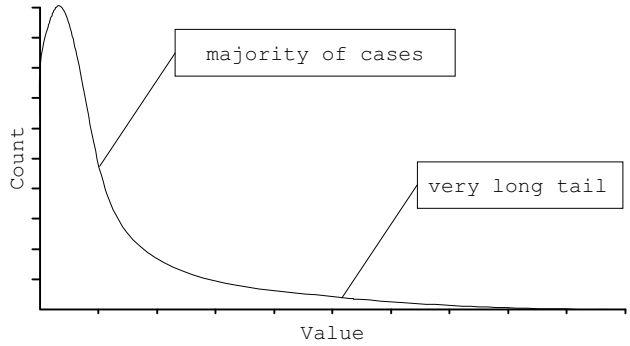
Sometime when you try to transform your data it just won't approach normality. This is often the case with distributions such as *uniform* distributions:



which are common in disease surveillance (i.e. a disease with no seasonal variation), *multimodal* distributions:



which are common biological distributions or when two groups are being (inappropriately) analysed together (e.g. hormone levels in mixed sex groups), or *J-Shaped* (very heavily skewed) distributions:



which are typical of survival times (e.g. AIDS diagnosis progressing to death) or counts (e.g. of parasites). A right skew is shown here but *J-shaped distributions* may also be skewed to the left. These extreme degrees of skew may be difficult to correct using transformations.

ANOVA techniques do **not** produce reliable results with these sorts of distribution because the *parameters* used in its calculation (means and variances) do **not** accurately reflect the distribution of the data.

Another problem with ANOVA

ANOVA and small datasets or small group sizes

All statistical tests require an adequate sample size if results are to be *generalised* from sample to population. ANOVA techniques require an adequate sample size in each *strata*. This means that there must be a *sufficient number* of cases in both the PGCE and SOCIAL WORK groups. A sufficient number is usually taken to mean more than twenty (20) cases in this context but this really depends on the magnitude of difference that it is important to be able to detect. This is the final assumption that needs to be met if ANOVA results are to be trusted. Data that do not meet this assumption are called *sparse data*.

Non-parametric statistics

Non-parametric statistics are a set of statistical methods that make fewer assumptions about the distribution of data. They are called *non-parametric* because they do not assume that data can be described by a small number of *parameters* such as the mean and standard deviation which completely describes the normal distribution. They may also be used reliably with small sample / group sizes.

There are problems with using non-parametric techniques. Often, no parameters (e.g. mean and standard deviation) are estimated so it can be difficult to estimate the strength and direction of an effect or calculate confidence intervals. There are no reliable methods for calculating sample sizes for use with non-parametric statistics. Non-parametric statistics are also *conservative*. This means that they may falsely attribute small (but real) differences in the data to random variation.

Parametric techniques (such as ANOVA) are preferred to non-parametric techniques when the data meets the required assumptions.

Non-parametric statistics In ANALYSIS

ANALYSIS performs two non-parametric tests as part of the MEANS command. These are the *Wilcoxon Rank-Sum Test* (a.k.a. *Mann-Whitney U-Test*) and the *Krustall-Wallis Test*. Which test is performed depends on the number of groups being analysed. The Wilcoxon Rank-Sum Test is performed for two groups and the Kruskal-Wallis Test if there are more than two groups. Both test the hypothesis that the sample **medians** are equal.

Differences in medians

A worked example of the Wilcoxon Rank-Sum Test

This section presents a worked example of how the Wilcoxon Rank-Sum Test is calculated. As demonstration data we will use GHQ scores from five SOCIAL WORK cases and six PGCE cases. These data are shown below:

Cohort	GHQ	Cohort	GHQ
SOCIAL WORK	2	PGCE	24
SOCIAL WORK	17	PGCE	9
SOCIAL WORK	21	PGCE	18
SOCIAL WORK	13	PGCE	13
SOCIAL WORK	22	PGCE	6
		PGCE	1

The data from both groups is combined, sorted, and given a rank (1 for the lowest value, 2 for the next lowest, and so on). It there are any ties (more than one value the same) the average rank is assigned to each observation:

Social Work	Ranks		PGCE
		1.0	1
2	2.0		
		3.0	6
		4.0	9
13	5.5	5.5	13
17	7.0		
		8.0	18
21	9.0		
22	10.0		
		11.0	24
	33.5		

The ranks of the observations in the group with the smallest sample size are summed to produce a statistic called the *U-Statistic* or the *sum of the ranks*:

$$U = 2 + 5.5 + 7 + 9 + 10 = 33.5$$

On the basis of the *null hypothesis* that the distributions of ranks in both groups are the same, the distribution of the *U* statistic can be calculated and a p-value for the test can be derived from statistical tables. Fortunately ANALYSIS performs all of these calculations for us.

Note that with groups that have an even number of cases the median is calculated as the mean of the middle pair of cases. In this example, the median GHQ score for the PGCE group is:

$$(9 + 13) / 2 = 11$$

Non-parametric tests in ANALYSIS

An example of ANALYSIS output

The following table was produced using the command MEANS GHQ COHORT /N. Important information has been highlighted:

```
COHORT            Obs      Total       Mean   Variance    Std Dev
PGCE               33        384     11.636     38.614      6.214
SOCIAL WORK        50        422      8.440     42.660      6.531
Difference                             3.196

COHORT        Minimum     25%ile     Median     75%ile    Maximum       Mode
PGCE            0.000      8.000     10.000     17.000     24.000      8.000
SOCIAL WORK     0.000      2.000      8.000     14.000     22.000      0.000

                                ANOVA
                    (For normally distributed data only)
          The p value is equivalent to that for the Student's T Test,
                        since there are only 2 samples.

Variation          SS   df          MS  F statistic    p-value
Between       203.104    1     203.104        4.946   0.027183
Within       3325.956   81      41.061
Total        3529.060   82

                Bartlett's test for homogeneity of variance
    Bartlett's chi square =   0.094  deg freedom =  1   p-value = 0.758942

              The variances are homogeneous with 95% confidence.
     If samples are also normally distributed, ANOVA results can be used.


Mann-Whitney or Wilcoxon Two-Sample Test (Kruskal-Wallis test for two groups)

Kruskal-Wallis H (equivalent to Chi square) =       4.882 <- Test Statistic
                         Degrees of freedom =           1
                                    p value =    0.027140 <- p-value
```

What you should quote in reports

If you are using non-parametric techniques such as the Wilcoxon Rank-Sum test to analyse your data you should quote the median for each group together with the name of the test used, the test statistic (e.g. *U* or Kruskal-Wallis H), the degrees of freedom, and p-value. A typical quote might be:

> *'The PGCE group had higher GHQ-28 scores than the social work trainee group (Kruskal-Wallis H, PGCE median = 10.000, Social Work median = 8.000, chi-square = 4.882, df = 1, $p < 0.05$)'*

Non-parametric tests and group size

Non-parametric techniques are better at dealing with sparse data than parametric techniques such as ANOVA. The non-parametric tests presented here will work reliably with group sizes of five or more.

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

Exercise 9 - Doing it!

The two hypotheses behind the **Stress and the Student Social Worker Study**were:

1. Students undertaking professional training exhibit higher stress levels than undergraduate students.

2. Students undertaking professional training in social work exhibit higher stress levels than those undertaking postgraduate teacher training.

Use the techniques and commands presented thus far to test these hypotheses.

Two continuous variables

Working with different variable types

We have used techniques to analyse data of two different types:

Type	Description
Categorical data	Cases belong to a defined *category* or group such as MARITAL status, SEX, and LIVING situation. A particularly common type of categorical variable is a *binary categorical* variable where cases can belong to one of two alternative groups. The variables GHQCASE, which we created earlier, and SEX are examples of *binary categorical* variables in the **Stress and the Student Social Worker** dataset.
Count or *continuous* data	Unmodified numerical measures. There are two types of continuous data. *Discrete* continuous data is limited to whole numbers or *integers*. A count of cases over time would be a *discrete* variable. Discrete variables in the **Stress and the Student Social worker** dataset are GHQ, ROSENBERG, MEE, MDP, and MPA. True *continuous* data is measured on a continuous scale. Examples are height, weight, and blood pressure.

The statistical technique we use to analyse the data depends upon the types of data that we need to analyse:

1st Variable	2nd Variable	Technique
Categorical	Categorical	Relative risks and confidence intervals (two-by-two tables) Odds ratios and confidence intervals (two-by-two tables) Chi-square statistic and p-values (contingency tables) Fisher's Exact Test (two-by-two tables)
Continuous	Categorical	ANOVA (normally distributed data only) Wilcoxon rank-sum test (non-normal or sparse data) Kruskal-Wallis test (non-normal or sparse data)

As you would expect, there are statistical techniques for analysing data that is stored as two continuous variables. These techniques are known as *correlation* and *regression*.

Why use correlation and regression analysis?

Instead of worrying about new techniques we could stick to what we already know. When faced with the need to analyse two continuous variables we could recode one or both of the variables into groups and perform a contingency table analysis or use ANOVA (or non-parametric equivalent) techniques. The problems with this approach are:

1. When we recode data we are losing a lot of information. Recoded data often hides the variability of the data and can lead us into making unwarranted assertions about our data.

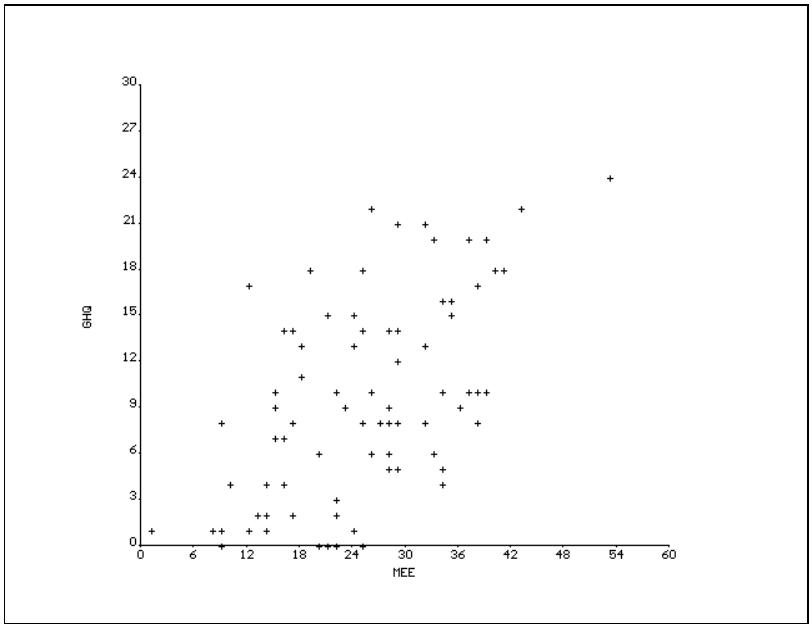
2. The choice of cut-points for group membership is often arbitrary. The choice of cut-point can make a difference to the results of any analysis. Variables such as GHQ, where it is valid to choose a cut-point of five, are rare - the GHQ-28 score was developed over a long period to ensure that this cut-point was not arbitrary.

To avoid these problems it is best to use techniques that can cope with the raw continuous data.

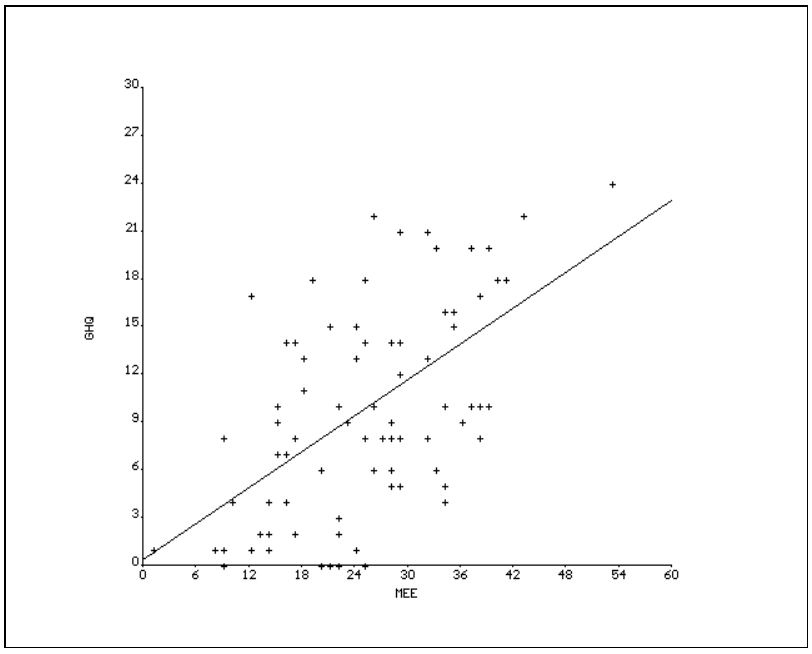
Examining associations

Using scatter plots to examine associations

The chart below shows a scatter plot of the emotional exhaustion (MEE) variable and the GHQ-28 score (GHQ) variable for each case:



There appears to be an association between emotional exhaustion (MEE) and the GHQ-28 score. It is a *positive association* because as MEE increases so does GHQ. The relationship is also called *linear* because the observed points cluster (more or less) around a straight line:



A scatter plot is the first step in studying the association between two continuous variables.

The strength of an association

Assessing the strength of an association

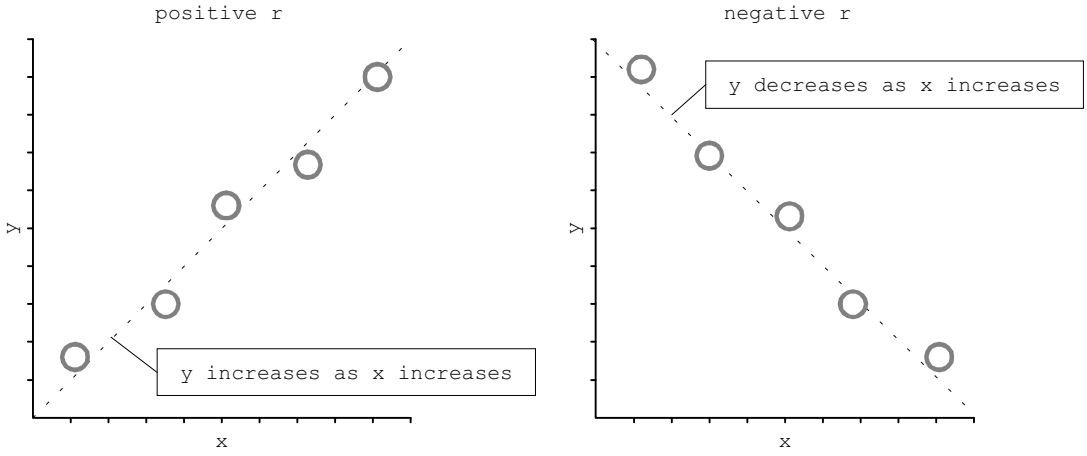
The strength of an association can be measured by calculating the *Pearson correlation coefficient*:

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{(N-1)S_X S_Y}$$

where N is the number of cases and S_X and S_Y are the standard deviations of the two variables. The absolute value of r is an indication of the strength of the *linear relationship*.

The largest possible *absolute* value (the value regardless of sign) of r is one (1). This only occurs when all points fall upon a straight line.

When the line has a positive slope (i.e. y increases as x increases) the value of r is positive and when the line has a negative slope (i.e. y decreases as x increases) the value of r is negative:



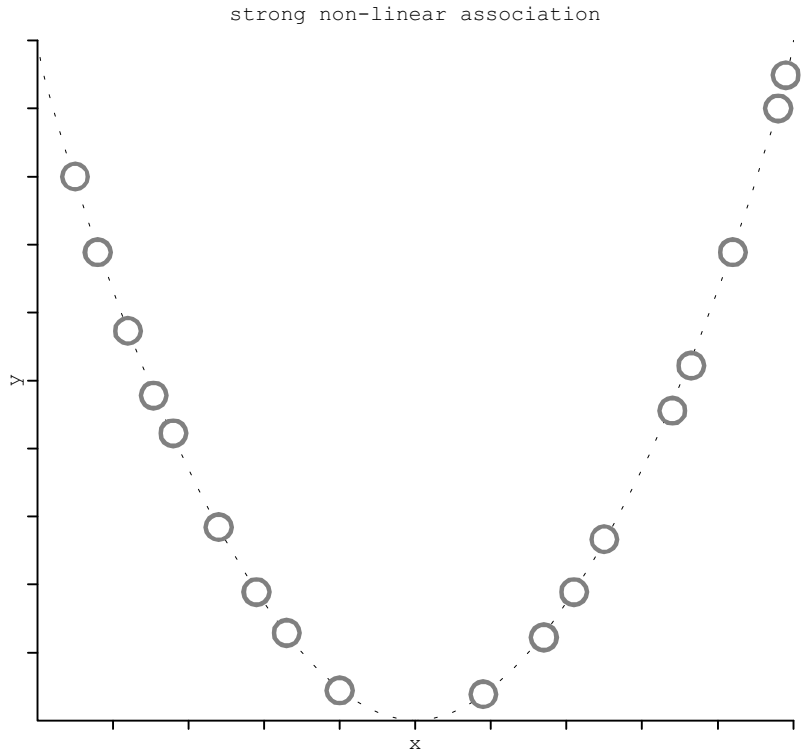
A value of zero (0) indicates no *linear* association.

Note that the calculation of r uses means and standard deviations and therefore requires that the distribution of both variables are approximately normal. You may need to transform one or both variables to meet this assumption (see pages 78 - 79).

Non-linear associations

The correlation coefficient and non-linear associations

You should note that Pearson correlation coefficient measures the strength of a *linear* association. It is possible to have a very strong association between two variables but a very small correlation coefficient. The chart below shows such a relationship:



which yields a small Pearson correlation coefficient (r = 0.2). In this case it is not appropriate to measure the strength of association using the Pearson correlation coefficient - other techniques (which are not covered in this book) must be used. The Pearson correlation coefficient must only be used to assess the strength of a *linear* (as in **straight line**) association.

Because the correlation coefficient only tests the strength of a linear association, it is important to examine the nature of the association using scatter plots before calculating the correlation coefficient. If a strong association is observed but it is non-linear (i.e. cannot be well summarised using a straight line) then it may be possible to transform one or both of the variables to arrive at a more linear (straight line) association.

The calculation of the Pearson correlation coefficient is complicated and best left to purpose designed computer programs such as ANALYSIS.

Transformation and non-linear associations

Why transform data?

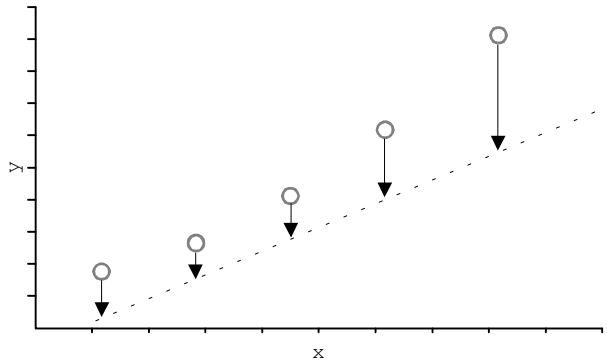
Transformation can also be useful when using correlation and regression analysis. You would apply a transformation to one or both variables in order to:

1. Make a skewed distribution more symmetrical (i.e. make a skewed variable more normal).
2. Make the variances (standard deviations) similar.
3. Make an association more linear.

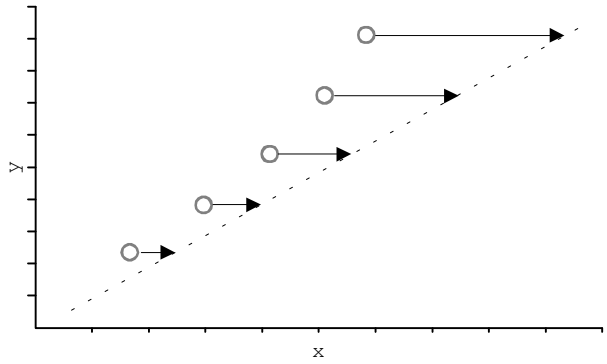
The first two applications have already been covered (see pages 64 - 69). All applications are important but, with correlation and regression analysis, the goal of making an association more linear is the most important.

How transformation can straighten an association

You can usually make an association more linear by applying a transformation to one of the variables. In this situation:



a transformation of the y-variable (such as a logarithmic or square-root transformation) will move the higher *y* data values more than the lower y data values bringing the *y* data values closer to a straight line. A similar effect might be obtained by applying a transformation to the x-variable. In this situation:



a transformation of the x-variable (such as squaring the x-variable) will move the higher *x* data values more than the lower *x* data values bringing the *x* data values closer to a straight line.

Which variable? Which Transformation?

Which variable should be transformed?

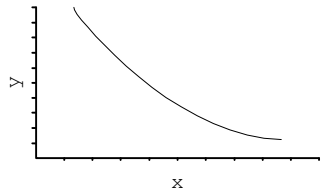
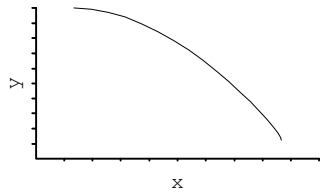
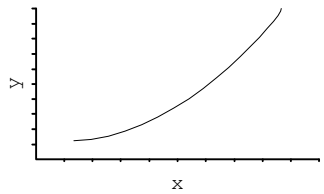
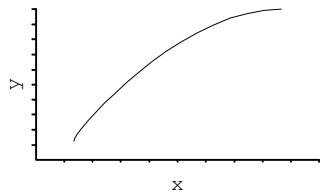
It is not particularly important which variable you choose to transform. Your choice should be guided by the three goals of transformation:

1. Make a skewed distribution more symmetrical (i.e. make a skewed variable more normal).
2. Make the variances (standard deviations) similar.
3. Make an association more linear.

Sometimes you will need to transform only one variable (either x or y) or both variables to achieve these goals. You will need to experiment with different transformations of the two variables to achieve these goals. You should bear in mind that the third aim (to make an association more linear) can only be performed on associations where one variable increases (or decreases) as the other variable increases. In situations where a variable first decreases (or increases) and then increases (or decreases), it will not be possible to make the association linear.

Guidelines for transformation

There are four types of non-linear associations that respond well to transformation. These are:

Association Type	Transform *X* Only	Transform *Y* Only
	$\sqrt{x}$ $\log(x)$ $-1/x$ $-1/\sqrt{x}$	$\sqrt{y}$ $\log(y)$ $-1/y$ $-1/\sqrt{y}$
	x^2 x^3	y^2 y^3
	x^2 x^3	$\sqrt{y}$ $\log(y)$ $-1/y$ $-1/\sqrt{y}$
	$\sqrt{x}$ $\log(x)$ $-1/x$ $-1/\sqrt{x}$	y^2 y^3

The sample transformations may not apply if both the *x-variable* and the *y-variable* are to be transformed.

Scatter plots and correlation coefficients

Producing scatter plots and correlation coefficients in ANALYSIS

Start ANALYSIS again and instruct ANALYSIS to read the SSSW dataset by typing:

```
read sssw.rec
```

Examine the relationship between the GHQ and MEE variables with the command:

```
scatter mee ghq
```

You can instruct ANALYSIS to draw the *best-fit* straight line by specifying the '/r' option to the SCATTER command:

```
scatter mee ghq /r
```

You can instruct ANALYSIS to calculate the Pearson correlation coefficient with the command:

```
regress ghq = mee
```

ANALYSIS displays the Pearson correlation coefficient and 95% confidence limits:

```
Correlation coefficient: r  = 0.58             <- correlation coefficient
                         r^2= 0.33
95% confidence limits:   0.41 < R < 0.71       <- confidence limits
```

The interpretation of the confidence limits is similar to using confidence limits for relative risks and odds ratios. Remember that *r* must lie between -1 and +1 and that a value of *r* = 0 indicates no linear association.

You may also have noticed that ANALYSIS calculated another measure called *r^2* (*r-squared*). This is the square of the correlation coefficient and is called the *coefficient of determination*: This is a measure of how well the data corresponds to the best-fit straight line:

```
Correlation coefficient: r  = 0.58
                         r^2= 0.33             <- coefficient of determination
95% confidence limits:   0.41 < R < 0.71
```

It is sometimes interpreted as how much of the variability in the *y-variable* (e.g. GHQ) can be 'explained' by the variability in the *x-variable* (e.g. MEE) and is expressed as a proportion (e.g. r^2 = 0.33 or 33%).

Note that the order of the variables for the SCATTER and REGRESS commands are reversed. The SCATTER command uses the form:

```
scatter x-variable y-variable
```

and the REGRESS command uses the form:

```
regress y-variable = x-variable
```

In correlation and regression analysis the *y-variable* is called the *response* or *dependant* variable (this is analogous to the *outcome* variable in two-by-two table analysis) and the *x-variable* is called the *explanatory* or *independent* variable (this is analogous to the *exposure* variable in two-by-two table analysis). Before using these techniques you should be clear which variable is dependant and which is independent. The choice does not effect the correlation coefficient but does effect other regression statistics. In practice, it may be difficult to decide which is which. In this example, we are assuming that the GHQ-28 score is dependent upon (explained by) the level of emotional exhaustion.

Interpreting the correlation coefficient

What different values of the correlation coefficient mean

Interpretation of the correlation coefficient is not straightforward. Strengths of linear association vary depending on the field of study. What is considered a strong association in one field may be considered weak in another. The only really useful guide to interpreting the correlation coefficient is a thorough literature review of the field of study.

In general terms, a correlation coefficient that is close to zero indicates no linear association and a correlation coefficient that is close to one (or minus one) indicates a linear association. The table below shows general guideline values for assessing the strength of a linear association but it must be stressed that the correct interpretation depends upon the field of study:

Range of correlation coefficient	Interpretion
-1.0 to -0.8	strong negative linear association
-0.8 to -0.5	moderate negative linear association
-0.5 to -0.3	weak negative linear association
-0.3 to +0.3	no linear association
+0.3 to +0.5	weak positive linear association
+0.5 to +0.8	moderate positive linear association
+0.8 to +1.0	strong positive linear association

If you examine the correlation coefficient (and its confidence limits) for the association between the GHQ and MEE variables:

```
Correlation coefficient: r  = 0.58              <- correlation coefficient
                         r^2= 0.33
95% confidence limits:   0.41 < R < 0.71        <- confidence limits
```

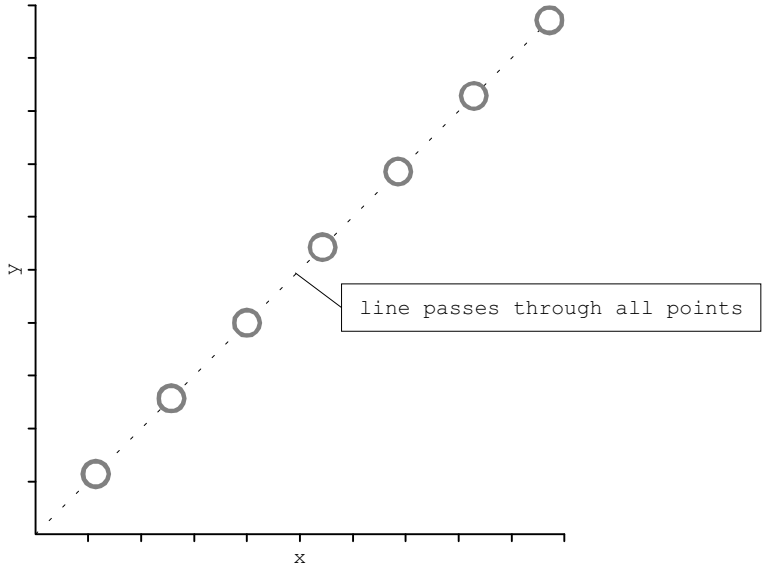
and compare them with the guideline values in the table, you should see that there is a weak to moderate positive linear association between the two variables.

The interpretation of the confidence limits for the correlation coefficient is similar to that used with the relative risk or odds ratio except that zero is the null value. A confidence interval that includes zero indicates that the observed association may be due to random variation.

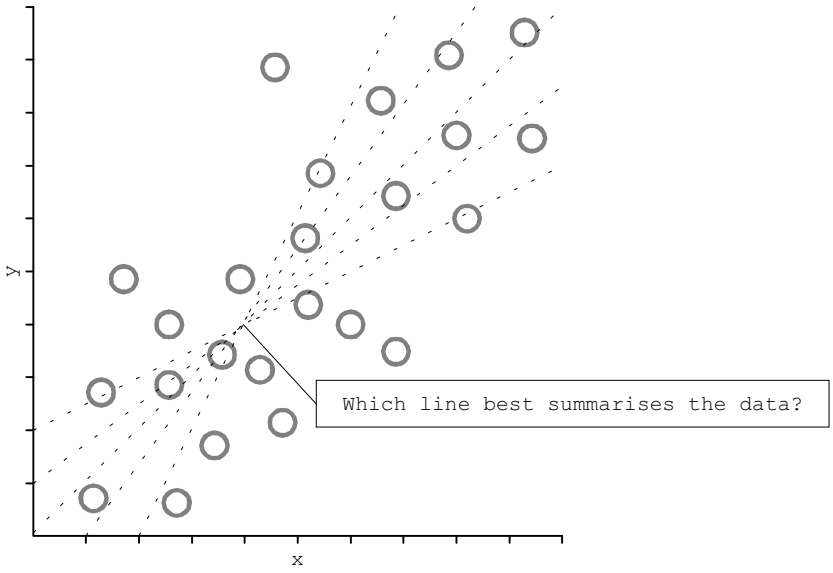
The regression line

Summarising associations with straight lines

If there is a linear association between two variables you can use a straight line to *summarise* the data (in the same way as you can use the mean and standard deviation to summarise a single normally distributed continuous variable). When the correlation coefficient (r) is +1 or -1 the line that best summarises the data is the line that passes through all observations:



but when the variables are less strongly correlated there are many possible lines that could be chosen to represent the data:



The most common method of *fitting* a line to data is the methods of *least squares*. This methods results in a line that minimises the sum of the squared vertical distances from the data point to the line.

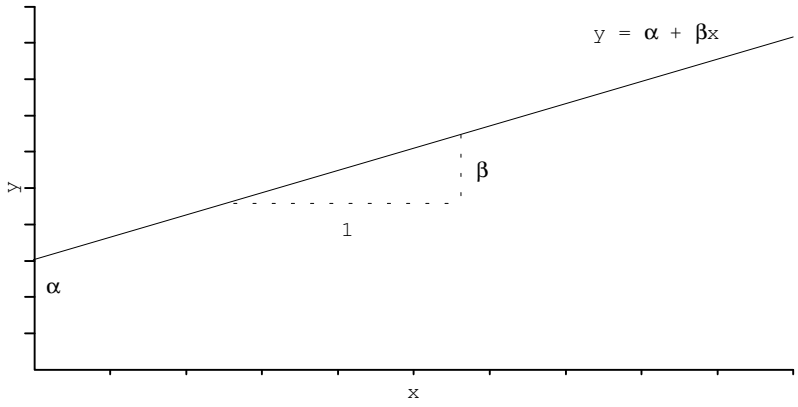
The regression line

The method of least squares

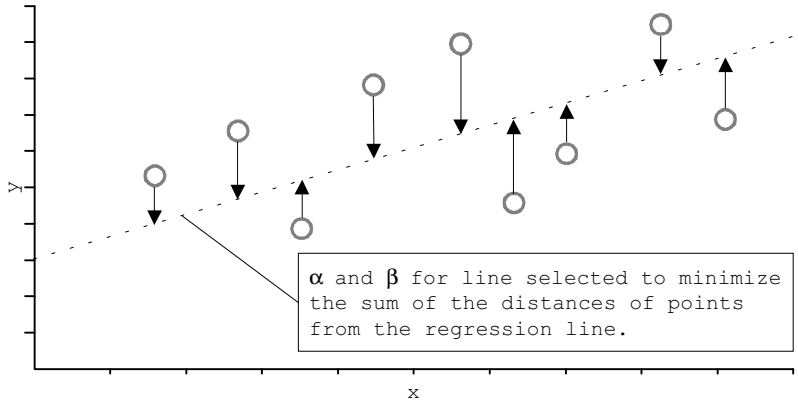
Any straight line drawn on a chart can be represented by the equation:

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

where y is the *response* (or *dependant*) variable and x is the *explanatory* (or *independent* variable). The equation tells us how these two variables are related. The term α is the *intercept*, the point at which the line cuts the y-axis (i.e. the value of y when $x = 0$). The term β is the *slope* of the line, the increase (or decrease) in y per unit increase in x:



We could draw a line by eye using a transparent ruler but this would be prone to error. A better approach is to use the method of *least squares*. Using this method we can calculate values for α and β that minimise the vertical distances between the data points and the line. The method is known as *least squares* because it minimises the sum of the squares of the vertical distances:



The method of least squares is available in most computer packages (including ANALYSIS) and is often called *linear regression* or *ordinary least squares (OLS) regression*. The line found by this method is called the *regression line*. The regression line *estimates* the mean value of y for any given value of x. The line always passes through the point defined by the mean value of x and the mean value of y. The slope, β, is usually referred to as the *regression coefficient* or the *β-coefficient*.

Before using regression techniques you should be clear which variable is dependant (y) and which is independent (x). The choice does not effect the correlation coefficient but does effect the intercept and regression coefficient.

Linear regression

The REGRESS command in ANALYSIS

You can use the REGRESS command in ANALYSIS to perform linear regression:

```
regress ghq = mee
```

ANALYSIS calculates and displays the Pearson correlation coefficient, 95% confidence limits for the Pearson correlation coefficient, r-squared, the y-intercept (α) and the regression coefficient (β):

```
Correlation coefficient: r  = 0.58
                        r^2= 0.33
95% confidence limits:   0.41 < R < 0.71

Source             df   Sum of Squares   Mean Square   F-statistic
Regression          1        1177.6686     1177.6686         40.57
Residuals          81        2351.3917       29.0295
Total              82        3529.0602

ß Coefficients

                          ß            95% confidence                 Partial
Variable     Mean    coefficient     Lower       Upper    Std Error    F-test
MEE         24.9036    0.3762896    0.260495   0.492084    0.059079   40.5680
Y-Intercept            0.3398722
```

In this case the *Y-Intercept* (α) is equal to 0.3398722 and the *slope* (β-*coefficient*) is equal to 0.3762896. The equation for the fitted line is:

```
GHQ = 0.3398722 + (0.3762896 * MEE)
```

For each unit increase in the MEE score the GHQ score increases by 0.3762896. The correlation coefficient (r) is equal to 0.58 which suggests that MEE score is moderately predictive of GHQ score. The coefficient of determination is 0.33 suggesting that 33% of the variability in the GHQ variable can be explained by the variability in the MEE variable.

There is a formal statistical test that can be used with linear regression. This is the same F Statistic that is used with ANOVA techniques. ANALYSIS calculates the F statistic but does not calculate the associated p-value. The F statistic for this regression is equal to 40.57 which, when compared to the F distribution (at $df_1 = 1$ and $df_2 = 82$) in a set of statistical tables (see page 92), is highly significant ($p < 0.001$) suggesting that GHQ and MEE are associated with each other and that MEE is predictive of GHQ.

It is possible to specify more than one independent variable with linear regression. This procedure is called *multiple linear regression* and is beyond the scope of this book.

What You Should Quote in Reports

If you are using linear regression techniques to analyse your data you should quote r, the F-statistic, degrees of freedom, and p-value:

> *'We found MEE score to be moderately predictive of GHQ-28 score ($r = 0.58$, $F = 40.57$, $df=81$, $p < 0.001$)'*

What the regression line does not explain

Predicted values

The equation for the association between the MEE score and the GHQ score is:

```
GHQ = 0.3398722 + (0.3762896 * MEE)
```

Using this equation we can make a *prediction* of a subject's GHQ score given their MEE score. For example, we can predict the GHQ score for subjects with an MEE score of 25 using the equation:

```
GHQ = 0.3398722 + (0.3762896 * MEE)
GHQ = 0.3398722 + (0.3762896 * 25)
GHQ = 9.7471122
```

Which we round up to the whole number value 10 (it is impossible to have a fractional score). Examine the GHQ score for those cases with a MEE score of 25 using the following commands:

```
select mee = 25
freq qhq
```

Note that **no** cases have a GHQ score of 10. This may due to random variation (which is always present). It may also be because MEE score is only a moderately good predictor of GHQ score or because the relationship between GHQ and MEE is not best summarised by a straight line.

Note also that the mean GHQ score is the same as our predicted GHQ score:

```
GHQ   ¦  Freq  Percent   Cum.
------+----------------------
 0    ¦     1   25.0%    25.0%
 8    ¦     1   25.0%    50.0%
14    ¦     1   25.0%    75.0%
18    ¦     1   25.0%   100.0%
------+----------------------
Total ¦     4  100.0%

Sum                =     40.00
Mean               =     10.00  <- Mean GHQ for MEE = 25
Standard deviation =      7.83
```

This is exactly what we would expect as the method of least squares causes the regression line to estimate the mean value of *y* for any given value of *x*.

What the regression line does not explain

Residual values

The difference between the values predicted by the regression line and the actual data are called the *residuals*. It is important that you examine the residuals as it will help you decide how well the linear (straight line) model fits the data and whether you need to apply a transformation to one or both of the variables.

Issue the following commands:

```
define resid ##.####
resid = ghq - (0.3398722 + (0.3762896 * mee))
```

to calculate the residuals. If we plot the residuals against the original x-variable (MEE):

```
scatter mee resid
```

we should see a random plot. This would indicate that there is a relatively straightforward association between MEE score and GHQ score and that the straight line summarises much of the structure in the data with the difference between the actual and predicted values being due to random variation.

There should be no structure in a plot of the residuals against the original *x-variable*. The *residual plot* should be a random cloud of points. If the residual plot shows some structure (other than randomness) this indicates that the straight line does not summarise the data well and that you may need to apply a transformation to one or both of the variables to make the association more linear.

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10 or on the inside back cover of this book.

Exercise 10 - Correlation and regression

Use the SCATTER command to investigate the association between the various continuous variables in the dataset.

Use the REGRESS command to asses the strength of any associations you find.

Calculate and plot residual values for each of the associations you find.

Solutions to practical exercises

Exercise 1 - Examining a dataset

There are 152 cases in the SSSW.REC dataset. The number of cases in a dataset is shown on the first line of the status information at the top of the ANALYSIS screen:

```
Dataset:  A:\SSSW.REC (152 records)                    Free memory: 250K
Criteria: All records selected
```

The command to SELECT only those people who are LIVING alone is:

```
select living = 1
```

Counting the records shown by a BROWSE or LIST command shows that 17 records are selected comprising of 8 records in the ACADEMIC group, 5 records in the PGCE group, and 4 records in the SOCIAL WORK group.

To SELECT only those persons who are **not** White Europeans you should first clear the SELECTion by issuing the SELECT command without specifying a SELECTion criteria:

```
select
```

and then issuing the command:

```
select ethnic <> 10
```

Counting the records shown by a BROWSE or LIST command shows that 42 records are selected comprising of 18 records in the ACADEMIC group, 2 records in the PGCE group, and 22 records in the SOCIAL WORK group.

To SELECT only those persons who are both West-Indian and female you should first clear the SELECTion by issuing the SELECT command without specifying a SELECTion criteria:

```
select
```

and then issuing the command:

```
select ethnic = 2 and sex = 2
```

Counting the records shown by a BROWSE or LIST command shows that 3 records are selected all of which are in the SOCIAL WORK group.

To continue to work with all of the cases in the dataset you should issue the SELECT command without specifying a SELECTion criteria:

```
select
```

Solutions to practical exercises

Exercise 2 - Producing charts

Issue the command:

```
bar marital
```

to produce a BAR chart of MARITAL status. The most common marital status is 2 (single). The two groups 3 (Divorced) and 4 (Separated) have less than 10 persons.

Issue the command:

```
pie age
```

to produce a PIE chart of AGE. 50.7% of the study population was over 25 years old.

A HISTOGRAM is the most appropriate chart for the GHQ-28 (GHQ) score. The PIE chart is too difficult to read because there are two many 'slices' and because the data is ordered (which is difficult to see with a PIE chart). The LINE chart is useful and clear but because LINE charts are often used to show trends over time a LINE chart may be a little confusing for others to read (especially if you include it in a report that also uses LINE charts to show time-series data).

The command to produce a SCATTER plot of the GHQ-28 (GHQ) variable and the emotional exhaustion (MEE) variable is:

```
scatter ghq mee
```

You can specify a regression line be adding '/r' to the SCATTER command:

```
scatter ghq mee /r
```

Solutions to practical exercises

Exercise 3 - Summarising distributions

The command to examine the distribution of the distribution of the ROSENBERG variable is:

```
freq rosenberg
```

which produces the following output:

```
ROSENBERG ¦  Freq  Percent   Cum.
----------+----------------------
10        ¦    10    6.6%     6.6%
11        ¦     4    2.6%     9.3%
12        ¦    12    7.9%    17.2%
13        ¦    11    7.3%    24.5%
14        ¦     5    3.3%    27.8%
15        ¦     6    4.0%    31.8%
16        ¦     4    2.6%    34.4%
17        ¦     9    6.0%    40.4%
18        ¦    13    8.6%    49.0%
19        ¦     8    5.3%    54.3%
20        ¦    17   11.3%    65.6%   <- most common ROSENBERG score
21        ¦    13    8.6%    74.2%
22        ¦     6    4.0%    78.1%
23        ¦     4    2.6%    80.8%
24        ¦     8    5.3%    86.1%
25        ¦     1    0.7%    86.8%
26        ¦     5    3.3%    90.1%
27        ¦     2    1.3%    91.4%
28        ¦     2    1.3%    92.7%
29        ¦     3    2.0%    94.7%
30        ¦     3    2.0%    96.7%
31        ¦     2    1.3%    98.0%
32        ¦     1    0.7%    98.7%
33        ¦     2    1.3%   100.0%
----------+----------------------
    Total ¦   151  100.0%

Sum                 =   2825.00
Mean                =     18.71
Standard deviation  =      5.66
```

48 cases had a ROSENBERG score of less than or equal to 15 (i.e. 10 + 4 + 12 + 11 + 5 + 6 = 48). This represents 31.8% of the study population.

The most common ROSENBERG score is 20 with 17 cases (11.3% of the study population).

Only 5 cases had a ROSENBERG score greater than 30 (2 + 1 + 2 = 5). This represents 3.3% of the study population (100.0% - 96.7% = 3.3%).

Note that only 151 cases are shown in the frequency table. This is because there is one cases with a missing ROSENBERG score.

Solutions to practical exercises

Exercise 3 - Summarising distributions

The command to examine the distribution of the ETHNIC variable is:

```
freq ethnic
```

which produces the following output:

```
ETHNIC |  Freq  Percent   Cum.
-------+----------------------
 1     |     1    0.7%     0.7%
 2     |     6    3.9%     4.6%
 3     |     9    5.9%    10.5%   <- Indian
 4     |     2    1.3%    11.8%
 5     |     3    2.0%    13.8%   <- Bangladeshi
 6     |     1    0.7%    14.5%
 7     |     2    1.3%    15.8%
 9     |     1    0.7%    16.4%
10     |   110   72.4%    88.8%   <- White European
11     |    17   11.2%   100.0%
-------+----------------------
 Total |   152  100.0%

Sum                 =   1379.00   <- Not valid
Mean                =      9.07   <- Not valid
Standard deviation  =      2.54   <- Not valid
```

5.9% of the study population were Indian (coded as ETHNIC = 3).

2.0% of the study population were Bangladeshi (coded as ETHNIC = 5).

27.6% (100% - 72.4% = 27.6%) of the study population were not White European (coded as ETHNIC = 10).

It is **not** valid to use the sum, mean, and standard deviation to summarise the distribution of the ETHNIC variable because it is a categorical variable and these summary measures can only be applied usefully to continuous variables.

You can see a visual representation of a frequency table using the BAR and PIE commands:

```
bar ethnic
pie ethnic
```

There are too many categories of the ETHNIC variable for a PIE chart to be used successfully.

Solutions to practical exercises

Exercise 4 - Summarising distributions

The commands to produce the seven figure summaries for the GHQ, ROSENBERG, MEE, MDP, and MPA variables are:

```
means ghq all /n
means rosenberg all /n
means mee all /n
means mdp all /n
means mpa all /n
```

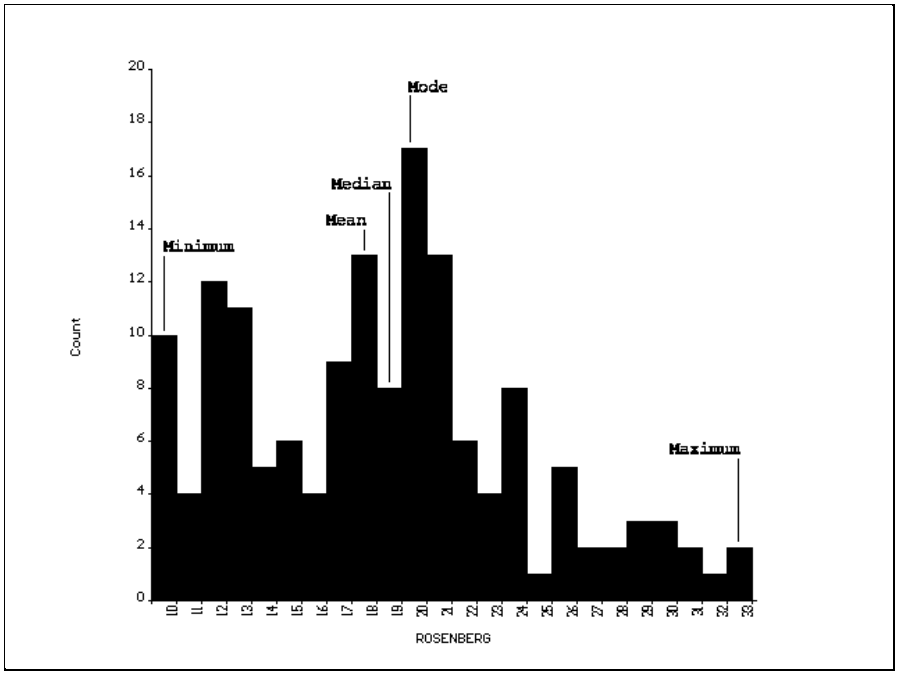
Your completed table should look like this:

	GHQ	ROSENBERG	MEE	MDP	MPA
number of cases	152	151	83	83	83
mode	0.000	20.000	29.000	4.000	34.000
median	7.000	19.000	25.000	5.000	34.000
mean	7.789	18.709	24.904	6.072	33.181
minimum	0.000	10.000	1.000	0.000	5.000
maximum	27.000	33.000	53.000	28.000	46.000
standard deviation	6.719	5.665	10.071	4.999	7.469

Visual representations of the distribution of these variables can be produced using the HISTOGRAM command:

```
histogram ghq
histogram rosenberg
histogram mee
histogram mdp
histogram mpa
```

The mode, median, mean, minimum, and maximum of the ROSENBERG variable are marked on this graph:



Solutions to practical exercises

Exercise 5 - Recoding variables and two-by-two tables

The first step in this exercise is to examine the distribution of the ROSENBERG variable:

```
means rosenberg all /n
```

which produces the following output:

```
Minimum     25%ile     Median     75%ile     Maximum       Mode
 10.000     14.000     19.000     22.000      33.000     20.000
```

And then use the upper quartile as a cut-point for a newly defined variable:

```
define rg <AAAAAAAAAA>
recode rosenberg to rg 10-21="LOW NORMAL" 22-hi="HIGH"
```

You should always check the result of any RECODE command using the BROWSE or LIST commands:

```
browse rosenberg rg
```

And then to use the TABLES command to investigate the association between PROFESSION and your new variable:

```
tables profession rg
```

which produces the following output:

```
                            RG
PROFESSION ¦HIGH      LOW NORMAL ¦ Total
-----------+---------------------+------
         + ¦        14        69 ¦    83
         - ¦        25        43 ¦    68
-----------+---------------------+------
     Total ¦        39       112 ¦   151

                    Single Table Analysis
Odds ratio                                              0.35
Cornfield 95% confidence limits for OR      0.15 < OR < 0.80

Relative risk of (RG=HIGH) for (PROFESSION=+)           0.46
Greenland, Robins 95% conf. limits for RR   0.26 < RR < 0.81
               (Biometrics 1985;41:55-68)
        Ignore relative risk if case control study.

                        Chi-Squares   P-values
                        -----------   --------
        Uncorrected:          7.72     0.00544784 <---
        Mantel-Haenszel:      7.67     0.00560444 <---
        Yates corrected:      6.72     0.00952995 <---
```

Showing that those in the PROFESSIONal group are **less** likely to have high ROSENBERG scores (RR = 0.46, 95% C.I. 0.26 - 0.81, Yates chi-square = 6.72. $p < 0.01$). You can also use the TABLES command to investigate the association between AGE, SEX, MARITAL, and ETHNIC and your new variable:

```
tables age rg
tables sex rg
tables marital rg
tables ethnic rg
```

Showing that those in the under-25 age group are **more** likely to have high ROSENBERG scores (RR = 3.02, 95% C.I. =1.58 - 5.75, Yates chi-square = 12.19, $p < 0.0005$).

Solutions to practical exercises

Exercise 6 - Recoding, subsets, and two-by-two tables

Use the MEANS command to find the quartiles of the distribution of the MEE and MDP variables (as was done in the previous exercise):

```
means mee all /n
means mdp all /n
```

and then use the DEFINE and RECODE commands to create the variables holding the grouped data:

```
define meecase <AAAAAAAAAA>
recode mee to meecase 1-32="LOW NORMAL" 33-hi="HIGH"

define mdpcase <AAAAAAAAAA>
recode mdp to mdpcase 0-8="LOW NORMAL" 9-hi="HIGH"
```

and remember to use the BROWSE or LIST command to check the results of the RECODE commands:

```
browse mee meecase mdp mdpcase
```

Then use the TABLES command to investigate the association between COHORT and your new variables:

```
tables cohort meecase
tables cohort mdpcase
```

Showing that PGCE students were **more** likely to have high MEE scores than SOCIAL WORK students (RR = 2.46, 95% C.I. 1.15 - 5.28, Yates chi-squares = 4.59, $p < 0.05$). This does **not** support the researchers' hypothesis.

Solutions to practical exercises

Exercise 7 - Analysis of variance (ANOVA)

The commands needed to complete this exercise are:

```
means ghq cohort /n
means rosenberg cohort /n
means mee cohort /n
means mdp cohort /n
means mpa cohort /n
```

The PGCE group had a higher mean GHQ-28 score than the social work trainee group (PGCE mean = 11.636, Social Work mean = 8.440, F = 4.946, df = 1, P < 0.05).

The PGCE group had a higher mean MEE score than the social work trainee group (PGCE mean = 30.091, Social Work mean = 21.480, F = 17.447, df = 1, P < 0.005).

The PGCE group had a lower mean MDP score than the social work trainee group (PGCE mean = 4.697, Social Work mean = 6.980, F = 4.313, df = 1, P < 0.05).

Note that these finding may **not** be valid as some of the ANOVA assumptions may be violated (see pages 61-62 and Exercise 8 on page 63).

Solutions to practical exercises

Exercise 8 - Testing the ANOVA assumptions

The commands needed to complete this exercise are:

```
means rosenberg cohort /n
means mee cohort /n
means mdp cohort /n
means mpa cohort /n
```

By comparison of means, median and modes the variables ROSENBERG, MEE, MDP, and MPA all meet the ANOVA assumptions. Comparison of variances or standard deviations using Bartlett's test shows that the MDP variable fails to meet the ANOVA assumptions.

To examine the distributions of these variable they must first be grouped:

```
define rosenlook <AAAAAAAA>
recode rosenberg to rosenlook by 5

define meelook <AAAAAAAA>
recode mee to meelook by 5

define mdplook <AAAAAAAA>
recode mdp to mdplook by 5

define mpalook <AAAAAAAA>
recode mpa to mpalook by 5
```

before issuing the LINE commands to examine the underlying distributions:

```
line rosenlook cohort
line meelook cohort
line mdplook cohort
line mpalook cohort
```

Examination of these charts shows that the MEE and MPA variables are approximately normally distributed (MPA is slightly skewed to the left) and that the ROSENBERG and MDP are severely non-normal. This means that it may **not** be valid to use ANOVA techniques with the ROSENBERG and MDP variables.

Solutions to practical exercises

Exercise 9 - Doing it!

You’re on your own for this exercise. The descriptive and analytical techniques and the ANALYSIS commands necessary for this exercise (and the analysis of the data arising from the Stress and the Student Social Worker Study) have all been introduced and rehearsed in the previous exercises. Good luck . . .

Solutions to practical exercises

Exercise 10 - Correlation and regression

You will have to issue many SCATTER and REGRESS commands to investigate each possible pair of variables to complete this exercise:

SCATTER commands	REGRESS commands
scatter ghq rosenberg /r	**regress rosenberg ghq**
scatter ghq mee /r	**regress mee ghq**
scatter ghq mdp /r	**regress mdp ghq**
scatter ghq mpa /r	**regress mpa ghq**
scatter rosenberg mee /r	**regress mee rosenberg**
scatter rosenberg mdp /r	**regress mdp rosenberg**
scatter rosenberg mpa /r	**regress mpa rosenberg**
scatter mee mdp /r	**regress mdp mee**
scatter mee mpa /r	**regress mpa mee**
scatter mdp mpa /r	**regress mpa mdp**

In this exercise we have ignored considerations about which variable is the *independent* or *explanatory* variable and which variable is the *dependent* or *response* variable (see page 80). Associations (none of which are particularly strong) exist between the following variables:

y-variable	x-variable	r	95% C.I.
ROSENBERG	GHQ	0.33	0.18 - 0.46
MEE	GHQ	0.58	0.41 - 0.71
MEE	ROSENBERG	0.29	0.08 - 0.47
MDP	MEE	0.29	0.08 - 0.48

We will concentrate on the association between the GHQ and ROSENBERG variables. The command:

```
regress rosenberg ghq
```

produces the following output:

```
                          ß            95% confidence                 Partial
Variable     Mean    coefficient     Lower       Upper    Std Error    F-test
GHQ          7.7947    0.2762311    0.148804   0.403658    0.065014   18.0524
Y-Intercept           16.5554701
```

Residuals may be calculated and displayed using the following commands:

```
define rosenresid ##.####
rosenresid = rosenberg - (16.5554701 + (0.2762311 * ghq))
scatter ghq rosenresid
```

The scatter plot is a random cloud of points indicating that the straight line:

```
scatter ghq rosenberg /r
```

summarises the data well and that no transformation need be applied to either variable.

Statistical tables

Statistical tables

The following pages show partial statistical tables for the chi-square distribution, the sum of ranks in the smallest groups in the Wilcoxon rank sum test (*U*), and the *F*-distributions with examples of how to use them.

ANALYSIS automatically calculates p-values for most procedures but does not do so for the *F* under linear regression. If you intend to use ANALYSIS to perform linear regression you will have to refer to the table on page 92 to obtain a p-value. If you intend to use ANALYSIS to perform multiple linear regression you will need to purchase a set of statistical tables. A suitable set of tables is:

Neave HR, *Elementary Statistical Tables*, Routledge, 1981

Statistical tables

Percentage points of the chi-square distribution

In the following table ROSE20PLUS indicates a score of twenty or more for the Rosenberg Self-Esteem Scale and GHQCASE indicates a score of five or more for the GHQ-28 questionnaire:

```
                          GHQCASE
ROSE20PLUS ¦CASE     NOT CASE ¦ Total
-----------+-----------------+------
         + ¦       49      20 ¦    69
         - ¦       42      41 ¦    83
-----------+-----------------+------
     Total ¦       91      61 ¦   152
```

and yields a Pearson's chi-square statistic of 6.53. The degrees of freedom for this table are:

```
(rows - 1) * (columns - 1) = (2 - 1) * (2 - 1) = 1
```

From the table below we can see that for a chi-square of 6.53 on 1 degree of freedom the p-value is between p = 0.05 and p = 0.01:

	p-value			
d.f.	**0.05**	**0.01**	**0.005**	**0.001**
1	3.84	6.63	7.88	10.83
2	5.99	9.21	10.60	13.82
3	7.81	11.34	12.84	16.27
4	9.49	13.28	14.86	18.47
5	11.07	15.09	16.75	20.52
6	12.59	16.81	18.55	22.46
7	14.07	18.48	20.28	24.32
8	15.51	20.09	21.96	26.13
9	16.92	21.67	23.59	27.88
10	18.31	23.21	25.19	29.59
11	19.68	24.73	26.76	31.26
12	21.03	26.22	28.30	32.91
13	22.36	27.69	29.82	34.53
14	23.68	29.14	31.32	36.12
15	25.00	30.58	32.80	37.70
16	26.30	32.00	34.27	39.25
17	27.59	33.41	35.72	40.79
18	28.87	34.81	37.16	42.31
19	30.14	36.19	38.58	43.82
20	31.41	37.57	40.00	45.32

This would be reported as:

> *'We found an association between a score of twenty or higher on the Rosenberg Self-Esteem scale and a score of five or higher on the GHQ-28 scale (chi-square = 6.53, p < 0.05).'*

Note that only a partial table is shown here as ANALYSIS calculates and reports a p-value automatically.

Statistical tables

Critical ranges for the Wilcoxon Rank Sum Test

The Worked Example of the Wilcoxon Rank-Sum Test shown a page 71 yielded a *U*-statistic of 33.5 with sample sizes for the two groups of five and six. The table below shows 33.5 to be within the critical range of 18 to 42 at p = 0.05 for group sizes of five and six:

	p-value		
n_1, n_2	**0.05**	**0.01**	**0.001**
5, 5	17 to 38	15 to 40	
5, 6	18 to 42	16 to 44	
5, 7	20 to 45	17 to 48	
5, 8	21 to 49	17 to 53	
5, 9	22 to 53	18 to 57	15 to 60
5, 10	23 to 57	19 to 61	15 to 65
5, 11	24 to 61	20 to 65	16 to 69
5, 12	26 to 64	21 to 69	16 to 74
5, 13	27 to 68	22 to 73	17 to 78
5, 14	28 to 72	22 to 78	17 to 83
5, 15	29 to 76	23 to 82	18 to 87
5, 16	31 to 79	24 to 86	18 to 92
5, 17	32 to 83	25 to 90	19 to 96
5, 18	33 to 87	26 to 94	19 to 101
5, 19	34 to 91	27 to 98	20 to 105
5, 20	35 to 95	28 to 102	20 to 110

n_1, n_2 = sample sizes of two groups
Significant if sum of ranks (*U*) in smallest group is on boundary or outside of range

This would be reported as:

> *'We found no association between group membership and GHQ-28 score (Wilcoxon Rank-Sum Test, U = 33.5, p > 0.05)'.*

Note that only a partial table is shown here as ANALYSIS calculates and reports a p-value automatically.

Statistical tables

Percentage points of the *F* distribution

The following output:

```
Source             df   Sum of Squares   Mean Square   F-statistic
Regression          1        1177.6686     1177.6686         40.57
Residuals          81        2351.3917       29.0295
Total              82        3529.0602

ß Coefficients

                           ß            95% confidence                  Partial
Variable     Mean    coefficient     Lower       Upper    Std Error     F-test
MEE         24.9036    0.3762896    0.260495   0.492084    0.059079   40.5680
Y-Intercept            0.3398722
```

was produced by the command:

```
regress ghq = mee
```

Degrees of freedom df_1 is the number of independent variables (1) and the degrees of freedom df_2 is the number of cases minus the number of independent variables minus one (81).

From the table below we can see that the p-value for *F* of 40.57 on 1 and 81 degrees of freedom (we have to look at the values for $df_2 = 60$ and $df_2 = 120$) is $p < 0.001$:

df_2	p -value	$df_1 = 1$
10	0.05	4.96
	0.01	10.04
	0.001	21.04
12	0.05	4.75
	0.01	9.33
	0.001	18.64
14	0.05	4.60
	0.01	8.86
	0.001	17.14
16	0.05	4.49
	0.01	8.53
	0.001	16.12
18	0.05	4.41
	0.01	8.29
	0.001	15.38
20	0.05	4.35
	0.01	8.10
	0.001	14.82

df_2	p - value	$df_1 = 1$
25	0.05	4.24
	0.01	7.77
	0.001	13.88
30	0.05	4.17
	0.01	7.56
	0.001	13.29
40	0.05	4.08
	0.01	7.31
	0.001	12.61
60	0.05	4.00
	0.01	7.08
	0.001	11.97
120	0.05	3.92
	0.01	6.85
	0.001	11.38
∞	0.05	3.84
	0.01	6.63
	0.001	10.83

This would be reported as:

> *'We found an association between the Maslach Emotional Exhaustion scale and the GHQ-28 score (F = 40.57, df = 81 p > 0.001)'.*

It may also be appropriate to quote the Pearson correlation coefficient in this situation.

Note that you do not need to use this table to assess the significance of *F* with the MEANS command as ANALYSIS calculates a p-value automatically.