Analysing Data

A practical primer using EpiInfo

Mark Myatt, Sue Ritter,
Esther Hudes, Deborah Bain & Alison Graves

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

This edition published 2000 for use solely by the Center for AIDS Prevention Studies, University of California, San Francisco.

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

ISBN 1-873937-46-6

Note: The previous edition of this text has been modified by Esther Hudes, Deborah Bain, and Alison Graves of the Center for AIDS Prevention Studies, UCSF, for use by the International Visiting Scientists Program. The original data set used for this publication, SSSW.REC, was replaced by a data set compiled from data collected during the San Francisco Men's Health Study, SFMHSHIV.REC. The text has also been altered so that it reflects EpiInfo version 6.0. Details of the SFMHS is given on page 9 of this publication.

‘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

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 of HIV infection among men in San Francisco. 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 HIVSTAT
COMMAND	ANALYSIS commands referred to in the text.	the REGRESS command
MODULE	Modules of the EpiInfo 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 age 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 is a public domain database and statistics package called EpiInfo and can be downloaded from the following website: http://www.cdc.gov/epo/epi/epiinfo.htm. This site also provides instructions on downloading the software and a phone number for technical assistance.

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.

The example dataset will be on your computer hard drive in the directory C:\CAPS as C:\CAPS\SFMHSHIV.REC.

For the purposes of this workbook we will be using the ANALYSIS module of EpiInfo. To use EpiInfo, double-click on the EpiInfo icon in the C:\CAPS directory. You will need to create a shortcut icon if one does not exist. ANALYSIS can read both EpiInfo and dBase format files. Using the mouse or arrow keys, go to the pull-down menu under Programs. Choose ANALYSIS of data.

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, Version 6: 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

The San Francisco Men's Health Study is a prospective epidemiological study designed to learn more about the natural history of Acquired Immune Deficiency Syndrome (AIDS). The study was funded by the National Institutes of Allergy and Infectious Disease (NIAID) under a contract with the University of California, Berkeley.

Several organizations cooperated in carrying out the San Francisco study. The Survey Research Center of the University of California, Berkeley, was responsible for questionnaire design, sampling, recruitment of participants, interviewing, coding and data management. The Children's Hospital of San Francisco was the location for the interviewing and provided the facilities for the physical examinations, the various clinical tests performed, and the collection of specimens, some of which were sent to NIAID. The Irwin Memorial Blood Bank of San Francisco performed most of the blood analysis and also stored frozen lymphocytes. Research on the specimens retained locally was carried out at the University of California, San Francisco.

The target population was defined as follows:

1. Residents of 19 census tracts in San Francisco. The tracts are 162-171, 204-207, and 210-214.
2. Men, currently unmarried, aged 25-54.
3. English speaking -- to avoid the cost of questionnaires in other languages for very few cases.

The sample was a stratified two-stage sample of all households within the target area. All eligible males in each selected housing unit were taken into the sample. There were 1043 interviews completed during the first wave of data collection. For details of the sample design and weighting, see the Survey Research Center Technical Report entitled "Sampling Methods and Wave 1 Field Results of the San Francisco Men's Health Study".

Work on the project began in September, 1983. The first year was devoted to the design of the project and the first wave of interviews. The study plan called for reinterviewing participants several additional times at six-month intervals.

The initial wave of interviews (N=1043) was conducted from 1 May 1984 to 11 April 1985. The final wave of interviews (N=527) began 5 January 1992 and ended 14 May 1993. In this text we will only use data from the initial wave. We will also ignore the complex survey structure and the sampling weights.

Further details of the study can be found in:

Winkelstein, W Jr; Samuel, M; Padian, N; Wiley, J; Lang, W; Anderson, R; Levy, J. **The San Francisco Men's Health Study: III. Reduction in Human Immunodeficiency Virus Transmission among Homosexual/Bisexual Men, 1982-86.** AJPH, 1987 June, 76(9):685-689.

Lang, W; Anderson, RE; Perkins, H; Grant, RM; Lyman, D; Winkelstein, W Jr; Royce, R; Levy, JA. **Clinical, immunologic, and serologic findings in men at risk for acquired immunodeficiency syndrome. The San Francisco Men's Health Study.** JAMA, 1987 Jan 16, 257(3):326-30.

Winkelstein, W Jr; Lyman, DM; Padian, N; Grant, R; Samuel, M; Wiley, JA; Anderson, RE; Lang, W; Riggs, J; Levy, JA. **Sexual practices and risk of infection by the human immunodeficiency virus. The San Francisco Men's Health Study.** JAMA, 1987 Jan 16, 257(3):321-5.

Variable names, types, and ranges

The structure of the SFMHSHIV.REC dataset

Data from this study is stored in an EpiInfo data file called SFMHSHIV.REC. The following table lists the *names*, *types*, and *ranges* of variables in the SFMHSHIV.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
ID	Subject ID number	Number	#####	54541 thru 59383
AGE	Age	Number	##	Continuous (22-55)
RACE	Racial background	Number	#	1 = White 2 = Hispanic 3 = Black 4 = Other
EDUC	Education level	Number	#	1 = High school diploma or less 2 = College diploma or any college 3 = Graduate degree or classes
DEPRESD	Felt depressed in past week	Number	#	1 = Rarely or none 2 = Some or a little 3 = Occasionally 4 = Most or all the time
SLEEP	Sleep was restless in the past week	Number	#	1 = Rarely or none 2 = Some or a little 3 = Occasionally 4 = Most or all the time
ENJOYLF	Enjoyed life in the past week	Number	#	1 = Rarely or none 2 = Some or a little 3 = Occasionally 4 = Most or all the time
SMOKER	Smokes cigarettes now	Number	#	1 = Yes 2 = No
SMOKEAMT	Packs per day	Number	#	1 = Less than ½ pack 2 = ½ pack to one pack 3 = 1-2 packs 4 = 2 or more packs
SEXPREF	Sexual orientation	Number	#	1 = Homosexual or bisexual 2 = Heterosexual
NPARTNER	Number of male sex partners in past 2 years	Number	###	Count (0-998)
NPARTC	Number of male sex partners in past 2 years (grouping of the variable NPARTNER)	Number	#	0 = 0 1 = 1 2 = 2-9 3 = 10-49 4 = 50 or more
ANONSEX	Anonymous male sex partners in past 2 years	Number	#	1 = Yes 2 = No
MARIJUAN	Used marijuana in past 2 years	Number	#	1 = Yes 2 = No
POPPERS	Used poppers in past 2 years	Number	#	1 = Yes 2 = No
COCAINE	Used cocaine in past 2 years	Number	#	1 = Yes 2 = No
HALUCIN	Used halucinogens in past 2 years	Number	#	1 = Yes 2 = No
DOWNERS	Used downers in past 2 years	Number	#	1 = Yes 2 = No
UPPERS	Used uppers in past 2 years	Number	#	1 = Yes 2 = No
OTHDRUG	Used other drugs in past 2 years	Number	#	1 = Yes 2 = No
OTHDRT	Name of other drugs used in past 2 years	Character	<AAAAAAAAA>	MDA ETHYL CHL HEROIN
HEIGHT	Height in cm	Number	###	Continuous (160-196)
WEIGHT	Weight in kg	Number	###	Continuous (48-118)
CD4	CD4+ cell count	Number	####	Count (47-7107)
HIVSTAT	HIV test results	Number	#	1 = Positive 2 = Negative
HIVLOAD	HIV viral load	Number	#######	Continuous (398-3311311)
ANYDRUG	Used any drug in past 2 years	Number	#	1 = Yes 2 = No

Categorical and continuous variable

The two broad groups

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

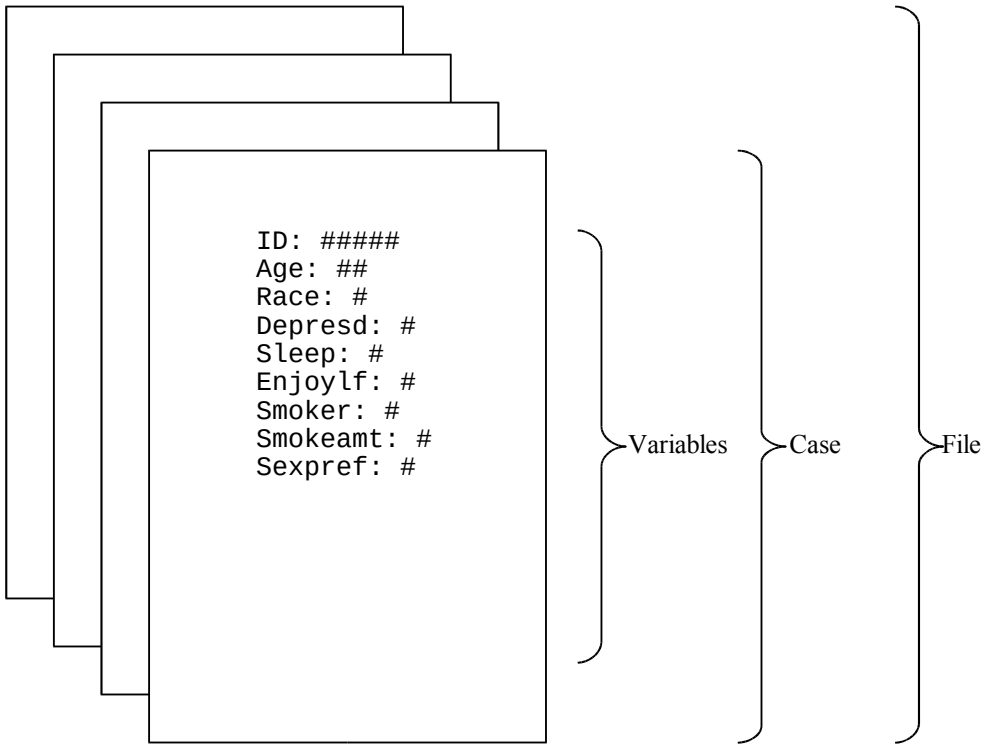
> **Qualitative/Categorical variables** that hold *codes* that indicate membership of a defined category or group. The variables RACE, EDUC, and SMOKEAMT are examples of categorical variables in the **San Francisco Men's Health Study** dataset. Categorical variables can be stored as either numerical or non-numerical data. Even when stored as numerical data, the actual numerical values are typically not meaningful. A particularly common type of categorical variable is a *binary categorical variable* where cases can belong to one of two alternative groups. The variables SMOKER and HIVSTAT are examples of binary categorical variables in the **San Francisco Men's Health Study** dataset. Categorical variables can be *ordered* (as with SMOKEAMT) or *non-ordered* (as with RACE).
>
> **Quantitative/Numerical variables** that hold data that is measured and stored on an *ordered numerical scale*. The numerical information such variables hold is typically meaningful. There are two types of quantitative/numerical variables. *Discrete variables* hold data limited to distinct values, usually, but not always, whole numbers (integers). This includes data such as scores (scales) and counts. The variables NPARTNER, CD4, and HIVLOAD are examples of discrete variables in the **San Francisco Men's Health Study** dataset. *Continuous variables* hold data that vary in a continuous range not constrained to distinct values. The variables HEIGHT and WEIGHT are continuous variables. In this workbook we will use the word "continuous" loosely to describe any quantitative/numerical variables with many values.

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 ten variables for the first three cases in the San Francisco Men's Health Study:

ID	Age	Race	Educ	Depresd	Sleep	Enjoylf	Smoker	Smokeamt	Sexpref
54541	25	1	2	1	1	4	2	.	2
54544	49	1	3	4	1	2	2	.	2
54550	43	1	1	1	1	1	1	3	2

Note that there is no data in the SMOKEAMT variable for the first two cases. This is because the study subjects whose data is held in these two records were non-smokers (SMOKER = 2).

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. On the EpiInfo screen, choose ANALYSIS of data from the Programs pull-down menu (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 **San Francisco Men's Health Study** is stored in an EpiInfo file called SFMHSHIV.REC. Type:

```
read sfmhshiv.rec
```

and press ENTER . Alternatively, you can type read and press enter. You will be given a list of possible files with the ending ".REC". Use the arrow keys to highlight the SFMHSHIV.REC file 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:  C:\CAPS\SFMHSHIV.REC (1043 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 age race educ sexpref npartner anonsex
```

The AGE, RACE, EDUC, SEXPREF, NPARTNER, and ANONSEX 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 1043 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 npartner hivstat hivload
```

and the command:

```
browse npartner hivstat hivload
```

both allow you to examine the contents of the NPARTNER, HIVSTAT, and HIVLOAD 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 sexpref = 2
```

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

```
Dataset:  C:\CAPS\SFMHSHIV.REC (1043 records)                    Free memory: 250K
Criteria: sexpref = 2
```

Type the command:

```
list id age sexpref race
```

If we page through the output screen, we can see that only heterosexuals are listed. Now suppose we want to work with the homosexual/bisexual subjects (SEXPREF = 1). If we type:

```
select sexpref = 1
```

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

```
Dataset:  C:\CAPS\SFMHSHIV.REC (1043 records)                    Free memory: 250K
Criteria: (sexpref = 2) AND (sexpref = 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:  C:\CAPS\SFMHSHIV.REC (1043 records)                    Free memory: 250K
Criteria: All records selected
```

Issue another SELECT command to SELECT the homosexuals/bisexuals:

```
select sexpref = 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. 'SEXPREF = 1') used with the SELECT command can be a complex *expression* using a combination of *inequality operators*:

Operator	Meaning	Example
=	equal to	`sexpref = 1` `othdrt = "heroin"`
<>	not equal to	`sexpref <> 1`
<	less than	`age < 35`
>	greater than	`age > 35`
<=	less than or equal to	`age <= 35`
>=	greater than or equal to	`age >= 35`

and *Boolean* operators:

Operator	Meaning	Example
and	both expressions must be true	`sexpref = 1 and age < 35`
or	either of the expressions may be true	`sexpref = 1 or age < 35`
not	the expression must not be true	`sexpref = 1 and not age < 35`

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 age less than 35?') 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
`sexpref = 2`	Heterosexuals
`sexpref = 2 and smoker = 1`	Heterosexuals who smoke
`sexpref = 2 and smoker = 1 and race = 1`	White heterosexuals who smoke
`sexpref = 2 or smoker = 1`	Heterosexuals and smokers
`sexpref = 2 or (smoker = 1 and race = 1)`	Heterosexuals and white smokers

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 quotes (‘) 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 EpiInfo program screen.

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 **San Francisco Men's Health Study** dataset by typing:

```
read sfmhshiv.rec
```

Suppose we want to examine the distribution of the RACE (racial background) variable. We could produce a BAR chart by typing:

```
bar race
```

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 AGE variable using the HISTOGRAM command:

```
histogram age
```

You can produce a PIE chart of the EDUC (education) variable using the PIE command:

```
pie educ
```

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

```
line age
```

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

```
line age sexpref
```

shows separate lines for heterosexuals and homosexuals/bisexuals. The labels for the lines (1 and 2) are taken from the values of the SEXPREF variable (i.e. 1 = homosexual/bisexual, 2 = heterosexual).

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 height weight
```

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 height weight /r
```

Chart types and their uses

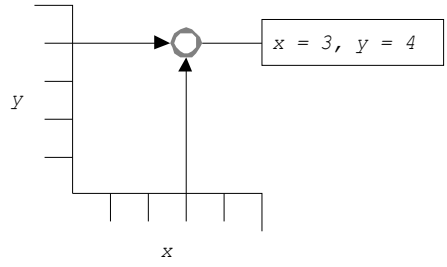
Analysis chart types and their uses

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

Chart	Examples	Description
PIE	`pie educ` `pie race`	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. sexual preference). 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 educ`
	`bar race`
	`bar sleep`
		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 age`
	`histogram height`
	`histogram weight`
		Used to display the distribution of values in a (continuous or count) quantitative/numerical 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 age` `line age sexpref`	Used to display the distribution of values in a (continuous or count) quantitative/numerical 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 height weight`	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.

Exercise 1 - Examining a dataset

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

SELECT only those persons who are SMOKERs. Use BROWSE or LIST to see how many records are selected in each of the HIVSTAT groups (positive and negative).

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

SELECT only those persons who are **not** White. Use BROWSE or LIST to see how many records are selected.

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

SELECT only those persons who are both Hispanic and homosexual/bisexual. Use BROWSE or LIST to see how many records are selected in each of the HIVSTAT groups (positive and negative).

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

Exercise 2 - Producing charts

Produce a BAR chart of RACE. Which is the most common racial background?

Produce a PIE chart of EDUCation. What percentage of the study population either has a graduate degree or has attended graduate school?

Produce a HISTOGRAM, PIE chart, and LINE plot of the AGE variable. Which chart do you think is the most appropriate and why? What was the highest and lowest age recorded for any person?

Produce a SCATTER plot of the HEIGHT and WEIGHT variables. Produce the same chart again but with a regression line. Do the points seem to lie close to the 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 race
```

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

```
freq race depresd sleep
```

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

```
 RACE |  Freq  Percent   Cum.
------+----------------------
    1 |   897   86.6%    86.6%
    2 |    52    5.0%    91.6%
    3 |    52    5.0%    96.6%
    4 |    35    3.4%   100.0%
------+----------------------
Total |  1036  100.0%


      Total        Sum       Mean   Variance    Std Dev    Std Err
       1036       1297      1.252      0.492      0.701      0.022

    Minimum     25%ile     Median     75%ile    Maximum       Mode
      1.000      1.000      1.000      1.000      4.000      1.000

Student's "t", testing whether mean differs from zero.
T statistic = 57.447,  df =  1035   p-value = 0.00000
```

In this case the summary statistics (*sum*, *mean*, *standard deviation, etc.*), with the exception of the mode, are meaningless because RACE is a *categorical* variable (i.e. it contains codes that indicate membership of mutually exclusive categories). The quoted value for the mean (1.252) does **not** point somewhere between being White and Hispanic. 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.

Exercise 3 - Summarising distributions

Use the FREQ command to examine the distribution of the NPARTC variable (number of male sexual partners). For how many subjects was the number of male sexual partners equal to zero? What was the most common NPARTC value? For how many subjects was the number of male sexual partners greater than or equal to 50?

Use the FREQ command to examine the distribution of the SMOKEAMT variable. Note that the variable is only defined for smokers. What percentage of subjects who smoked cigarettes smoked less than ½ pack per day? What percentage smoked ½ pack to one pack per day? What percentage smoked 2 or more packs per day?

The table below was produced using the command FREQ SMOKEAMT. 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:

```
SMOKEAMT |  Freq  Percent   Cum.
---------+----------------------
       1 |    50   12.4%    12.4%
       2 |    87   21.6%    34.0%
       3 |   192   47.6%    81.6%
       4 |    74   18.4%   100.0%
---------+----------------------
   Total |   403  100.0%


      Total        Sum       Mean   Variance    Std Dev    Std Err
        403       1096      2.720      0.819      0.905      0.045

    Minimum     25%ile     Median     75%ile    Maximum       Mode
      1.000      2.000      3.000      3.000      4.000      3.000

Student's "t", testing whether mean differs from zero.
T statistic = 60.321,  df =   402   p-value = -0.00000
```

What problems might there be in using the sum, mean, variance, standard deviation, and standard error (the first line of statistics) to summarise this distribution? Does the Student's "t" make sense?

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

Summarising distributions

Categorical variables

Variables such as RACE, EDUC, and DEPRESD are stored as a set of mutually exclusive codes. Such variables are known as *categorical* variables. The RACE variable, for example, is coded:

Code	Meaning
1	White
2	Hispanic
3	Black
4	Other

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:

```
RACE   |  Freq  Percent   Cum.
-------+----------------------
     1 |   897   86.6%    86.6%       <- Largest (modal) group
     2 |    52    5.0%    91.6%
     3 |    52    5.0%    96.6%
     4 |    35    3.4%   100.0%       <- Smallest group
-------+----------------------
 Total |  1036  100.0%


      Total        Sum       Mean   Variance    Std Dev    Std Err
       1036       1297      1.252      0.492      0.701      0.022

    Minimum     25%ile     Median     75%ile    Maximum       Mode
      1.000      1.000      1.000      1.000      4.000      1.000

Student's "t", testing whether mean differs from zero.
T statistic = 57.447,  df =  1035   p-value = 0.00000
```

Beware of the summary statistics presented below this frequency table (with the exception of *Total* and *Mode*) as they are meaningless for categorical variables such as RACE, EDUC, and DEPRESD.

"Continuous" variables

Some variables, such as AGE, NPARTNER, CD4, and HIVLOAD, hold quantitative data with many values. 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 HEIGHT variable by issuing the command:

```
freq height
```

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 HEIGHT variable. Important information has been highlighted:

```
HEIGHT|  Freq  Percent   Cum.
------+----------------------
  160 |    14     1.4%     1.4%  <- minimum value
  163 |    11     1.1%     2.4%
  165 |    37     3.6%     6.0%
  168 |    56     5.4%    11.4%
  170 |    92     8.9%    20.3%
  173 |   138    13.3%    33.6%
  175 |   134    12.9%    46.6%
  178 |   143    13.8%    60.4%  <- most common (modal) & middle (median) values
  180 |   129    12.5%    72.9%
  183 |   113    10.9%    83.8%
  185 |    79     7.6%    91.4%
  188 |    48     4.6%    96.0%
  191 |    17     1.6%    97.7%
  193 |    15     1.4%    99.1%
  196 |     9     0.9%   100.0%  <- maximum value
------+----------------------
Total |  1035  100.0%           <- number of observations (cases)


      Total        Sum       Mean   Variance    Std Dev    Std Err
       1035     183408    177.206     48.289      6.949      0.216

    Minimum     25%ile     Median     75%ile    Maximum       Mode
    160.000    173.000    178.000    183.000    196.000    178.000

Student's "t", testing whether mean differs from zero.
T statistic = 820.394,  df =  1034   p-value = 0.00000
```

The highlighted information is known as the *seven figure summary* and can be used to summarise the distribution of quantitative/numerical 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 (s)	$\sqrt{\frac{\sum (x-\bar{x})^2}{(n-1)}}$	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.

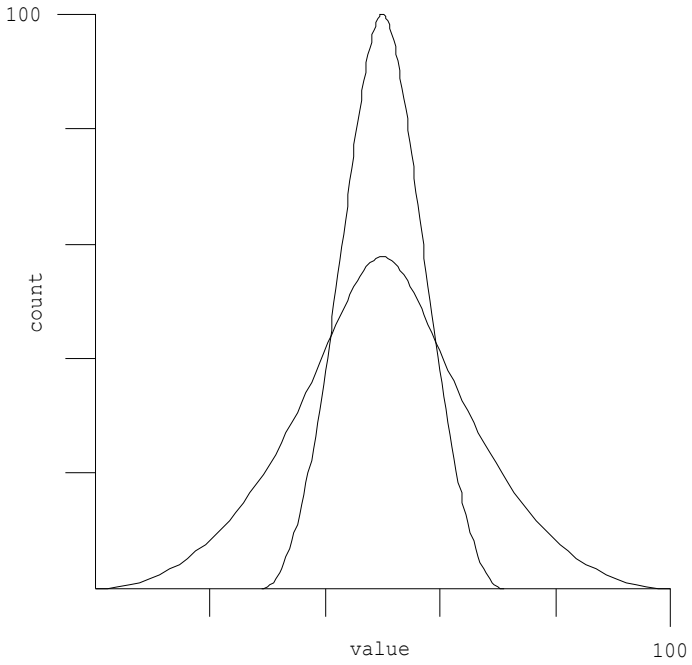
Summarising distributions of continuous data

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* (8) 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.

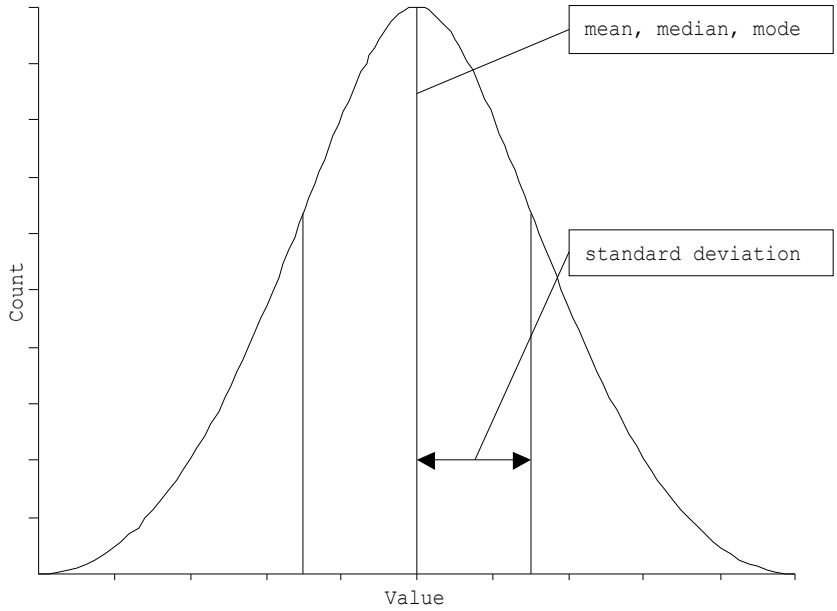
The range and the standard deviation

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 65 - 70).

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 DESCRIBE command for dealing with continuous data. The DESCRIBE command produces summary measures. Issue the command:

```
describe height
```

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

```
Variable        Obs        Total       Mean   Variance    Std Dev
HEIGHT         1035   183408.000    177.206     48.289      6.949
```

You can instruct ANALYSIS to display the cut-point values for the quartiles and the minimum, maximum, and modal values by adding "/all" to the end of the command line:

```
describe height /all
```

which gives the following ouput:

```
Variable        Obs        Total       Mean   Variance    Std Dev
HEIGHT         1035   183408.000    177.206     48.289      6.949

Variable     Minimum     25%ile     Median     75%ile    Maximum       Mode
HEIGHT       160.000    173.000    178.000    183.000    196.000    178.000
```

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 race
bar race
```

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 age
line age
```

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 age sexpref
```

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

A more useful histogram would be for age in categories, rather than age itself (see page 34).

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10.

Exercise 4 - Summarising distributions

Use the DESCRIBE command to produce the seven figure summaries for the AGE, NPARTNER, HEIGHT, WEIGHT, and CD4 variables and complete the following table:

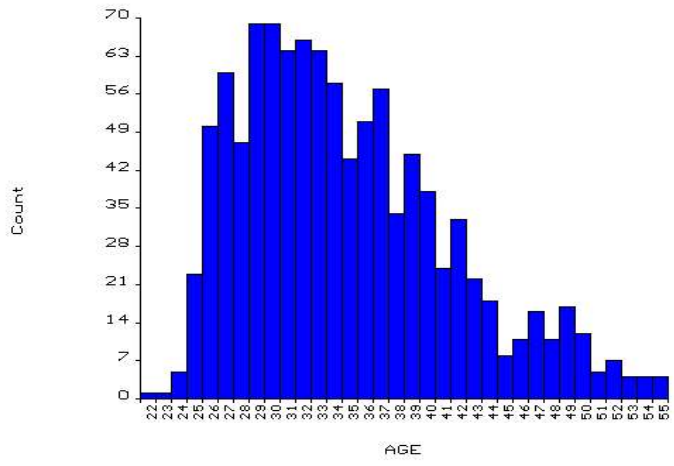
	AGE	NPARTNER	HEIGHT	WEIGHT	CD4
number of cases					
mode					
median					
mean					
minimum					
maximum					
standard deviation					

Use the HISTOGRAM command to graph the distributions of the AGE, HEIGHT, and WEIGHT variables.

The command:

```
histogram age
```

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. AGEGROUP) you first need to tell ANALYSIS what kind of variable AGEGROUP will be using the DEFINE command. Variable types that could be used for the new variable are:

Name	Typical Values	Variable Type	DEFINE 'Code'
AGEGROUP	1, 2	Number	#
AGEGROUP	Y, N	Logical (Yes/No)	<Y>
AGEGROUP	+, -	Character	<A>
AGEGROUP	"YOUNG", "OLD"	Character	<AAAAA>

Issue the command:

```
define agegroup #
```

to create a new number variable called AGEGROUP. The number of # signs defines the length of the new variable. Use the ANALYSIS help system to discover how other codes can be used with the DEFINE command.

Press F4 or type BROWSE. You should see a blank column at the right hand side of the data file which has the heading AGEGROUP (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 age to agegroup lo-29=1 30-39=2 40-49=3 50-hi=4
```

to RECODE the variable AGE into the variable AGEGROUP. The RECODE command must be issued **after** defining the new variable. Your data file now has the original variable AGE and a new variable AGEGROUP. The choice of *cut-points* for group membership is often *arbitrary* and can make a difference to the results of analysis. 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 age agegroup
browse age agegroup
```

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

```
define htgroup <AAAAAA>
if height <= 173 and height <> . then htgroup = "SHORT"
if height >= 174 and height <= 182 then htgroup = "MEDIUM"
if height >= 183 then htgroup = "TALL"
```

In this example the lower and upper quartiles (found using the DESCRIBE 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 HTGROUP 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 weight
```

It might be easier to examine the distribution of the variable if it were grouped into 10kg categories. Issue the commands:

```
define wtgroup ###
recode weight to wtgroup 40-49=45 50-59=55 60-69=65 70-79=75 80-89=85
90-99=95 100-109=105 110-119=115
```

to group the values of the WEIGHT variable into six groups (or *intervals*). Please note that the RECODE command should be typed on a single line. You should not press the [Enter] key after you have entered the text "80-89=85". Type the entire command before pressing the [Enter] key.

Use the BROWSE or LIST command to check that the commands have had the desired effect. Issue the command:

```
histogram wtgroup
freq wtgroup
```

It is easier to examine the distribution of the data, particularly the proportion of subjects in the tails of the distribution (i.e. those subjects with either low or high weights).

WTGROUP	Freq	Percent	Cum.	
45	1	0.1%	0.1%	<- Low weights
55	60	5.8%	5.9%	
65	288	27.9%	33.9%	
75	374	36.3%	70.1%	
85	197	19.1%	89.2%	
95	79	7.7%	96.9%	
105	18	1.7%	98.6%	
115	14	1.4%	100.0%	<- High weights
Total	1031	100.0%		

Creating and recoding variables

Rules for recoding continuous data

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 weight to wtgroup 40-50=45 50-60=55 60-70=65 ...
   ```

   is *ambiguous* for cases where WEIGHT = 50 or WEIGHT = 60.

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

   ```
   define wtdiff ###
   recode weight to wtdiff 40-49=45 50-59=55 60-64=62 65-69=67 70-79=75
   80-84=82 85-89=87 90-99=95 100-110=105 111-120=115
   histogram wtdiff
   ```

   give a wrong picture of the underlying distribution of the WEIGHT 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 wtfew ###
   recode weight to wtfew 40-79=60 80-120=100
   histogram meefew
   ```

   also give a wrong picture of the underlying distribution of the WEIGHT 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 provides a quick way of recoding data into equal width intervals using the BY keyword. The commands:

```
define wtquick <AAAAAAAAAA>
if weight <> . then recode weight to wtquick by 10
```

will recode the WEIGHT variable into groups of width ten. Note that the recode line begins with the expression "if weight <> ." so that missing values of WEIGHT will continue to be missing for the WTQUICK variable.

Examine the distribution of the WTQUICK variable using the FREQ and HISTOGRAM commands.

```
freq wtquick
histogram wtquick
```

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.

With the **San Francisco Men's Health Study** data, it is possible to explore whether non-white race is associated with HIV infection status. The RACE variable is distributed as follows:

```
 RACE |  Freq  Percent   Cum.
------+----------------------
    1 |   897   86.6%    86.6%
    2 |    52    5.0%    91.6%
    3 |    52    5.0%    96.6%
    4 |    35    3.4%   100.0%
------+----------------------
Total |  1036  100.0%
```

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

```
define white #
```

and then use the RECODE command to create the two groups:

```
recode race to white 1=1 2,3,4=2
```

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

```
browse race white
```

Now that we have two variables WHITE and HIVSTAT 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 non-white race is associated with HIV infection status:

```
tables white hivstat
```

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 **San Francisco Men's Health Study** data, the command:

```
tables poppers hivstat
```

produces the following table:

```
                    HIVSTAT
    POPPERS |          1          2 | Total
------------+----------------------+------
          1 |        348        272 |   620
          2 |         60        353 |   413
------------+----------------------+------
      Total |        408        625 |  1033
```

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

```
              a =  348 = number exposed and ill
              b =  272 = number exposed and not ill
              c =   60 = number unexposed and ill
              d =  353 = number unexposed and not ill
          a + b =  620 = number exposed
          c + d =  413 = number unexposed
          a + c =  408 = number ill
          b + d =  625 = number not ill
  a + b + c + d = 1033 = number in study (excluding those with missing values)
```

In this example *exposed* means use of poppers within the past two years and *ill* means being infected with HIV. Make sure that you can identify these figures in the table before proceeding.

Two-by-two tables

The crude risk

The *crude 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:

```
                    HIVSTAT
    POPPERS |         1          2 | Total
------------+----------------------+------
          1 |       348        272 |   620
          2 |        60        353 |   413
------------+----------------------+------
      Total |       408        625 |  1033
```

The crude risk can be calculated as:

```
(a + c) / (a + b + c + d) = 408 / 1033 = 0.395 = 39.5%
```

Each person in the **San Francisco Men's Health Study** ran a risk of 39.5% of being infected with HIV. This is the same as saying that 39.5% of people in the study can be considered as 'cases' as defined by their HIV status.

The crude 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 a crude 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 a crude 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 HIV status among popper users and non-popper users using 20 cases (a + c) and 20 controls (b + d) the crude 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 crude risk would appear to be:

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

Both of these estimates differ from the *true* crude risk among the 1033 persons in the study. It is **not** valid to calculate a crude 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:

POPPERS	HIVSTAT 1	2	Total
1	348	272	620
2	60	353	413
Total	408	625	1033

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

```
a / (a + b) = 348 / (348 + 272) = 348 / 620 = 0.561 = 56.1%
```

Each person in the study who used poppers in the last 2 years ran a risk of 56.1% of being defined as a ‘case’ based on their HIV status. This is the same as saying that 56.1% of popper-users 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 crude 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:

```
                    HIVSTAT
    POPPERS |         1          2 | Total
------------+----------------------+------
          1 |       348        272 |   620
          2 |        60        353 |   413
------------+----------------------+------
      Total |       408        625 |  1033
```

It is possible to calculate the risk of being a 'case' for popper-users as:

```
a / (a + b) = 348 / (348 + 272) = 348 / 620 = 0.561 = 56.1%
```

and the risk of being a 'case' for **non**-popper-users as:

```
c / (c + d) = 60 / (60 + 353) = 60 / 413 = 0.145 = 14.5%
```

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)) = (348 / 620) / (60 / 413) = 0.561 / 0.145 = 3.86
```

A popper-user was 3.86 times as likely to be a 'case' as a non-popper-user.

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, where subjects have been selected without the investigator knowing either exposure or outcome status.

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:

```
                    HIVSTAT
    POPPERS |          1          2 | Total
------------+----------------------+------
          1 |        348        272 |   620
          2 |         60        353 |   413
------------+----------------------+------
      Total |        408        625 |  1033
```

In this example the *odds* of being a 'case' among popper-users can be calculated as:

```
a / b = 348 / 272 = 1.28
```

The odds of being a 'case' among non-popper-users can be calculated as:

```
c / d = 60 / 353 = 0.170
```

The odds ratio can be calculated as:

```
(a / b) / (c / d) = (348 / 272) / (60 / 353) = 1.28 / 0.170 = 7.53
```

Use of poppers was associated with a 7.53 fold increase in the odds of being HIV infected. 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 **San Francisco Men's Health Study** data because it is a cross-sectional study. It would **not** be valid to use the relative risk if a certain number of HIV-positive cases and HIV-negative controls had been selected for the study (e.g. 500 cases and 500 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:

POPPERS	HIVSTAT 1	2	Total
1	348	272	620
2	60	353	413
Total	408	625	1033

of POPPERS by HIVSTAT closely we can extract the following information:

exposed and ill	poppers = 1 **and** hivstat = 1	a	**348**
exposed and not ill	poppers = 1 **and** hivstat = 2	b	**272**
unexposed and ill	poppers = 2 **and** hivstat = 1	c	**60**
unexposed and not ill	poppers = 2 **and** hivstat = 2	d	**353**
exposed	poppers = 1	a + b	**620**
unexposed	poppers = 2	c + d	**413**
ill	hivstat = 1	a + c	**408**
not ill	hivstat = 2	b + d	**625**
total records	sum of all the cells in the table	a + b + c + d	**1033**

Identify these figures in the displayed table.

We can use this information to calculate various risk measures:

crude risk	(a + c) / (a + b + c + d)	408 / 1033	**0.39**
risk among exposed	a / (a + b)	348 / 620	**0.56**
risk among unexposed	c / (c + d)	60 / 413	**0.15**
relative risk	(a / a + b) / (c / c + d)	(348/620) / (60/413)	**3.86**
odds ratio	(a / b) / (c / d)	(348/272) / (60/353)	**7.53**

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* (note that Relative Risk and Risk Ratio are two names for the same risk measure):

```
Odds ratio                                                          7.53
Cornfield 95% confidence limits for OR                 5.41 < OR <  10.50

RISK RATIO(RR)(Outcome:HIVSTAT=1; Exposure:POPPERS=1)               3.86
95% confidence limits for RR                            3.03 < RR <  4.93
```

and will calculate the crude 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 poppers hivstat
```

This command now produces the following output:

```
                          HIVSTAT
    POPPERS |              1              2 | Total
------------+---------------------------------+------
          1 |            348            272 |   620  <- Counts
            >          56.1%          43.9% > 60.0%  <- Row percentages
            |          85.3%          43.5% |        <- Column percentages
          2 |             60            353 |   413
            >          14.5%          85.5% > 40.0%
            |          14.7%          56.5% |
------------+---------------------------------+------
      Total |            408            625 |  1033
            |          39.5%          60.5% |
```

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 POPPERS = "1" and HIVSTAT = "1" is equivalent to the exposure specific risk or attack rate (56.1%) and the row percentage of the column total for HIVSTAT = "1" is the crude risk (39.5%). 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 exposure-variable outcome-variable
```

The risk or exposure variable (e.g. POPPERS) should **always** be the first variable you specify with the TABLES command. The outcome variable (e.g. HIVSTAT) 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, crude 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 **San Francisco Men's Health Study** data (using the command TABLES POPPERS HIVSTAT) for the association between use of poppers within the past 2 years and HIV status:

```
RISK RATIO(RR)(Outcome:HIVSTAT=1; Exposure:POPPERS=1)                    3.86
95% confidence limits for RR                              3.03 < RR <    4.93
```

The observed relative risk is 3.86. The true relative risk is likely to lie somewhere between 3.03 and 4.93. The true relative risk is unlikely to be equal to one. It is reasonable to state that the use of poppers was associated with being HIV-positive.

The following relative risk and 95% confidence limits have been calculated from the **San Francisco Men's Health Study** data for the association between being white and being HIV-positive:

```
RISK RATIO(RR)(Outcome:HIVSTAT=1; Exposure:WHITE=1)                      1.07
95% confidence limits for RR                              0.85 < RR <    1.36
```

The observed relative risk is 1.07. The true relative risk is likely to lie somewhere between 0.85 and 1.36. It is **not** unlikely that the true relative risk is equal to one. It is **not** reasonable to state that white race was associated with being HIV-positive.

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:

POPPERS	HIVSTAT 1	2	Total
1	348	272	620
2	60	353	413
Total	408	625	1033

The *null hypothesis* is that the use of poppers within the past 2 years is **not** associated with being HIV-infected. The *alternative hypothesis* is that the use of poppers within the past 2 years **is** associated with being HIV-infected. 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' (*crude risk*) is:

```
(a + c) / (a + b + c + d) = 408 / 1033 = 0.395 = 39.5%
```

If the null hypothesis were true, the *expected* number of people who are HIV-positive and have used poppers within the past 2 years would be 39.5% of 620 (i.e. the risk of being a case would **not** vary with exposure), or:

```
Expected[a] = 620 * (408 / 1033) = 244.88
```

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

```
Expected[a] = 620 * (408 / 1033) = 244.88

Expected[b] = 620 * (625 / 1033) = 375.12

Expected[c] = 413 * (408 / 1033) = 163.12

Expected[d] = 413 * (625 / 1033) = 249.88
```

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

POPPERS | HIVSTAT 1	2 | Total
1	348	272	620
2	60	353	413
Total	408	625	1033

EXPECTED

POPPERS | HIVSTAT 1	2 | Total
1	244.88	375.12	620.00
2	163.12	249.88	413.00
Total	408.00	625.00	1033.00

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

POPPERS | HIVSTAT 1	2 | Total
1	103.12	-103.12	00.00
2	-103.12	103.12	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:

POPPERS | HIVSTAT 1	2 | Total
1	10633.73	10633.73	21267.47
2	10633.73	10633.73	21267.47
Total	21267.47	21267.47	42534.94

before dividing them by the *expected* values:

POPPERS | HIVSTAT 1	2 | Total
1	43.42	28.35	71.77
2	65.19	42.56	107.75
Total	108.61	70.91	179.52

The overall total of this table (179.52) 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* (8) 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 179.52 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 the use of poppers and HIV status. If the probability of obtaining 179.52 under the null hypothesis is large then the null hypothesis might be true and we would not assume that there is an association between the use of poppers and HIV status.

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:

	HIVSTAT		
POPPERS	1	2	Total
1	348	??	620
2	??	??	413
Total	408	625	1033

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
```

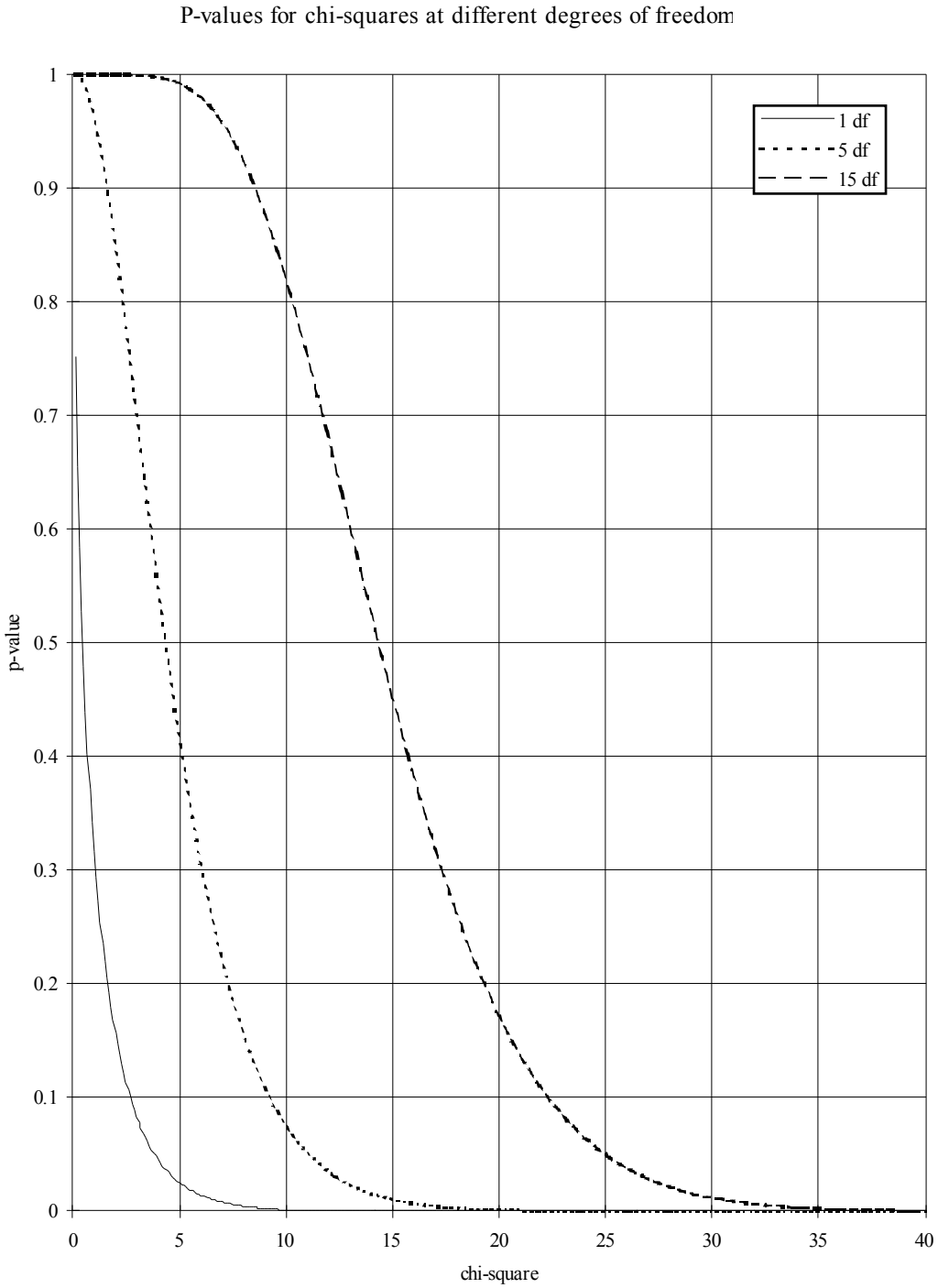
The chart on the next page shows the probability of observing chi-square values at different degrees of freedom. Look at the chi-square distribution for one degree of freedom. The probability of getting a chi-square value of 179.52 from a 2-by-2 table is very small. This probability is referred to as 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 0.0001. The probability that the observed association arose by chance is less than 0.0001 (i.e. less than one in ten 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 studies 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 variable in one direction only (i.e. does use of poppers increase the probability of being HIV-positive?).

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:



The value of 179.52 on the x-axis (chi-square on *x-axis*) is off the chart, but you can still 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 SMOKER and being a HIVSTAT:

	HIVSTAT		
SMOKER	1	2	Total
1	187	242	429
2	221	382	603
Total	408	624	1032

This table yields a chi-square statistic of 5.05 with a p-value of 0.025. 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:

	HIVSTAT		
SMOKER	1	2	Total
1	561	726	1287
2	663	1146	1809
Total	1224	1872	3096

which yields a chi-square statistic of 15.15 (3 * 5.05) with a P-value of0.0000993. 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:

```
RISK RATIO(RR)(Outcome:HIVSTAT=1; Exposure:SMOKER=1)                    1.19
95% confidence limits for RR                              1.02 < RR <   1.38
```

for the original table and:

```
RISK RATIO(RR)(Outcome:HIVSTAT=1; Exposure:SMOKER=1)                    1.19
95% confidence limits for RR                              1.09 < RR <   1.30
```

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 POPPERS HIVSTAT. Important information has been highlighted:

```
                    HIVSTAT
    POPPERS |          1          2 | Total
------------+----------------------+------
          1 |        348        272 |   620
          2 |         60        353 |   413
------------+----------------------+------
      Total |        408        625 |  1033

                          Single Table Analysis

Odds ratio                                                    7.53 <- Odds Ratio
Cornfield 95% confidence limits for OR         5.41 < OR <   10.50 <- Confidence Limits
Maximum likelihood estimate of OR (MLE)                       7.51
Exact 95% confidence limits for MLE            5.44 < OR <   10.50
Exact 95% Mid-P limits for MLE                 5.50 < OR <   10.37
Probability of MLE >=  7.51 if population OR = 1.0     0.00000000

RISK RATIO(RR)(Outcome:HIVSTAT=1; Exposure:POPPERS=1)         3.86 <- Relative Risk
95% confidence limits for RR                   3.03 < RR <    4.93 <- Confidence Limits
               Ignore risk ratio if case control study

                        Chi-Squares   P-values
                        -----------   --------
        Uncorrected:       179.52     0.00000000 <---            <- Chi-square/p-value
        Mantel-Haenszel:   179.35     0.00000000 <---
        Yates corrected:   177.78     0.00000000 <---            <- 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 race hivstat
```

to investigate the association between race and HIV status. The following table is produced:

```
                HIVSTAT
RACE        |     1     2 | Total
------------+-------------+------
         1  |   354   537 |   891
         2  |    22    30 |    52
         3  |    20    29 |    49
         4  |     8    26 |    34
------------+-------------+------
     Total  |   404   622 |  1026

            Chi square =       3.85          <- Chi-square
     Degrees of freedom =          3
               p value = 0.27773826          <- p-value
```

If you examine the statistics that follow the table you will see that RACE does not appear to be associated with being HIVSTAT. The p-value that ANALYSIS reports for the chi-square statistic is 0.27773826 (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:

```
                HIVSTAT
RACE        |     1     2 | Total
------------+-------------+------
         1  |   354   537 |   891      <- Counts
            > 39.7% 60.3% > 86.8%      <- Row percentages
            | 87.6% 86.3% |           <- Column percentages
         2  |    22    30 |    52
            > 42.3% 57.7% >  5.1%
            |  5.4%  4.8% |
         3  |    20    29 |    49
            > 40.8% 59.2% >  4.8%
            |  5.0%  4.7% |
         4  |     8    26 |    34
            > 23.5% 76.5% >  3.3%
            |  2.0%  4.2% |
------------+-------------+------
     Total  |   404   622 |  1026
            | 39.4% 60.6% |
```

If you examine the row percentages, you will see that the proportion of HIV-positives for are similar for **all** values of the RACE variable. In this case HIVSTAT is said to be *unrelated* to RACE (i.e. HIVSTAT and RACE are **not** associated with each other). This observation can be quantified with the chi-square and above p-value.

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 popper use within the past two years to be associated with HIV infection (RR = 3.86, 95% C.I. 3.03 - 4.93).'*

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

> *'We found popper use within the past two years to be associated with HIV infection (Yates chi-square = 177.78, p < 0.0001).'*

Some may want both:

> *'We found popper use within the past two years to be associated with HIV infection (RR = 3.86, 95% C.I. 3.03 - 4.93, Yates chi-square = 177.78, 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 race and HIV infection (chi-square = 3.85, df = 3, p = 0.28).'*

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.

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

Use the RECODE or IF ... THEN commands to recode the CD4 variable into two groups ('1LOW' and '2NORMAL'). Use the lower quartile of the distribution of the CD4 variable for the '1LOW' group and the upper three quartiles for the '2NORMAL' group. You can find the quartiles of the distribution of the CD4 variable using the DESCRIBE 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 your new variable and HIVSTAT.

Use the TABLES command to investigate the association between SEXPREF, SMOKER, and POPPERS and your new variable.

Tables with subsets of data

Selecting cases for ANALYSIS

With data from the **San Francisco Men's Health Study,** it is possible to explore whether Blacks and Hispanics have different risks of being HIV-infected. The RACE variable records which group each individual case belongs to:

```
RACE  |  Freq  Percent   Cum.
------+----------------------
   1  |   897   86.6%    86.6%
   2  |    52    5.0%    91.6%
   3  |    52    5.0%    96.6%
   4  |    35    3.4%   100.0%
------+----------------------
Total |  1036  100.0%
```

Use the SELECT command to instruct analysis to ignore the data from the "white" (RACE = 1) and "other" (RACE = 4) individuals:

```
select race <> 1
select race <> 4
```

and use the TABLES command to investigate the association between RACE and HIVSTAT:

```
tables race hivstat
```

which produces the following two-by-two table:

```
                        HIVSTAT
      RACE  |          1          2 | Total
------------+-----------------------+------
          2 |         22         30 |    52
          3 |         20         29 |    49
------------+-----------------------+------
      Total |         42         59 |   101

                           Single Table Analysis

Odds ratio                                                          1.06
Cornfield 95% confidence limits for OR                 0.44 < OR <  2.55
Maximum likelihood estimate of OR (MLE)                             1.06
Exact 95% confidence limits for MLE                    0.45 < OR <  2.53
Exact 95% Mid-P limits for MLE                         0.48 < OR <  2.37
Probability of MLE >=  1.06 if population OR = 1.0            0.52005279

RISK RATIO(RR)(Outcome:HIVSTAT=1; Exposure:RACE=2)                  1.04
95% confidence limits for RR                           0.65 < RR <  1.65

               Ignore risk ratio if case control study

                         Chi-Squares   P-values
                         -----------   --------

        Uncorrected:         0.02     0.87920065
        Mantel-Haenszel:     0.02     0.87979558
        Yates corrected:     0.00     0.96012694
```

Note that the SELECT command has excluded those cases where RACE = 1 or RACE = 4 from the analysis (the '< >' operator means 'not equal to ...').

Examine the relative risk and its associated confidence limits. The relative risk is greater than 1 indicating a higher risk of being HIV-infected for Hispanics than for Blacks. The confidence interval includes one indicating that the observed association (or increase in risk) is likely to be due to random variation. The p-value for Yates chi-square is 0.96 which also indicates that the observed association (or increase in risk) is likely to be due to random variation.

The data do **not** support the hypothesis that Blacks and Hispanics have different risks of being HIV-infected (RR = 1.06, 95% C.I. 0.65 - 1.65, Yates chi-square = 0.00, p = 0.96).

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10.

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

Use the RECODE or IF ... THEN commands to recode the HIVLOAD variable into two groups ('1HIGH' and '2LOW'). Use the lower three quartiles of the distribution of this variable for the '2LOW' group and the upper quartile for the '1HIGH' group. You can find the quartiles of the distribution of a variable using the DESCRIBE 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 RACE and your new variable. Does the evidence support the hypothesis that Blacks are at greater risk of high HIV load relative to Hispanics?

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. 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.

If you haven't already done so, clear the race selections you made earlier by typing:

```
select
```

At the top of the screen should be a message saying "All records selected."

For the next analysis we will compare the "White" group to the "Non-white" group by combining "Hispanic," "Black," and "Other" into one category. To do this, type:

```
define white #
recode race to white 1=1 2,3,4=2
```

Now have a look at the mean HEIGHT for each category of the WHITE variable by issuing the command:

```
means height white /n
```

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

```
WHITE             Obs      Total       Mean   Variance    Std Dev
1                 892     158459    177.645     45.991      6.782
2                 136      23720    174.412     56.229      7.499
Difference                           3.233

WHITE         Minimum     25%ile     Median     75%ile    Maximum       Mode
1             160.000    173.000    178.000    183.000    196.000    178.000
2             160.000    168.000    175.000    180.000    193.000    173.000
```

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 a line chart. Issue the command:

```
line height white
```

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

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

```
means height white /n
```

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

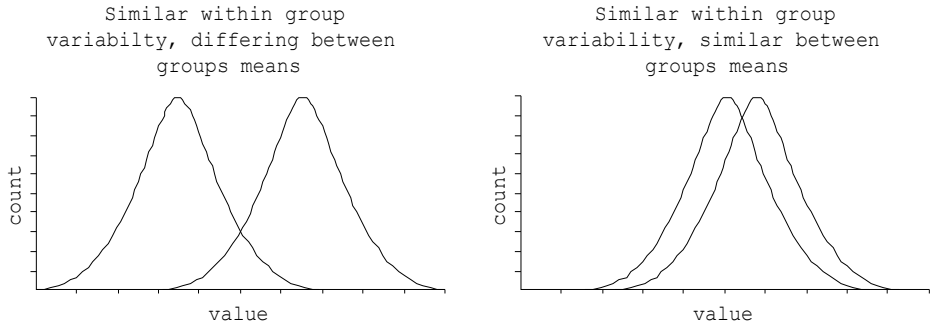
Differences in means

Analysis of variance

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

1. Variability of the data **within** each group around the group mean
2. Variability of the means **between** groups

If HEIGHT doesn't vary much between individuals within the a WHITE group (e.g. white participants have similar heights, and non-white participants have similar heights) 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:

$$SSW = \sum_{i=1}^{k} (N_i - 1) S_i^2$$

where S_i^2 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 = (892 - 1) * 45.991 + (136 - 1) * 56.229 = 48568.896
```

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 overall mean. For our data:

```
SSB = 892 * (177.645 - 177.217)^2 + 136 * (174.412 - 177.217)^2 = 1233.451528
```

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:

$$F = \frac{between \text{ groups means square}}{within \text{ groups means square}} = \frac{1233.452}{47.338} = 26.054$$

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 20 values of HEIGHT for the WHITE group and 18 values of HEIGHT for the non-WHITE group:

WHITE	193	180	178	188	183	185	178	183	183	180
	180	180	178	178	165	175	193	168	173	175
NON-WHITE	178	173	168	165	168	173	180	163	163	173
	183	163	165	168	173	165	183	165		

The means and variances for this data are:

Group	Mean	Variance
BOTH	175.395	68.894
WHITE	179.800	49.326
NON-WHITE	170.500	46.618

The within groups sum of squares is:

```
SSW = (20 - 1) * 49.326 + (18 - 1) * 46.618 = 1729.700
```

The within group degrees of freedom is:

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

The within group mean square is:

```
1729.700 / 36 = 48.047
```

The between groups sum of squares is:

$$SSB = 20 * (179.800 - 175.395)^2 + 18 * (170.500 - 175.395)^2 = 819.379$$

The between group degrees of freedom is:

```
2 - 1 = 1
```

The between groups mean square is:

```
819.379 / 1 = 819.379
```

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

```
F = 819.379 / 48.047 = 17.054
```

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 $p < 0.05$.

ANOVA in ANALYSIS

An example of ANALYSIS output

The command:

```
means height white /n
```

produces the following summary statistics and *ANOVA table*:

```
WHITE             Obs      Total       Mean   Variance    Std Dev
1                 892     158459    177.645     45.991      6.782
2                 136      23720    174.412     56.229      7.499
Difference                           3.233

WHITE         Minimum     25%ile     Median     75%ile    Maximum       Mode
1             160.000    173.000    178.000    183.000    196.000    178.000
2             160.000    168.000    175.000    180.000    193.000    173.000

                                ANOVA
                    (For normally distributed data only)

Variation          SS   df          MS  F statistic    p-value
Between      1233.340    1    1233.340       26.054   0.000000  <- F-Statistic/p-value
Within      48569.285 1026      47.338
Total       49802.625 1027

               Bartlett's test for homogeneity of variance
    Bartlett's chi square =   2.481  deg freedom =  1   p-value = 0.115240

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

In this case the p-value is less than 0.0001 which is evidence that the observed difference in the means of the two groups is unlikely to have arisen by chance.

Issue the command:

```
means height white /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 White and Non-white groups had differing mean heights (White mean = 177.645, Non-white mean = 174.412, F = 26.054, df = (1,1026), p < 0.0001)'*

Practical exercise

Variables and variable names

For a full explanation of variables, their names, and their codes refer to the table on page 10.

Exercise 7 - Analysis of variance (ANOVA)

Use the MEANS command to test the null hypothesis that the means of the WEIGHT and AGE variables for White and Non-white study participants are the same.

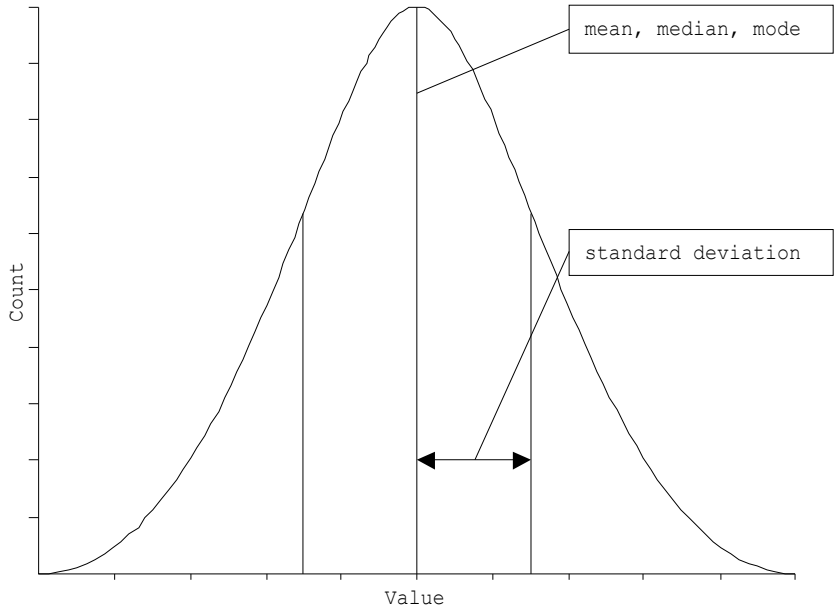
Problems with ANOVA

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 White group and the Non-white 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 HEIGHT WHITE:

```
                    MEANS of HEIGHT for each category of WHITE

WHITE             Obs      Total       Mean   Variance    Std Dev
1                 892     158459    177.645     45.991      6.782
2                 136      23720    174.412     56.229      7.499
Difference                           3.233

WHITE         Minimum     25%ile     Median     75%ile    Maximum       Mode
1             160.000    173.000    178.000    183.000    196.000    178.000
2             160.000    168.000    175.000    180.000    193.000    173.000
```

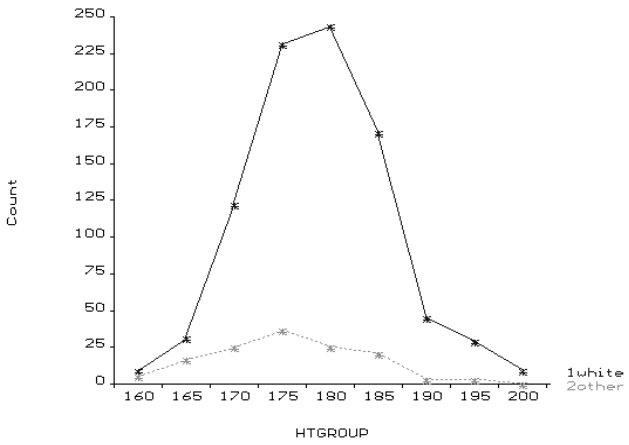
We can examine the assumption of *normality* by comparing the mean, mode and median within each group. If the data is *normally* distributed we would expect the mean, median, and mode to take similar values. In our data this is true for both groups. The assumption that the variances are equal can be 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 =   2.481  deg freedom =  1   p-value = 0.115240

            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 HEIGHT has been grouped to better show the underlying distribution:



This also shows that the HEIGHT data in the both groups is approximately normally distributed. This means that it is 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.

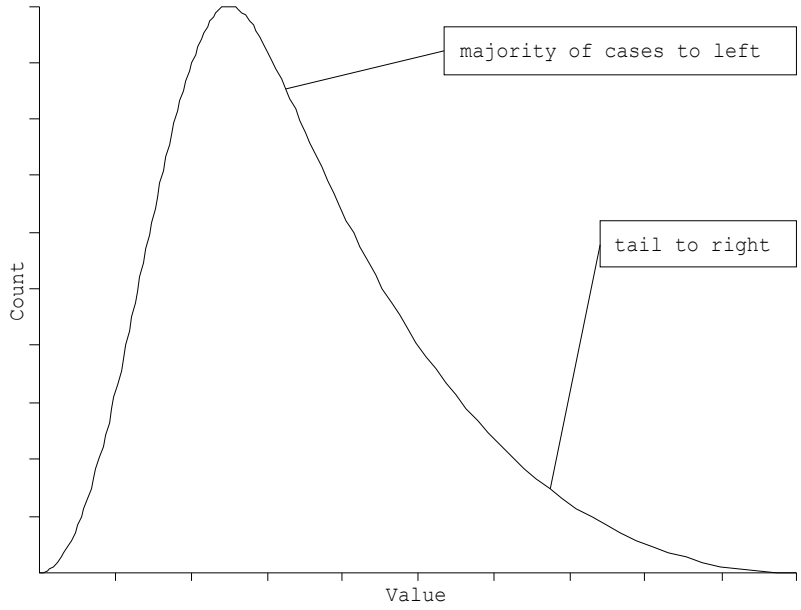
Exercise 8 - Testing the ANOVA assumptions

Use the MEANS command to test the ANOVA assumptions for WEIGHT and AGE grouped for the White and Non-white study participants. Test the ANOVA assumptions 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.

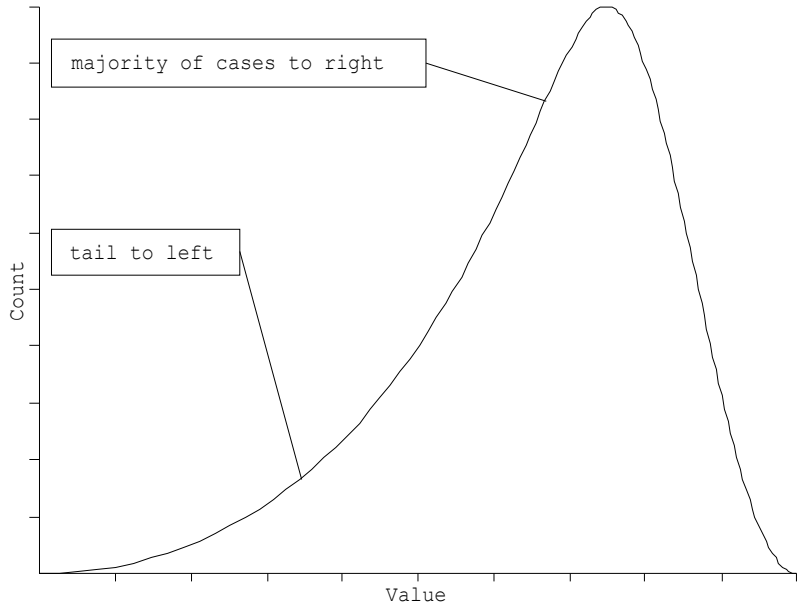
When ANOVA is unsuitable

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. 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 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.

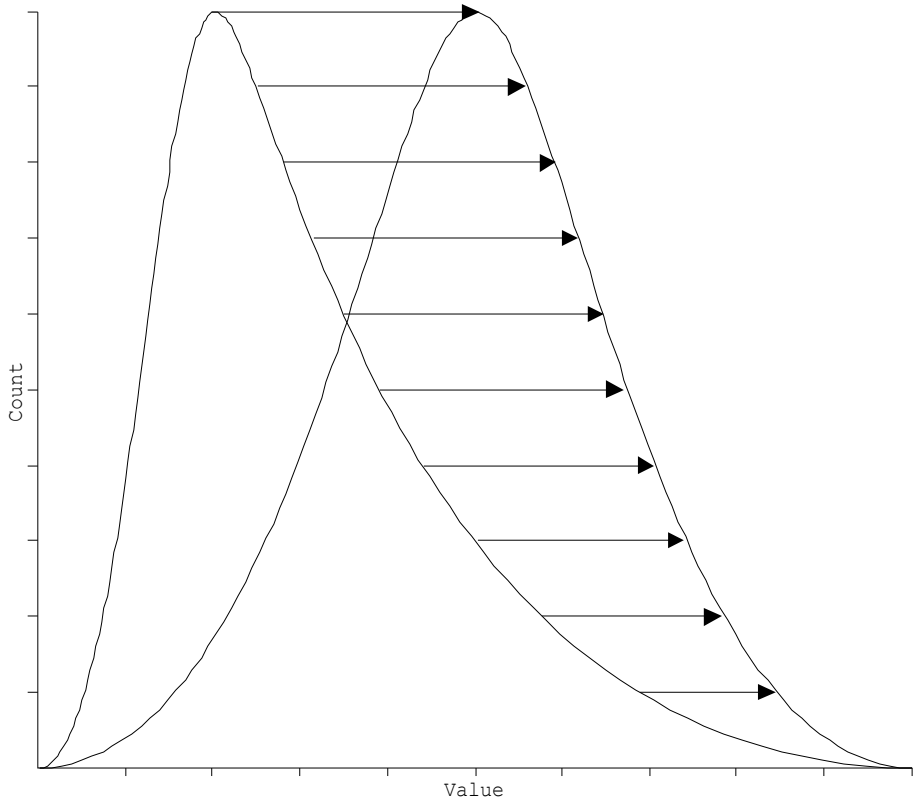
What can be done?

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.

Checking the data

Checking the data for skew and unequal variance

Issue the command:

```
means npartner white
```

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

WHITE	Obs	Total	**Mean**	Variance	Std Dev
1	895	35890	**40.101**	10005.390	100.027
2	138	2520	**18.261**	1524.895	39.050
Difference			**21.840**		

WHITE	Minimum	25%ile	**Median**	75%ile	Maximum	**Mode**
1	0.000	1.000	**10.000**	30.000	998.000	0.000
2	0.000	0.000	**4.000**	16.000	300.000	**0.000**

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 =   1.3E+02  deg freedom =  1   p-value = 0.000000

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

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 LOGPART to hold the logarithm of the NPARTNER variable:

```
define logpart #.####
```

and assign a value to this variable:

```
logpart = log(npartner + 1)
```

The variable LOGPART now contains the logarithm of the NPARTNER variable. We need to add 1 to NPARTNER 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. Issue the commands:

```
select
select logpart <> . and white <> .
means logpart white /n
```

The first line clears any previous selections, and the second and third lines eliminate any cases for which LOGPART and WHITE are missing. You should do this before any analysis using two or more variables.

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

WHITE	Obs	Total	Mean	Variance	Std Dev
1	895	891	0.996	0.559	0.747
2	138	106	0.770	0.430	0.656
Difference			0.226		

WHITE	Minimum	25%ile	Median	75%ile	Maximum	Mode
1	0.000	0.301	1.041	1.491	3.000	0.000
2	0.000	0.000	0.699	1.230	2.479	0.000

The standard deviations in the two groups are also similar. Note that the p-value for Bartlett's test is borderline at p=0.05.

```
                Bartlett's test for homogeneity of variance
Bartlett's chi square =   3.796  deg freedom =  1   p-value = 0.051368

            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. However, since p is very close to 0.05, we may still want to use the non-parametric Kruskal-Wallis (Wilcoxon-Mann-Whitney).

ANOVA and transformed data

What is the antilogarithm of the mean of the logarithms?

We have tried to transform the data to meet the assumptions of the ANOVA technique. If the assumption of equal variance had been met, we could have used this new transformed data to test our hypothesis. As a demonstration, issue the command:

```
means logpart white /n
```

which produces the following output:

```
WHITE             Obs      Total       Mean   Variance    Std Dev
1                 895        891      0.996      0.559      0.747
2                 138        106      0.770      0.430      0.656
Difference                           0.226

WHITE         Minimum     25%ile     Median     75%ile    Maximum       Mode
1               0.000      0.301      1.041      1.491      3.000      0.000
2               0.000      0.000      0.699      1.230      2.479      0.000

                                ANOVA
                    (For normally distributed data only)

Variation          SS   df          MS  F statistic    p-value    t-value
Between         6.097    1       6.097       11.259   0.000821   3.355423
Within        558.281 1031       0.541
Total         564.378 1032
```

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 White group the 'average' number of partners is $10^{0.996} = 9.908$, and for the Non-white group it is $10^{0.770} = 5.888$. 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

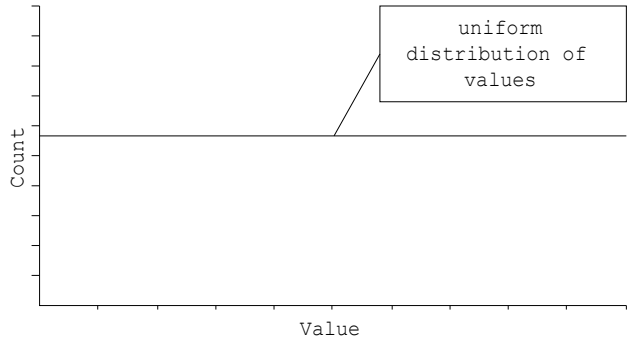
If you are using ANOVA techniques to analyse your data and have applied a logarithmic transformation to your data you should quote the geometric means for each group together with the *F*-statistic, degrees of freedom, and p-value. A typical quote might be:

> *'The White study participants had a higher geometric mean numbers of partners compared to Non-whites (White = 9.908, Non-white = 5.888, F = 11.259, df = 1, P < 0.001)'*

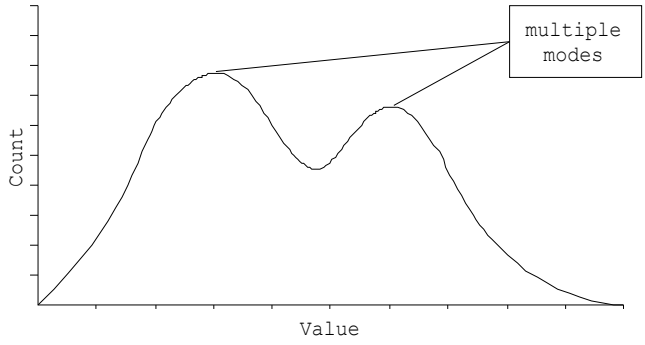
Can't transform! Won't transform!

Difficult distributions

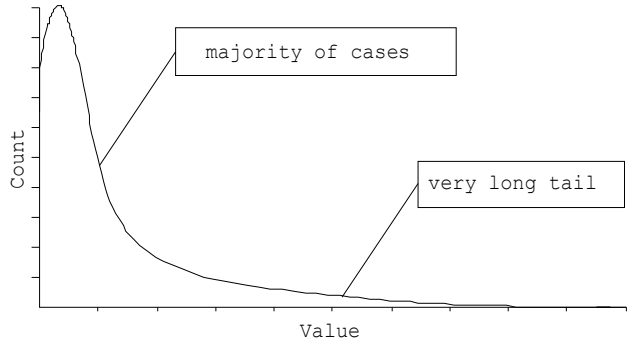
Sometime when you try to transform your data it just won't approach normality. This is often the case with distributions such as *uniform* distributions:



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 the White and Non-white 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 HEIGHT values from five White study participants and six Non-white study participants. These data are shown below:

WHITE	NONWHITE
193	178
168	163
178	173
170	168
180	185
	183

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:

WHITE	Ranks		NON-WHITE
		1.0	163
168	2.5	2.5	168
170	4.0		
		5.0	173
178	6.5	6.5	178
180	8.0		
		9.0	183
		10.0	185
193	11.0		
	32.0		

The ranks of the observations in the group with the smallest sample size (in this case the White group) are summed to produce a statistic called the *W-Statistic* or the *sum of the ranks*:

```
W  = 2.5 + 4 + 6.5 + 8 + 11 = 32
```

On the basis of the *null hypothesis* that the distributions of ranks in both groups are the same, the distribution of the *W* 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 HEIGHT for the Non-white group is:

```
(173 + 178) / 2 = 175.5
```

Non-parametric tests in ANALYSIS

An example of ANALYSIS output

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

```
WHITE             Obs      Total       Mean   Variance    Std Dev
1                 895      35890     40.101  10005.390    100.027
2                 138       2520     18.261   1524.895     39.050
Difference                           21.840

WHITE         Minimum     25%ile     Median      75%ile    Maximum       Mode
1               0.000      1.000     10.000      30.000    998.000      0.000
2               0.000      0.000      4.000      16.000    300.000      0.000

                                  ANOVA
                       (For normally distributed data only)

Variation          SS   df          MS  F statistic    p-value    t-value
Between     57028.862    1   57028.862        6.423   0.011411   2.534415
Within    9153729.558 1031    8878.496
Total     9210758.420 1032

                Bartlett's test for homogeneity of variance
    Bartlett's chi square =   1.3E+02  deg freedom =  1   p-value = 0.000000

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


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

Kruskal-Wallis H (equivalent to Chi square) =      11.376
                        Degrees of freedom =           1
                                  p value =    0.000744
```

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. *W* or Kruskal-Wallis H), the degrees of freedom, and p-value. A typical quote might be:

> *'The White study participants had more male sexual partners in the previous two years than the Non-white study participants (Kruskal-Wallis H, White median = 10.0, Non-white median = 4.0, chi-square = 11.376, df = 1, p < 0.001)'*

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!

Use the techniques and commands presented thus far to test the hypotheses that:

1. Whites have a higher number of male sexual partners than Non-whites.

2. Higher educational status is associated with a lower CD4 count.

Two continuous variables

Working with different variable types

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

Type	Description
Qualitative/ Categorical Data	Cases belong to a defined *category* or group such as RACE, EDUCational status, or SEXPREF. A particularly common type of categorical variable is a *binary categorical* variable where cases can belong to one of two alternative groups. The variables SEXPREF and SMOKER are examples of *binary categorical* variables in the **San Francisco Men's Health Study** dataset.
Qualitative/ Numerical Data	Unmodified numerical measures. There are two types of numeric data. *Discrete* data is limited to discrete *integers*. A count of cases over time would be a *discrete* variable. *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.

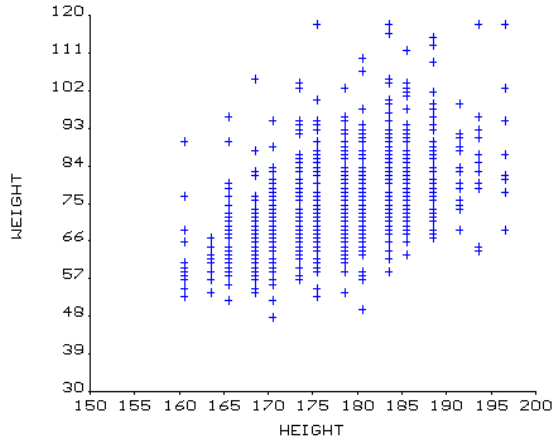
2. The choice of cut-points for group membership is often arbitrary. The choice of cut-point can make a difference to the results of any analysis.

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

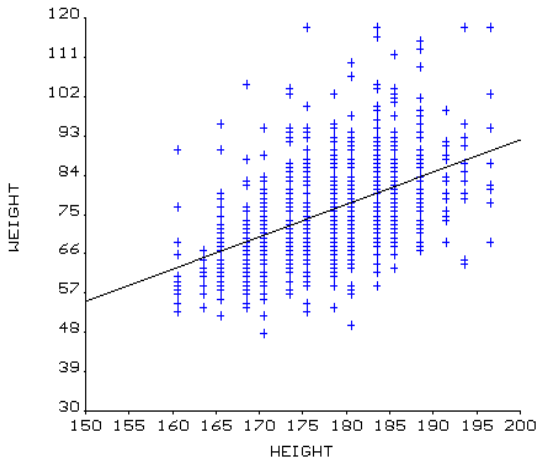
Examining associations

Using scatter plots to examine associations

The chart below shows a scatter plot of the HEIGHT and WEIGHT variables for each case:



There appears to be an association between HEIGHT and WEIGHT. It is a *positive association* because as HEIGHT increases so does WEIGHT. 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

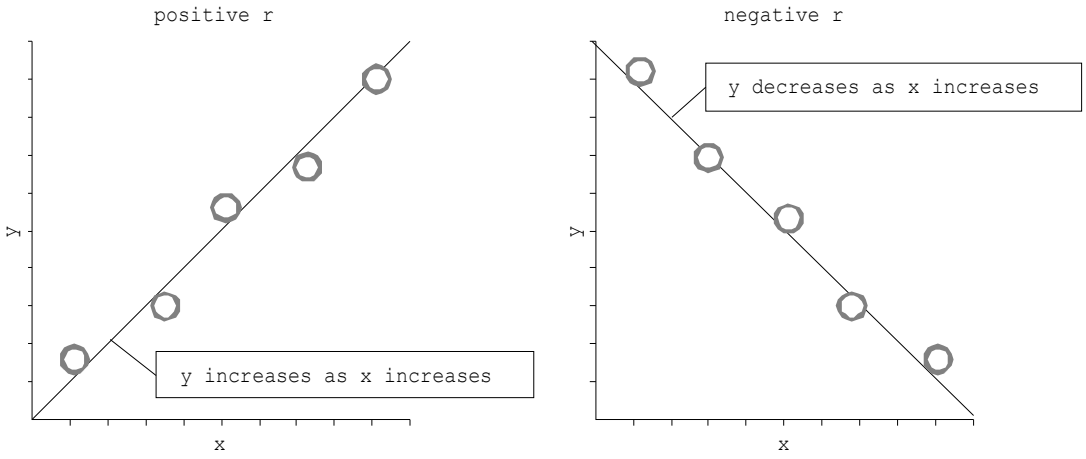
The strength of an association can be measured by calculating the *Pearson correlation coefficient*:

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

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:



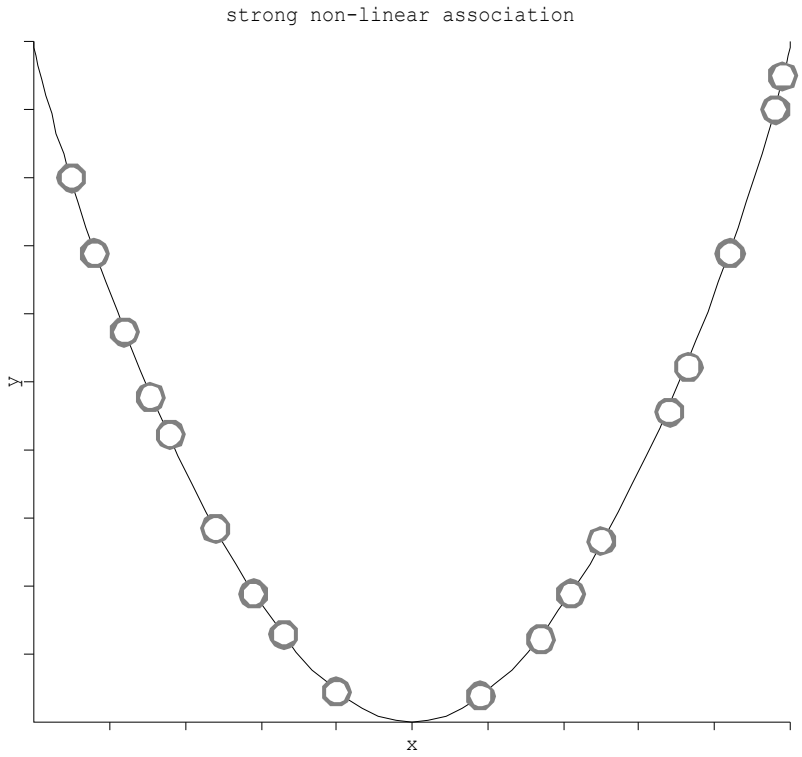
A value of zero (0) indicates no *linear* association.

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 79 - 80).

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).

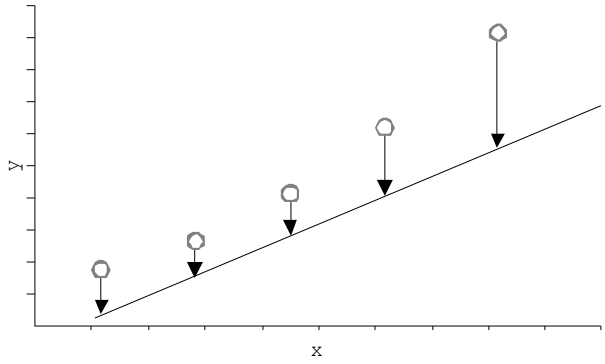
1. Make the variances (standard deviations) similar.

1. Make an association more linear.

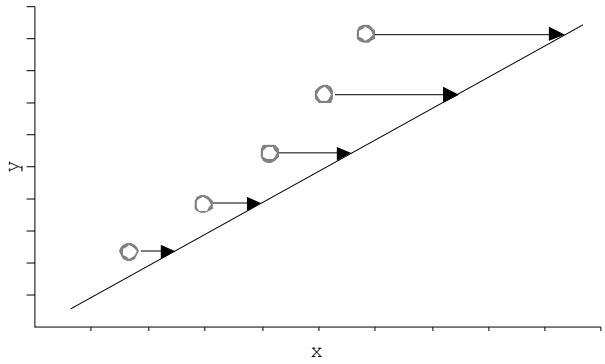
The first two applications have already been covered (see pages 65 - 70). All applications are important but, with correlation and regression analysis, the goal of making an association more linear is the most important.

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).

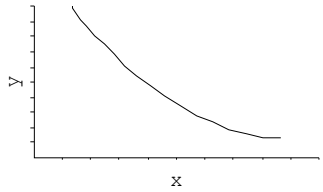
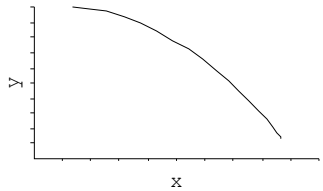
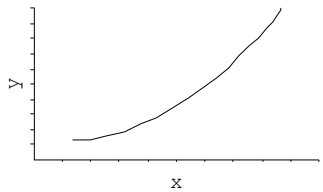
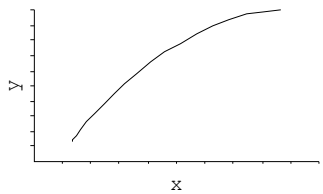
1. Make the variances (standard deviations) similar.

1. 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 San Francisco Men's Health Study dataset by typing:

```
read sfmhshiv.rec
```

Examine the relationship between the HEIGHT and WEIGHT variables with the command:

```
scatter height weight
```

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

```
scatter height weight /r
```

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

```
regress weight = height
```

ANALYSIS displays the Pearson correlation coefficient:

```
Correlation coefficient: r   = 0.47                <- correlation coefficient
                         r^2 = 0.22
95% confidence limits:   0.16 < r^2 < 0.28
```

You may also have noticed that ANALYSIS calculated another measure called *r^2* (*r-squared*) and 95% confidence limits. 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   = 0.47
                         r^2 = 0.22                <- coefficient of determination
95% confidence limits:   0.16 < r^2 < 0.28         <- confidence limits
```

To determine the confidence limits for the Pearson correlation coefficient, take the square root of the confidence limits for r^2. 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.

The coefficient of determination is sometimes interpreted as how much of the variability in the *y-variable* (e.g. WEIGHT) can be 'explained' by the variability in the *x-variable* (e.g. HEIGHT) and is expressed as a proportion (e.g. r^2 = 0.22 or 22%).

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 WEIGHT of an individual is dependent upon (explained by) his HEIGHT.

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

Examine the correlation coefficient (and its confidence limits) for the association between the HEIGHT and WEIGHT variables:

```
Correlation coefficient: r   = 0.47                <- correlation coefficient
                         r^2 = 0.22
95% confidence limits:   0.16 < r^2 < 0.28        <- confidence limits
```

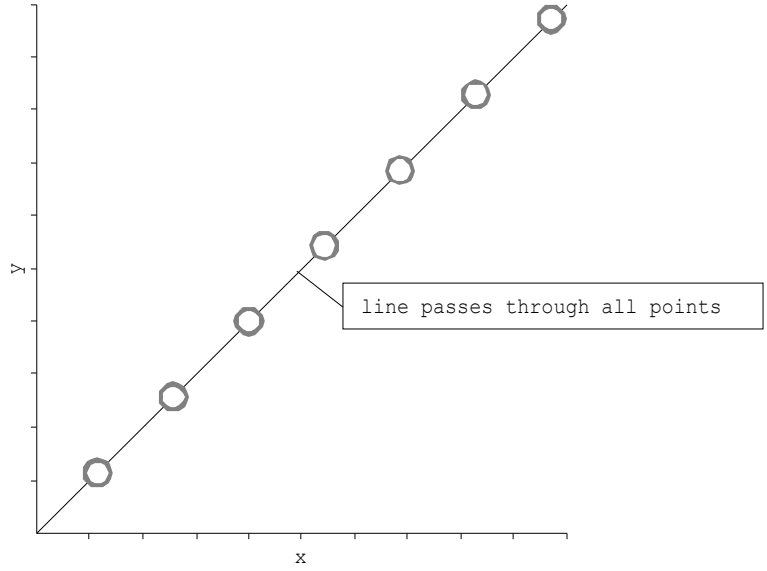
In this case, the correlation coefficient is 0.47 with 95% confidence limits of 0.40 - 0.53 (these are calculated by taking the square root of the confidence interval for r^2 reported by ANALYSIS). If you compare these numbers with the guideline values in the table, you should see that there is a weak 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. In our example, as the confidence interval does not include 0, there does appear to be a relationship between height and weight.

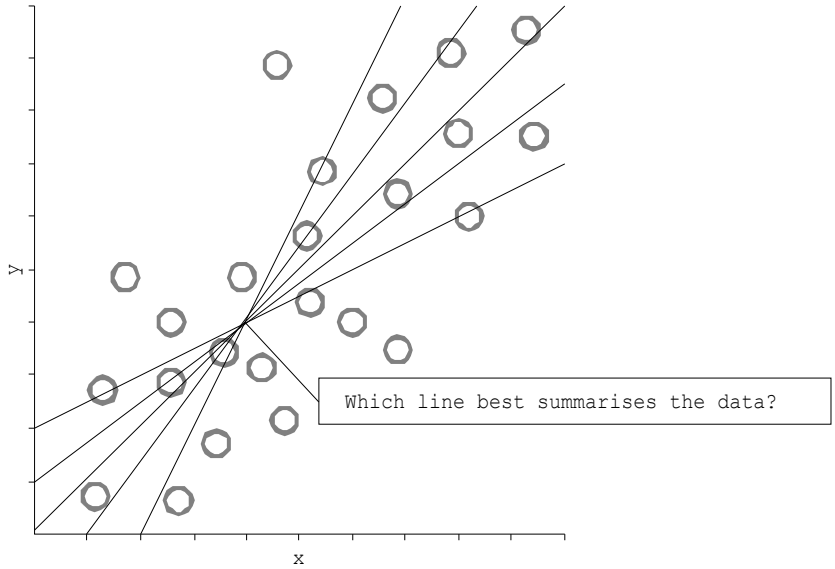
The regression line

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.

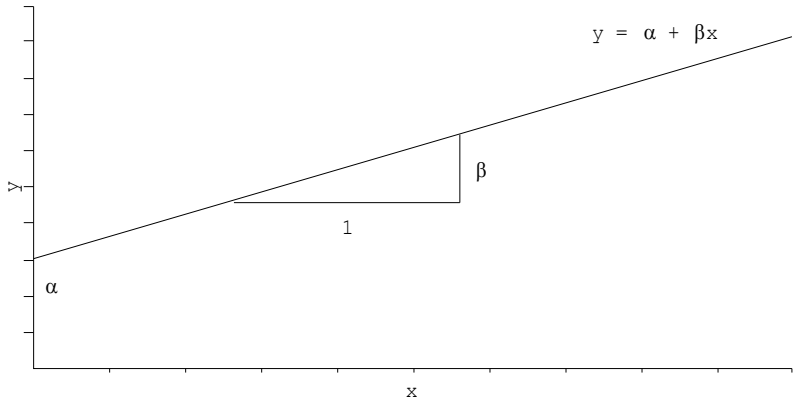
The regression line

The method of least squares

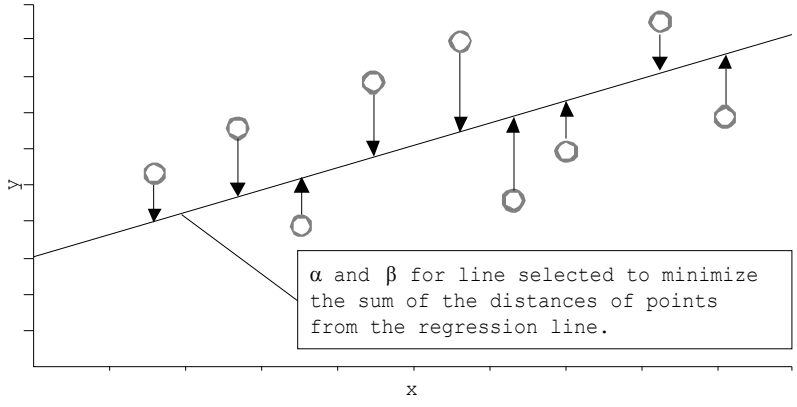
Any straight line drawn on a chart can be represented by the equation:

$$y = 9 + 9x$$

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 9 is the *intercept*, the point at which the line cuts the y-axis (i.e. the value of y when $x = 0$). The term 9 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 9 and 9 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, 9, is usually referred to as the *regression coefficient* or the 9-*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. First clear all selections, and then select all cases for which there are no missing values for WEIGHT and HEIGHT:

```
select
select weight <> .
select height <> .
```

Then issue the following command:

```
regress weight = height
```

ANALYSIS calculates and displays the Pearson correlation coefficient, 95% confidence limits for the Pearson correlation coefficient, r-squared, the y-intercept (9*)* and the regression coefficient (9):

```
Correlation coefficient: r   = 0.47
                         r^2 = 0.22
95% confidence limits:   0.16 < r^2 < 0.28

Source             df   Sum of Squares   Mean Square   F-statistic
Regression          1       28800.4259    28800.4259        288.78
Residuals        1028      102523.1469       99.7307
Total            1029      131323.5728

B Coefficients

                          B            95% confidence                 Partial
Variable     Mean    coefficient     Lower       Upper    Std Error    F-test
HEIGHT     177.2301    0.7605781    0.672751   0.848405    0.044757  288.7820
Y-Intercept           -59.7487821
```

In this case the *Y-Intercept* (9) is equal -59.75and the *slope* (9-*coefficient*) is equal to 0.76. The equation for the fitted line is:

```
WEIGHT = -59.75 + (0.76 * HEIGHT)
```

For each unit increase in HEIGHT, WEIGHT increases by 0.76 kg. The correlation coefficient (*r*) is equal to 0.47 which suggests that HEIGHT is moderately predictive of WEIGHT. The coefficient of determination is 0.22 suggesting that 22% of the variability in the WEIGHT variable can be explained by the variability in the HEIGHT 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 288.78 which, when compared to the *F* distribution (at $df_1 = 1$ and $df_2 = 1029$) in a set of statistical tables, is highly significant ($p < 0.00001$) suggesting that WEIGHT and HEIGHT are associated with each other and that HEIGHT is predictive of WEIGHT.

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 HEIGHT to be moderately predictive of WEIGHT (r = 0.47, F = 288.78, df=1028, p < 0.00001)'*

What the regression line does not explain

Predicted values

The equation for the predicted association between HEIGHT and WEIGHT is:

```
PREDWT = -59.75 + (0.76 * HEIGHT)
```

We can define the predicted weight in our sample with the commands:

```
define predwt ###.##
predwt = -59.75 + 0.76 * height
```

We can examine the actual and predicted weight for the relevant sample of 1030 records by typing the commands:

```
describe weight predwt /all
```

which gives the following output:

```
Variable        Obs        Total       Mean   Variance    Std Dev
WEIGHT         1030    77300.000     75.049    127.623     11.297
PREDWT         1030    77193.220     74.945     27.946      5.286

Variable     Minimum     25%ile     Median     75%ile    Maximum       Mode
WEIGHT        48.000     67.000     73.000     81.000    118.000     73.000
PREDWT        61.850     71.730     75.530     79.330     89.210     75.530
```

Note that the mean of WEIGHT and PREDWT are equal, except for rounding error.

We can use the equation to make a prediction of a person's WEIGHT given their HEIGHT.

```
select height = 175
describe height weight predwt /all
```

which gives the following output:

```
Variable        Obs        Total       Mean   Variance    Std Dev
HEIGHT         133    23275.000    175.000      0.000      0.000
WEIGHT         133     9787.000     73.586    114.320     10.692
PREDWT         133     9742.250     73.250      0.000      0.000

Variable     Minimum     25%ile     Median     75%ile    Maximum       Mode
HEIGHT       175.000    175.000    175.000    175.000    175.000    175.000
WEIGHT        53.000     67.000     71.000     78.000    118.000     71.000
PREDWT        73.250     73.250     73.250     73.250     73.250     73.250
```

We see that mean WEIGHT is similar to the predicted PREDWT for this group. This is so because the selected HEIGHT is close to the mean HEIGHT of 177.23.

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 = weight - (0.76 * height - 59.75)
```

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

```
scatter height resid
```

we should see a random plot. This would indicate that there is a relatively straightforward association between HEIGHT and WEIGHT 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.

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 assess 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 1043 cases in the SFMHSHIV.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:\SFMHSHIV.REC (1043 records)                    Free memory: 250K
Criteria: All records selected
```

The command to SELECT only SMOKERs is:

```
select smoker = 1
```

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

```
select
```

and then issuing the command:

```
select race <> 1
```

(Note that the command:

```
select not (race = 1)
```

serves the same purpose.)

To SELECT only those persons who are both Hispanic and homosexual/bisexual you should first clear the SELECTion by issuing the SELECT command without specifying a SELECTion criteria:

```
select
```

and then issuing the command:

```
select race = 2 and sexpref = 1
```

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 race
```

to produce a BAR chart of RACE groups. The most common race group is 1 (White), with close to 900 people. The other three groups have less than 100 people each.

Issue the command:

```
pie educ
```

to produce a PIE chart of EDUCation status. 56.7% had a college degree or any college-level coursework, while 33.0% of the study population had a graduate degree or any graduate-level coursework.

A HISTOGRAM is the most appropriate chart for the AGE variable. 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 HIVLOAD variable and the CD4 variable is:

```
scatter hivload cd4
```

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

```
scatter hivload cd4 /r
```

Solutions to practical exercises

Exercise 3 - Summarising distributions

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

```
freq npartc
```

which produces the following frequency distribution:

```
NPARTC |  Freq  Percent   Cum.
-------+----------------------
   0   |   235   22.6%    22.6%
   1   |    68    6.5%    29.1%
   2   |   213   20.5%    49.6%
   3   |   320   30.8%    80.4%
   4   |   204   19.6%   100.0%
-------+----------------------
 Total |  1040  100.0%
```

235 subjects had zero male sexual partners in the previous 2 years. This represents 22.6% of the study population.

The most common NPARTC value is 3 with 320 subjects (30.8% of the study population) having 10-49 male sexual partners in the previous 2 years.

204 cases had an NPARTC value greater than or equal to 50. This represents 19.6% of the study population.

Note that only 1040 cases are shown in the frequency table. This is because there are 3 cases with a missing number of partners.

Solutions to practical exercises

Exercise 3 - Summarising distributions

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

```
freq smokeamt
```

which produces the following output:

```
SMOKEAMT    |  Freq  Percent   Cum.
------------+----------------------
          1 |    50   12.4%    12.4%
          2 |    87   21.6%    34.0%
          3 |   192   47.6%    81.6%
          4 |    74   18.4%   100.0%
------------+----------------------
      Total |   403  100.0%


      Total        Sum       Mean   Variance    Std Dev    Std Err
        403       1096      2.720      0.819      0.905      0.045

    Minimum     25%ile     Median     75%ile    Maximum       Mode
      1.000      2.000      3.000      3.000      4.000      3.000

Student's "t", testing whether mean differs from zero.
T statistic = 60.321,  df =   402   p-value = -0.00000
```

12.4% of the study population smoked less than ½ pack per day (coded as SMOKEAMT = 1).

21.6% of the study population smoked ½ to one pack per day (coded as SMOKEAMT = 2).

18.4% of the study population smoked 2 or more packs per day (coded as SMOKEAMT = 4).

It is **not** valid to use the sum, mean, and standard deviation to summarise the distribution of the SMOKEAMT 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 smokeamt
pie smokeamt
```

Solutions to practical exercises

Exercise 4 - Summarising distributions

The commands to produce the seven figure summaries for the AGE, NPARTNER, HEIGHT, WEIGHT, and CD4 variables are:

```
describe age npartner height weight cd4 /all
```

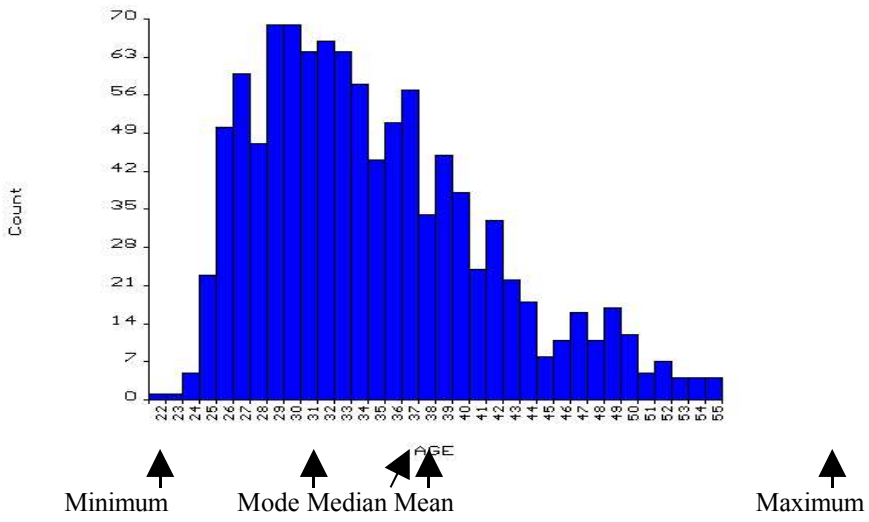
Your completed table should look like this:

	AGE	NPARTNER	HEIGHT	WEIGHT	CD4
number of cases	1042	1040	1035	1031	978
mode	29.000	0.000	178.000	73.000	778.000
median	34.000	10.000	178.000	73.000	873.000
mean	34.773	37.476	177.206	75.088	919.299
minimum	22.000	0.000	160.000	48.000	47.000
maximum	55.000	998.000	196.000	118.000	7107.000
standard deviation	6.724	95.267	6.949	11.363	438.520

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

```
histogram age
histogram height
histogram weight
```

The mode, median, mean, minimum, and maximum of the AGE 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 CD4 variable:

```
describe cd4 /all
```

which produces the following output:

```
Variable        Obs        Total       Mean   Variance    Std Dev
CD4             978   899074.000    919.299 192300.046    438.520

Variable      Minimum     25%ile     Median     75%ile    Maximum       Mode
CD4            47.000    631.000    873.000   1137.000   7107.000    778.000
```

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

```
define cd4group <AAAAAAA>
recode cd4 to cd4group lo-631="1LOW" 632-hi="2NORMAL"
```

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

```
browse cd4 cd4group
```

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

```
set percents = on
tables hivstat cd4group
```

which produces the following output:

```
                                CD4GROUP
HIVSTAT     |1LOW    2NORMAL | Total
------------+----------------+------
          1 |     195     193 |   388
            >  50.3%  49.7% > 39.7%
            |  79.3%  26.4% |
          2 |      51     539 |   590
            >   8.6%  91.4% > 60.3%
            |  20.7%  73.6% |
------------+----------------+------
      Total |     246     732 |   978
            |  25.2%  74.8% |

                Single Table Analysis

RISK RATIO(RR)(Outcome:CD4GROUP=1LOW; Exposure:HIVSTAT=1)                5.81
95% confidence limits for RR                              4.39 < RR <    7.70

              Ignore risk ratio if case control study

                          Chi-Squares   P-values
                          -----------   --------

        Uncorrected:        215.30     0.00000000 <---
        Mantel-Haenszel:    215.08     0.00000000 <---
        Yates corrected:    213.10     0.00000000 <---
```

Showing that those in the HIV-infected group (HIVSTAT = 1) group are **more** likely to have low CD4 counts (RR = 5.81, 95% C.I. 4.39 – 7.70, Yates chi-square = 213.10, p < 0.000001). You can also use the TABLES command to investigate the association between SEXPREF, SMOKER, and POPPERS and your new variable:

```
tables sexpref cd4group
tables smoker cd4group
tables poppers cd4group
```

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 HIVLOAD variable (as was done in the previous exercise):

```
describe hivload /all
```

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

```
define hivgroup <AAAAA>
recode hivload to hivgroup lo-87096="2LOW" 87097-hi="1HIGH"
```

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

```
browse hivload hivgroup
```

Then use the TABLES command to investigate the association between RACE and your new variable:

```
tables race hivgroup
```

Showing that there is no difference in HIV load between Hispanics and Blacks. (RR = 2.70, 95% C.I. 0.62 – 11.72, Yates chi-squares = 1.06, p = 0.304). Note that the relative risk indicates a 2.7-fold higher risk of high HIV load for Hispanics, but the confidence interval includes 1. This is likely due to the small number of Blacks and Hispanics who are HIV-infected. Therefore, the data do **not** support the hypothesis that Blacks are at greater risk of high HIV load relative to Hispanics.

Solutions to practical exercises

Exercise 7 - Analysis of variance (ANOVA)

The commands needed to complete this exercise are:

```
select
select race <> 4
means weight race /n
means age race /n
```

The White, Hispanic, and Black race groups had different mean weights (White mean = 75.244, Hispanic mean = 72.769, Black mean = 79.176, F = 4.309, df = (2,987), P = 0.014).

The White, Hispanic, and Black race groups did not have differenct mean ages (White mean = 34.876, Hispanic mean = 33.500, Black mean = 35.500, F = 1.295, df = (2,998), P = 0.274).

Note that these finding may **not** be valid as some of the ANOVA assumptions may be violated (see pages 62-63 and Exercise 8 on page 64).

Solutions to practical exercises

Exercise 8 - Testing the ANOVA assumptions

The commands needed to complete this exercise are:

```
select
means weight race /n
means age race /n
```

By comparison of means, median and modes, as well as comparison of variances or standard deviations using Bartlett's test, both variables meet the ANOVA assumptions.

To examine the distributions of these variables they may first be grouped:

```
define wtgroup <AAAAAAAAAA>
if weight <> . then recode weight to wtgroup by 5

define agegroup <AAAAAAAA>
if agegroup <> . then recode age to agegroup by 5
```

before issuing the LINE commands to examine the underlying distributions:

```
line wtgroup race
line agegroup race
```

Examination of these charts shows that the both variables are approximately normally distributed. This means that it is valid to use ANOVA techniques with the WEIGHT and AGE 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 San Francisco Men's Health 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 npartner height /r	**regress height npartner**
scatter npartner weight /r	**regress weight npartner**
scatter npartner cd4 /r	**regress cd4 npartner**
scatter npartner hivload /r	**regress hivload npartner**
scatter height weight /r	**regress weight height**
scatter height cd4 /r	**regress cd4 height**
Scatter height hivload /r	**regress hivload height**
scatter weight cd4 /r	**regress cd4 weight**
scatter weight hivload /r	**regress hivload weight**
scatter cd4 hivload /r	**regress hivload cd4**

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 81). An association exists only between HEIGHT and WEIGHT. See pages 81-87 for the results of this regression and residual values.

y-variable	x-variable	r	95% C.I.
WEIGHT	HEIGHT	0.47	0.40 - 0.53

The command:

```
regress weight height
```

produces the following output:

```
                          ß            95% confidence                  Partial
B            95% confidence                Partial
Variable     Mean    coefficient     Lower       Upper    Std Error    F-test
HEIGHT     177.2301    0.7605781    0.672751   0.848405    0.044757  288.7820
Y-Intercept           -59.7487821

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 wtresid ##.####
wtresid = weight - (-59.75 + (0.76 * HEIGHT))
scatter height wtresid
```

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

```
scatter height weight /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 group in the Wilcoxon rank sum test (*W*), 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 103 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

The following table compares use of MARIJUANa to HIVSTATus:

```
                           HIVSTAT
MARIJUAN   |       1       2 | Total
-----------+----------------+------
        1  |     368     521 |   889
        2  |      40     104 |   144
-----------+----------------+------
    Total  |     408     625 |  1033
```

and yields a Pearson's chi-square statistic of 8.15. 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 9.62 on 1 degree of freedom the p-value is between p = 0.001 and p = 0.005:

	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 use of marijuana and being HIV-infected (chi-square = 9.62, p < 0.005).'*

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 on page 72 yielded a *W*-statistic of 32 with sample sizes for the two groups of five and six. The table below shows 32 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 (*W*) in smallest group is on boundary or outside of range

This would be reported as:

'We found no association between White race and height (Wilcoxon Rank-Sum Test,W = 32, 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       28800.4259    28800.4259        288.78
Residuals        1028      102523.1469       99.7307
Total            1029      131323.5728

B Coefficients

                           B            95% confidence                 Partial
Variable      Mean     coefficient     Lower       Upper    Std Error    F-test
HEIGHT     177.2301     0.7605781    0.672751   0.848405    0.044757  288.7820
Y-Intercept            -59.7487821
```

was produced by the command:

```
regress weight = height
```

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 (1028).

From the table below we can see that the p-value for *F* of 288.78 on 1 and 1028 degrees of freedom (we have to look at the values for $df_2 = \infty$) 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 height and weight ($F = 288.78$, $df = 1028$ $p < 0.001$)'.

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.