Statistical Testing in Practice with StatsDirect

Research without Tears or Algebra

Cole Davis

Statistical Testing in Practice with StatsDirect

Research without Tears or Algebra

Cole Davis

Llumina Press

ISBN: 978-1-60594-450-0 (PB)
978-1-60594-451-7 (HC)
978-1-60594-452-4 (EB)

Printed in the United States of America by Llumina Press

Library of Congress Control Number: 2009913282

Contents

Section 1: Background Knowledge

Section 2: Statistical Testing

Section 3: Beyond the Tests

Section One:

Background Knowledge

Chapter One

Introduction

Why This Book was Written

As an applied researcher, I was once asked to produce a training course on computerised statistics for that numerous body of individuals, researchers who do not understand or feel comfortable with statistics. I then found similar needs among the ranks of business people, accountants, the medical profession, civil servants, local government officials and students struggling with academic texts.

I looked for books which could ameliorate this situation and found that many were full of equations, satisfying a statistician's concern with mathematical rigour but repelling the ordinary reader. Others told you more than you needed to know as a beginner, so much so that there was no way of knowing what was important and what was peripheral. The few better books either failed to address the data analysis requirements of the real world – primarily dealing with experimental situations - or told the user how to calculate the statistics by hand. Yet other books attempted to teach both statistics and the use of complex industrial software packages at the same time: one researcher suggested that MSc and PhD students are dropped in at the deep end like this "because it's supposed to be hard".

This is intended to be a short, highly practical book, designed to offer the reader a quick and easy introduction to the analysis of data. Fundamental concepts are presented in the first section – which should be read for a proper understanding of statistical testing – but other ideas are introduced where they logically arise. Each statistical test is accompanied by worked examples which are pertinent to applied research and to basic business problems.

In order to deal with the many practical problems of data analysis without learning how to use tricky information technology at the same time, all the exercises in this book refer to StatsDirect. An excellent but easy-to-use software package, StatsDirect is both inexpensive and allows easy transfer of data to Microsoft Excel spreadsheets. StatsDirect is very rich in features, so the book demonstrates some of the most basic statistical procedures to stop you losing your way.

Who is This Book for?

The primary audience is obviously those needing to interpret and analyse data.

Members of this audience may include business personnel who want to add that extra value to what they do. As will be seen, this may include ascertaining whether or not a trend can be said to be significant. It may also involve working out which factors have the most influence on business outcomes.

Other readers may be those involved in applied research, such as interviewing, focus groups and survey questionnaires (also of value to marketers). Social or educational researchers, for example, may want to quantify their findings. Other readers may wish to add to their own market value by acquiring additional skills.

Other potential beneficiaries may be those undertaking research for the first time – for example, as part of a postgraduate project – or returning to research after a break. The case of one of the researchers who spoke to me on these issues is illustrative of both. She said that not only was it a nightmare to learn statistics at the same time as learning a difficult software package, but this flurry of activity had an almost inevitable outcome: she couldn't remember a thing!

People wanting to 'break into' data analysis may do well to start here. If they subsequently want to transfer their understanding to the use of more difficult software, SPSS users could do worse than to advance to *SPSS 12 Made Simple* by Paul R. Kinnear and Colin D. Gray, 2004 (Psychology Press: Hove, East Sussex).

The Presentation of Concepts in This Book

Coherence, rather than comprehensiveness, is the key to this book. As an applied researcher, I do not feel constrained by the requirements of academic boards, so I am avoiding squiggles and citations of formulae. (Ok, I do use numbers and decimals; into all gardens, some rain must fall.)

I do not think it practical, however, to protect the reader from all statistical jargon: this will be met when using any piece of statistical software, let alone when reading the reports of others, and any other books on statistics. Where appropriate, I introduce alternative names for various concepts; this variegation is reflected in different software packages and books.

Each chapter contains a 'controversy' section, or *'Talking Point'*. These tend to raise issues where the received wisdom may be erroneous or impractical.

The Background Knowledge section should be read in its entirety in order to make full sense of the practical work to come. As statistical testing is primarily about inferential statistics, my introduction to descriptive statistics is speedy and highly selective. But it does have a bearing on the preliminary analysis of data, which should be carried out before running tests. The section on the analysis of differences is fairly conventional, apart from a slight leaning towards non-parametrics as a more useful tool for data analysis 'in the field'. Although possibly a little controversial in its title, the section on qualitative (or categorical) analysis is fairly standard, but a broader set of practical usages is introduced than is usually the case in more academic texts. The section on relationships between data - Correlations, Regression and Factor Analysis - goes beyond some introductory texts in introducing both multiple regression and factor analysis. We need to face the fact that more and more use is made of multiple regression these days and, as will become obvious from the business exercise, it has clear practical advantages. It is often the sorry fate of people faced with PhD projects (amongst others) to find themselves face-to-face with a problem requiring some form of factor analysis technique (I use Principal Components Analysis here) and to discover that 'simple introductions' to these techniques tend to be anything but simple. I hope that the minimalist treatment given here will meet the need for immediate action and will make it easier to absorb more in-depth texts because you will already have acquired a grasp of the essentials.

After covering survival analysis, traditionally used in health research, but with other uses, the book culminates with a section which includes a short set of exercises. Unlike academic tests, exercises do not accompany each section: as practical research is not accompanied by a guide saying 'this problem is one which involves a t-test', it makes more sense to use your overall knowledge from the book to tackle these problems once you have read it through.

The chapter on the presentation of statistics for non-academic purposes is designed to satisfy organisations with practical concerns; readers with academic needs can use this book as a speedy way of gaining a fundamental understanding of statistics, but will need to move to more specific works of reference before writing academic reports. A brief discussion of some other types of statistical tests follows.

Some data sets used for exercises will be reused and built up as the book goes on in order to avoid the need for either lengthy data entry or internet downloads. On the whole, small data sets have been chosen. Through most of the book, two sets of worked examples accompany the tests, with themes which should be of interest to various audiences. It should be noted, however, that statistical principles are applicable to a broad range of disciplines and that readers are encouraged to follow all the exercises with StatsDirect in order to absorb the basic principles.

Who Shouldn't Read This Book?

Those already comfortable with statistics and in need of extending their knowledge to, for example, multivariate or biological statistics should be looking elsewhere. Those wishing to study for an academic qualification in statistics should really be going to a text specifically targeting that qualification. And, no offence intended, statisticians and mathematicians will probably hate this book – most squiggles are probably typographic errors.

Neither is research design a primary focus of this book. Although some references to this will be made in the theoretical section and a little more at the start of practical sections, they are primarily there to facilitate sensible use of the tests.

Using the Book

This is a short book, and I strongly recommend that you read all of it, preferably completing the exercises as well. If this can't be achieved, I still consider it absolutely necessary to read the theoretical section first, before selecting a chapter and launching straight into a project. Understanding is all.

Readers who later wish to learn more about statistics, including more advanced methods, will be able to do so, confident that they have a grasp of the basic issues involved. Similarly, those who wish to progress to expensive and complicated statistics packages such as SPSS or SAS should be able to do so after having gained useful core experience of data analysis.

A Brief Note About the Author

I was never 'good at mathematics', attaining the school O level (approximate equivalent to the English GCSE) after some retakes. Learning univariate (single variable) statistical tests when studying psychology as an undergraduate at the Open University, I had to do these by hand in those days, spending hours working out formulae without generally seeing the point of the mathematics. As soon as I bought my first computer – a Hitachi which ran with two clunking disk drives and no hard drive – I put the formulae on a spreadsheet and found that the real point was to know which tests to use on what data and what they were useful for, without worrying overmuch about the intermediary equations.

Working on the MSc Occupational Psychology at Birkbeck College, University of London, I deepened my knowledge of applied research and became one of those stressed people attempting to learn about factor analysis for a project at the same time as trying to master SPSS for DOS (the 'black screen' version, using typed commands). Since then, I have become involved in a wide range of research projects, within diverse organisations, extending my understanding *en route*. At the time of writing, I work as a free-lance quantitative and qualitative researcher.

Thanks

I would like to thank Dr George Clegg, whose experience of applied statistics has covered both academia and the defence industry. Having asked a lot of hard questions about what I intended to do, he became the book's first reader. He is the *eminence grise* behind the project.

The main non-statistical reader has been Kerin Freeman, the editor and writer based in New Zealand. I am also very grateful for readings and comments by Elinor Hannay in Australia and Diana Kealey in Manchester. Also, I would like to acknowledge the encouragement at different stages from Iain Buchan of the University of Manchester and StatsDirect, and Ruth Hawthorn of the National Institute of Careers Education and Counselling in London (sorry, I know you wanted a course).

I appreciate their contributions, but remain, of course, responsible for any shortcomings.

Talking Point

You do not need mathematics any more than I do to use statistics effectively.

Chapter Two

Descriptive Statistics Introduced.

The focus of this book is on inferential statistics, where we make generalisations from limited amounts of data. Some knowledge of descriptive statistics is essential, however, for three reasons: it is useful in itself, it is essential preparatory work before using statistical tests, and it provides concepts underlying the use of statistical tests.

Descriptive statistics refer to both quantities and also the *shape* of the data. Before getting involved in this, we need to define the word *statistic*: as a singular noun, a statistic is a number which represents or summarises data.

The Limitations of Absolute Data

There are times when you can just say it as it is. We have: 99 red balloons; 20,000 drug addicts; 101 Dalmatians.

Some uncontroversial statistics - figures representing groups of data – can also be used. A common one is the *range*, formed from the maximum and minimum values. When given as a single statistic, this is Max minus Min: so if the highest value is 206 and the lowest value is 186, then the Range statistic would be 20.

Problems emerge, however, when we compare groups of data, or *data sets*. If we look at, for example, the earnings of individuals in countries of different sizes, a direct comparison may be misleading. We therefore tend to use ‘averages’: the description is then along the lines of the average person earning a lot in this country, and very little in that one, regardless of the comparative sizes.

The Average (or Central Tendency)

Although I am trying to avoid mathematics wherever possible, one particular statistic is a key to understanding statistics as a discipline. Relax, however: almost everyone will have heard of the concept of averages.

An *average* has another title, one which is longer but perhaps more meaningful: it is a measure of *central tendency*. Roughly, it defines the middle of the data being examined.

The word ‘average’, however, is rather a layman’s term when applied to statistics. If a newspaper article claims that the average wage or salary is X pounds/dollars/roubles, what is meant? Let us look at three definitions of central tendency: the mean, the mode and the median.

The *mean* is an average calculated by adding together the numbers involved and then dividing the resulting number by the number of items, as in the simple example of this data set of five numbers: **2, 3, 3, 4, 8**. The sum, $\sum$, = 2 + 3 + 3 + 4 + 8 = 20. The number of items, N = 5. The mean is, therefore, $\sum$ / N: 20/5 = 4.

The *mode* is the number that comes up most often. In the previous example, this would be the number 3.

When the list is spread, as above, from the biggest to the smallest, then the *median* is the value sitting in the middle of the string of numbers. We count inwards from the string of numbers in our example, discounting first the 2 and the 8, then the outer 3 and the 4, and we are left with the number 3 as the median.

To illustrate the potential use of these measures of central tendency, let us return to the newspaper article referring to 'average' pay. I don't know which statistic any given newspaper is referring to and I suspect that the reader (and possibly the writer) doesn't either. Without making a judgement, however, let us look briefly at the possibilities.

The *mean* takes the wages of all the paid workers in the city, adds them together and divides the total by the number of the workers, as if to see what one worker would have if all were equally paid. The strength of this statistic, that it takes into account the multimillionaire through to the poorest paid, can also be its weakness. Distortions created by, for example, one or two billionaires, could lead to a rather unrepresentative statistic (see the following section for further discussion of this issue). The mean is used a lot within statistical calculations.

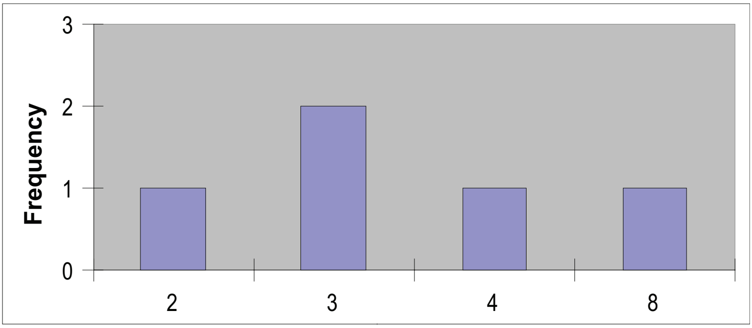
The *mode* may deal with this problem: most wage-payers may, for example, be administrative workers and office workers. Their wage will become the average when using the mode, but how informative will this be when trying to describe the earnings of the workforce in general?

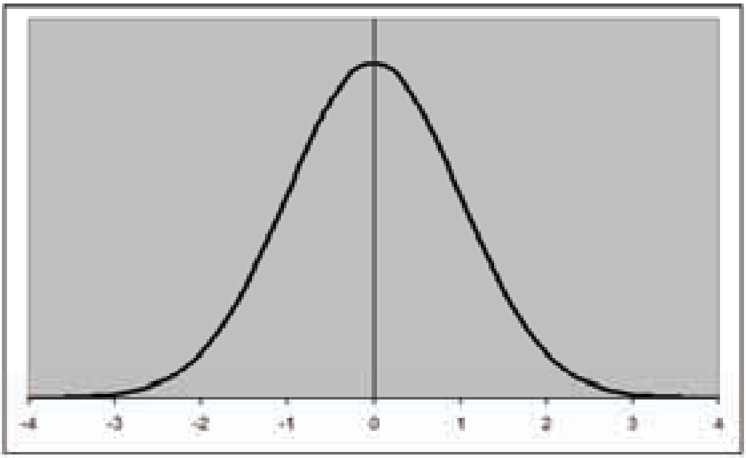
The *median* may find a central value, perhaps a middle-manager's salary. This is useful, as are the other statistics, but it may not indicate the most common wage (as the mode would) nor would it take into account to any extent the many people on poor pay and the considerable purchasing power of the very rich.

My reasons for concentrating on the concept of central tendency or the ‘average’ are both to reflect the importance of the concept and to demonstrate more broadly that data may be subject to different interpretations, of differing degrees of usefulness depending upon the context.

The Distribution of Data

Averages are just part of what is known as the distribution of data. Data can be shown in a histogram; we again use our string of numbers: 2, 3, 3, 4, 8.





In cases where we have more data – often where the data represents a natural population, for example, using automobile maximum speeds or intelligence test results – we get a distribution. One common type of distribution is the ‘normal distribution’, the famous bell curve (an idealised symmetrical one is shown below).

If you look up the *Help* file of StatsDirect and look up *Distributions/Normal Distributions*, you will see various shapes which are all normal distributions. There are other types of distribution, for example, binomial, Poisson and even random.

The consistency or otherwise of a distribution is particularly relevant when choosing between parametric and non-parametric tests. The appropriate use of descriptive statistics will be referred to when we look at test usage.

Regardless of whether or not the distribution is normal, various statistics are derived from the distance around the mean. Although these statistics have a considerable effect on calculations made by statistical software and are cited regularly in textbooks on statistics, we need not be overly concerned with them here. What is important is the general concept of *variance*. Variability of data around the mean – movement away from the central tendency - can be the tell-tale effect that we think we are examining or, as will be seen, it can be other factors.

Talking Point

All manner of additional statistics are churned out both by statistical packages and by books on statistics. One example is *kurtosis*, a statistic representing the degree of flatness of the curve surrounding a distribution histogram. The author has never used this in his research and it is unlikely that you will.

Another point: there are times when descriptive statistics are all you need. Inferential statistics should only be used when there is something you want to find out (infer) from the data.

Chapter Three

Inferential Statistics Introduced

Inferring can mean jumping to conclusions. I hope to show that we can logically view some data and come to *reasonable* conclusions about them. In this chapter, I ask the reader to bear with me while learning some key concepts. Soon after this, we are ready to practise using the tests.

Samples from Populations

Generally, we tend to examine collections of data which are *samples* from a wider *population*. 'Population' can have its usual meaning, the total number of people in a town or country, but in research terms, this often means a group of interest to us, or target group. This could be employees or students, or more narrowly, bank clerks or biology students.

As it is usually impractical to observe, question or test a whole population, for example, biology students in the whole of Canada, we usually limit ourselves to *samples* from the target population. Samples could comprise bank clerks from a selection of banks in a small town, or biology students from three colleges in different regions. Given limitations in resources, samples can be even more restricted than this.

Some rules of thumb have been proposed for sample size in terms of being representative of a population (Roscoe, J.T. (1975) *Fundamental research statistics for the behavioral sciences*. New York: Holt, Rinehart and Winston): most research can be appropriately covered by sample sizes of between about 30 and 500, with 500 representing a population of millions. Simple, tightly controlled studies (for example, with matching pairs of participants) can have as few as 10 to 20 participants. The breaking down of samples into subsamples (e.g. males/females, supervisors/workers) requires at least thirty participants per category. Multivariate techniques and also multiple regression require a sample size several times as large as the number of *variables* (variables are discussed in the chapter on the Analysis of Differences), preferably 10 times as many. The author would suggest that relatively small samples may be used for surveys, as long as it is recognised that there are dangers of being rather unrepresentative and also that very real (if small) effects may be missed.

In this book, small samples of data have been chosen, with often fairly artificial characteristics, in order for the reader to enter data quickly.

Looking for an Effect

If we measure an entire population, what we see is what is out there. Assuming reliable measurement, descriptive statistics would be sufficient for illustrating any perceived phenomenon. Such a phenomenon is the *effect*.

We are never sure, however, about just how representative a sample is of the population from which it is drawn. The reason for using statistical tests, at least at this introductory level, is to find out about the likely existence of an *effect*, or its absence, when using a sample of data. Essentially, we are interested in whether or not there are real differences between two or more sets of data or, in correlational designs, whether or not there are real relationships between them. These are *effects*.

Significance

For those of you who have previously been introduced to statistics and have been baffled by the 'p values' or 'null hypotheses'/ 'alternative hypotheses', this section is for you! **All readers should read this section carefully.**

It should be noted that I referred to *real* differences or relationships between data. With any samples, we cannot be sure of whether or not an effect is meaningful. The perceived effect could, in fact, be a chance fluctuation in the sample data or the influence of a different and perhaps unexpected effect.

The search for meaningfulness is the point of *significance testing*. Is the perceived phenomenon a fluke or not? Let us say, for example, that the same sample of people have their typing speeds measured on a Monday and on a Tuesday. You want to know if the day of the week matters - *does Monday differ from Tuesday* in terms of employee performance – but, you are also aware that extraneous factors (such as the journey in to work on a given day, or illness) could also affect results. The use of statistics here would be to see if there is a significant difference between the typing speeds on the two days.

Academic texts refer to the *null hypothesis*. The null hypothesis states that any perceived effect is, in fact, a matter of chance or a non-relevant factor. If, in our example, any differences in performance are likely to be down to a cold or a traffic jam, then *the null hypothesis is accepted*. In everyday terms, the result is *not significant*.

If, however, the effect is clear-cut – regardless of a few people having a bad day, there is a clear difference between the performances on Monday and on Tuesday – then, academically speaking, *the null hypothesis is rejected*, or *the alternative hypothesis* (the alternative to chance fluctuation) *is accepted*. In everyday terms, the result is *significant*.

I have mentioned these definitions of hypotheses, as you will encounter them in text books, academic reporting and, sometimes, in statistical software results. When reporting in applied research, however, and for your own sanity, the fluke / chance / interfering factors / null result can be referred to as *not significant*. The 'real' effect can be referred to as a *significant* effect or result.

But why have I discussed these issues in what appears to be the wrong order, starting with the non-significant (null hypothesis) result and only then the sought-after significant result? Essentially, tests of significance are concerned with the likelihood of an effect being the result of extraneous factors. These calculations of variance about the mean, like the computers running them, do not share your enthusiasm for that significant effect; they are seeking out the *probability of a chance result.*

There is a lot of theory about probability, starting from simple branching – heads and tails – through to mind-exploding calculations. All that you need to consider here is the question, 'is the effect significant or a matter of irrelevant fluctuation?'

Which takes us to the '*p value'*. The p value of a test is the measure of significance. It tells us *the likelihood of a result being insignificant!*

The percentage of the p value is the calculated chances of your test result being a fluke. Let us say, for example, that the precise p value is .03. This means that there is a three in a hundred chance that the result has emerged from irrelevant fluctuations. It therefore seems rather likely that your result is significant.

Try not to get carried away, however, by starting to talk about 97% success rates or anything like that. Stick to the .03. What it really means is that, according to the statistical calculations, if you tried the test on a hundred samples, then there is a 3% chance that the data could yield a fluke result. Yes, it looks good, but your result could still be that three-in-a-hundred irrelevance.

A corollary of this is that if you used 100 samples, there is a likelihood that you could end up with three misleading results. There is a practical side to this: if you run a series of tests, a temptation when dealing with a complicated data set, the chance of some of them being fluke results increases considerably. This is why replication of results is often recommended in academic journals and - if the result is at all important to your business - it may well be sensible in practice. (See the example of reliability testing using the Pearson Test within the chapter on correlations.)

A high p value, for example, .337 (the highest number is 1), is very suggestive of a fluke. .333 informs us that there is a one in three chance of your results being due to a random or irrelevant fluctuation. Although you could replicate this, it does not seem worth it. But, what level of significance is worth considering?

Generally, we do not refer to a precise p value. Computers often and researchers almost always, refer to the p value in more general terms: p is smaller than something. Commonly quoted p values are $p < 0.05$, $p < 0.01$ and $p < 0.001$ respectively. The chances of a fluke result are calculated as being less than five in a hundred (5%), less than one in a hundred and less than one in a thousand; $p < .02$ and $p < .005$ are not unheard of. All refer to the likelihood of any variance being a matter of chance or unexpected factors.

Although it is quite common to merely ‘read off’ the p values from the computerised results, some would argue that the researcher should decide upon an acceptable level of significance *before* using the statistical tests. The point is that there should be an acceptable level of risk (or academic rigour). While $p < .05$ may be acceptable to a psychologist using a sample of twenty students, where it may be difficult to find a significant result and where the experiment can be replicated easily next year, a business may, in some cases, want the chance of an error to be less than $p < .01$. An aircraft manufacturer, however, may want $p < .001$, with replications. In any case, the question is, should the researcher decide on what is acceptable *after* looking at the results?

A final point when deciding upon which level of significance is acceptable is the question of *one-tailed* and *two-tailed* hypotheses. A one-tailed hypothesis means that you know from the start, because of the nature of the exercise, in which direction the effect is going; a two-tailed means that you can not be sure in which direction a significant result would run. (These terms relate to variance from the mean to only one end of the distribution versus possible variance to both sides.) If you were sure, for example, that the typing speeds would be worse on Mondays than on Tuesdays (and this was indeed the result), if the statistical test indicates that $p < .01$, you can accept that there was a less than 1% chance of a fluke result. If, however, you weren't committed to which day was likely to result in faster typing, a chance result was twice as likely and, in the interests of rigour, you should be accepting a two-tailed result: the p value doubles and you accept $p < .02$ two-tailed. (In academic reporting, you would refer to the result being significant at $p < .01$ one-tailed or significant at $p < .02$ two-tailed. In reporting in non-academic reports, you would just quote the level of significance as being smaller than .01 or .02).

Talking Point

If you already have the data for a whole population, you don't need to infer: a population is already representative of itself. You may, however, use inferential tests to compare (or contrast) one part of the population with another. They may differ in more ways than just the measure in which you are interested, so the test can allow for fluctuations.

Chapter Four

On The Nature Of Numbers

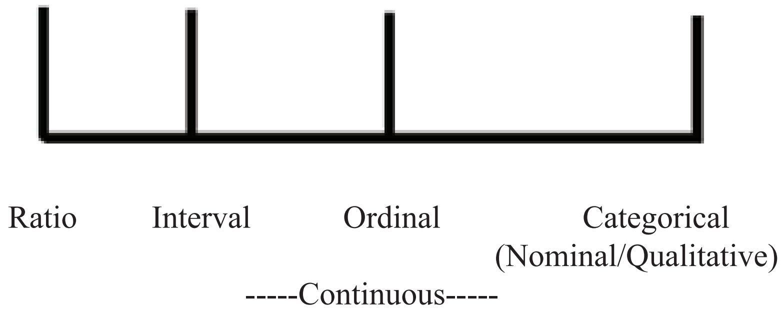
An important concern in quantitative research is that of *the criterion problem*. The criterion is the subject of measurement. If the criterion, that which is being measured, is inaccurate or inappropriate for the test being used, then no estimations of significance can be considered valid. One has built on a house of cards. Although the accuracy of measurement is more a matter of research design and is therefore not our primary concern in this book, the issue of data being appropriate for test purposes must be discussed.

Different tests carry different assumptions about the data they analyse. Such assumptions become particularly important when deciding whether or not to use parametric tests (to be discussed shortly) and in considering the quantification of qualitative statistics (yes!) later in the book.

The nature of numbers is fundamental to using statistical tests. Running tests with inappropriate data can lead to a mass of apparently significant (or insignificant) results which are, in fact, meaningless. This sort of error is easily made and computer software packages in many cases will not be any more aware of the problem than you are. This is, of course, a good reason for examining data before testing, including using descriptive charts.

You must ensure that the data you input is appropriate. In most cases, this includes using the same type of numbers in all data sets used.

One way of conceptualising the different types of data used is to consider its texture. Is it finely chopped or coarse?



Continuous data exists as a run of information which has 'natural' proportions. Distances, seconds and prices (where unbanded) are of this kind. Ratio data runs from zero (e.g., 0, 1 dollar, 2 dollars, 3 dollars…), while interval data can start above zero (as in typing speeds per minute: 30 wpm, 31 wpm, 32 wpm, 33 wpm, etc.). In terms of statistical calculations, there is no difference between ratio and interval types of continuous data. Continuous data is usually the grist for parametric tests.

At times, you may be unsure as to whether or not you are dealing with truly continuous data. Ask yourself if doubling (or halving) a given amount will give a meaningful result. IQ results, for example, appear to be interval data, but are you sure that somebody with an IQ of 140 has exactly twice the measured intelligence as somebody tested as having an IQ of 70? Similarly, the idea of doubling or halving a Likert Scale from a questionnaire (see below) is a dubious one:

1	2	3	4	5	6
Very Unsatisfied	Un-satisfied	In-different	Satisfied	Very Satisfied	Ecstatic

More coarsely grained numbers are usually subsumed within *ordinal numbers*. IQ and most Likert scales truly belong here (unless the latter are very well calibrated, but Item Response Theory is well outside the remit of this book and is not usually come across in everyday questionnaire work). The same would be true of an arbitrary set of prices (£1.40, £2.99), ranges of pay (£15,000-£20,000, £21,000-£30,000) and various other measures which are not strictly continuous.

Ordinal data is so-called because it can be ordered, or ranked. There is a quantitative difference between numbers - one is bigger than another - even if they are not 'smooth' and cannot be arithmetically manipulated like continuous numbers. This sort of data is generally examined by non-parametric tests (see below), which puts them into a rank order of magnitude.

At the far end of our data continuum is *categorical* data, otherwise known as *nominal* or *qualitative* data. My old physics teacher referred to the concept as 'elephants and telegraph poles' when referring to entities which do not mix. I can't remember the precise context, or almost anything else I was taught in physics, but this concept of exclusive differentiation stuck. Categorical data can be *dichotomous* (male/female, or yes/no). It can also include classifications arbitrarily decided by the researcher: examples could include types of employee (manual, clerical, managerial), people holding specific viewpoints (hostile to a concept, approving, indifferent) or even agglomerations of ranges of numbers (well-paid, comfortable, in dire straits).

Each category must be exclusive: each observation is only counted in the *frequency* of one category.

This qualitative choice raises an issue which should be considered throughout your data analysis or research. There needs to be a theory behind your decision-making. I do not mean academic theories (although these may be considered in order to make sense of what you are doing), but some form of rationale. Without this, you may, in effect, be doing something similar to just putting data into the computer and hoping that what emerges is meaningful – which it usually isn't.

(Mixed data, including dichotomous, or binary data, can be subjected to logistic regression, but that takes us outside this introductory volume.)

Parametric Versus Non-Parametric Data

In both tests of differences between data sets and also tests involving correlation (relationships between data), the test user will sometimes discover a choice between using parametric and non-parametric tests.

Parametric tests are used when certain *assumptions* are made about the parameters (the limits and nature) of the data. These tests should only be made if the assumptions are met. Data should be continuous, as in being naturally proportioned, should form a recognised distribution, and should contain homogeneity of variance. Essentially, homogeneity of variance means that numbers in the data sets should be of the same proportion (*not*, for example, 2, 3, 4, 5, 3 tested against 23, 28, 33, 46, 30).

As a rule of thumb, experimental and quasi-experimental data (see Analysis of Differences chapter), when continuous, may be suitable candidates for parametric tests. A strict test of data suitability, looking for normal distribution, is provided by StatsDirect: the Shapiro-Wilk test.

If the assumptions are not met but you are still using measurable data (i.e., of different numerical quantities, but not categorical data), then you should use non-parametrics. Non-parametric tests are designed to ignore the lop-sidedness of data. If, for example, we had the numbers 16, 15, 13, 32, 4, 4, 3, a parametric test would certainly be much affected by the far-flung number (assuming that 32 is an outlier rather than an error). A non-parametric test would rank the data:

32 (1st), 16 (2nd), 15 (3rd), 12 (4th), 4 (5th equal), 4 (5th equal), 3 (7th).

As would be usual when using statistical tests, you would check your descriptive statistics first. Is the outlier (32, in this case) a legitimate part of your data set? If it is a matter of poor data entry, one can alter this as appropriate. If you decide that this was an uncharacteristic performance (maybe the respondent had an unfair advantage) you may consider deleting that record (although such a decision should be carefully considered and recorded). If the outlier is legitimate, a non-parametric test can use the data without being unduly influenced by its extreme nature.

Talking Point

Many books recommend the use of parametric tests wherever possible. They are considered to be ‘more powerful’ than their non-parametric alternatives and should, therefore, be able to detect significant results which may otherwise be missed. The choice, however, is most dependent upon conditions. As most data analysis conditions in everyday research are not text-book examples, non-parametric tests are more likely to be used. In any case, when confronted with data suitable for parametric tests, non-parametric tests often produce very similar results.

Chapter Five

Data Analysis in Practice

H*ere, we make our first reference to using StatsDirect.*

When practising, people find different ways of approaching problems. What is suggested below is just one way of structuring your work.

Decide What it is You Want to Measure First - Have a Rationale

Ideally, you will have designed the data collation yourself, thus meeting your knowledge needs and making it easy to analyse. Often, however, you are given the data and it is left to you to analyse it. In either case, you need to decide what you want to discriminate between, categorise or compare. As I hope will be quite clear, just inputting data and hoping the 'results' are significant is likely to mislead you and everybody else.

Look at the Descriptive Statistics

StatsDirect breaks this into two, descriptive statistics (numeric) and graphics (pictorial).

Open the ‘Test’ worksheet in StatsDirect (if it isn’t already open, use File/Open File/Test). Select two or three data sets and then use Analysis/Descriptive/Univariate Summary. You will generally want to report the means for your data sets. Large differences between the means often indicate differences between data sets.

Then use Graphics to look at your data. Box & Whisker charts are useful for analysis of differences: in significantly different data sets, these will tend to diverge. The Box & Whisker contains a set of the statistics we have already described. The ‘whiskers’ at each end represent the Minimum and Maximum values and the bar with a diamond represents the Median. The ‘box’ is based on the quartile system, which extends the ‘median’ idea of the central spread of the data: the box represents central data, but unlike the median, it covers half of all of the total data for the variable.

Chart Maker or Scatter are useful for examining potential correlations.

Decide on the Test to be Used

These issues will be described in more detail in the rest of the book. For deciding if parametric tests are applicable for a data set, however, use Analysis / Parametric/Shapiro-Wilk, a test for examining normal distribution.

Record, Expand and Report

Decisions made, including the omission of data, should be recorded. Raw data and cross-tabulations may be transferred to Microsoft Excel for the creation of additional charts. Reporting of data is discussed later in the book.

Talking Point

As will become obvious later, it is important to examine data in a graphical format before getting carried away with the wildly significant (or depressingly insignificant) results of your statistical test.

Section TWO:

Statistical Testing

Chapter Six

The Analysis Of Differences

Deciding on the Question

People thinking about everyday data may think of research design as a rather abstract concept. Design needs to be considered, however, every time we ponder which question should be asked (if any). The question of design must be posed before an appropriate test can be selected. It is presumed here, however, that we have already decided that we want to test for significant differences between data sets.

Unrelated Versus Related Design

On each occasion, we need to look at the constituents of the sample.

In *unrelated design*, different participants/records are used in each condition. If we are testing for differences between typing speeds on Mondays and Tuesdays, this would mean one group of staff being tested on a Monday, and a different group of staff being tested on a Tuesday. Although this would get rid of order effects (e.g., people getting bored with tests on the second occasion) or practice effects (getting better with practice), there is an obvious problem: any test is going to have to take into account a range of individual differences and is going to differ from a test applied to related designs.

Related design endeavours to get rid of these individual differences by one of two methods. The *same* people can be used in all conditions; in our typing example, the same people are tested on both Monday and Tuesday. This eliminates individual differences entirely, although not order/practice effects or the possibility of having a ‘bad day’.

Another related design method is the *pairing (or matching)* of different participants for particular characteristics. In our typing example, although different people are tested on each day, each person’s results are paired with the results of another person with, say, the same pre-tested typing speed, maybe also controlled for gender and seniority. This being the case, we eliminate practice/order effects while also eliminating some individual differences. In terms of testing, the paired or matched design is usually subsumed under the related design category.

Here are some alternative terms you may come across:

Unrelated Design	*Related Design*
Different subjects/participants	Same subjects/participants
Between-subjects Design	Within-subjects Design
Unpaired	Paired
Unmatched	Matched
	Repeated Measures (the same over time)
	Panel Data (the same people over time, but this term is used in business statistics/econometrics)

Two or More Conditions

Our discussion of Monday and Tuesday dealt with two conditions. Another example of a two condition test would be the comparison of 'before and after' treatment (e.g., pre-operation, post-operation). In our typing example, however, we may wish to extend our analysis to every day of the week (giving five *conditions*). Another example would examine the differences between a range of different medical treatments (T1, T2, T3).

Data Type

This section deals with quantitative data, where, at the very least, each observation can be compared numerically. For nominal (categorical) data, see the chapter on Qualitative Research.

A Note on Research Design Terminology

In our typing example, we are running an *experiment* (even if not strictly controlled under laboratory conditions). We are manipulating *variables*, types of phenomena which are changeable (variable). We are actively manipulating an *independent variable*; in the typing example, we are interested in the effect of the day of the week; it could be, however, the type of medical treatment, or a single treatment effect. We observe the effect of manipulating the independent variable by looking at changes in the *dependent variable,* that which is being measured. The dependent variable can be typing speeds, time of recovery from illness, the number of relapses, etc. The independent variable is varied according to the will of the experimenter, independent of real life if you like, with the dependent variable being the data dependent on such variations. The terms independent variable and dependent variable should, strictly speaking, only be used in relation to experimental data but are widely used by some writers in other contexts.

Most ‘real world’ analyses of differences between data sets are *quasi-experimental*. We do not manipulate variables ourselves but take them from records or from observations without planned allocation of participants into groups; and, if we choose to use pairing/matching (comparing people with similar relevant attributes), this would be done through taking details from records rather than actually selecting groups of people for an experiment. In a quasi-experimental version of our typing speeds study, the effect of the day of the week would be called a *predictor* (rather than an independent variable) and the typing speed would be the *criterion* (rather than the dependent variable).

The *predictor* and the *criterion* are a similar pair of terms, which should be used in non-experimental research, although these are often interchangeable with independent and dependent variables in the literature. Both terms are used in the following example (ignore the results: we would, of course, want more than 5 records in each condition).

Independent Variable / Predictor: Day of the Week

	Condition 1: Monday	**Condition** 2: Tuesday
Dependent	40	50
Variable /	30	40
Criterion:	45	45
Typing speed	60	50
(in wpm)	45	43

Because this book is largely about non-experimental situations, the predictor/criterion terminology will be used throughout the rest of the book.

And now, to the tests!

Same Subjects, Two Conditions

Wilcoxon – A Non-Parametric Test

In this example, the same customers have filled in a 5-point scale describing their attitudes toward a company's milk chocolate and also that of a rival company. Lower scores indicate hostility, higher scores approval. As this scale has not undergone preliminary calibration, a non-parametric test has been chosen. Being a small data set, we settled for a .05 level of significance; but as we did not know which product would be preferred, we opted for the more rigorous two-tailed hypothesis.

		Predictor: Milk Chocolate Product	
		Condition 1: Ours	**Condition** 2: Theirs
Criterion:	person 1	4	5
Attitude	person 2	3	3
	person 3	2	4
	person 4	4	5
	person 5	3	5
	person 6	4	2
	person 7	3	3
	person 8	5	4
	person 9	3	5
	person 10	4	5
	person 11	3	5
	person 12	2	4
	person 13	2	5

Open File/New Workbook and then enter the numbers given again, with ‘Ours’ and ‘Theirs’ at the top of each column of numbers. As you will use this data again and add to it, save the file (File/Save) and give it a name (e.g., choc.sdw).

For those not familiar with manipulating data, data in more than one non-adjacent column can be selected by holding down the Control (Ctrl) button during selection.

If we look at our descriptive data (Analysis/Descriptive/ Univariate Summary), we will see a difference of 1 between the means, quite a lot when considering a five point scale used by only 13 people, and a different median. Select Graphics/Box & Whisker to view the differences between the data.

We want to ascertain, however, that the more positive result for ‘Theirs’ is significant, so we use the Wilcoxon test. We use Analysis/Non-Parametric/Wilcoxon. If we had been considering a one-tailed test of significance, the p value would have been .021: $p < .05$. The two-tailed test (StatsDirect uses the term two-sided), however, gives a value of .042, still $p < .05$. We can report the difference as being significant, $p < .05$ two-tailed.

Paired t-test (also known as Related t-test) – a Parametric Test

Students who didn’t pass an examination initially, all achieving less than the 40 point pass mark, retake their examinations with generally improved results (‘First’). They retake a second time (‘Second’) in an attempt to improve their grades.

		Predictor: Examination Retakes	
		Condition 1: First	**Condition** 2: Second
Criterion:	**person 1**	52	60
Score	person 2	53	34
	person 3	47	38
	person 4	40	52
	person 5	48	54
	person 6	45	55
	person 7	52	36
	person 8	47	48
	person 9	51	44
	person 10	38	56

Again, it is suggested that you save the input of these numbers (e.g., retake1), as you will add to the data shortly. Check that the data can be considered to have a normal distribution. This is confirmed by using Analysis/Parametric/Shapiro-Wilk. Then check the descriptive and graphical information: there is an increase in the mean score (0.4) and an increase in the median, but the Box and Whisker plot looks ambiguous.

We now use Analysis/Parametric/Paired t. The large p values tell us that the difference is not significant.

Same Subjects, More Than Two Conditions

Friedman – A Non-Parametric Test

This extends the Wilcoxon test. We use it here to expand on our work with both our examples from the Wilcoxon test and from the Paired t-test.

		Predictor: Milk Chocolate Product		
		Condition 1: Ours	**Condition** 2: Theirs	**Condition** 3: Ours Adapted
Criterion:	person 1	4	5	5
Attitude	person 2	3	3	3
	person 3	2	4	2
	person 4	4	5	2
	person 5	3	5	2
	person 6	4	2	2
	person 7	3	3	1
	person 8	5	4	3
	person 9	3	5	2
	person 10	4	5	4
	person 11	3	5	2
	person 12	2	4	3
	person13	2	5	1

Re-opening the file containing survey reactions to our chocolate products, add the column of figures in the third column (the other columns are unchanged). Our company, worried by the onslaught of the rival product, has produced a new improved variety.

Use the mouse to select all three columns and use Analysis/Non-Parametric/Friedman. The initial screen gives a p value of .0003 ($p < .001$), a highly significant result.

Now, using the 'Multiple Comparisons' option, press Calculate. As before, the relationship between the first two conditions is significant at the level of p < .05 (the approximate p value here being .02). There are also significant differences between ‘Ours Adapted’ and both of the other conditions.

A glance at the means in the descriptive data indicates that this new product is perceived as being worse than its predecessor as well as the rival product. We can also check to see if this trend is significant, using the Cuzick Trend test (under Non-Parametrics). If you don’t want to bother with specifying group scores (SUM from the descriptive data) during this process, it is suggested that you copy the columns in the workbook in the following order (following the trend) - ‘Ours Adapted’, ‘Ours’, ‘Theirs’ – and then use the Cuzick test. Oh dear! The trend is highly significant: back to the drawing board.

Predictor: Examination Retakes

Condition 1:	**Condition** 2:	**Condition** 3:
First	Second	Third

Re-opening the file containing retake examination results, enter the following ten numbers alongside the other two columns of data.

62, 56, 40, 37, 62, 56, 68, 55, 68, 60.

Our candidates have tried again and we are back using the Friedman test. The initial screen gives a p value of .0156 (p < .02). Something is significant there.

Now, using the ‘Multiple Comparisons’ option, press Calculate. As before, the relationship between the first two conditions is not significant, but there are significant differences between the scores from the third retake when compared with either of the preceding scores. This does not tell us the reason, however; perhaps some other effect is in play.

Talking Point

It is a commonplace truism that correlations (to be discussed later in this volume) do not prove 'cause and effect'. As the above example suggests, however, neither do quasi-experimental structures.

Different Subjects, Two Conditions

Mann-Whitney – A Non-Parametric Test

A chain of estate agents regularly receives complaints about the endowment policies it recommends for the paying-off of mortgages. It is now examining the effects of different types of warning notice upon the level of complaints made about each office. Some offices have a wall-poster describing the situations in which endowments are appropriate for mortgages, others pass their clients leaflets as part of a sales pack.

	Predictor: Warning Method			
	Condition 1: Wall-Poster		**Condition** 2: Leaflet	
Criterion:	office 1	5	office 11	6
Number of complaints (per branch)	office 2	4	office 12	15
	office 3	16	office 13	4
	office 4	6	office 14	4
	office 5	7	office 15	6
	office 6	22	office 16	7
	office 7	8	office 17	16
	office 8	9	office 18	7
	office 9	9	office 19	5
	office 10	8	office 20	4

In each condition, we are told by the Shapiro-Wilk test that ‘the sample is unlikely to be from a normal distribution’, so we should be using a non-parametric test. After checking descriptives, we use the Mann-Whitney test to find out if the differences between the means are significant. Use Analysis/Non-Parametric/Mann-Whitney and find large p values: there is no significant difference between the data sets and, therefore, no significant difference between the use of wall-posters and of leaflets in this scenario. It is suggested that you save this data set (e.g., ‘mortgage’).

Unpaired t-test (Also Known as Independent t-test) – A Parametric Test

In research into the effects of different office conditions, one group of office workers is given a lengthy proof-reading task to do within enclosed offices; the other group does the same task in an open-plan office.

	Predictor: Office Conditions			
	Condition 1:		**Condition** 2:	
	Own Office		Open Plan	
Criterion:	person 1	80	person 15	56
Task	person 2	68	person 16	69
score	person 3	77	person 17	73
	person 4	78	person 18	70
	person 5	85	person 19	61
	person 6	82	person 20	65
	person 7	79	person 21	59
	person 8	76	person 22	60
	person 9	77	person 23	53
	person 10	83	person 24	61
	person 11	84	person 25	62
	person 12	82	person 26	71
	person 13	81		
	person 14	80		

(Notice that in this study, there are different numbers of participants in the different conditions; only ‘same subject’ studies are required to have the same numbers.)

Each piece of data appears to be normal, using the Shapiro-Wilk test, so we can use a parametric test. The descriptive statistics show different means and we use Analysis/Parametric/Unpaired to use the t-test, which shows a very high significance level. According to this (fictional) data, being in an open plan office seriously impairs performance on a proof-reading task. Please save this file (‘office’).

Different Subjects, More Than Two Conditions

Kruskal-Wallis – A Non Parametric Test

If we open the file with data on mortgages, previously used with the Mann-Whitney, we can add some data to another column: 8, 4, 3, 12, 4, 4, 9, 8, 32, 6.

This data represents complaints following the use of yet another method of warning, incorporating the necessary warnings into the sales talk. We use Analysis/Non-Parametric/Kruskal-Wallis which can be considered an extension of the Mann-Whitney test. The large p value indicates an insignificant result. Again, there is no significant difference between warning methods. This suffices. The use of ‘calculate’ to examine individual relationships in this situation (also known as ‘dredging’) can uncover apparently ‘near-significant’ data, which can mislead: ‘hunting’ for tenuous results is likely to multiply (or inflate, if you like) the probability of chance results.

One-Way ANOVA – A Parametric Test

If we open the file about office conditions, we can put the following 16 numbers into a third column: 70, 70, 73, 80, 81, 75, 75, 73, 81, 76, 75, 75, 73, 71, 72, 67.

Although there is a clear productivity advantage in everybody having their own offices, this is expensive. If we were to introduce cubicles to break up the open plan office, would this produce better results than open plan? The new figures represent (fictional) test scores from workers using cubicles.

To examine our three conditions – open plan, enclosed offices, cubicles, use:

Analysis/Analysis of Variance/One Way

The higher the number for the F ratio in the results, the higher the variance: the variable under examination contributes to the overall effect, as opposed to the effect being caused by unknown variables (‘error’). A large F ratio, 36.86, gives a high significance level. So far, so good, but what about the relationships between one of our variables and another? We can look at a multiple comparison option (e.g., Tukey), where we see that all three relationships are significant, although one suspects that the effect is less between Own Office and Cubicles, which would be unsurprising, given the nature of the compromise. How about a trend test? If we look at the means (use Descriptive/Univariate Summary), we will find means of 79.42 for Own Office scores, 63.33 for Open Plan and 74.19 for Cubicles. It looks like Cubicles is a sensible compromise between the two other office systems, but is the trend significant? For ease of testing, put the columns in order of magnitude (e.g., Own Office, Cubicles and Open Plan) and then run Analysis/Non-Parametric/Cuzick Trend Test. As we have aligned the columns in order, we can save ourselves time and opt for ‘No’ when asked to ‘specify group scores’ (group scores are ‘sum’ in the descriptive statistics). Cuzick demonstrates a significant trend with little room for doubt.

This table of tests of difference is not exhaustive, but refers to tests used in this chapter.
n.b. *Non-parametric tests can be used with ‘parametric’ data.*

Test	Design	Conditions	Data
Wilcoxon	**Same or Paired Subjects**	2	Non-parametric
Paired t-test		2	Parametric
Friedman		3 or more	Non-parametric
Mann-Whitney	**Different subjects**	2	Non-parametric
Unpaired t-test		2	Parametric
Kruskal-Wallis		3 or more	Non-parametric
1-way ANOVA		3 or more	Parametric
Cuzick Trend	Flexible – tests for **trend**	3 or more	Non-parametric
Shapiro-Wilk	Diagnostic – tests for normal **distribution**	1 at a time	Any data, to see if suitable for parametric tests

Chapter Seven

Qualitative Research

There will be many occasions on which you may have gathered a lot of information but it does not appear to be quantifiable. People hold different impressions of a government policy; consumers fall into different 'types' of buyers or social class; different categories of situation or behaviour emerge from incident records. The data is nominal: such phenomena may not be assigned numerical values, as membership of one such *category* is not necessarily 'better' or weightier than another. We can count the *frequency* of their occurrence, however.

The statistical analyses here are concerned with comparing what is observed with what may be predicted. If our predictions are founded on chance, what may happen at random, then we are interested in whether or not actual observations differ significantly from predicted observations.

It should be noted, however, that in order to do this accurately, data must be both *exclusive and exhaustive*: all observations from a sample must be allocated to a category in the analysis and each observation can only be allocated to one category.

Dichotomies – the Sign Test

Let us select a case of a simple 'heads or tails' (dichotomous) event. People walk through a park and they can choose to fork left around a clump of trees or they can fork right around the same clump. Assume that the view is pretty much the same in either direction and that the clump looks rather uninviting. If we observe 30 individuals and find that 17 go left and 13 go right, it is likely that any statistical test is likely to find the difference between the observed 17:13 and the predicted 15:15 to be insignificant. (N.b., in the case of a bigger difference, where something significant *might* be happening but you had no way of knowing beforehand in which direction it would happen, then the more rigorous two-tailed level of significance would be preferred to one-tailed.)

If one seeks a comparison with the parameters of the tests for difference, then dichotomous tests are looking at the significance of differences between two conditions within one (within-subjects) variable. Condition 1 is 'Yes' or 'Heads' or 'Left'; Condition Two is 'No' or 'Tails' or 'Right'. Each condition only contains one number, the *frequency*, the number of times each condition has been observed.

Let us take our example with 30 people finding their way past the clump in the tree. Open Analysis/Exact Tests on Counts/Sign. Firstly, enter 30 as the total observation. As a matter of interest, enter 15 – in other words, a 'fifty-fifty' scenario: $p > 0.9999$ is, of course, $p = 1$, which would be predicted as the random scenario. Now we try our 17:13 scenario; leaving 30 in the sample box, enter 17. $P = 0.58$, still insignificant (we select a two-tailed significance level because, even had there been a significant result, we had no reason to expect the direction of the effect to be right rather than left or vice-versa). Again, out of interest, enter 13 in the box instead of 17; the result is the same, as in a single sample, both results are the two sides of the same coin.

Another example of the Sign test:

Would you consider buying X product?
Yes: 63 / 100 No: the rest.

P <.02 two-tailed (compare this to the statistic given in the Sign test). If we had been sure that the product was superior beforehand, we could have used the one-tailed level of analysis: if we had been looking for p < .01 (only a one-in-a-hundred possibility of being chance), then this would have been satisfied.

More Than Two Conditions – the Chi Square Goodness of Fit Test

In the case of 40 people being observed turning left or right at the end of a supermarket aisle, we may expect product placement to influence the outcome. Let us say that the difference is 26:14 and we believe that an organised promotion is working well for one side, against a random shelf arrangement on the other. Then the Sign test would tell us whether or not 26:14 is significantly different from chance. (Because we already have a good idea of the direction of the effect, the one-tailed level of significance would be acceptable.)

As it might be more realistic to bring in 'straight ahead' as a direction, let us extend our analysis to three conditions. (A methodological point: yes, I know some people may reverse back up the aisle and I could have four conditions, but the relatively small numbers, with all too clear a difference in the frequency count, would distort the test. It is perfectly valid to exclude these observations as long as the rationale is clear and the decision is recorded for future scrutiny. It would not be valid, however, if I just did it to 'get significance'.) Let us have a bigger sample:

Left	*Right*	*Forward*
77	66	47

(Total: 190) On the worksheet, put the three numbers in a column. Select: Analysis/Non-parametric/Chi-Square Goodness of Fit and select 'Grouped Frequencies'. The expected frequencies of 63.33 per category represent the total of 190 divided by the 3 conditions (there are times when you might change the 'expected' frequencies, for example, in studies involving genetics, but for the purposes of this study, let us not go there). There are significant differences: $p < .05$

Another example would be a poll of 105 people asking which social ill is the *worst* (note that this is exclusive). Crime = 35; Global warming = 37; Immigration = 33. (We could have had a fourth or fifth: Insurance companies? Advertising?) The expected frequency is 35 per category (105/3) and as suggested by the closeness of all the categories to the expected frequency, any difference is insignificant; $p = 0.892$ The good citizens will be bolting their doors and calling for tougher immigration legislation as the fresh water diminishes and the tides rise…

The Relationship Between Variables – the Chi Square Test of Association

These tests - found within Analysis/Chi-Square Tests – allow a test of the relationship between variables. As well as the points about nominal data and exclusivity mentioned previously, it should also be noted that there should be at least 20 observations in the sample, with at least 5 in each category.

Let us extend our supermarket example: we may want to find out if gender interacts with the nature of the goods on offer - or perhaps men and women have a tendency to walk in different directions?

	Males	Females	Use the 2 by k option (without trend)
Left	40	37	
Right	31	35	$p = .7673$; non-significant
Forward	25	22	

Another example for examining the significance or otherwise of the *interaction* between variables is that of a product's advertising profile and an intention to buy that product. Here, we can use the 2 by 2 option.

		Have you seen the advertisement?	
		Yes	**No**
	Yes	18	12
Do you intend to buy the product?	**No**	12	30

If we have been prepared to accept p < .02, then p = .0153 is a significant result. As the results show, there are clear differences between the observed results (above) and the expected results calculated by the statistical method.

The r by c option can be used from worksheet or from input data to use a wider range of variables. It should be noted, however, that in spite of the advantages of being able to look for differences between observed and predicted values (the bigger difference, the better), larger groupings of variables are likely to be less and less meaningful when it comes to interpretation.

Talking Point

There are times when data is 'unquantifiable'. When the data is unreliable, no degree of mathematical sophistication can fail to be undermined. Similarly, if there is no way of establishing a sensible 'starting point' or rationale for analysis, then quantification is pointless.

I would merely point out, however, that there are times when those who declare a topic to be immeasurable may be arguing from a viewpoint of personal preference rather than methodological certainty.

Let me take their side for a while, however. A hermeneutic approach is applied to an interview. That interview may give insights into what may lie behind wads of data. It may also be important to find out if such a viewpoint is shared - returning eventually to quantification in order to demonstrate this - and as a starting point for empirical investigations in new directions.

Without a qualitative focus, how could one decide upon which of hundreds of thousands of potential experiments are worth conducting? Essentially, there should be a relationship between statistical work and an interpretative focus. Should these be divorced, however?

This test table refers to *frequencies of observations* within *categories* of a sample.

Number of Variables	Number of Conditions	Focus	Test
One	2 (dichotomy)	Differences	Sign Test
One	More than 2	Difference	Chi Square Goodness of Fit (differences)
Multi-variable	2 or more	Interaction between variables	Chi Square Test of Association

Chapter Eight

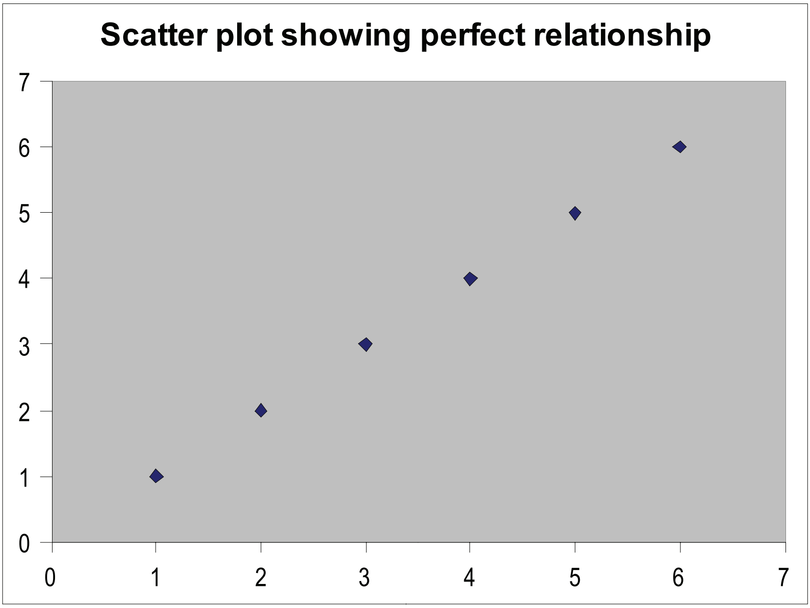
Correlations, Regression and Exploratory Factor Analysis

Correlation

Returning to data with measurable differences, we may wish to examine the relationship between one variable and another, its *correlation*. A correlation is summarised by a statistic known as the *correlation coefficient*. This statistic runs from +1 (a perfect positive correlation), through 0 (completely random) to –1 (a perfect negative relationship). The following numbers and scatter plots will illustrate these.

1	1
2	2
3	3
4	4
5	5
6	6

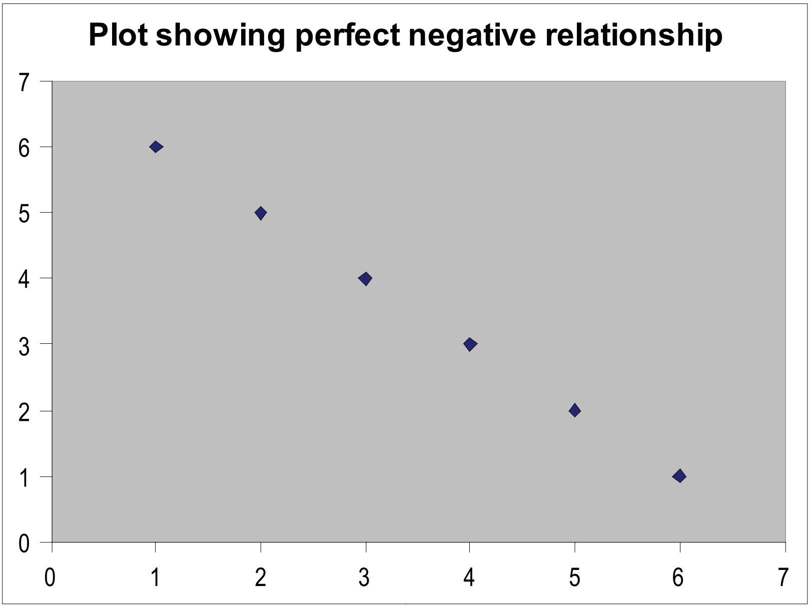
These two columns of numbers are exactly the same. For example: the same people scoring the same on a measure on two occasions. This can be represented visually using a *scatter plot* from the Excel Chart Wizard or via StatsDirect. To use StatsDirect, select Graphics/ Scatter. As we are only looking at one relationship, opt for just one 'series'.



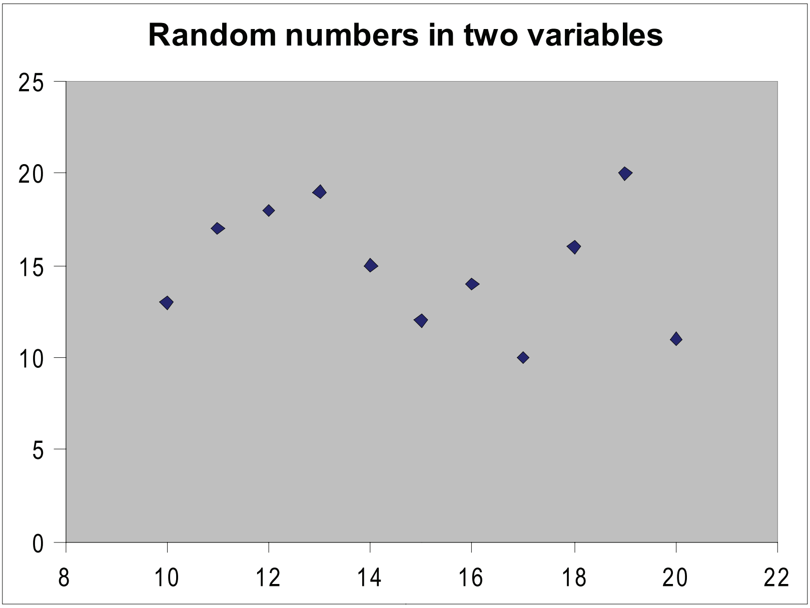
Although one would not expect a perfect relationship in the course of research, a similar slope of this type would indicate a positive relationship between two variables.

1	6
2	5
3	4
4	3
5	2
6	1

Now we juxtapose the inverse, a negative relationship; the higher the score on one of the measures, the lower the score on the other one. This is a perfect *negative* relationship.



Again, a similar slope, this time rising from right to left, would be indicative of a negative relationship between two variables.



No sensible line could go through this set of relationships. The numbers here were randomly generated. With more numbers involved, a globular cluster is quite typical.

The random numbers used were:

61	16
29	90
22	18
41	10
73	31
36	86
80	65
38	99
49	19
0	57

Now let us use one of the tests.

Two Conditions Tested for a Relationship (Correlation) – Spearman, a Non-Parametric Test

As discussed previously, a non-parametric test does not concern itself overmuch about the nature of the data. This is just as well here, as we are firstly going to ask it about the perfect positive relationship, the perfect negative one and the pair of variables with random scores. Select Analysis/Non-parametric/Spearman Rank Correlation. (Also reachable by Analysis/Regression & Correlation/Spearman Rank Correlation.)

On each occasion, choose one of the sets of numbers given above. In the case of the positive perfect relationship, the coefficient (Rho in the case of this test) = 1.

The negative perfect relationship has Rho = -1. Our randomly drawn numbers give a coefficient of 0.2 (i.e., pretty close to zero).

There is a continuum of correlation coefficients:

1 (perfect positive) - - - - - - - - **0** (random) - - - - - **–1** (perfect negative)

Now, let us approach real life. In a survey, people may express confidence in the government, running from very high at rating 1 to very low at rating 5. They may also be rated in terms of confidence in the future, from not confident at 1 to very confident at rating 5.

The scale 'Confidence in the Government' (suggesting that this is at low ebb) is:

5, 1, 2, 4, 4, 3, 5, 5, 3, 4

The scale 'Confidence in the Future' (here, people seem unsure) is:

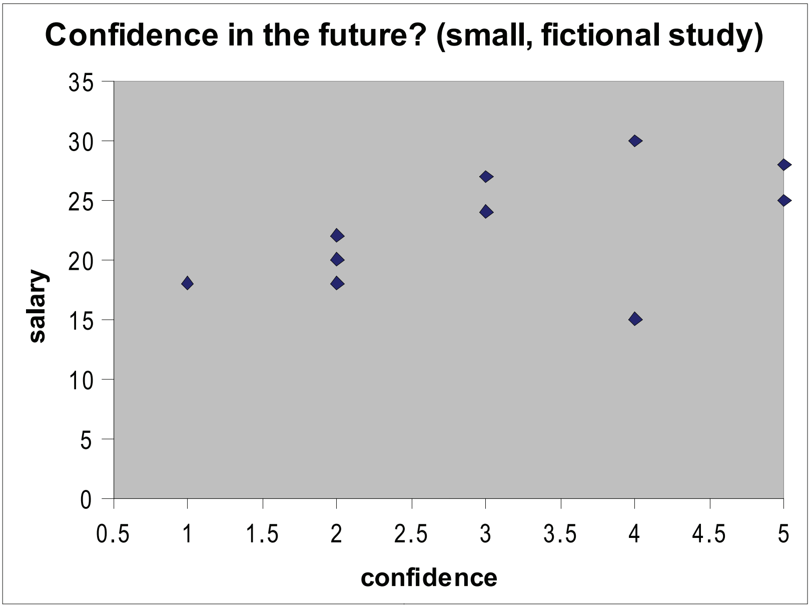
5, 4, 5, 3, 1, 2, 4, 3, 2, 2

Having seen a fairly bitty scatter plot chart, the Spearman test coefficient turns out to be .06, obviously small. The p level is 0.89; any effect is clearly a matter of chance.

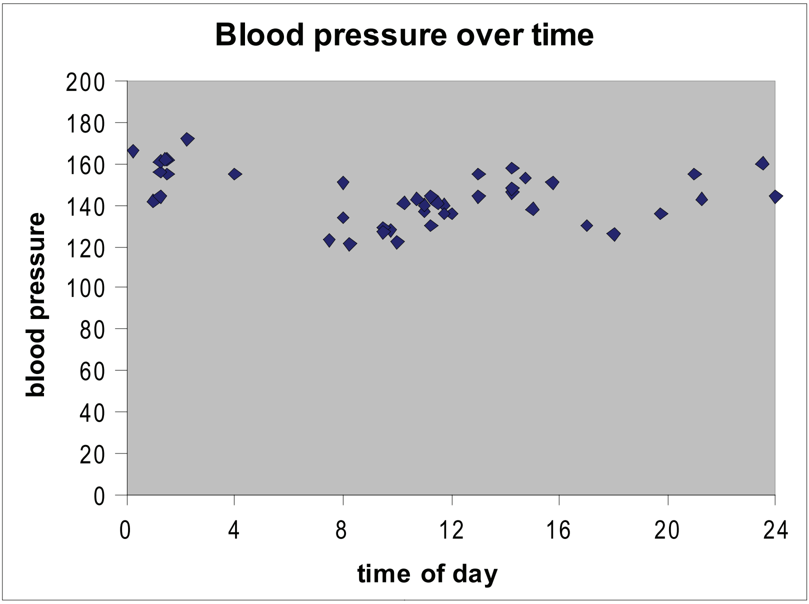
Keep the 'Confidence in the Future' ratings and compare them with the following salaries for each person (in thousands): 28, 30, 25, 27, 18, 20, 15, 24, 18, 22;

assuming that the researcher had predicted beforehand that salary would be positively associated with confidence in the future and that given the small numbers, $p < .05$ was an acceptable level (5 in 100 chance of a fluke finding), then such findings could be accepted at a one-tailed level of analysis ($p < .05$ one-tailed but is greater than that at the two-tailed level).

Did you check the scatter graph for this example? A slope can be discerned.

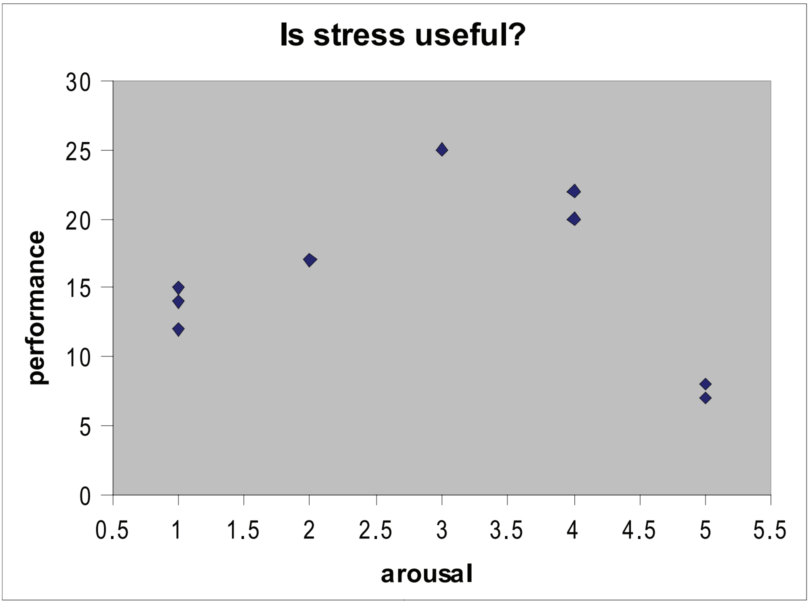


A cautionary note: Why do I insist upon looking at scatter plots in addition to using a test? The reason is that the correlational tests here (non-parametric and parametric) are *linear*. They assume a relationship that runs in one direction, whether positive or negative. If you run a test without looking at a scatter plot as well, there is the danger of assuming significance or lack of it from test results which are really fictitious. Below are two examples of non-linear correlations. One is a real-life error by the author. I examined a friend's blood pressure readings against time (thanks, you know who, for allowing me to reproduce this evidence of my impulsive nature). The coefficient was –0.16, with a two-tailed p level of .3 – a surprise to both of us as we had expected some sort of pattern to emerge. Then, I remembered…



As is suggested by the graph's wave-like formation, there is a tendency towards higher blood pressure readings in the early afternoon and also at night. So there is an effect, but it is not a linear one. While the graph is informative, I should not have used a linear test; the correlations were of course meaningless.

Another non-linear example may be familiar to students of stress and sports fans.



This (fictional) grading of performance against high and low graded stress levels is indicative of what is known as the Yerkes-Dodson Law. Although high stress may damage performance, some degree of stimulation is seen as necessary. Again, this non-linear effect would emerge from a graph but could be erroneously seen as either significant or otherwise if subjected to Spearman's test or a parametric alternative. Essentially, 'curvilinear' relationships should not be subjected to these tests.

Two Conditions Tested for a Relationship (Correlated) – Pearson, a Parametric Test

For this statistical test, select Analysis/Regression & Correlation/Simple Linear Regression & Correlation. As will be seen when we look at regression analysis, there is a close relationship between the two methods. For the moment, however, we are just interested in correlations; Pearson's Product Moment

Correlation provides the correlation coefficient – *r* - required here.

If you reopen the examination retakes file, you should find the following numbers in the third column: 62, 56, 40, 37, 62, 56, 68, 55, 68, 60. If you remember, this represented a significant improvement in examination scores. Suppose now there was the thought that this result may have been a lucky accident: we want to test *reliability* (whether or not a measure is consistent over time). As 'test-retest', however, would be inappropriate because of practice effects, we use a parallel test, which asks different questions about the same taught material.

The new scores are: 65, 55, 39, 43, 66, 54, 73, 58, 72, 64. If we use the scatter plot to examine the relationship between these two variables, we see a positive slope. When you again select Analysis/Regression & Correlation/Simple Linear Regression & Correlation, StatsDirect asks you to 'select data for outcome' and 'select data for predictor': these are important for regression, to be covered shortly, but if we are just using this test for correlation purposes, it does not matter which way round the variables go. The correlation coefficient *r* is .9871, showing a very high significance level; clearly, the examination methodology demonstrates high reliability. (Note also, for the next section, that a statistic called r^2 ['R Squared'] is .97436.)

For an example of a non-significant result using Pearson's correlation, try comparing the first two variables in the examination retake file (data repeated here):
52, 53, 47, 40, 48, 45, 52, 47, 51, 38 __ 60, 34, 38, 52, 54, 55, 36, 48, 44, 56

Effect Size

Another statistic is worth considering at this stage. Note that when we looked at the significant relationship, $\boldsymbol{r^2}$ was cited as

= .97436. If you use Tools/Calculator, you will, of course, note that this is indeed the square of the correlation, .9871*.9871 Some people mistake the correlation coefficient (here *r*) as a measurement of the size of the relationship; it is, in fact, an assessment as to whether or not an effect is 'real' (i.e., significant) as opposed to an irrelevant mixture of variables. It is the *effect size* (here $\boldsymbol{r^2}$) which shows how far the effect accounts for the *variance*, the swing away from the mean.

In the current example, .974 rather than .987 seems of little import, but what about a case where the coefficient *r* is (a still quite high) .63? Here, the effect $\boldsymbol{r^2}$ .3969 contributes no more than 40% of the variance. Quite often, therefore, a significant result can represent a negligible effect size; .33, for example, gives only 11% of the variance. Cohen (1988) suggests the following categories for effect size (***d***):

0.2 to 0.5 = small effect size; 0.5 to 0.8 = medium; > 0.8 = large.

Various ideas emerge from this. If the rest of the variance is random 'noise', could the research model be improved, made more reliable? Is there another factor at play which could contribute to our understanding of the effect? Is the significant result meaningful in real-world terms, and should we invest in it? These questions become less academic when we look at multiple regression and correlation (MRC).

As a building block, however, we first need to consider regression as a concept.

Linear Regression (Two Conditions)

When we looked at the scatter plot charts for significant relationships between two sets of data, a slope could be ascertained, representing the relationship. A line can be drawn through the incline (assuming it makes sense when considering the real world context), taking us into the realms of prediction. By looking for where the X variable meets the intercept (the line), we can see what value is expected from the Y variable.

Non-Parametric Linear Regression

Use when the data is rather rough-hewn.

Taking a simple (fictional) example, ratings for some individuals with debt-related problems (high ratings indicate worse problems) - 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5 - and the rated likelihood of their looking for help (5 meaning they have declared themselves certain to seek help): 3, 1, 1, 2, 1, 2, 3, 2, 3, 2, 3, 4, 4, 5, 5 Select Analysis/Regression & Correlation/ Non-Parametric Linear Regression; use the seeking help ratings as Outcomes and debt levels as Predictors: i.e., will debt levels be a useful guide to the likely take-up of counselling services?

Looking at the intercept, one can see that people with greater problems are likely to seek help, but also that this is by no means certain at an intermediate level; perhaps information about services needs to reach the intermediate problem group before they get into further trouble…

Simple Linear Regression (Parametric Linear Regression)

Another example of regression can be found on the commercial data set below. For the moment, let us just look at Sales and Price. Is there a meaningful relationship between these variables, and can we make predictions about how well a product will sell if we change the price? (Before anybody blames me for their financial demise, these are fictional figures. ‘Pile High, Sell Cheap’ works in some markets and not in others…)

Select Analysis/Regression & Correlation/Simple Linear & Correlation

Choose 'Sales' as Outcome (Y axis) and Price as Predictor (X axis)

Shop	Sales	Price	Instore Advertising	Street Promotion	Local Radio
1	8600	24.99	2180	6400	12000
2	9100	18.99	2200	7800	11500
3	9400	24.99	2220	6800	12400
4	9500	24.99	2160	7000	13500
5	9800	18.99	2220	6500	13200
6	10700	18.99	2170	5000	13500
7	11200	18.99	2280	6800	13200
8	11400	18.99	2500	7200	13500
9	11400	18.99	2200	6000	13500
10	11700	18.99	2250	7400	12900
11	3800	30.99	2190	5000	11000
12	4900	30.99	2250	7500	12000
13	6100	24.99	1180	5400	11900
14	6500	30.99	2250	6000	12500
15	6900	30.99	2170	8200	12100
16	7300	30.99	2180	6500	14000
17	7400	20.99	2255	6100	12200
18	7600	30.99	2250	6800	12300
19	7800	18.99	2200	6000	13300
20	8100	20.99	2240	6900	10000
21	11800	30.99			

The Correlation Coefficient (Pearson's Product Moment r) = -0.61 ; the negative is fine, representing an inverse relationship: lower prices, higher sales. p < .01 The effect size - r2 - is .37; we shall look at the implications of this shortly.

Now take the 'plot regression' option. Apart from anything else, we want to make sure that the relationship is linear. It is linear. Another problem emerges, however: an outlier. One coordinate is well away from the other data and is theoretically dubious, selling at the highest price and yet is also selling well; let's not go there.

During data exploration, removing data just because it is inconvenient to you is unforgivable. We are now, however, in the business of making predictions, so it is reasonable to remove the outlier to improve the prediction model. We are interested in the generality of usual behaviour. In future calculations using this data set, do not use this outlier (conveniently placed at the bottom).

Now run the Regression again without the outlier (i.e., 20 items). The correlation is now –0.77, with an effect size of 0.59. With much more of the variance accounted for, we have a more valid tool for prediction.

Now we get to the point of this exercise: *interpolate X to Y*. In other words, put in a single value for Price and see its likely effect on Sales. For example, if we put up the price to 35, sales are likely to be a little over 4,500; if we drop the price to 15, sales are likely to rise to above 11,000; price = 25, sales = 7,800. Let's pass the calculator to the CEO…

Multiple Regression – multiple predictor variables against one criterion variable

Returning to our table, we may ask if price by itself is the only significant factor in determining the number of sales. Multiple regression allows one to build a *model* for effective prediction. We are interested in two main issues: does the addition of extra variables make an appreciable difference to predictions? And, if so, are some variables more useful than others?

Select Analysis/Regression & Correlation/Multiple Linear and ignore the invitations to use weights (too heavy at this stage of our development). Select *Sales* as the *Outcome (Y axis)* – remember just to choose the first 20, without the outlier – then select *Price, Instore Advertising, Street Promotion and Local Radio* as *Predictors*. The immediate screen is quite informative: Price, in an inverse way (r = -0.82), is seen to be significant (i.e., not a chance effect), but so is Local Radio.

To find out the answer to our first question, whether or not this makes an appreciable difference to our predictive model, select *Analysis of Variance* (and press 'calculate').

There is a significant effect, with the Multiple Correlation Coefficient R = 0.89 (our simple price/sales regression was –0.77). The effect size R^2 is .79 (referred to by StatsDirect as 79.5% of the variance), but it may be better to pay attention to the adjusted effect size of .74, or 74% of the variance. As the effect size of the simple regression was 59%, this model is a more effective one. In other words, we would do well to consider other factors in addition to price when predicting sales.

Before trying anything else, it may be worth checking for statistical safety. Calculate '*Residual plots*'. Residuals, essentially errors, are nicely spread around on both sides of the axis when correlated with Y and with our favoured variables (1 for prices and 4 for local radio) – this is fine, as they should be randomly spread. The almost straight line in the bottom graph is also fine, as a straight line on a Normal plot for residuals indicates that the errors are spread randomly. Also calculate '*Collinearity*': essentially, high collinearity means that variables are over-correlated against each other, possibly meaning that they may be measuring the same underlying construct; in such a case, we probably need to get rid of unnecessary variables. (Rather than fishing for results, one should only include variables that 'make sense'.) As it is, all variables show a high tolerance factor (good) and a low variance inflation factor (1 is good). Press '*Help*' if you want more guidance on these measurements.

Returning to the significance levels, it looks like a worthwhile move to refine our model by removing Instore Advertising and Street Promotion as less influential factors. So, now we go back and perform multiple regression with just price and local radio as predictors: *Analysis of Variance* gives a multiple correlation coefficient of 0.85, and an adjusted effect size of .69 (69%). Our two factors account for considerably more of the variance than the simple model. We do not seem to have lost much by removing the others, although one could always look for another relevant factor.

Having found a useful model, however, let us predict with it. Opt for *Prediction*. Use the left-hand text box to adjust values for the Predictor variables. If we alter the Price to 11.99 and put up our expenditure on local radio to 20,000 – using the left-hand box as a drop-down menu - then, after pressing 'calculate', the screen shows our sales soaring to over 18,000. The intervals cited underneath show the possibility of this varying somewhat…

Multiple Correlation

One can also study the pattern of relationships between more than two variables. Bear in mind, however, that the more correlations you run that appear to be significant, the greater the chance that some of them are, in fact, the product of chance results. If each test has, say, a one in a hundred chance of being a fluke, then this rises as more tests are conducted. One highly conservative method to counter this is the Bonferroni technique: you multiply a correlation's p value by the number of comparisons to get an adjusted p value. If we take our sales example using all the variables there, the use of multiple regression will show us that the radio promotion p value is p = 0.0076; our comparisons numbering 10, we get 0.076. This is rather harsh and – this is the real world entering our calculations – almost certainly wrong; the Bonferroni is particularly fierce when applied to a large number of tests. If, however, we just look at sales, price and radio promotion, as recommended in our multiple regression example, we get a less impressive p value of 0.01, but we are only using 3 comparisons, giving an adjusted p value of 0.03.

Leaving aside the Bonferroni method, let us examine a *correlation matrix* involving our sales example: select Analysis/ Regression & Correlation/Principal Components/Correlation and use all 5 variables. If offered the opportunity to 'reverse scale option', say 'No' at the moment: negative correlations are meaningful in this example (high sales, low price); on other occasions, you may opt for reverse scaling where obverse relationships are irrelevant. Now ignore the initial screen (components, Eigenvalues, etc.): this will be discussed when we look at factor analysis. Opt for *Correlation Matrix* and 'Calculate'. (You may wish to copy the matrix to a workbook for clarity.)

Two common features of correlation matrices may be seen. Firstly, where a variable is matched against itself, a perfect correlation (1) is observed. Secondly, the two halves of the correlation matrix either side of the line of 1's are mirror images of each other, so you only need to pay attention to one of these triangles of data (the lower half is probably easier to look at).

In our example, we can see a large negative correlation between Sales and Price and a fairly large positive correlation between Sales and Radio promotion (supporting our previous decision to concentrate on these factors in multiple regression prediction). With regard to the effect size – how far the correlations account for the variance of the effect – you can work this out by squaring the correlation coefficient. -.767 * -.767 = .588: Price x Sales accounts for half the variance. 0.534 * .534 = .285: Price x Radio promotion is also influential. The other effect sizes would be much less impressive even if they were significant; it should be noted that the apparent paradox of statistically significant variables with negligible effect sizes is a commonplace and you will, at some point, have to make some difficult decisions about what is 'significant' when it comes to application in the real world.

Situations can be less clear than in our example. Several correlations, with a range of decent sized coefficients, may render 'eyeballing' of the correlation matrix a hopeless enterprise. Some subjective decisions can be made with the assistance of Exploratory Factor Analysis.

Factor Analysis – a Data Reduction Methodology

Exploratory Factor Analysis

Factor analysis is both a generic name for the extraction of underlying variables from data and for a specific technique. Here, we consider factor analysis in general. It should be noted that the reader may find this section more difficult than the earlier ones; believe it or not, you will find far more complex explanations of factor analysis elsewhere …

The basic point of exploratory factor analysis is to take several correlations and reduce a bulky conglomeration of variables into hopefully meaningful components, or *factors*. Factor analytic techniques try to find a structure in the relationships between variables, reducing the number of variables into a smaller number of components. If, for example, survey respondents declare a liking for several variables, you may want to find if there are a few common attitudes which account for many of the responses.

As usual, there are real world considerations. Avoid the poor practice of lumping together all the variables from a large study; put in those variables which you think are likely to be relevant. Also, note that the results should be inter-correlated to some extent: if there is no relationship, then there will be no common factors. On the other hand, too high a correlation between two variables may indicate that they measure the same thing; one should be removed before a factor analysis is conducted. There is serious debate about whether or not factors are merely statistical artefacts rather than meaningful entities. In practical terms, this can be resolved to some extent; as is usual when using statistical techniques, a lot depends upon having a sensible rationale for action rather than just pressing the button to see what happens.

There is a lot of literature on the process of factor analysis. The following, very brief explanation is not necessary for conducting the tests but may be helpful when you are confronted by reports on factor analysis. Firstly, a correlation matrix is formed from the data (as we looked at in the previous section). Secondly, factors are extracted, for example, using the Principal Components Analysis (PCA) technique. Thirdly, if you were using the *technique* called Factor Analysis, the factors would be *rotated*; as we do not use factor analysis here, rotation need not concern us, but a brief discussion can be found in the technical appendix later in this section.

Principal Components Analysis – A Factor Analytical Technique

PCA is a precursor to factor analysis the technique, but is now the most widely used method of factor analysis. The data below consists of 15 rows, representing observations (individuals, shops or other objects) and the 12 variables, A to L. The latter could be measures of business data, individuals' performance on different tests, etc. We want to see if there are any common underlying variables, how many there are, and how far each contributes to the variance.

A	B	C	D	E	F	G	H	I	J	K	L
82	92	61	28	25	16	79	54	60	43	28	61
41	44	67	51	81	45	62	61	32	57	54	67
41	72	71	74	79	30	85	57	40	51	41	39
93	84	58	43	42	15	70	70	76	69	14	20
86	92	48	54	36	26	72	69	67	59	14	26
39	53	51	75	48	42	94	84	69	58	19	39
54	45	89	54	56	42	89	60	49	47	35	40
88	98	58	27	35	11	70	62	72	57	15	33
78	97	69	53	41	12	72	67	54	46	26	42
66	92	63	76	60	28	56	43	21	39	80	83
72	95	55	79	39	35	60	53	32	53	60	74
71	87	65	73	70	13	80	65	64	65	20	33
76	80	52	67	60	16	67	52	53	54	51	56
47	61	76	45	79	36	56	63	56	55	38	51
84	81	56	73	61	48	47	39	14	34	75	87

Three points about the raw data. There must be at least 100 observations. There should also be 5 times as many subjects (rows) as variables; these rules are ignored here for brevity's sake. Also, it is helpful to place together variables that you think are related.

Using the data above, select Analysis/Regression & Correlation/Principal Components/Correlation. You will then be asked about ‘scale reversal’. Saying ‘yes’ removes negative signs from correlations, which may be helpful when negatives are merely an artefact of survey coding, for example mixtures of 1-5 and 5-1 scales, and, therefore, irrelevant. In this case, however, say ‘no’ and leave the negatives in.

The main screen appears as follows:

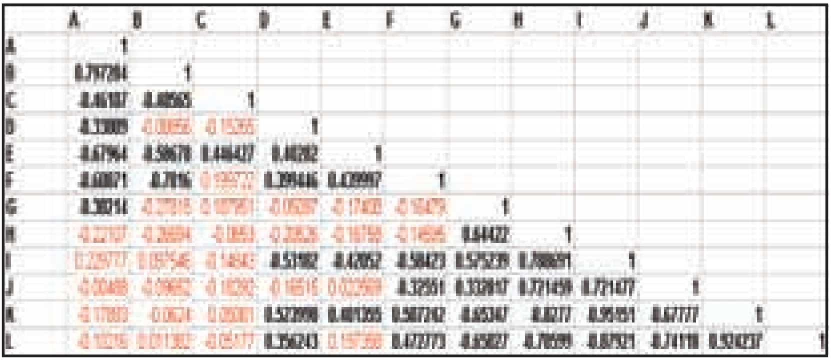
```
                    Principal components (correlation)
Component      Eigenvalue (SVD)   Proportion    Cumulative
1              5.279569           44%           44%
2              3.22758            26.9%         70.89%
3              1.200434           10%           80.9%
4              0.836621           6.97%         87.87%
5              0.724655           6.04%         93.91%
6              0.259211           2.16%         96.07%
7              0.183885           1.53%         97.6%
8              0.137727           1.15%         98.75%
9              0.075948           0.63%         99.38%
10             0.040318           0.34%         99.72%
11             0.032293           0.27%         99.99%
12             0.001759           0.01%         100%
```

Components 1 to 12 on the left are generated by the software and are not of importance, except in showing the order of size of potential components, from largest to smallest. Any components could conceivably be considered to be factors and we need a way of deciding on a cut-off. As will be discussed, this can be a matter of judgement, dependent on known attributes, but the reading of Eigenvalues is a useful statistically based cut-off procedure: generally speaking, we only take account of components with Eigenvalues of 1 or more, so in this case, there would appear to be three discernible factors.

‘Proportion’ shows each component’s contribution to the variance; as will be noted by the ‘Cumulative’ column, the three largest factors account for almost 81% of the variance.

In addition to the main screen, opt for ‘Correlation Matrix’ (and ‘Calculate’).

Where there are several variables in a correlation matrix, the StatsDirect matrix may be cramped: if you copy this to the workbook, the matrix can be seen clearly.



	A	B	C	D	E	F	G	H	I	J	K	L
A	1											
B	0.797284	1										
C	0.46187	0.40565	1									
D	0.33089	-0.08656	-0.15365	1								
E	0.67964	0.59678	0.446427	0.40282	1							
F	0.60071	0.7016	0.199772	0.29446	0.439917	1						
G	0.38751	-0.27815	0.107951	-0.05287	-0.17403	-0.16479	1					
H	-0.22157	-0.26694	-0.1853	-0.2826	-0.16759	-0.14595	0.84422	1				
I	0.229777	0.05746	-0.14642	0.57182	0.42052	0.58423	0.575239	0.708091	1			
J	-0.00488	-0.09682	-0.10392	-0.16516	0.022569	0.32551	0.332817	0.721459	0.721477	1		
K	-0.17803	0.0624	0.20081	0.527998	0.481365	0.587742	0.65347	0.8277	0.95051	0.67777	1	
L	-0.10216	0.011382	-0.05177	0.356243	0.197356	0.472773	0.65027	0.70599	0.87921	0.74118	0.924237	1

The upper half of the computer-generated matrix has been removed from the chart above, as a whole matrix mirrors about the leading diagonal and can play tricks with the eyes, providing a mirage of imaginary clusters. The cells have also been conditionally formatted; any figure above .3 or below -.3 is in a darker font. Positive correlations of less than .3 and negative correlations larger than -.3 tend not to be meaningful in this context.

Previously, I recommended the practice of placing variables which are particularly likely to be related to each other into adjacent columns of the raw data. This allows easy viewing of any correlation clusters. Guessing which variables are closely related before subjecting the data is always a good idea. As has been discussed, you should have some rationale behind your data selections, but, in the case of the correlation matrix, failure to align the variables will require an even more intense eyeballing of the data. When looking at your own data, however, the nature of the factors should become clearer.

In this case, we can see that variables K and L are highly correlated, with strong negative relationships with several other variables; often a construct may be partially defined by opposites. A and B are also highly correlated, with a grouping of positive relationships. The guidance from the Eigenvalues, seeing that we have two very influential factors, seems to be borne out.

Other clusters occur as well. The Eigenvalues suggest that a third factor is likely to have explanatory power. When dealing with variables in real life, however, we are likely to have more of a clue as to which factor and how important this is. Again, such real world considerations are likely to inform us whether or not the smaller fourth or fifth factors have any meaning or are merely the mathematical constructs implied in the list of components with Eigenvalues below 1.

As you might suspect, in spite of all the calculations, exploratory factor analysis has a considerable element of subjectivity, an art as well as a science if you like.

Confirmatory Factor Analysis

As a relatively advanced technique, CFA is not going to be covered in any detail here.

Essentially, Exploratory Factor Analysis, as above, looks at data in a 'bottom-up' way and attempts to create a model of underlying factors.

Confirmatory Factor Analysis works 'top-down'. It uses data to confirm or test factorial models. One can test models against each other, for example. One can test the reliability of a model; in the example given, you might try out new sets of data to see if the three- or four-factor model is the more useful, or if the entire model is misleading.

Factor Analysis – A Short Technical Appendix

Factors – also known as latent variables, dimensions, or core constructs.

Eigenvalue – represents the total amount of variance contributed to by a factor. Deciding to use Eigenvalues of 1 and above is known as *the Kaiser Criterion*.

Eigenvalues, however, are a *guide*, particularly for when the divide between potential factors is less than clear. Depending upon the meaningfulness of the potential categories, one could have argued for four factors in our example. In real world decision-making, the number of factors should be influenced by theory and/or empirical evidence. All other things being equal, however, a theory with fewer factors should be preferred in borderline cases.

Rotation – used in the technique called ‘Factor Analysis’. Whereas the positions of clusters on X and Y axes make a lot of sense in univariate (one-variable) statistics, the immediate position of variables in multivariate statistics is not particularly meaningful. The point is that the initial position of a set of factors on graphical ‘X Y’ axes may not lead to sensible clusters: a three dimensional concept, as it were, is placed on a two-dimensional chart, rendering the direction of the axes largely irrelevant. The constellations of stars could be a useful metaphor: these appear to be real formations but are, in fact, individual stars at diverse distances merely clustered according to the viewer’s perspective. Rotation is a legitimate way of adjusting the position of the clusters to achieve what textbooks refer to as ‘*simple structure*’, a tight and meaningful separation of factors. (Simple structure is an elusive concept; before running a factor analysis, you are also presented with the rather complex question of which type of rotation to choose, orthogonal or oblique, etc!)

Talking Point

We have already established that apparently strong correlations can be coincidental. Do too many tests and that five-in-a-hundred fluke is increasingly likely to occur.

Another perceived weakness of correlations is that they do not prove cause and effect. This is, of course, a problem. As an example, let us say that we have established a reliable relationship between confidence in the government and high levels of spending. Are we sure that perceived stability leads to higher spending? Or, could it be that high disposable incomes lead to political complacency? Or, is there a *mediating factor* which needs to be taken into account, such as unemployment?

One way of sorting out the problem is to run experiments or quasi-experiments. It is not always practical to do so, however. *Triangulation* of methods can be also be used; different aspects of a problem may be subjected to different forms of analysis to see if the original theory may be disproven. Confirmatory Factor Analysis, for example, could be used to test the given relationship, looking at different variables to examine the direction of an effect or to provide more explanatory models. Methods such as Structural Equation Modelling are also used to examine the direction of correlational effects.

This table of tests of relationships is not exhaustive, but refers to tests used in this chapter.

N.b. *Non-parametric tests can be used with 'parametric' data.*

Purpose	Number of Conditions	Data	Test
Correlation	2	Non-parametric	Spearman
	2	Parametric	Pearson
	More than 2	Parametric	Multiple Correlation
Prediction	2	Non-parametric	Non-parametric linear regression
	2	Parametric	Simple linear regression
	More than 2	Parametric	Multiple Regression
Significance levels	More than 2	Non-parametric	Bonferroni Test
Data reduction (exploratory factor analysis)	Multi-variable (must have several conditions)	Parametric	Principal Components Analysis (PCA)

Chapter Nine

The Time until Events: Survival Analysis

Here, we are interested in the area of statistics known as *Survival Analysis*.

Survival Analysis is concerned with how long it takes for an outcome, or *event*, to take place among a range of individuals. It focuses on the interval between the starting point - for example, the completion of a course of treatment – and the event; this interval is known as the *survival time, observation period* or *follow-up period.*

Traditionally, the events of interest have been the deaths of patients after a therapeutic intervention, hence the term 'survival analysis', but this range of techniques can be used much more widely. We can look at many other outcomes, positive and negative, such as employees resigning after the completion of a reorganisation initiative, staff promotions after a training course or the consequences of changes in taxation laws on such events as students graduating, university drop-outs and couples being married.

We are interested in learning about the survival time. Do events tend to accelerate after a particular period of time? Are there phases during which events tend to cluster? What proportion of individuals are affected in a particular phase?

It is also possible that we may wish to contrast different samples. Does the group on a special training programme truly have a different promotion rate from a control group on a traditional course? Will drop-out rates for an employment rehabilitation scheme differ according to different levels of support?

Other ways of measuring event rates could be considered, for example the graphical representation of averages or using the predictive power of linear regression. Survival analysis, however, has a range of advantages. It provides a far more informative analysis of the process under investigation. Non-linear patterns are not a problem; note, for example, that if one is looking at mortality, then there will inevitably be sizeable shifts from the average at both the beginning and the end of the follow-up period. Crucially, however, survival analysis accounts for missing information, known as *censored events*.

Censored information is that which the researcher can only guess at. This includes cases where the event fails to occur until after the follow-up period has expired; the fact, for example, that a cancer sufferer does not die within the survival time does not mean that he may not die from cancer after observations have been concluded. Similarly, individuals who have withdrawn from the study or with whom researchers have lost contact (*loss to follow-up*), may or may not experience the pertinent event. Even when we are certain that events will have taken place (we are all dead in the long run, said Keynes), we can not assume that we know when.

If we omit censored data, or count it as having the same duration as the latest recorded event, then we underestimate the effect under investigation. Although we could negate the time distortions by just focusing on events, we would lose the considerable richness of data provided by looking at duration.

Survival analysis measures the occurrence of events in terms of the duration of time in which they take place. It also takes into account censored data. This is the case when data is missing both during and after the monitoring period; both of these are called ‘right censored data’. One should avoid introducing ‘left censored data’, where complications precede the survival time; some procedures can be put in place to deal with this, but we can’t be dealing with it here.

Some assumptions about the data are required by survival analysis. Continuous data is required; this is usually days in the case of the Kaplan-Meier Survival Function. Longer gaps such as years should be measured by the Follow-Up Life Table. Each record must be a different individual; that is, you don’t include the same person more than once. Censoring should be random; you must not exclude individuals from the survival period because they seem to be a particularly high (or low) risk. The event should be categorical; either they’re dead or they’re alive, cured or not, promoted or not, lapsed or not.

The Kaplan-Meier Survival Function

The beauty of the Kaplan-Meier is the intuitiveness of the survival plot. Its steepness shows whether or not the probability of an event – e.g. promotion – is high or low, whether or not this accelerates or decelerates over time and differences between groups of patients under different conditions.

Firstly, however, we must prepare the data. This is ordered chronologically. One column of data holds the time number, often the number of days in which the person was observed. This is either the day in which the event occurred or the last observation day before the person became censored; individuals who go to the end of the period without the event occurring are given the number of the last day of the observation.

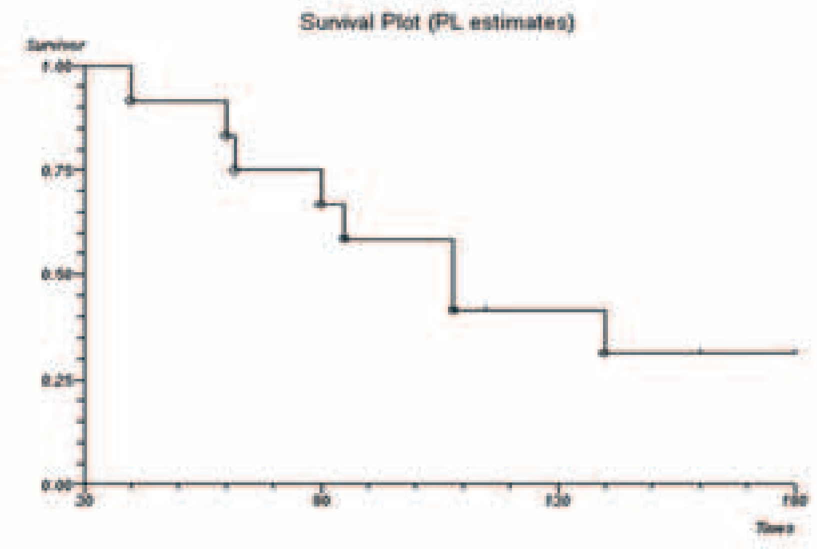
In each case, another column deems the event to have occurred (1) or to have been censored (0). If more than one treatment condition is to be examined, then another column is prepared, with a number for the group (1, 2, 3, etc). If there is more than one group, ordering of the data should be chronological.

Let us start with a very simple example. A management course has been completed among a group of staff at a similar level. We are interested in subsequent promotions. The cohort is deliberately small here, as we are interested in introducing this form of analysis rather than concerning ourselves with statistical significance or organisational meaningfulness.

Days	Promotions	We have a cohort of 12.
40	1	The first promotion takes place on the fortieth day.
60	1	
62	1	
80	1	
85	1	
108	1	Subsequent promotions take place until day 108.
108	1	
115	0	Censored data; perhaps the person has left the organization.
140	1	Another promotion on day 140.
160	0	Censored data again on day 160.
180	0	The remaining 2 of our cohort are not promoted on the last
180	0	day of observation, day 180.

After copying the first two columns onto a StatsDirect workbook, select Analysis/Survival Analysis/Kaplan-Meier. When prompted to ‘select data for Times’, select the Days column and press OK. When prompted to ‘select data for Death/Events or Censored {0}’, select the Promotions column and press OK. As we have not separated our cohort into groups, press ‘Cancel’ when asked for ‘Group Identifier’.

Opt for ‘No’ to ‘save estimates and CIs’. (Although it is often useful to declare Confidence Intervals, i.e. that you are 95% confident that results are within certain bounds, it may not be particularly useful here.) After sending the details to a report sheet, select ‘show event markers’ as a survival plot option; although you may not want this in the future, it helps here in our introduction.



The event markers are the nodules at the bottom of each ‘stair rung’; there are only 7 to be seen, as two events occur on the same day (108); as can be seen, a steeper drop occurs about midway along the time line. The censored events appear as upright marks along the horizontal lines; Kaplan-Meier takes these into account in making calculations but makes no assumptions about the effect, the rate of promotions.

The calculations of the likelihood of an event happening can be seen to have been worked out on the plot. It will be noted that the upright, or y axis, runs from 0 to 1, from zero to 100% of the individuals in the study. At 1, we have 100% of our cohort unaffected. After the first event, after 40 days, it looks like a bit over 90% of the cohort are unaffected, when we look at the figures ('S' in the table shown), we find that this is 91.667%. If we call it 92%, then we can say that the probability of promotion after 40 days is approximately 8% (100%-92%). Note that in real life, we would not make such extrapolations so early in such a small sample.

Then we look at the middle of the upright (y) axis and look across to the intersect. This is the median survival time, 108 days. This is only a 'point estimate' of duration, however; it may not be all that helpful to give this. The confidence interval median survival time may be more helpful, giving the most central figures. One way is to read off the days on the x axis relating to 75% and 25% on the y axis, or cite one of the statistics given in the read-out, the Brookmeyer-Crowley, a robust non-parametric statistic but which should not be used where events tend to happen on the same day, and the Andersen statistic.

Kaplan-Meier survival estimates (columns beyond 'S' have been omitted)

Time	At risk	Dead	Censored	S [Survival]	
40	12	1	0	0.916667	[92%]
60	11	1	0	0.833333	[84%]
62	10	1	0	0.75	[75%]
80	9	1	0	0.666667	[67%]
85	8	1	0	0.583333	[58%]
108	7	2	0	0.416667	[42%]
115	5	0	1	0.416667	

140	4	1	0	0.3125	[31%]
160	3	0	1	0.3125	
180	2	0	2	0.3125	

Median survival time = 108

Andersen 95% CI for median survival time = 69.506329 to 146.493671

Brookmeyer-Crowley 95% CI for median survival time = 80 to 180

It is perfectly acceptable, however, to use the ‘S’ figures, looking at the graph to maintain common sense, when we want to describe the effect. Here (with a bigger sample), we could say that after 62 days, there would appear to be a 25% likelihood of promotion (100% - 75%), and a 68% (100% - 42%) likelihood after 108 days. Note that the likelihood doesn’t change after day 115, a censored event.

Now we move to a more complicated sample. After completion of a new rehabilitation programme for young offenders, we have 2 samples: members of the experimental group undertook the programme, with the control group undergoing standard procedures within the youth justice system. The follow-up period is 90 days. In each case, we have measured days until the event, committing a fresh criminal offence, or the last record.

Experimental group (15 participants):
27, 29 (C), 40, 54, 60, 72, 83, 84, 88, 90 (C), 90 (C), 90 (C), 90 (C), 90 (C), 90 (C).

Control group (15 participants):
16, 17, 18, 20, 22, 25, 25, 30 (C), 53, 71, 84, 86, 90 (C), 90 (C), 90 (C).

Expt = 1; Ctrl = 2	Days	Reoffend	We enter the data together, for comparison
2	16	1	on the plot, to examine the data as a whole.
2	17	1	
2	18	1	
2	20	1	Offences start to occur in the control group
2	22	1	after the first two weeks post-rehabilitation.
2	25	1	
2	25	1	
1	27	1	One offence in the experimental group.
1	29	0	Two individuals have either withdrawn from
2	30	0	the programme or otherwise lost to follow-up.
1	48	1	
2	53	1	
1	54	1	Some more offences occur as the weeks go by,
1	60	1	in both groups.
2	71	1	
1	72	1	
1	83	1	
1	84	1	
2	84	1	

2	86	1	
1	88	1	
1	90	0	Six members of the experimental group survive
1	90	0	the observation period without offending (n.b.
1	90	0	Kaplan-Meier, during the calculations, does not
1	90	0	ignore the possibility of a future offence).
1	90	0	
1	90	0	
2	90	0	Three members of the control group also
2	90	0	survive the follow-up period without offending.
2	90	0	

Select Analysis/Survival Analysis/Kaplan-Meier. When prompted to ‘select data for Times’, select the Days column. When prompted to ‘select data for Death/Events or Censored {0}’, select the Relapse column. For the moment, press ‘Cancel’ when asked for ‘Group Identifier’, as it seems reasonable to look at the data overall. As before, opt for ‘No’ to ‘save estimates and CIs’; this time, we’ll ignore ‘show event markers’ as a survival plot option.

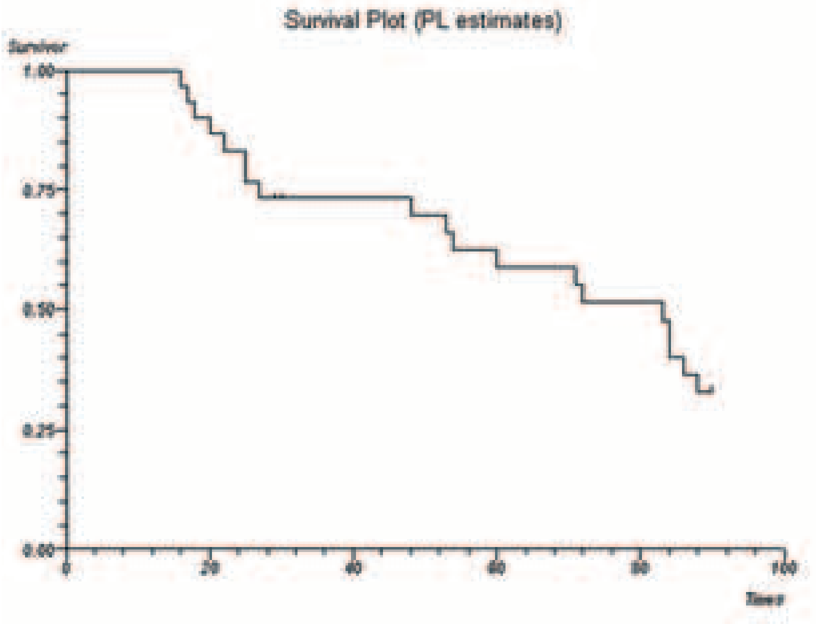
```
Kaplan-Meier survival estimates

Time      At risk    Dead      Censored     S
16        30         1         0            0.966667
```

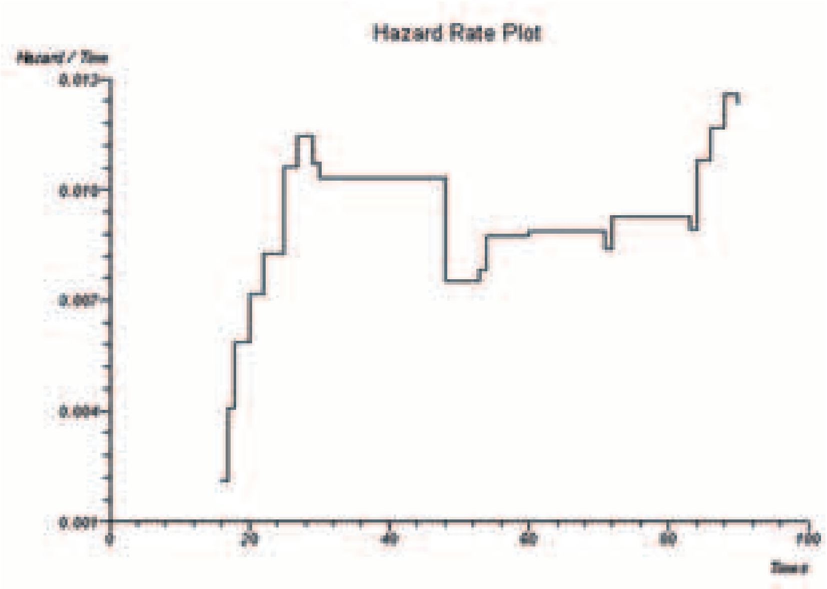
17	29	1	0	0.933333
18	28	1	0	0.9
20	27	1	0	0.866667
22	26	1	0	0.833333
25	25	2	0	0.766667
27	23	1	0	0.733333
29	22	0	1	0.733333
30	21	0	1	0.733333
48	20	1	0	0.696667
53	19	1	0	0.66
54	18	1	0	0.623333
60	17	1	0	0.586667
71	16	1	0	0.55
72	15	1	0	0.513333
83	14	1	0	0.476667
84	13	2	0	0.403333
86	11	1	0	0.366667
88	10	1	0	0.33
90	9	0	9	0.33

One thing to note is the number of censored events at the end of the row of figures. In general, a high number of these can mean that a lot of individuals do indeed fail to experience the event in question; but an alternative explanation can be that the follow-up period is just not long enough to make sufficient observations. We can say, however, that there is a 51% survival likelihood after 72 days (49% are likely to relapse by then).

On the other hand, 33% are likely to continue without offending up to or beyond 90 days. This is also clear in the survival plot, which also indicates something else. A steep re-offending trend appears in the third week after follow-up, with a fairly pronounced trend towards the end of the follow-up period.



The hazard rate plot, which shows the relative acceleration of risk, demonstrates this.



Let us now compare the two groups. This time, when using Kaplan-Meier and asked to 'select data for Group Identifier', select the group column (here, 'Expt = 1; Ctrl = 2').

Kaplan-Meier survival estimates

Expt1Crtl2 = 2

Time	At risk	Dead	Censored	S
16	15	1	0	0.933333
17	14	1	0	0.866667
18	13	1	0	0.8
20	12	1	0	0.733333
22	11	1	0	0.666667
25	10	2	0	0.533333
30	8	0	1	0.533333
53	7	1	0	0.457143

71	6	1	0	0.380952
84	5	1	0	0.304762
86	4	1	0	0.228571
90	3	0	3	0.228571

Median survival time = 53
Andersen 95% CI for median survival time = 8.959823 to 97.040177
Brookmeyer-Crowley 95% CI for median survival time = 22 to 84

Mean survival time (95% CI) [limit: 90 on 86] = 52.504762 (35.664489 to 69.345034)

Expt1Crtl2 = 1

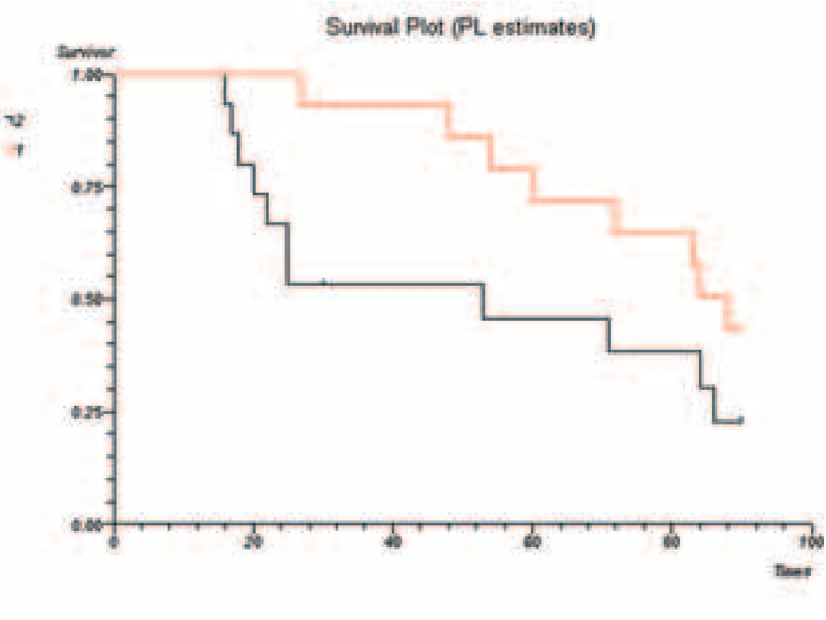
Time	At risk	Dead	Censored	S
27	15	1	0	0.933333
29	14	0	1	0.933333
48	13	1	0	0.861538
54	12	1	0	0.789744
60	11	1	0	0.717949
72	10	1	0	0.646154
83	9	1	0	0.574359
84	8	1	0	0.502564
88	7	1	0	0.430769
90	6	0	6	0.430769

Median survival time = 88
Andersen 95% CI for median survival time = 78.962093 to 97.037907

Brookmeyer-Crowley 95% CI for median survival time = 60 to 90

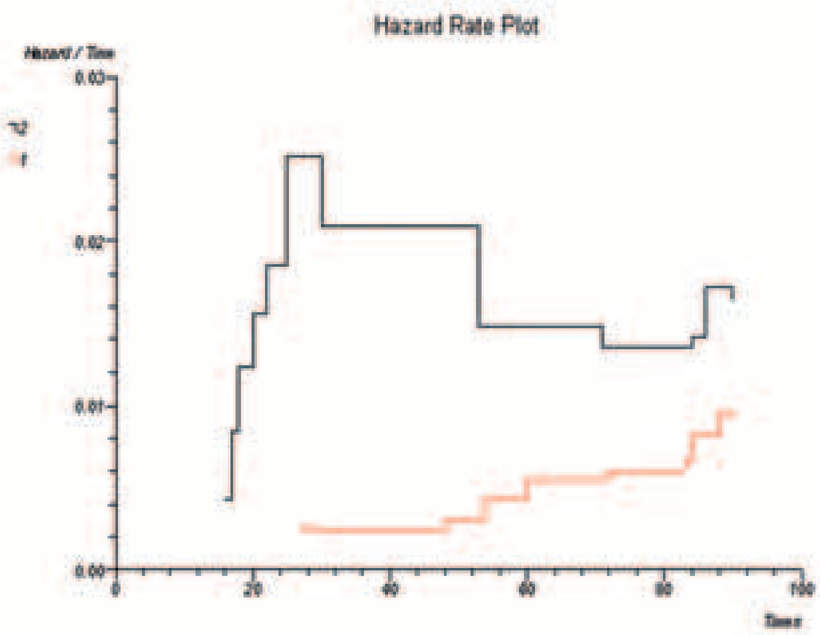
Mean survival time (95% CI) [limit: 90 on 88] = 75.676923 (65.005153 to 86.348693)

We note at the bottom of each table the number of people who have lasted up to or beyond the 90 day observation period: three in the control group (Expt1Crtl2 = 2) and six in the experimental group (Expt1Crtl2 = 1). There are also clear differences between the groups in the rest of the survival data. Little more than 50% of the control group are likely to survive (S) without offending after 30 days (S = 53%; there is 47% chance of an event, an offence, at this time). In the meantime, there is very little movement among the experimental group. By the end of follow-up, only 23% of the controls are likely to survive without incident (77% reoffending likelihood), compared to 46% of the experimental group. The other statistics show a similar story (note that the median survival time is generally deemed to be of more use in survival analysis than the mean).



The survival plot shows an interesting story. Group 2, the control group, reoffend very quickly, within about a week. If this is usual criminologically, then fine, it shows that the new rehabilitation scheme would appear to help participants to avoid this slough; it would be worth checking, however, in case something unusual affected this particular group of young people at this time. There would appear to be a clear difference between the groups over time, although there is a similar decline towards the end, which does rather suggest that the observation time should have been somewhat longer.

The hazard plot shows a clear increase in risk among the control group in the early period. Increased risk appears at the end for both groups. At the very least, we should be concerned about possible vulnerability of all young ex-offenders during this period of time.



Tests of significance in Survival Analysis

Before seeing whether or not the differences in our example are significant, we need to consider the tests of significance offered by StatsDirect.

Peto's log-rank test is best when testing a survival curve throughout its entire course and is more sensitive when the two groups show consistently similar patterns (the first set of data on the Survival sheet of StatsDirect's test file responds in such a way).

The Peto-Prentice Wilcoxon test is considered to be a very reliable test for use in Survival Analysis.

The Gehan-Breslow Wilcoxon test is considered to be more sensitive to differences between groups at early stages. This test can, however, tend towards Type Two errors (judging results not to be significant when they are – a 'False Negative'; the opposite is a Type One error, a 'False Positive', seeing significance when it should be absent).

The Tarone-Ware Wilcoxon test is preferred where survival curves intersect or move away from each other.

Select Analysis/Survival Analysis/Log-Rank & Wilcoxon. The order of data selection is a little different here. First, 'select data for Group Identifier'; here this is Expt1Ctrl2. Then, 'select data for times'; here, Days. Then, 'select data for Deaths/Events or Censored {0}'; here, Relapse. Opt for 'Cancel' when asked for to 'select data for Strata'.

You would use stratification if you had different conditions to contrast (e.g. young with drug problems and without drug problems would be considered stratified data). If dealing with stratified data, you would enter data like so, with stratified data on the right (the variable is Drg1Nodrg2: 1 = drug history, 2 = no drug history).

Expt = 1; Ctrl = 2	Days	Relapse	Drg1Nodrg2
2	16	1	1
2	17	1	2
1	18	0	2
2	20	1	1
etc	etc	etc	etc

Returning to this study, however, the most pertinent tests of significance are the Peto-Prentice and Gehan-Breslow. Both show significant between-group differences, at p < .05

(If we look for interest's sake at the less appropriate tests, the differences appear, unsurprisingly, to be insignificant under the Log-Rank (Peto) test; p < .1 This reflects the differences between the patterns of the two groups. Tarone-Ware, however, shows p < .05 , suggesting the potential dangers of dredging.)

The Follow-Up Life Table

Generally speaking, if you can use the Kaplan-Meier, with its richness of data, you should. However, in spite of being a non-parametric method, Kaplan-Meier is designed to look at continuous data. If we are looking at intervals, such as months, quarters and years, then a Follow-Up Life Table should be considered. The minimum sample size for a life table, in terms of the number of participants starting the study, is 30, although some authorities recommend at least 100; life tables are typically used with thousands.

As an example, let us say that we are interested in how long convicted drink-drivers remain unconvicted after regaining their licences. The initial sample is 200 in number. For the sake of brevity, we use a rather short observation period.

Years	Reconvicted	Withdrawn	The first column carries the intervals (these
0	60	18	do not have to be 1,2,3, etc).
1	35	3	The second column shows actual events.
2	20	5	The third has ‘Lost to follow-up’/censored.
3	15	2	N.B. the table does not contain the whole sample;
4	5	1	36 were not convicted in the follow-up period.

Select Analysis/Survival Analysis/Follow-Up Life Table

At the prompts, select the Years column for ‘Times’, the Reconvicted column for ‘Deaths’ and Withdrawn for ‘Withdrawals’. You are then asked the number who started the study; StatsDirect offers by default the total given within the table, but we enter 200 as the actual sample, as we have survivors who are not included in the columns.

Follow-up life table

Interval	Deaths	Withdrawn	At risk	Adj. at risk	P(death)
0 to 1	60	18	200	191	0.314136
1 to 2	35	3	122	120.5	0.290456
2 to 3	20	5	84	81.5	0.245399
3 to 4	15	2	59	58	0.258621
4 up	5	1	42	*	*

Interval	P(survival)	Survivors (lx%)	SD of lx%	95% CI for lx%
0 to 1	0.685864	100	*	* to *
1 to 2	0.709544	68.586387	12.986565	61.485147 to 74.651522
2 to 3	0.754601	48.66503	10.569952	41.230443 to 55.685715
3 to 4	0.741379	36.722691	9.874568	29.650539 to 43.801528
4 up	*	27.225444	9.661451	20.757746 to 34.076007

The first part of the table does not just reflect our input. As in the Kaplan-Meier, the risk is adjusted. Instead of working out reconvictions (Deaths) as a fraction of the entire sample (in the first interval, 60/200 = 0.3), the table adjusts the sample as if half of the withdrawn sample were reconvicted, giving a slightly higher 'death'/reconviction rate (60/191 = 0.314). When this sum is subtracted from 100, we get the survival rate for that particular interval (0.686).

To look at the cumulative effects, rather than survival rates for a particular period, we need to look at the Confidence Intervals on the right (95% CI, i.e. p < .05%) and describe them year by year. Therefore we can say with a degree of confidence that 1 year after the operation, between 61.5% and 74.7% of the drivers are likely to have continued to drive without reconvictions (the point estimate of the individual survival rate being 68.6%), falling to between 41.2% to 55.7% after 2 years, and so on.

Talking Point

The statistical methods referred to in this chapter are often referred to by the overall title of Survival Analysis. This reflects its most familiar usage in health statistics, the tracking of the survival rate of patients over time; this is made manifest in StatsDirect's use of 'deaths' when events are referred to in analysis readouts. Having said that, I feel that it is more useful to consider it as the study of the time until events, as we are not necessarily discussing life and death and can be measuring events which are positive and not necessarily dramatic. Other terms are indeed used, including 'time to event analysis' and, in sociology and economics, 'duration analysis'.

This table of survival analysis techniques only refers to tests used in this chapter.
N.b. *Non-parametric tests can be used with 'parametric' data.*

Test	Number of groups	Data	Purpose
Kaplan-Meier	1 or more	Non-parametric. Large and small sample. Data must be continuous (i.e. not less frequent than days).	Tracking events over time
Follow-up Life Table	1	Non-parametric. Large samples. Interval data (e.g. months, quarters, years).	Tracking events over time

Log-rank & Wilcoxon tests	2 or more	Non-parametric. For usage differences between Peto Log-Rank, and variations of Wilcoxon tests Peto-Prentice, Gehan-Breslow and Tarone-Ware, see earlier notes in this section.	Testing significance of group differences under Kaplan-Meier

You will find that StatsDirect offers other tests, extending survival analysis to multivariate data (Cox's Regression, Wei-Lachin). The program also provides tests relating to parametric distributions (Probit, Logit, etc). I feel, however, that these methods are not appropriate for an introductory text; indeed, they require the assistance of a statistician.

Section Three:

Beyond The Tests

Chapter Ten

Exercises

Question 1

40 tasters of two fruit drink products, OurDrink and RivalDrink, were asked which they would *prefer* to give their children. No ‘abstaining’ was allowed, from drinking or preference. 26 of the tasters answered in favour of OurDrink. What test should be used and what is the outcome? Accept a significance level of $p < .05$

Question 2

Wanting to find if women were more likely to take a different view of a policy than men, ratings of the policy were examined. What method was appropriate?

Question 3

Assuming a proven relationship between scores on an admission test and subsequent pass rates on a course, what method should we use to choose applicants?

Question 4

You are looking at the relationship between gender and management level in an industry. The numbers at each hierarchical level are as follows.

	ChiefExecs	**SeniorTeam**	**Middle**	**FirstLine**
Female	6	11	20	30
Male	8	18	16	32

What method should we use? Are there significant differences?

Question 5

Three different tax policies are currently being considered. Each politician has been asked for the policy they are most in favour of. There are three political parties. Of the Red Party, 100 favoured Policy A, 230 favoured Policy B, and 150 favoured Policy C. The Blue Party favoured Policies A, B and C as follows: 600, 400, 100. The Green Party favoured Policy A with 180 votes, with 120 for Policy B, and 100 for C.

What design should be adopted? Are the test results significant?

Question 6

The correlation matrix for a mass of data includes many correlations at .9. What should you do?

Question 7

Is a correlation coefficient of .2 significant?

Answers to exercises

Answer 1

Use the Sign Test. The 26:14 ratio would be significant if a one-tailed level of significance was adopted. But do we really have preconceptions about whether or not people would be inclined to like one product more than the other? I would argue for the two-tailed level of significance. If we adopt this, then we are unable to consider the result to be significantly different from chance.

Answer 2

Mann-Whitney examines the differences between two sets of individuals on rating scales. (With great confidence in the rating scales, the unpaired t-test could be used.)

Answer 3

Regression. The individual's admission score would be used to predict his or her likely grade. Given the likelihood of plenty of data, simple regression would be preferred to non-parametric regression.

Answer 4

Chi-square. Not significant.

Answer 5

	Policy A	**Policy B**	**Policy C**
Red Party	100	230	150
Blue Party	600	400	100
Green Party	180	120	100

Chi-square. Clearly significant

Answer 6

Check for collinearity. It is likely that some variables are saying the same thing. Examine the content of the items and consider removing items which have over-similar meanings. You do want some similarities but not duplication. After such a cull with a lowering of collinearity, you may consider a type of factor analysis to reduce data further, then going into more in-depth examination of the correlations.

Answer 7

Especially when dealing with large amounts of data, this sort of correlation coefficient can indeed be accompanied by an acceptable p value. More to the point is how useful is an effect size of .04? In some cases, 4% of the variance is important, in others not.

Chapter Eleven

Reporting In Applied Research

Please note that the recommendations in this particular chapter are for people analysing data in applied settings, not for people who are preparing academic research papers. The latter are usually able to refer to their college's or professional body's specifications; failing that, most books on research methods should help.

In the discussion below, I shall refer to three levels of sophistication of the likely readers of your research. Such a reader may be the commissioner of research, or a line manager within your own organisation or an external body. The sophisticated reader will know as much as you do about the use of statistical methods, and probably more. The intermediate reader is likely to, at least, remember about significance levels from their own studies in the past. The unsophisticated reader will not necessarily understand the difference between statistical significance and the everyday use of the word; p levels will mean nothing.

Written Reporting

Firstly, in reporting significant results, we dispense with the null hypothesis. Although I explained the term while explaining methodology, the lay reader is not concerned with such apparent paradoxes. You merely report that the result is 'significant' (or 'not significant'). If your audience is relatively sophisticated, you may put it as being 'significant ($p < .05$)' or 'significant to level $p < .05$'; with an intermediate level of sophistication, the first time such a citation is made, you may explain that this means that the likelihood of the results being a matter of chance are only 5 in a hundred. You would only consider using references to 'two-tailed' or 'one-tailed' hypotheses for the sophisticated readers; even for such an audience, a discussion of whether or not the effect was expected in a particular direction would be helpful.

Following on from this point is effect size. The term effect size would only be used with a sophisticated audience. It should be noted that until very recently, the concept (or proportion of variance) was rarely mentioned in introductory books. The intermediate reader could be told that this is a 'large', 'medium-sized' or 'small' effect; do not bother with $\mathbf{r}^2$ (or *d* or any other references to the names of actual statistics). Be even more sparing with unsophisticated audiences; mentions of particularly large (or small) effects are sufficient.

Thirdly, only cite the statistical tests used on a regular basis with sophisticated audiences. With the intermediate reader, the occasional mention of Friedman, or whatever, should only be given in parentheses – "the result was significant to $p <.05$ (Friedman)" just to impress a little. Don't cite at all with the lowest level of reader.

Fourthly, unlike in academic research, the reader does *not* want to know how you organised your data (unless this is strictly of relevance to the organisation). You should still of course, record your methods so that they can be replicated as necessary. In the case of removing outliers, however, the sophisticated reader may be informed; they are likely to understand the relevance or otherwise of such data.

In general terms, just provide the reader with relevant results, ones which have a bearing on the purpose of the research. It should not resemble an academic thesis, should not be long and, while not an entertainment, should be readable and cogent.

Verbal Reporting

If using slide-show presentations, try to keep the information interesting. At the least, have clauses sliding in from the side. Don't have large slabs of text if you can possibly avoid this.

A good idea is to have one idea, with a chart and possibly the relevant statistic (bearing in mind the audience), on one page.

Talking Point

Clear categorisation and knowing what to omit are central to good report-writing. Unfortunately, however, there are two ways of considering this point. Making things clearer to the audience is clearly one. Playing to the audience is another: there can be a tension between the audience's needs and objectivity. A current maxim in UK research circles (circa 2006) suggests that, in many cases, researchers are not helping to develop evidence-based policy so much as producing policy-based evidence! May your ideals go with you…

Chapter Twelve

A Taste Of Further Statistical Methods

This book is for beginners, but nevertheless a lot of useful research can be conducted with the methods discussed here. I offer here, however, a taste of what more advanced methods can do. In order not to create confusion, only a few types of test will be considered, building upon what you have already learned.

Cluster Analysis

Viewed from the perspective of a correlation matrix, the columns - the variables - were examined by way of factor analysis (we used the Principal Components Analysis technique). Factor analysis reduces the number of variables to a smaller set of underlying dimensions. In business, for example, one can work out if particular types of attitudes underlie a range of different sales figures.

Cluster analysis is another form of data reduction technique, but here the focus is upon the rows of the matrix, which represent records: observations of phenomena or the individual participants. Here, data is reduced to *objects* or groups of people. In the range of sales figures, we may discover that particular groups of people – retired or students, maybe – tend towards different purchasing patterns. So, in an inversion of factor analysis, we look not at core constructs but at the behaviour of clusters of individuals or groups of data.

Logistic Regression

At times, you are going to find that much of your research data consists of zeros and ones (yes/no). Although categorical (or qualitative) methods have their place, when you have relatively rich data sets including multiple predictors, you are likely to want to use methods which can sensitively weigh up the relative contributions of different variables. Logistic regression has a range of uses and allows the sophisticated measurement of binary data. You will find Logistic Regression in StatsDirect.

Multidimensional Scaling

MDS allows the variables of factor analysis and the objects of cluster analysis to be viewed together in interaction. This has a wide range of applications and allows both visual representation and some reflection of the magnitude of differences.

Meta-analysis

Analysis of a range of findings together. StatsDirect contains tools for this.

Talking Point

Although other useful tests are available, many researchers rarely operate beyond the statistical foothills. If you mastered all of the techniques in this book, which I hope proved enjoyable and useful to you, this may be all that you need; but having gained such skills, you should be able to benefit from more advanced training as required.

Index of Tests

Index of Tests

Index of Concepts

Index of Concepts

Addenda

Some changes have been made to StatsDirect 3, either minor improvements in accuracy or additional features. These are discussed here.

(1) References to the Shapiro-Wilk test for Normality
At times within the book, when looking for a normal distribution for the purposes of using a parametric test, you are referred to the Shapiro-Wilk test, using the procedure Analysis / Parametric / Shapiro-Wilk. In the new version, you use *Analysis / Parametric / Normality*. If any of the tests have significant p values, then you should take this as indicating non-normal distribution; no one test is foolproof for demonstrating normality. The new version of StatsDirect usefully includes a chart to go with the tests. See the StatsDirect Help guide by pressing F1 for more details.

(2) 'Decide on the test to be used' (page 30) – see point (1) on normality / Shapiro-Wilk.

(3) Wilcoxon test (page 40)
The book cites p values of .021 one-tailed and 0.042 two-tailed. You should now see 0.0225 (if rounded to three decimal points, .023) and 0.0449 (round to .045). As you will have guessed, the respective critical values remain p < .05.

(4) Paired (aka Related) t-test (page 41).
Normality and Shapiro-Wilk; use *Analysis / Parametric / Normality*. See addendum (1).

(5) Mann-Whitney (page 45).
Normality and Shapiro-Wilk; use *Analysis / Parametric / Normality*. As in the book, we have non-normal data. In both subsamples, we see that the skewness, Shapiro-Wilk and Shapiro-Francia tests are all significant. Note also the wandering around of

the data from the line on the charts. See item (1) for a general discussion.

(6) Unpaired (aka Independent) t-test (page 46).
Normality / Shapiro-Wilk: use *Analysis / Parametric / Normality* – see addendum (1). The sharp-eyed will notice some evidence of skew in the first subsample. Generally, t-tests are considered quite robust, but I think you should also try Mann-Whitney on the data (which also shows a very high level of significance).

(7) Chi Square Goodness of Fit Test (pages 51 to 52).
This procedure works somewhat differently now. In the first example, the book merely tells you to enter the following data onto the worksheet:
77
66
47

This has now changed. You now enter two columns, as follows:

77	0.3333333
66	0.3333333
47	0.3333333

To get the second set of figures, you divide 1 by the number of data sets (here 3). These, as you might guess, are assuming equal probability for all categories. Although a nuisance, this allows more advanced users to change these assumptions, when expected outcomes are assumed not to be equally likely. **NB do not stint on the number of figures in the second column if you intend to use the Monte Carlo simulation to improve p value accuracy. I would suggest that you keep the seven figures after the decimal point.**

Optionally, you could have a third column, with category names, like so:

77	0.3333333	Left
66	0.3333333	Right
47	0.3333333	Forward

As in the book, select Analysis / Nonparametric / Chi-square Goodness of Fit. You are first asked to select ‘observed counts’; choose the left-hand column and press OK. Then you select the second column for ‘select cell probabilities or expected counts’. Then for category names, either select these and press ‘OK’ or if you do not have category names, press ‘Skip’. The p value for this should be 0.0263 – which fits into the critical value given in the book: p < .05 For the moment, I suggest just pressing ‘Skip’ to return to the workbook. We will deal with the further test when we get to the next exercise, where it gets more interesting (well, a bit).

(8) Chi Square Goodness of Fit Test, second exercise (page 52). We enter the following data onto the worksheet (adding category labels if you want), following the same procedure.

35	0.3333333
37	0.3333333
33	0.3333333

The outcome is p = 0.892 – the same as in the book. However, StatsDirect goes further, offering more accurate p values via a Monte Carlo simulation. For this, use the new dialog box at the top. Each time you press ‘Calculate’, you get a new suggested p value. Different ‘seeds’ are used to start each calculation, so there is no point in my printing values, as they will be different. What you will find after doing this several times is that the figures vary between about 0.0903 and 0.0904 after rounding, so it is best to round to two figures, i.e. 0.90 The confidence intervals provided by the simulation will also suggest how to round the figures. The rounded simulation statistic – here .90 – should be quoted, as it is to be considered to be more accurate than our original 0.89

Press ‘Skip’ when you finally want to return to the data sheet.

(If you return to the previous exercise to do this, with 77, 66 and 47, you will find that the simulation differs very little from the original p value, except in the rounding, so again you would round it to two decimal places: p = .03)

(9) Chi Square Test of Association, second exercise (page 53). The book's p value, p = .0153, is still correct. However, it should be read from the Yates-corrected Chi Square statistic.

It should also be noted that Fisher's exact test is also offered. This is for where the total sample is below 20 or one of the expected frequencies is less than 5.

(10) Pearson (page 63). The book gives r as .9871 and r squared as .97436; these are now .97082 and .942491 respectively.

(11) Tests of significance in survival analysis (pages 96 to 97). Peto's Log Rank test no longer appears to be available.