SECOND EDITION

Essential Statistics for the Pharmaceutical Sciences

Philip Rowe

WILEY

Essential Statistics for the Pharmaceutical Sciences

Essential Statistics for the Pharmaceutical Sciences

Second Edition

Philip Rowe

Liverpool John Moores University, UK

WILEY

This edition first published 2016 © 2016 by John Wiley & Sons, Ltd.

Registered Office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

The right of the author to be identified as the author of this work has been asserted in accordance with the UK Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author(s) have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Rowe, Philip, author.
Essential statistics for the pharmaceutical sciences / Philip Rowe. – Second edition.
p. ; cm.
Includes index.
ISBN 978-1-118-91338-3 (cloth) – ISBN 978-1-118-91339-0 (pbk.)
I. Title.
[DNLM: 1. Research Design. 2. Statistics as Topic. 3. Pharmacology–methods. QV 20.5]
RS57
615′.1072–dc23
2015015316

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

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Cover images: © Ma-k/iStockphoto, © FotografiaBasica/iStockphoto, © Polygraphus/iStockphoto

Set in 10.5/12.5pt Minion by SPi Global, Pondicherry, India

1 2016

To

Carol, Joshua and Nathan

for continued support

Contents

Preface

At whom is this book aimed?

Statisticians or statistics users?

The starting point for writing this book was my view that most existing statistics books place far too much emphasis on the mechanical number crunching of statistical procedures. This makes the subject seem extremely tedious and (more importantly) diverts attention from what are actually vital and interesting fundamental concepts. I believe that we need to distinguish between 'Statisticians' and 'Statistics users'. The latter are the people at whom this book is aimed – those thousands of people who have to use statistical procedures without having any ambition to become statisticians.

There is any number of student programmes which include an element of statistics. These students will have to learn to use at least the more basic statistical methods. There are also those of us engaged in research in academia or industry. Some of us will have to carry out our own statistical analyses and others will be able to call on the services of professional statisticians. However, even where professionals are to hand, there is still the problem of communication; if you don't even know what the words mean, you are going to have great difficulty explaining to a statistician exactly what you want to do. The intention is that all of the above should find this book useful.

As a statistics user, what you really need to know is:

- Why are statistical procedures necessary at all?
- How can statistics help in planning experiments?
- Which procedure should I employ to analyse the results?
- What do the statistical results actually mean when I've got them?

This book is quite happy to treat any statistical calculation as a black box. It will explain what needs to go into the box and it will explain what comes out the other end. But do you really need to know what goes on inside the box? This approach isn't just lazy or negative. By stripping away all the irrelevant bits, we can focus on the aspects that actually matter. This book will try to concentrate on the issues listed above – the things that statistics users really do need to understand.

To what subject area is the book relevant?

All the procedures and tests are illustrated with practical examples and data sets. The cases are drawn from the pharmaceutical sciences and this is reflected in the book's title. However, pretty well all the methods described and the principles explored are perfectly relevant to a wide range of scientific research, including pharmaceutical, biological, biomedical and chemical sciences.

At what level is it aimed?

The book is aimed at everybody from undergraduate science students and their teachers to experienced researchers.

The first few chapters are fairly basic. They cover data description (mean, median, mode, standard deviation and quartile values) and introduce the problem of describing uncertainty due to sampling error (Standard Error of the Mean and 95% Confidence Interval for the mean). These chapters are mainly relevant to first year students.

Later chapters then cover the most commonly used statistical tests with a general trend towards increasing complexity. The approach used is not the traditional one of giving equal weight to a wide range of techniques. As the focus of the book is the issues surrounding statistical testing rather than methods of calculation, one test (the two-sample *t*-test) has been used to illustrate all the relevant issues (Chapters 7–11). Further chapters then deal with other tests more briefly, referring back to general principles that have already been established.

What has changed since the first edition of this book in 2007?

My motivation for producing a second edition has very little to do with the arrival of any new statistical methods that are likely to have broad applicability for working pharmaceutical scientists – there are precious few.

So, why a new edition? I provide statistical advice to researchers in diverse areas of pharmaceutical science (and beyond) and the change I have noticed is an increased familiarity and confidence with the use of statistical packages. This brings both opportunities and pitfalls.

Opportunities

There are several statistical methods that I considered covering in the first edition but I concluded, at that time, that very few researchers would have the confidence to tackle them. Hopefully we have now moved on. For this edition I have added analysis of covariance, logistic regression, measures of agreement (e.g. Cronbach's Alpha

and Cohen's Kappa) and survival analysis. Many of these are more advanced than the topics in the first edition, but with some clear explanatory material (which I hope I have supplied) and relatively easy to use statistical packages, most pharmaceutical scientists should be perfectly capable of applying them.

Pitfalls

On the negative side, powerful statistical packages also offer new and improved methods to make a complete fool of yourself. Where I have seen examples of this over the last seven years I have tried to include warnings in this new edition.

Other new material

Apart from the completely new topics listed earlier, I have also filled in a number of gaps from the first edition. Many of these additions concern studies that generate simple dichotomous outcomes (e.g. Yes/No or Success/Failure). I have added the use of the Relative Risk, Odds Ratio and Number Needed to Treat (RR, OR and NNT) as descriptors of the extent of change in a dichotomous outcome. I have also described Fisher's and McNemar's tests as additions to the simple chi-square test which was included in the first edition.

Finally, when you teach statistics to various groups of students, year in, year out, you inevitably have the occasional light-bulb moment, when you realise that there is actually a much better way to explain something than the awkward method you have used for the last 30 years. Some of these are scattered around the book.

Key point and pirate boxes

Key point boxs

Throughout the book you will find key point boxes that look like this:

Proportions of individuals within given ranges

For data that follows a normal distribution:

- About two-thirds of individuals will have values within 1 SD of the mean.
- About 95% of individuals will have values within 2 SD of the mean.

These never provide new information. Their purpose is to summarise and emphasise key points.

Pirate boxes

You will also find pirate boxes that look like this:

Switch to a one-sided test after seeing the results

Even today, this is probably the best and most commonly used statistical fiddle.

You did the experiment and analysed the results by your usual two-sided test. The result fell just short of significance (*P* somewhere between 0.05 and 0.1) There's a simple solution – guaranteed to work every time. Re-run the analysis, but change to a one-sided test, testing for a change in whatever direction you now know the results actually suggest.

Until the main scientific journals get their act into gear, and start insisting that authors register their intentions in advance, there is no way to detect this excellent fiddle. You just need some plausible reason why you 'always intended' to do a one-tailed test in this particular direction, and you're guaranteed to get away with it.

These are written in the style of Machiavelli, but are not actually intended to encourage statistical abuse. The point is to make you alert for misuses that others may try to foist upon you. Forewarned is forearmed.

The danger posed, is reflected by the number of skull and cross-bone symbols.

 Minor hazard. Abuse easy to spot or has limited potential to mislead.

 Moderate hazard. The well-informed (e.g. readers of this book) should spot the attempted deception.

 Severe hazard. An effective ruse that even the best informed may suspect, but never be able to prove.

A potted summary of this book

The book is aimed at those who have to use statistics, but have no ambition to become statisticians *per se*. It avoids getting bogged down in calculation methods and focuses instead on crucial issues that surround data generation and analysis (Sample size estimation, interpretation of statistical results, the hazards of multiple

testing, potential abuses etc.). In this day of statistical packages, it is the latter that cause the real problems, not the number-crunching.

The book's illustrative examples are all taken from the pharmaceutical sciences, so students (and staff) in the areas of pharmacy, pharmacology and pharmaceutical science should feel at home with all the material. However, the issues considered are of concern in most scientific disciplines and should be perfectly clear to anybody from a similar discipline, even if the examples are not immediately familiar.

Material is arranged in a developmental manner. The first six chapters are fairly basic, with special emphasis on random sampling error. The next block of five chapters uses the two-sample *t*-test to introduce a series of general statistical principles. Remaining chapters then cover other topics in (approximately) increasing order of complexity.

The book is not tied to any specific statistical package. Instructions should allow readers to enter data into any package and find the key parts of the output. Specific instructions for performing all the procedures, using Minitab or SPSS, are provided in a linked website (www.ljmu.ac.uk/pbs/rowestats/).

Statistical packages

There are any number of statistical packages available. It is not the intention of this book to recommend any particular one.

Microsoft Excel

Probably the commonest way to collect data and perform simple manipulations is within a Microsoft Excel (XL) spreadsheet. Consequently, the most obvious way to carry out statistical analyses of such data would seem to lie within XL itself. Let me give you my first piece of advice. Don't even consider it! The data analysis procedures within XL are rubbish – a very poor selection of procedures, badly implemented. (Apart from that, they are OK.) It is only at the most basic level that XL is of any real use (calculation of the mean, SD and SEM). It is therefore mentioned in some of the early chapters but not thereafter.

Other packages

A decision was taken not to include blow by blow accounts of how to perform specific tests using any package, as this would excessively limit the book's audience. Instead, general comments are made about:

- Entering data into packages;
- The information that will be required before any package can carry out the procedure;
- What to look for in the output that will be generated.

The last point is usually illustrated by generic output. This will not be in the same format as that from any specific package, but will present information that they should all provide.

Detailed instructions for Minitab and SPSS on the website

As Minitab and SPSS clearly do have a significant user base, detailed instructions on how to use these packages to execute the procedures in this book are available through the website (www.ljmu.ac.uk/pbs/rowestats/). These cover how to:

- Arrange the data for analysis.
- Trigger the appropriate test.
- Select appropriate options where relevant.
- Find the essential parts of the output.

About the website

Supplementary material, including full data sets and detailed instructions for carrying out analyses using packages such as SPSS or Minitab, is provided at:

www.ljmu.ac.uk/pbs/rowestats/

Part 1

Presenting data

1
Data types

This chapter will ...

- Set out a system for describing different types of data.
- Explain why we need to identify the type of data with which we are dealing.

1.1 Does it really matter?

To open a statistics book with a discussion of the way in which data can be categorised into different types probably sounds horribly academic. However, the first step in selecting a data handling technique is generally identifying the type of data with which we are dealing. So, it may be dry, but it does have real consequences.

We will discuss three types of data. These go under a variety of names. The names that this book will use are (with common alternatives in brackets):

- Interval scale (Continuous measurement data)
- Ordinal scale (Ordered categorical data)
- Nominal scale (Categorical data)

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

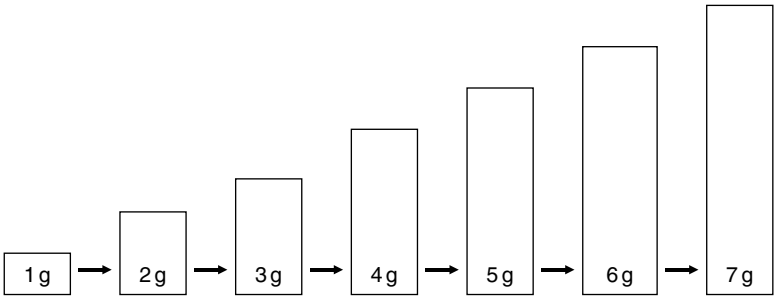


Figure 1.1 Interval scale data – a series of weights (1–7 g)

1.2 Interval scale data

The first two types of data that we will consider are both concerned with the measurement of some characteristic. 'Interval scale' (or what is sometimes called 'Continuous measured') data includes most of the information that would be generated in a laboratory. These include weights, lengths, timings, concentrations, pressures etc. Imagine we had a series of objects weighing 1, 2, 3 and so on up to 7 g, as in Figure 1.1.

Now think about the differences in weights as we step from one object to the next. These steps, each of one unit along the scale, have the following characteristics:

1. *The steps are of an exactly defined size*. If you told somebody that you had a series of objects like those described above, he or she would know exactly how large the weight differences are as we progressed along the series.

2. *All the steps are of exactly the same size*. The weight difference between the 1 and 2 g objects is the same as the step from 2 to 3 g or 6 to 7 and so on.

Because these measurements have constant sized steps (intervals), the measurement scale is described as a 'Constant interval scale' and the data as 'Interval scale'. Although the weights quoted in Figure 1.1 are exact integers, weights of 1.5 or 3.175 g are perfectly possible, so the measurement scale is said to be 'Continuous'.

1.3 Ordinal scale data

Again measurement is involved, but the characteristic being assessed is often more subjective in nature. It's all well and good to measure nice neat objective things like blood pressure or temperature, but it's also a good idea to get the patient's angle on

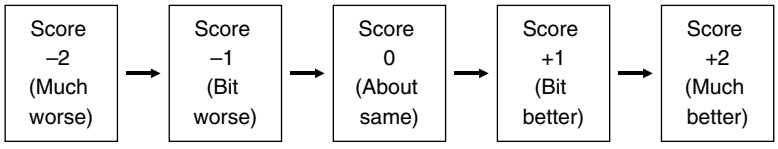


Figure 1.2 Ordinal scale data – scores for patient responses to treatment

how they feel about their treatment. The most obvious way to do this is as a score, of (say) –2 to +2 with the following equivalences:

–2 = Markedly worse
–1 = A bit worse
0 = About the same
+1 = A bit better
+2 = Markedly better

In this case (Figure 1.2) all we know is that if one patient reports a higher value than another, they are more satisfied with their outcome. However, we have no idea how much more satisfied he/she might be.

Since we have no idea how large the steps are between scores, we obviously could not claim that all steps are of equal size. In fact, it is not even necessarily the case that the difference between scores of –2 and 0 is greater than that between +1 and +2. So, neither of the special characteristics of a constant interval scale apply to this data.

The name 'Ordinal' reflects the fact that the various outcomes form an ordered sequence going from one extreme to its opposite. Such data is sometimes referred to as 'Ordered categorical'. In this case the data is usually discontinuous; individual cases being scored as –1 or +2 and so on, with no fractional values.

1.4 Nominal scale data

In this case there is no sense of measuring a characteristic; we use a system of classifications, with no natural ordering. For example, one of the factors that influences the effectiveness of treatment could be the specific manufacturer of a medical device. So, all patients would be classified as users of 'Smith', 'Jones', or 'Williams' equipment. There is no natural sequence to these; they are just three different makes.

With ordinal data we did at least know that a case scored as (say) +2 is going to be more similar to one scored +1 than to one scored 0 or –1. But, with nominal data, we have no reason to expect Smith or Jones equipment to have any special degree of similarity. Indeed the sequence in which one would list them may be entirely arbitrary.

Quite commonly there are just two categories in use. Obvious cases are Male/Female, Alive/Dead or Success/Failure. In these cases, the data is described as "Dichotomous".

Data types

Interval scale: Measurements with defined and constant intervals between successive values. Values are continuous.
Ordinal scale: Measurements using classifications with a natural sequence (lowest to highest) but with undefined intervals. Values are discontinuous.
Nominal scale: Classifications that form no natural sequence.

1.5 Structure of this book

The structure of this book is largely based upon the different data types. Chapters 3 to 16 all deal with the handling of continuous measurement data, with Chapters 17 to 20 focusing on categorical data; and then Chapter 21 covers ordinal data.

1.6 Chapter summary

When selecting statistical procedures, a vital first step is to identify the type of data that is being considered.

Data may be:

- Interval scale: Measurements on a scale with defined and constant intervals. Data is continuous.
- Ordinal scale: Measurements on a scale without defined intervals. Data is discontinuous.
- Nominal scale: Classifications that form no natural sequence.

2
Data presentation

This chapter will ...

- Describe the use of numerical tables, bar charts, pie charts, pictorial symbols and scattergrams.
- Consider which type of data is appropriate for each method of presentation.
- Stress the importance of considering the type of readership at which a presentation is targeted. (Scientifically literate or general public?)
- Assess the strengths and weaknesses of each method.

Statistical analyses allow complex data to be summarised objectively and clearly in one or two numbers. However, used in isolation, such analyses can be horribly misleading and this book will emphasise the value of the pictorial representation of data. A clear graph or bar chart will often alert you to some important aspect of the data that would be missed if we relied solely on one or two dry, numerical statistics.

This chapter will not attempt to identify any method of data presentation as a universal panacea. Selecting the best method depends upon the type of data involved (interval, ordinal or nominal) and the nature of the readership being addressed (scientifically trained or the general public?). Roughly speaking, the various methods

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

A picture is worth 1000 words. Data should be assessed both pictorially and statistically

- Pictures often reveal unsuspected aspects of the data.
- Statistics provide an objective test of what the data really does (or does not) demonstrate.

described are ordered according to their 'Friendliness' (i.e. accessibility for a non-scientific readership), starting with the least friendly.

2.1 Numerical tables

We have three different anti-emetic drugs to be used in conjunction with a chemotherapy regime. One is our current standard medicine and then we have two new candidates. We want to look at the degree of nausea reported when they are used. Each anti-emetic is administered to 30 patients and they assess nausea on a four-point ordinal scale (1 = None, 2 = Slight, 3 = Moderate, 4 = Severe).

The results are shown in Table 2.1.

Presenting the data as a numerical table has both good and bad aspects:

- *Good:* The full details of the original data are available. No doubt we will have done our own analysis of the data, but others may want to analyse the same data in a different way or may wish to combine this data with that from other studies in a meta-analysis. Reporting the data as a numerical table makes such re-analyses possible.

- *Bad:* These tables are rather forbidding and lacking in immediacy. If your readership is highly numerate – for example colleagues at a scientific conference – they will not be put off by this table. However, if you showed this type of table to a lay audience, their eyes would glaze over and all higher intellectual functions would face imminent shutdown.

Table 2.1 Number of patients reporting varying degrees of nausea following use of three different anti-nausea drugs

	Current standard	Candidate one	Candidate two
1 (None)	3	4	8
2 (Slight)	6	6	12
3 (Moderate)	19	17	9
4 (Severe)	2	3	1

Even with a numerate audience there is still the problem of immediacy. About the only thing that emerges quickly is that most patients have suffered moderate levels of nausea, absence of nausea being quite rare. But, the more important issue is the comparison of one anti-emetic with another. If you look carefully enough it is possible to see that Candidate One barely differs from the current standard but Candidate Two may be a useful improvement, however such niceties certainly don't stand out immediately. The stacked bar chart discussed in the next section gets the message over more dramatically.

Numerical tables

Good: Raw data available for further analysis.
Bad: Unfriendly and poor immediacy.

2.2 Bar charts and histograms

Any of the three types of data (interval, ordinal or nominal) can be reported as a bar chart. But the ease of doing so varies.

2.2.1 Simple bar charts

2.2.1.1 Nominal data A series of patients are offered four different formulations of cough medicine. They are asked simply to indicate which of the four they most favour. The outcomes – numbers preferring product A, B, C or D – form nominal scale data. Nominal data is always discontinuous and generally falls into a small number of natural and distinct categories. It is therefore ideal for presentation as a bar chart, as in Figure 2.1. Detailed instructions for producing all the figures in this chapter using SPSS are provided on the accompanying website (www.ljmu.ac.uk/pbs/rowestats/).

Note that the horizontal scale represents nominal scale data. It does not represent a continuous scale of measurement. To emphasise the discrete nature of the categories, spaces are left between the bars.

How to produce the graphs in this chapter

Detailed instructions for producing all the figures in this chapter using SPSS are provided on www.ljmu.ac.uk/pbs/rowestats.

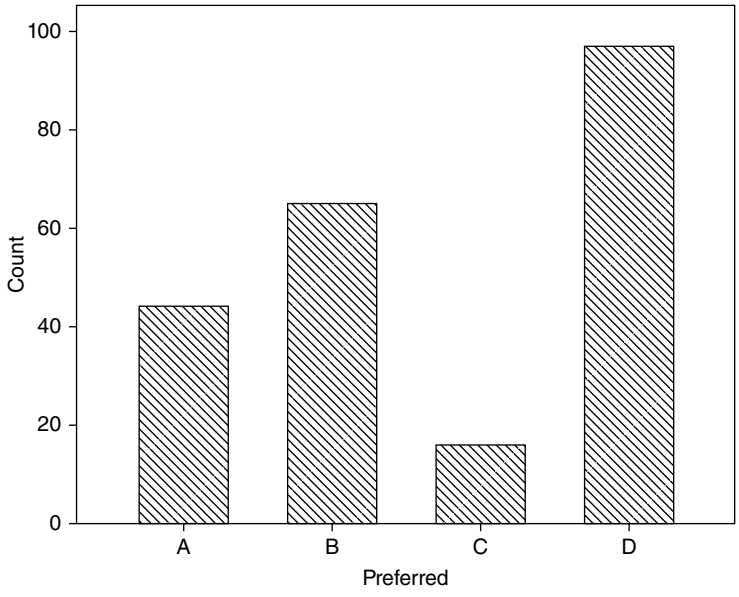


Figure 2.1 Simple bar chart using nominal data – numbers of subjects preferring formulation A, B, C or D of a liquid medicine

2.2.1.2 Ordinal data Ordinal data is most commonly collected using a scale of measurement with a small number of possible values, so it is usually appropriate for use in bar charts. The example below (Figure 2.2) concerns a Likert scale measure for opinion on the time a treatment requires, using a five-point scale. Ordinal data forms a natural basis for a simple bar chart, especially as it is intuitive to see the left-hand end of the horizontal axis as the low end of a scale of measurement, with higher values as we move to the right.

As with Figure 2.1, gaps have been left between the bars, to emphasise that the horizontal scale is not a continuous scale of measurement; there were no opinions recorded at any points in between the five values shown.

A simple bar chart is adequate to describe a single set of outcomes. However, if we want to compare two (or more) sets of outcomes, we are going to need something fancier. Sections 2.2.2 and 2.2.3 describe two ways to do it.

2.2.2 Three-dimensional bar charts

A group of volunteers were each shown a tablet and asked to choose one word that best expressed their opinion of its appearance. They could choose from a list of five ('Dangerous', 'Powerful', 'Neutral', 'Safe' or 'Calming'). All tablets were identical apart from their colour, which was either red or green. The results are presented as a three-dimensional bar chart in Figure. 2.3. Patients are influenced by the colour, with red being seen as powerful or dangerous while green is safe or calming. That difference stands out immediately in the bar chart.

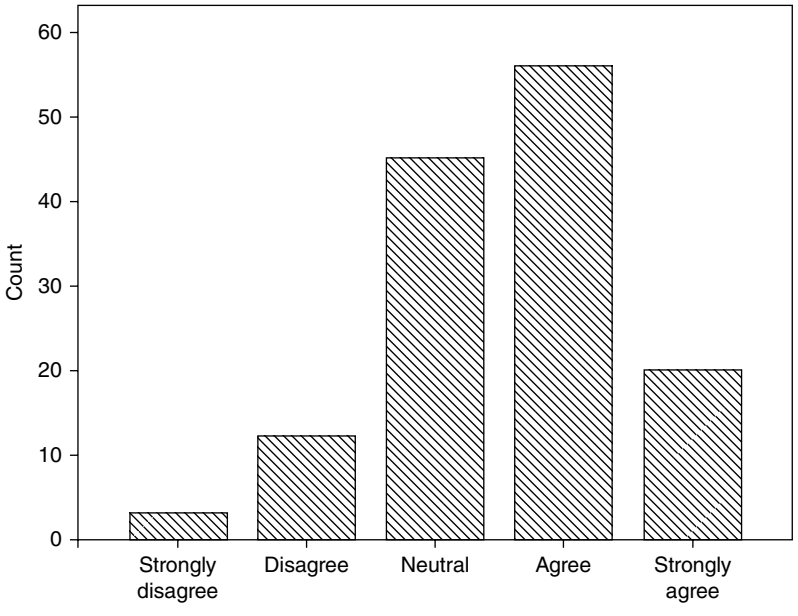


Figure 2.2 Bar chart based on ordinal data – opinions on the statement 'The treatment took too long'

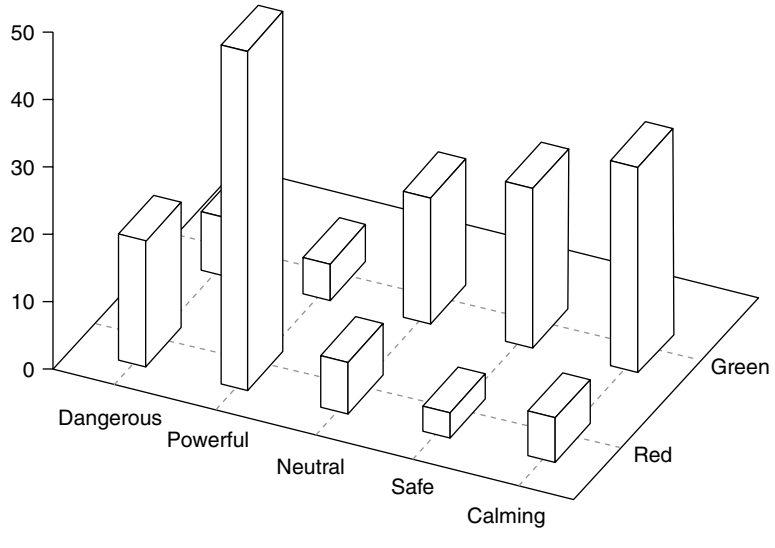


Figure 2.3 Three-dimensional bar chart of subjects' impressions of tablets that are identical in all aspects other than colour

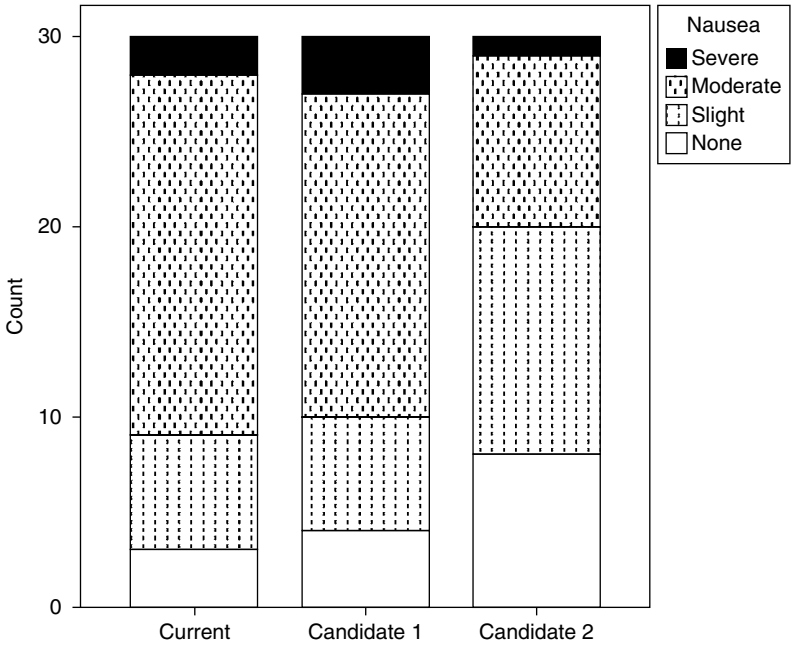


Figure 2.4 Stacked bar chart for levels of nausea with different anti-emetics

2.2.3 Stacked bar chart

We might try to present the data from Table 2.1 concerning levels of nausea with various anti-emetics as a three-dimensional bar chart, but it doesn't actually work out too well, because some shorter bars get hidden behind taller ones. However, stacked bars can be used to portray the data quite effectively. It is immediately obvious from Figure 2.4 that anti-nausea drug Candidate One has produced very little change relative to the existing standard. But Candidate Two is visibly different; there is a majority with No or only Slight nausea.

Stacked bar charts could be used with nominal data, but they work especially well with ordinal data such as these nausea grades.

2.2.4 Histograms

Trying to present interval type data as a bar chart is less straightforward. This type of data is usually measured on a continuous scale and so there are large numbers of different values. However we can artificially convert it to a more limited number of categories by breaking it up into bands.

For example, we have some observations of patients' temperatures five days following surgery. We could classify each individual into one of the following bands based on their temperatures:

36.8 – 37.0°C
37.1 – 37.3°C
and so on
38.6 – 38.8°C

Note that these bands must fulfil three requirements:

- No gaps. If we had bands of 36.5–36.7 and then 36.9– 37.1°C, we could not allocate a temperature of 36.8°C to a category.
- No overlaps. If we had bands of 36.5–36.7 and then 36.7–36.9°C, an individual with a value of 36.7°C could be allocated to two possible categories.
- All bands of equal width. If the first few bands covered a range of 0.3°C, but then we went to bands covering 0.6°C, the later categories would contain greater numbers of individuals. The increased heights of these bars would have nothing to do with these temperatures being commoner; it would just be an artefact of the way we had categorised the data.

The temperatures are then presented as in Figure 2.5.

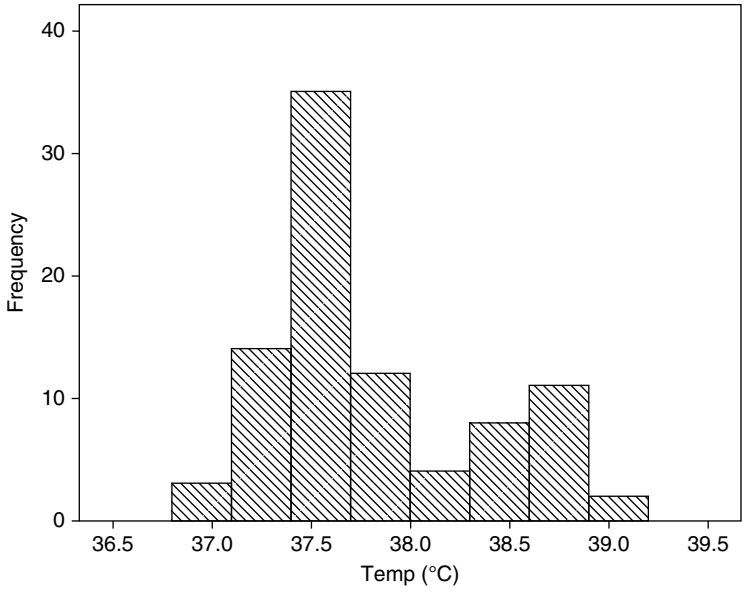


Figure 2.5 Histogram of patients' temperatures

The chart suggests that there is a distinct sub-population with elevated body temperatures – presumably they have become infected whereas the others have not.

Unlike all of the previous cases, the horizontal axis does now represent a continuous scale of measurement which contains no sudden breaks. To emphasise the continuous nature of the scale, we do not leave gaps between the bars.

Where the scale of measurement is essentially continuous, but has been artificially broken into bands, the resultant chart is given the special name of a 'Histogram'. Note that the term 'Histogram' should only be used in this context. Figures 2.1–2.4 are bar charts, but not histograms.

Histograms

A histogram is a bar chart using data that was originally on a continuously varying scale, but which has been subdivided into ranges to render it in a classified format. No gaps are left between the bars.

2.2.5 A general assessment of bar charts and histograms

Bar charts are probably somewhat less intimidating than numerical tables, for a non-scientifically oriented audience. However, they are still far from perfect. They are much better than numbers in terms of their immediacy. Not just patterns in single sets of data but also contrasts between sets of data are far more easily appreciated. The one loss is that we no longer have access to the exact data. It is possible to add numbers to the bars, but in many cases this is awkward. For example in Figure 2.3, several of the bars are partially hidden and we would have to write the number somewhere else on the chart and have an arrow connecting it to the appropriate bar. This clutters the diagram and much of the simplicity and clarity will be lost.

Bar charts and histograms

Good: Excellent immediacy for all main messages and reasonably friendly.
Bad: Often difficult to include exact values without loss of clarity.

2.3 Pie charts

2.3.1 Simple pie charts

We would almost never use pie charts to present data that was interval or ordinal in nature. Such data falls on a scale with a low and a high end, which is more naturally expressed in a bar chart. Pie charts are circular and simply don't match the needs of

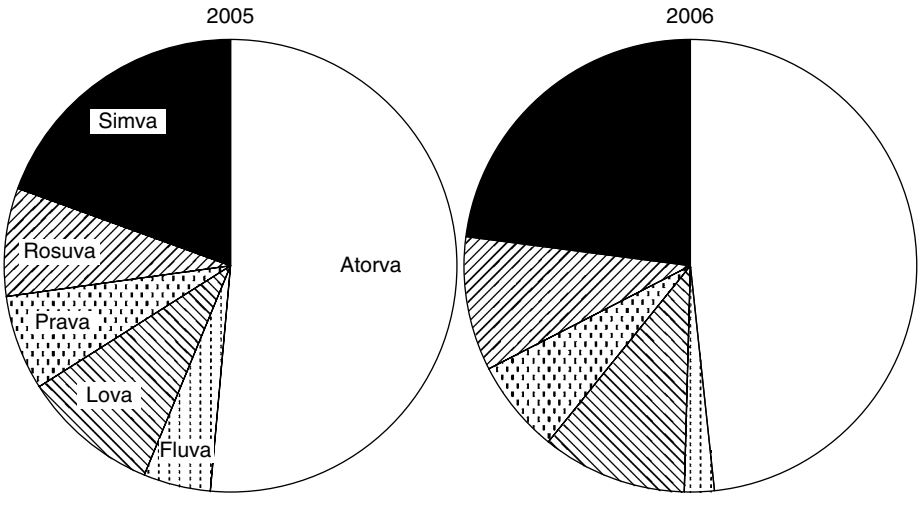


Figure 2.6 Pie chart of prescriptions for various statins in 2005 and 2006 (Atorvastatin, Fluvastatin, Lovastatin, Pravastatin, Rosuvastatin and Simvastatin)

measurement data; the high and low ends of the scale are not apparent. Pie charts are useful for nominal type data where there is no logical sequence to the categories.

Figure 2.6 shows numbers of patients treated with a variety of different cholesterol-lowering statin drugs in a group of United States hospitals in the years 2005 and 2006.

Simple messages, such as the predominant use of Atorvastatin and Simvastatin are conveyed with excellent immediacy. The style of presentation is also pretty unthreatening to even a non-numerate audience; people quite like the mental image of a pie being sliced into larger or smaller portions.

Unfortunately, pie charts don't convey changes in patterns as effectively as bar charts. Differences are easily seen in Figures 2.3 and 2.4, but in Figure 2.6 we have to check backwards and forwards between the two pie charts. Eventually, you may have noticed that the use of Simvastatin increased in 2006, but it most likely didn't hit your eye immediately. (Simvastatin lost its US patent in 2006 and became available in a cheaper generic form, hence the increased use.)

As with barcharts, the original numerical data is lost, unless we are prepared to add a lot of clutter to what are currently nice, clear figures.

Unfortunately you may meet a certain academic snobbery concerning pie charts. Some folk seem to see them as 'pretty pictures' that any Tom, Dick or Harry could understand.

2.3.2 Exploded pie charts

If one of the main points we want to convey is the increase in the use of Simvastatin in 2006, then Figure 2.6 is rather weak, but Figure 2.7 is a little better. By exploding the relevant slice we can ensure that it gets noticed.

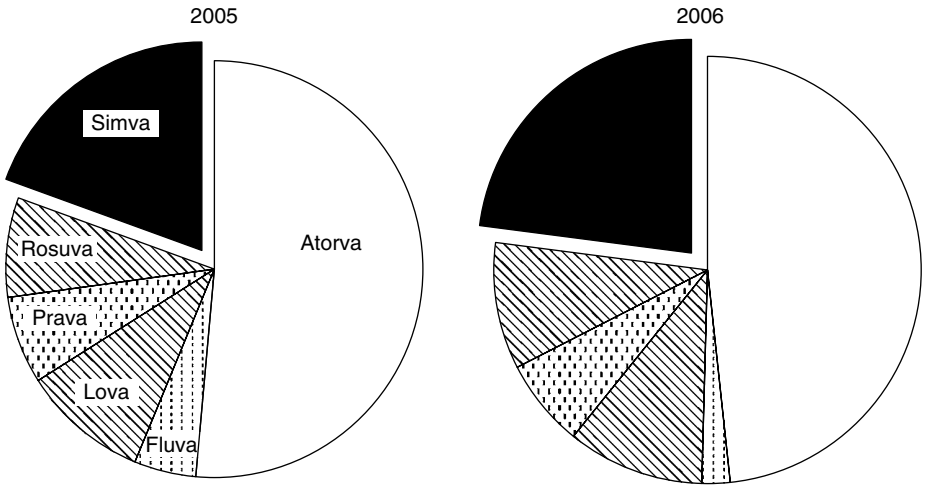


Figure 2.7 Exploded pie chart of prescriptions for various statins in 2005 and 2006 emphasising increased use of Simvastatin in 2006

Pie charts

Good: Friendly. Excellent immediacy for conveying which categories have the highest frequencies.
Bad: Only appropriate for nominal type data. Less immediate identification of changes in patterns. (Exploded slices may help.) Difficult to include exact values without loss of clarity. Academic snobbery.

2.4 Scatter plots

2.4.1 Dependent versus independent variable

All the data presentation methods we've looked at so far are appropriate for cases where there is just one measured value (parameter) being reported. Not uncommonly, two parameters will have been determined and we want to look at the relationship between them. For this we normally use a scatter plot.

At a number of places in this book we will meet the distinction between a 'Dependent' and an 'Independent' variable. If we find that two parameters (A and B) are related, then the question is how we would interpret that relationship. Is the value of A controlled by that of B or vice versa. For example, in a pharmacokinetic trial, the patients' body weights and their clearances of theophylline would probably be related to one another. (Clearance describes the efficiency with which a drug is

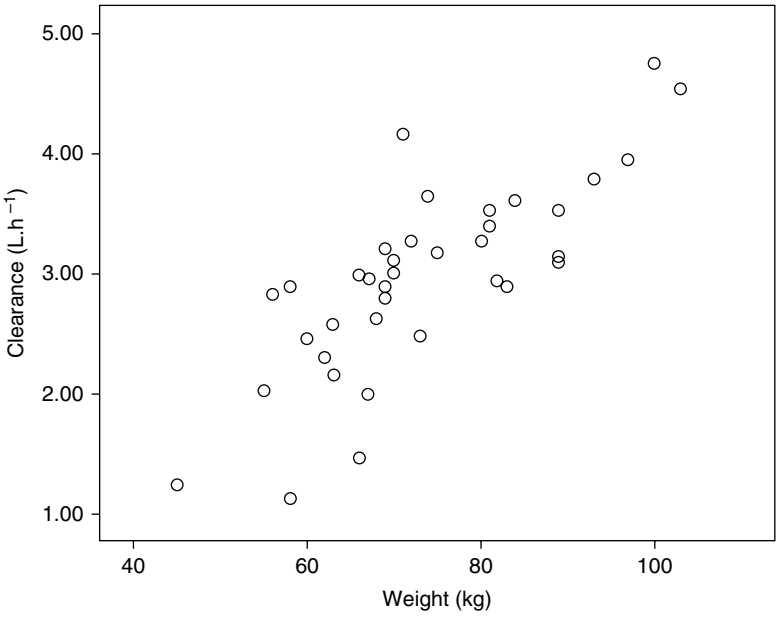


Figure 2.8 Scattergram of theophylline clearance versus body weight

eliminated from the body.) We could reasonably assume that it was the clearance that was controlled by the body weight and not vice versa. In this case clearance is the dependent variable and body weight is the independent. We then always plot the dependent variable up the vertical axis and the independent along the horizontal. It is also customary to describe this as 'Plotting clearance against body weight'. (Note the order – it's dependent against independent, not the other way round.) Figure 2.8 shows the data as a scatter plot.

Dependent and independent variables

The dependent variable should be plotted up the vertical (y) axis and the independent along the horizontal (x) axis.

We say that 'The dependent variable is plotted versus the independent'.

The vertical and horizontal axes may also be referred to as 'y' and 'x'. Other terms that are used (albeit less frequently, since nobody can ever remember which is which) are the 'Ordinate' and 'Abscissa' (Figure 2.9).

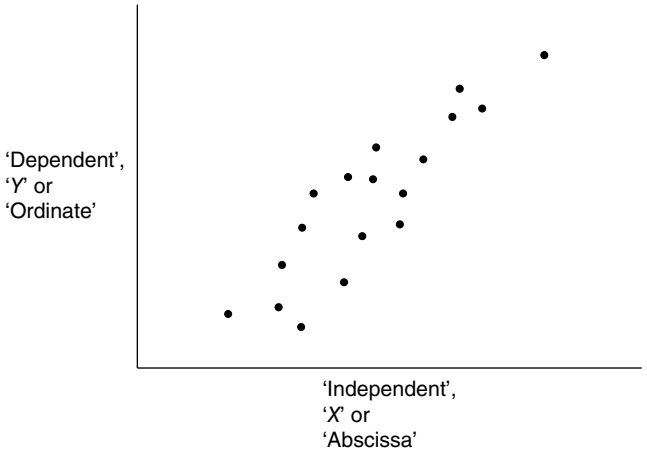


Figure 2.9 The vertical and horizontal axes may be referred to as 'Dependent and Independent', '*y* and *x*' or the 'Ordinate and Abscissa'

2.4.2 Scatter plots for data where there is no identifiable dependency within the data

In some cases, two parameters may show a clear relationship to one another, but neither can be identified as being dependent upon the other. This commonly arises when both parameters are dependent upon some third factor that causes them to vary together.

An example of the latter would be the volume of distribution and clearance of a drug. (Volume of distribution describes the apparent space into which a drug distributes when it spreads from the blood out into the tissues.) These two parameters are linked because both are dependent upon body weight. But there is no sense in which volume of distribution is dependent upon clearance or vice versa. In such a case we could equally well plot the data as volume versus clearance (Figure 2.10) or as clearance against volume.

2.4.3 Darell Huff and the 'Gee Whiz graphs'

Back in the 1950s Darrel Huff drew attention to one of those tricks long beloved by presenters of misleading data. The basic idea is that you convert a disappointingly shallow graph into one that shoots up is a pleasingly dramatic way. To achieve this we stretch the vertical scale. Stretching the vertical axis could of course lead to a graph that was excessively tall, but the real secret is to use only a small part of the available range of figures.

Consider some figures for improvement in cure rates for Fick–Tishus disease following the introduction of MeToomycin. Figures 2.11 and 2.12 describe the results. They look very different at first glance, but actually convey exactly the same results.

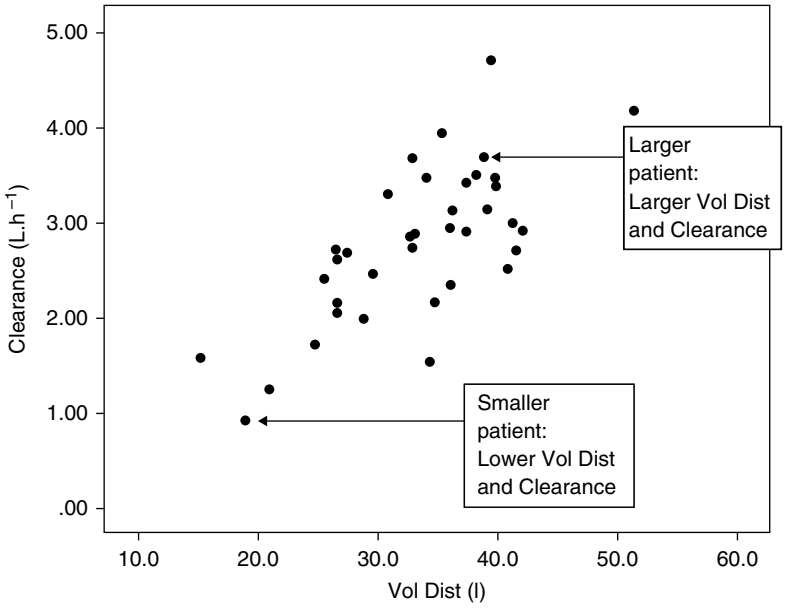


Figure 2.10 Scattergram of data where there is no identifiable dependency. Here, Clearance happens to be plotted against Weight, but the reverse pattern would be equally acceptable. Vol Dist = Volume of Distribution

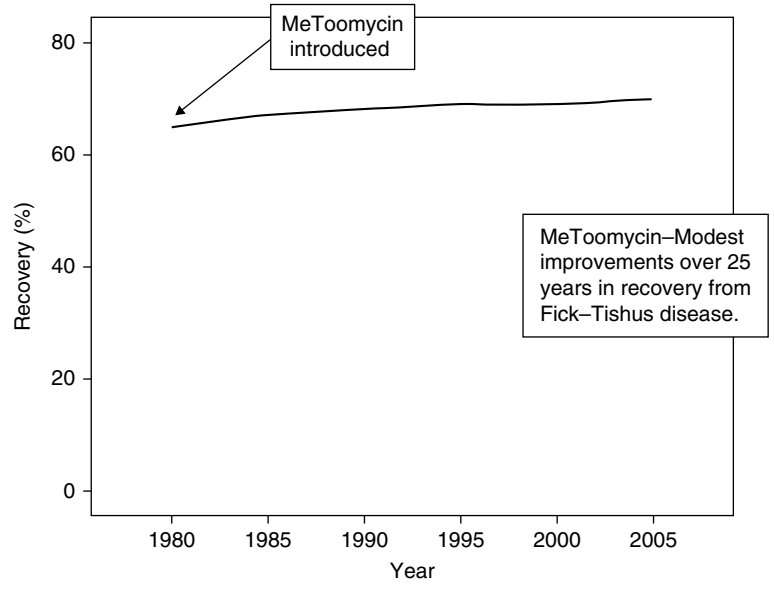


Figure 2.11 Line graph of recovery rates from Fick–Tishus disease since introduction of MeToomycin (honest but boring)

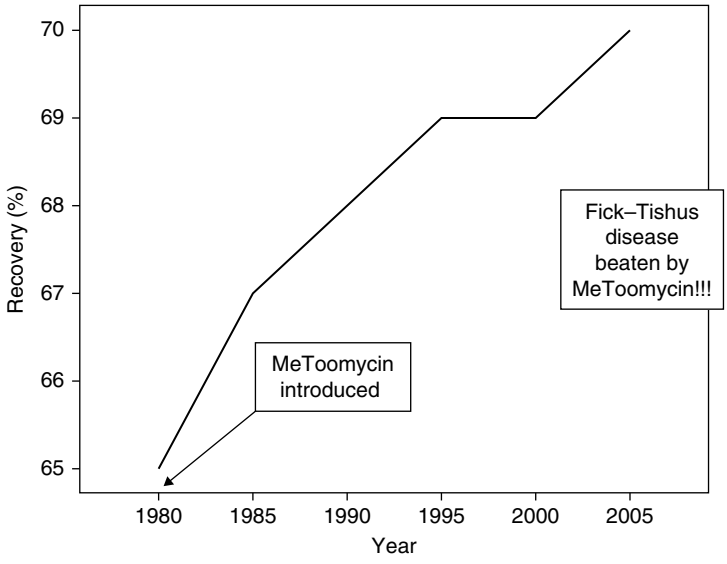


Figure 2.12 Deliberately misleading line graph that exaggerates the change in recovery rates by including only a small part of the range of values on the vertical axis

Both could be subjected to criticism. Figure 2.11 is rigorously honest, but 90% of it is boring blank space. Linked to this, there is also the problem that if we wanted to read off what the recovery rates were in any given year, it would be difficult to do so with any accuracy.

Figure 2.12 is far worse, being deliberately dishonest. A very small range of values has been stretched out to form the vertical axis, exaggerating the apparent increase in recoveries. The real crime is then the 'Gee whiz' headline designed to help the more gullible reader rush out and stock up on MeToomycin.

Apart from the abuse of the vertical axis in Figure 2.12, there is also the question that the graph only starts from 1980. We are given no idea what was going on before then. For all we know, general improvements in patient care may have been allowing a steady improvement in recovery rates for the last 20 years and the introduction of the alleged wonder drug might have had no impact whatsoever.

The pathetic becomes dramatic

A once brilliant scheme, now faded.

Even the most modest increase (or decrease) can be made to look impressive by quietly suppressing the zero on the vertical axis and expanding a small part of the scale. The problem is that Darrell Huff, in the best statistics book ever written ('How to lie with statistics') blew the cover on this one 50 years ago. There is no way you will get away with it in any reputable journal. However, if you're writing a polemical article for a popular magazine or newspaper, you can still fool many of the people much of the time.

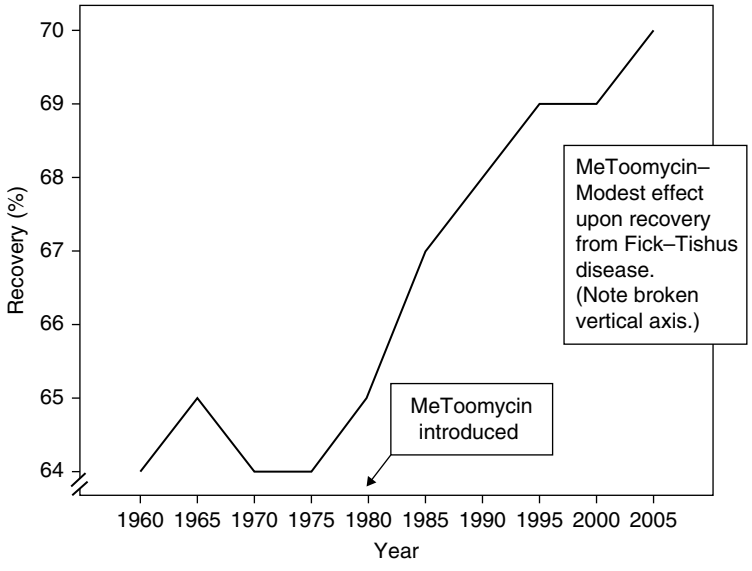


Figure 2.13 Fair and informative line graph of MeToomycin data

An acceptable presentation of the data is shown in Figure 2.13. It is similar to Figure 2.12, but with a more modest headline and a clear indication that the vertical axis is incomplete. If data for the period prior to the introduction of MeToomycin is available, it would be useful to include it. From this figure we can see that there was no consistent progress with this disease during the 20 years prior to the introduction of MeToomycin. We would also gather that even after its introduction, progress was less than miraculous.

2.5 Pictorial symbols

We have progressed from the least friendly mode of data presentation (numerical tables) to the much friendlier bar charts and pie charts. There is one more step we can take in our journey to nirvana – pictorial symbols. Figure 2.14 shows the utilisation of Wundadrug in a large district hospital.

From a strictly objective stand point these is absolutely nothing wrong with this figure. The meaning of the symbols is clearly defined. (One tablet symbol equals 10 000 tablets dispensed.) The escalating use of the drug is immediately obvious and we get an accurate sense of the scale of increase.

So what's wrong with it? Why would you be laughed at if you included it in a presentation to a 'learned' society? The answer is almost certainly its utter clarity. Any member of the ordinary public could understand it and therein lies the problem. The most important task for all academics is to convince the public we are much cleverer than them and, in that respect, Figure 2.14 is a disaster. Joe Bloggs on the number 13 bus could understand it just as easily as Professor Halfmoons from the Institute of Advanced Obscurantism.

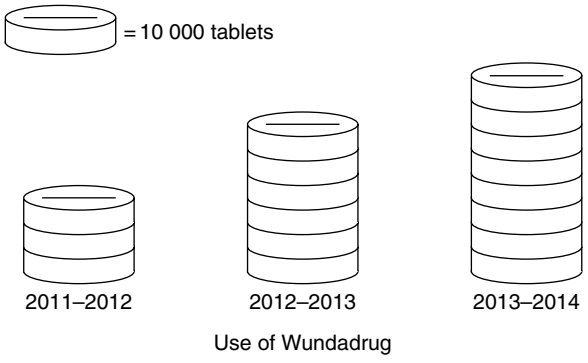


Figure 2.14 Pictorial representation of change in Wundadrug usage

While you may never be able to use pictorial symbols in the academic world, they can be a valuable way to make a point to the general public. Using Figure 2.14, you could achieve what is normally impossible – convey quantitative data to an audience that would instinctively run a mile from anything with numbers in it.

2.6 Chapter summary

Data should always be explored graphically as well as statistically. A picture is worth 1000 words.

Numerical tables allow readers to access the primary data for re-analysis, but are unfriendly for less numerate readers and fail to convey the main features of the data with any great immediacy.

Bar charts can be used for any type of data. They are reasonably friendly and convey the main aspects of the data with excellent immediacy. They usually result in the loss of access to the primary data. If data from a continuously varying scale is rendered into classes based upon ranges, a bar chart of such data is then called a histogram.

Pie charts should only be used for nominal data. They are very friendly and convey which classes occur most commonly with great immediacy. Changes that have occurred do not necessarily show up very clearly (exploded charts may help). Access to the primary data is generally lost.

Scatter plots are used to illustrate the relationship between two measured parameters. Where one parameter can be identified as being dependent upon the other, the dependent should be plotted up the vertical (y) axis and the independent along the horizontal (x). Beware of salespersons who make minor increases look disproportionately large by using only part of the y axis.

Pictorial symbols offer a unique opportunity to smuggle quantitative information into public information, without scaring your readers. Compare the response you would get with Figure 2.14 to what you might expect from the same data presented as a numerical table.

Detailed instructions for producing all the figures in this chapter using SPSS are provided on www.ljmu.ac.uk/pbs/rowestats.

Part 2

Interval-scale data

3

Descriptive statistics for interval scale data

This chapter will ...

- Review the use of the mean, median or mode to indicate how small or large a set of values are and consider when each is most appropriate.
- Describe the use of the standard deviation to indicate how variable a set of values are.
- Show how quartiles can be used to convey information similar to that mentioned above, even in the presence of extreme outlying values.
- Discuss the problem of describing ordinal data.

3.1 Summarising data sets

Experiments and trials frequently produce lists of figures that are too long to be easily comprehended and we need to produce one or two summary figures that will give the reader an accurate picture of the overall situation.

With interval scale (continuous measurement) data, there are two aspects of the figures that we should be trying to describe:

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

- How large are they?
- How variable are they?

To indicate the first of these, we quote an 'Indicator of central tendency' and for the second an 'Indicator of dispersion'.

In this chapter we look at more than one possible approach to both of the above. It would be wrong to claim that one way is universally better than another. However, we can make rational choices for specific situations if we take account of the nature of the data and the purpose of the report.

Descriptive statistics

Indicators of central tendency: How large are the numbers?
Indicators of dispersal: How variable are the numbers?

3.2 Indicators of central tendency: Mean, median and mode

The term 'Indicator of central tendency' describes any statistic that is used to indicate an average value around which the data is clustered. Three possible indicators of central tendency are in common use: the mean, median and mode.

3.2.1 Mean – Ten batches of vaccine

The usual approach to showing the central tendency of a set of data is to quote the average. However, academics abhor such terms as their meanings are far too well known. We naturally prefer something a little more obscure – the 'Mean'.

Our first example set of data concerns a series of batches of vaccine. Each batch is intended to be of equal potency, but some manufacturing variability is unavoidable. A series of ten batches have been analysed and the results are shown in Table 3.1.

The sum of all the potencies is 991.5; and dividing that by the number of observations (ten) gives an average or mean activity of 99.15 Units.mL^{-1}.

The arithmetic is not open to serious question, but what we do need to consider is whether the figure we quote will convey an appropriate message to the reader. Although it may not strictly be justified, many readers will view that figure of 99.15 Units.mL^{-1} as indicating a typical figure. In other words, a batch with an activity of 99.15 Units.mL^{-1} is neither strikingly weak nor abnormally potent. A visual representation of the data is useful in testing whether this really is the case (see Figure 3.1).

Table 3.1 Potency of ten batches of vaccine

Potency (Units.mL^{-1})
106.6
97.9
102.3
95.6
93.6
95.9
101.8
99.5
94.9
103.4
Mean = 99.15

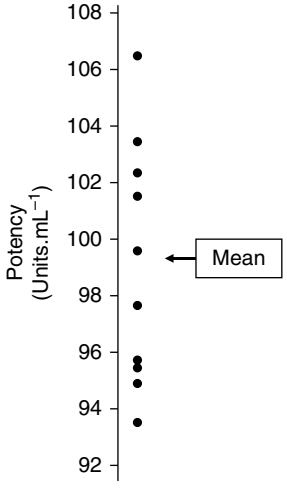


Figure 3.1 The mean satisfactorily indicates a typical potency among ten batches of vaccine

The mean is nicely in the middle of the bunch of values and does indicate a perfectly typical value. In this case the simple mean works fine and there is no need to consider more obscure alternatives. However, this is not always the case.

3.2.2 Median – Time to open a child-proof container

Fifteen patients were provided with their drugs in a child proof container of a design that they had not previously experienced. A note was taken of the time it took each patient to get the container open for the first time. The second column of Table 3.2 shows the results.

Table 3.2 Ranked times taken to open a child-proof container and calculation of the median

Rank	Time (s)	
1	2.2	
2	3.0	
3	3.1	
4	3.2	
5	3.4	
6	3.9	
7	4.0	
8	4.1	← The median value
9	4.2	
10	4.5	
11	5.1	
12	10.7	
13	12.2	
14	17.9	
15	24.8	
	Mean = 7.09 s	

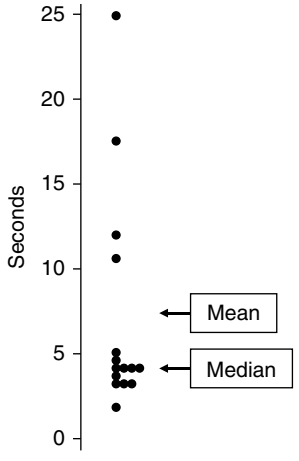


Figure 3.2 Median satisfactorily indicates a typical time taken to open a novel child-proof container. Mean is distorted by outliers

The mean is shown as 7.09 s, but again, we need to ask about the message that may be conveyed. Is this a representative figure? Figure 3.2 shows that it definitely is not.

Most patients got the idea more or less straight away and took only two to five seconds to open the container. However, four seem to have got the wrong end of the stick and ended up taking anything up to 25 s. These four have contributed a

disproportionate amount of time (65.6 s) to the overall total. This has then increased the mean to 7.09 s. We would not consider a patient who took 7.09 s to be remotely typical; they would be distinctly slow.

This problem of mean values being disproportionately affected by a minority of outliers arises quite frequently in biological and medical research. A useful approach in such cases is to use the median. To obtain this, the results shown in Table 3.2 are in ranked order (quickest at the top to slowest at the bottom) and their ranking positions are shown in the first column. Then we want to find the middle individual. This is the one ranked eighth, as there are seven patients slower and seven faster than this individual. The median is then the time taken by this eighth-ranking individual, that is 4.1 s.

Figure 3.2 shows that a patient taking 4.1 s is genuinely typical. So, in this case, the median is a better indicator of a representative figure.

3.2.2.1 Should we automatically use the median in such cases? It would be an over-generalisation to suggest that in every case where the data has outliers, the median is automatically the statistic to quote. We need to keep an eye on what use is to be made of the information. If we were dealing with the cost of a set of items and the intention was to predict the cost of future sets of such items, the mean would be appropriate even if there were outliers. The only way to predict the cost of a future batch of items would be to multiply the number of items by our estimate of the mean individual cost. For those purposes, the median would be useless.

3.2.2.2 The median is robust to extreme outliers The term 'Robust' is used to indicate that a statistic or a procedure will continue to give a reasonable outcome even if some of the data is aberrant. Think what would happen if the very slow individual who took 24.8 s to crack the child-proof container had instead taken a week to get it open. This character, who is currently ranked 15th, would still be ranked 15th and the median, which is established by the eighth ranker, would be quite unchanged at 4.1 s. In contrast, the mean would be hugely inflated if somebody took a week and so it is not considered robust.

This resistance to the effects of a few outlying values is the reason that the median is considered to be robust and the mean much less so.

The median

The middle ranking value. Especially useful with data containing occasional highly outlying values. It is 'Robust' (resists the influence of aberrant data).

In the right hands, this robustness is a useful characteristic and can allow us to indicate a representative figure even if a few bizarre figures have crept in. The danger

is always that of potential abuse, where somebody wants to use the median to hide the embarrassing handful of cases that spoilt their otherwise beautiful set of results.

Use the median to abolish those wretched outliers

Put this one up against a reasonably untutored readership and you should get away with it.

Simply quote the median for a group of observations, but make no mention of the outliers and under no circumstances publish the full set of individual results. That way, your median will be dominated by the majority of the data points and the outliers should nicely disappear from view.

3.2.2.3 Calculating the median where there is an even number of observations In the previous example (Table 3.2), the total number of timings is 15. With any odd number of data points, we can identify a single middle-ranking individual. However, with even numbers of observations there is no middle individual. In those cases, we identify the middle pair and then use the average of their values. An example of six timings is shown in Figure 3.3.

One slightly awkward consequence is that although the timings were apparently made to the nearest whole number of seconds, we can end up with a median value that contains a fraction.

3.2.3 Mode – A global assessment variable for the response to an anti-inflammatory drug

The condition of a series of patients with arthritis is recorded using a global assessment variable. This is a composite measure that takes account of both objective measures of the degree of inflammation of a patient's joints and subjective measures of their quality of life. It is set up so that higher scores represent better condition. The patients are then switched to a new anti-inflammatory product for three months

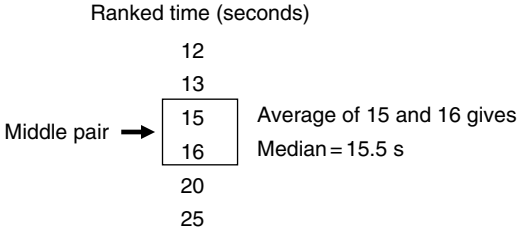


Figure 3.3 Calculation of a median with an even number of data points (timings in seconds)

Table 3.3 Individual changes in a global assessment variable following treatment with an anti-inflammatory (60 patients)

Score changes		
11	−9	−8
0	−9	2
−5	−15	−11
11	−13	−12
−13	−13	10
7	−18	−11
7	−13	9
−12	9	14
10	14	−9
−12	10	17
−10	−9	−14
6	11	−6
13	−11	13
−11	14	12
10	10	−6
−9	21	−9
9	6	2
8	−13	5
−12	−6	−7
10	−9	−12

and re-assessed using the same global measure. We then calculate the change in score for each individual. A positive value indicates an improvement and a negative one a deterioration in the patient's condition. Sixty patients participated and the results are shown in Table 3.3.

A histogram of the above data (Figure 3.4) shows the difficulty we are going to have.

Most of the patients have shown reduced joint inflammation, but there are two distinct sub-populations so far as side effects are concerned. Slightly under half the patients are relatively free of side effects, so their quality of life improves markedly, but for the remainder, side effects are of such severity that their lives are actually made considerably worse overall.

Mathematically, it is perfectly possible to calculate a mean or a median among these score changes and these are shown on Figure 3.4. However, neither indicator remotely encapsulates the situation. The mean (–0.77) is particularly unhelpful as it indicates a value that is very untypical – very few patients show changes close to zero. We need to describe the fact that, in this case, there are two distinct groups.

The first two sets of data we looked at (Vaccine potencies and Container opening timings) consisted of values clustered around some single central point. Such data are referred to as 'Unimodal'. The general term 'Polymodal' is used for any case with

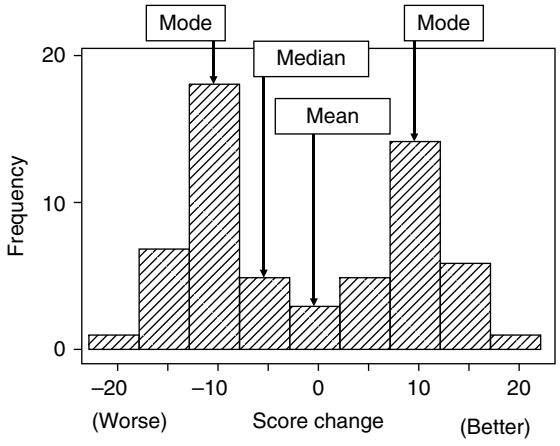


Figure 3.4 Individual changes in a global assessment score. Neither mean nor median indicates a typical value with bimodal data. Only modes achieve this

several clusterings. If we want to be more precise, we use terms such as bimodal, trimodal and so on to describe the exact number of clusters. The arthritis data might be described generally as polymodal or more specifically as bimodal.

With polymodal data we need to indicate the central tendency of each cluster. This is achieved by quoting the most commonly occurring value (the 'Mode') within each cluster. For the arthritis data, the two modes are score changes of –10 and +10. We would therefore summarise the results as being bimodal with modes of –10 and +10.

Unimodal and polymodal data

Unimodal: In a single cluster.
Polymodal: In more than one cluster (a general term).
Bimodal: Specifically in two clusters.
Trimodal: In three clusters.
etc.

Mode

A value that occurs with peak frequency. The only effective way to describe polymodal data.

3.2.4 Selecting an indicator of central tendency

There is a definite pecking order among the three indicators of central tendency described above. The mean is the industry standard and is the most useful for a whole range of further purposes. Unless, there are specific problems (e.g. polymodality or marked outliers), the mean is the indicator of choice. The median is met with quite frequently and the mode (or modes) tends only to be used when all else fails.

3.3 Describing variability – Standard deviation and coefficient of variation

3.3.1 Standard deviation

We have two tableting machines producing erythromycin tablets with a nominal content of 250 mg. The two machines are made by the 'Alpha' and 'Bravo' Tableting Machine Corporations respectively. Five hundred tablets are randomly selected from each machine and their erythromycin contents assayed. The results for both machines are shown as histograms in Figures 3.5 and 3.6. To produce these histograms the drug contents have been categorised into bands 2 mg wide. (The curves superimposed onto the histograms are discussed in the next chapter.)

The two machines are very similar in terms of average drug content for the tablets – both producing tablets with a mean very close to 250 mg. However, the two products clearly differ. With the Alpha machine, there is a considerable proportion of tablets with a content differing by more than 10 mg from the nominal dose

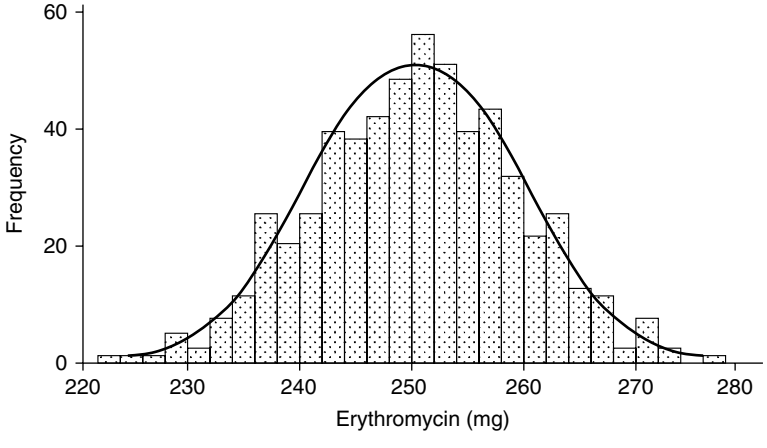


Figure 3.5 Histogram of erythromycin content of 500 tablets from an Alpha tableting machine (relatively variable product – large standard deviation)

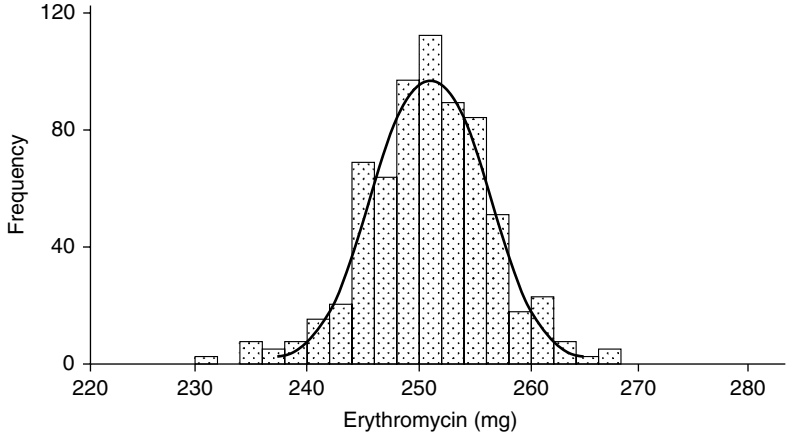


Figure 3.6 Histogram of erythromycin content of 500 tablets from a Bravo tableting machine (more consistent product – smaller standard deviation)

(i.e. below 240 or above 260 mg), whereas with the Bravo machine, such outliers are a lot rarer. An 'Indicator of dispersion' is required in order to convey this difference in variability.

The Standard Deviation (SD) is the most commonly accepted indicator of dispersion. This book generally discourages time-wasting manual calculations, but it is worth looking at an example of how the SD is calculated, because it makes clear what it reflects. Calculations of SDs for erythromycin content in samples of ten tablets from the two machines are shown in Table 3.4.

The first column shows the drug contents of ten individual tablets from an Alpha machine. The mean among these is 248.7 mg. The next column then shows the 'Deviation' of each individual tablet from the group mean. So, for example, the first tablet contained 249 mg of drug which is 0.3 mg more than the average. Hence the figure of 0.3 in the next column. The details of the rest of the calculation are not wildly interesting, but are presented for completeness. The next step is to take all the individual deviations and square them as in the third column. We then sum these figures and get 684.1. That is then divided by the number of observations minus one, yielding 76.01. Finally, we take the square root (8.72 mg) and that is the SD.

The key step in the calculation is the production of the second column – the individual deviations from the mean. The first machine produces rather variable tablets and so several of the tablets deviate considerably (e.g. –13.7 or +15.3 mg) from the overall mean. These relatively large figures then feed through the rest of the calculation, producing a high final SD (8.72 mg).

In contrast, the Bravo machine is more consistent and individual tablets never have a drug content much above or below the overall average. The small figures in the column of individual deviations then feed through the rest of the sausage machine, leading to a lower SD (3.78 mg).

Table 3.4 Erythromycin contents of ten tablets from an Alpha and a Bravo tableting machine and calculation of their Standard Deviations

Alpha Machine			Bravo Machine		
Erythro Content (mg)	Deviation from mean	Deviation squared	Deviation Content (mg)	Deviation from mean	Deviation squared
249	0.3	0.09	251	−0.1	0.01
242	−6.7	44.89	247	−4.1	16.81
252	3.3	10.89	257	5.9	34.81
235	−13.7	187.69	250	−1.1	1.21
257	8.3	68.89	254	2.9	8.41
244	−4.7	22.09	251	−0.1	0.01
264	15.3	234.09	252	0.9	0.81
249	0.3	0.09	255	3.9	15.21
255	6.3	39.69	244	−7.1	50.41
240	−8.7	75.69	250	−1.1	1.21
Mean		Total	Mean		Total
248.7		684.1	251.1		128.9
Sum of squared deviations = 684.1			Sum of squared deviations = 128.9		
Divide by $n-1$: 684.1/9 = 76.01			Divide by $n-1$: 128.9/9 = 14.32		
Take square root: = $\sqrt{76.01}$ = 8.72 mg (SD)			Take square root: = $\sqrt{14.32}$ = 3.78 mg (SD)		

3.3.1.1 Reporting the SD – The '±' symbol The ± symbol – reasonably interpreted as meaning 'more or less' – is used to indicate variability. With the tablets from our two machines, we would report their drug contents as:

Alpha machine: 248.7 ± 8.72 mg (± SD)

Bravo machine: 251.1 ± 3.78 mg (± SD)

Since it is conceivable that some statistic other than the SD has been quoted, it is useful to state this explicitly. When a result is simply quoted as one figure ± another figure, we would normally assume that it is the SD that has been provided.

The figures quoted above succinctly summarise the true situation. The two machines produce tablets with almost identical mean contents, but those from the Alpha machine are two to three times more variable.

3.3.1.2 Units of SD The SD is *not* a unitless number. It has the same units as the individual pieces of data. Since our data consisted of erythromycin contents measured in milligrams, the SD is also in milligrams.

The Standard Deviation (SD)

The general purpose indicator of variability (dispersion) for interval scale data.

3.3.2 The coefficient of variation

It is often easier to comprehend the degree of variability in a set of data if we express it in relative terms. A typical example of this is when we want to express the precision of an analytical method.

3.3.2.1 The precision of an HPLC analysis for blood imipramine levels expressed as the coefficient of variation An HPLC method for the measurement of blood levels of the antidepressant drug imipramine has been developed and as part of its validation we want to determine the reproducibility of its results. So, we analyse aliquots of the same blood sample on eight separate occasions. The mean result is 153 ± 9.33 ng.mL^{-1} (± SD). The figure of ±9.33 ng.mL^{-1} is our measure of assay variability. However, in isolation, this figure tells us precious little. To judge whether the method is acceptably precise we need to express the variability in a way that relates it to the mean amount being measured. The obvious way is as a percentage. The imprecision in the method is:

$$\pm 9.33 / 153 \times 100\% = \pm 6.1\%$$

We can now see that, for most purposes, the method would be acceptably precise.

What we have just calculated is referred to as the Coefficient of Variation (CoV).

The Coefficient of Variation

$$\text{Coefficient of Variation} = \frac{\text{SD}}{\text{Mean}}$$

Expresses variation relative to the magnitude of the data.

The result could have been expressed as either a fraction (0.061) or a percentage (6.1%). Because the coefficient of variation is a ratio, it is unitless (unlike the SD).

3.4 Quartiles – Another way to describe data

The only real alternative to the Mean and SD as descriptors for sets of measurements, is the system of 'Quartiles'. This enjoys a certain vogue in some research areas, but there are many others where you will almost never see it used.

We have already seen that the median is a value chosen to cut a set of data into two equal-sized groups. Quartiles are an extension of that idea. Three quartiles are chosen, so as to cut a data set into four equal-sized groups.

3.4.1 Quartiles – drug half-lives

The elimination half-lives of two synthetic steroids have been determined using two groups, each containing 15 volunteers. The results are shown in Table 3.5, with the values ranked from lowest to highest for each steroid.

Look at steroid number one first. Among the ranked half-lives, we highlight the half-lives that are ranked as fourth, eighth and 12th and these will provide the three quartile values. We have chopped the set of data into four equal-sized groups. There are three half-lives shorter than Q1, three between Q1 and Q2, three between Q2 and Q3 and three greater than Q3. We then check back for the actual values of the three highlighted cases and they are found to be:

Q1 = 5.4 h

Q2 = 7.8 h

Q3 = 10.0 h

Note that we always rank from the lowest to highest value and so Q1 always takes a lower value than Q3.

Table 3.5 Ranked half-lives for two steroids

Steroid One		Steroid Two		
Rank	Half-life (h)	Rank	Half-life (h)	
1	3.9	1	4.4	
2	4.0	2	4.5	
3	4.5	3	4.5	
4	*5.4*	4	*5.5*	← 4th ranked values = Q1
5	6.4	5	5.8	
6	6.5	6	5.9	
7	7.2	7	6.1	
8	*7.8*	8	*6.6*	← 8th ranked values = Q2 (Median)
9	8.6	9	7.2	
10	9.2	10	7.2	
11	9.3	11	7.3	
12	*10.0*	12	*7.8*	← 12th ranked values = Q3
13	10.6	13	8.5	
14	11.1	14	8.6	
15	15.8	15	9.1	

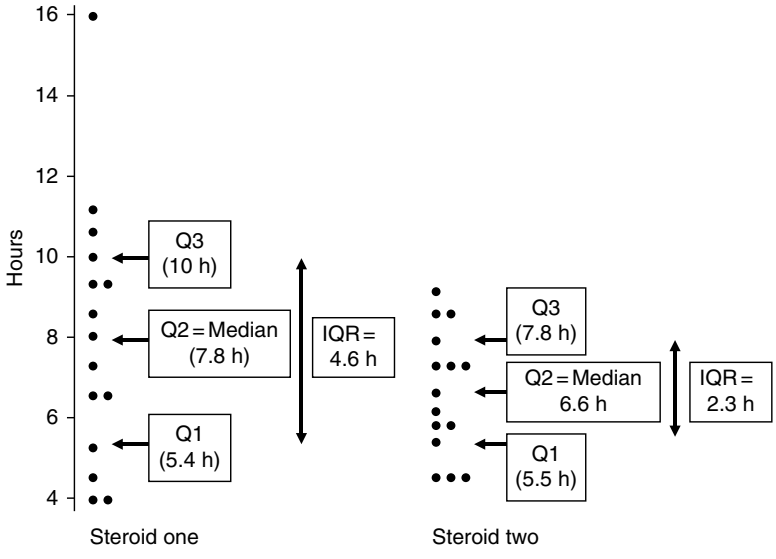


Figure 3.7 Steroid half-lives. The median (second quartile) indicates generally longer elimination half-lives for Steroid 1 relative to Steroid 2. The interquartile range (IQR) indicates greater variability for the first steroid

A similar exercise for steroid two yields quartiles values of 5.5, 6.6 and 7.8 h.

The quartiles are indicated in Figure 3.7.

3.4.1.1 The second quartile (Median) as an indicator of central tendency The second quartile has half the observations above it and half below and is therefore synonymous with the median. We have already established that the median is a useful and robust indicator of central tendency, especially when there are some extreme outlying values.

Q2 or Median = 7.8 h for the first steroid and 6.6 h for the second.

The median values suggest that somewhat longer half-lives are seen with the first steroid.

3.4.1.2 The inter-quartile range as an indicator of dispersion The inter-quartile range is defined as the difference between the upper and lower quartiles (Q3 – Q1). So:

For steroid 1, Inter-quartile range = 10.0 – 5.4 = 4.6 h.

The fact that the inter-quartile range does reflect dispersion can be appreciated if you compare steroid one with number two. With steroid two the half-lives are visibly less disperse and Q1 and Q3 are closer together:

For steroid two, inter-quartile range = 7.8 – 5.5 = 2.3 h.

The inter-quartile range for the half life of steroid two is thus only half that for steroid one, duly reflecting its less variable nature.

Just as the median is a robust indicator of central tendency, the interquartile range is a robust indicator of dispersion. Take the longest half-life seen with steroid one (15.8 h) and consider what would have happened if that individual had produced a half-life of 100 h (or any other extreme value). The answer is that it would make absolutely no difference to the inter-quartile range. The value of 15.8 h is already something of an outlier, but it had no undue effect on the inter-quartile range.

The Standard Deviation is much less robust. If there was an odd individual with an extremely long half-life for either steroid, then its value would deviate massively from the group mean and the SD would be inflated.

Median and inter-quartile range are robust indicators of central tendency and dispersion

The median (second quartile) and inter-quartile range can be used as an alternative method for describing the central tendency and dispersion of a set of measured data. Both are robust and can be useful where there are occasional extreme values.

3.4.2 Other quantiles

Having met the median and quartiles which divide sets of data into two or four ranges, we can then extend the principle to any system where $n - 1$ points are chosen to divide our data into n ranges. These methods are collectively referred to as

Quantile systems

Quantile systems divide ranked data sets into groups with equal numbers of observations in each group. Specifically:

- 3 *Quartiles* divide data into four equal-sized groups.
- 4 *Quintiles* divide it five ways.
- 9 *Deciles* divide it ten ways.
- 99 *Centiles* divide it 100 ways.

'Quantile' systems. Thus, the median and quartiles are specific examples of quantiles. Quantile systems that cut data into more than four ranges are really only useful where there are quite large numbers of observations.

The only other quantile systems that are used with any regularity are quintiles, deciles and centiles. There are four quintiles, which divide data into five ranges, nine deciles for ten ranges and 99 centiles that produce 100 ranges. The ninth decile is thus equivalent to the 90th centile and both indicate a point that ranks 10% from the top of a set of values.

3.5 Describing ordinal data

Summarising ordinal data is a challenge; there is no one method that is universally appropriate. To illustrate the challenge we will consider some comparative studies.

3.5.1 The mean is not generally considered appropriate

The mean is the default indicator of central tendency for interval scale data, however, there are objections to its use with ordinal data.

One objection to using the mean is that it generally produces a non-integer value whereas the original ordinal scale consists of a series of discrete categories with integer values. The calculated mean does not correspond to any of these real-world values.

A more worrying problem arises from the fact that within ordinal scales the step sizes between the available scores are not necessarily of equal significance. On an ordinal scale some of the steps might be viewed as considerably more important than others. For example in a treatment trial we could have an ordinal scale of outcomes where: 5 = Great improvement; 4 = Moderate improvement; 3 = Unchanged; 2 = Deteriorated; 1 = Died; and a comparative trial might then give the results in Table 3.6.

The mean scores for the control and active groups are respectively 3.49 and 3.66 which might suggest some slight superiority for the active treatment. However there is a strong case for querying this apparent superiority. The ordinal scale allows the same

Table 3.6 Outcomes from a comparative trial – Pitfalls with using the mean to describe ordinal data

Outcome	Control	Active treatment
1 – Died	0	4
2 – Deteriorated	8	7
3 – Unchanged	11	7
4 – Moderate improvement	19	8
5 – Great improvement	5	18

one point difference between Died and Deteriorated and between Moderate and Great improvement. It is not something we can ever measure objectively, but most patients would probably see the difference between Died and Deteriorated as far more important than that between Moderate and Great improvement. So, the active treatment may generate a higher mean score, but this fails properly to reflect the importance of those four fatalities seen with the active treatment. Is the active treatment really better?

3.5.2 The median will not always work either – Treating post-operative pain

In the next case we compare two treatments for post-operative pain. We use a five point scale where higher scores represent increasing levels of pain. The results are shown in Table 3.7 and Figure 3.8.

The best starting place is usually a bar chart of the responses. Figure 3.8 shows that treatment one is producing less cases with high pain scores (levels four or five) and is therefore the more effective.

We have seen objections to using the mean to summarise ordinal data so perhaps we could use the median instead? However, Figure 3.9 shows a problem that can quite easily arise. There are 25 subjects in each group, so to calculate the median, we need to identify the 13th highest ranked individual in each group. Unfortunately, for both treatment groups, this 13th ranked individual falls within the category of pain level three (Mild pain) and thus the median is the same for both groups. The median fails to illustrate the lower pain levels with treatment one. In general, the median is probably the best indicator of central tendency for ordinal data, but this example shows that it is not a universal remedy.

3.5.3 The mode too unstable for general use

One final alternative might be to quote the modes for the two groups, which (from Figure 3.8) are three and four. The modes reflect the difference between the treatments quite nicely in this case. However, it would be impossible to give a general recommendation for use of the modes. A particular problem is that modes are very unstable, that is if we repeated this relatively small trial, it is quite possible that one or other group would produce a different mode.

Table 3.7 Ordinal measures of post-operative pain under two treatment regimes

	Treatment one	Treatment two
1 – None	3	1
2 – Virtually none	8	5
3 – Mild	11	7
4 – Moderate	3	10
5 – Severe	0	2

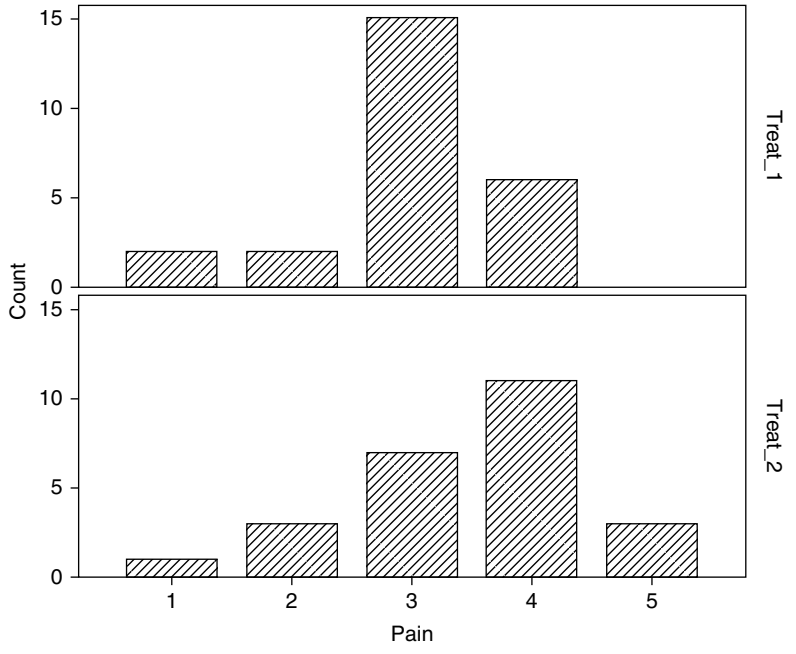


Figure 3.8 Outcomes with two treatments for post-operative pain

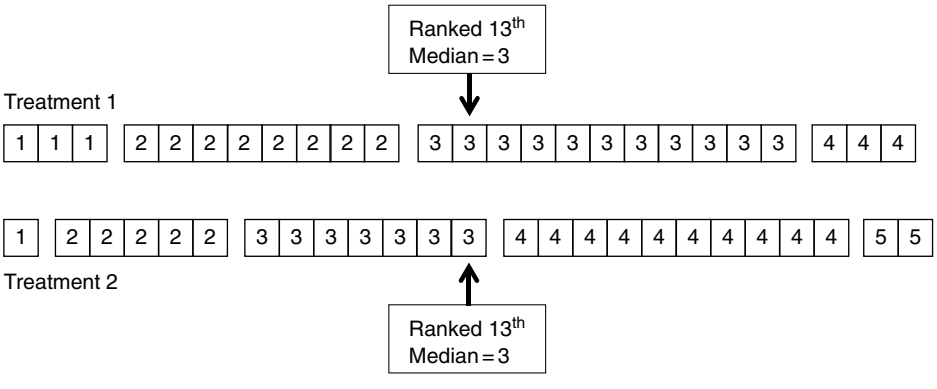


Figure 3.9 Calculating the median pain level in two treatment groups. Arrows indicate the 13th most highly ranked individual for both groups – the median values

3.5.4 Dispersion in ordinal data

Bar charts will give a visual impression of the variability in a set of ordinal data. The only statistic that can really be quoted as a measure of dispersal is the interquartile range.

3.5.5 In summary – How can we describe ordinal data?

There is no single universal answer. Probably the best solution is to focus on simple or stacked bar charts such as Figures 2.4 and 3.8. These usually reflect the overall pattern without producing false impressions or hiding genuine features.

We may then be able to summarise the data with an indicator of central tendency. However, every data set needs to be considered individually and the choice of an appropriate indicator will vary from one data set to another. The mean is widely frowned upon, but in some cases may be the only way forward. The median is the best starting point, but is certainly not an automatic choice in all cases.

The interquartile range can be used to express data variability among ordinal data.

3.6 Using computer packages to generate descriptive statistics

3.6.1 Excel

The one thing at which Excel does not excel is statistical analysis. The (very basic) level of material covered in this chapter is about at (if not beyond) the limit of its capabilities. It can be used to generate means, medians, SDs and quartiles, but while the first three are OK, the quartile values generated are somewhat unconventional and will not be pursued further. The mean, median and SD of a data set can be generated by using either worksheet functions or the Data Analysis tool.

To use worksheet functions, enter the tablet erythromycin contents from Table 3.4 into the first two columns of an Excel spread-sheet (assume titles in A1 and B1 and the two sets of data in A2:A11 and B2:B11), and then enter the following formulae into A12:A14 and B12:B14:

A12 =Average(A2:A11)

A13 =StDev(A2:A11)

A14 =Median(A2:A11)

B12 =Average(B2:B11)

B13 =StDev(B2:B11)

B14 =Median(B2:B11)

Notice that the formulae must be preceded by an equals sign, as shown. The appropriate means, medians and standard deviations will be displayed. (You might also want to try setting up a spread-sheet to perform the full calculations for the SD as shown in Table 3.4, just to reinforce exactly how this is derived.)

To use the Data Analysis tool, enter the data as above and then proceed through the menus Tools then Data Analysis then select Descriptive Statistics. In the box labelled 'Input Range', enter A2:B11 and tick the box for 'Summary Statistics'. The mean, median and standard deviation will be shown for both data sets, but you will probably need to widen the columns to make the output clear.

When you open the Tools menu, Data Analysis may not appear on the menu, in which case it needs to be installed. To do this, go back to the Tools menu and select Add-Ins … and tick the box for Analysis ToolPak.

3.6.2 Statistical packages

3.6.2.1 General approach in this book The website associated with this book (www.ljmu.ac.uk/pbs/rowestats/) gives detailed instructions for generating means, medians, standard deviations and quartiles using Minitab and SPSS. For all statistical procedures, the book will provide general instructions that will certainly be appropriate for SPSS and Minitab, but also for most other packages. These will cover those aspects that are fairly standard:

- The pattern in which data should be entered.
- The information that will need to be entered in order to run the routine.
- What information to look for in the output.

The one thing that is completely variable between packages is navigation through menu structures to trigger a particular routine, so this is not covered here.

To illustrate what to look for, generic output is provided. This is not formatted in the same manner as that from any specific package, but contains details that all packages ought to generate.

3.6.2.2 Obtaining descriptive statistics for the erythromycin data The data are generally entered into a column. Once the menu structure has been navigated to

Table 3.8 Generic output of descriptive statistics for the erythromycin data shown in Table 2.4

Descriptive statistics: Erythromycin							
	n	Mean	SD	SEM	Median	Q1	Q3
Alpha	10	248.7	8.72	2.76	249.00	241.50	255.50
Bravo	10	251.1	3.78	1.20	251.00	249.25	254.25

select the Descriptive Statistics routine, the only additional information to be supplied is which column(s) contain the relevant data. Table 3.8 shows generic output.

The number of observations (*n*) and the statistics referred to in this chapter should be fairly obvious. The second quartile (Q2) is generally not shown as it is the same as the median. The Standard Error of the Mean (SEM) may not be familiar at the moment, but is an important statistic that will be described in Chapter 5.

3.7 Chapter summary

When choosing a descriptive statistic we need to be aware of whether the data contains extreme outlying values and whether it contains a single cluster (unimodal) or several (polymodal) We should also think about what use is intended for the statistic we quote.

The mean, median and mode are all indicators of central tendency and tell us about the magnitude of the figures within our data.

- The mean is synonymous with the average.
- The median is the middle ranking value.
- The mode (or modes) is/are the most frequently occurring value either overall or within each cluster of values.

For most purposes there is a pecking order with the mean being most useful and therefore the prime candidate, then the median and finally the mode (or modes) only used in desperation. So long as the data is unimodal and not unduly affected by extreme values, the mean is the obvious choice. Where the data is badly affected by outliers, the median may provide a better impression of a typical value. Only the modes can properly describe polymodal data.

The Standard Deviation (SD) is an indicator of dispersion. It tells us about the variability among the figures within our data. The Coefficient of Variation describes relative variability by expressing the SD as a ratio to the mean.

The three quartile values indicate the figures that appear 25%, 50% and 75% of the way up the list of data when it has been ranked. The second quartile is synonymous

with the median and can act as an indicator of central tendency. The interquartile range (difference between first and third quartile) is an indicator of dispersion. The median and interquartile range are 'Robust' statistics which means that they are more resistant to the effects of occasional extreme values than the mean and SD.

With ordinal data, bar charts need to be used to give a general impression of the central tendency and dispersion of the results. The first choices of statistics to represent such data are the median and interquartile range. However, each data set needs to be considered individually and descriptors chosen to ensure that they successfully reflect the essential features of the data set.

Detailed instructions are provided for the calculation of the mean, median and SD (But not quartiles) using Microsoft Excel. Readers are referred to www.ljmu.ac.uk/pbs/rowestats for detailed instructions on generating all these descriptive statistics (including quartiles) using Minitab or SPSS. Generalised instructions that should be relevant to most statistical packages are provided in this book.

4

The normal distribution

This chapter will ...

- Describe the normal distribution.
- Suggest visual methods for detecting data that does not follow a normal distribution.
- Describe the proportion of individual values that should fall within specified ranges.
- Explain the terms 'Skewness' and 'Kurtosis'.

4.1 What is a normal distribution?

Many of the things we measure show a characteristic distribution, with the bulk of individuals clustered around the group mean and then cases become steadily rarer as we move further away from the mean.

In the previous chapter, Figure 3.5 showed a sample of 500 tablets from an 'Alpha' tableting machine. The bulk of the tablets were clustered in the central region with between 240 and 260 mg erythromycin. Out at the extremes (less than 230 or more than 270 mg), cases were very rare. Five hundred is quite a large sample and yet the histogram was still a long way from a perfectly smooth shape. With larger and larger samples we would gradually move towards a graph that

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

followed the superimposed curve. The idealised curve in Figure 3.5 is a normal distribution. It looks like a cross-section through a bell.

When a curve is superimposed on the data from the Bravo machine (Figure 3.6), it has different proportions (taller and narrower) because these tablets are less variable and cluster more tightly around the mean. However, both sets of data follow normal distributions.

Normal distributions can vary in appearance

Normal distributions may be squat and broad if there is a large Standard Deviation (SD) or tall and thin with a small SD.

4.2 Identifying data that are not normally distributed

4.2.1 Does it matter?

Many of the statistical techniques that we are going to look at in later chapters only work properly if the data being analysed follow a normal distribution. The requirement is more than just a statistical nicety. In some cases, we can get quite ludicrous results by applying methods that assume normality to severely non-normal data. An unfortunate side-effect of the very term 'Normal' distribution is a subliminal suggestion that such distributions are what we would normally expect to obtain and anything else is, in some way, abnormal. That sadly is not the case.

4.2.2 How to spot non-normal data

There are a number of statistical tests that will allegedly help in deciding whether data deviate from a normal distribution. However, in the author's experience the results of such tests tend to be difficult to interpret in real life situations. Unfortunately an exposition of the problems of testing for a normal distribution requires knowledge of statistical significance and power and necessary sample size that are not fully covered until Chapter 10. That material has therefore been placed in an appendix to this chapter. Once you have read Chapter 10 you should be able to understand the material in the appendix.

Rather than using one of these dubious statistical tests, what will be suggested is a simple visual inspection of histograms of data and specific features that indicate non-normality have been highlighted. There are three visual characteristics that any true normal distribution will possess. Strictly speaking, possession of all three of these does not guarantee normality, but it is rare to find data sets that fit all three

criteria and still manage to be sufficiently non-normal to cause serious practical difficulties. The three characteristics are:

- The data are unimodal.
- The distribution is symmetrical.
- The frequencies decline steadily as we move towards higher or lower values, without any sudden, sharp cut-off.

By way of illustration, the three cases below show infractions of each of the criteria in turn. You do need to be aware that unless samples are very large, histograms will not exactly follow the classical, pleasing to the eye bell curve. What you need to check for are obvious and gross deviations from a normal distribution.

4.2.2.1 Problem 1. Data with a polymodal distribution If we give a series of patients a standard oral dose of the anti-tuberculosis drug isoniazid, obtain a blood sample from each individual six hours later and determine the isoniazid concentrations of those samples, the results will probably look like Figure 4.1. The data are bimodal, because the metabolism of isoniazid is genetically controlled and we all fall into one of two groups – fast or slow metabolisers. The fast metabolisers form the cluster at the low end of the concentration scale and the slow metabolisers form a distinct group with higher levels.

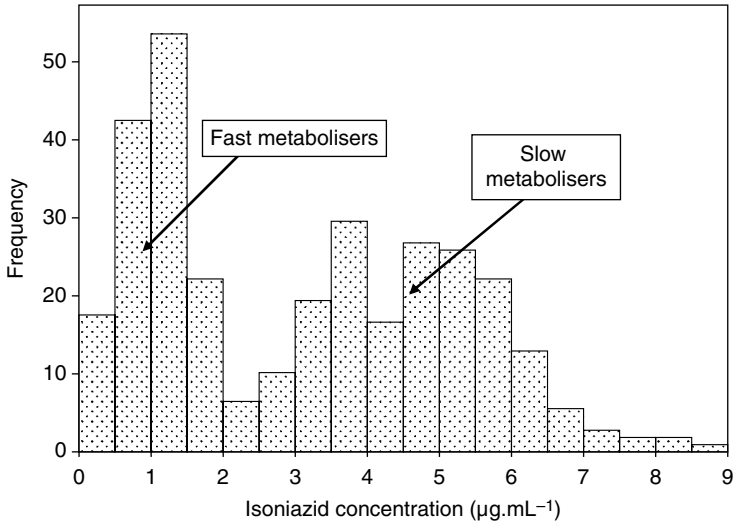


Figure 4.1 Bimodal data. Isoniazid concentrations (µg.mL^{-1}) six hours after a standard oral dose

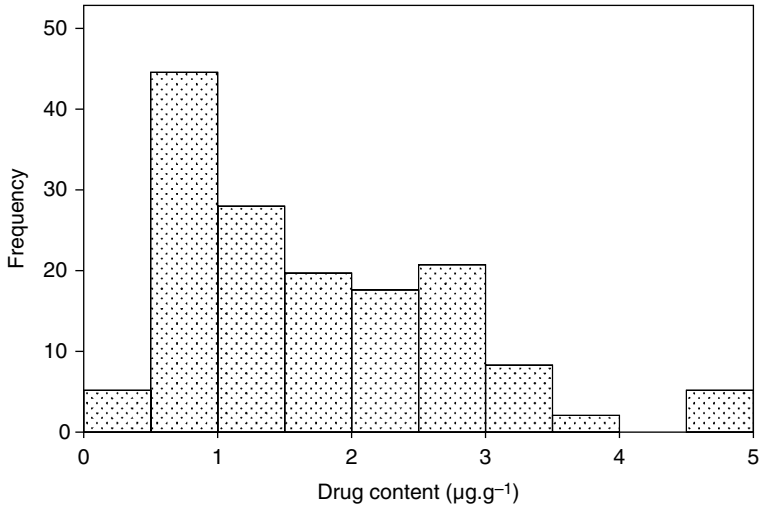


Figure 4.2 Skewed data. Drug content ($mg.g^{-1}$ of dried plant tissue) of a series of individual plants

Any attempt to apply statistical methods that rely upon a normal distribution to data like this, is likely to end in nonsensical conclusions.

4.2.2.2 Problem 2. Data that are severely skewed A candidate anti-cancer drug is being extracted from plant leaves. Specimens of the plant have been obtained from a wide range of geographical regions and each plant has been analysed for its content of the drug. Figure 4.2 shows the results. The pattern is far from symmetrical. Instead we have a long tail of results on one side, not balanced on the other. This is referred to as a 'Skewed' distribution. Again, any attempt to analyse such data using statistical methods that assume a normal distribution is likely to lead to real practical difficulties.

The skewing seen in Figure 4.2 is referred to as 'Positive skew', because the long tail extends towards high values. It is also possible to encounter Negative skew. Figure 4.3 shows generalised examples of both types of skew.

Skew often arises when results are constrained in one direction, but not the other. The drug concentrations for example (Figure 4.2) cannot be less than zero, but can extend virtually without limit to the right. Practical experience shows that positive skew is quite common in data from biological and medical research – negative skew less so. This is largely because many endpoints cannot fall below zero, but it is rarer for values to push up against some upper limit.

4.2.2.3 Problem 3. Data that are sharply truncated above and/or below the mean Patients report the amount of disability they are suffering as a result of gout. A 'Visual analogue scale' is used to gather the information. A scale is printed with

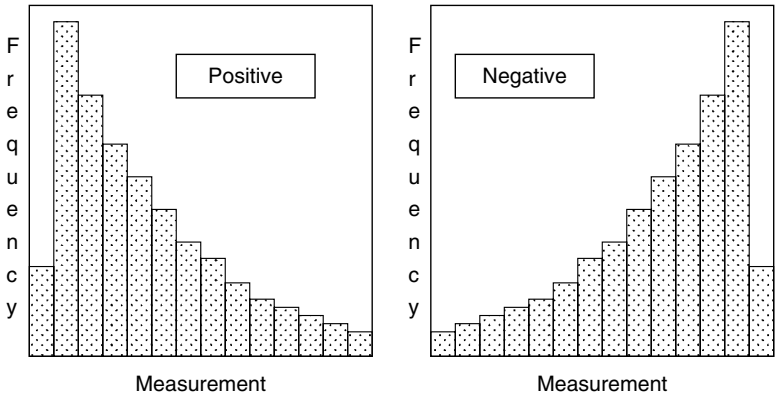


Figure 4.3 Positive and negative skew

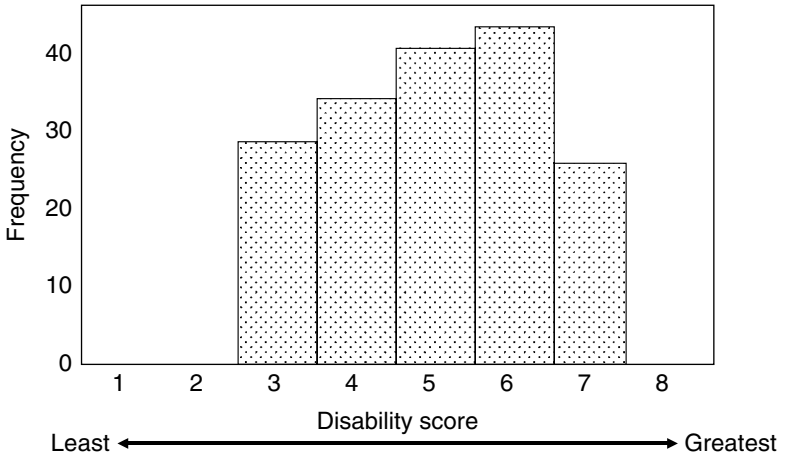


Figure 4.4 Data with sudden cut offs. Patients' self assessment scores for degree of disability due to gout

labels along its length reading 'Minimal disability', 'Moderate disability' and 'Severe disability'. Patients make a mark on the scale at a point that they feel described their situation. The scale is then divided into eight equal lengths and each patient's mark on the scale converted to a score of 1 to 8, with 8 being the greatest degree of disability. The final result is an ordinal scale of measurement. A histogram of the scores is shown in Figure 4.4.

For whatever reason, the patients have only used a limited, central part of the available scale. No patient has considered their condition to be at either extreme. For this to be a proper normal distribution, there should be gradual declines in frequencies below 3 and above 7. Instead we see these sudden, sharp cut-offs.

Ordinal scores on scales with a limited number of possible values are notorious for producing patterns that are wildly non-normal and we tend to avoid analysing such data using methods where a normal distribution is a pre-requisite.

Identifying data that is not normally distributed

With many statistical routines, we must avoid data sets that are markedly non-normally distributed. If data has all of the characteristics below, it is unlikely to be so non-normal as to cause any practical problems:

- Unimodal
- Symmetrical
- No sharp cut-offs at high or low values.

4.3 Proportions of individuals within 1SD or 2SD of the mean

Because the shape of the normal distribution is exactly mathematically defined, a predictable proportion of individuals falls within any given distance of the mean.

4.3.1 Approximately two-thirds lie within 1SD of the mean

For example, assume that, in a large group of subjects, the mean clearance of a drug is 10 $L.h^{-1}$ with a standard deviation of 1.5 $L.h^{-1}$. We can then define a range of clearances that fall within one SD of the mean. One SD below the mean is 8.5 $L.h^{-1}$ and one SD above is 11.5 $L.h^{-1}$.

It is a general property of normally distributed data that 68% of individuals fall within 1 SD of the mean. So, assuming that the clearance data follow a true normal distribution, 68% of individuals should fall within the range calculated above. With 68% in this central range, the remaining 32% must constitute the two tails of individuals with relatively high or low clearances. Because the normal distribution is symmetrical, this 32% is distributed as 16% in each of the two tails. This is shown in Figure 4.5. 68% may not seem like a particularly memorable proportion, but (rather conveniently) near enough two-thirds of individuals fall within 1 SD of the mean.

4.3.2 Approximately 95% lie within 2SDs of the mean

It is then possible to extend this principal and calculate the proportion of individuals who fall within any given number of SDs of the mean. In standard tables, you can look up how many people should fall within 1, 2, 3, 1.5 or any other number of SDs

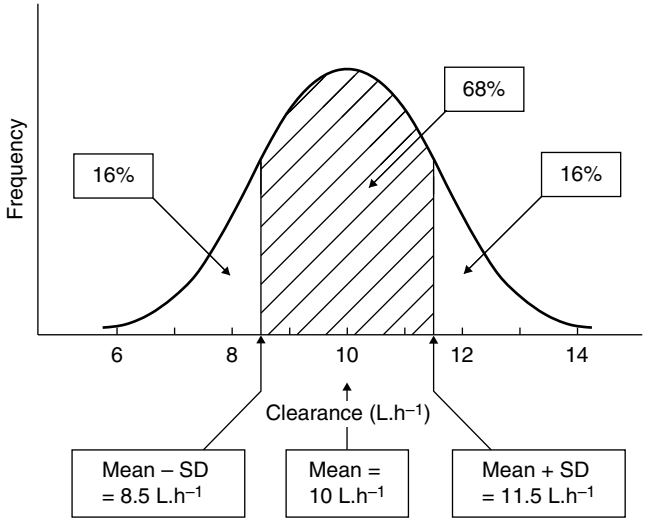


Figure 4.5 Approximately two-thirds of individuals with a drug clearance within 1 SD of the mean (mean = 10.0 ± 1.5 $L.h^{-1}$)

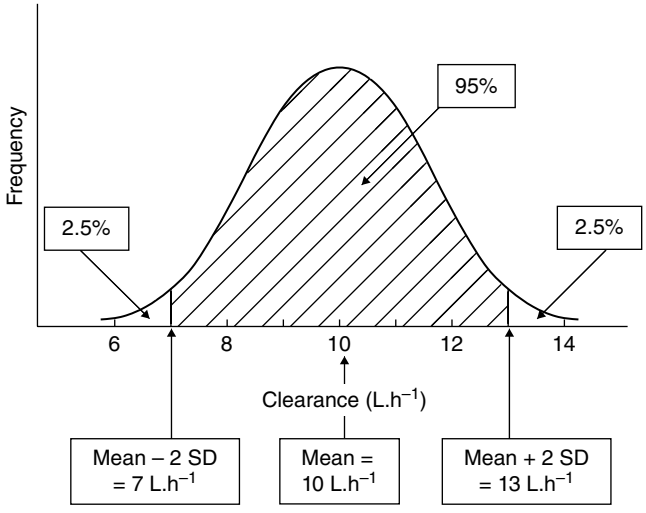


Figure 4.6 Approximately 95% of individuals with a drug clearance within 2 SDs of the mean (mean = 10.0 ± 1.5 $L.h^{-1}$). Often quoted as the 'Normal range'

of the mean. However, the only other case of special interest is that approximately 95% of individuals fall within the range ±2 SDs from the mean. Taking our clearance values, 2 SDs (3 $L.h^{-1}$) below the mean would be 7.0 $L.h^{-1}$ and 2 SDs above would be 13.0 $L.h^{-1}$. Figure 4.6 shows the 95% of individuals falling into the wider range ±2 SDs. The remaining (approximately 5%) of individuals are then found in the two small, extreme tails.

Proportions of individuals within given ranges

For data that follows a normal distribution:

- About two-thirds of individuals will have values within 1 SD of the mean.
- About 95% of individuals will have values within 2 SDs of the mean.

4.3.3 'Normal range'

For practical purposes people frequently find it convenient to identify a 'Normal range' for a particular parameter. For example, clinically, we may measure all sorts of ions and enzymes in a patient's blood and then we want to be able to spot any abnormal values that may be of importance in diagnosis and treatment. To make it easier to spot these interesting values, clinicians like to establish a Normal Range for each substance measured. Then, all they have to do is check for values outside the official Normal Range.

However, if the parameter is normally distributed in ordinary healthy people, then there is no nice clear cut point at which the values suddenly become 'Abnormal'. There will be perfectly healthy people with values out in the high and low tails of the distribution. A common approach is an arbitrary declaration that the middle 95% of values are 'Normal'. A normal range can then be arrived at by taking the mean plus or minus two SDs. Assuming the parameter to be normally distributed this will then include about 95% of individuals.

The problem of course, is that the remaining 5%, who may be perfectly healthy are suddenly condemned as 'Abnormal'. It is very easy to condemn such a system as arbitrary, however, in many cases there is no better system available and it is difficult to proceed without agreeing some sort of normal range.

'Normal ranges'

Normal ranges are frequently based on the mean plus or minus two SDs. These need to be treated with some scepticism, but are often the only pragmatic approach available.

4.4 Skewness and kurtosis

Deviations from a normal distribution were described visually in Section 4.2. You may meet references to a more formal method of describing non-normality based on two statistics called 'Skewness' and 'Kurtosis'.

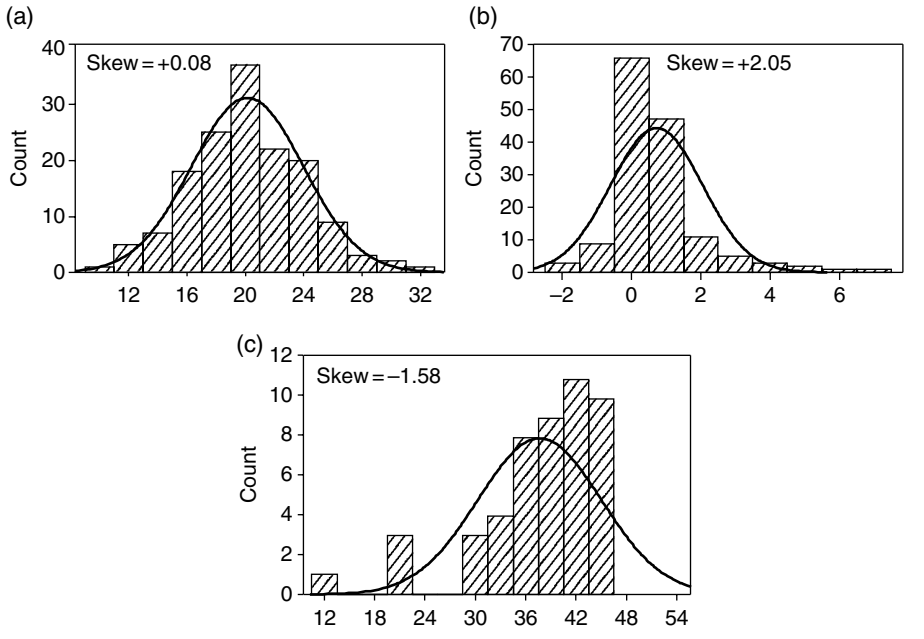


Figure 4.7 Histograms of data sets showing various levels of skewness: (a) non-skewed, (b) strong positive skew, (c) strong negative skew

4.4.1 Skewness

Skewness (Both positive and negative) has been mentioned in a general way (Section 4.2), but this did not attempt to describe how the degree of skewness of a data set can be reported as a numerical value. Skewness is reported on a scale where zero represents a perfectly symmetrical distribution and positive or negative values indicate positive or negative skew. There are no upper or lower limits to skewness values. Figures 4.7 (a) to (c) show histograms of various data sets. The ideal normal distribution (based upon the mean and SD of the particular data set) is indicated by the superimposed curve.

Real-world samples never adhere perfectly to the classic normal distribution, but Figure 4.7 (a) shows a distribution that is as symmetrical as you are likely to see (Skewness = +0.08). With Skewness of +2.05, part (b) shows obvious asymmetry with an elongated tail at the high end of the scale. Part (c) shows a clear case of the rarer negative skew (Skewness = –1.58).

4.4.1.1 Does skewness matter? Chapter 21 will emphasise that it is very dangerous to apply statistical procedures that require a normal distribution to badly skewed data; skewness can greatly reduce the ability to detect experimental effects.

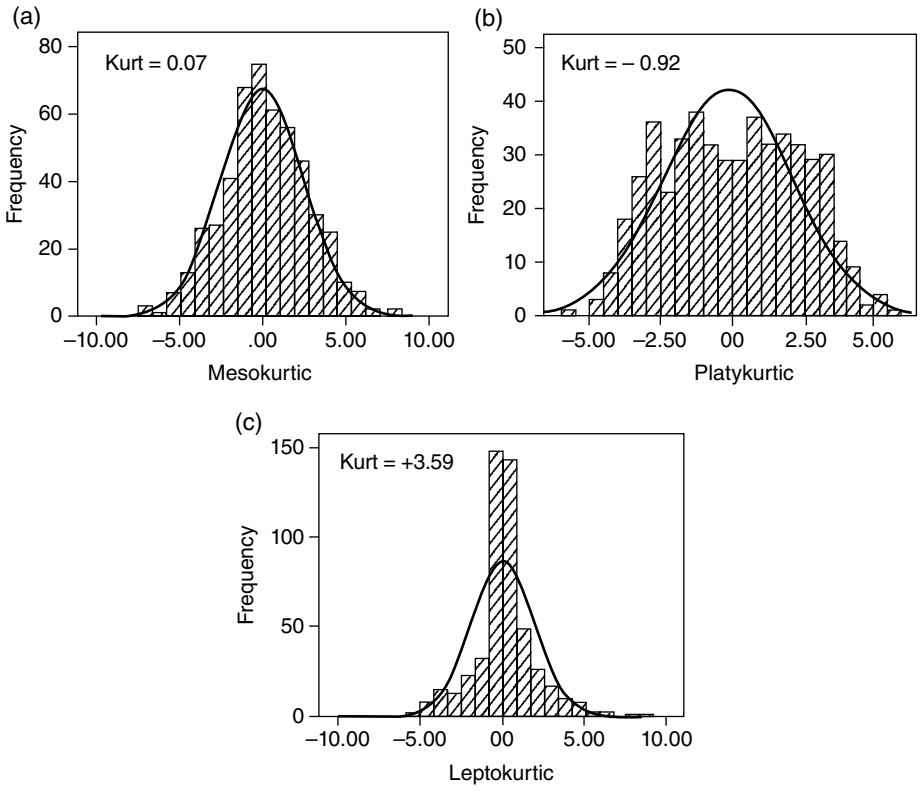


Figure 4.8 Histograms of data sets showing various levels of kurtosis: (a) normal distribution 'Mesokurtic', (b) negative kurtosis 'Platykurtic', (c) positive kurtosis 'Leptokurtic'

4.4.2 Kurtosis

Kurtosis describes the relative proportions of the peak, shoulders and tails of a distribution. It is most easily understood by looking at Figures 4.8 (a) to (c).

- Part (a) shows data from a true normal distribution with kurtosis close to zero (Kurtosis = 0.07). This is described as 'Mesokurtic'. The data follows the idealised normal distribution curve very closely.

- Part (b) shows negative kurtosis (–0.92). Here the shoulders are too high and wide to fit the ideal normal distribution. Alternatively you could say that the bars in the centre of the graph are too low and those out in the tails are also too low. The histogram looks as if we have taken hold of the two shoulders and stretched the graph laterally. Such graphs are sometimes described as 'Broad shouldered' or 'Short tailed'. The technical term is 'Platykurtic'.

Table 4.1 Characteristics of negative and positive kurtosis

	Shoulders	Peak	Tails
Negative kurtosis	High and wide	Low	Short
Positive kurtosis	Low and narrow	High	Long

- In part (c) we see positive kurtosis (+3.59). We now have the opposite pattern to (b). The shoulders are narrow and low, but the bars in the centre and out in the tails are higher than we would expect for a normal distribution. The graph now looks as if it has been squeezed inwards at the shoulders. The terms 'Narrow shouldered' or 'Long tailed' may be seen – technically 'Leptokurtic'.

The characteristics of negative and positive kurtosis are summarised in Table 4.1.

4.4.2.1 Do kurtosis problems matter? Generally speaking, deviations of kurtosis from the ideal, cause less damage than that arising from skewness. If data show strong negative kurtosis, the likelihood of successfully detecting experimental effects will be less than would be predicted from the SD of the data and tests are then described as 'Conservative'. Positive kurtosis has the opposite effect – increased risk of detecting an apparent effect (even if no such effect is genuinely present) – testing is overly 'Liberal'.

4.4.3 Using statistical packages to determine skewness and kurtosis

Most statistical packages can be used to obtain skewness and kurtosis values as part of their descriptive statistics routine, however in many packages (e.g. SPSS and Minitab) it is necessary to use options to change default settings so that these statistics are added. These statistics can also be obtained in MicroSoft XL using the formulae '=SKEW(*Cell range*)' or '=KURT(*Cell range*)'.

Skewness and Kurtosis

These statistics can be used to quantitate deviations from a normal distribution. Skewness describes the extent of any asymmetry and kurtosis describes any deviation from the ideal relative proportions of the peak, shoulders and tails of a normal distribution.

4.5 Chapter summary

Data that follow a normal distribution are clustered around the mean, with declining frequencies as we move to values further away. A graph of these frequencies looks like a cross-section through a bell. It is symmetrical.

The shape depends upon the variability within the data. With a low SD, the data are all clustered tightly around the mean and a histogram will be tall and thin. With more scattered data (higher SD), the distribution is low and wide.

Many statistical routines are liable to produce misleading results if applied to data that depart severely from a normal distribution. It is recommended that a check for gross non-normality should be made by producing a histogram of the data and checking that the distibution is unimodal, symmetrical and free from sharp cut-offs at either high or low values.

Data where the distribution is not symmetrical are described as 'Skewed'. In positive skew, there are outlying extreme values all (or most) of which are above the mean. In negative skew, the outliers are below the mean. Positive skew is quite common in biological and medical data.

In a normally distributed data set, about two-thirds of individual values will lie within one SD of the mean and about 95% within two SDs.

The 'Normal range' for a value is frequently equated with the range mean ±2 SDs. This is pragmatically useful, but the 5% of values outside this range can be overly simplistically interpreted as evidence that the individuals concerned are 'Abnormal.'

Skewness and kurtosis are statistics that may be quoted to quantitate any deviation from a normal distribution. Skewness reflects any lack of symmetry and kurtosis describes the relative proportions of the peaks, shoulders and tails of the distribution. For a true normal distribution, both these statistics should take a value of zero.

An appendix to this chapter explains the pitfalls of trying to use formal tests of whether data follows a normal distribution. Such tests are blighted by the twin perils of: (a) small data sets that provide too little power and so cannot detect even quite severe non-normality and (b) large samples that provide too much power and are liable to detect trivial deviations from normality.

4.6 Appendix: Power, sample size and the problem of attempting to test for a normal distribution

4.6.1 The nature of tests for normal distribution

In Section 4.2.2 it was pointed out that any consideration of the problem of formal tests for normal distribution required the use of concepts (power and necessary sample size) that were, at that stage, unfamiliar. However, once you have read as far as Chapter 10 you should be equipped to understand the problem.

For many statistical tests it is a pre-requisite that the data are consistent with a normal distribution (especially *t*-tests and analyses of variance). Various statistical tests are available which can allegedly be used to test for normality. These include the Shapiro–Wilk, Ryan–Joiner, Anderson–Darling and the wonderfully named Kolmogorov–Smirnov tests. However, Section 4.2 recommends a simple visual inspection of the data rather than any of these formal tests.

All the available tests have a null hypothesis that the data are normally distributed and a significant outcome is then evidence of non-normality. These tests are

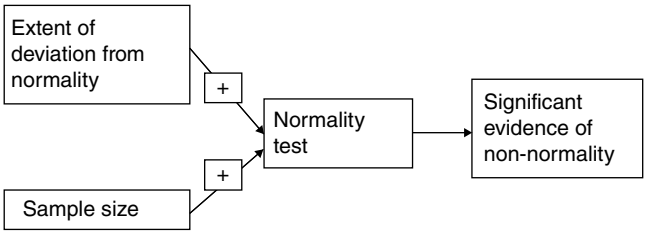


Figure 4.9 Aspects of the data that influence the outcome of a test for non-normal distribution

therefore not strictly speaking tests for normality; they test for non-normality. The likelihood of a significant result depends upon the extent of any non-normality that may be present and the amount of data (see Figure 4.9). Major deviations from normality may be detectable from quite small data sets but minor deviations will only achieve formal significance if large amounts of data are available.

4.6.2 How can these tests mislead us?

There are two obvious ways to arrive at inappropriate conclusions.

4.6.2.1 Failure to detect serious non-normality in a small data set Even gross non-normality may not produce a significant result if the data set is small. This would not matter too much if investigators could be relied upon to interpret non-significance properly. If they simply said 'There was no significant evidence one way or the other.' no great harm would be done. But in reality, a non-significant result tends to be taken as a licence to assume that the data are normally distributed and that *t*-tests and ANOVAs can therefore be relied upon.

4.6.2.2 Trivial non-normality within a large data set leading to a significant outcome ANOVAs and *t*-tests are pretty robust and moderate non-normality won't

Non-significant results

A non-significant result does not prove that the data is normal. The correct conclusion is 'We have not proved the existence of non-normality'. This may be because:

- The data is normally distributed.
- The data is non-normal but we have too few data points to be able to detect the non-normality.

The danger here is that you will seduced into using a *t*-test or ANOVA when you should not.

distort their outcomes to an extent that need be of practical concern. However, if the data set is large enough, even the most trivial departure from normality will be detected. Such an outcome probably shouldn't deter us from using a *t*-test and so on, but if you have gone to the trouble of carrying out such a test, are you just going to ignore the result?

Significant results

Chapter 10 emphasises the difference between statistical and practical significance. Very few data sets perfectly follow a normal distribution. If your data set is big enough you are likely to detect statistically significant 'Non-normality', but you need to ask whether the degree of non-normality is practically significant.

The danger here is that you will be deterred from using a *t*-test or ANOVA when it would be perfectly reasonable to do so.

4.6.3 Two example data sets

Set out below are two data sets that illustrate the twin perils of normality testing.

4.6.3.1 Markedly non-normal, but data set too small to achieve significance Table 4.2 sets out a very small data set. With so little data, the usual approach advocated – a histogram – is not going to be practical, but a dot plot (Figure 4.10) is shown.

There are too few data to say anything with certainty, but there is a very strong smell of positive skew. We have a cluster of low values and a couple of high ones. It is particularly noteworthy that the two highest values cover a wider range than the lowest five. The apparent skewness in these data could cause an appreciable loss of statistical power if they were included in a *t*-test or ANOVA. However, if we subjected this data to the Anderson–Darling test in Minitab the *P* value would be 0.107. These very dubious data fail to produce significant evidence of non-normality.

4.6.3.2 Large data set that produces a significant result despite being only marginally non-normal The next data set is very long (100 observations) and is therefore placed at the end of the appendix (Table 4.3).

Table 4.2 A very small data set that may suffer non-normality

Small data set
8.53
25.24
88.10
21.12
41.83
10.45
Mean 32.5 ± 29.7

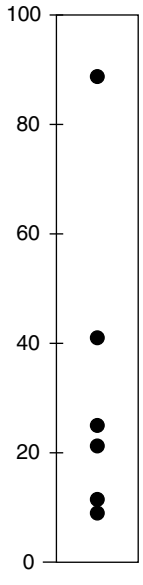


Figure 4.10 A very small data set with a strong suspicion of non-normality due to positive skew

Table 4.3 A large data set (n = 100) that deviates only very slightly from a normal distribution

Large data set				
10.22	10.92	11.49	10.86	10.49
10.96	9.13	7.66	11.15	12
9.35	9.92	10.96	10.49	12.88
12.36	10.59	11	10.81	10.4
12.84	11.91	10.97	10.19	10.79
8.44	15.56	8.66	10.46	11.62
10.93	11.31	9.12	8.96	10.32
13.64	9.5	8.3	10.84	9.52
9.36	10.33	9.83	8.23	10.27
9.26	10.19	9.64	9.37	9.34
9.89	12.08	10.11	9.35	6.19
7.47	9.81	14.54	8.75	11.14
8.59	10.05	9.41	6.87	9.13
11.21	10.33	11.79	14.71	9.96
8.8	12.56	12.05	8.65	11.09
14.2	9.31	7.17	10.39	9.46
9.35	12.27	10.29	9.52	9.45
11.55	10.38	10.8	9.8	9.12
9.93	8.93	8.19	12.23	8.68
8.96	8.24	10.25	11.94	9.67

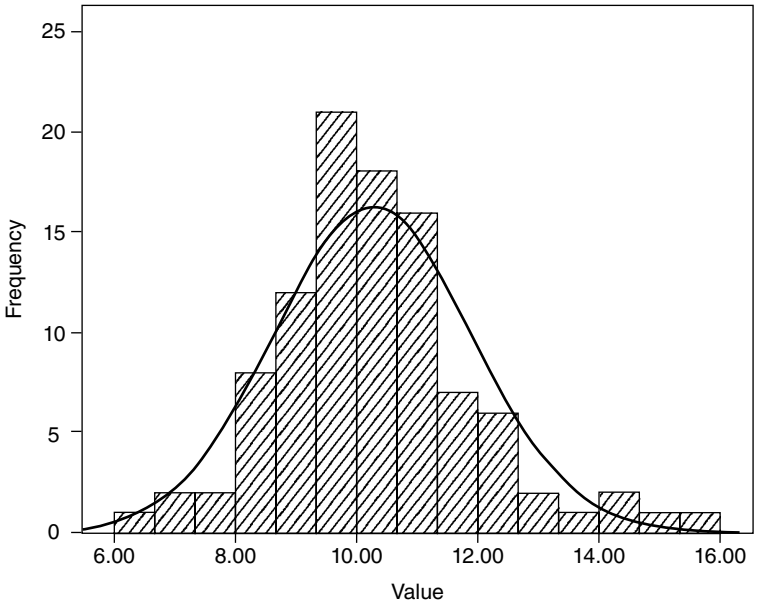


Figure 4.11 Histogram showing data that adheres very closely to a normal distribution, but which triggers a statistically significant finding of non-normality

Figure 4.11 shows that the data are about as close to normal distribution as you will ever see in the real world. For practical purposes, a *t*-test or ANOVA would cope perfectly satisfactorily with such a data set.

However, there is a very slight positive skew to the data – so slight that you would hardly even notice the handful of excess data points in the right-hand tail. Unfortunately, this marginal non-normality does get picked up as statistically significant due to the size of the data set (Anderson–Darling test $P = 0.028$).

So, it's Scylla and Charybdis. On one side, if you have too little data you may miss serious non-normality, but on the other, use too much and you may detect trivial non-normality. It is almost impossible to know when you are in the Goldilocks zone – not too much and not too little, but just right. Author's advice remains: Inspect a histogram and if the data looks as if it could be from a normal distribution, then treat it as such and robust procedures like *t*-tests and ANOVAs are unlikely to be very misleading.

5

Sampling from populations: The standard error of the mean

This chapter will ...

- Distinguish between samples and populations.
- Describe the way in which sample size and the SD jointly influence random sampling error.
- Show how the Standard Error of the Mean (SEM) can be used to indicate the quality of a sampling scheme.

5.1 Samples and populations

Statisticians carry on endlessly about two particular terms – 'Population' and 'Sample'. There is in fact good reason for the emphasis placed on these concepts and we need to be clear about the distinction.

5.1.1 Population

One of the hallmarks of good research is that there should be a clear definition of the target group of individuals (or objects) about whom we aim to draw a conclusion. The term 'Population' is used to cover all the individuals who fall

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

within that target group. The experimenter can define the target group as widely or narrowly as they see fit. Some experiments may be very general, with the intention that our conclusions should be applicable to the entire human race or all the cats or all the dogs on the planet. In other cases, the target may be defined much more narrowly, perhaps all the females aged 55–65 with moderate to severe rheumatoid arthritis, living in North America. The size of the population will vary accordingly. At one extreme (all humans) there might be 7000 million of them whereas the group with arthritis (defined above) might contain only a few million.

Population

The complete collection of individuals about whom we wish to draw some conclusion.

While the sizes of populations may vary, they all tend to be too large for us to be able to study them in their entirety. Of course it is possible to define a population so tightly that the numbers become manageable. Left-handed, red-headed, males aged 95–100, living in Inverness, Scotland might well constitute a population small enough for us to be able to study all its members. However, with all due respect to these sinistral, Scottish redheads, it is unlikely that anybody would be very interested in such a narrowly constituted group. As a general rule, any population that is worth studying will be too large to study.

5.1.2 Sample

Since it is usually impossible to study a whole population, real science is carried out on smaller samples randomly selected from the larger population. The sample is unlikely to be of any direct interest to the reader. However, it is of indirect interest – the hope is that we can use the sample as a tool to throw light on the nature of the general population.

Sample

A random selection of individuals from the population we wish to study.

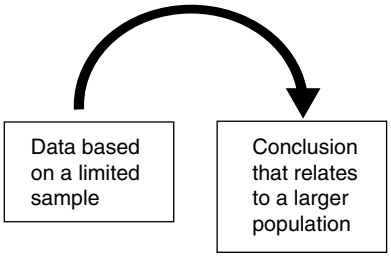


Figure 5.1 The great leap of faith required when drawing a conclusion about a population, based only on limited data from a smaller sample

5.2 From sample to population

However, there must be some doubts about the validity of using data that relates to one thing (a sample) to draw conclusions about something else (the population). Is this leap from sample to population justified? As Figure 5.1 implies, we need to maintain a healthy scepticism about such extrapolations.

If our sample is in fact not representative of the underlying population, then instead of throwing light on the situation, it will positively mislead us. So, we will next consider the various and nefarious ways in which a sample might misrepresent the population.

From sample to population

The purpose of a sample is to estimate properties of the broader population, such as its mean value, but we need to be alert to the possibility that a sample may misrepresent the population.

5.3 Types of sampling error

There are two distinct types of error that may arise – Bias and Random error.

5.3.1 Bias: Systematic error

It is all too easy to take samples which are virtually guaranteed to produce an average value that is predictably above (or predictably below) the true population mean. In some cases the problem is blindingly obvious and in others more subtle. A couple of examples follow:

- If we want to determine average alcohol consumption among the citizens of a given city, we might recruit our subjects by giving out questionnaires in a bar in

the town centre. The average at which we would arrive is pretty obviously going to be higher than the true average for all citizens of the city – very few of the town's teetotallers will be present in the bar.

- A drug causes raised blood pressure as a side-effect in some users and we want to see how large this effect is. We might recruit patients who have been using the drug for a year and measure their blood pressures. In this case the bias is less blindingly obvious, but equally inevitable. We will almost certainly understate the hypertensive effect of the drug as all the most severely effected patients will have been forced to withdraw from using the drug. We will be left with a select group who are relatively unaffected.

Bias or Systematic error

A consistent form of mis-estimation of the mean. Either most such samples would over-estimate the value or most would under-estimate it.

The principal characteristics of bias are:

- If we were to repeat the same sampling procedure several times, we could pretty much guarantee that we would make an error in the same direction every time. With the drink survey, we would always tend to over-estimate consumption; and with the hypertension study, we would consistently under-estimate the side-effect.

- Bias arises from flaws in our experimental design.

- We can remove the bias by improving our experimental design (always assuming that we recognise the error of our ways!).

5.3.2 Random error

Now take an essentially sound experimental design. We want to determine the average elimination half-life of a new anti-diabetic drug in type II diabetics aged 50–75, within Western Europe. We recruit several hospitals scattered throughout Western Europe and they draw up a census of all the appropriately aged, type II diabetics under their care. From these lists we then randomly select potential subjects.

Although this design is unlikely to produce a biased estimate, it is still very unlikely that our sample will produce a mean half-life that exactly matches the true mean value for the whole population. All samples contain some individuals

that have above average values but their effect is more or less cancelled out by others with low values. However, the devil is in that expression 'more or less'. In our experiment on half-lives, it would be remarkable if the individuals with shorter than average half-lives exactly counterbalanced the effect of those with relatively long half-lives. Most real samples will have some residual over- or under-estimation.

But what we have just described is random error and there is an important difference from bias. There is no longer any way we could predict whether error would take the form of over or under estimation. Indeed, if we carried out such a survey repeatedly, we would probably suffer very similar numbers of over- and under-estimates.

Random error

Any given sample has an equal chance of under- or over-estimating the population mean value.

The characteristics of random error are:

- Over- and under-estimation are equally likely.
- Even the best designed experiments are subject to random error.
- It is impossible get rid of random error.

5.3.3 This rest of this book is concerned with random error only – not bias

For the rest of this book we will be concerned solely with the problem of random error. Bias just shouldn't be there – good experimental design should see it off. But random error is ever with us and the science of statistics is there to allow us to draw conclusions despite its presence. Statistics won't make random error disappear; it just allows us to live with it.

Life's certainties

In this world, nothing is certain but death, taxes and random sampling error (modified from Benjamin Franklin's original).

5.4 What factors control the extent of random sampling error when estimating a population mean?

To illustrate the points made in this section we will refer back to the two tableting machines introduced in Section 2.3. One machine (Alpha) produced more variable tablets than the other (Bravo). Based on Figures 3.5 and 3.6, we will assume that the true long-term average vitamin content of the tablets corresponds to their nominal content of 250 mg.

Two factors control the extent of random sampling error. One is almost universally recognised, but the other is less widely appreciated.

5.4.1 Sample size

If we take random samples of tablets made on an Alpha machine, the accuracy of the sample means will depend on how well any high and low values happen to cancel out. Two samples are shown in Table 5.1.

In both samples an odd outlying high value (marked **) has crept in. Because the first sample is small, the outlier has displaced the sample mean considerably above the true population mean. There is also an odd value in the second sample, but it is now only one observation among 12, and so the sample mean remains much closer to the true value.

Most of us recognise that the larger a sample is, the better it is likely to reflect the true underlying situation. However, it doesn't matter how big a sample is, we must always anticipate some random sampling error.

Table 5.1 Erythromycin contents (mg) for two random samples of tablets (both from an Alpha machine)

	Sample 1 ($n = 3$)	Sample 2 ($n = 12$)
	246	258
	253	249
	270**	258
		249
		270**
		253
		237
		246
		259
		248
		242
		258
Mean	256.33 ± 12.34	252.25 ± 8.93

There's a constant tension between the statistician who wants the most precise (i.e. largest) possible samples and the experimenter who wants to keep the experiment as quick and cheap (i.e. small) as possible. Real-world experimental design always ends up as a compromise between these conflicting priorities.

Sample size (Statistician's point of view)

Big samples: Good
Small samples: Bad

5.4.2 Variability within the data

The second factor that influences random error is more easily overlooked. Table 5.2 shows a typical result, if we take two random samples of tablets – one for an Alpha and the other for a Bravo machine. The secret is to remember that the tablets from the Alpha machine are more variable than the Bravo ones.

Because the tablets produced by the Alpha machine are so variable, an outlying value can be high or low enough to do quite a lot of damage. In the above sample from the Alpha machine there happens to be some very low values (e.g. 235 and 240 mg). Consequently our sample mean for this machine differs rather badly from the true population mean. (At 247.33 mg it is 2.67 mg below the assumed true population mean of 250 mg.) However, extreme values are rarer among tablets produced by the

Table 5.2 Erythromycin contents (mg) for two random samples of tablets (one from an Alpha and the other from a Bravo machine)

	Alpha	Bravo
	252	254
	240	246
	243	247
	243	251
	250	254
	242	247
	257	251
	253	250
	251	242
	246	250
	235	247
	256	248
Mean ±SD	247.33 ±6.85	248.92 ±3.45

more consistent Bravo machine and so it is no surprise that the sample is not spoiled by any similar cluster of high or low values. As a result, the sample mean for this machine remains closer to the true population mean. (At 248.92 mg, it is only 1.08 mg below the ideal figure.)

So, the other general rule is that samples means based on highly variable data are themselves rather variable and may provide a poor reflection of the true situation. With non-varying data, almost all samples will be pretty precise.

Variability in the data (Everybody's point of view)

Big SDs: Bad
Small SDs: Good

Unlike sample size, the SD is not primarily under the control of the experimenter. Tablets from the Alpha machine simply are rather variable and there's very little we can do about it. The only thing we can do is to make sure that we don't increase the variability within our data by adding unnecessary measurement error. If we use imprecise measurement techniques, the final SD among the data we collect will be even greater than the intrinsic variability of the parameter in question. In this particular case, it's bad enough that the tablets are so variable, but if we then used an imprecise analytical technique, the figures will end up being even more scattered.

5.5 Estimating likely sampling error – The SEM

We have identified the two relevant factors:

- Sample size
- SD within the data

and are now in a position to estimate likely sampling error for any given experimental situation.

At one extreme, if we take a small sample from a population of highly varying data, we must accept that our sample mean may be wildly misleading. In contrast, a large sample from non-varying data will produce a sample mean that is virtually guaranteed to be very close to the true population mean.

5.5.1 Standard error of the mean

Frustratingly, we cannot calculate exactly how much sampling error will be present in any individual sample. The error is random in nature and with any given sample we may over or underestimate the true situation, or (if it was an unusually lucky sample) we might be almost bang on. What we can estimate is a typical amount of sampling error that would occur with a given sampling regime. This is referred to as the Standard Error of the Mean (SEM). The purpose of the SEM is to provide an estimate of the likely sampling error that should be anticipated, given the size of the sample and the variability among the data being sampled. Thus the SEM is best viewed as a measure of the general quality of a sampling scheme, rather than necessarily telling us anything about a particular sample.

Standard Error of the Mean (SEM)

Reflects the quality of a sampling scheme based upon the sample size and the SD among the data being sampled. A high SEM tells us that the sampling scheme is weak – liable to generate large random sampling errors. A low SEM indicates a more precise scheme where random errors will be small.

5.5.2 The technical definition of SEM

A common approach in statistics is to ask 'What would happen if we were to repeat a sampling procedure many times?' In this case, the question we ask is 'What would happen if we were to take repeated samples and calculate the mean of each sample?' Fortunately, we don't actually have to take real repeated samples. We can calculate what would happen if we did, based on the fact that we know sampling error is dependent upon sample size and SD. An example of hypothetical repeated re-sampling is shown in Figure 5.2. Note that the horizontal axis represents the mean values of samples, not the individual values that go into the samples. The sample means mainly cluster around the true population mean, but there are a few outlying results badly above and below. These sample means themselves form a normal distribution. We can indicate the variability among these sample estimates in the usual way – we quote their SD. This SD among a large number of hypothetical sample means is then given the special name of the Standard Error of the Mean (SEM).

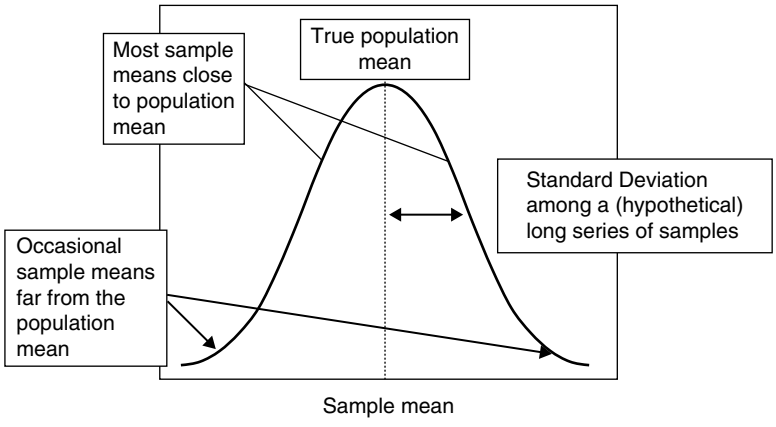


Figure 5.2 The SEM is the SD among hypothetical repeated sample means

Definition of the SEM

The SD among (hypothetical) repeated sample means from the same population.

5.5.3 The SEM in action

The SEM is shown in action in Figure 5.3. This contrasts what would happen if we were to take either small or large samples from the same population. In the first frame we would see small samples that are not very precise and the sample means would be badly spread out. Hence, the SEM would be large. With the larger samples (second frame), the means would be more tightly clustered and the SEM smaller. The SEM has thus achieved what we wanted. It has acted as an indicator of the quality of the two sampling schemes – large errors are quite likely with the smaller samples, but big samples are less error prone.

We could construct a diagram similar to Figure 5.3, to compare the situation with small and large SDs. In that case, the spread of sample means (and hence the SEM) would be greatest when there was a large SD.

5.5.4 Calculation of the SEM

The calculation of the SEM is based upon the SD and the sample size, as shown in Figure 5.4.

The form of the diagram used in Figure 5.4 will be used throughout this book. The idea is to avoid equations, while still identifying the aspects of the data that influence a particular outcome and also indicating the direction of their influence.

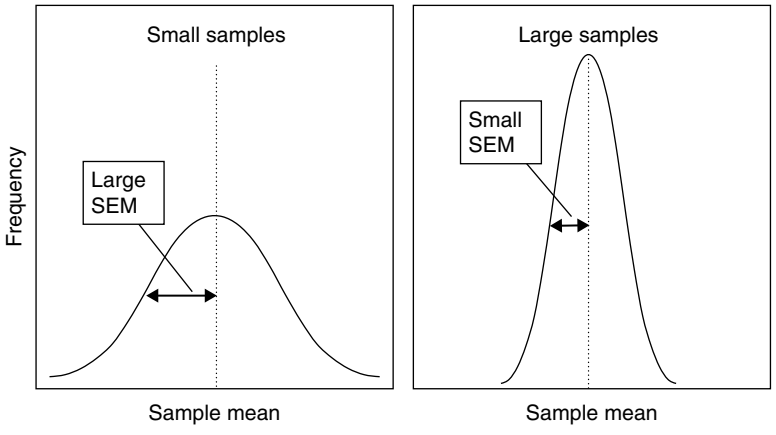


Figure 5.3 The SEM in action – small and large samples

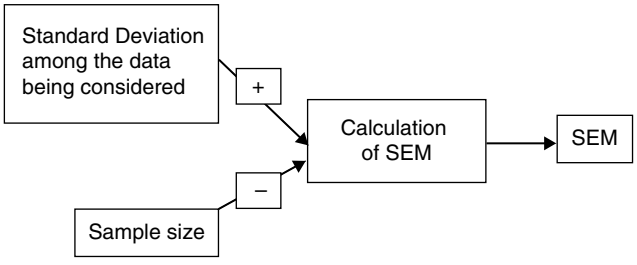


Figure 5.4 Calculating the SEM. Signs indicate whether increases in an input would increase (+) or decrease (–) the SEM

Thus, Figure 5.4 tells us that the aspects governing the SEM are the SD and the sample size. Then the plus sign indicates that the SD is positively related to the SEM. In other words the greater the SD, the greater the SEM will be. The minus sign indicates a negative relationship; the greater the sample size, the lower the SEM will be.

In this particular case the mathematical equation underlying the block diagram is so simple that we could have used it without causing excessive consternation to even the least numerate

$$\text{SEM} = \frac{\text{SD}}{\sqrt{n}}$$

$\sqrt{n}$ being the square root of the sample size.

While the actual equation is reasonably digestible in this case, many of the other relationships at which we will be looking are far more complex and the block diagrams are much more accessible. In fact, we don't need to worry about even this

equation, as stats packages invariably report the SEM as part of their descriptive statistics routines (see Table 3.6).

5.5.5 Obtaining SEMs for the samples of erythromycin contents in Tables 5.1 and 5.2

It would be no great hardship to calculate the SEMs manually, as below.

For Table 5.1 (Contrasting small versus large sample sizes):

For n = 3 : SEM = $12.34/\sqrt{3}$
= 12.34/1.73
= 7.13 mg

For n = 12 : SEM = $8.93/\sqrt{12}$
= 8.93/3.46
= 2.58 mg

For Table 5.2 (Contrasting more variable versus less variable data):

Alpha SEM = $6.85/\sqrt{12}$
= 6.85/3.46
= 1.98 mg

Bravo SEM = $3.45/\sqrt{12}$
= 3.45/3.46
= 0.996 mg

However, computers are quicker (and more reliable). Excel does not offer the SEM as a standard worksheet function, but it is included in the output from the Data Analysis tool (see Section 3.6.1). Both Minitab and SPSS include it in their Descriptive Statistics routines.

In Table 5.1 we see a contrast in sample size. The first sample contained only three observations and is therefore thoroughly unreliable; its SEM is given as 7.13 mg. The second sample is larger and more reliable and so its SEM is considerably lower (2.58 mg).

In Table 5.2 there is a contrast in data variability: The SEM for Alpha is 1.98 mg but for Bravo it is only 0.996 mg. The sample of tablets from the Alpha machine is liable to be distorted by somewhat outlying values and so the SEM is relatively high. In contrast, the mean based on the tablets from the Bravo machine will be more reliable as it is less likely to contain badly outlying values and the SEM is correspondingly lower.

5.6 Offsetting sample size against SD

Because sample quality is adversely affected when the SD is large, it is difficult to work with such data. However, it is not impossible. Figure 5.4 shows us that the SEM can be controlled to any required value by adjusting the sample size. Even if the data are very variable, a suitably low SEM can be achieved by increasing the sample size. In contrast, with data of low variability, it is possible to enjoy the luxury of small sample sizes and still obtain a respectable SEM.

Offsetting SD and sample size

Sample sizes can be adjusted to restrict the SEM to any pre-selected value, by taking account of the size of the SD.

5.7 Chapter summary

Scientific data generally consist of randomly selected samples from larger populations. The purpose of the sample is to estimate the mean and so on of the population.

A sample may mis-estimate the population mean as a result of bias or random sampling error. Bias is a predictable over- or under-estimation, arising from poor experimental design. Random error arises due to the unavoidable risk that any randomly selected sample may over-represent either low or high values. Random error is equally likely to result in under- or over-estimation.

The extent of random sampling error is governed by the sample size and the SD of the data. Small samples are subject to greater random error than large ones. Data with a high SD are subject to greater sampling error than that with low variability.

The SEM is used to indicate the quality of a sampling scheme by calculating the extent of random sampling error that would typically arise with a particular sampling scheme. It is calculated by taking account of the sample size and the SD. The technical definition of the SEM is that it is the SD that would be found among a hypothetical long series of sample means drawn from the relevant population. Examples are given where the SEM is greater for a small sample than a large one and also where it is greater for data with a high SD than where it is low.

6

95% Confidence interval for the mean and data transformation

This chapter will ...

- Describe the concept of a confidence interval (C.I.).
- Explain what is meant by '95% confidence'.
- Show how other levels of confidence could be used.
- Show that the width of a confidence interval is dependent upon sample size and SD and the level of confidence required.
- Describe one-sided C.I.s.
- Describe the C.I. for the difference between two means.
- Emphasise that data needs to adhere reasonably well to a normal distribution, if it is to be used to generate a 95% C.I. for the mean.
- Show how data transformations may be used to convert skewed data to a normal distribution.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

6.1 What is a confidence interval?

We have already established that a mean derived from a sample is unlikely to be a perfect estimate of the population mean. Since it is not possible to produce a single reliable value, a commonly used way forward is to quote a range within which we are reasonably confident the true population mean lies. Such a range is referred to as a 'Confidence interval' (C.I.).

The mean derived from the sample remains the best available estimate of the population mean and is referred to as the 'Point estimate'. We add and subtract a suitable amount to the point estimate to define upper and lower limits of an interval. We then state that the true population mean probably lies somewhere within the range we have now defined.

6.2 How wide should the interval be?

We will use some data where we have measured the quantity of imipramine (an antidepressant) in nominally 25 mg tablets. Nine tablets have been randomly selected from a large batch and found to contain the amounts shown in Table 6.1. We want to calculate a confidence interval for the population mean based on this sample.

The obvious question is how wide should the interval be? Clearly, the wider the interval is, the greater our confidence that it will include the true population mean. For example, interval (a) in Figure 6.1 only covers a range of about 0.2 mg. Since we know, from the SEM, that a sampling error of 0.3 mg is perfectly credible, we would have very little confidence that the true mean will fall within any such narrowly defined interval.

In contrast, interval (c) is so wide as to make it almost inconceivable that the population mean would not be included. While this high level of confidence is

Table 6.1 Imipramine content of nine randomly selected tablets

Imipramine (mg)
24.7
25.8
26.7
25.5
24.6
25.0
26.2
25.2
24.2
Mean 25.32 ± 0.81 mg (SEM = 0.27 mg)

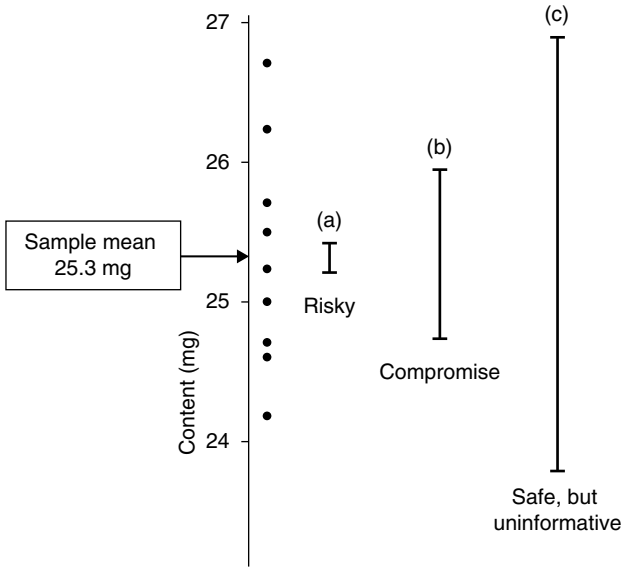


Figure 6.1 How wide a C.I. for mean imipramine content in tablets? See text for details about intervals (a), (b) and (c)

re-assuring, the price we have paid is that the interval is now singularly unhelpful. The information that the mean probably falls within a range which covers all of the individual data points, is hardly novel.

6.2.1 A compromise is needed

In reality we need to use an interval that is wide enough to provide reasonable confidence that it will include the true population mean, without being so wide as to cease to be informative [e.g. interval (b) in Figure 6.1]. The generally accepted compromise is the so-called '95% confidence interval'.

6.3 What do we mean by '95%' confidence?

The concept of being 95% confident that an interval includes the population mean is difficult to comprehend in terms of a single case. After all, an interval is either 100% right (it does include the population mean) or it's 100% wrong. It cannot be 95% correct! However, the concept makes perfectly good sense if viewed in terms of a long series. Imagine a scientist who regularly makes sets of measurements and then expresses the results as 95% C.I.s for the mean. The width of the intervals will be calculated so that 19 out every 20 will include the true population mean. In the remaining 20th case, an unusually unrepresentative sample generates an interval that fails to include it.

6.3.1 Other levels of confidence

It is possible to calculate other confidence intervals, for example 90% or 98% C.I.s. If we routinely used 90% C.I.s, we would have to accept being wrong on 10% of occasions, whereas with 98% C.I.s, we would be wrong only 2% of the time. But there is a downside to using higher levels of confidence, such as 98%; the intervals have to be wider to provide that extra level of assurance. In the real world, intervals for anything other than 95% confidence are not commonly met.

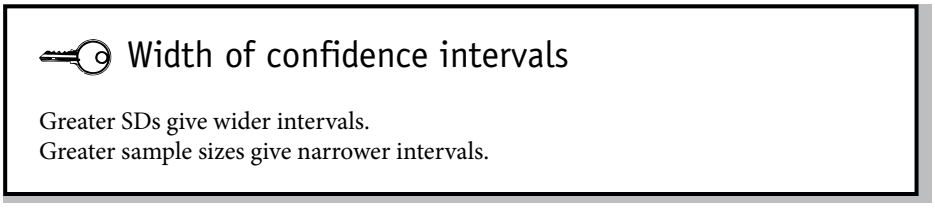
6.4 Calculating the interval width

We have already identified the factors that govern the reliability of a sample, that is Variability within the data and the sample size (Chapter 5). It is these same factors that influence the width of a 95% C.I.

Variability in the data: We know that high variability tends to make samples less reliable. So, if the SD is large, our sample will be relatively unreliable and the true population mean might be considerably higher or lower than our sample suggests. We will therefore have to set the interval relatively wide to ensure that we include the real population mean.

Sample size: With larger samples we will obtain estimates that are closer and closer to the truth, so we can afford to make the interval narrower.

Figure 6.2 summarises the situation. The plus sign indicates that an increase in SD will lead to a wider interval and the minus sign that increasing the sample size will reduce its width.



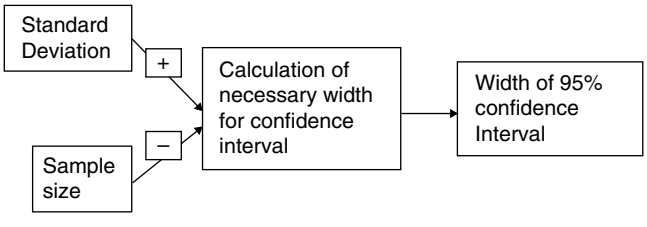


Figure 6.2 Calculation of the width of the 95% C.I. for the mean

Table 6.2 Generic output from calculation of 95% C.I. for mean imipramine content in tablets

95% C.I. for mean: Imipramine					
n	Mean	SD	SEM	95% C.I. lower limit	95% C.I. upper limit
9	25.32	0.807	0.269	24.70	25.94

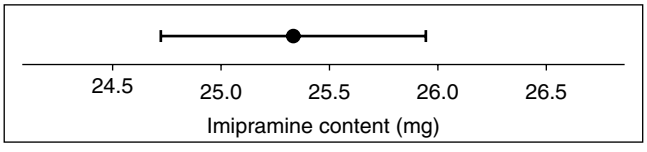


Figure 6.3 95% C.I. for the mean imipramine content among the population of tablets, based upon our sample

6.4.1 Using statistical packages to obtain 95% C.I.s

With most packages you will simply enter the data into a column, call up the appropriate routine and identify the column containing the data. You will then be supplied with a point estimate for the mean and the upper and lower limits of the confidence interval. Good old Excel and its Data Analysis tool half does the job; it provides a figure that needs to be added to/subtracted from the mean to obtain the limits of the interval.

For the imipramine tablets, Table 6.2 shows confidence limits of 24.70 and 25.94 mg.

So, we can state, with 95% confidence, that if we returned to this batch of tablets and took larger and larger samples, the mean imipramine content would eventually settle down to some figure no less than 24.70 mg and no greater than 25.94 mg. This can be conveniently presented visually as in Figure 6.3. The dot indicates the point estimate and the horizontal bar represents the extent of the 95% C.I.

6.5 A long series of samples and 95% C.I.s

Figure 6.4 shows simulated samples of nine tablets taken from a large batch, for which the true mean imipramine content is 25.0 ± 1.0 mg (±SD). Each horizontal bar represents the 95% C.I. from one of the samples. Out of the 30 samples, we would expect 95% to produce intervals that include the true population mean and the remainder (one or two cases) will be unusually misleading samples that lead to false C.I.s. That is pretty much what we see. There is just the one sample with a very low mean that produced a misleading outcome. Notice that even with the one interval that is technically incorrect, the true population mean is only marginally outside the interval. These C.I.s are almost never seriously misleading.

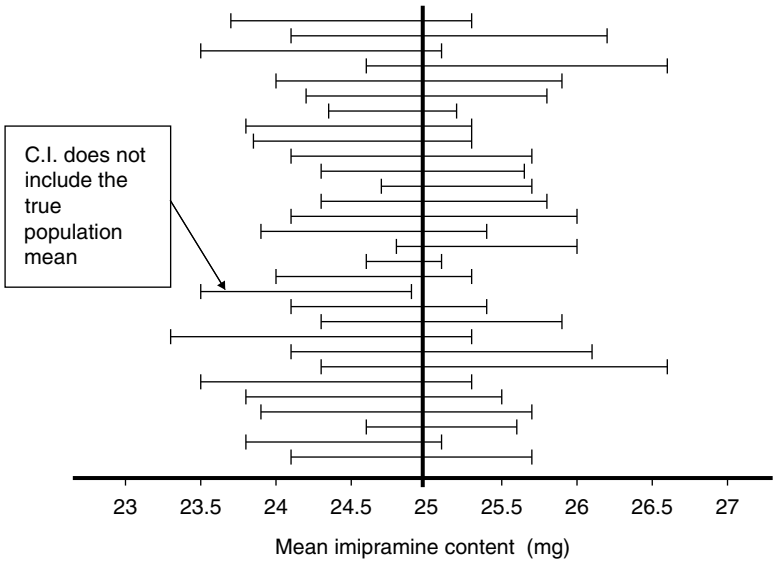


Figure 6.4 Simulation of repeated estimates of the 95% C.I. for mean imipramine content using 30 samples. Population mean is known to be 25.0 mg

95% C.I. is a good compromise

95% C.I.s have stood the test of time, providing a good compromise. They are narrow enough to be informative, but are almost never seriously misleading.

6.6 How sensitive is the width of the C.I. to changes in the SD, the sample size or the required level of confidence?

6.6.1 Changing the SD

Figure 6.5 shows what happens to the width of the C.I. if we change the SD of the data, but keep the sample size constant at nine. We had already anticipated that as the SD increases, the C.I. would have to be widened. The relationship is simply linear in nature. Doubling the SD doubles the width of the C.I.

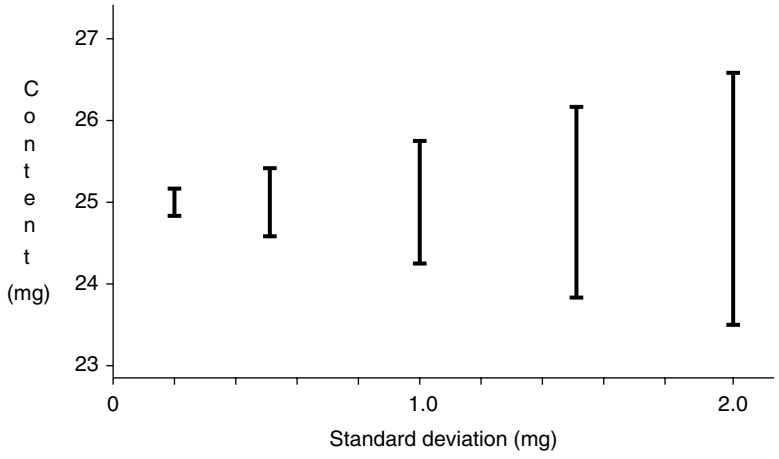


Figure 6.5 95% C.I.s for mean imipramine content. Population mean = 25 mg. Sample size = 9. SD varies between ±0.2 and ±2.0 mg

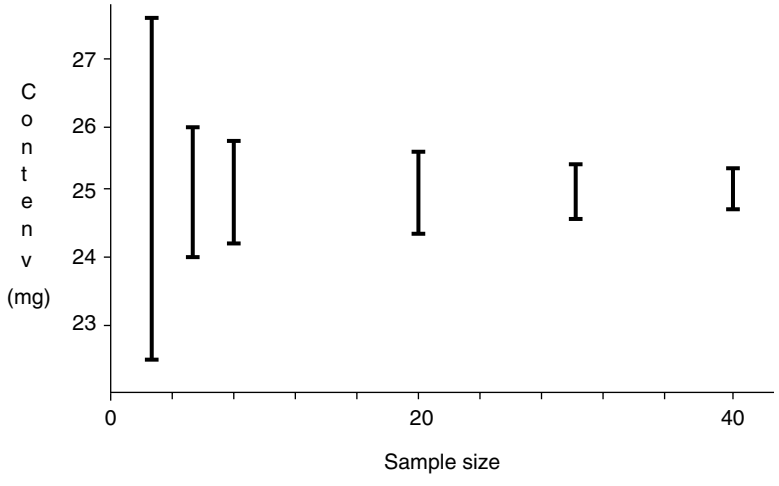


Figure 6.6 95% C.I.s for mean imipramine content. Population mean = 25 mg. SD = ±1.0 mg. Sample size varies between 3 and 40

6.6.2 Changing the sample size

Figure 6.6 looks at varying sample size while the SD is kept constant at ±1 mg. The relationship is dramatically sensitive at small sample sizes. With samples of less than about ten, any change in sample size greatly alters the width of theC.I. However, beyond ten, it is more of a struggle to reduce the width of the C.I.

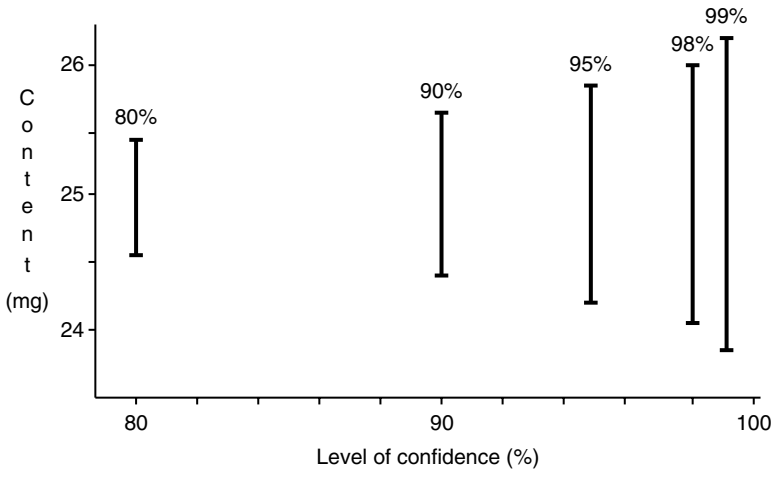


Figure 6.7 Confidence intervals for mean imipramine content. Population mean = 25 mg. SD = ±1.0 mg. Sample size = 9. Level of confidence varies between 80 and 99%

6.6.3 Changing the required level of confidence

Finally, Figure 6.7 shows the influence of the level of confidence required. The SD and sample size are kept constant, but the interval is calculated to provide anything between 80 and 99% confidence. The width of the interval does increase if greater levels of confidence are required, but the relationship is not dramatic. The usual standard is 95% confidence and any move up to 98 or down to 80% confidence, produces C.I.s that are not radically different.

The most striking conclusion from this section is that the biggest problems arise with very small samples. These lead to embarrassingly wide C.I.s and a relatively modest investment in a larger sample size would pay off handsomely.

The danger of small samples

C.I.s become wider with greater SDs, narrower with larger sample sizes and wider if higher levels of confidence are required. Most dramatically – Small samples give horribly wide 95% C.I.s.

6.7 Two statements

The solution arrived at in this chapter is basically that we abandon any attempt to draw a single definitive conclusion and instead we make two statements:

1. Some conclusion about the population.

2. An assessment of the reliability of the first statement.

Our conclusion was that the population mean lies within a stated range and then we can say that 95% of such statements are correct. This sets a pattern that we will see throughout statistics. All of the procedures at which we will be looking, include this crucial stage of providing an objective assessment of the reliability of any conclusion that might be drawn.

Assessing the reliability of our conclusions

Statistics never draws any absolute conclusions. It will offer an opinion and then back that up with a measure of that conclusion's reliability. It is a very grown up science, putting behind it the dubious certainties of childhood.

6.8 One-sided 95% C.I.s

The procedure we have followed so far actually entails making two claims as we have established both maximum and minimum limits for the mean. It has been emphasised that there is a 5% chance of error. That overall 5% hazard is split between two smaller risks – the true mean could be greater than the maximum we quoted or less than the minimum. We refer to such a procedure as 'Two-sided' or 'Two-tailed'.

Two-sided 95% confidence intervals – Two claims

1. The true population mean is no *less* than some stated figure. (2.5% chance that this is false.)

2. The true population mean is no *greater* than some stated figure. (2.5% chance that this is false.)

However, there are circumstances where only one of the above claims is actually of practical relevance. For example if we have manufactured a batch of crude drug for sale, the purchaser will demand some statement as to its minimum purity, but will hardly insist that we guarantee that it doesn't contain more than a particular amount of the authentic material. What we can do in such a case is to calculate what is initially a 90% C.I. for true drug content, which will generate two statements:

Drug content is no less than this figure (5% risk of error).

Drug content is no greater than this other figure (5% risk of error).

If we then only make the relevant claim and simply make no comment about possible values in the opposite direction, our total chance of error is only 5% and we again have 95% assurance that we are telling the truth. We then refer to this as a 'One-sided 95% C.I.'.

If we apply this approach to the drug purity problem, in a long series of instances there will be three types of outcome. The true drug content may be:

1. Below the minimum figure we quoted and we will have misled the customer and caused some annoyance. (5% of all cases.)

2. Within the two limits of the 90% C.I. that was initially calculated. Customer is OK about that. (90% of all cases.)

3. Above the upper limit of our initial C.I. We never made any claim as to maximum content and the customer certainly won't complain. (5% of all cases.)

Whilst initially it might seem perverse to calculate a 90% C.I. and then quote it as a 95% C.I., it is in fact perfectly fair so long as we quote only one of the confidence limits.

One-sided 95% C.I.s – One claim

Either:

The true population mean is no *less* than some stated figure (5% chance that this is false).

or:

The true population mean is no *greater* than some stated figure (5% chance that this is false).

6.8.1 Using statistical packages to obtain one-sided C.I.s

Most packages will produce one-sided intervals. There are two possible approaches:

- Find an option that allows you to select a one-sided interval in place of the default two-sided version. In this case you still request 95% confidence.
- Stay with the two-sided version, but change the confidence level to 90%. Then quote only the limit (maximum or minimum) that matches your requirements.

Using Excel, you would take the latter approach.

6.8.2 An example of a one-sided 95% C.I. – the purity of tetracycline

We have produced a batch of tetracycline and taken eight random samples of material from the batch. The results of analyses of the material are shown in Table 6.3. The potential purchaser requires a statement as to the minimum content of authentic material.

Using either of the approaches shown above should yield a lower limit for a one-sided 95% confidence interval of 75.7% tetracycline. This is the figure we can quote to a potential purchaser, knowing that there is only a 5% risk that the true content might be any lower.

6.8.3 Visual presentation of one-sided C.I.s

Figure 6.8 shows a useful way to present one-sided C.I.s. The idea is to emphasise the fact that we have established a definite lower limit, but are making no comment concerning how great the value might be. The figure also shows a normal two-sided 95% C.I. for the same data; it places limits both above and below the mean.

Table 6.3 Tetracyline content in eight samples taken from a single batch

Tetracyline (%w/w)
75.7
77.7
78.4
77.5
73.9
77.6
75.5
77.1

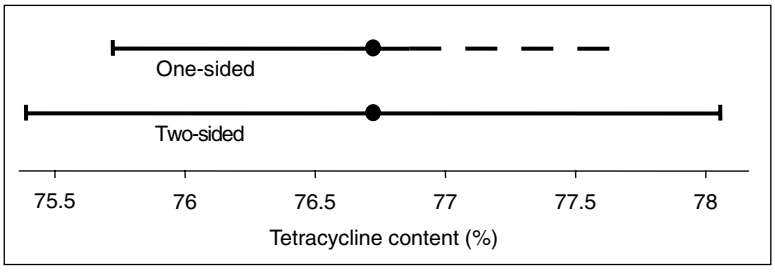


Figure 6.8 One- and two-sided 95% C.I.s for the mean tetracycline content for a batch of product (data in Table 6.3)

6.8.4 A one-sided confidence limit is closer to the mean

Figure 6.8 shows that when a one-sided C.I. is calculated, the one limit that is calculated is closer to the mean than the corresponding limit for a two-sided C.I. This is related to the amount of risk we are taking. With the one-sided C.I., we are prepared to take a 5% risk that the true mean might be below the indicated figure, but with the two-sided C.I., we can only allow a 2.5% chance of error in that direction.

One-sided C.I.s

One-sided C.I.s are a natural way to provide assurance that a value is no greater than (or less than) some critical value.

6.9 The 95% C.I. for the difference between two treatments

Experimental work often concerns comparisons between two differently treated groups of people (or animals or objects). Frequently, we will determine the mean value for some measured endpoint in each group and then look at the difference between these. But, our two mean values will almost certainly be based upon samples and each will be subject to the usual random sampling error. If we are estimating the difference between two imperfectly defined values, the calculated difference will also be subject to sampling error. There is an obvious role for a 95% C.I. in expressing the difference between two means.

Figure 6.9 shows a typical case, where we are comparing plasma cholesterol levels in two groups.

The error in estimating the difference will depend upon the imprecision of the two samples. If we have reliable estimates for mean cholesterol in both groups, then our estimate of the difference will also be reliable, but if the initial figures are imprecise, the calculated difference will be similarly woolly. We already know that the

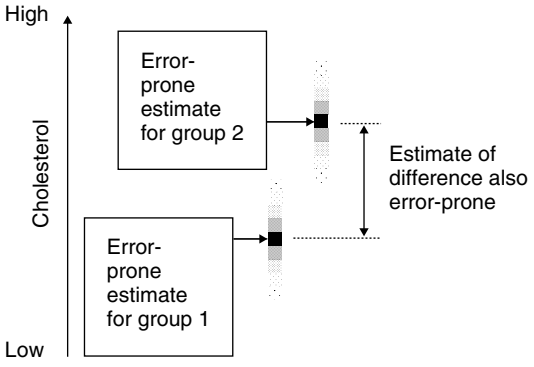


Figure 6.9 Uncertainty when estimating the difference between two sample means

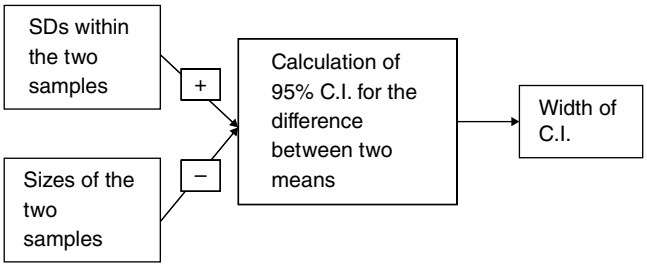


Figure 6.10 Calculation of the width of the 95% C.I. for the difference between two means

precision of the two samples depends upon their size and the variability of the data within each sample.

So, Figure 6.10 shows that the calculation of a 95% C.I. for the difference between two means will depend upon the SDs within the two samples (greater SDs bring more uncertainty and a wider interval) and the sizes of our samples (larger samples give greater precision and a narrower interval).

Any decent statistical package (not XL) will perform the calculation of such a confidence interval, but it will generally be presented as a so-called '*t*-test'. We won't consider the use of computers to produce this calculation any further at this juncture, but it lies at the heart of the next chapter.

Confidence interval for the difference between two means

A 95% C.I. for the difference between two means can be calculated by taking account of the size of the two samples and their SDs.

6.10 The need for data to follow a normal distribution and data transformation

Many of the statistical methods we will look at are based on an assumption that the data follows a normal distribution. The standard method of calculating a 95% C.I. is among these. It is generally accepted that if the distribution of data is approximately (but not perfectly) normal, the resulting 95% C.I. will not be unduly misleading. The term 'Robust' is used to express this ability to perform acceptably even with moderately imperfect data. However, when data is grossly non-normal, even robust methods can lead to silly conclusions. In this section we will look at a problem case and see how we can circumvent the difficulty.

6.10.1 Pesticide residues in foxglove leaves

Crops of foxglove leaves have been collected for the extraction of the cardiac drug digoxin. The crops were from widely scattered sites. All have been analysed for contamination by a pesticide. The results are shown in the first column of Table 6.4.

Table 6.4 Pesticide residue concentration in 20 crops of foxglove leaves

Residue (ng per g of leaf)	
Conc.	Log of conc.
31.9	1.503
89.8	1.953
31.8	1.502
105.6	2.024
8.5	0.929
2.1	0.322
1.5	0.176
94.4	1.975
2.1	0.322
2.7	0.431
21.4	1.330
12.1	1.083
5.7	0.756
267.6	2.428
88.4	1.946
7.4	0.869
1.7	0.230
1141.8	3.058
0.5	–0.301
1.9	0.279

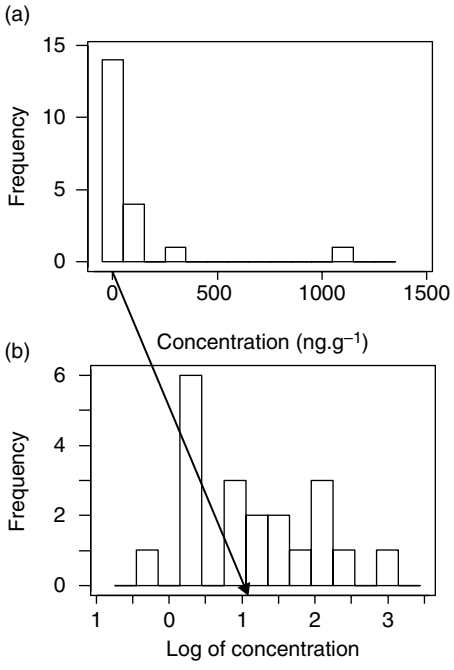


Figure 6.11 Log transformation of pesticide concentrations: (a) untransformed data, (b) log transformed data

Figure 6.11a shows a histogram of this data and they are obviously not remotely normally distributed. There is a very strong positive skew arising because a few sites have values far above the main cluster and these cannot be balanced by similarly low values, as these would be negative concentrations.

We could simply turn a blind eye to this non-normality and proceed to generate a 95% C.I. in the usual way and the results would be a mean of 95.9 ng.g^{-1} with confidence limits of –23.1 to +215.0 ng.g^{-1}. These are shown in Figure 6.12a. The result is clearly nonsensical – the true population mean could not possibly take the sort of negative value implied by the lower limit of the C.I.

Non-normal data

Data deviating dramatically from a normal distribution can produce severely misleading or even nonsensical confidence intervals.

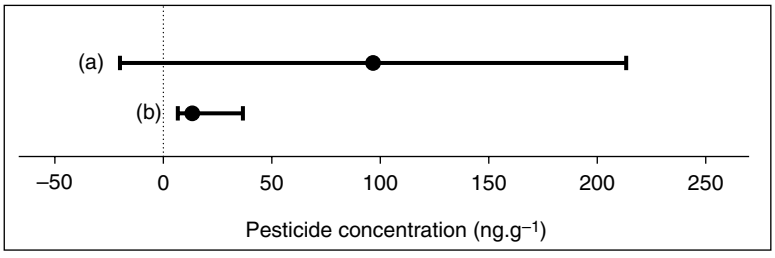


Figure 6.12 95% C.I. for mean pesticide content calculated (a) directly or (b) via log transformation

6.10.2 The log transform

A well established solution to the problem of strong positive skew is to transform the data by taking logarithms of the original values. The log transformed data are shown in the second column of Table 6.4. (The logs used in this case are simple base 10 logs.) Figure 6.11b shows that with log transformation we have a more symmetrically distributed data set. The reason for this is that the main cluster of values is centred around 10 $ng.g^{-1}$. With the untransformed data, this main cluster lies at the extreme left-hand side of a scale going from 0 to 1500 $ng.g^{-1}$. However, when we transform the data, a value of 10 is transformed to log (10) which equals 1. Such values are then close to the middle of a scale going from –1 to +3. The arrow on Figure 6.12 shows the main cluster of values moving to the centre of the distribution, thereby improving its symmetry.

We then calculate a mean and confidence interval using these log transformed values. The mean is 1.141, but as this was calculated from the logs of the data, we need to take the antilog to get the real mean. The mean is then Antilog (1.141) = 13.8 $ng.g^{-1}$. Similarly, the limits for the confidence interval (0.723 and 1.558) need to be transformed back to their antilogs (5.28 and 36.14 $ng.g^{-1}$). These are then shown in Figure 6.12b. The results now make sense with both limits being positive figures.

Log transform

The log transform will often convert badly positively skewed data to a reasonably normal distribution.

6.10.3 Arithmetic and geometric means

The mean value obtained varied according to the method of calculation. By the direct method it was 95.9 $ng.g^{-1}$ but using the log transformation it was only 13.8 $ng.g^{-1}$. We distinguish between these two values as being the 'Arithmetic' and 'Geometric' means respectively.

Arithmetic and geometric means

Arithmetic mean is calculated directly from the data.
Geometric mean is calculated via log transformation of the data.

With normally distributed data the arithmetic and geometric means would be equal, but with positively skewed data the arithmetic mean is always greater than the geometric mean.

6.10.4 Log transform gives asymmetrical limits

The 95% C.I. (Figure 6.12b) obtained via the log transform is markedly asymmetrical and this is characteristic of C.I.s obtained in this way.

6.10.5 Other transformations

A whole range of other transformations have been recommended for particular purposes. Two that are met with on occasions are described below.

6.10.5.1 Log transform with an added constant If the data contains zero or negative values, it is impossible to convert them to their logs. The usual solution is to add a fixed value to each data point. Values of 0.5 or 1.0 are commonly used. So, for example if the data contains some zero values, the transformation we used might be $\log (x + 0.5)$ where x is the original measured value.

6.10.5.2 Square-root transform Data that consists of counts tends to be positively skewed, unless the counts are very large. An example would be bacterial colonies on a growth medium. Technically, such data is said to follow a Poisson distribution. This can be converted to a normal distribution by taking the square-root of each count.

6.11 Chapter summary

A C.I. for the mean is derived from sample data and allows us to establish a range within which we may assert that the population mean is likely to lie. In the case of 95% C.I.s, such statements will be correct on 95% of occasions. In the remaining 5% of cases, particularly misleading samples will produce intervals that are either too high or too low and do not include the true population mean.

The width of the confidence interval will depend upon the SD for the sample, the size of the sample and the degree of confidence required. The width is especially dependent upon sample size – small samples leading to very wide intervals.

One-sided C.I.s can be used to specify a value that the population mean is unlikely to exceed (or be less than).

A 95% C.I. for the difference between two sample means can be calculated by taking account of the size of the two samples and the standard deviations within each sample.

Data that is to be used to calculate a C.I. for the mean should be at least approximately normally distributed. Severely non-normal data can lead to misleading intervals. Data that is markedly positively skewed can sometimes be restored to normality by log transformation thereby allowing the calculation of a geometric mean and 95% C.I.

7

The two-sample *t*-test (1): Introducing hypothesis tests

This chapter will ...

- Introduce hypothesis tests.
- Describe the use of the two-sample *t*-test to determine whether there is a real difference in a measured (interval scale) endpoint between two sets of observations.
- Describe null and alternative hypotheses.
- Describe the use of the two-sample *t*-test to generate a 95% C.I. for the difference between the two means.
- Show how the confidence interval is used to diagnose whether there is sufficient evidence of a difference.
- Set out the characteristics that data should fulfil if it is to be subjected to a two-sample *t*-test.
- Introduce the statistical use of the term 'Significance'.
- Explore the aspects of the data that will determine whether a significant result is obtained.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

- Introduce the synonymous terms 'False positive' and 'Type I error'.
- Show the use of 'Alpha' to report the risk of a false positive.

7.1 The two-sample *t*-test – an example of an hypothesis test

7.1.1 What do these tests do?

At first sight, data may seem to suggest that a drug treatment changes people's cholesterol levels or that high salt consumption is associated with high blood pressure and so on. However, we always need to use a statistical test to determine how convincing the evidence actually is. Such tests are known generically as 'Hypothesis tests'.

The first experimental design we are going to consider involves the measurement of the same interval scale endpoint in two groups of people, rats, tablets (or whatever). We calculate the mean value of the endpoint in each group and then want to test whether there is convincing evidence of a difference between the two mean values. The procedure we use is a two-sample *t*-test, the term 'Two-sample' reflecting the fact that we are comparing two distinct samples of individuals.

The test is also known as the 'Independent samples *t*-test'. The use of numerous apparently distinct names for exactly the same test is a handy device used by statisticians to keep everybody else in a constant state of uncertainty.

7.1.2 An example – Does rifampicin change the rate of elimination of theophylline?

It is known that the antibiotic rifampicin increases the amount of drug metabolising enzyme present in the liver and consequently increases the rate of elimination of a wide range of other drugs. This experiment is designed to detect whether rifampicin affects the metabolic removal of the anti-asthma drug theophylline. Any such interaction could be of real practical importance. A marked increase in the elimination of theophylline would result in inadequate treatment of the patient's asthma.

In the experiment, there are two groups of ten subjects. With the first group, each individual was pre-treated with oral rifampicin (600 mg daily for 10 days). The other group acted as a control, receiving only placebo pre-treatment. All subjects then received an intravenous injection of theophylline (3 $mg.kg^{-1}$). A series of blood samples were obtained after the theophylline injections and were analysed for drug content. The efficiency of removal of theophylline was reported as a clearance value.

Any increase in clearance would be interpreted as evidence that rifampicin had increased the ability of the liver to eliminate this drug. The results are shown in Table 7.1 and Figure 7.1. (The units of clearance are ml of blood cleared of drug every minute per kg of body weight.)

Table 7.1 Clearance of theophylline ($ml.min^{-1}.kg^{-1}$) for control subjects and for those pre-treated with rifampicin

	Control	Treated
	0.81	1.15
	1.06	1.28
	0.43	1.00
	0.54	0.95
	0.68	1.06
	0.56	1.15
	0.45	0.72
	0.88	0.79
	0.73	0.67
	0.43	1.21
	0.46	0.92
	0.43	0.67
	0.37	0.76
	0.73	0.82
	0.93	0.82
Mean	0.633	0.931
SD	0.216	0.202

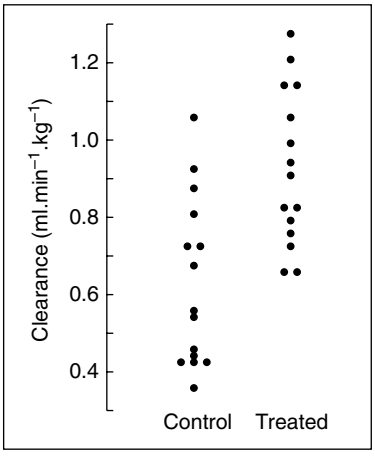


Figure 7.1 Theophylline clearance ($ml.min^{-1}.kg^{-1}$) in controls and in subjects pre-treated with rifampicin

These samples suggest that the mean clearance is about 50% greater in the treated group compared to the controls.

7.1.3 Diagrammatic representation of when the test is used

All the tests described in this book will be represented in the general form seen in Figure 7.2.

On the right-hand side of the figure we see the type of outcome data generated. In the current case, this is theophylline clearances which would be recorded as an interval variable. The significance of the term 'Unpaired' will be explained in Chapter 13. On the left-hand side, there will be one or more factors – things that may influence the outcome. In this case there is only one factor – the use / non-use of rifampicin. Data concerning this factor would be recorded using nominal data – labels such as 'Control' or 'Rifampicin-treated'. Where a factor is recorded using nominal data, there will be a restricted number of possibilities. In the current case, there are just two – Control and Treated and so we say there are two 'Levels' within this factor.

7.1.4 Null and alternative hypotheses

If rifampicin had no effect on the liver, the mean clearance in two very large groups of control and treated subjects would be virtually identical. However, with limited samples like these, both samples will probably yield imperfect estimates and their two means would almost certainly differ even in the absence of any real drug effect.

An apparent difference even in the absence of any real treatment effect

Samples are always subject to random error and control and treated samples are unlikely to produce identical means, even when a treatment has absolutely no real effect.

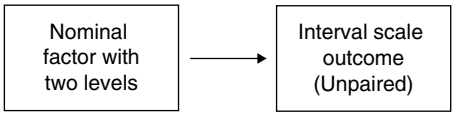


Figure 7.2 Diagrammatic representation of an experimental structure where use of the two-sample *t*-test is appropriate

There are therefore two possible explanations for the difference in clearance between our two samples; it could have arisen by sheer chance or it could be indicative of a real and consistent effect. The two possible explanations are formally called 'Null' and 'Alternative' hypotheses. Remember that at this stage, these are simply two competing theories. As yet, we have not decided which should be accepted.

- *Null hypothesis:* There is no real effect. The apparent difference arose from random sampling error. If we could investigate larger and larger numbers of subjects, the mean clearances for the two groups would eventually settle down to the same value. In stat speak 'The difference between the population mean clearances is zero'.

- *Alternative hypothesis:* There is a real effect and this is what caused the difference between the mean clearances in the samples. If we investigated larger and larger numbers of subjects, the results would continue to confirm a change in clearance. Also translatable into stat speak as 'The difference between the population mean clearances is *not* zero'.

All hypothesis tests start by setting out appropriate null and alternative hypotheses such as those above.

Null hypothesis

The statistical wet blanket: Whatever interesting feature has been noted in our sample (a change in an endpoint or a relationship between two sets of data etc.) is assumed not to be present in the general population. The apparent change or relationship is claimed to be due solely to random sampling error.

Alternative hypothesis

The logical alternative to whatever the null hypothesis claims. It will claim that the effect or relationship seen in the sample is perfectly real, i.e. it is present in the wider population.

To follow the analysis, it is particularly important to understand the mechanism implied by the null hypothesis. Figure 7.3 will be used to describe these alleged events. A vertical scale shows theophylline clearances. Each circle (open or closed)

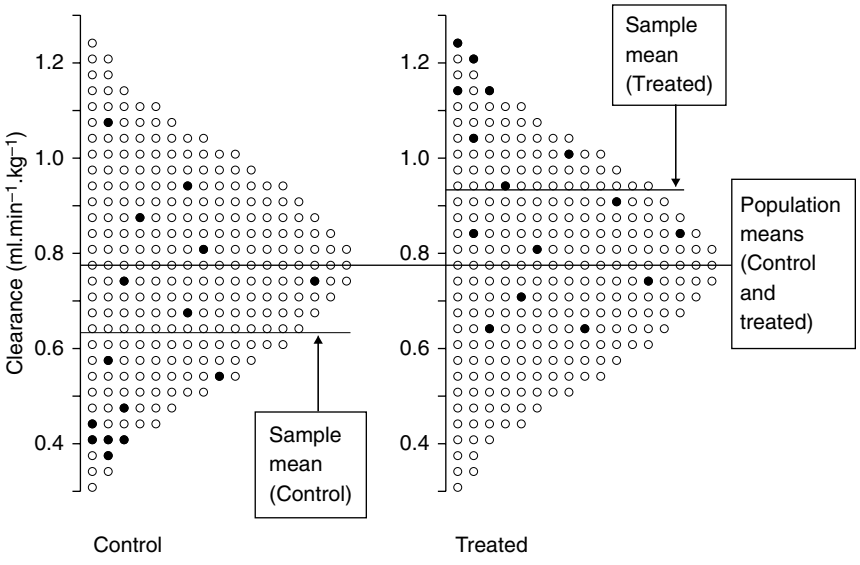


Figure 7.3 Null hypothesis: Random sampling error produced the appearance of an effect, although none was really present

represents a member of the population. The data form normal distributions (but these are shown turned at right angles to the usual orientation).

According to the null hypothesis, the two populations have identical means (as shown). We have then selected some individuals for inclusion in our random sample (filled circles). The null hypothesis assumes that our samples were both somewhat unrepresentative, leading to an apparently lower mean among the controls than among the treated subjects. At the end of the experiment, all we would actually see are the two samples with their differing means.

The alternative hypothesis (Figure 7.4) is much simpler. There really is a difference – the mean clearance among pre-treated individuals is greater than that among the controls – and the difference between our samples merely reflects that fact.

7.1.5 Using the *t*-test to generate the 95% C.I. for the difference between the two treatments

The function of the *t*-test is to assess the credibility of the null hypothesis. Its principal function is to generate a 95% C.I. for the difference between the two treatments. (We have already met the idea of a 95% C.I. for the difference between means in Section 6.9.)

From this point on, you can pretty well kiss Excel goodbye. For reasons that will be clarified in Chapter 10, its implementation of the two-sample *t*-test is highly undesirable.

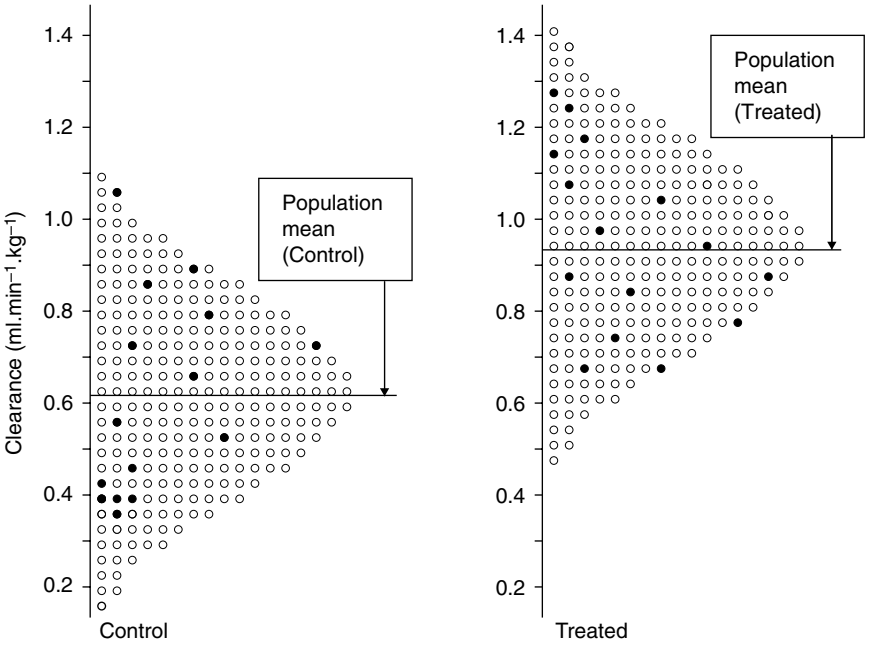


Figure 7.4 Alternative hypothesis: Rifampicin really has increased clearance values – No need to assume sampling error

In most packages, you will enter the data from both samples into a single column of a data sheet and then set up a second column that contains labels indicating which sample each value belongs to. With the theophylline clearance data, you would place all thirty clearance values in one column and then in another column, the first 15 cells would contain a code indicating a result from a control subject and the next 15 would contain a code indicated a rifampicin-treated subject. In some packages you will have to use numbers to code the two groups, in which case '0' is commonly used for controls and '1' for treated individuals, but any two numbers could be used. Many packages allow meaningful text labels such as 'Rif_treated' and 'Control'. To perform the test you will simply identify which column contains the data and which the treatment codes. In some packages, there is an option to enter the data for each sample in separate columns and then identify the two columns of data.

Table 7.2 shows generic output.

7.1.6 Is the null hypothesis credible?

Parts of the output may not be immediately familiar, but will be explained later. The key parts at the moment are the point estimate for the difference in clearance, expressed as clearance for treated subjects minus that for the controls (+0.298)

Table 7.2 Generic output from a two-sample *t*-test comparing theophylline clearances in rifampicin-treated and control subjects

Two-sample *t*-test	
Mean (Rif_treated)	0.931
Mean (Control)	0.633
Difference (Rif_treated – Control)	+0.298
95% C.I. Difference	+0.142 to +0.455
P	0.001

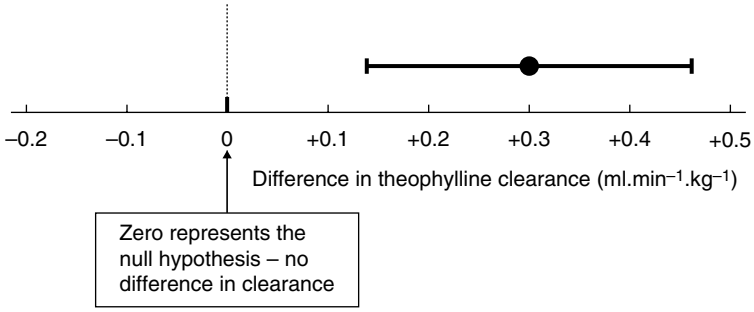


Figure 7.5 95% C.I. for the difference in theophylline clearance between control and rifampicin-treated subjects

and the 95% C.I. for this difference. The limits for the C.I. are +0.142 and +0.455 ml.min^{-1}.kg^{-1}. It is useful to represent this pictorially, as in Figure 7.5.

Figure 7.5 tells us that if we were to collect more and more data, the difference between the two mean clearances would not necessarily remain at the point estimate of +0.298 ml.min^{-1}.kg^{-1} that our small samples yielded. Ultimately we might find the true difference is actually as small as +0.142 or as large as +0.455 ml.min^{-1}.kg^{-1}.

The key point is that it's pretty obvious the null hypothesis (with its claim of zero difference between the population mean clearances) is difficult to believe. Figure 7.5 highlights the zero difference posited by the null hypothesis and this lies well outside the 95% C.I. Since the null hypothesis now looks so shaky, the balance shifts in favour of accepting the alternative hypothesis that rifampicin really does cause a change in theophylline clearance.

7.1.7 General interpretation of the results of a two-sample *t*-test

The general interpretation of the results that we might obtain from a two-sample *t*-test is shown in Figure 7.6:

Two of the confidence intervals in Figure 7.6, (a) and (c), would be interpreted as providing evidence of a difference. In the upper case (a) we see a convincing decrease

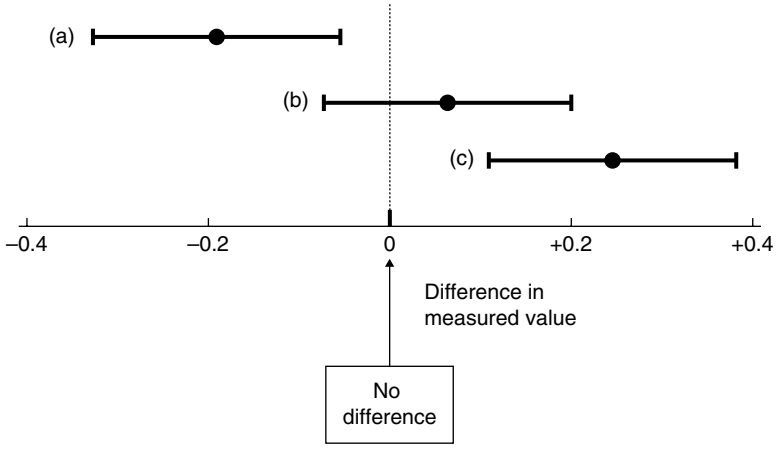


Figure 7.6 General interpretation of the results of a two-sample *t*-test

and in (c) a convincing increase. In the middle case (b), we would have to conclude that the evidence is not adequate to permit any positive conclusions; there is a real possibility that the treatment has no effect on whatever we were measuring, since a zero difference is included within the confidence interval.

Is there evidence of an effect?

If the C.I. *includes* zero, the null hypothesis that the treatment produces no effect is credible. Nothing has been proved.

If the C.I. *excludes* zero, we have worthwhile evidence that there is an experimental effect.

7.2 Significance

Where the confidence interval for a difference excludes zero, the data provides considerable support for the existence of a real difference and the results are granted the status of being 'Significant'. The original choice of this word was very canny. It tells us that the evidence is worthy of note and should be added to whatever other evidence is available. Unfortunately, many slip into the lazy habit of simply assuming that if the current set of results are significant, then the case is proven – Period. The correct interpretation of statistical significance is discussed more fully in Chapter 12.

When zero is within the 95% C.I., the results are described as 'Non-significant'. The use of 'Insignificant' as the opposite of significant looks uneducated.

Significant and Non-significant

If the evidence is 'Significant', it is strong enough to merit being added to whatever body of knowledge already exists. It does not mean that we should blindly accept the current result.

'Non-significant' implies that the evidence is weak and will have little influence upon our thinking.

7.3 The risk of a false positive finding

7.3.1 What would happen if there was no real effect?

A major requirement of any statistical test is that it should provide protection against concluding that there is a real effect, in circumstances where none really exists. This protection is considered below.

If the truth were that rifampicin actually had no effect at all (true difference in clearance = zero) and we repeatedly carried out the previous experiment, then we would anticipate the sort of results seen in Figure 7.7. Out of 20 experiments, 19 would be successful – the 95% C.I. will include the true mean change (none) and we would correctly conclude that there is no convincing evidence of an effect. In the one remaining experiment, we get an unfortunate coincidence of high values in one sample and low values in the other leading to apparently convincing evidence of a difference. (In this case, there seems to be a reduction in clearance.) In that one case, we would make a false diagnosis. Typically, when there is no real effect present, we will be fooled on just one out of 20 occasions (5%).

Cases where there is no real effect but we erroneously conclude that there is are called 'False positives'. Another name for these is a 'Type I error'.

'False positives' or 'Type I errors'

If there is no real effect of the treatment we are investigating, but we happen to obtain particularly misleading samples we may wrongly conclude that there is adequate (Significant) evidence of an effect. In that case, we have generated a 'False positive' or 'Type I error'.

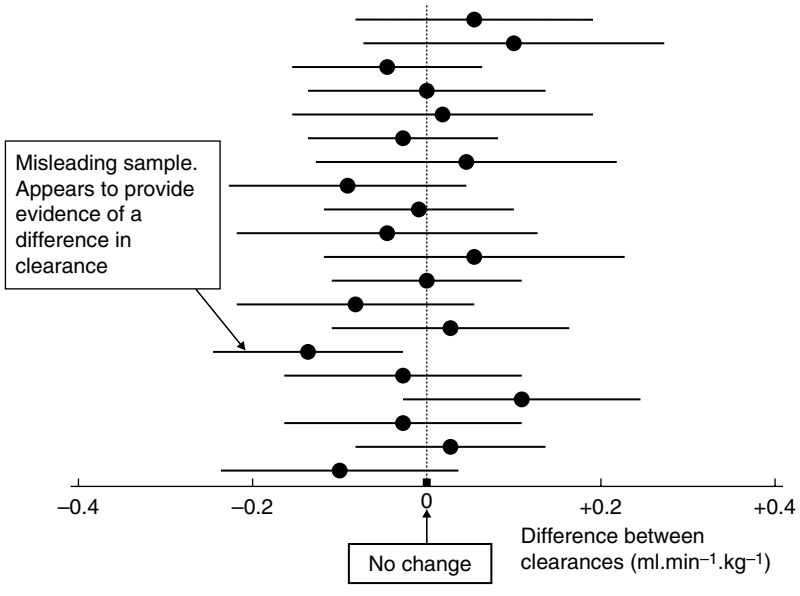


Figure 7.7 Repeated determinations of the 95% C.I. for the difference in theophylline clearance between control and rifampicin-treated subjects, assuming that there is actually no real effect

7.3.2 Alpha

The Greek letter alpha 'α' is used to represent this small residual risk of a false positive. Alpha obviously depends upon what confidence interval is inspected. In the case of the standard 95% C.I., the remaining risk is 100–95 = 5%. But, if we were particularly anxious to avoid the risk of a false positive, we might calculate a 98% C.I. and only declare a positive finding if that wider interval excluded zero. In that case Alpha would be only 2%.

α – The rate of false positives

Whenever we investigate a situation where there is no real difference present, alpha (α) is the risk that we will generate a false positive finding. With the usual 95% C.I.s, alpha is 5%.

7.4 What aspects of the data will influence whether or not we obtain a significant outcome?

Figure 7.6 made the point that we are looking to see if the 95% C.I. crosses (and therefore includes) the zero line. Two things influence the likelihood of a significant outcome and these are shown in Figures 7.8 and 7.9.

Figure 7.8 shows the influence of the size of the experimental effect. If the mean clearances differ to only a very small extent (as in the two lower cases), then the 95% confidence interval will probably overlap zero, bringing a non-significant result. However, with a large effect (as in the two upper cases), the confidence interval is displaced well away from zero and will not overlap it.

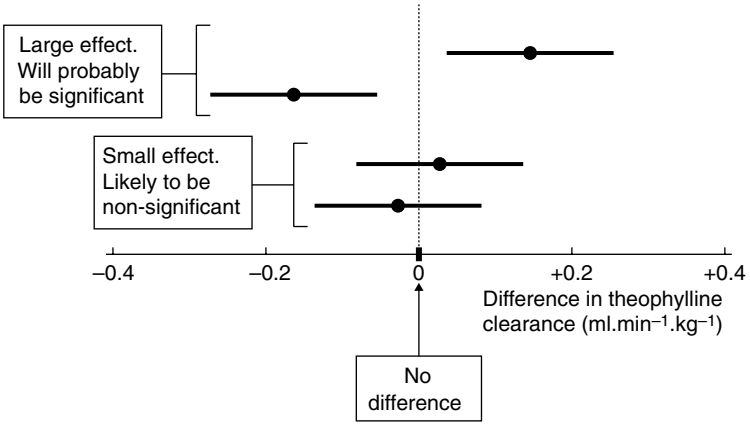


Figure 7.8 The size of the experimental effect and the likelihood of obtaining a significant outcome

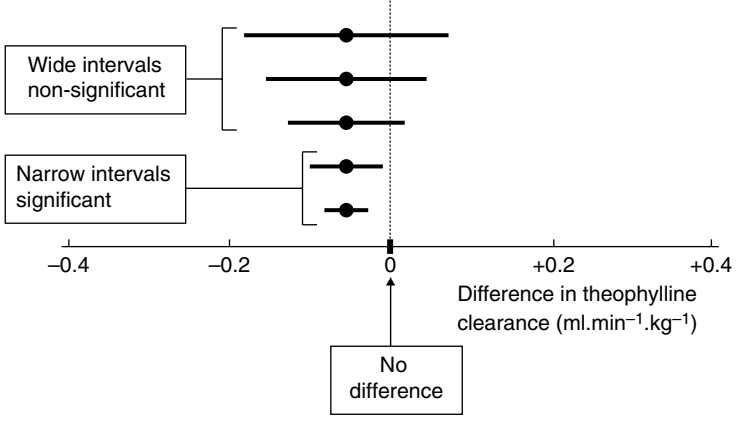


Figure 7.9 Width of the confidence interval and likelihood of obtaining a significant outcome

Figure 7.9 shows the effect of the width of the confidence interval. We know that the interval width depends upon the variability (SD) of the data being considered and the number of observations available. So, if the theophylline clearances are fairly consistent and large numbers of subjects have been studied, the interval will be narrow (as at the bottom of the figure) and the result significant. At the other extreme, if the clearances are highly variable and the numbers of subjects small, the interval will be wide, as at the top of the figure, and the result non-significant.

In summary, whether or not a significant outcome arises depends upon:

- The size of the experimental effect.
- The width of the confidence interval, which in turn depends upon:
  - the variability in the data;
  - the sample size.

This is summarised in Figure 7.10.

In this figure, the plus and minus signs are used to indicate the effect each aspect has on the chances of a significant outcome.

- A large experimental effect will increase the chances of a significant result. Hence the plus sign.
- Large SDs will widen the confidence interval and increase the danger of overlap with zero and so significance becomes less likely. Thus, this gets a minus sign.
- Large sample sizes will make the confidence interval narrower and so reduce the danger of overlap – hence significance more likely – this gets a plus sign.

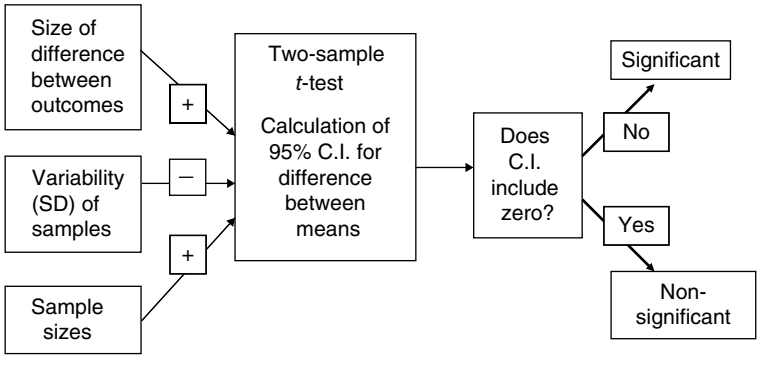


Figure 7.10 Factors influencing the outcome of a two-sample *t*-test

Aspects taken into account by the two-sample *t*-test

The two-sample *t*-test takes account of the apparent size of the experimental effect, the SDs of the samples and the sample sizes. It then combines all of these to determine whether the data are statistically significant.

An important aspect of the two-sample *t*-test is the ability to offset the three relevant aspects against one another. A couple of examples are set out below:

- If an experimental effect is present but is small in size, this makes it harder (but not impossible) to detect. What we would need to do is to strengthen the other two factors as much as possible. The amount of control we have over the variability in the data is pretty limited, so the main recourse would be to larger sample sizes. It doesn't matter how small the effect is, it is always possible to detect it if we can obtain sufficiently large samples.

- In the more comfortable scenario of a large experimental effect and data with low SDs, we would have the luxury of being able to use small samples and yet still obtain a significant result.

7.5 Requirements for applying a two-sample *t*-test

7.5.1 The two sets of data must approximate to normal distributions and have similar SDs

The data within each of the two samples should be drawn from populations that are normally distributed and have equal SDs. We have to be aware that the data we are dealing with are small samples. Judgements can be tricky. Even if populations adhere

Assumption of normal distributions and equal SDs

The mathematical basis of the two-sample *t*-test assumes that the samples are drawn from populations that:

- Are normally distributed;
- Have equal SDs.

perfectly to the above requirements, small samples drawn from them are unlikely to have perfect, classic, bell-shaped distributions or to have exactly coincident SDs.

The test is quite robust, but where there is strong evidence of non-normality or unequal SDs, action is required. We may be able to convert the data to normality, for example by the use of a logarithmic transformation (Introduced in Chapter 6). Alternatively, we may have to use an alternative type of test ('Non-parametric') which does not require a normal distribution. These are discussed in Chapter 21.

Where the problem is solely one of unequal SDs, there is a minor variant of the test 'Welch's approximate *t*' that does not have this requirement. In fact, in Minitab, this is the default test and if you want the classical *t*-test, you will have to select it (See www.ljmu.ac.uk/pbs/rowestats for details).

Two-sample *t*-test is robust but there are limits

The two-sample *t*-test is robust, that is it will still function reasonably well with samples that are somewhat non-normally distributed. However, where data is severely non-normal, even this test will start to produce inappropriate conclusions. In that case, either transform the data to normality or use a non-parametric test.

7.6 Performing and reporting the test

To perform the test you will need two variables. One must be an interval scale measure that records the outcomes (Clearances in our example) and the other must be a nominal variable with just two values which is used to record the group to which a particular result belongs.

The website associated with this book (www.ljmu.ac.uk/pbs/rowestats/) gives detailed instructions for performing the test using SPSS or Minitab.

In your methods section you should state what statistical package was used and the name of the procedure as is referred to in that particular programme (e.g. 'Two-Sample T-Test' in Minitab or 'Independent-Samples T-Test' in SPSS). You should also describe any option that you selected which differs from the defaults; for example, the option to assume equal variances for the two samples in Minitab.

In the results section, it is useful to include a figure that shows visually whether there is much evidence of any difference in mean values between the two groups and which also gives an impression of the distribution of the data. If a *t*-test is to be used, the reader needs to see that there is no strong evidence of non-normal distribution in the samples. In the present case, Figure 7.1 would fulfil these requirements.

When reporting the outcome of the *t*-test the key results are the means and SDs for the two samples and the limits for the 95% C.I. for the difference between the two group means. You should also report the *P* value (This is described in the next chapter).

In the case of SPSS you should report the results of Levene's test for equality of variances and (based on that) state whether the results quoted are those where equal variances are assumed or not assumed.

7.7 Chapter summary

We've met our first 'Hypothesis test'.

The two-sample *t*-test is used to determine whether two samples have convincingly different mean values or whether the difference is small enough to be credibly explained away as the result of random sampling error.

The data in each sample are assumed to be from populations that followed normal distributions and had equal SDs.

We create a null hypothesis that there is no real experimental effect and that for large samples the mean value of the endpoint is exactly the same for both groups. According to this hypothesis, random sampling error is responsible for any apparent difference and with extended sampling the apparent difference would eventually evaporate.

The alternative hypothesis claims that there is a real effect and that there is a difference between the mean values for the two populations. If we took larger and larger samples the difference would continue to make itself apparent.

A 95% C.I. for difference between the two means is used to determine whether the null hypothesis is reasonably credible. If the 95% C.I. includes zero, the results are compatible with the null hypothesis and are declared 'Non-significant'. If the 95% C.I. excludes zero, we have worthwhile evidence against the null hypothesis and declare the results 'Significant'.

The outcome of the test is governed by:

- Size of the difference between the sample means (Greater difference means significance more likely).
- SDs within the data (Greater SDs means significance less likely).
- Sample sizes (Greater sample sizes means significance more likely).

If we investigate a situation where there is no real difference and test our data by constructing a 95% C.I., there is 95% assurance that our interval will include the true treatment effect of zero and so we will draw the appropriate conclusion that the data is non-significant. This leaves a residual 5% risk that unusually misleading samples will apparently indicate significant evidence in favour of an effect. Such events are referred to as 'False positives' or 'Type I errors'. The degree of risk of such an event is represented by the Greek letter alpha – α. If a 95% C.I. is being used to test for significance, $\alpha = 5\%$.

8

The two-sample *t*-test (2): The dreaded *P* value

This chapter will ...

- Explain the information conveyed by *P* values.
- Describe how *P* values can be used to determine statistical significance.
- Describe how *P* values should be reported.
- Review how useful (or otherwise) *P* values actually are.

8.1 Measuring how significant a result is

In the last chapter, we saw how experimental results could be tested for statistical significance. The outcome was simply dichotomous – the results were either 'Significant' or 'Non-significant'. However, one could imagine a hypothetical series of outcomes such as those seen in Figure 8.1.

Outcomes (a) to (d) and (f) to (i) would all be judged significant and yet there are clearly considerable differences among them. Outcomes such as (d) or (f) are just barely significant, with zero only slightly outside the confidence interval, whereas for (a) or (i), the interval is well away from zero and the data is very clearly significant. The ancient statisticians who founded the methods of hypothesis testing hit upon a way to quantitate the level of significance. The underlying logic takes some thinking about, so concentrate hard …

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

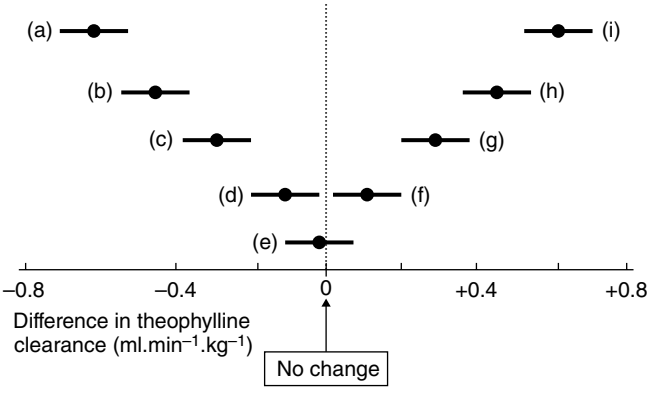


Figure 8.1 Various hypothetical outcomes for the effect of rifampicin on the clearance of theophylline

8.2 *P* values

We need to consider hypothetically what would happen if we investigated a treatment that had absolutely no effect. In particular, we calculate the risk that we might obtain a result as impressive as the one we actually observed. That likelihood is then quoted as the *P* value.

The *P* value can then be used as a measure of the strength of the evidence – the lower the *P* value, the more unlikely it is that such results would arise by sheer chance and so the stronger the evidence. An outcome such as (e) in Figure 8.1 would be very likely to arise with a treatment that has no real effect, so it has a high *P* value and provides virtually no evidence of an effect. Result (d), would arise relatively rarely as the interval fails to include zero, however such 'near misses' are not that rare and its *P* value would be a small percentage (less than 5%) – moderately strong evidence of an effect. The likelihood that a treatment which doesn't produce any real change would produce an outcome such as (a) or (i), would be extremely low, so its *P* value would be some minute fraction of 1%. Such a result would provide very strong evidence in favour of an effect.

A strict understanding of *P* needs to take account of two things:

- When considering an effect of a given size, we need to describe the likelihood of events that are not only exactly equal in size to the one we observed but also those more extreme. So, for outcome (d), we want to describe the likelihood that in, the absence of any real treatment effect, we might obtain that exact result *or something even more extreme* such as (a), (b) or (c).

- When determining the risk of an event such as (d), we must remember that outcome (f) would also arise with equal likelihood and would provide equally impressive evidence of an effect (albeit in the opposite direction).

When determining *P* we must therefore calculate the likelihood of an appropriately large difference *in either direction*.

Our final definition of the *P* value is therefore:

P value

P is the likelihood that with a treatment that has no real effect, we would observe an effect (positive or negative) as extreme as (or more extreme than) the one actually observed.

8.3 Two ways to define significance?

For no particular reason, the founding fathers of statistics decided to declare results 'Significant' if the associated *P* value was less than 5% ($P = 0.05$). At this point you might feel we have a problem – we now seem to have two criteria of significance:

- If the 95% confidence interval for the size of the experimental effect excludes zero, the result is significant.
- If the *P* value is less than 0.05, the result is significant.

However, although these two criteria sound very different, they are in fact mathematically equivalent; they require exactly the same standard of proof. With any set of data, the outcome will either be significant by both criteria or non-significant by both. We will never be faced with the dilemma of one criterion indicating significance but not the other.

P values and significance

Results are rated as significant if *P* is less than 0.05. This criterion will always produce a verdict concordant with that based upon inspection of whether the 95% C.I. for the effect size includes zero.

8.4 Obtaining the *P* value

Pretty well all statistical packages will include a *P* value as part of their out-put for the *t*-test. It was included as the last line in the generic output in Table 7.2, although its meaning was not explained at that stage.

Whatever package is used, the theophylline clearance data from the previous chapter should produce a *P* value of 0.001. This can be read as:

> 'If rifampicin actually had no effect on theophylline clearance, there would be only a 0.1% (one in 1000) chance that random samples would suggest an effect as great as (or greater than) the difference we actually observed.'

As this data had already been declared significant on the basis of the 95% C.I. for the size of the treatment effect, we would expect the *P* value to agree with this. The achieved *P* is indeed less than 0.05 and therefore significant.

Reporting result in terms of the *P* value

The data provided significant evidence of a change in clearance ($P = 0.001$).

8.4.1 Very low *P* values

Where data is overwhelmingly significant, some statistical packages have a nasty habit of generating silly *P* values, such as '$P = 0.000000012$'. Values as low as this are highly suspect, particularly as they would require our experimental data to adhere to a perfect normal distribution to an extent that we could never guarantee. The solution adopted in many statistical packages is a little brutal – They just round the result to a fixed number of decimal places and report in a format such as '$P = 0.000$'. This is also open to criticism as it could be read as a claim that there is a zero probability that such a result would arise by chance. This is never the case. If packages produce such output, it is best reported as '$P < 0.001$'

8.5 *P* values or 95% confidence intervals?

Traditionally statistics books used to emphasise *P* values. This was probably because the manual calculation of whether or not *P* is less than 0.05 is quite manageable, but for many statistical tests, calculation of the exact limits of the 95% C.I. can be pretty arduous. However, modern computer-based statistical packages remove that problem. Our choice of approach should now be based upon a consideration of what is best, not what involves the minimum of calculation steps. There are considerable arguments for preferring an inspection of the 95% C.I. These mainly arise because the *P* value only answers the question 'Is there an effect?', whereas the 95% C.I. answers that question, and additionally the question 'How great is the effect?'.

Chapter 10, in particular, will raise questions which can be answered only by inspecting the 95% C.I. for the size of the experimental response. In these cases, the *P* value is at best useless and at worst downright misleading.

The *P* value may have passed its glory days but it's not yet dead and buried:

- It has some value in quantifying just how strong our evidence is. A *P* value < 0.001 tells us the evidence is very strong, whereas $P = 0.04$ indicates that the data is barely significant. Inspection of the C.I. provides a visual (but not quantitative) impression of the same information. An interval that stops just short of zero is significant (but only just) whereas an interval that is far removed from zero indicates much stronger evidence.

- For some of the statistical routines that we will look at later, there is no single C.I. that could encapsulate the overall significance of a data set, so computer packages only report the *P* value not a C.I.

This book will always emphasise the use of 95% C.I.s wherever possible but accepts that, in some circumstances, *P* is all we are going to get.

P values – a dubious legacy?

In the past there has been an over-emphasis on *P* values. They are going to be around for a long time yet, but for many tests, *P* is less informative than the 95% C.I. and it is the latter that should be focused upon.

8.6 Chapter summary

P values provide a way to express how significant a finding is.

The correct interpretation of *P* is:

> *If the treatment we investigated actually produces no effect, the chances that we would obtain a result as impressive as (or more impressive than) the one we actually obtained, is* P.

If the *P* value is less than 5% ($P < 0.05$) we interpret the result as statistically significant. The verdict will always be the same as that obtained by observing whether the 95% C.I. for the difference between outcomes included zero.

Very low *P* values should not be quoted as unjustifiably small figures. It is better to report these as being less than some appropriately small figure (e.g. $P < 0.001$).

The *P* value gives no information regarding the size of the experimental effect and therefore, for many purposes, it is much less useful than the 95% C.I. for the size of the effect. Where the latter is available, it should be emphasised rather than the *P* value.

9

The two-sample *t*-test (3): False negatives, power and necessary sample sizes

This chapter will ...

- Describe 'False negatives' (Type II errors) where we fail to detect a difference when one is actually present.
- Show how we use 'Beta' to report the risk that we may fail to detect a difference and 'Power' to report the level of assurance that we will detect it.
- Consider the aspects of an experiment that determine its power.
- Expound the need to plan rationally how large our samples should be.
- Show how to calculate the necessary sample sizes for an experiment that will be analysed by a two-sample *t*-test.
- Suggest that experiments should be planned to be large enough (and therefore sufficiently sensitive) to detect any difference that would be of practical significance, but that resources should not be squandered trying to detect trivially small differences.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

9.1 What else could possibly go wrong?

In the last two chapters we established that, in a situation where there is actually no experimental effect, we will correctly declare that the evidence is unconvincing 95% of the time. This leaves us with a small (5%) proportion of cases where the sample misleads us and we declare the evidence significant. We referred to these cases as 'False positives' or 'Type I errors'. In this chapter we consider a new and quite different type of error.

9.1.1 False negatives – Type II errors

One of the factors that feed into the calculation of a two-sample *t*-test is the sample size. If we investigate a case where there is a real difference, but use too small a sample size, this may widen the 95% confidence interval to the point where it overlaps zero. In that case, the results would be declared non-significant. This is a different kind of error. We are now failing to detect an effect that actually is present. This is a 'False negative' or 'Type II error'.

'False negative' = 'Type II error'

Failure to detect a difference that genuinely is present.

9.1.2 The beta value

The Beta value (β) does a similar job to the alpha value. It reports the rate of type II errors. However, matters are a little more complicated, because the chances of detecting an experimental effect very much depend on how big it is (see Figure 7.9). If a large effect is present, it will be easy to detect, but very small effects are much more elusive. Think about the clearance data in the last two chapters. If rifampicin had a minute effect upon theophylline clearance – say it reduced average clearance by 1% – this would be almost impossible to detect and a Type II error would be practically inevitable. In contrast, a very large change in clearance would be much easier to detect and a Type II error would be less likely.

Consequently, Beta has to be defined in terms of the risk of failing to detect some stated size of experimental effect.

Beta

If a difference *of a stated size* is present, then Beta defines the risk of failing to detect it.

Table 9.1 Errors and truths

	Difference is not actually present	Difference is actually present
No difference detected	Correct	Type II error False negative (β)
Difference is detected	Type I error False positive (α)	Correct

The different types of error that can arise can be neatly summarised as in Table 9.1

Two of the possible outcomes are satisfactory. In the top left-hand box, there is no effect and we declare the evidence to be non-significant, and in the bottom right hand box an effect is present and we declare the evidence to be significant. The two types of error are seen in the other two boxes. In the top right case, we fail to detect a difference and in the lower left, we declare the evidence to be significant when there is no real difference.

9.2 Power

If Beta is the chance that we will fail to detect a difference, then Power is simply the opposite; it is the chances that we will successfully detect the difference.

Power

If a difference *of a stated size* is present, then Power defines the likelihood that we will detect it and declare the evidence statistically significant.

Power and Beta are then simply related as:

$$\text{Power} = 100\% - \beta$$

So, if Beta is 10%, Power will be 90%.

9.2.1 The power curve

Power will be related to sample size. Figure 9.1 shows the underlying logic using the investigation of altered theophylline clearance as an example. At the top of the figure, small samples produce wide 95% C.I.s which are almost certain to overlap zero, so there is little hope of significance. At the bottom of the figure, large samples produce narrow intervals and significance is practically guaranteed.

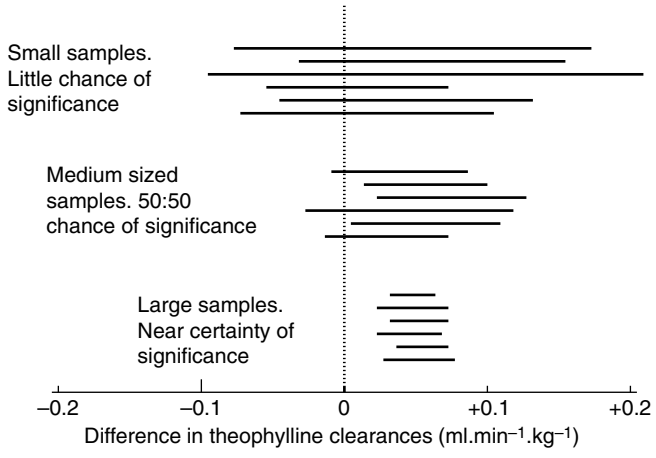


Figure 9.1 Effect of sample size upon the width of the 95% C.I. for the difference in outcomes and hence on the likelihood of statistical significance

However there is an intermediate situation where the outcome is unpredictable. With medium-sized samples, we may obtain data that happens to slightly understate the size of the difference in clearances and the C.I. crosses zero (non-significant), but the next pair of samples may suggest a greater difference and the result is then significant.

So, power increases with sample size. The relationship between sample size and power for our rifampicin/theophylline experiment is illustrated in Figure 9.2. This assumes that the true difference between the placebo and rifampicin treated subjects is 0.3 $ml.min^{-1}.kg^{-1}$ and that the SD among theophylline clearances is 0.21 $ml.min^{-1}.kg^{-1}$.

The figure shows that, with very small samples (less than five) there would be little chance of detecting the change in clearance, but with larger samples (20 or more) we are almost certain to detect the effect. With a sample size of five our chances of achieving a significant outcome are about 50:50. The sample size used in the experiment introduced in Chapter 7 was ten, which provided a power of about 86%.

9.2.2 Aspects of the data that determine the power of an experiment

Power is influenced by four aspects of any experiment.

9.2.2.1 Size of difference that we are trying to detect As described in Section 9.1.2, large differences are easier to detect than small ones and so, the greater the difference we are trying to detect, the greater the power of our experiment.

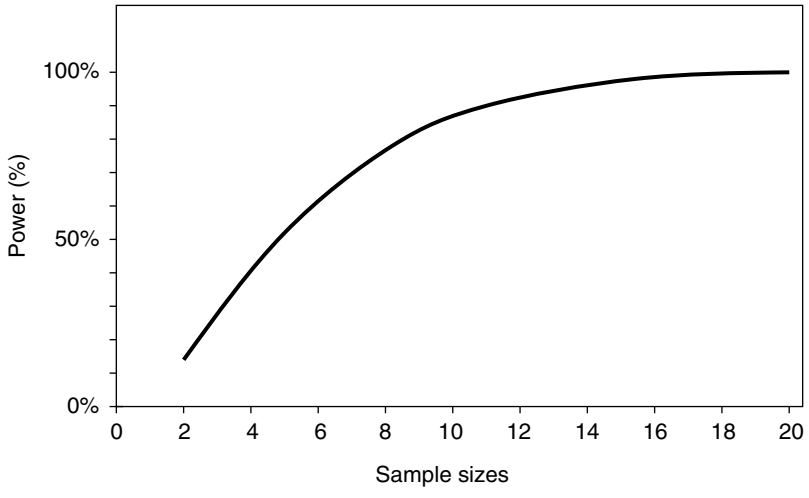


Figure 9.2 Power curve for the theophylline clearance experiment. (Assumes the true difference is an increase of 0.3 ml.min^{-1}.kg^{-1} and the SD among clearances is ±0.21 ml.min^{-1}.kg^{-1})

9.2.2.2 Sample size In a previous section we saw that sample size has a strong influence upon power. Greater sample sizes give greater power.

9.2.2.3 Variability of the data We know that the width of the 95% C.I. depends upon the variability of the data (Section 6.9). If data vary greatly, the confidence interval will be wider and more likely to overlap zero, implying non-significance. Greater SDs bring less power.

9.2.2.4 Standard of proof demanded (α) A formal calculation of power technically needs to take into account the standard of proof being required for a declaration of significance. The usual criterion is that the 95% confidence interval excludes a zero effect ($\alpha = 0.05$). If an experiment was designed to achieve a higher standard of proof (e.g. $\alpha = 0.02$), a 98% C.I. will have to be used and this will be wider than the standard 95% C.I. The wider interval is then more likely to cross the zero line and so power will be lower. So, requiring a lower risk of a false positive (reducing alpha) will lead to less power.

The alpha value was included in the above discussion, simply for the sake of completeness, but from this point on, we will ignore the question of standards of proof and simply assume that the normal standard is used (i.e. alpha = 0.05; a 95% C.I. is used or significance accepted if $P < 0.05$). Figure 9.3 shows the inputs into a formal calculation of the power of an experiment and the effects they would have.

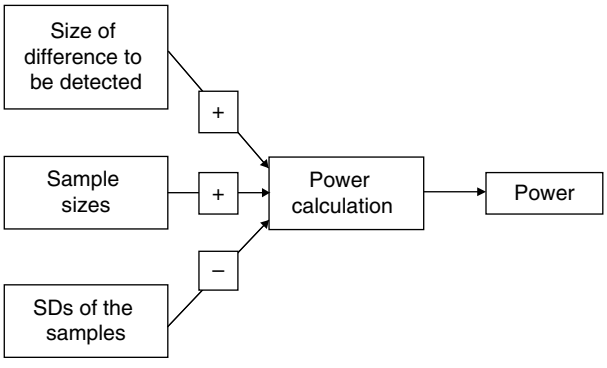


Figure 9.3 Calculation of the power of an experiment that will be analysed by a two-sample *t*-test

Main factors affecting the power of a study

- The greater the sample size, the greater the power.
- The greater the experimental effect that we are trying to detect, the greater the power.
- The greater the variability in the data, the less the power.

9.3 Calculating necessary sample size

9.3.1 The need for rational planning

In day to day practical experimentation, sample sizes probably owe more to tradition than logic. PhD students use five rats not on the basis of any rational calculation but simply because their supervisor always used five. (And they in turn used five because their supervisors used five and so *ad infinitum*.) There is now increasing recognition that hereditary malpractice ought to be replaced by soundly based optimisation.

It is a sin to perform an experiment that is either too small or too large. Experiments that are excessively large are obviously a waste of resources. An answer could have been arrived at, with less expenditure of time, effort and cash. If animals or humans were the subjects of the experiment, then there is also the ethical issue of unnecessary inconvenience or suffering. Experiments that are too small are perhaps even worse. If numbers are inadequate then the statistical analysis will almost certainly produce a non-significant result and we will be none the wiser. An effect may yet be present, but our pathetic little experiment could never

have detected it. In that case all the expense and patient inconvenience were completely wasted. It is therefore important to plan the sizes of samples so they are big enough but not too big. The general impression is that while examples of both under- and over-powered experiments exist, much the commonest problem is under-powering.

9.3.2 Factors influencing necessary sample size

The road to proper sample size planning begins here. In the same way that we can re-arrange an algebraic equation, we can also re-arrange one of our block diagrams. If we take Figure 9.3 which shows the factors that influence power, we can re-arrange it (Figure 9.4) to show the factors that influence necessary sample size. Sample size has been moved to the right-hand side of the 'equation' and power to the left.

The logic of these pluses and minuses is:

- Size of experimental effect to be detectable. Large experimental effects are relatively easy to detect and small sample sizes will suffice. But, smaller effects require larger sample sizes. The relationship is negative.

- Standard deviation. Greater variability in the response will make it harder to detect an effect and so we will need larger sample sizes – a positive relationship.

- Power required. Power increases with sample size, so if we want more power we will need larger samples – another positive relationship.

Figure 9.4 implies that, if we want to be able to calculate necessary sample size, we will need to supply values for each of the factors on the left of the diagram. At this point it would be useful to make a few additional comments about these three factors.

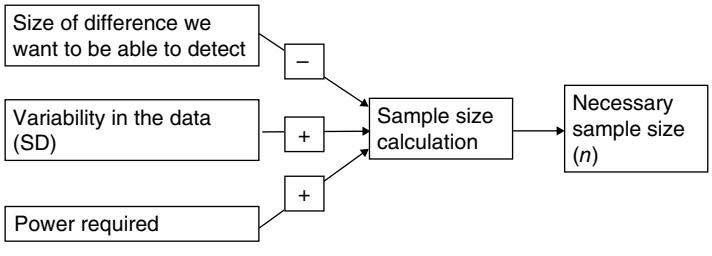


Figure 9.4 Calculation of necessary sample size for an experiment that will be analysed by a two-sample *t*-test

Main factors affecting necessary sample size

- Size of difference to be detected.
- Standard deviations in the samples.
- Power to be achieved.

9.3.3 Size of experimental effect we want to be able to detect

Necessary sample size will be governed by the smallest difference we want to be able to detect. The latter is not based on statistical considerations. It can only be defined by somebody with expertise in the particular field. They should be able to say what sort of size of effect needs to be detectable.

There is always a temptation to be a super-perfectionist and insist that if there is even the slightest difference, you want to know about it. However, if you set the detection limit to an excessively low value, the necessary sample size will probably escalate to something unmanageable. We have to accept that we cannot detect infinitesimally small effects and we must draw the limit at some reasonable point.

It is useful to consider 'What is the smallest experimental effect that would be of *practical* significance?' It is pointless trying to detect any effect much smaller than this figure. If we do try to detect a smaller effect, our experiment will be unnecessarily large and the extra cost and effort will merely enable us to detect effects that are too small to matter anyway! However, it is also important not to stray in the opposite direction. If we set the detection limit too high, this may yield a satisfyingly small sample size, but our experiment will now be too insensitive and may fail to detect a difference that is large enough to be of real practical significance.

In clinical trials, the smallest practically significant change in outcome is usually called the 'Clinically Relevant Difference' (CRD).

Size of effect to be detected

The figure we choose for the smallest difference we want to be able to detect, should match the smallest effect that would be of real practical significance.

No bigger. No smaller.

9.3.4 The SD among individual responses to the treatment

When a statistician who has been asked to plan the size of an experiment, demands to know what the SD is going to be, the response is often quite shirty. 'How the devil would I know, I've not done the experiment yet.' is not untypical. The objection is theoretically valid, but in practice, rarely a real problem. There are several obvious solutions which include:

- Look in the literature for other experiments measuring the same endpoint and see how variable other people found it to be.
- Conduct a pilot experiment to estimate the SD and then plan the main experiment.
- Start the experiment and take a peek after the first few cases, use these to estimate the SD and then work out how big the experiment needs to be.

9.3.5 How powerful we want our experiment to be

Like the detection limit, this is also something that cannot be calculated statistically. It really depends upon the resources available and how critical the experiment is. It is usual to aim for a power somewhere between 80 and 95%. Always remember that if we go for greater power, a greater sample size will be required. If resources are plentiful and the experiment is critical then it might be perfectly reasonable to plan to achieve 95% power. If data is either very difficult or expensive to generate then, you might be forced to settle for 80% power.

9.3.6 An example – the theophylline clearance experiment

To calculate an appropriate sample size for the rifampicin/theophylline experiment, we need to establish suitable values for the three inputs to the power calculation.

9.3.6.1 Minimum effect to be detectable We talk to some 'experts' and they decide that, if theophylline clearance is changed to an extent of plus or minus 20%, they would want to know about it. (The implication is that if some smaller difference were present, but our experiment failed to detect it, it would be considered no serious matter.) A typical textbook mean clearance for theophylline (under normal circumstances) is 0.67 $ml.min^{-1}.kg^{-1}$. So, a 20% change would equate to an increase or decrease of about 0.13 $ml.min^{-1}.kg^{-1}$.

9.3.6.2 SD among theophylline clearances A review of the literature concerning the clearance of theophylline gave conflicting results, but a 'guesstimate' of SD = ±0.25 $ml.min^{-1}.kg^{-1}$ was agreed upon.

9.3.6.3 Power to be achieved Data for an experiment of this type is likely to be acquired quite slowly and will be expensive to obtain, so we are unlikely to want to try for a very high power such as 90 or 95%. Furthermore, on ethical grounds, we don't want to expose unnecessarily large numbers of subjects to this experiment, so we settle for 80% power.

9.3.7 Using a statistical package to calculate necessary sample size

You may need to shop around for a statistics package that does calculations of necessary sample size. SPSS does not have this functionality as part of its standard package – you need access to an add-on (SamplePower®). Minitab is the winner in this regard – it does include suitable calculations for all simple experimental designs. Instructions for calculating sample sizes are included on the website for this book, www.ljmu.ac.uk/pbs/rowestats/.

Whatever package you use, you will have to provide values for the three inputs shown in Figure 9.4. Generic output is shown in Table 9.2.

The number calculated (60) is the number per group, so 120 subjects would have to be recruited and then divided into control and treated groups. It is obviously only possible to use exact whole numbers of subjects – fractional patients are impossible (unless some terrible accident has occurred). Generally, no exact number of subjects will provide precisely 80% power, so your package should report the smallest whole number that will provide at least the requisite power. If your package reports a fractional number, then in order to provide at least adequate power, it should always be rounded up to the nearest integer – we do not round to the nearest integer if that would mean rounding downwards. The package also reports the actual power that will be achieved having rounded the number of subjects upwards.

9.3.8 Was our experiment big enough?

In our real experiment we used only 15 subjects per group, not 60, and yet, despite this, we got a clearly significant result. The main reason for this is that the sample size of 60 was calculated as being adequate to reveal a change of 0.13 $ml.min^{-1}.kg^{-1}$,

Table 9.2 Generic output for calculation of necessary sample size for the rifampicin/theophylline experiment

Sample size for two-sample *t*-test	
Assumed SD	0.25
Difference to detect	0.13
Power required	0.8
Sample size	60
Power achieved	0.806

whereas the actual difference found was almost 0.3 $ml.min^{-1}.kg^{-1}$. Because the real effect was so large, it was easy to detect and our small experiment proved adequate. (More by good luck than good management.)

Notice that this still does not justify conducting such a limited experiment. If the difference had been (say) 0.2 $ml.min^{-1}.kg^{-1}$, that would been big enough for us to need to know about it, but our experiment would almost certainly have failed to detect it.

9.3.9 Sensitivity of necessary sample size to the three controlling factors

So far we have identified which factors influence necessary sample size. In the rest of this chapter we will see how great an effect each of these factors has.

9.3.9.1 Varying the minimum difference in clearance to be detectable In our sample size calculation, the minimum effect to be detectable was set at a change of 0.13 $ml.min^{-1}.kg^{-1}$. What would happen if we varied this detection limit while maintaining everything else constant? Figure 9.5 shows the results. We already anticipated that necessary sample size would increase if we tried to detect small changes, but this figure shows that the relationship is an extremely sensitive one. In fact the necessary sample size depends approximately on the square of the detection limit. So if we tried to make the experiment ten times more sensitive (i.e. reduce the detection limit by a factor of ten), the necessary sample size would increase 100-fold!

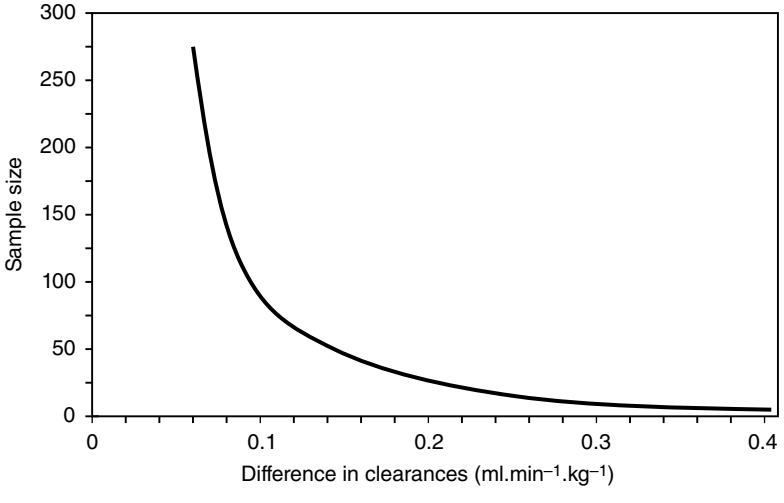


Figure 9.5 Necessary sample size when varying the minimum difference to be detected (SD = 0.25 $ml.min^{-1}.kg^{-1}$ and Power = 80%)

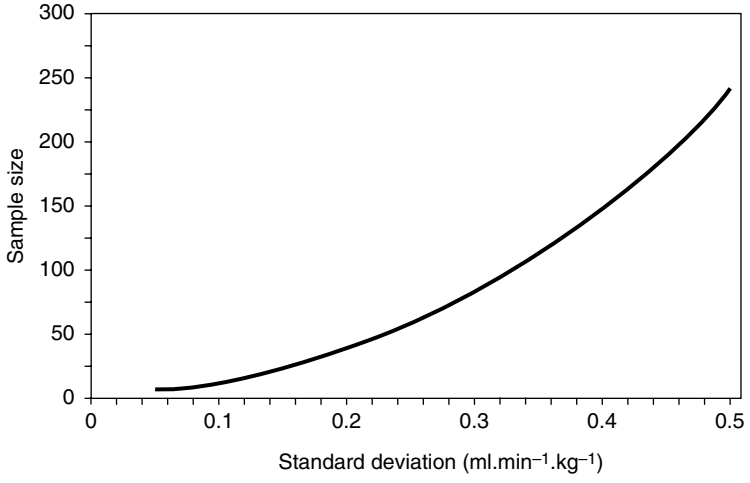


Figure 9.6 Necessary sample size when varying the SD for the samples. (Difference to be detectable = 0.13 ml.min^{-1}.kg^{-1} and Power = 80%)

This serves to emphasise how important it is not to be overly ambitious. Any attempt to detect experimental effects that are trivially small, could lead to astronomically large numbers.

9.3.9.2 Varying the SD among theophylline clearances We can do a similar exercise, varying the assumed SD among the clearances, while holding everything else constant. Figure 9.6 shows the results.

SD also has a powerful effect upon necessary sample size. This is also approximately a square relationship with sample size depending on the square of the SD. So, if we could halve the variability in our data, the sample size would fall to about one-quarter.

It does therefore follow that it is very important to reduce the SD as much as possible. Theophylline clearances are intrinsically variable and there is nothing we can do about that. However, random measurement errors may be inflating the SD and we should certainly be attempting to eradicate all unnecessary sources of variability.

9.3.9.3 Varying the planned power of the study Finally, we can vary the planned power for the study, while holding everything else constant. Figure 9.7 shows the results.

As we anticipated, achieving increased power will require greater sample sizes. However, the effect is quite modest. Within the commonly used range of power from 80 to 95%, necessary sample size does not vary greatly. If we were to try to raise the power beyond 95%, then sample sizes would start to increase uncomfortably, but it would be unusual to attempt to achieve such high power figures.

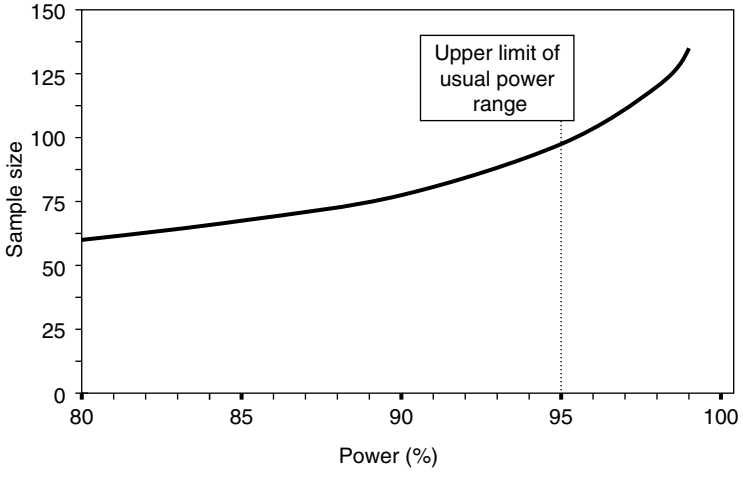


Figure 9.7 Necessary sample size when varying the planned power. (Difference to be detectable = 0.13 $ml.min^{-1}.kg^{-1}$ and SD = 0.25 $ml.min^{-1}.kg^{-1}$)

Factors influencing necessary sample size

Necessary sample size shows the following pattern of dependencies:

- Highly sensitive to minimum effect to be detected.
- Highly sensitive to the standard deviations within the samples.
- Modestly sensitive to the planned power of the experiment.

9.3.10 Calculated sample size sometimes unrealistic

The very formal approach outlined above is the ideal and would certainly be required in (say) a clinical trial intended to support a marketing application. However, in the real world, we sometimes dutifully go through the requisite steps, only to find that the required sample size is way beyond our resources. That does not mean we should automatically abandon the whole idea.

In Section 9.3.7 we calculated that sample sizes of 60 were apparently necessary for our rifampicin/theophylline experiment. If we had carried out that calculation prior to performing the experiment (as indeed we should), we would recognise that such numbers are quite impractical. A way forward, would be to decide how big the samples could realistically be and then investigate whether it is worth continuing.

Let us assume that the maximum sample size practically achievable is 15. Power, difference to detect and sample size are all interrelated and if we can fix

any two of them, we can calculate the remaining value. So, we could fix sample size at fifteen and power at 80% and calculate what difference in clearance would be detectable. The result would be that we would have 80% power to detect a change of 0.265 $ml.min^{-1}.kg^{-1}$.

You now need to consider whether it is realistically likely that the effect might be this great. If you decide that this is much greater than anything that is realistically likely to emerge, then you are unfortunately at a dead end. However, if it looks like something that might emerge, then it may still be worth continuing with the work. You just need to be aware that it is a relatively insensitive experiment; it is only capable of detecting a relatively large experimental effect – smaller (but possibly practically significant) differences may escape detection.

Minitab can be used for the approach outlined above; it contains a particularly neat routine to calculate any one of power, difference or sample size from values for the other two. The instructions for sample size calculation on www.ljmu.ac.uk/pbs/rowestats include the alternative approach of calculating the detectable difference.

9.4 Chapter summary

If we fail to detect a difference that truly is present, this constitutes a false negative or type II error. If a difference of a stated size is present, 'Beta' is the risk that we will fail to detect it. Conversely, 'Power' is the probability that we will detect it.

The main factors affecting power are:

- The greater the sample size, the greater the power.
- The greater the size of the difference that we are trying to detect, the greater the power.
- The greater the variability in the data, the lower the power.

The design of all experiments, should involve a rational calculation of how large the samples need to be. For an experiment that will be analysed by a two-sample *t*-test, the main influences upon necessary sample size are:

- Increasing the minimum difference that we want to be able to detect dramatically decreases sample sizes.
- Increasing the variability of the data dramatically increases sample sizes.
- Increasing the planned power of the experiment moderately increases sample sizes.

When calculating sample sizes, the figure for the smallest difference to be detectable should generally match the smallest difference that would be of practical significance.

10

The two-sample *t*-test (4): Statistical significance, practical significance and equivalence

This chapter will ...

- Show that a statistically significant difference is not synonymous with a practically significant difference.
- Introduce equivalence limits and the equivalence zone.
- Explain how to test for a practically significant difference between two measured outcomes.
- Explain how to test for equivalence (i.e. show that two measured outcomes do not differ to an extent that is of practical significance).
- Explain how to test for non-inferiority (i.e. show that one product or process is at least as good as another).

10.1 Practical significance – Is the difference big enough to matter?

Figure 7.9 showed the interplay between various aspects of the data in determining the outcome of a two-sample *t*-test. Two of these factors were the extent of the difference observed and the sample sizes. It is perfectly possible for a study to produce

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

a statistically significant outcome even where the difference is very small, so long as the sample size is correspondingly large. In principal, there is no lower limit to the size of experimental effect that could be detected, if we were prepared to perform a big enough experiment. But, this can cause problems, as the next example demonstrates.

10.1.1 Detection of a trivially small effect upon haemoglobin levels

We want to check whether a widely used antidepressant might be having an effect upon haemoglobin levels (possibly by interfering with vitamin absorption). We have already accumulated a mass of basic clinical data from routine monitoring of patients taking various antidepressants. We simply ask the IT department to extract haemoglobin data for at least 200 patients taking the suspect medicine and from another 200 (plus) taking any anti-depressant other than the one that is suspect. These later will be our controls.

Controls: $n = 220$ Mean haemoglobin = 157.61 ± 11.85 $g.L^{-1}$ (±S.D.)
Suspect drug: $n = 235$ Mean haemoglobin = 155.10 ± 12.19 $g.L^{-1}$ (±S.D.)

There is some evidence of lower haemoglobin levels.

Subjecting this data to a two-sample *t*-test, we find that the point estimate for the difference between mean haemoglobin levels (expressed as Suspect drug – Control) is –2.51 $g.L^{-1}$ with a 95% confidence interval of –0.294 to –4.728 $g.L^{-1}$. As zero is excluded from the interval, there is significant evidence of a difference in haemoglobin levels.

The danger is that word 'Significant'. For the uninitiated there will be a temptation to interpret 'Significant' as indicating a difference big enough to be of practical importance. However, statistical significance is only an indication of whether or not there is a change, and tells us nothing about how big any difference might be. If we want to know whether the effect is big enough to matter, we need to look at the actual size of the change and compare this to what is known about the clinical significance of variations in haemoglobin levels.

Statistical and practical significance

Statistical significance: Is there a difference?
Practical significance: Is there a difference big enough to matter?

10.1.2 'Equivalence limits' and 'Zone of equivalence'

As a first step towards a formal test of practical significance we need to establish just how large any change would need to be, for it to have practical consequences. This decision is not based on any statistical considerations – it depends upon the judgement of an expert in the particular field.

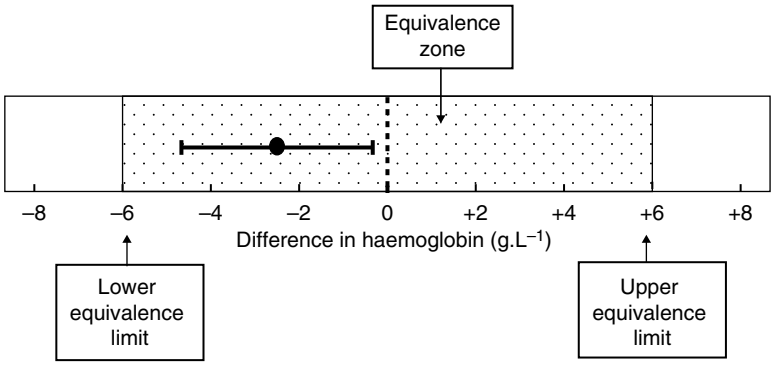


Figure 10.1 Confidence interval for the difference in haemoglobin levels in relation to the equivalence zone

Experts suggest that we can be confident that any change of less than 6 $g.L^{-1}$ will have no practical consequences. This allows us to establish lower and upper 'Equivalence limits' at –6 and +6 $g.L^{-1}$. For the purposes of this discussion, we will assume that these patients tend to be somewhat anaemic, so a reduction of greater than 6 $g.L^{-1}$ would be detrimental and a similar sized increase would be beneficial. Between these possibilities, there is a 'Zone of equivalence'. (We are confident the patient will be neither better nor worse.)

Equivalence limits

By how much would the parameter in question need to change before there is a realistic possibility of practical consequences? They are not calculated by any statistical procedure; they are based upon expert opinion.

We could then present the results of our haemoglobin experiment as a nice, intuitive diagram. (Figure10.1)

Notice that the horizontal axis is the difference in haemoglobin levels in the two groups. The whole of the 95% C.I. for the difference arising from use of the suspect drug lies within the zone of equivalence. Even looking at the greatest effect the suspect drug might realistically have (A change of –4.7 $g.L^{-1}$), this is still too small to be of any consequence.

Haemoglobin levels

- The results are *statistically* significant. The 95% C.I. excludes zero, so there is evidence of a difference between the two groups.
- The results are not *practically* significant. The difference is too small. None of the 95% C.I. extends beyond the equivalence zone.

10.1.3 General interpretation of practical significance

The general way in which we would determine the practical significance of a change in haemoglobin levels is summarised in Figure 10.2, assuming that decreases in haemoglobin would be detrimental and increases beneficial.

The first three cases are unambiguous:

a. Is detrimental and definitely of practical significance. We are confident that even the minimum effect (a reduction of 7.5 $g.L^{-1}$), is large enough to cause practical harm.

b. Definitely has no practical significance. The whole C.I. is within the equivalence zone.

c. is beneficial and definitely of practical significance. The most pessimistic view would be that there was an increase of 7 $g.L^{-1}$, which is still enough to be useful.

In the remaining cases, the C.I. crosses between zones and the result is ambiguous. We are limited as to what conclusions can be drawn and judgement has to be exercised.

d. The only definite thing we can say is that it is not doing any good, since none of the 95% C.I. extends into the zone of practical benefit. Most of the C.I. is in the zone of practical detriment, so we would almost certainly want to avoid whatever caused this effect, even if it is not yet proven to be doing practical harm.

e. This definitely does no practical harm, and may be beneficial. The C.I. is quite wide and a larger experiment should narrow it, with reasonable prospects that it would then emerge as unambiguously beneficial.

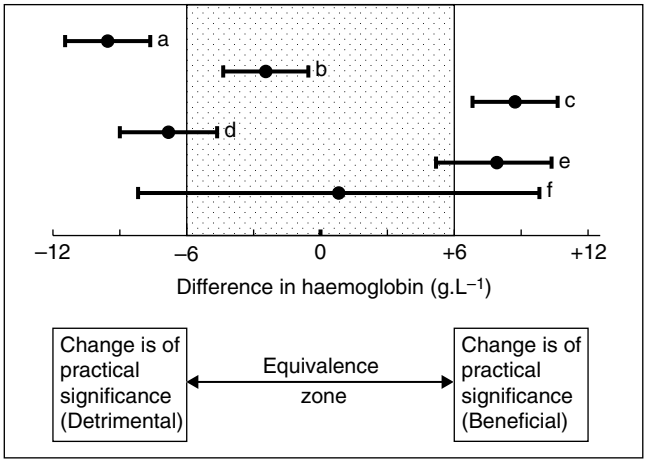


Figure 10.2 Determining the practical significance for differences in haemoglobin levels

f. Is completely ambiguous. We can say nothing definite. The treatment could be beneficial, harmful or have no noticeable effect. The experiment appears to have been of very low power – sample size probably too small.

Practical significance

To demonstrate that a treatment produces an effect great enough to be of practical significance, we must show that the 95% C.I. for the size of the treatment effect lies entirely outside the equivalence zone.

10.1.4 Unequal sample sizes

In the previous example, the two samples differed slightly in size. This does not invalidate the *t*-test; the calculation of the test can incorporate such inequalities without difficulty. The only reason we normally used balanced designs (equal numbers in both groups) is that for any total number of observations, statistical power is greatest if they are split equally. In this case the imbalance is so small that any loss of power will be trivial. However, if imbalance leads to one of the groups being quite small, the loss of power may make the whole exercise futile.

Balanced samples

To maintain statistical power, try to avoid major imbalances between the sizes of the two samples being compared.

10.2 Equivalence testing

Traditional statistical testing tends to focus on attempts to demonstrate that a measured value changes according to what treatment is used. This is called 'Difference testing'. However, there are occasions when we want to show that a measured value does not change. This, we would call 'Equivalence testing'.

Typical cases where we might wish to demonstrate equivalence include changing a pharmaceutical formulation of a drug. The important factor might be that the new preparation delivers the same amount of drug as the old one. We want patients to be able to switch from the old to the new preparation without their effective dose changing.

There is a problem however – it is impossible to demonstrate that there is absolutely no difference between two treatments or procedures. Take the case above – two drug

preparations. Could we ever demonstrate that they deliver exactly the same amount of active material? The answer must be 'No'. Any experiment we conduct will be finite in size and can therefore only detect differences down to a certain size. However large an experiment we conduct, there could always be some tiny difference just too small for the current experiment to detect. It is therefore impossible to demonstrate that two preparations behave absolutely identically.

Since we can't demonstrate exact equality, we turn instead to practical equivalence. We try to demonstrate that any possible difference between two outcomes is too small to be of practical significance.

Equivalence testing is conducted in three stages:

1. Determine an equivalence zone, based on expert knowledge of the likely effect of a change of any given size.

2. Determine the 95% C.I. for the difference between the two products/methods or whatever else is being tested.

3. Check to see if the whole of the 95% C.I. for any difference fits entirely within the equivalence zone. If it does, then we have a positive demonstration of practical equivalence. If either end of the interval extends beyond the equivalence zone, then equivalence cannot be assured.

Equivalence testing

Equivalence testing is designed to show that there are no differences big enough to be worth worrying about. To demonstrate equivalence, the whole of the 95% C.I. for the size of any difference should lie entirely within the equivalence zone.

10.2.1 An example of equivalence testing – comparing two propranolol formulations

Propranolol is a drug used to reduce blood pressure. It is known that only a limited proportion (about one-third) of the dose swallowed actually survives intact and gets into the patient's blood system. When changing the formulation of such a drug, one would want to be assured that the proportion being successfully absorbed did not suddenly change. The best marker of absorption is the so-called AUC, that is the Area Under the Curve for the graph of blood concentration versus time. This is measured in units of $\mu g.h.L^{-1}$. Table 10.1 shows the AUC values when a series of individuals take either the old or the new preparations.

Table 10.1 Areas Under the Curve (AUCs) for old and new formulations of propranolol

AUCs for Old preparation (μg.h.L^{-1})				AUCs for New preparation (μg.h.L^{-1})			
555	493	569	500	505	746	431	556
495	272	418	365	595	525	362	497
736	592	673	667	549	665	585	675
699	379	544	623	679	490	727	566
377	734	604	552	559	645	524	564
573	709	649	573	604	593	559	684
692	752	527	681	474	596	403	674
270	756	571	688	515	678	577	532
596	690	633	196	547	432	613	532
508	457	475	531	546	581	434	431
582	369	472	461	611	501	554	570
668	500	495	552	601	506	525	550
538	425	659	497	648	621	608	655
600	689	603	493	412	518	729	
686	697	588	522	483	537	586	
Mean = 557.8 ± 124.6 μg.h.L^{-1}				Mean = 563.2 ± 84.7 μg.h.L^{-1}			

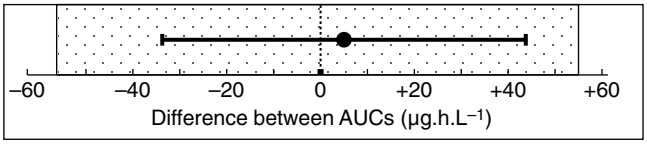


Figure 10.3 95% C.I. for difference between AUCs of two propranolol formulations (New – Old). Shaded area is the equivalence zone

It is decided that the new and old preparations are unlikely to show any clinically significant difference in effectiveness if we can be assured that their AUCs do not differ by more than 10%. Since the old preparation produced a mean AUC of 556 μg.h.l^{-1}, and 10% of this is approximately 55 μg.h.L^{-1}, the equivalence zone is set to cover a change of –55 to +55 μg.h.L^{-1}.

Next we generate a confidence interval by carrying out a two-sample *t*-test. The point estimate for the difference (Defined as AUCnew – AUCold) is 5.4 μg.h.L^{-1} with a 95% confidence interval for the difference between the two preparations of –33.6 to +44.3 μg.h.L^{-1}.

This is shown in Figure 10.3 along with the equivalence zone.

The whole of the 95% C.I. fits within the equivalence zone. So, we have positively demonstrated that if there is any difference between the two preparations, it is too small to be of practical consequence. The two formulations have been demonstrated to be effectively equivalent.

10.2.2 Difference versus equivalence testing. Same test – different interpretations

Notice that difference and equivalence testing use exactly the same statistical calculation. In both cases, a two-sample *t*-test is used to generate a 95% C.I. for difference. It's the interpretation that differs.

When testing for …

- Statistically significant difference: Is zero excluded from the C.I.?
- Practically significant difference: Is the whole of the C.I. outside the equivalence zone?
- Equivalence: Is the whole of the C.I. within the equivalence zone?

10.2.3 How *not* to test for equivalence

A common misapplication of statistics is to test for equivalence by performing a test for difference, obtaining a non-significant result and then claiming that because there is no evidence of any difference, the two things must be equivalent. The fallacy of this approach is immediately visible if we think about a possible result for our comparison of propranolol formulations. Figure 10.4 shows a case where the result of a test for difference would be declared statistically non-significant, since a zero difference is included within the 95% C.I. However, the formulations are not necessarily equivalent, since there is a real possibility that the new formulation may be producing AUC values of up to 80 μg.h.L^{-1} greater than the old.

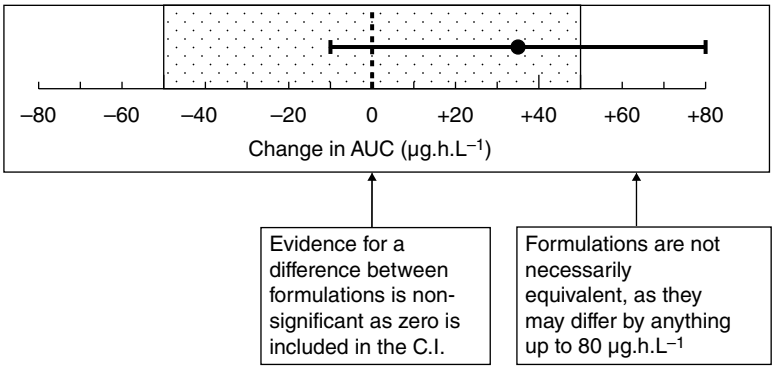


Figure 10.4 Results that should *not* be claimed as providing evidence of equivalence

Make that embarrassing difference disappear

If your target journal sends your paper to a statistically savvy reviewer, you won't get away with this one, but don't worry, many of them do not.

Your new treatment or analytical method (or whatever) is supposed to produce the same results as the old one, but the wretched thing obviously doesn't. Don't despair, just follow these simple instructions:

- Set up a comparison of the new versus the old using a small sample size that you know to be suitably underpowered.
- Carry out a statistical test for difference and get a non-significant result.
- Swear blind that as you didn't show a difference, the new system must produce the same results as the old one.
- Should this fail, simply repeat with a smaller and smaller sample sizes, until you force the difference to stop being statistically significant.

'Absence of evidence is not evidence of absence'

The failure to find a difference does not necessarily imply that none exists. Maybe you didn't look very hard.

10.3 Non-inferiority testing

10.3.1 Change in dialysis method

With equivalence testing there are two possibilities that we need to dispose of – the difference between outcomes might be either (a) greater than some upper limit or (b) less than a lower limit. The propranolol tablets provided an example of this. It would be potentially dangerous for the patient if the amount of drug absorbed suddenly increased or decreased and we have to exclude both possibilities. However, there are many cases where our only worry is that a particular value might (say) decrease. In that case, presumably an increase in the value would be perfectly acceptable. (We would either welcome an increase or at least be neutral about it.) The question now being asked is 'Is the new product/method and so on at least as good as the old one?' This is referred to as 'Non-inferiority testing'.

At least as good as ...

Non-inferiority testing is used to see whether a product/process is at least as good as some alternative.

Take, as an example, the use of dialysis in renally compromised patients. We have an established method, but want to evaluate a possible alternative which would be simpler and cheaper, if it is clinically acceptable. A key endpoint is plasma urea concentration. If changing the dialysis method were to lead to an increase in urea levels, this would be unacceptable. On the other hand, a decrease in urea levels would be a bonus, but is not essential. Clinical judgement is that a difference in urea levels of ±1 mM is the smallest change likely to cause a noticeable effect in the patient.

A series of possible outcomes from a comparison of the two methods are shown in Figure 10.5. In this case, the shaded area is not an equivalence zone as this is no longer relevant. What is now shown is the 'Failure zone'. Differences in this zone would be indicative of a rise in urea concentrations too great to be acceptable. The condition for acceptance of the new treatment is that no part of the confidence interval enters the failure zone starting at +1 mM.

Notice the difference. For equivalence testing we have to show that any change lies between an upper and a lower limit, but with non-inferiority testing we only have to show that there is no change beyond a single (upper or lower) limit.

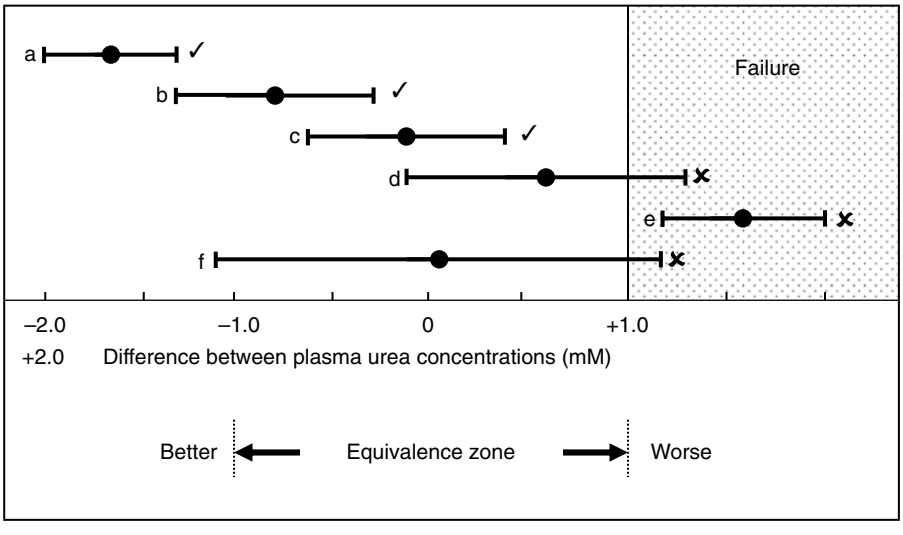


Figure 10.5 Interpretation of non-inferiority testing. Comparison of new dialysis method with old. Difference in plasma urea concentration (mM)

The top three cases (a–c) would all pass a test of non-inferiority, as they preclude an increase in plasma urea great enough to cause any appreciable deterioration in the patients' conditions. With the remaining three cases however we could not, at this stage, claim to have demonstrated non-inferiority. In all three cases a deterioration large enough to be of concern remains a possibility. Treatment (e) looks like a pretty hopeless case and could probably be discounted as a candidate. The other two – (d) and (f) – might be acceptable, but larger trials would be required in order to narrow the confidence intervals and then either (or both) of them might be established as non-inferior.

Non-inferiority testing

Can we demonstrate that the new product/procedure is at least as good as that with which it is being compared? No part of the confidence interval for the difference between the treatments should enter the territory indicating a practically significant deleterious change.

10.4 *P* values are less informative and can be positively misleading

All of the above procedures – demonstration that a change is large enough to be of practical significance or demonstration of equivalence or non-inferiority – are entirely dependent upon the use of the 95% C.I. for the size of the difference in treatment outcomes. *P* values would be pointless. (Unless we wanted to cheat!)

However, *P* values are not only less informative than the 95% C.I.; they can even be downright misleading, if taken in isolation. Table 10.2 gives the results of two trials of candidate antipyretic drugs (A and B). A beneficial effect would consist of a reduction in temperature and clinical judgement is that a reduction of at least 0.5 °C, would be required, if patients are likely to feel any benefit. It is also assumed that a rise of 0.5 °C would be detrimental.

On the basis of the *P* values alone, Drug A appears to be worth further development as it is produces a statistically significant effect, whereas there is no significant evidence that B produces any effect and the temptation would be to drop the latter. However, in Figure 10.6 we assess the same data using the 95% C.I.s and combine

Table 10.2 Results of two trials of potential antipyretics

Drug	Sample size	Difference between mean temperatures (°C)	*P*
A	200	–0.247	0.009 Sig
B	15	–0.537	0.074 Non-sig

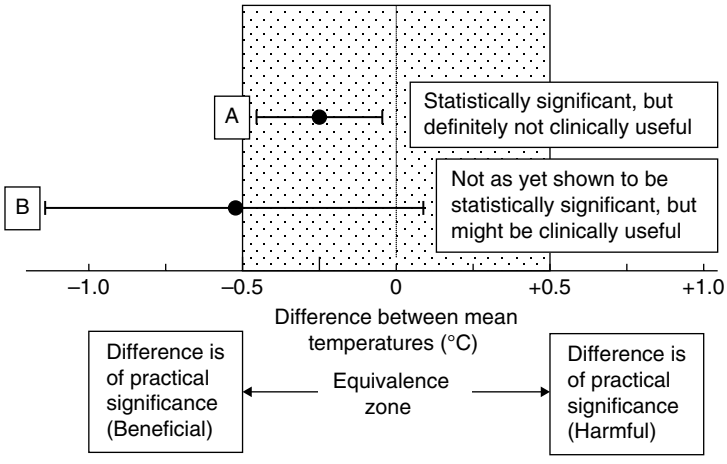


Figure 10.6 Correct interpretation of effectiveness of two possible antipyretic preparations A and B

these with our assumption that a practically significant temperature change would have to be of at least 0.5 °C. This suggests exactly the opposite conclusion. Drug A does produce an effect, but this is positively demonstrated to be too small to be of any clinical value. With B, the experiment is very small and the confidence interval correspondingly wide. Consequently the interval overlaps zero and the result is not statistically significant, however a large part of the confidence interval is in the area that would be of practical benefit. There is therefore still a reasonable chance that B may produce an effect, and furthermore this may be big enough to be useful. The excessively wide C.I. almost certainly arose because the sample size was much too small. Increasing the sample size would lead to a narrower interval which might be statistically significant. (This is not guaranteed, but is reasonably likely.) If we are going to expend money/effort on further development of either of these candidates, B is the better bet, not A.

The superiority of the 95% C.I.

To demonstrate practical superiority, equivalence or non-inferiority, we must inspect the 95% confidence interval for the difference between mean values The *P* value is completely irrelevant to all these questions and is potentially misleading.

The implementation of the two-sample *t*-test offered in Excel is very unfortunate as it provides a *P* value but no confidence interval for the difference.

10.5 Setting equivalence limits prior to experimentation

Equivalence limits should be set before experimental work begins. It is very difficult to decide objectively where equivalence limits should be placed, if you already know the experimental results and what the consequence would be of placing the limits at any particular level.

For the sake of your own credibility it is also very useful if you can register the decision with some independent party, prior to experimentation. That way, everybody else can be assured that the placing of the limits has not been inappropriately influenced.

Make the limits do the work

If you push this one too far, it'll stand out like a sore thumb, but within reason, you might get away with it.

Your new product/method …

- worked, but not very well or …
- is obviously different from the old one, when it should have been the same or …
- is annoyingly inferior to the old one.

However, there is still the question of where we're going to put those pesky equivalence limits. In the first case, we'll adjust them nice and close to zero, then even our pathetic effort will cause a 'practically significant' difference. For the latter two, we'll invent some reason to set them generously wide and hey presto, we are 'equivalent' or 'non-inferior'.

What you need is a journal that still hasn't twigged that they should be getting authors to declare these things in advance – that is pretty well all of them.

Anthony Grosso provides an interesting discussion of this problem in relation to anticoagulant trials (2009, *Brit. J. Clin. Pharmacy* **1**, 245–246).

Now read the appendix to Chapter 4

You have now met all the concepts that you need in order to understand the appendix to Chapter 4 which explains why you should be very wary of statistical tests that allegedly determine whether data follows a normal distribution.

10.6 Chapter summary

A demonstration of statistical significance provides evidence that the treatment being investigated does make a difference, but it tells us nothing about whether that effect is large enough to be of practical significance. Large experiments/trials are liable to detect trivially small effects as being statistically significant.

Upper and lower 'Equivalence limits' define the smallest increase or decrease that would of practical relevance. Establishing the numerical value of these limits is based solely upon expert opinion. The range of values between the two equivalence limits is the 'Equivalence zone'.

Demonstrating that a difference is of practical significance: To demonstrate that a treatment produces a change that is of practical significance, we need to show that the 95% C.I. for the difference lies entirely outside the equivalence zone.

Demonstrating equivalence: It is impossible to demonstrate that there is absolutely no difference between two procedures or treatments, but we may be able to show that there is no difference large enough to matter ('Equivalence testing'). We need to demonstrate that the whole of the 95% C.I. for the size of any difference lies within the equivalence zone. A non-significant result arising from a test for difference is not an adequate demonstration of equivalence.

Demonstrating non-inferiority: To demonstrate that a procedure or treatment is at least as good as an alternative (without needing to show that it is any better) requires 'Non-inferiority' testing. We calculate a 95% confidence interval for the difference between the two treatments. If this interval precludes any deterioration large enough to matter, we may claim evidence of non-inferiority.

Equivalence limits should always be fixed prior to undertaking experimental work and if possible the selected values should be registered at the same early stage.

11

The two-sample *t*-test (5): One-sided testing

This chapter will ...

- Introduce the use of one-sided tests where an experiment was designed to find out whether an endpoint changes in a specified direction.
- Describe the special form of null hypothesis used in one-sided tests.
- Describe the use of one-sided confidence intervals to test such null hypotheses.
- Show that, in marginal cases, data may be non-significant when assessed by a two-sided test, and yet be significant with the one-sided version of the same test.
- Review the age-old trick of switching from a two-sided to a one-sided test thereby converting disappointing non-significance to the much coveted significance.
- Show that said trick is not acceptable because it raises the risk of a false positive to 10% instead of the standard 5%
- Set out the correct protocol for using one-sided tests.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

11.1 Looking for a change in a specified direction

Sometimes, we may want to use a *t*-test in a way that differs from our previous approach. Say, for example, we are considering the use of urinary acidification to hasten the clearance of amphetamine from patients who had overdosed. An initial trial in rabbits is used to test the general principal. One group of rabbits would receive ammonium chloride to induce a lower urinary pH and another group would act as controls. All rabbits would receive a test dose of radio-labelled drug, the clearance of which would be studied over the next few hours. In this case, the question posed should be 'Is there an *increase* in clearance?' rather than the standard 'Is there a *difference* in clearance?' The former constitutes a one-sided question.

It often perplexes people that they cannot simply consider how an experiment was performed in order to tell whether a one- or two-sided test should be used. The fact is that the decision depends upon what question the experiment was designed to answer. If the purpose was to look for any old change – use a two-sided test. If your only interest lay in checking for a change in a particular, specified direction, a one-sided test may be appropriate.

One- and two-sided questions

Is there a *change* in clearance? Answer will be 'Yes', if there is either a marked increase or a marked decrease. This is a *two-sided* question.

Is there an *increase* in clearance? Answer will be 'Yes' only if there is a marked change in the specified direction. This is a *one-sided* question.

11.1.1 Null and alternative hypotheses for testing one-sided questions

If, with our amphetamine clearance experiment, we are going to ask the one-sided question, we need to modify our null and alternative hypotheses. For two-sided testing, the null hypothesis would be that there is no difference in clearance and the

Hypotheses for a one-sided test looking for an increase in amphetamine clearance

Null hypothesis: For a very large group (Population), the mean clearance among actively treated rabbits is either equal to or less than that in controls.
Alternative hypothesis: For a very large group (Population), the mean clearance among actively treated rabbits is greater than that in controls.

alternative, that there is. For the one-sided test (looking for greater clearance) we want our alternative hypothesis to be 'There is an *increase* in clearance'. The null hypothesis then has to cover all other possibilities – 'Clearance is either unchanged or reduced'.

In Section 6.8, we saw that we can generate a one-sided 95% confidence interval by calculating a 90% confidence interval and then using just one limit and ignoring the other. We can then say that there is only a 5% chance that the true mean value lies beyond the one limit that is being quoted.

In this case the null hypothesis is that the difference between the two groups is equal to or less than zero (see Figure 11.1). The parts of the graph that correspond to the null hypothesis (no change or a decrease) are shown shaded in the figure. To achieve statistical significance we need to show that the difference does not fall in this range. That will allow us to dismiss the null hypothesis. As can be seen in the figure, it is the position of the lower limit that settles the matter; the position of the upper limit is of no relevance to this question. So, when we are testing for an increase in an endpoint, we calculate the lower confidence limit. If we were testing for a reduction in the outcome, we would calculate the upper confidence limit.

The steps are thus:

1. Make a firm decision that we are only testing for evidence of an increase in clearance.

2. Calculate the mean clearance for both groups.

3. Generate a 90% C.I. for the difference between these mean clearances.

4. Only quote the confidence limit that acts as a test of the null hypothesis (in this case the lower limit).

5. Check whether this lower confidence limit is above zero. If it is, we have evidence of increased clearance.

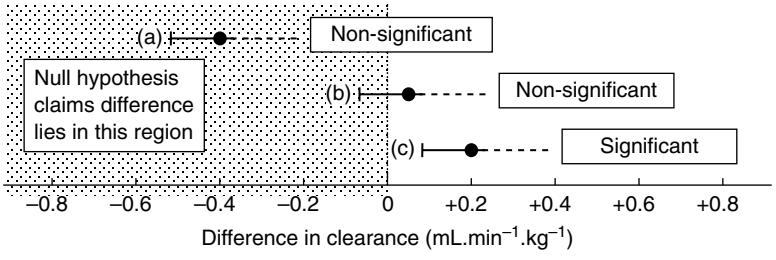


Figure 11.1 Interpretation of the C.I. for the difference between mean clearances, when performing a one-sided test for evidence of an increased value

Figure 11.1 shows how we would interpret the various possible outcomes of a one-sided test for an increase.

In Figure 11.1, only case (c) would be interpreted as a significant result, as it alone excludes all the territory claimed by the null hypothesis. With cases (a) and (b), a reduced or unchanged clearance is possible and the null hypothesis cannot be rejected. It might be thought that case (a) looks as if it means something significant, but the question being addressed is 'Is there evidence of an *increase* in clearance?' and case (a) in no way supports any such notion.

Performing one-sided tests

One-sided test for an *increase* in the measured value: Quote the lower confidence limit and declare the result significant if it is above zero. (A reduced/unchanged value has been successfully excluded.)
One-sided test for a *decrease* in the measured value: Quote the upper confidence limit and declare the result significant if it is below zero. (An increased/unchanged value has been successfully excluded.)

11.2 Protection against false positives

With all statistical tests one aim is to ensure that, where there is no relevant effect, we will make false positive claims on only 5% of occasions. Consider what would happen if there was actually no effect on clearance and we carried out 20 trials, each analysed by a one-sided test (testing for an increase). Bearing in mind that the actual procedure is to calculate a 90% C.I., but then convert it to what is effectively a 95% C.I. by ignoring one of its limits, we can predict the likely outcomes as in Figure 11.2.

- **18 times out of 20, the 90% confidence interval will span zero.** The lower limit is below zero, so the null hypothesis cannot be rejected. The result is correctly declared as non-significant.

- **1 time in 20, the entire interval will be below zero.** The lower limit is again below zero and the result is correctly interpreted as non-significant.

- **1 time in 20, the entire interval will be above zero.** The lower limit is now above zero and the result will be falsely interpreted as providing significant evidence of increased clearance.

Thus we have achieved our goal. If there is no real difference, there is only a 5% chance that we will falsely declare significant evidence of an increase in clearance.

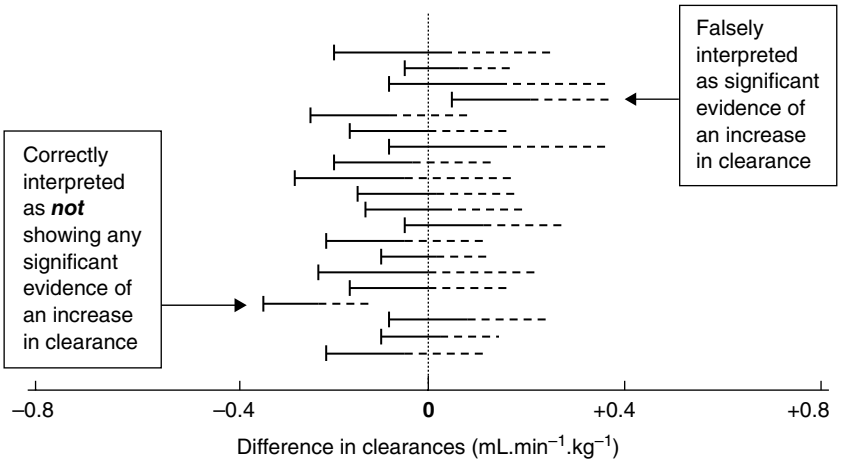


Figure 11.2 20 repetitions of one-sided 95% C.I.s for change in clearance when there is no real treatment effect (one-sided tests for evidence of an increase)

Protection from false positives

Where a treatment has no real effect, a properly conducted, one-sided test will expose us to the standard 5% risk of a false positive conclusion that there is an effect in the direction tested for.

11.3 Temptation!

11.3.1 Data may be significant with a one-sided test even though it was non-significant with a two-sided test

Table 11.1 shows a set of results for an investigation of the effects of urinary acidification on the clearance of amphetamine.

We could test this set of results using two different approaches:

- A one-sided test looking for increased clearance as discussed earlier.
- A normal two-sided test looking for any change in clearance.

The confidence limits for the difference are:

One-sided test (lower limit only)	+0.021
Two-sided test (lower and upper limits)	–0.031 to +0.582

Table 11.1 Clearance ($mL.min^{-1}.kg^{-1}$) of an amphetamine with and without urinary acidification

Control	Acidification
1.84	1.37
1.19	0.25
1.02	1.26
0.71	1.87
0.98	2.14
1.27	1.42
1.27	1.26
0.68	0.97
1.43	1.78
1.55	1.54
1.28	1.51
1.03	1.98
1.21	2.06
1.57	1.22
1.14	1.67
Mean 1.211	1.487
± S.D. ±0.312	± 0.482

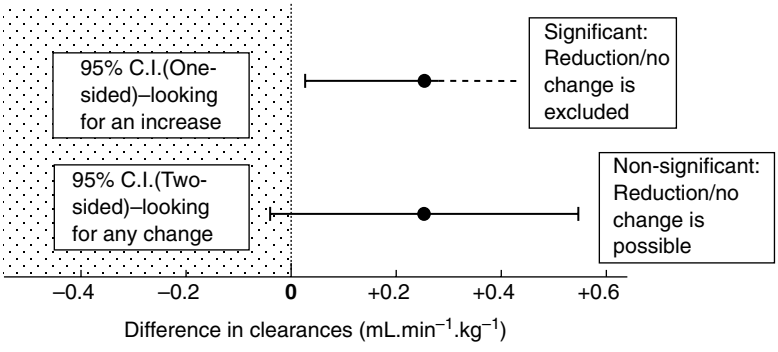


Figure 11.3 A one-sided test (for an increase) and a two-sided test (for any difference) in clearance

These are shown diagrammatically in Figure 11.3.

For the one-sided test, the confidence limit began life as a part of a 90% C.I., which is narrower than the 95% C.I. used for the two-sided test. In this case, the difference in width just happens to make a critical difference – The two-sided test isn't significant, but the one-sided test is.

Significance with a one-sided, but not a two-sided test

In marginal cases, results that are not significant when tested by a two-sided test, may just achieve statistical significance if re-tested with a one-sided test.

There are references above to both a significant and non-significant conclusion. However, if the question was whether there was an increase in clearance and there had been a prior decision to carry out a one-sided test, then the correct conclusion is that there is significant evidence of an increase. It should perhaps also be said that the results may be statistically significant, but the size of change is probably not of practical significance.

The *P* value for the one-sided test is 0.038, which supports our previous conclusion that it produced a significant result. The two-sided test would yield a *P* value exactly twice as large, that is 0.076 (Non-significant).

11.3.2 Possible cheat

This obviously raises the possibility of abuse. If we initially performed the experiment intending to carry out a two-sided test and obtained a non-significant result, we might then be tempted to change to the one-sided test in order to force it to be significant.

11.3.3 Chances of a false positive would be raised to 10%

The problem with such an approach is illustrated in Figure 11.4 in which we consider a case where there is no real experimental effect and we insist on looking at the results first and then performing a one-sided test in whatever direction suits our purposes. A 90% confidence interval is initially calculated, so there is a 10% likelihood that the interval will not include the true value of a zero change. This will be made up of a 5% risk that it will suggest an increase and another 5% that it will suggest a reduction. In this case, we have already seen the results and can guarantee that the one-sided test we carry out will be in the 'appropriate' direction to give a significant result. So, in total, we now have a 10% chance of a false positive.

Improper use of one-sided tests

If a one-sided test is improperly selected after the data has been seen, we raise the risk of a false positive from the stated level of 5% to a real level of 10%.

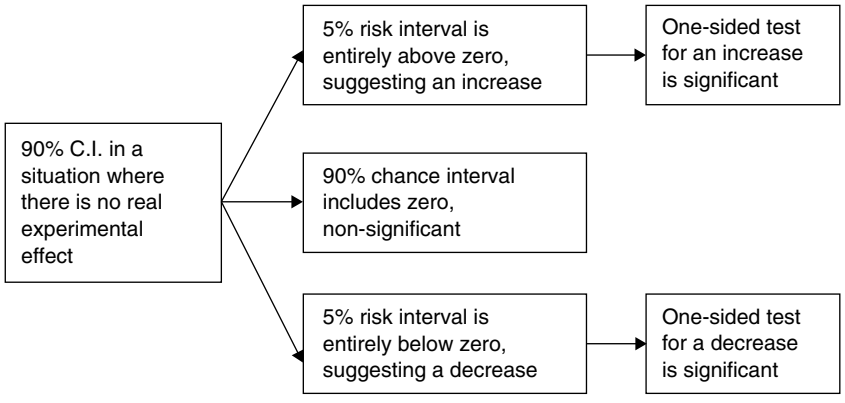


Figure 11.4 A 10% chance of a false positive outcome if we abuse one-sided testing by choosing the direction of testing after the results are known

11.3.4 One-sided tests must be conducted to a strict protocol if they are to be credible

A properly conducted one-sided procedure would involve us committing ourselves to the direction of testing in advance. If we had committed ourselves to testing only for (say) an increase in clearance then the 5% of cases apparently showing a reduction in clearance (lowest branch in Figure 11.4) would not be declared significant and our false positive rate would remain where it should be – 5%.

The fair way to perform one-sided tests

If a one-sided test is to be performed, the following steps must be taken *in the order shown:*

1. Decide that the test is to be one-sided and which direction of change will be tested for.
2. Perform the experiment.
3. Analyse the data according to the pre-specified plan.

Because the cheat outlined below is so easy to work, one-sided testing has fallen into considerable disrepute. That is a shame, because one-sided tests do have a perfectly respectable and useful role, if used appropriately. We can only hope that journal editors will, before too long, offer authors the opportunity to record our intentions in advance, to stop colleagues sniggering behind our backs when we use one-sided tests.

Switch to a one-sided test after seeing the results

Even today, this is probably the best and most commonly used statistical fiddle. Powerful – Undetectable – C'est magnifique!

You did the experiment and analysed the results by your usual two-sided test. The result fell just short of significance. There's a simple solution – guaranteed to work every time. Re-run the analysis, but change to a one-sided test, testing for a change in whatever direction you now know the results actually suggest. If *P* was originally anything between 0.05 and 0.1, this manoeuvre will exactly halve the *P* value and significance is assured.

Until scientific journals get their act into gear, and start insisting that authors register their intentions in advance, there is no way to detect this excellent fiddle. You just need some plausible reason why you 'always intended' to do a one-tailed test in this particular direction, and you're guaranteed to get away with it.

Table 11.2 Generic output from a one-sided two-sample *t*-test for *higher* clearances with urinary acidification

Two-sample *t*-test (One-sided option selected: Treated > Control)	
Mean (Treated)	1.487
Mean (Control)	1.211
Difference (Treated – Control)	+0.212
95% confidence limit for difference (Lower)	+0.021
P	0.038

11.4 Using a computer package to carry out a one-sided test

For clarity, the test has been described in terms of generating a 90% C.I. and using just one limit, but with most statistical packages the practical approach is to select an option for the direct performance of a one-sided test. You will also have to indicate the direction of change that is to be tested for. The programme will then generate a 95% confidence limit (either upper or lower, as requested) and a *P* value should also be produced. Typical output is shown in Table 11.2.

11.5 Chapter summary

One-sided testing can be used where the purpose of the experiment is to test for changes in one specified direction.

In the case of a one-sided test for an increased value, the null hypothesis is that the value is either unchanged or reduced and the alternative is that it is increased.

One-sided tests are conducted by generating the appropriate confidence limit for a one-sided 95% C.I. (This may be achieved by calculating a 90% C.I. and discarding one of the limits.) When testing for an increase, the lower confidence limit is used and vice versa.

If the confidence limit excludes the possibilities proposed by the null hypothesis, the outcome is statistically significant.

With a properly conducted one-sided test, the risk of an accidental false positive when investigating a treatment that has no real effect is held at the usual 5%.

Data may be non-significant with a two-sided test, and yet significant with a one-sided test. This can be abused to convert a non-significant finding to apparent significance. Such abuses raise the risk of false positives from 5 to 10%.

One-sided tests should only be used if a firm decision had already been made to do so and the direction of change for which we would test had also been decided upon, before the data was generated.

12

What does a statistically significant result really tell us?

This chapter will ...

- Demonstrate that a statistically significant result does not lead to a fixed level of confidence that there really is a difference in outcome between two treatments.
- Suggest that significance tells us that we need to increase our level confidence that there is a difference, but that what we end up believing depends upon the strength of the prior evidence.
- Show that treatments for which there is already a sound basis of support can be accepted as virtually proven, given new, statistically significant evidence.
- Suggest that intrinsically unlikely treatments should only be accepted if confirmed by several significant findings.

12.1 Interpreting statistical significance

Just how strong is the evidence when we obtain a statistically significant result? It has already been emphasised on a number of occasions that it is not proof absolute. Even when significance is demonstrated, there is still a residual possibility that we

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

could have obtained unrepresentative and therefore misleading samples. But it is not satisfactory simply to say that some doubt remains, people want to know how much doubt/certainty we are left with.

12.1.1 You cannot simply interpret a confidence interval as providing any set level of proof

There is a common misconception that because methods like the two-sample *t*-test are based on 95% confidence intervals, then 95% of all significant results are true positives and 5% are false. If only it were so simple! The following examples will show that sadly it isn't. We need to imagine two research workers. One works in early phase clinical trials and it is her job to test whether candidate drugs really do have pharmacological activity in humans. All the substances have already been extensively tested in other mammals and found to be active. Clearly there will be a few substances that are active in rats and dogs but not in humans, but in the great majority of cases they will also work in us. We will also consider somebody looking for pharmacological activity among traditional herbal remedies for which there is no existing scientific evidence of activity. Among such materials there will be a proportion that are effective, but it will probably be a small minority. Let's assume the following:

- Among the substances proven to be effective in other mammals, 90% are also genuinely effective in humans.
- Among the traditional remedies, 10% are genuinely effective in humans.
- Both researchers carry out statistical testing using the usual standard of significance (95% CIs for difference exclude zero or equivalently $P<0.05$).
- Both design their experiments to achieve 90% power.
- All products are either completely inactive or they have a level of activity that exactly matches the figure used to plan experimental size.

In Figure 12.1, 200 compounds that have already shown activity in other species should include 90% (180) that are truly active in humans and 10% (20) that lack activity. When the 180 genuinely active molecules are subjected to experiments with 90% power, 162 trials will successfully detect that activity but 18 will fail to do so. Among the 20 inactive compounds, there will be the usual 5% rate of false positives (One case) with the remaining 19 being correctly judged as non-significant. There are thus a total of 163 positive findings of which only one is false.

Figure 12.2 does the same analysis for 200 traditional herbal remedies. The big difference is that there are now 180 inactive products which throw up a far greater

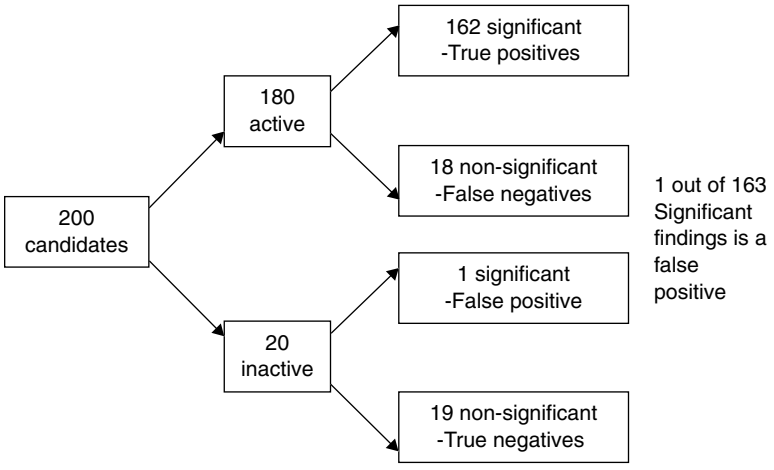


Figure 12.1 Investigating products where there is already a high level of evidence of activity

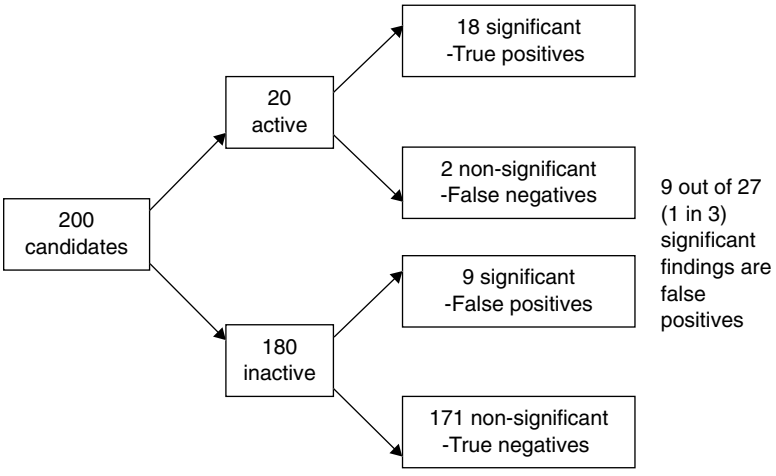


Figure 12.2 Investigating products where activity is unlikely

number (nine) of false positives. Among the total of 27 positive findings, fully one-third are false positives.

So, there is no simple answer to the question as to how much assurance a significant result actually provides. It depends upon what is usually referred to as the 'Prior likelihood'. With the situation depicted in Figure 12.1, any given candidate molecule is already known to have a high prior likelihood of activity and obtaining a significant result very nearly guarantees true activity (One chance in 163 that it is a false positive.)

However with the traditional remedies, even if a particular material does come up with a significant result, we still need to be very cautious – there remains a one-third chance that it is a false positive – not truly active.

Look at all the previously available evidence as well as today's *P* value

Two experiments may produce exactly the same *P* value, but that does not mean that they necessarily lead to the same level of certainty that there is a true difference in outcome.

- If previous evidence (or basic scientific principles) already suggests a difference is very likely, a significant result will give a high level of confidence that there is a true difference.
- If prior information suggests that a real difference is unlikely, even a significant result will still leave considerable doubt.

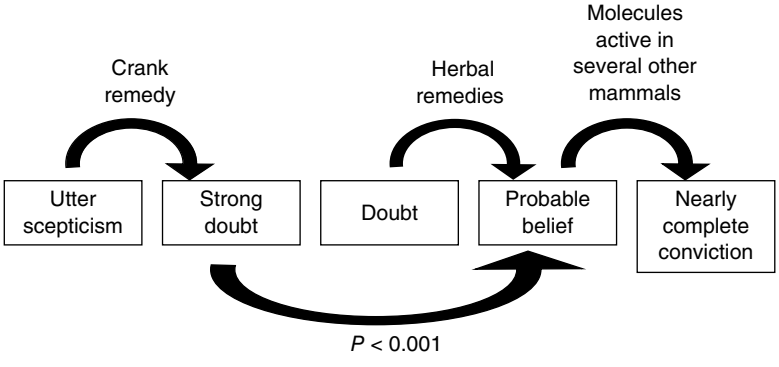


Figure 12.3 How a statistically significant result changes our beliefs about the existence of a treatment effect

12.1.2 Statistical results tell us how to modify our existing beliefs

The results of statistical tests are best interpreted in terms of how much they change our view rather than expecting them to tell us what we should end up believing. Figure 12.3 illustrates the idea. With a molecule already known to be active in other mammals we start out knowing that it is likely to be active in humans. A significant finding then boosts our faith in it to something approaching certainty (only one chance in 163 that it might ultimately be proved inactive). In contrast, with any

individual traditional remedy our starting point is that we doubt whether it will be active, but once it produces a significant result, our faith in it is boosted a stage. However, the odds that it is genuinely active are still only 2:1 on, so we could hardly say that we have anything near complete conviction – "Probable belief" is about as far as we could go.

This is where *P* values are of value. A marginally significant result might take us one step up the ladder of belief. But if we get a result where the *P* value is extremely small (usually reported as $P < 0.001$), then that will provide a greater boost to our confidence – strong doubt might be converted into probable belief.

Significant results

A statistical result does not tell us what we should believe. It tells us how much we should change what we believe.

Non-significant: Insufficient evidence to require any change
Significant ($P < 0.05$): Increase credence to a useful extent.
Highly significant ($P < 0.001$): Increase credence markedly.

12.2 Starting from extreme scepticism

Figure 12.3 suggests that in the case of a crank remedy, where there is no rational expectation of therapeutic effectiveness, a significant outcome would still leave us in strong doubts as to its activity.

12.2.1 Comparing unlikeliness

It might seem unfair that we could obtain a significant result and yet still remain unconvinced that a treatment actually works. However, statistical significance only tells us that the results we obtained would be unlikely to arise if the null hypothesis were true. That does not automatically mean that we should start believing the alternative hypothesis. What we should do is to compare the relative likelihood of the two hypotheses.

If a quack medicine produced a greater effect than a placebo control and $P = 0.01$, then …

- It is hard to believe the null hypothesis, because the results we obtained would be unlikely to arise on that basis (one chance in 100 that such results would arise by sheer chance). But …

- It is also hard to believe the alternative hypothesis, because it posits activity where none would logically be expected. (No exact figure is available but, for many of the more ludicrous 'treatments', there is probably far less than the one chance in a 100 seen above.)

Neither theory is attractive and all we can do is choose the one that is least unlikely; it is probably easier to believe that this was just a chance result. We remain unconvinced.

Inherently unlikely treatments

If an experimental assessment of a highly improbable treatment produced a statistically significant result in favour of activity, the rational conclusion would be that you still don't believe it works.

12.2.2 Demonstrating the truth of apparently highly unlikely theories

It might appear that we are reverting to good old-fashioned prejudice. If we already believe in something and we obtain a significant result, we continue to believe in it. But, if we were sceptical to start with, we remain sceptical even after a significant result! However it's not that bad; we are just acknowledging that we are not starting from a blank sheet of paper. A significant result always moves us up one notch of credence (as in Figure 12.3). Even with the crank remedy, a significant result would leave us less sceptical and if a series of well planned experiments continued to produce significant evidence of activity for this snake oil, then each of those trials would take our belief one notch up the scale. After three or four such trials we would eventually have to accept that the wretched stuff does actually work.

12.3 Bayesian statistics

The general approach used in this book is usually referred to as 'Frequentist'. There is an alternative (Antagonistic?) school of thought labelled 'Bayesian'. In its fully developed form Bayesian statistics uses a methodical and fully quantitative form of the less formal approach set out in this chapter.

There would be three steps to a Bayesian analysis for the effectiveness of a product:

- Determine the 'Prior Likelihood' that the product is effective. This is quantitative and is based on a scientific appraisal of the inherent likelihood that such a product would work and/or empirical evidence of its effectiveness that was already available before the current trial.

- Use the new evidence that has now become available from the trial to calculate how much we need to modify our current view.

- Use exact mathematical rules to combine the Prior Likelihood with the new evidence to produce a Posterior Likelihood. This is then a direct measure of the credibility that the product is effective.

The undoubted advantage of a Bayesian approach is that we do end up with a measure of how likely it is that the product will be effective. With a frequentist approach, we get a measure of the unlikeliness of the null hypothesis, given the actual results observed, but as the first part of the chapter points out that alone is no direct guide as to the credibility of the alternative hypothesis.

This book is not going to pursue the Bayesian approach any further. To do so would trigger a whole new book. There is also another problem; all the frequentist methods described in the book are available, right now 'Off the shelf', using reasonably friendly standard software. At the moment, Bayesian analyses tend to be more bespoke – tailored to each individual study. Maybe someday reasonably standardised Bayesian procedures may become more easily available, but that day has not arrived yet.

Bayesianism is undoubtedly philosophically superior to frequentism, we just need it implemented in a way that we mere mortals can use.

12.4 Chapter summary

The interpretation of statistical significance must involve not only looking at the *P* value for the current experiment, but also taking stock of the previously available evidence as to whether two treatments are likely to give differing outcomes.

In a situation where there is already a strong empirical or rational basis for anticipating a difference, a significant result will leave us almost convinced that there is a real difference.

Where there is less reason for anticipating an effect, a significant result will still increase our belief that there we are seeing a real difference, but some caution is still appropriate.

In a situation where any difference in outcome is wildly unlikely, but we obtain statistically significant evidence in favour of an effect, we should remain sceptical and see whether a difference is confirmed in further experiments.

A model is proposed whereby statistically significant results always increase the credibility of a treatment effect, but precisely how convinced we become depends upon our prior assessment of the likelihood of a difference in outcomes.

Bayesian statistics formalise the approach outlined above.

13

The paired *t*-test: Comparing two related sets of measurements

This chapter will ...

- Describe the difference between paired and unpaired data.
- Demonstrate that the paired *t*-test has greater power than the two-sample *t*-test when dealing with paired data.
- Explain the source of its superior power.
- Show how the paired *t*-test is performed.
- Explore the merits of paired versus unpaired experimental designs.

13.1 Paired data

There are many cases where we are faced with two columns of measured values and, as with examples in previous chapters, we want to see whether values are generally higher in one column than the other. Thus far, the situation is familiar. However, the data in the two columns may be related – they form natural pairs. In that case, the paired *t*-test provides a superior alternative to the two-sample *t*-test. An example follows.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

13.1.1 Does a weight-loss drug really work?

An oral drug allegedly causes weight loss. We recruit 30 subjects who have received medical advice that they should loose weight. They all receive two periods of treatment each lasting three months. In one period patients receive placebo tablets and during the other they are given the active product.

Each patient's weight is recorded at the end of both treatment periods.

The results are shown in Table 13.1.

Table 13.1 Effects of an alleged weight-reducing drug on subjects' weights

Subject number	Weight after placebo (kg)	Weight after active (kg)	Change in weight (active – placebo) (kg)
1	115.4	112.7	–2.7
2	118.9	113.8	–5.1
3	98.6	89.6	–9.0
4	108.3	93.1	–15.2
5	120.2	115.3	–4.9
6	115.0	112.3	–2.7
7	125.1	124.0	–1.1
8	120.4	115.6	–4.8
9	132.6	131.5	–1.1
10	100.8	98.9	–1.9
11	111.9	111.4	–0.5
12	105.3	103.0	–2.3
13	111.8	101.3	–10.5
14	98.0	91.3	–6.7
15	113.7	112.6	–1.1
16	117.1	118.2	+1.1
17	121.7	116.1	–5.6
18	123.6	127.5	+3.9
19	130.0	120.2	–9.8
20	128.0	117.3	–10.7
21	109.3	116.4	+7.1
22	117.7	113.8	–3.9
23	105.2	104.9	–0.3
24	120.3	123.4	+3.1
25	125.5	119.6	–5.9
26	114.3	106.4	–7.9
27	124.0	122.2	–1.8
28	123.3	123.8	+0.5
29	112.7	111.2	–1.5
30	113.8	106.6	–7.2
Mean	116.08	112.47	–3.62
SD	8.94	10.43	4.78

13.2 We could analyse the data by a two-sample *t*-test

We could analyse this data using the two-sample *t*-test. Comparing the mean weights at the end of the two treatment periods, we would obtain the following result:

Point estimate for difference between mean weights = –3.62 kg
95% C.I. for difference = –8.64 to +1.41 kg

The confidence interval includes zero, so the result is non-significant. The *P* value (0.155) confirms non-significance.

13.3 Using a paired *t*-test instead

13.3.1 Data variability

Table 13.1 also shows, for each subject, the difference between their two weighings. This has been calculated as weight after taking the active drug minus that after the placebo. In that way a negative figure indicates a weight loss.

Figure 13.1 shows (left panel) the weights at the end of both treatment periods. Each line connects an individual patient's post-placebo weight to that after active treatment. The vertical arrows emphasise the wide spread among these weights (placebo and actively treated weights cover ranges of approximately 30 and 40 kg respectively). However, the individual changes in weight (right panel) are markedly less spread out (a range of about 20 kg).

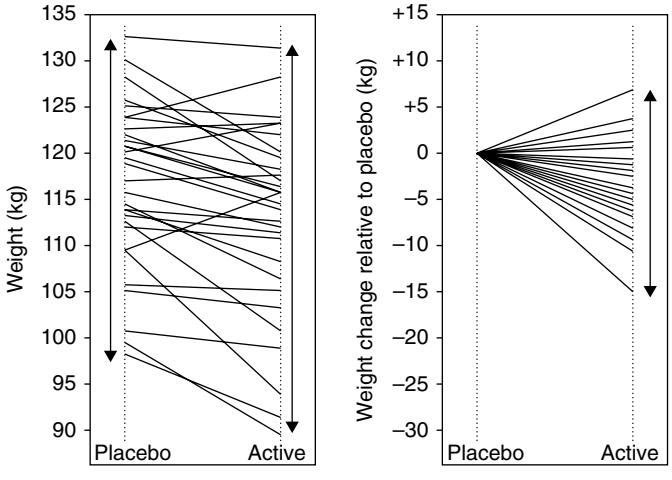


Figure 13.1 Greater variability among weights than among weight changes

The difference in variability reflects the fact that some individuals may be much bigger than others, but those who start largest generally end largest and the smallest end smallest. Consequently, the changes in weight do not show the extreme variation seen among the actual weights.

Changes less variable than the actual values

Where a series of individuals have widely differing values for an endpoint, we may yet find that a treatment induces a relatively constant change in all individuals. In such a case, the actual values will have large SDs but the SD among the changes will be smaller.

13.3.2 Reducing the variability we have to contend with

We know from Chapter 7 that data variability is a spoiler for *t*-tests and the large SDs for the weights in the first two columns of Table 13.1 are a prime contributor to the non-significant outcome of the two-sample *t*-test.

It would be attractive to be able to base a statistical test solely on the column of weight changes, as these are considerably less variable. This is exactly what the paired *t*-test does.

13.4 Performing a paired *t*-test

13.4.1 Null and alternative hypotheses for the paired *t*-test

The calculation of a paired t-test is based on the changes in weight. The actual weights are, of course, necessary in order to calculate the changes, but once this has been done, only the changes are used.

The null hypothesis for the paired *t*-test will claim that, for a large group of subjects ('the population'), the mean weight change would be zero. It accepts that some individuals gain and others lose weight on the drug treatment, but it assumes that in a large enough group, the gains and losses would cancel out. This implies that any apparent effect seen in our sample must be due to random sampling error.

The alternative hypothesis is that the drug does have an effect and that however large a group we looked at, we would continue to see the effect.

Null and alternative hypotheses for a paired *t*-test

Null: The mean weight change in a large population of subjects would be zero.
Alternative: The mean weight change in a large population of subjects would *not* be zero.

13.4.2 Calculation

To perform a paired *t*-test we simply calculate a 95% C.I. for the mean among the weight changes (using final column of Table 13.1) and check to see whether the interval includes the value posited by the null hypothesis (zero).

Most statistical packages include this test. The two sets of data are usually entered into separate columns, with paired valued lying side by side. You then indicate the two relevant columns. With most implementations, you will not need to calculate the individual changes – this should be carried out as part of the procedure. Generic output is shown in Table 13.2.

A useful diagrammatic way to present the data is shown in Figure 13.2. The dots show that a large majority of subjects experienced a negative weight change (only five gained weight) which informally suggests a significant effect. A zero mean change, as claimed by the null hypothesis, is clearly excluded by the 95% C.I., so the result is formally statistically significant. The *P* value (<0.001) confirms the significance.

Table 13.2 Generic output from a paired *t*-test comparing weights after active and placebo drug treatment

Paired *t*-test		
	n	Mean
Active	30	112.47
Placebo	30	116.08
Differences	30	–3.62
95% C.I. for mean difference:		–1.83 to –5.40
P:		0.000

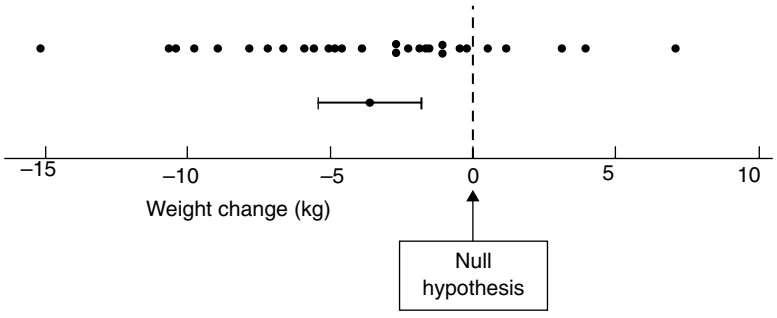


Figure 13.2 95% C.I. for the mean effect of the weight-loss drug

Paired *t*-test is just a special use of a 95% C.I. for the mean of a single set of values

To perform a paired *t*-test we simply calculate a 95% C.I. for mean individual change and check whether the interval includes zero.

13.4.3 Different order of calculation in a paired *t*-test

In the two-sample *t*-test the two steps were:

1. Calculation of the mean for each sample
2. Calculation of the 95% C.I. for the difference between the two means

For the paired *t*-test the order is reversed.

3. Calculation of the change that has occurred within each pair of results.
4. Calculation the 95 C.I. for the mean among these changes.

Order of calculation

Two-sample *t*-test: Means, then difference of the means
Paired *t*-test: Differences, then mean of the differences

13.5 What determines whether a paired *t*-test will be significant?

If you look at Figure 13.3, the outcome of the test will depend on two things:

- *How far from zero is the mean value for the individual changes?* If there is an approximate balance of positive and negative weight changes, the mean will be close to zero and the interval will probably overlap zero making significance unlikely. With large changes in a consistent direction, the interval will be displaced well away from zero and the result should be significant.

- *How wide is the interval?* Small samples and/or high variability among the individual weight changes will make for a wide interval that is likely to cross zero, robbing us of significance.

The logic is very similar to that for the two-sample *t*-test, with one crucial difference. For a two-sample *t*-test, the variability that would have to be considered is that among the two sets of weighings (the first two columns of data in Table 13.1). With a paired *t*-test, what matters is the variability among the individual weight changes (final column of Table 13.1). So, with a paired *t*-test, the initial observations might be hideously variable, but if all individuals show similar changes, we can still obtain statistical significance.

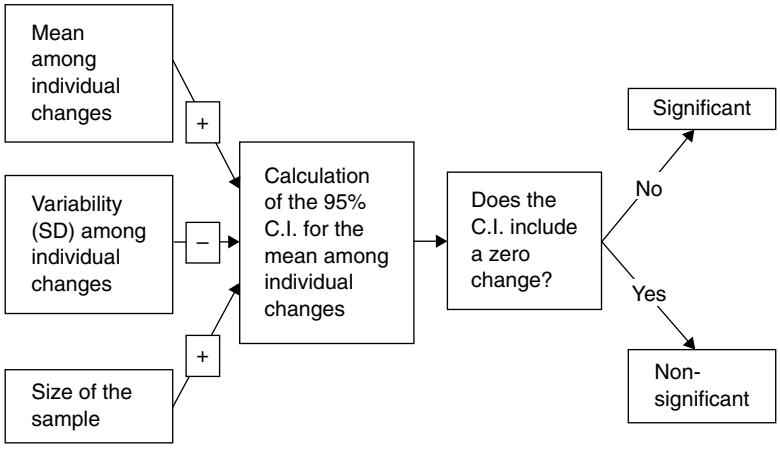


Figure 13.3 Factors influencing the outcome of a paired *t*-test

13.6 Greater power of the paired *t*-test

When we initially applied a two-sample *t*-test to the first two columns of data in Table 13.1 (Section 13.2), the result was non-significant ($P = 0.155$), but the result of the paired *t*-test was clearly significant ($P < 0.001$). This is a typical example of the greater power of the paired *t*-test.

Where data form natural pairs (as in the current example), it is very frequently the case that individuals who have the highest values under the first set of experimental conditions also have the highest values under the alternative conditions. As a result, the variability in the two original sets of data is greater than that among the individual changes that occur. This is the reason why the paired *t*-test is more powerful (often much more powerful) than the two-sample *t*-test.

13.7 Applicability of the test

13.7.1 The paired *t*-test is only applicable to naturally paired data

The whole logic of the paired *t*-test that we just performed was founded on the fact that we could calculate a change for each individual participant and then use those changes to calculate the rest of the test. With the theophylline/rifampicin experiment (Table 7.1), the data also consisted of two columns of data, but in that case there was no pairing. The first figure in the first column and that in the second column were derived from different individuals and it would have made no sense to start calculating the difference between these two figures and then move down to the second number in each column and so on.

It is perhaps worth noting that the enhanced power of the paired *t*-test is dependent upon the data being genuinely paired. The paired test has no superiority where data lacks natural pairing. So, if you have a set of data that is not genuinely paired and your two-sample *t*-test is non-significant, don't bother trying to concoct some artificial basis for claiming that the data is paired, just to allow a switch to the paired test. That particular manoeuvre is not only unjustified, but also ineffectual. (It deserves a pirate box with zero pirate flags.)

Greater power of the paired *t*-test

Where data is genuinely paired, the paired *t*-test is likely to be considerably more powerful than the two-sample *t*-test and should be employed.

If data is not naturally paired, the use of a paired test is unjustified (and entirely pointless).

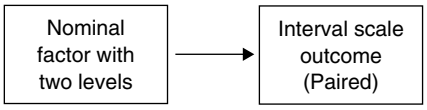


Figure 13.4 Diagrammatic representation of an experimental structure where use of the paired samples *t*-test is appropriate

13.7.2 Diagrammatic representation of the paired *t*-test

The structure of studies suitable for analysis by a paired *t*-test is shown in Figure 13.4. It is very similar to that for the two-sample *t*-test. The outcome would again be recorded as a measurement on an interval scale and there is just one factor being studied which has two levels – active and placebo. The sole difference is that the outcome data is collected in a paired pattern whereas Figure 7.2 indicated an unpaired pattern.

13.8 Choice of experimental design

Many experiments could be carried out either as paired or unpaired studies. For example the rifampicin/theophylline experiment (Table 7.1) was performed on an unpaired basis – 15 people got one treatment and a separate group of 15 received the other. This is referred to as a 'parallel groups' trial. We could have used a paired structure, with 15 subjects receiving one treatment on one occasion and the other treatment at some other time (a 'cross over' trial). The paired alternative would almost certainly have been a lot more powerful. However, it does not follow automatically that we should always be looking for a paired experimental design. The following points need to be borne in mind.

13.8.1 In favour of paired designs – greater power

The use of a paired design produces data that can be analysed by the more powerful paired *t*-test, whereas data from an unpaired experiment can only be analysed by the less powerful two-sample *t*-test.

13.8.2 Against paired designs

13.8.2.1 Greater practical difficulties In a paired design, each subject has to be studied twice. This may be slower to implement, especially if you need to leave a significant period of time between the two stages of the study. With human studies, there is also the problem that people may be less likely to volunteer, if they know they will be experimented upon twice instead of just once.

13.8.2.2 Greater problems in the case of data loss If we used an unpaired design with ten subjects in each group, we might find ourselves unable to obtain a measurement from one of our subjects. In that case, we would be left with ten observations in one column, and only nine in the other. With a two-sample *t*-test, we would still be able to use all of the data obtained. However, if we were performing a paired study with ten subjects (and presumably a paired *t*-test), we would not only lose that data point, but additionally its accompanying paired value would become useless and would have to be discarded.

13.8.2.3 The paired design may be logically impossible Some tests are destructive. For example, if we want to know whether a candidate drug causes liver enlargement in mice, we would need to compare organ weights after placebo and drug treatment. But, as the only practical way to determine liver weight is to kill the mouse and remove its liver, a paired experiment is going to be tricky.

Greater complexity of paired experimental designs

Paired experimental designs (which can be analysed by the more powerful paired *t*-test) may be more problematical than simple unpaired designs.

13.9 Requirement for applying a paired *t*-test

13.9.1 The column of individual changes should be consistent with a normal distribution

The paired *t*-test is just a special application of the 95% C.I. for the mean and the requirement for normally distributed data, described in Chapter 5, applies in this case too. Just be aware that the confidence interval is calculated using the column of individual changes in weight (etc.), so it is these that need to be normally distributed. The subjects' placebo and active treated weights could be skewed or bimodal or any other horrible distribution – that wouldn't matter. We just need the changes to be normal. As previously, we do not expect small samples to form perfect classic normal distributions, but if the column of individual changes shows extreme signs of non-normality, we should not trust a paired *t*-test. (See Chapter 21 for possible solutions.)

Requirement for performing a paired *t*-test

The individual changes must be consistent with a normal distribution. There is no need for the original observations to be normally distributed.

13.10 Sample sizes, practical significance and one-sided tests

Chapters 9–11 explained sample size calculations, practical significance and one-sided tests in detail and the point was made that these were general concepts that could be applied to a wide variety of tests. This section briefly reviews their application to the paired *t*-test.

13.10.1 Sample size calculations

The factors that influence necessary sample size for a paired *t*-test are very similar to those encountered with the two-sample test, but notice that we take into account variability among the individual changes in the endpoint, not the two sets of initial observations. Below, are reasonable values for the three relevant factors for our experiment on the effect of the weightloss drug.

- *Size of average change to be detectable:* Assume any weight change of less than 2 kg is of no medical or aesthetic relevance, so set the detection limit to this figure.

- *Variability in data:* Remember that it is the variability in the weight changes that matters, not the variability among the actual weights. We know that the sort of subject at whom the treatment is aimed, will be constantly adopting and abandoning all manner of diets, so apart from anticipated variability among responses to the drug, there is also likely to be a lot of apparently random background noise. We allow for an SD among weight changes of ±4 kg.

- *Power:* 90% power is considered adequate.

If we feed these values into any statistical package that includes sample size calculations (e.g. Minitab), we will be told that a sample of 44 is required. But remember that this is the amount of data we want to end up with and furthermore that paired designs can waste a lot of data if there are drop-outs, so realistically we need to start with 50 (or so) participants. This is another case where we were lucky to get away with what was in fact an underpowered experiment. Although

the numbers were inadequate, fortunately the drug caused a weight loss (3.62 kg) considerably greater than the minimum we wanted to be able to detect (2 kg) and significance was attained.

13.10.2 Practically significant change, equivalence and non-inferiority testing

In Chapter 10, we saw the use of the two-sample *t*-test to answer questions such as 'Is the change big enough to matter?' or 'Can we demonstrate that there is no change of any consequence?' or 'Is this new product/procedure at least as good as the old one?' The paired *t*-test generates a 95% C.I. for the treatment effect that can be combined with appropriate equivalence limits to answer precisely the same kinds of questions. The results are interpreted exactly as previously.

13.10.3 One-tailed testing

As with the two-sample *t*-test, there may be circumstances where the question we want to answer concerns possible changes in some pre-determined direction. For example, with our weight-change experiment, the question posed might very reasonably have been 'Does the drug lead to a *loss* of weight?', since that would presumably be the motivation for using the drug. That question would be one-sided. As usual, the test could be converted to one-sidedness either by calculating a 90% confidence interval and ignoring one limit or (if your statistical package allows) selecting a one-sided option. The usual rules apply – one-sided testing is fair enough so long as the decision to do so was made before the data was seen.

Table 13.3 Distinctions between the paired and two-sample *t*-tests

	Paired *t*-test	Two-sample *t*-test
Methodology	1. Calculate the difference for each pair of values. 2. Calculate C.I. for the mean of these differences.	1. Calculate the mean for each column. 2. Calculate C.I. for the difference between these two means.
Use with unpaired data?	Unjustified and pointless.	Correct procedure.
Use with paired data?	Correct procedure.	Possible, but a poor choice – lacks power.
Use if unequal numbers of observations in the two columns?	Impossible – data evidently not paired!	No problem, except slight loss of power.

13.11 Summarising the differences between paired and two-sample *t*-tests

Table 13.3 summarises the features that distinguish the paired from the two-sample *t*-test.

13.12 Chapter summary

The paired *t*-test is used where data form natural pairs. Its classic use arises when we've observed the same individual (human or otherwise) under two different circumstances (e.g. before versus after treatment).

For each individual we calculate the difference in the measured value, under the two circumstances. These individual changes are then used to calculate a 95% C.I. for the mean effect. If the interval excludes zero, the result is statistically significant.

The paired *t*-test offers the greatest advantage over the two-sample *t*-test when values are much higher in some individuals than in others, but all individuals show roughly the same change. In such cases, the two-sample test would be degraded by the extreme variation between individuals, but the paired test would only have to cope with the lesser variation among the individual changes.

It is always worth considering the use of a paired experimental design, as it will allow the use of the more powerful paired *t*-test. However, paired experiments can present practical difficulties that may outweigh this statistical advantage.

For a valid paired *t*-test, the individual changes in the measured parameter should be approximately normally distributed.

General statistical methods such as sample size estimation, determination of practical significance and one-sided testing can be applied to the paired *t*-test in the same manner that we have already seen for the two-sample *t*-test.

14

Analyses of variance: Going beyond *t*-tests

This chapter will ...

- Describe how experimental design is described in terms of 'Factors' and 'Levels'.
- Show the use of the one-way analysis of variance to analyse experiments where there is only one experimental factor, but it takes three or more levels.
- Describe the use of 'Follow-up tests' to discover which treatments differ from which others.
- Describe the two-way analysis of variance where two experimental factors have been investigated in parallel.
- Explain the concept of 'Interaction' between factors.
- Describe the distinction between 'Fixed' and 'Random' factors

14.1 Extending the complexity of experimental designs

With *t*-tests, we can compare two sets of measured (Interval scale) data. There are other experimental designs which will require a comparison of more than two sets of data and that is when we need an 'Analysis of Variance' ('AoV' or 'ANOVA').

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

Traditional statistics books always get their knickers in a frightful twist trying to explain ANOVAs. It is difficult to imagine why, because they are actually quite minor extensions of the two-sample *t*-test.

14.1.1 Factors and levels

A 'Factor' is something that either spontaneously varies or which we can manipulate as part of an experiment and which may alter the endpoint we are measuring. In the Rifampicin/Theophylline experiment (Chapter 7), the factor was Rifampicin. We then say that the factor has a number of 'Levels'. This is the number of different possibilities for that factor. There were only two levels for Rifampicin – It was either administered or withheld. In the weight-loss experiment in the previous chapter, there was again just one factor (Drug) and it also had two levels (Placebo or Active). In fact, for any experiment that can be analysed by a *t*-test there is always one experimental factor for which there are just two levels – the simplest of all experimental designs.

'Factors', 'levels' and analyses of variance

A 'Factor' is something that can take two or more categorical values. We want to know whether variation between these categorical values causes changes in the relevant measured outcome. Each different possibility within a factor is a 'Level'.

t-tests are used for the simplest possible experimental designs, that is a single factor with just two levels. Analyses of variance are used for any design of greater complexity.

14.2 One-way analysis of variance

14.2.1 Single experimental factor

In our first step up the ladder of complexity we will stick with just one experimental factor, but consider cases where that factor has more than two levels.

As an example, we want to improve the chemical synthesis of a drug. It is already known that finely divided platinum is a good catalyst for the reaction, but we want to investigate the potential use of other related metals. We therefore perform the synthesis of the drug under fixed conditions of pressure, temperature and so on, but vary the catalyst added. The metals that we wish to investigate are Platinum, Palladium, Iridium and Rhodium and an alloy of Palladium and Iridium. Here we have one factor (Which catalyst is being used) which has five levels (Five different metals).

Table 14.1 Effect of catalyst on yield (percentage of theoretical maximum)

	Platinum	Palladium	Iridium	Palladium/ Iridium	Rhodium
	11.3	15.4	12.1	13.1	12.0
	10.7	17.0	12.2	13.7	11.6
	9.8	18.4	13.1	13.5	9.1
	10.4	17.5	11.8	14.0	11.9
	11.5	18.8	10.4	14.2	11.3
Mean	10.74	17.42	11.92	13.70	11.18
SD	0.69	1.34	0.98	0.43	1.20

Five replicate syntheses are carried out using each catalyst. The yields obtained (expressed as percentages of the theoretically achievable yield) are shown in Table 14.1.

14.2.2 Do not perform multiple *t*-tests

A seductively simple way to analyse the data would be to say that Platinum is our current standard catalyst, so it makes sense to treat Platinum as a control and compare it against each alternative in turn, using a series of two-sample *t*-tests. However, this would involve performing four separate tests. We know that for every statistical test performed, there is a 5% risk of a false positive. With a simple experiment like the weight-loss study (Chapter 13), only one test is performed and the chance of a false positive is an acceptable 5%. However, in this case we would be performing four tests each carrying a 5% risk and the chances of hitting a false positive at some stage would be a lot more than 5%. (The risk is not quite as bad as the simplistic $4 \times 5\% = 20\%$, but it is well in excess of 15%.) The general term for such repeated analyses with their attendant risk of false positives is 'Multiple testing' and it's a problem that can raise its ugly head in a wide range of statistical scenarios. (Chapter 23 considers other forms of multiple testing.)

Multiple testing

When results from several treatments are to be compared, multiple *t*-tests should not be used – increased risk of false positives.

Instead we need a single test that will consider the whole data set and deliver a single verdict. The appropriate test is called the 'One-way analysis of variance'. The 'One-way' part of the name reflects the fact that there is still only one factor under investigation (In this case, which catalyst to use).

One-way analysis of variance

Is used when:

- The endpoint is a measured value (generally on an interval scale).
- There is one experimental factor.
- The factor has three or more levels.

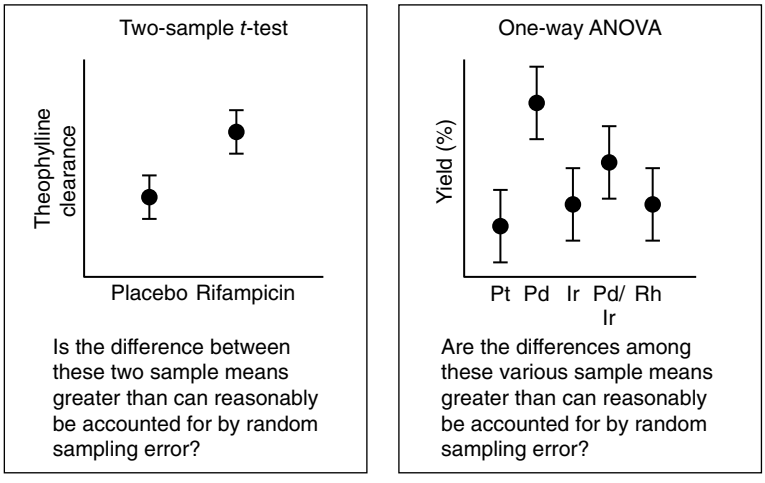


Figure 14.1 The one-way analysis of variance as a minor extension of the two-sample *t*-test

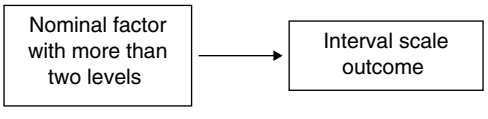


Figure 14.2 Diagrammatic representation of experimental designs where the use of a one-way analysis of variance is appropriate

Figure 14.1 shows that the application of a one-way ANOVA to this data is only a minor extension of what we already did with the rifampicin/theophylline clearances and a two-sample *t*-test.

Figure 14.2 provides a diagrammatic representation of when the test is appropriate. Notice the similarity with that for the two-sample *t*-test (Figure 7.2) apart from the increase in the number of levels.

14.2.3 Null and alternative hypotheses

The null hypothesis must deny that any of the catalysts differs from any of the others.

Null and alternative hypotheses

Null: In the long term, the mean yield of product would be the same for all five catalysts.
Alternative: In the long term, at least one of the catalysts would produce a different mean yield from one of the others.

The mechanism assumed by the null hypothesis is also just an extension of that discussed in relation to the *t*-test. It is assumed that the five catalysts are in reality indistinguishable, but within these small samples, random sampling error has led to an illusion of variability in their effectiveness. Presumably, the effectiveness of some catalysts have been overestimated and/or that of others understated.

14.2.4 What governs whether a significant result will emerge?

Since the function of the one-way ANOVA is so similar to the of the two-sample *t*-test, it's no great surprise that the aspects that govern its outcome are also virtually identical. They are shown in Figure 14.3:

In considering this diagram, we need to keep a close eye on the term 'Variation'. There are two types:

14.2.4.1 Intra-group variation The five individual results listed under 'Platinum' in Table 14.1 were all gathered under conditions that were intended to be, as far as possible, identical. However, each time we repeat the experiment, some variation always creeps in. Variation within a group of replicates is called 'Intra-group variation'.

14.2.4.2 Inter-group variation There are also overall differences between groups as reflected by the mean yields for each type of catalyst. This is termed 'Inter-group variation'. There will always be a degree of inter-group variation, even if all the catalysts are exactly equally effective, because of random sampling error. However, if there are real differences in the effectiveness of the various catalysts, inter-group variation will be boosted beyond the level we would expect to arise from random sampling error alone.

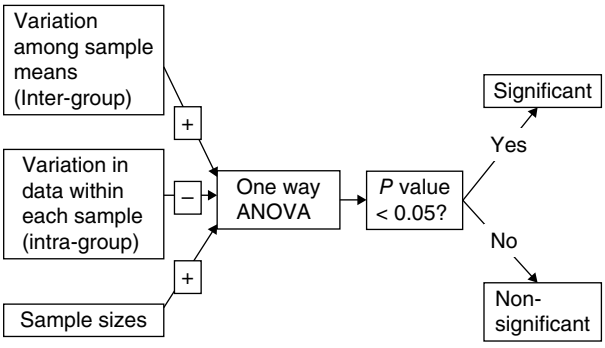


Figure 14.3 Aspects of the data influencing the outcome of a one-way ANOVA

14.2.4.3 Plus and minus signs in Figure 14.3 The plus and minus signs in Figure 14.3 reflect the effect of each aspect on the likelihood of a significant outcome and are assigned as follows:

- Differences between the sample means (Inter-group variation): Small differences may only reflect random sampling error, but if the sample means differ widely, a significant conclusion is more likely.

- Variability in the data within samples (Intra-group variation): If the data within each sample is highly variable, random sampling error will be increased and we will be less convinced that any apparent differences between catalysts are real.

- Sample sizes: Large samples should produce less random error and we will be more confident that any inter-group variation is due to real differences among the catalysts.

14.2.5 Performing a one-way analysis of variance

With most statistics packages, data that is to be subjected to a one-way analysis of variance is entered into two columns in a similar way to that seen with a two-sample *t*-test (Section 7.1.4). One column contains a series of codes indicating what catalyst was used and the other column contains the corresponding experimental results. In the first five rows, the results are labelled as being due to the use of Platinum (Pt), the next five are due to Palladium (Pd) and so on. The general appearance will be as in Table 14.2.

To carry out the test, you will need to indicate which column contains the codes and which the actual data. The associated website (www.ljmu.ac.uk/pbs/rowestats/) gives details of how to use SPSS or Minitab to perform all the tasks described in this chapter. The output varies considerably between packages, but almost all will provide that shown in Table 14.3, as a minimum:

Table 14.2 Generalised method for entering data into statistics packages in preparation for a one-way analysis of variance

Column 1 Catalyst	Column 2 Yield
Pt	11.3
Pt	10.7
Pt	9.8
Pt	10.4
Pt	11.5
Pd	15.4
Pd	17.0
Pd	18.4
Pd	17.5
Pd	18.8
Ir	12.1
Ir	12.2
etc.	etc.

Table 14.3 Generic output for one-way analysis of variance of effect of catalyst on reaction efficiency

One-way analysis of variance

Endpoint: Yield
Factor: Catalyst

Source	DF	SS	MS	*F*	*P*
Catalyst	4	148.062	37.016	38.37	0.000
Error	20	19.296	0.965		
Total	24	67.358			

The reporting of analyses of variance is traditionally an ugly business, including a mass of intermediate working that is of limited value in most circumstances. So, if we ignore all the dross, the *P* value is given as '0.000' (i.e. <0.001) which is statistically, clearly significant.

14.2.6 Follow up tests – Interpreting significance

A *P* value well below 0.05 means that we have strong evidence against the null hypothesis. However, even if we are happy to exclude the null hypothesis, Figure 14.4 shows that we are still left with a wide range of possible alternatives. It could be [see part (b) of Figure 14.4] that just one catalyst is different and all the others are indistinguishable, or (c) they might break up into two groups, or (d) they might all be

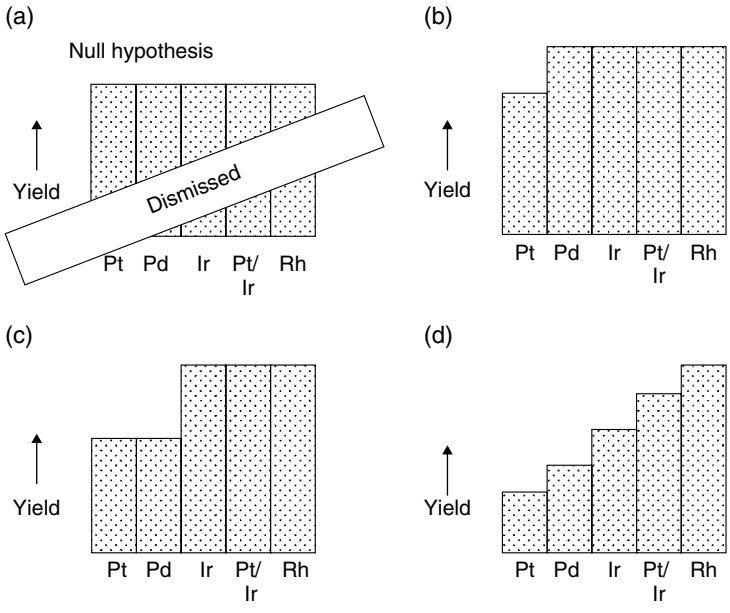


Figure 14.4 Interpreting a significant result from a one-way ANOVA

different from each other, and so on. We are left with several dozen possibilities. The significant result from the ANOVA is rather frustrating – it tells us there are some differences in there somewhere, but it's not telling us where they lie.

The means and SDs for each catalyst (Table 14.1), suggest that the main features are palladium producing higher yields than any other metal, with palladium/iridium a good second and little to choose between the others. However, we really need a more objective assessment and this is where 'Follow-up' tests come in. They are called 'Follow-up' tests because traditionally they are only applied after an ANOVA has proved significant, although there is no strict need to follow that sequence. There are innumerable follow up tests available, but the two outlined below will cope with most situations.

14.2.6.1 Dunnett's test With Dunnett's test, we select one of the treatments as a 'Control' or 'Reference' group. All other groups are then compared against the control. In our case we might declare Platinum to be the control and then compare all other treatments against it – a total of four comparisons.

14.2.6.2 Tukey's test Here, all groups are compared to all other groups in every possible pairing. With our data set, we would have to compare the first treatment against four others, the next against three others and so on, giving a total of 4+3+2+1=10 comparisons.

Table 14.4 Generic output from Tukey's test, showing confidence limits for the differences between pairs of catalysts (percentage points)

Tukey's test

Confidence Intervals: 99.28%

	Lower limit for difference	Upper limit for difference	Sig?
Palladium – Palladium/ Iridium	+1.86	+5.58	Sig
Palladium – Iridium	+3.64	+7.36	Sig
Palladium – Rhodium	+4.38	+8.10	Sig
Palladium – Platinum	+4.82	+8.54	Sig
Palladium/ Iridium – Iridium	–0.08	+3.64	NS
Palladium/ Iridium – Rhodium	+0.66	+4.38	Sig
Palladium/ Iridium – Platinum	+1.10	+4.82	Sig
Iridium – Rhodium	–1.12	+2.60	NS
Iridium – Platinum	–0.68	+3.04	NS
Rhodium – Platinum	–1.42	+2.30	NS

14.2.6.3 Performing Tukey's test With our experiment, a case could be made for performing either a Dunnett's or a Tukey's test. The general rule, that choices of statistical methodology should be made in advance of seeing the data, includes the selection of a follow up test. Let's assume that a decision had been made, that we would use Tukey's test.

In most statistical packages, the implementation of the analysis of variance includes an option to select a Tukey's test. The format of the output varies enormously, but (as in Table 14.4) should include a list of confidence intervals for the difference between each possible pair of catalysts. Each line of output shows the difference calculated as the yield with the first metal minus that with the second. The results are shown ordered according to yield: from Palladium (highest) to Platinum (lowest).

The output begins with Palladium and it is contrasted with all four other catalysts. In each case the figures are positive (Palladium is superior) and the C.I. for the difference excludes zero, so all comparisons are statistically significant. Palladium is demonstrably superior to all the competition, even its nearest rival – Palladium/ Iridium.

Next, there is a block comparing Palladium/Iridium to all other catalysts except palladium (already considered). In this case, the gap between Palladium/Iridium and its nearest rival (Iridium) is just a little too small to be statistically significant, however the alloy was shown to be superior to the other two metals.

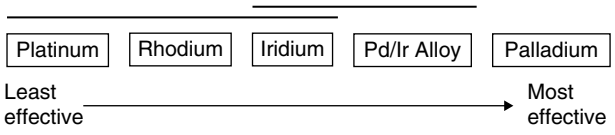


Figure 14.5 Groups of catalysts that are not statistically distinguishable

For all the remaining comparisons among Iridium, Rhodium and Platinum the confidence interval includes zero and they are non-significant.

Initially, this seems to be a confusing jumble of information, but Figure 14.5 should make sense of it. The metals are ordered according to their effectiveness (This time going from least to most effective). For the first three metals (Pt, Rh and Ir), all comparisons among them were non-significant, so a horizontal bar indicates that these three metals are not statistically distinguishable. Another bar shows our failure to demonstrate a significant difference between Iridium and the alloy, but notice that neither of these bars covers all four metals, because the contrast between Platinum and Rhodium at one extreme and the Pd/Ir alloy at the other was great enough to be significant. Palladium is not covered by any bars, as it is distinguishable from all other metals.

14.2.6.4 Practical significance The ANOVA only tested statistical significance. However, Tukey's test reports confidence intervals for the sizes of the various differences, so we can also assess whether any increase in yield that might be achieved by a change of catalyst would be big enough to be of practical significance.

Dunnett's and Tukey's follow up tests

The Analysis of Variance only tells us whether there are any differences among the available treatments. It does not tell us either:

- Which treatment differs from which other or
- The extent of the difference between any pair of treatments.

Follow-up tests rectify both of these shortcomings.
Dunnett's: Compares one control group against all others
Tukey's: Compares all groups against all others.

14.2.6.5 Haven't we just resorted to multiple testing? Earlier (Section 14.2.2) we rehearsed arguments suggesting that it would be wrong to carry out multiple t-tests, as this would increase the risk of false positives. Doesn't the same objection apply to

Tukey's test, since it incorporates ten individual comparisons? The answer is No. Tukey's test is designed to produce a 'Test-wide error rate' of 5%. What this means is that there is a total risk of 5% that we might produce one (or more) false positive findings. This is achieved, by having each individual comparison performed to a higher standard of proof, guaranteeing that any one comparison will carry much less than a 5% risk. By the time we've performed all ten comparisons, we will have accumulated a total risk of 5%. In many statistical packages, the output for Tukey's test includes a statement of the level of confidence used for each individual comparison. With our catalyst experiment, 99.28% confidence intervals are used (See top of Table 14.4). The chances of a false positive arising from any given contrast is therefore 100 – 99.28 = 0.72%. By the time we have performed all ten comparisons, the total risk accumulates to the standard (and acceptable) 5%.

14.2.6.6 Performing Dunnett's test In the real world, it would be naughty to start doing a Dunnett's test at this stage, since we have already seen the data and had previously committed ourselves to Tukey's test. However, just so you can see the procedure, we'll perform Dunnett's test with Platinum as control.

If you select the option for Dunnett's test in your statistical package, you will additionally have to indicate that platinum is to act as the control. The generic output (Table 14.5) shows the difference of each metal in turn from Platinum.

Platinum is demonstrated to be inferior to Palladium and the alloy, but there is no statistically significant evidence of a difference from Rhodium or Iridium.

If your package indicates the level of confidence used it should be 98.47%. This means that the individual error rate for each comparison is 1.53%. This is higher than the corresponding figure (0.72%) for the Tukey's test. The reason for this is that we are now performing only four comparisons instead of ten and so individual comparisons do not need to be performed to quite such high standards. When planning experiments, it is worth remembering that as the number of treatments increases,

Table 14.5 Generic output from Dunnett's test, showing confidence limits for the differences between platinum and the other catalysts (percentage points)

Dunnett's test

Control: Platinum
Confidence Intervals: 98.47%

	Lower limit for difference	Upper limit for difference	Sig?
Rhodium – Platinum	–1.21	+2.09	NS
Iridium – Platinum	–0.47	+2.83	NS
Palladium/ Iridium – Platinum	+1.31	+4.61	Sig
Palladium – Platinum	+5.03	+8.33	Sig

there is a very steep increase in the number of comparisons that Tukey's test would make and the individual comparisons will have to be carried out to correspondingly high standards of proof.

Test-wide error rate remains at 5%

When Tukey's or Dunnett's test make multiple comparisons, each of these is performed to a higher than normal standard of proof, so that the cumulative risk of generating any false positives remains at the usual 5%.

14.2.7 Balanced data

The data from our catalyst experiment is 'Balanced', that is there are exactly equal numbers of replicates for each catalyst. This is not a requirement for the one-way ANOVA and if a small amount of data loss occurs, the analysis can still go ahead. For any given number of observations, the power of the ANOVA will be greatest with a balanced data set and this is also true for Tukey's test. The only circumstance where power will be greater with an imbalanced data set is where a Dunnett's test is planned. Here, the control group is of special importance because it is used in all the comparisons. For this test, it is worthwhile trying to generate some extra data for the control group.

14.2.8 Requirements for performing analyses of variance

The data requirements are similar to those for the two-sample *t*-test. Each set of data should be drawn from a normally distributed population and they should all have the same SD. Small samples never exactly fit these requirements, but gross deviations from the ideal can cause real difficulties. In severe cases, the data needs to be transformed to normality (Section 6.10) or an alternative method such as the Kruskal–Wallis test can be substituted for the one-way ANOVA (Chapter 21).

14.3 Two-way analysis of variance

14.3.1 Investigating two experimental factors simultaneously

The final step up the ladder of complexity is the simultaneous consideration of more than one experimental factor. With our drug synthesis experiment, where we considered different catalysts, we might also want to vary the method of mixing during the reaction. In the experiment presented in Table 14.1, mixing was achieved by stirring, but we suspect that this may be inadequate and that ultrasonication might

Table 14.6 Factors and levels to be considered

Factor	Levels
Catalyst – five levels	Pt, Pd, Ir, Pd/Ir, Rh
Mixing – two levels	Stirring, Ultrasonication

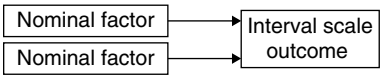


Figure 14.6 Diagrammatic representation of the type of experimental design where a two-way analysis of variance would be applied

break up aggregated material more effectively. Table 14.6 shows the factors and levels that we will now consider.

Our experiment will then investigate all possible combinations of these two factors, that is $5 \times 2 = 10$ combinations. When we use all combinations of two (or more) factors it is called a 'Full factorial experiment'.

The appropriate statistical test for two experimental factors and a full factorial design is the two-way analysis of variance. Figure 14.6 provides a diagrammatic representation of when the two-way analysis of variance is used.

Two-way analysis of variance

Is used when:

- The endpoint is a measured value (generally on an interval scale).
- There are two experimental factors.
- All possible combinations of the two factors have been studied (full factorial experiment).

We used five replicates of each combination of the two factors and the results are shown in Table 14.7

14.3.2 Interaction

When we analyse the data, we will obviously be looking to see whether altering the catalyst or altering the mixing method changes the yield. However, with more than one factor, we will also need to check for something new – interaction. This is most easily explained by looking at the results of our experiment. We start by calculating the mean result for each set of five replicates and these are shown in Table 14.8

Table 14.7 Effects of catalyst and mixing method on yield (percentage of theoretical maximum)

	Platinum	Palladium	Iridium	Palladium/ Iridium	Rhodium
	11.1	15.9	10.9	14.1	13.0
	11.7	18.9	13.4	13.9	12.3
Stirred	9.8	18.9	10.5	14.2	11.0
	12.1	15.5	12.2	13.5	9.9
	9.3	17.2	10.9	14.0	10.6
	12.8	20.0	14.8	16.3	13.1
	13.8	19.9	11.9	17.2	13.4
Ultrasonic	12.3	19.5	15.3	16.0	13.8
	12.4	20.5	14.3	15.1	13.6
	12.0	18.3	15.2	17.0	12.8

Table 14.8 Mean yield for each combination of catalyst and stirring method (%)

	Pt	Pd	Ir	Pd/Ir	Rh
Stirred	10.80	17.28	11.58	13.94	11.36
Ultrasonic	12.66	19.64	14.30	16.32	13.34

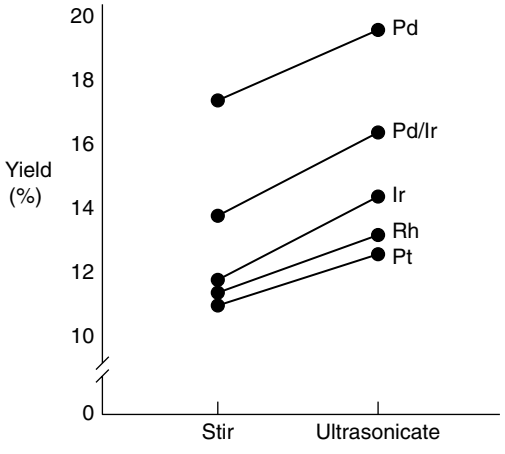


Figure 14.7 Effect of catalyst and mixing method on yield

These can then be presented graphically – see Figure 14.7.

Changing the mixing method does seem to have an effect on the yield; for each metal, the yield is always higher with ultrasonication than with stirring. Furthermore, the increase in yield when we switch to ultrasonication is fairly consistent for all five

metals: an increase of around 2–3%. Because the increase in yield is about the same for all five metals, we say that there the effects of the catalyst and stirring are simply 'Additive' and there is no interaction between the two factors. If ultrasonication caused a much greater increase in yield with some metals than with others, interaction would be present.

Interaction

Interaction is present when the effect produced by changing the level of one factor is dependent upon the level of another factor.

14.3.2.1 Graphical check for interaction We can make a visual check for interaction by looking for parallel lines in a graph like Figure 14.7. Two possible outcomes are shown in Figure 14.8. In the left-hand graph, the boost in yield that results from the use of ultrasonication is exactly the same for all metals (Parallel lines – No interaction), but on the right hand side, there is only a tiny change with Platinum compared to that seen with the Palladium/Iridium alloy (Non-parallel lines – Interaction).

The points plotted in Figure 14.7 are of course only sample means and they are subject to the usual random sampling error. We cannot arrive at any definitive conclusions just by looking at such a graph. If the points for ultrasonication were only slightly higher than those for stirring, the difference could be nothing more than sampling error. Similarly, slight non-parallelism, may not be real interaction, but just more sampling error. We need a two-way analysis of variance to see whether such features within the data are sufficiently clear cut to convince us they're real.

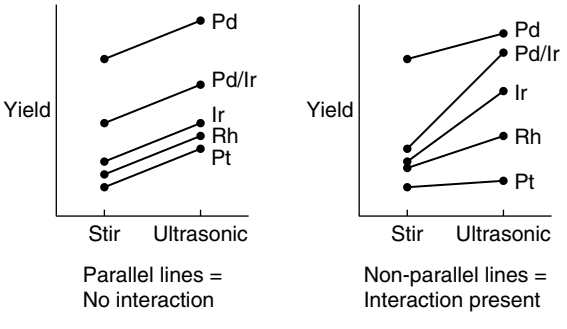


Figure 14.8 Visual check for interaction

14.3.3 Null hypotheses

In this case, we will be testing three null hypotheses:

- The long-term mean yield would be the same with all five catalysts.
- The long-term mean yield would be the same with both mixing methods.
- There is no interaction between the two factors.

These are independent hypotheses. We could end up accepting them all, rejecting them all or accepting some and rejecting others.

14.3.4 Performing a two-way ANOVA

We enter the data into a stats package in a similar way to that seen in Table 14.2, except that we now need an additional column to contain codes for the mixing method. (The additional column can be seen in Table 14.9.)

The detailed method for conducting a two-way analysis of variance varies from one stats package to another, but with all packages the output, should include all the usual ANOVA rubbish, along with the three *P* values that we really want (Table 14.10).

Table 14.9 Generalised method for entering data into statistics packages in preparation for a two-way analysis of variance

Column 1 Catalyst	Column 2 Mixing	Column 3 Yield
Pt	Stir	11.1
Pt	Stir	11.7
Pt	Stir	9.8
Pt	Stir	12.1
Pt	Stir	9.3
Pd	Stir	15.9
Pd	Stir	18.9
.	.	.
.	.	.
Pt	Ultra	12.8
Pt	Ultra	13.8
Pt	Ultra	12.3
Pt	Ultra	12.4
Pt	Ultra	12.0
Pd	Ultra	20.0
Pd	Ultra	19.9
etc.	etc.	etc.

Table 14.10 Generic output for two-way analysis of variance of effect of catalyst and mixing method on reaction efficiency

Two-way analysis of variance
Endpoint: Yield
Fixed factors(s): Catalyst, Mixing
Random factor(s):

Source	DF	SS	MS	*F*	*P*
Catalyst	4	300.931	75.2327	67.41	0.000
Mixing	1	63.845	63.8450	57.20	0.000
Interaction	4	1.186	0.2965	0.27	0.898
Error	40	44.644	1.1161		
Total	49	410.606			

Table 14.11 Increases in yield due to the use of ultrasonication instead of stirring (%)

Metal	Increase in yield (%)
Pt	1.86
Pd	2.36
Ir	2.72
Pd/Ir	2.38
Rh	1.98
Mean	2.26%

14.3.5 Interpretation of a two way ANOVA in the absenceof interaction

The three rows to look at are those beginning 'Catalyst', 'Mixing' and 'Interaction'. The first two give the results for the effects of changing the catalyst and for changing the mixing method. The third row reports on the possibility of interaction between these two factors. Catalyst and mixing method are confirmed as clearly significant ($P < 0.001$), but a P value of 0.898 provides no evidence of interaction and we can reasonably go ahead with some fairly sweeping and simple interpretations of the effects of both factors.

So far as mixing is concerned, we can say that ultrasonication is always superior to stirring, whatever catalyst is being used. Also, since there is no evidence of interaction, we can produce a single estimate of the extent of the improvement we achieve by using ultrasonication. From Table 14.8 we can calculate the gain with Platinum as 12.66 – 10.80 = 1.86% and so on. See Table 14.11.

As there is no significant evidence of interaction, it is reasonable to assume that all the figures in Table 14.11 are error prone estimates of one common figure. We will therefore take an average of all of them and our estimate is that the extent of the superiority of ultrasonication over stirring is an increase in yield of about 2.3%.

The results for the catalysts essentially confirm what we already knew, with palladium being the most effective.

14.3.6 Balanced data

Like the single factor experiment (Table 14.1), this experiment is also balanced – there were equal numbers of replicates for each combination of factors. The classic two-way ANOVA does demand a balanced data set. It can therefore be a pain if a piece of data is lost. However, there is a technique called a General Linear Model that will achieve the same ends as the two-way ANOVA, without the need for perfectly balanced data. Many statistical packages offer the technique.

Summary

- There is no significant evidence of interaction between the two factors ($P = 0.898$).
- There is significant evidence that yield varies according to the mixing method used ($P < 0.001$). Whatever metal is used, ultrasonication produces a yield around 2.3% greater than that achieved by stirring.
- There is significant evidence confirming our previous conclusion that yield varies according to which catalyst is used ($P < 0.001$).

14.3.7 Two-way ANOVA with interaction

In the next example we investigated the effects of punch design and compression force on the tensile strength of tablets made from hydroxymethylpropylcellulose (HPMC). Tablets are produced using low or high compression force (10 or 20 kN) and three different designs for the metal punches that form the tablets. A full factorial design therefore required six combinations of the two factors. Each combination was studied using six replicates. The results are shown in Table 14.12 and Figure 14.9.

Table 14.12 Effects of punch design and compression force on tablet strength (MPa)

	Punch A		Punch B		Punch C	
	4.00		6.49		7.92	
	4.79		8.35		8.15	
Low force	5.50	Mean	6.40	Mean	7.60	Mean
	5.15	5.208	6.37	7.092	9.64	8.753
	5.26		6.86		9.98	
	6.55		8.08		9.23	
	4.90		9.63		14.60	
	4.90		9.67		13.81	
High force	5.04	Mean	8.72	Mean	13.20	Mean
	6.93	5.787	8.08	9.075	14.62	13.863
	7.13		10.25		13.77	
	5.82		8.10		13.18	

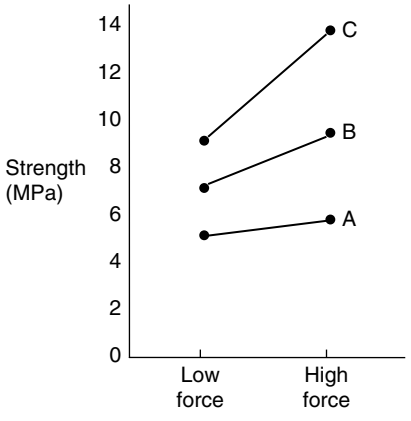


Figure 14.9 Effect of punch design (punches A, B, C) and compression force on tablet strength – quantitative interaction

Figure 14.9 strongly suggests interaction, with the higher compression force always increasing tablet strength, but its advantage over the lower force is fairly marginal with punch A and much more marked with punch C.

These results were analysed by a two-way ANOVA and the main results are shown in Table 14.13.

The statistical analysis confirms the presence of interaction ($P < 0.001$). We will therefore need to be more circumspect in our interpretation. We can still rank the various levels of the factors. Strength is always greatest with the use of the higher force and it always increases as we go from Punch A to B and then C. What we can no longer do is to make any across the board assessment of the extent of these superiorities. For example using high compression force increases

Table 14.13 Generic output for two-way analysis of variance of effect of punch designs and compression force on tablet strength

Two-way analysis of variance

Endpoint: Strength
Fixed factors(s): Force, Punch
Random factor(s):

Source	DF	SS	MS	*F*	*P*
Force	1	58.854	58.854	74.21	0.000
Punch	2	203.412	101.706	128.23	0.000
Interaction	2	32.286	16.143	20.35	0.000
Error	30	23.794	0.793		
Total	35	318.346			

Summary

- There is significant evidence of interaction between the two factors ($P < 0.001$).
- There is significant evidence that strength varies according to compression force ($P < 0.001$). High pressure always produces the greater strength.
- There is significant evidence that strength varies according to punch design ($P < 0.001$). Punch C produces the greatest strength and punch A the least.
- It is not possible to make any general statement about how much extra strength arises from the use of the higher compression force nor about the extent of the difference that arises from using (say) punch C instead of B.
- This is an example of quantitative interaction.

strength by only about 0.5 MPa with punch A, but the increase is in excess of 5 MPa with punch C. This type of interaction, where the effect of changing from one treatment to another always produces a change in one direction, but the extent of the change is variable is called 'Quantitative interaction'.

14.3.8 Quantitative versus qualitative interaction

In the tableting example there was interaction and as a result we could no longer make any general statement about the extent of the increase in tablet strength brought about by changing the compression force. However, we were at least able to comment on the direction of change – greater force always caused greater strength to some degree.

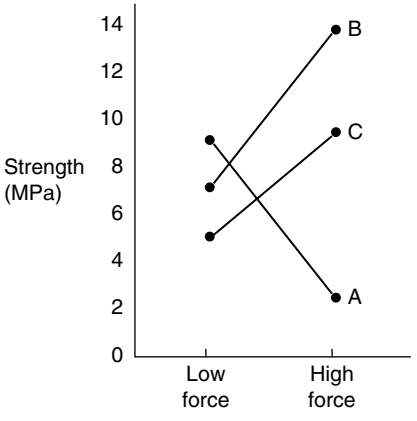


Figure 14.10 Qualitative interaction between punch design and compression force

The real nightmare is the sort of hypothetical result suggested in Figure 14.10.

We now have two punch designs (B and C) where increased pressure leads to stronger tablets, but punch A responds in exactly the opposite manner. The (very) non-parallel lines again indicate interaction, but now in a much nastier form. With this sort of messy outcome, we can't even produce any across-the-board comment on the direction of change that will arise from increased compression pressure, let alone the size of the change.

If what we saw in Figure 14.9 was 'Quantitative interaction' then Figure 14.10 may be said to show 'Qualitative interaction'.

Interpreting results in the presence of interaction

Type of interaction	General statement of the *direction* of difference when changing from one treatment to another?	General statement of the *size* of difference when changing from one tre atment to another?
None	Yes	Yes
Quantitative	Yes	No
Qualitative	No	No

14.3.9 Diagnosing the nature of interaction

If evidence of interaction is significant, a graph such as Figure 14.9 or 14.10 will illustrate the nature of the interaction (Quantitative or Qualitative). The *P* value will only tell you whether interaction is present, not what form it takes.

14.4 Fixed and random factors

14.4.1 Two contrasting scenarios

14.4.1.1 First scenario: Test of a generalised theory A rare eye condition is treated with an expensive ointment dispensed in tubes containing five grams of product. The amount being consumed seems excessive and there is a suspicion that patients are just squeezing the tubes in the middle, leaving much of the medicine in the lower end of the container (as many do with toothpaste). This is confirmed when a couple of patients return some of their 'empty' containers.

An alternative design of tube is available which is claimed to be less wasteful, although it is not immediately obvious that this would be so. An experiment is designed to test whether there should be a general policy to change to the alternative tubes. Three patients are randomly selected from a national register of sufferers and asked to return their used tubes. Initially they continue using the original design of tube, but then switch to the alternative. All returned tubes are analysed for residual ointment.

The results, including mean wastage for each patient, are shown in Table 14.14 and the mean values are also represented in Figure 14.11.

It is fairly clear that wastage for Patients A and C declined when they switched to the new tubes. For B, the outcome is less clear cut (Slight reduction or no change?).

Using common sense, what would we conclude from these results? Note that the question to be answered is:

Should there be a general switch to the new tubes?

We need to keep in mind that these three randomly selected patients are intended to represent a wider population of patients. Our null hypothesis will be that within the general population, the overall effect of switching to the new container would be no change in wastage. The assumption being that the new design of tube would cause some patients to increase and others decrease their wastage and that overall, these gains and losses would cancel out. It would also have to be assumed that our sample happened to over-represent those patients who showed reduced wastage. Figure 14.12 shows this diagrammatically.

If the effects of changing tube design were as suggested in Figure 14.12, it would not be especially unlikely that a random sample might happen to include two individuals showing reductions and one who was little changed. On the flimsy basis of just three patients, we really cannot be confident that a general switch to the new tubes would necessarily improve matters.

14.4.1.2 Second scenario: Test of a much more limited theory A clinic has just three patients with the condition mentioned previously. This clinic has also noticed apparent wastage and wants to know whether *their three patients* should switch to the new tubes. The clinic carries out essentially the same experiment as that described above, but using the three patients under their care. The results are exactly the same as those in Table 14.14.

Table 14.14 Ointment wastage (g) with old and new designs of tube

		Patient A	Patient B	Patient C
Old Design		1.92	1.93	2.26
		1.74	1.52	2.53
		2.24	2.17	2.16
		2.74	2.44	2.27
		1.80	1.68	2.40
			1.91	2.29
			1.76	2.09
			2.20	
	Mean	2.088	1.950	2.284
New Design		1.95	1.42	0.62
		2.35	1.54	1.13
		1.14	1.77	1.61
		1.32	1.81	1.11
		1.04	1.77	1.68
		1.69	1.89	0.53
		1.16	1.59	1.09
			1.93	1.21
			2.38	1.24
	Mean	1.521	1.790	1.135

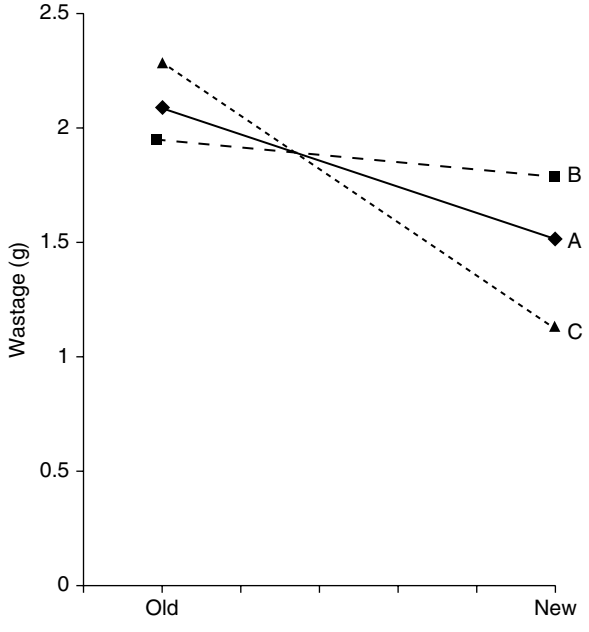


Figure 14.11 Ointment wastage (g) when dispensed in Old or New designs of tube for three patients: A, B and C

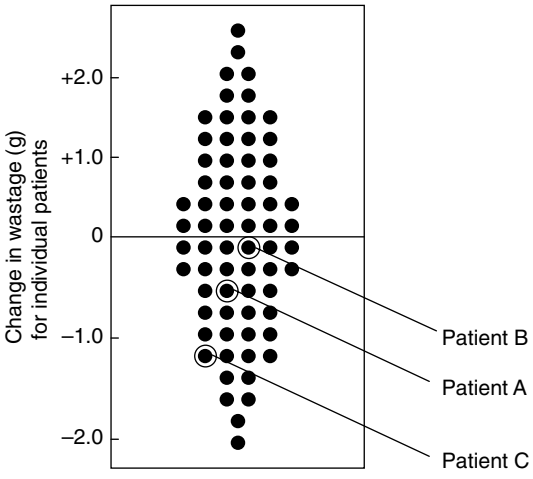


Figure 14.12 Assumed mechanism underlying the null hypothesis when considering whether there is any generalised advantage in changing tube design for ointment

Using common sense, what would we conclude from these results? In this case the question being asked is:

*Should this clinic, **treating these three specific patients**, switch to the new tubes?*

Our null hypothesis is now that there would be no change in average wastage if these three patients switched tubes. We are no longer using the three subjects to represent a wider population; we are specifically and exclusively interested in these three. Two of the patients showed fairly clear reductions in wastage and the other certainly showed no signs of increased wastage. That is all we need to know; this clinic should change to the new tubes. If other patients would not benefit, that is no concern in the present scenario.

14.4.2 Formal statistical approach for the two scenarios

14.4.2.1 Distinguish fixed from random factors In both scenarios the endpoint is wastage which is a measured (Interval) value and there are two factors that may influence the extent of wastage:

- The individual patients
- The different designs of tube.

The natural statistical test would therefore be a two-way analysis of variance with an interaction term. The method of calculating analyses of variance can be modified to suit the two scenarios considered.

The key is to recognise that factors may be 'fixed' or 'random'. With fixed factors, the levels studied are pre-determined and represent the full set of possibilities that

are of interest, whereas with a random factor, the levels are a random selection from some wider population. The distinction will probably be clearer in the context of our two scenarios.

- *Tube design:* In both scenarios, the design of tube is a fixed factor. The two designs studied were the two in which we were specifically interested. They were not some random selection from a catalogue of products.

- *Patients:* In the first scenario, the patients studied were a random selection and therefore constitute a random factor. However, in the second situation, the three patients were precisely the ones we have to consider; they were not randomly selected from any wider group. In this case the patients constitute a fixed factor.

If you are still unclear about the distinction, ask yourself the following question: If the results of your experiment were lost in a disastrous fire and you had to repeat the work, would you use the same levels of the factor as previously?

- In both scenarios we would use the same designs of tube as previously – a fixed factor.

- In the first scenario, Patients A, B and C were just a random choice and there is absolutely no reason to use them again. Any randomly chosen subjects will do, so this is a random factor. In the second case, these patients are the only ones in our clinic and we would definitely use them again – now a fixed factor.

Random and Fixed factors

- Random – you selected subjects from a wider group. Any randomly selected patients are appropriate.

- Fixed – you knew in advance who/what you wanted to study. You were specifically interested in Patients A, B and C.

14.4.2.2 Using statistical packages to perform the analyses Most statistical packages make a default assumption that all factors are fixed and then there is a mechanism to allow factors to be declared as random when appropriate. Our common-sense analysis where the three patients were randomly selected suggested that the results did not convincingly demonstrate that a generalised switching of all such patients to the new container would necessarily reduce ointment wastage. Happily, the formal analysis treating the patients as a random factor would produce a non-significant P value of 0.174 for the design of tube. The common sense assessment and formal analysis are consistent.

In contrast, if the patients are treated as a fixed factor, as is appropriate for the second scenario, we would obtain a clearly significant P value of <0.001 for tube design. Common sense suggested that the data were convincing in this context and the formal analysis supports this.

14.4.2.3 One- and two-way analyses of variance It is only with two- (or more) way analyses of variance where we really need to worry about whether factors are fixed or random. For experimental designs studying only one factor, that factor may be fixed or random, but this will not affect statistical significance.

When does it actually matter?

- The distinction between fixed and random factors is relevant for two- (or more) way analyses of variance. Statistical significance will not depend upon this distinction in one-way analyses.

14.4.3 How do the outcomes differ for the two statistical approaches?

14.4.3.1 Greater power when analyses treat factors as fixed It is generally the case that a two-way ANOVA is more likely to produce a significant result when both factors are analysed as fixed than when one or both are treated as random. Note that if we change the way in which we treat one factor, it is the other factor that shows a change in significance. In our examples, changing our opinion as to whether the patients constitute a fixed or random factor, changes the significance of the tube design.

The reason why the form of the analysis affects the chances of a significant result are obvious enough if we consider our two scenarios.

In the first scenario, there were two levels of random sampling; we randomly selected three patients and then randomly selected a set of used tubes from each of those individuals. Each level of random sampling is subject to sampling error. (The three patients may not be exactly representative of the whole population of patients and each set of tubes may not perfectly represent wastage by that particular patient.) Both of these sources of imprecision have to be taken into account in the analysis, reducing the chances of statistical significance. In this case, the very small sample of patients ($n = 3$) is hopelessly lacking in precision.

In the second scenario, there is only the one level of random sampling. (The random set of used tubes from each patient.) There is no longer any additional sampling error from the selection of patients. With only the one source of random sampling error, the result is more likely to be judged significant.

14.4.3.2 Greater generalisability when analyses have treated factors as random If the patients have been treated as a random factor and the result is statistically significant, then we can generalise our conclusion. We could state that, if we swapped to the new tubes, not only would we expect to see reduced ointment wastage in the specific patients studied, but we could legitimately generalise the claim; the new tubes should reduce wastage in any group of patients.

However, if we treated the patients as a fixed factor and achieved statistical significance, then we can only legitimately claim that changing tube design would reduce wastage among the specific subjects studied. We can no longer extrapolate the conclusion to patients in general.

Investigators who do not appreciate the characteristics of fixed and random factors are in danger of drawing sweeping conclusions that go way beyond anything justified by the data. Because most stats packages default to an assumption that factors are fixed, the danger is that a factor will be analysed unthinkingly as fixed, but then the outcome will be interpreted in an unjustifiably general manner. Typically, we might take the analysis from the second scenario with its clearly significant result and then claim that we have proved that patients *in general* would benefit, instead of restricting ourselves to a conclusion about these three specific individuals.

Generalise (when you shouldn't)

Given the widespread ignorance of the distinction between fixed and random factors, you've got a fair chance of getting away with this one, but if you push it too far, it will be obvious that something is not quite right.

You have multiple observations from a small number of individuals or institutions. Analyse the results treating the people/institutions as a fixed factor to increase your chances of a significant result and then insinuate that your finding applies to all patients/institutions rather than restricting your conclusion to the specific cases studied.

14.4.4 Other examples

The examples of two-way analyses of variance quoted earlier in this book are restricted to fixed factors. Section 14.3 described the effects of different types of catalyst and methods of mixing on the efficiency of a chemical synthesis. The catalysts investigated were not randomly chosen. (We could have made a random selection of metals by throwing darts at a copy of the periodic table, but it would make little sense!) Only a limited number of metals were ever likely to be any use and these were precisely the ones studied. Similarly, the methods of mixing were the only two practical alternatives, not a random selection.

In the first scenario in this section, the patients were randomly selected. Individual subjects frequently constitute a random factor. It is only in rare instances (such as the second scenario) that we want to investigate specific individuals; it is much more likely that individual subjects/patients are selected to represent a larger population.

Another factor that is likely to be random is a sample of institutions. If you are going to study the effect of a change in some procedure in (say) hospitals, you may study several. If these are selected randomly to represent all such institutions, they must be analysed as a random factor. If they are not declared as random then a significant outcome should only lead to a claim that the change in procedure had an effect in those particular hospitals rather than in all hospitals.

14.5 Multi-factorial experiments

There is no logical reason why we need to stop at two experimental factors. When trying to optimise a drug synthesis we might need to define the best:

- Catalyst
- Mixing method
- Concentrations of reactants
- Temperature
- Pressure
- Reaction time.

A full factorial experiment could then involve hundreds of combinations and be completely unrealistic. Fortunately there are compromise designs called 'Fractional factorial designs' that can be used so long as we can reasonably assume that interactions will not be excessive. Statistical analyses are then available to tease out which factors really are significant and to determine what combination of levels produces the optimum outcome. The design and analysis of such experiments goes well beyond the scope of this book, but many statistical packages provides all the methodology for rational optimisation (Minitab is especially good at this) – all you need is a competent statistician to hold your hand.

14.6 Chapter summary

A 'Factor' is a categorical aspect of an experiment that we can alter to see if this changes the endpoint we are measuring. The various different possibilities for each factor are then referred to as 'Levels'. While *t*-tests are used with the simplest

experimental designs – a single experimental factor that has just two levels, for more complex designs, analyses of variance (ANOVAs) are called for.

The one-way analysis of variance is used where there is a single factor that can be set to three or more levels. It is not appropriate to analyse such data by repeated *t*-tests as this will raise the risk of false positives above the acceptable level of 5%. If the ANOVA produces a significant result, this only tells us that at least one level produces a different result from one of the others. It does not tell us which level differs from which other nor does it tell us anything about the extent of the differences in the endpoint.

'Follow-up' tests will rectify both of these short-comings. Tukey's test will look at the difference for every possible pair of levels of the factor. Dunnett's test will treat one level as a reference and then compare all other levels against that. A confidence interval is calculated for the difference between each pair of treatments. If an interval excludes zero, that comparison is statistically significant. The intervals are calculated to give each comparison less than a 5% risk of producing a false positive. In this way, the entire series of comparisons will accumulate a total 5% risk.

The two-way analysis of variance is used where two factors are being varied and all combinations of both factors have been studied. This ANOVA will test whether certain levels of each factor are consistently associated with high or low values for the endpoint. It will also test whether the effect of changing from one level to another within a factor is a constant increase/decrease or whether the effect seen depends upon the level of the other factor ('Interaction'). Where interaction is present, a graphical method can be used to clarify what form the interaction takes.

A distinction is drawn between experimental designs where the levels of a factor that are studied are specifically the ones of interest and those where we make a random selection from a wider pool. Factors are then referred to as 'Fixed' or 'Random'. The distinction is mainly relevant in two-way analyses of variance. When one factor is treated as fixed, the likelihood that the second factor will be declared significant is greater. However, it will also limit the generalizability of any conclusions concerning the second factor.

More complex experiments (where three or more factors are varied) are possible and can be analysed by multi-way ANOVAs. However, such experiments can produce an unmanageable proliferation of combinations of treatments. Expert advice should be sought at an early stage.

15

Correlation and regression – Relationships between measured values

This chapter will ...

- Describe positive and negative correlation.
- Show how the correlation coefficient (r) indicates both the nature and strength of any correlation between measured variables.
- Describe how to test the statistical significance of correlation.
- Emphasise that correlation is not necessarily proof of a cause and effect relationship.
- Describe the use of regression equations to predict the value of a dependent variable ('Response') from that of an independent one ('Predictor').
- Warn against the casual use of extrapolation.
- Demonstrate the 'Reverse calculation' of the independent from the dependent variable.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

- Show how several predictors can be combined to estimate the value of a response, using multiple regression.
- Describe the selection of a set of statistically significant predictors from a larger set of candidates.
- Explain how nominal variables can be incorporated into multiple regression analyses by the use of 'Dummy' or 'Indicator' variables.

15.1 Correlation analysis

In this chapter we will be looking at interrelationships among measured variables. The simplest question is whether there is a relationship between two sets of measurements and, if so, how strong is that relationship. It is these questions that can be answered by correlation analysis.

The general concept of correlation is familiar enough. Age and handgrip strength are both variables that can be measured on appropriate interval scales (one in years and the other in newtons). We are also sadly aware that, beyond a certain point, the inexorable increase in one value (age) is generally accompanied by an equally depressing decline in the other. It is this linkage between two measures that we recognise as correlation. Correlation does not require the linkage to be perfect. There will always be the odd 70 year old with a grip to match or beat most 30 somethings, but there's a clear general trend.

15.1.1 Positive and negative correlation

In the example quoted above, increases in one parameter were associated with decreases in the other. We refer to this as 'Negative correlation'. The term 'Negative' should not be taken to imply an absence of correlation, just a particular form of correlation. In the contrasting situation, where both parameters tend to increase together, we refer to 'Positive correlation'.

Positive and negative correlation

Positive: As one value increases the other also tends to increase.
Negative: As one value increases the other tends to decrease.

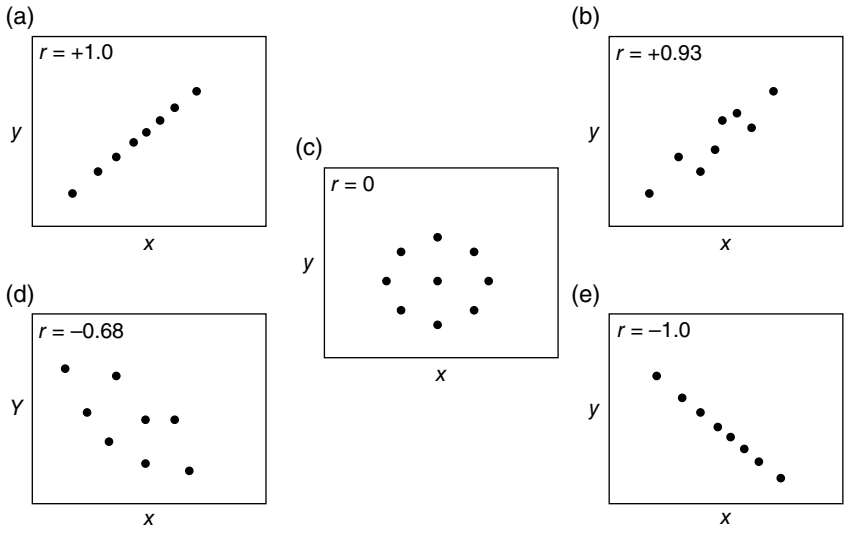


Figure 15.1 Information conveyed by the correlation coefficient (*r*)

15.1.2 The correlation coefficient (*r*)

Instances of correlation vary in strength. The standardisation graph for a colorimetric analytical method should show a very strong relationship between the measured absorption and the amount of analyte present. In contrast, in biological systems, where multiple factors tend to be at work, relationships are usually much vaguer. A graph of handgrip strength versus age would undoubtedly contain some pretty scattered points. The direction and strength of any correlation can be described by a statistic called the 'Correlation coefficient'. For reasons that escape the author, this is given the symbol '*r*'. (Note that a lower case *r* is used.)

The correlation coefficient can take any value between –1 and +1. Figure 15.1 shows examples of various types and strengths of correlation and the associated value of *r*. In this figure, the axes are simply labelled as *x* and *y*, so the graphs are general in nature.

Parts (a) and (b) both show positive correlation, so *r* takes a positive value. In (d) and (e) negative correlation leads to negative values of *r*. In (c), there is no relationship and *r* is exactly zero. Cases (a) and (e) represent the most extreme forms of correlation – perfect positive or perfect negative correlation. The associated *r* values are therefore also the most extreme values (+1 and –1). Cases (b) and (d) show partial correlation and the *r* values reflect the stronger correlation in (b) than in (d).

Correlation coefficient (*r*)

Describes the type and strength of correlation. It can take any value between –1 and +1.

The correlation described in this chapter is the 'Pearson correlation'. An alternative ('Spearman correlation') is described in Chapter 21. The Pearson correlation is used so frequently that it is common to see it referred to simply as 'Correlation'.

15.1.3 The need for significance testing of correlation coefficients

Imagine what would happen if two human characteristics were completely unrelated and we took a random selection of people and measured both characteristics in each person. In an ideal world, we would get a set of results that yielded a perfectly symmetrical graph like that seen in Figure 15.1 (c) and an associated *r* value of exactly zero. However in the real world, the points will almost always show some slight upward or downward trend and the *r* value (whilst close to zero) will take a small positive or negative value. Thus, finding a degree of correlation within a sample is not adequate justification for concluding that there must be correlation in the wider population. We need to go through the usual approach of setting up a null hypothesis and then assessing whether the evidence is consistent with such an hypothesis.

Null and alternative hypotheses

Null: Within the general population, the correlation coefficient between the two parameters is zero; that is they are uncorrelated.
Alternative: Within the general population the correlation coefficient is non-zero; that is there is a correlation of some type.

The null hypothesis assumes that if some apparent correlation is present in the sample, it arose by the mechanism shown in Figure 15.2. The underlying population may be distributed perfectly symmetrically, with no correlation at all, but the random sampling procedure must have selected points that create an impression of correlation (negative correlation in the example illustrated).

15.1.4 Factors that will influence the outcome of significance testing

There are two factors that will be taken into account when testing this null hypothesis – the size of the correlation coefficient and the sample size.

15.1.4.1 Size of the correlation coefficient Figure 15.3 (a) shows a sample with weak correlation and also indicates a hypothetical, uncorrelated population from which it

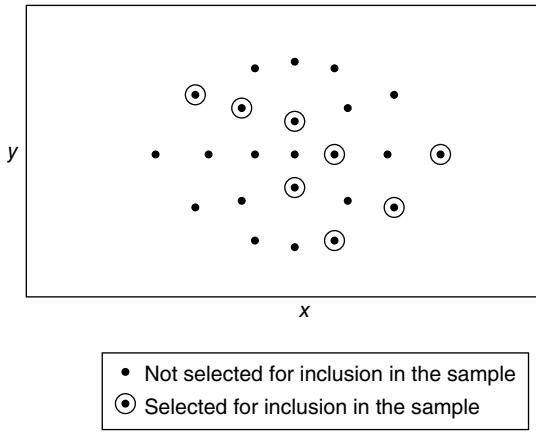


Figure 15.2 Mechanism assumed by the null hypothesis when testing for correlation

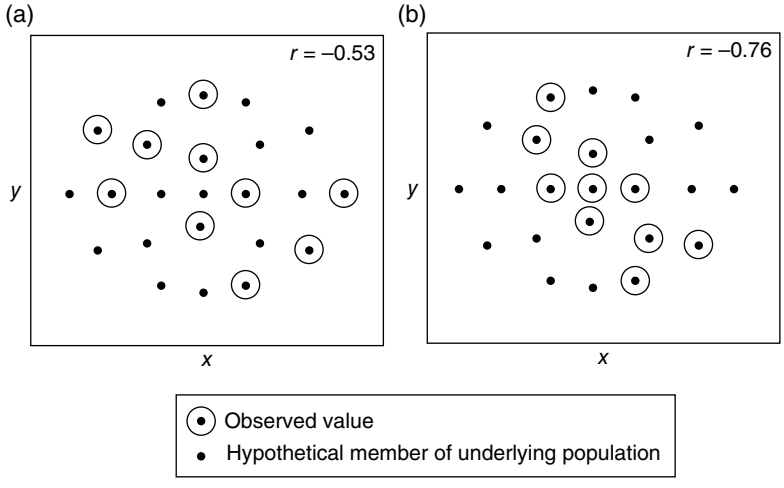


Figure 15.3 Influence of the size of the correlation coefficient on the outcome of significance testing

might have been drawn. It is not difficult to imagine that random sampling could quite easily lead to the vague appearance of correlation that we see in part (a). In that case we would have to accept that the null hypothesis is credible and there is inadequate evidence of correlation. However, in part (b) it would be much less likely that we would randomly select this more strongly correlated sample from an uncorrelated population. The null hypothesis becomes difficult to believe and this data does provide significant evidence of correlation. Note that the likelihood of significance depends upon what is called the 'Absolute' value of the correlation

coefficient. The closer r is to either –1 or +1, the greater the chances of significance; it is values around zero that are unlikely to prove anything.

15.1.4.2 Sample size Hopefully, the reader has by now recognised that the amount of data is an issue in all significance tests. Correlation testing is no exception. In Figure 15.4 (a), the data points are well aligned and the correlation coefficient would be quite high, but the null hypothesis is still perfectly credible – three randomly chosen points will quite frequently form something approaching a straight line. But, when ten points all confirm the same trend, as in Figure 15.4 (b), it is far more convincing.

Figure 15.5 summarises the situation. As usual, the various aspects can be offset against each other. So, for example, even very weak correlation can be statistically significant if enough observations are available. In contrast, small samples may yield a significant outcome, but only if correlation is very strong.

15.1.5 An example: The drug content in leaves and the height at which the leaves grew

We are planning the commercial collection of the leaves of a species of tree from which a drug will be extracted. One question we need to consider is whether it is worth using ladders to gain access to leaves at the tops of the trees or whether we would be better just collecting the easily accessed, low growing leaves and moving on to the next tree. If the leaves at the tops of the trees were a markedly

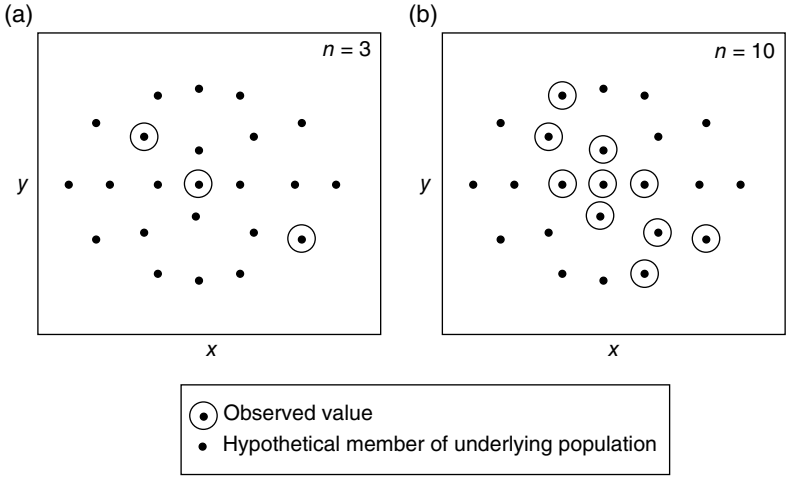


Figure 15.4 Influence of sample size on the outcome of significance testing for correlation

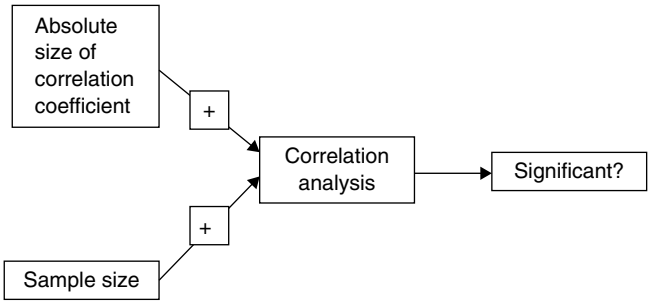


Figure 15.5 Factors influencing the outcome of significance testing for correlation

Table 15.1 Heights at which leaves were growing in the trees (m) and drug content (mg per 100 g) dry leaf

Height (m)	Drug conc. (mg per100 g)	Height (m)	Drug conc. (mg per 100 g)
1.70	1.66	3.23	1.27
2.31	1.34	3.29	0.85
2.89	1.27	3.46	1.16
1.30	1.61	3.95	1.14
3.21	1.17	1.70	1.25
1.84	1.73	2.92	1.49
3.27	1.17	2.67	1.17
4.21	1.19	3.02	1.16
1.32	1.93	2.37	1.75
3.67	1.10	2.64	1.36
2.78	1.37	4.25	1.00
3.71	1.19	1.90	1.48

richer source of the drug that those lower down, then the use of ladders might be worthwhile, but otherwise we would prefer to stay on *terra firma.*

We therefore collect a trial series of 24 leaves, recording the heights at which they were growing on the tree and also analyse them for drug content. The results are shown in Table 15.1.

15.1.6 Preliminary check for non-linearity

Using statistical packages to perform correlation analysis is so simple that it is tempting to wade straight in. Be advised – don't. Correlation analysis is a search for a straight line relationship. The danger is that two sets of data may be strongly related but in a non-linear manner. If a non-linear relationship is present, correlation analysis can be very misleading. It is vital that you always inspect a simple graph of

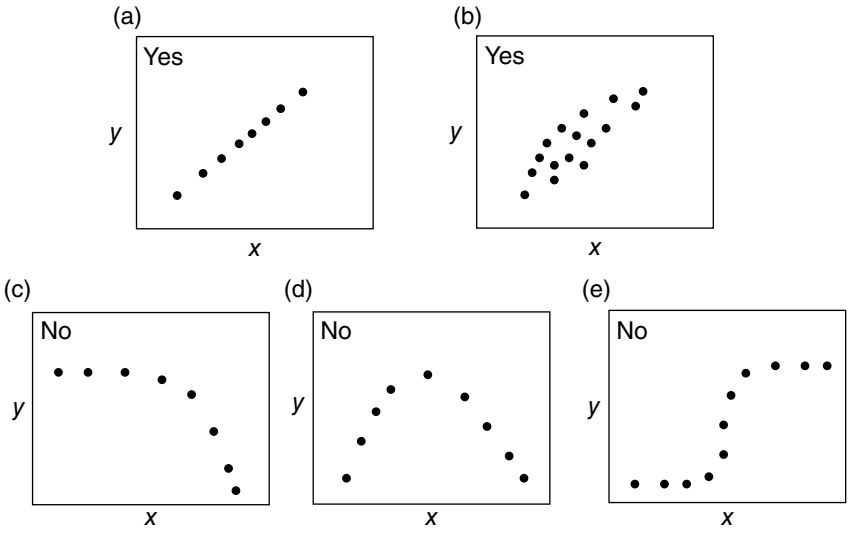


Figure 15.6 Is correlation analysis appropriate?

the data before proceeding to a statistical analysis. When we inspect the graph, we just need to satisfy ourselves that it doesn't provide unmistakable evidence of non-linearity. Figure 15.6 shows the types of patterns that are and are not acceptable. Parts (a) and (b) contain patterns where correlation analysis would be perfectly appropriate. Parts (c), (d) and (e) show strong evidence of non-linear relationships and correlation analyses would produce misleading results.

The leaf data (Figure 15.7) shows a pattern that could well be linear in nature and there is no objection to calculating a correlation coefficient.

Graphical inspection of the data

Data should be subjected to graphical inspection for any strong evidence of non-linearity, before launching into a formal correlation analysis.

15.1.7 Performing a correlation analysis

In computerised statistical packages the two sets of data are invariably entered into two columns and then we merely identify the columns to be analysed. The associated website (www.ljmu.ac.uk/pbs/rowestats/) gives details of how to use SPSS or Minitab to perform all the tasks described in this chapter. The output will include the correlation coefficient and a *P* value. Generic output for the height and drug content data is shown as in Table 15.2.

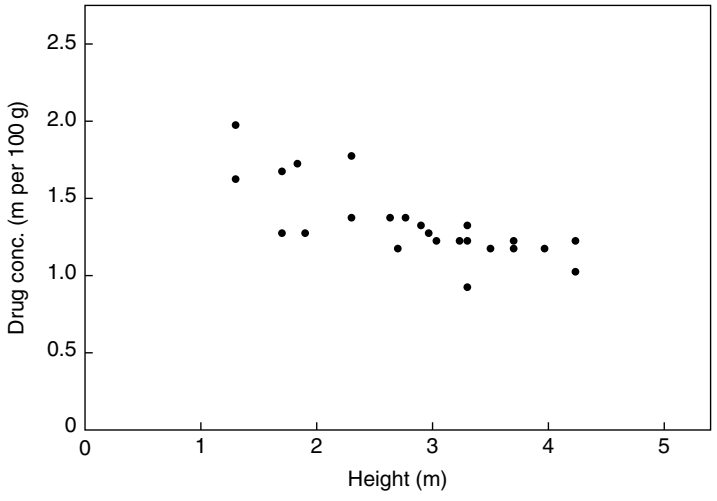


Figure 15.7 Drug concentration versus height at which leaves were growing

Table 15.2 Generic output for correlation analysis of height at which leaves were collected (m) and their drug content (mg per 100 g)

Correlation analysis
Correlation of Height and Drug: $r = -0.777$
$P = 0.000$

The r value is given as –0.777. The minus sign indicates negative correlation and a value of –0.777 tells us that there is quite a strong relationship. The results are also statistically significant ($P < 0.001$).

The practical conclusion is pretty obvious. There is certainly no evidence that the higher leaves contain extra drug that might lead us to risk life and limb up a ladder. Indeed there is statistically significant evidence of quite the opposite pattern. We'll stick to the nice easy leaves at the bottom, thank you very much!

15.1.8 A demonstration of correlation should not be assumed to imply a cause and effect relationship

The following data was obtained from the website of the United Kingdom (UK) Government's Office of National Statistics:

- Percentage of UK households owning a microwave oven (data available for seven years in the range 1991–2001).

- Numbers of deaths from liver disease (deaths per million population) for the same years.

A graph of this data shows a very strong (and certainly statistically significant) association between the two (Figure 15.8).

Rising microwave use is associated with rising deaths from liver disease. How one would interpret this trend, depends very much on what has gone before. If you have been exposed to a prolonged assertion that microwaves can interact detrimentally with liver cells, then you might be persuaded to start avoiding regular TV dinners. However, under more neutral circumstances, you would probably recognise that the points simply represent two independent time trends. The points on the graph are precisely in time order (Left point = 1991, Right = 2001). The graph then takes on its striking shape because during that decade there was a steady increase in microwave ownership (as with most consumer durables) and there was also a steady increase in fatal liver diseases. The reason for the rise in such deaths is not immediately apparent, but there is absolutely no reason to attribute it to microwave ovens. The decade saw changes in all manner of social habits, exposure to novel chemicals and drugs, increased foreign travel and goodness knows what else, any of which could be the real cause of the increased mortality.

In this case the disjunction between correlation and a causal relationship is fairly obvious, but the general principal that you can have correlation without a cause and effect relationship always needs to be borne in mind. The greatest danger arises in those cases where a causal relationship seems feasible.

Plant the suggestion and then produce the pretty graph

The secret with this one is doing things in the right order:

1. Argue some sort of theoretical case that A would cause B, then …
2. Point out that if you are right then A and B ought to be correlated and finally …
3. Produce the graph complete with impressive correlation. Naturally you will include a formal statistical analysis which will make it all look nice and official and nobody will argue.

Whether you get away with it depends largely on how effectively you complete stage 1. If that is done well, the overall effect can be highly seductive.

15.1.9 The magnitude of the correlation coefficient may depend on the range of data considered

It is useful to be aware that the magnitude of the correlation coefficient may vary according to the dynamic range of the data considered. Figure 15.9 shows a plot of some y values against their corresponding x values. The full data set [Within the

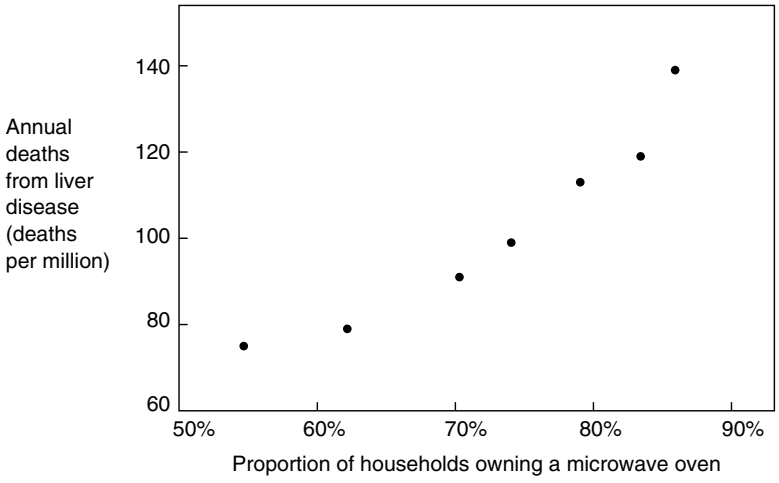


Figure 15.8 Microwave ovens and fatal liver diseases

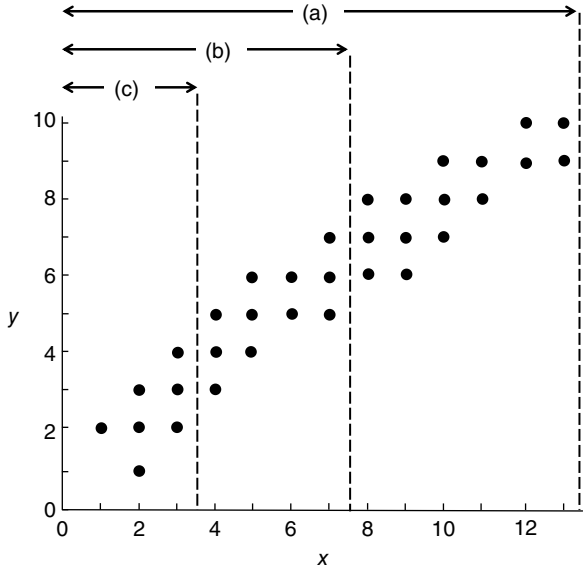


Figure 15.9 Three data sets (a, b and c) with differing ranges of x values and differing correlation coefficients

range indicated by the arrow labelled as (a)] is visibly strongly correlated; the points form a long, thin cloud. The correlation coefficient (r) for the full data set is consequently very high (+0.947).

However, if our data collection had failed to represent the higher values of x, we might end up with the data set indicated by arrow (b). The points still form a fairly long, thin cloud, but the relative dimensions have changed and the long dimension

is no longer as great compared to the short one. Correlation is therefore no longer as strong ($r = +0.874$).

If the range of x values that we could sample was very restricted as indicated by arrow (c), the cloud of points now becomes almost circular. The r value would now be only +0.484.

This means that (up to a point) we can manipulate the value of the correlation coefficient by controlling the range of data collected. A lazy data collector may fail to chase the full range of data and underestimate correlation and an overzealous one may artificially extend the dynamic range of the data to exaggerate the strength of correlation.

15.2 Regression analysis

15.2.1 An equation linking two measured variables

Correlation analysis only asks *whether* there is a relationship between two sets of data. Regression goes a step further and asks *how* are they related? More specifically it derives a mathematical equation that will allow us to predict one of the parameters if we know the value of the other.

Regression analysis

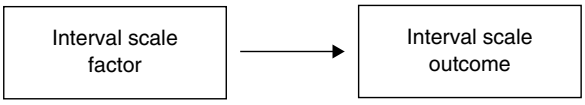
Regression analysis produces an equation by which the value of the *dependent* variable can be predicted from the *independent* variable.

The box above emphasises the fact that the equation operates in a specific direction. In order to undertake regression analysis, we have to decide which is the dependent and which the independent variable.

Figure 15.10 provides a diagrammatic representation of the circumstances where correlation or regression analyses are appropriate. The difference is based on whether we want to proceed to the stage of developing an equation to describe the relationship.

15.2.2 An example of regression: Fungal toxin contamination and rainfall

To illustrate regression, we will consider another data set, also from the natural products arena. A drug precursor molecule is extracted from a type of nut. The nuts are commonly contaminated by a fungal toxin that is difficult to remove during the purification process. We suspect that the amount of fungus (and hence toxin)



Is there a relationship? Correlation
What equation describes the relationship? Regression

Figure 15.10 Diagrammatic representation of circumstances where correlation or regression analyses are appropriate

Table 15.3 Rainfall at the growing site and concentration of fungal toxin in nuts

Rainfall (cm.week^{-1})	Toxin (μg per 100 g)
1.30	18.1
2.28	28.6
1.11	15.9
0.74	19.2
1.32	19.3
0.51	14.8
1.56	21.7
1.32	16.5
2.05	23.8
1.37	19.0

depends on rainfall at the growing site. We would like to be able to predict toxin concentration from rainfall in order to judge whether it would be worth paying additional rental charges for relatively drier sites. We analyse the toxin content in a series of batches of nuts and we also know the rainfall at the growing sites during the four months when the nuts are forming. The results are shown in Table 15.3.

As with correlation, the first job is to check for any obvious non-linearity in the data. Figure 15.11 shows that this is not a problem.

15.2.3 Identifying the line of best fit for the data – The least squares fit

We then want to find the best fitting straight line for the data. A possible line is shown on Figure 15.11. To determine how good a fit this line achieves, we determine the vertical distance (deviation) between each data point and the line. For example, the point corresponding to the lowest rainfall (0.51 cm.week^{-1}) deviates very slightly above the line, the next point is further above and the next deviates below the line and so on. The broken vertical lines indicate the deviation for each point. We then square each of these individual deviations and add them up. This provides what is referred to as the 'sum of squares'. The sum of squares acts as an inverse measure of goodness of fit – the lower the value, the better the fit.

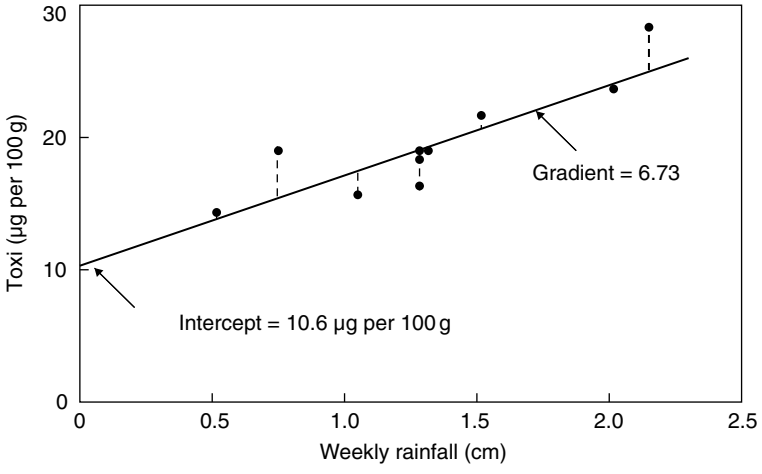


Figure 15.11 Regression line for fungal toxin and rainfall at the growing site

The line of best fit is then the one with the lowest sum of squares. (The 'Least squares fit') Fortunately we don't have to try endless different lines until we find the best. The line of best fit can be determined in a one-stage calculation.

Sum of squares and the 'Least squares fit'

Take all the vertical deviations of the points from the proposed line, square them and add them up. The lower the sum of squares, the better the fit. Select a line to achieve the 'Least squares fit'.

15.2.4 The line of best fit and the regression equation

The line shown in Figure 15.11 is in fact the line of best fit. It intercepts the vertical axis at a value of 10.6 µg per 100 g and it has a gradient of +6.73 (i.e. an increase in rainfall of 1 cm.week^{-1} is associated with an increase in toxin concentration of 6.73 µg per 100 g. So, the line corresponds to the relationship:

$$\text{Toxin concentration}\left(\mu\text{g per } 100\text{ g}\right) = 10.6 + 6.73 \times \text{rainfall}$$

This is referred to as the regression equation.

15.2.5 Performing a regression analysis using a statistical package

The rainfall and toxin data will be entered into two appropriately labelled columns. You will then have to indicate the relevant columns. But there is an important difference from correlation. With regression you must be careful to indicate correctly which is the dependent and which the independent variable. Unfortunately, statistical packages use a varied terminology. The toxin concentrations may be entered as the 'Dependent variable' or 'Response' and the rainfall may be the 'Independent variable' or the 'Predictor'.

Generic output is shown in Table 15.4. The order in which the various parts appear, depends upon the particular statistical package.

The first line tells us that the regression equation is:

$$\text{Toxin concentration}\left(\mu\text{g per 100 g}\right) = 10.6 + 6.73 \times \text{rainfall}\left(\text{cm.week}^{-1}\right)$$

Regression analyses produce a value called *R*-squared. This is in fact the square of the *r* value that we would get from a correlation analysis of the same data (often expressed as a percentage). Like *r*, it gives a measure of how closely the points fit to a straight line. Zero indicates random scatter and 100% a perfect fit to a straight line. Section 15.3.7 will explain in detail that unadjusted *R*-squared values are somewhat inflated and fail to take account of the complexity of the model being used. Consequently, a slightly reduced value '*R*-squared adjusted' (0.724 or 72.4%) is provided and is generally considered fairer. An *R*-squared of 72.4% indicates that rainfall is a pretty good predictor of toxin.

A lot of the output is concerned with the statistical significance of the regression equation. The null hypothesis is that within the general population there is actually no relationship between these two variables. This is essentially the same null hypothesis considered in correlation analysis. It is therefore no surprise that if a data set is

Table 15.4 Generic output for regression analysis of toxin concentration and rainfall

Regression analysis

Regression equation: Toxin = 10.6 + 6.73 × rainfall
R-Squared = 75.5% *R*-Squared (adjusted) = 72.4%

Analysis of overall equation

Source	DF	SS	MS	*F*	*P*
Regression	1	114.84	114.84	24.61	0.001
Residual Error	8	37.33	4.67		
Total	9	152.17			

Analysis of individual predictors

Predictor	Coeff.	SE Coeff.	T	*P*
Constant	10.6	1.961	5.39	0.001
Rainfall	6.73	1.356	4.96	0.001

subjected to both correlation and regression analyses the result is always the same. Either both analyses indicate significance or both non-significance – they never disagree. If you look at the next part of the output labelled 'Analysis of overall equation' there is a *P* value of 0.001, which is strongly significant.

There will also be an analysis of the significance of the predictor. With simple regression, this part of the output adds nothing of value.

15.2.6 Making predictions using the regression equation

Having obtained the regression equation, we might now have the chance to rent two agricultural locations where we could grow a crop of nuts. Enquiries show that the weekly rainfall at Sites A and B during the fruiting season are 2.05 and 1.25 cm. week^{-1} respectively.

We could therefore predict that nuts grown at these two sites would contain:

Site A

$$\begin{aligned} \text{Toxin} &= 10.6 + 6.73 \times \text{Rainfal} \\ &= 10.6 + 6.73 \times 2.05 \\ &= 10.6 + 13.8 \\ &= 24.4\ \mu\text{g per } 100\ \text{g} \end{aligned}$$

Site B

$$\begin{aligned} \text{Toxin} &= 10.6 + 6.73 \times \text{Rainfal} \\ &= 10.6 + 6.73 \times 1.25 \\ &= 10.6 + 8.4 \\ &= 19.0\ \mu\text{g per } 100\ \text{g} \end{aligned}$$

With its lower rainfall, site B is predicted to produce a slightly better crop (22% lower toxin load), but this modest advantage would have to be weighed against any other differences between the two sites.

15.2.7 Extrapolation

Using the regression equation we could predict that in an area with zero rainfall, nuts would contain $10.6 + 6.73 \times 0 = 10.6$ μg per 100 g. However, this is clearly nonsense. In an arid desert … no trees … no nuts … no toxin … no nothing. We have direct evidence of a reasonably linear relationship between toxin and rainfall over an observed range of 0.51–2.28 cm of rain per week. It is therefore reasonably safe to make predictions for any other site that has a rainfall figure within this range. This is referred to as 'Interpolation'. See Figure 15.12.

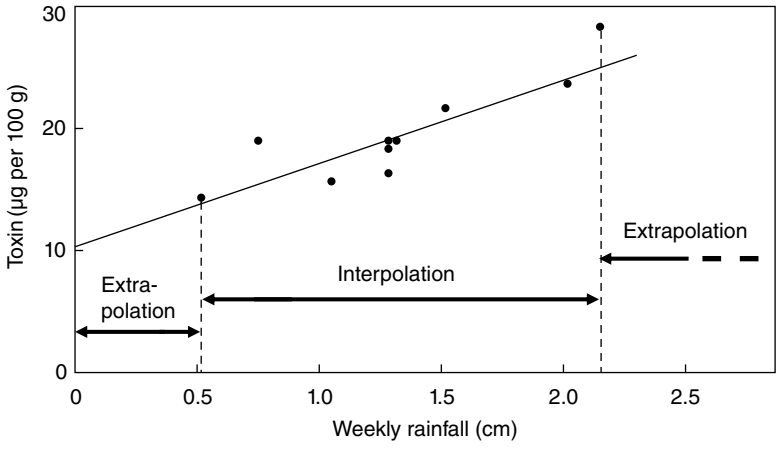


Figure 15.12 Interpolation and extrapolation

As soon as we move outside this range, we have no knowledge of whether the straight line relationship continues. In reality it is unlikely that it does. In the Atacama Desert, the trees would simply die and in Cherapunjee (rain in excess of 100 cm.week^{-1}) they'd get washed away. Attempts to make predictions in cases where the value of the independent variable is outside the range we have actually observed are referred to as extrapolation. The general rule is that extrapolation should be avoided unless there is a sound theoretical reason to believe that the linear relationship continues beyond the observed range. In reality that is rarely the case, although some instances do arise. An example of reasonable extrapolation is carbon^{14} dating. Nobody has ever observed C^{14} decaying for 5000 years, but once we have observed its behaviour for a couple of years, it is safe to predict its fate for the extra 4998 years.

Interpolation and Extrapolation

Interpolation: A prediction using a value of the independent variable that is within the observed range – Uncontroversial.
Extrapolation: A prediction using a value of the independent variable that lies outside the observed range. Extrapolation should be avoided unless there is sound reason to believe that the linear relationship extends beyond the observed range.

15.2.8 Reverse calculation

The normal purpose of regression is to be able to obtain a value for the dependent variable from the independent variable. However, there are times when we want to operate in the opposite direction. Classic examples are analytical methods that use a

calibration graph. For example, in a colorimetric method, we start with a set of standards and we can calculate a regression equation relating light absorption to amount of analyte in the usual way. However, in the next stage we will obtain readings of absorption for our unknown samples and will want to calculate their analyte contents. In that latter stage we are working backwards – calculating the independent variable (analyte concentration) from the dependent variable (absorption). The example below shows how this is done.

Para-amino benzoic acid (PABA) can be measured by a diazotisation reaction that yields a bright pink colour. The absorption is then measured at 510 nm. A series of standards yielded the results shown in Table 15.5.

A graph of this data (Figure 15.13) shows excellent linearity, so regression analysis is appropriate.

Table 15.5 Quantity of PABA and resultant absorption at 510 nm

PABA (μg)	Absorption
0	0.00
5	0.11
5	0.10
10	0.20
10	0.20
15	0.30
15	0.32
20	0.41
20	0.40
25	0.50
25	0.51

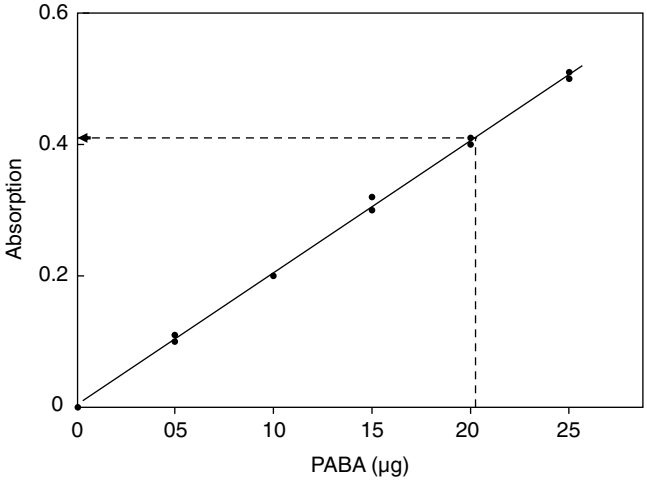


Figure 15.13 Reverse prediction of PABA concentration from absorption in a colorimetric analysis

The regression analysis should treat the PABA concentration as the independent variable and absorption as the dependent. The regression equation is:

$$\text{Absorption} = 0.0023 + 0.0202 \times \text{PABA}\left(\mu g\right)$$

We can then re-arrange the equation so that it does what we want – allow us to calculate the PABA concentration from absorption.

$$\text{PABA}\left(\mu g\right) = \frac{\text{Absorption} - 0.0023}{0.0202}$$

If we then have an unknown sample with an absorption of (say) 0.41, its PABA content must be:

$$\text{PABA} = \frac{0.41 - 0.0023}{0.0202}$$

$$\text{PABA} = \frac{0.4077}{0.0202}\mu g$$

$$\text{PABA} = 20.2\ \mu g$$

The procedure above seems rather cumbersome. Would it not be possible simply to carry out the regression, reversing the dependent and independent variables and get directly to an equation that predicts PABA concentration from absorption? We could do this, but the fitted line would be slightly different and would not be properly optimised. The correct procedure is the one shown above.

Reverse calculation

To predict the value of the independent from the dependent variable, the normal equation (predicting dependent from independent) is calculated initially and is then re-arranged.

15.3 Multiple regression

15.3.1 Using several predictors simultaneously

Multiple regression is a fairly complex subject with a number of potential pit-falls. This section is really only meant to give a taste of what it does. If you want to use it yourself, it would probably be a good idea to get some advice from a competent statistician.

In the previous example we used rainfall to predict toxin concentrations in crops of nuts. However, there may well be other aspects of the growing site that influence fungal growth. For example sunshine and wind might tend to dry the trees and reduce fungal growth. We could set up a whole series of regression equations each using one aspect of the growing site to make our predictions. However, each of these would be imprecise because if it used (say) wind speed then it would be ignoring the influence of rainfall and vice versa. What we want is a single equation that contains terms that reflect the influence of each relevant factor. The original regression equation contained a constant (a) and a term reflecting the influence of rainfall ($b \times$ rainfall):

$$\text{Predicted toxin conc.} = a + (b \times \text{Rainfall})$$

The term b is referred to as a 'coefficient. There was a positive relationship between rainfall and toxin, so the value of coefficient b was positive. In that way, the greater the rainfall the greater the predicted toxin concentration.

To make use of several predictors simultaneously, all we do is add extra terms to cover each of the additional factors. If we have information about rainfall, temperature, daily hours of sunshine and wind speeds, the equation becomes:

$$\begin{aligned}\text{Predicted toxin conc.} = a &+ (b \times \text{Rainfall}) + (c \times \text{Temp.}) \\ &+ (d \times \text{Sunshine}) + (e \times \text{Wind})\end{aligned}$$

We know that rainfall is positively related to toxin, but some of the other factors may show a negative relationship. For example we might suspect that higher wind velocities would cause drying and lower fungal growth. This can be accommodated simply by giving the various coefficients (b, c, d and e) positive or negative values as appropriate.

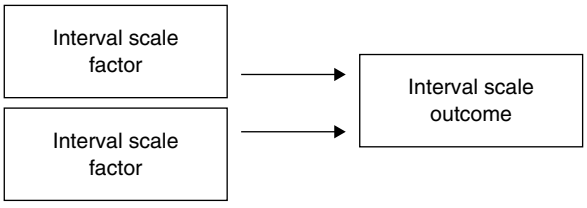
Multiple regression

Multiple regression allows the prediction of a measured outcome, using several predictors. A regression equation is developed containing a separate term for each predictor. These terms consists of a coefficient multiplied by the value of the particular predictor.

Figure 15.14 provides a diagrammatic representation of the use of multiple regression.

15.3.2 An example: Using several meteorological variables to predict fungal toxin levels

Table 15.6 shows the same data as Table 15.2, but also includes the additional data concerning temperature, sunshine and wind.



Two or more interval scale factors may influence an interval scale outcome. (Nominal factors can be incorporated–Section15.3.9.)

Figure 15.14 Diagrammatic representation of circumstances where the use of multiple regression is appropriate

Table 15.6 Rainfall, temperature, sunshine and wind speed at growing sites and concentration of fungal toxin in nuts

Rain ($cm.week^{-1}$)	Noon temp. (°C)	Sunshine ($h.day^{-1}$)	Wind speed ($km.h^{-1}$)	Toxin (μg per 100 g)
1.30	20.9	6.23	13.3	18.1
2.28	25.4	8.13	10.8	28.6
1.11	28.2	10.21	10.9	15.9
0.74	23.7	6.96	8.2	19.2
1.32	26.5	9.04	9.8	19.3
0.51	23.9	7.84	12.3	14.8
1.56	26.7	6.69	10.0	21.7
1.32	30.0	8.30	12.2	16.5
2.05	24.9	9.22	10.7	23.8
1.37	22.0	8.37	15.0	19.0

15.3.3 Performing multiple regression

15.3.3.1 Start with all the potential predictors In a statistical package, the data are entered into five columns and then you will have to indicate that toxin is the dependent variable/response and the four meteorological factors are the independent variables/predictors. Generic output is shown in Table 15.7.

The first thing to notice (under 'Analysis of overall equation') is that the regression equation is statistically significant ($P = 0.006$). However, all this tells us is that the equation as a whole probably does have real predictive power. The problem is that the effectiveness of the equation might be due solely to the fact that it takes rainfall into account (which we already know to be a useful predictor.) It is possible that some (or all) of the other factors we have added may be doing nothing to improve the accuracy of prediction. To test this we need to look at the section 'Analysis of individual predictors'. Here we find *P* values for each individual predictor.

Table 15.7 Generic output for regression analysis of toxin concentration using all potential predictors (rainfall, temperature, sunshine and wind)

Regression analysis

Regression equation:
Toxin = 31.6 + 7.07 × Rain – 0.420 × Temperature – 0.237 × Sun – 0.794 × Wind
R-Squared = 91.9% *R*-Squared (adjusted) = 85.3%

Analysis of overall equation

Source	DF	SS	MS	*F*	*P*
Regression	4	139.782	34.946	14.11	0.006
Residual Error	5	12.387	2.477		
Total	9	152.169			

Analysis of individual predictors

Predictor	Coeff.	SE Coeff.	T	*P*
Constant	31.6	7.105	4.45	0.007
Rain	7.07	1.003	7.05	0.001
Temp.	–0.420	0.2413	–1.74	0.142
Sun	–0.237	0.5086	–0.47	0.660
Wind	–0.794	0.2977	–2.67	0.045

Rain and wind are shown with *P* values of 0.001 and 0.045, so both are significant. The other two (Temperature and Sunshine) are apparently not significant.

15.3.3.2 Remove non-significant variables one at a time What we now need to do is start removing the non-significant factors from the regression equation. However, we do not immediately reject both temperature and sunshine. Multiple regression behaves rather strangely if two of the predictors are themselves correlated. In this case temperature and sunshine are (not surprisingly) positively correlated with an *r* value of +0.501. This means that there is an element of redundancy – the two sets of values largely reflect the same information. Consequently, it may well be true that we do not need to retain both factors. Unfortunately, when multiple regression analysis encounters two factors that are markedly correlated it tends to report that we don't need either! The correct way to proceed is to eliminate one of these factors and see what happens to the other. We will either find that the remaining factor is now miraculously revealed as significant (in which case we obviously retain it) or it will remain non-significant and we can get rid of it as well. The remaining question is which we should remove first – Temperature or Sunshine? From a purely statistical point of view, we would prefer to retain the factor that is closest to significance. Since Temperature had a *P* value of 0.142 it has a stronger claim to be retained than Sunshine ($P = 0.660$). In the absence of any other consideration we would probably drop Sunshine and try again. Occasionally our knowledge of the particular situation may lead us to believe that some factor is of special importance and ought to be retained. This might then over-ride purely statistical considerations. In this case,

Table 15.8 Generic output for regression analysis of toxin concentration with sunshine removed, leaving rainfall, temperature and wind as predictors

Regression analysis

Regression equation:
Toxin = 31.6 + 7.01 × Rain – 0.479 × Temperature – 0.822 × Wind
R-Squared = 91.5% *R*-Squared (adjusted) = 87.3%

Analysis of overall equation

Source	DF	SS	MS	F	P
Regression	3	139.242	46.414	21.54	0.001
Residual Error	6	12.927	2.155		
Total	9	152.169			

Predictor	Coeff.	SE Coeff.	T	P
Constant	31.6	6.625	4.76	0.003
Rain	7.01	0.9285	7.55	0.000
Temp.	–0.479	0.1919	–2.50	0.047
Wind	–0.822	0.2718	–3.02	0.023

there is no such special knowledge and we will repeat the regression, using rainfall, temperature and wind, but omitting sunshine. The output is now as in Table 15.8.

We now have the happy situation that the equation as a whole is significant ($P = 0.001$), as are all the contributory factors (Rain, Temperature and Wind) have significant P values (<0.001, 0.047 and 0.023, respectively). Temperature has been converted from non-significance to significance.

Removing non-significant factors from multiple regression equations

When two or more factors are shown to be apparently non-significant, do NOT remove them all simultaneously. Remove one factor at a time until all the remaining factors are statistically significant.

15.3.4 The final equation

The regression equation is:

$$\text{Toxin conc.} = 31.6 + (7.01 \times \text{Rain}) - (0.479 \times \text{Temperature}) - (0.822 \times \text{Wind})$$

The term for rainfall still has a plus sign, so higher rainfall will increase the predicted toxin concentrations. However, temperature and wind speed have minus signs, so high temperatures and strong winds are presumably associated with

drying and lower fungal growth and toxin concentrations. Biologically, this is all perfectly reasonable.

15.3.5 Forward selection of variables

In Section 15.3.3, we selected our predictors by starting with the full set and eliminating those considered irrelevant ('Reverse elimination'). It is possible to use 'Forward selection' where you start by finding the best single predictor and then look for the best additional variable to accompany it and so on. At each stage, the significance of all predictors is tested and you stop when you can no longer find a further variable that could be added and would achieve statistical significance.

There is no absolute case for preferring either approach, but for selection among a limited number of potential predictors, the author has generally found reverse elimination satisfactory. The one situation where forward selection is nigh on unavoidable is where there are very large numbers of potential predictors. To undertake reverse elimination you might have to start with a regression equation containing 100 terms. (However, see caution in Section 15.3.8 against trawling through excessive numbers of possible predictors.)

15.3.6 Using the equation to predict toxin contamination

Let's return to the two potential sites that we already compared on the basis of rainfall alone, but now take account of temperature and wind speed as well. The fuller data for these two sites are shown in Table 15.9.

In our initial prediction (based on rainfall alone), site B seemed better as it had a lower rainfall. However, we can now see that the other factors favour site A (higher temperatures and stronger winds). Taking everything into account, our predictions would now be:

Site A

$$\begin{aligned}\text{Toxin conc} &= 31.6 + (7.01 \times \text{Rain}) - (0.479 \times \text{Temp.}) - (0.822 \times \text{Wind}) \\ &= 31.6 + (7.01 \times 2.05) - (0.479 \times 26.1) - (0.822 \times 11.0) \\ &= 31.6 + 14.37 - 12.5 - 9.04 \\ &= 24.43\,\mu\text{g per } 100\text{ g}\end{aligned}$$

Table 15.9 Meteorological data for two potential growing sites

Site	Rain (cm.week^{-1})	Noon temp. (°C)	Wind speed (km/h^{-1})
A	2.05	26.1	11.0
B	1.25	22.5	9.1

Site B

$$\begin{aligned}\text{Toxin conc} &= 31.6 + (7.01 \times \text{Rain}) - (0.479 \times \text{Temp.}) - (0.822 \times \text{Wind}) \\ &= 31.6 + (7.01 \times 1.05) - (0.479 \times 22.5) - (0.822 \times 9.1) \\ &= 31.6 + 8.76 - 10.78 - 7.48 \\ &= 22.10\,\mu\text{g per } 100\text{ g}\end{aligned}$$

The effects of higher temperatures and wind speeds at site A have not quite offset the effects of the lower rainfall at Site B and the latter continues to be predicted to be the better site. However, the difference is now very small and we need to bear in mind that all these predictions are only approximate. There is effectively nothing to choose between the two sites.

15.3.7 Better fit with more predictors

Simple regression using just one predictor (Rainfall) produced an unadjusted *R*-squared value of 0.755 (Table 15.4). With multiple regression, the unadjusted *R*-squared value rose to the very high value of 0.915 (Table 15.8). The multiple regression equation would be expected to achieve a better fit, simply because additional predictors have been incorporated. This would be true even if the additional predictors consisted of random numbers. The adjusted *R*-squared value incorporates a penalty factor that reduces *R*-squared to an extent that depends upon the number of predictors – the greater the number of predictors, the greater the penalty. In this way, although unadjusted *R*-squared always increases with additional predictors, adjusted *R*-squared will only increase if an additional predictor improves the fit more than would be expected if the new predictor contained random data. It is for this reason, that adjusted *R*-squared is preferred to the unadjusted value.

15.3.8 Multiple testing

In Chapter 14, it was emphasized that multiple *t*-tests were not acceptable as they constituted multiple testing and would increase the risk of false positives. Multiple regression inevitably brings with it a risk of multiple testing as we are considering several possible factors and running a 5% risk of a false positive with each one. In the case studied above, we initially considered only four factors and three of these were retained as apparently significant, so it is unlikely that these were actually false positive findings. A further defence of our finding is that the three factors selected were all biologically plausible.

The real danger arises if we start with a very large number of possible factors and insist on picking out a small number of these. For example, if we were trying to relate the biological activity of a series of molecules to their chemical properties (Quantitative Structure Activity Relationships – QSAR) we could use modern computer software to generate literally hundreds of chemical descriptors for the

molecules in question. Let's assume that we start looking for relationships between all of these descriptors and the various molecules' biological activities and let's also assume that none of the descriptors actually has the slightest relationship to the endpoint. By the time we have trawled through (say) 100 descriptors, 1 in 20 cases will generate false positives and we would expect to find several 'Predictors' each carrying its own individual (and apparently comforting) *P* value of less than 0.05. We could then discard the 90 something descriptors that have been correctly identified as irrelevant and publish a multiple regression equation based on the handful of false positives.

When faced with any claim that a significant regression (or multiple regression) equation has been discovered, it is always worth asking how many potential predictors were initially considered.

☠☠ Check large numbers of possible 'predictors' and exploit the false positives

Some areas of research have been so bedevilled with this nonsense that you are unlikely to get away with it, but choose a virgin field and there may yet be some mileage.

What you need is a nice big data base, with lots of sets of values for factors that might be related to the matter in hand (You really need a minimum of 20 or 30 candidates). Trawl vigorously, rejecting 95% of the factors as non-significant. Publish the remainder, placing great emphasis on those lovely low *P* values which undoubtedly prove that this select band of flukes are genuine predictors.

15.3.9 Incorporating nominal factors into multiple regression models using indicator variables

15.3.9.1 Representing chemical structures by indicator variables It is possible to incorporate nominal (categorical) variables into a multiple regression model using 'Indicator' (sometimes called 'Dummy') variables. We will consider a problem in QSAR where we are trying to predict the activity of phosphodiesterase inhibitors from their chemical properties. Their inhibitory activity is expressed relative to caffeine (Caffeine = 1.0). The properties that will be used as predictors are the molecules' pK_a values and the substituent at a particular ring position (either remaining as hydrogen or substituted by fluorine). pK_a values constitute interval scale data and therefore fit naturally into a regression model, but the substituents do not.

We get around this problem by representing the structures as numerical values; zero and one are usually used, but this is not essential. In this case we will make an arbitrary decision to use zero to code for hydrogen and one where there is fluorination. The data is shown in Table 15.10.

Table 15.10 Phosphodiesterase inhibitory power of various compounds (relative to caffeine), their pK_a values and their substituents (Zero codes for hydrogen and one for fluorine)

Compound number	Activity	pK_a	Substituent
1	5.07	7.9	H = 0
2	2.42	8.5	H = 0
3	3.04	9.0	H = 0
4	0.78	9.6	H = 0
5	8.07	7.8	F = 1
6	5.01	8.0	F = 1
7	5.48	8.4	F = 1
8	4.63	8.9	F = 1
9	4.43	9.4	F = 1

15.3.9.2 Regression equation including an indicator variable The data in Table 15.10 is now available in numerical form and can be entered into a regression analysis. The resulting regression equation is:

$$\text{Activity} = 19.7 - (1.93 \times pK_a) + (2.21 \times \text{Substituent})$$

Both predictors are statistically significant and are therefore retained in the model.

So, if a new compound has a pK_a of 9.4 and contains a hydrogen atom then (bearing in mind that hydrogen is coded as zero) its predicted activity will be:

$$\begin{aligned}\text{Activity} &= 19.7 - (1.93 \times 9.4) + (2.21 \times 0) \\ &= 19.7 - 18.1 + 0 \\ &= 1.6\end{aligned}$$

The negative coefficient for pK_a means that higher pK_a values are associated with lower activity. The positive coefficient for the substituent means that any substance with a fluorine atom (carrying a numeric value of one) will be modelled as more active than one containing a hydrogen (coded as zero).

The numerical coding of hydrogen as zero and fluorine as one was largely arbitrary. Had we used the reverse coding, then the regression equation would have included a negative instead of positive coefficient for the substituent and we would still end up modelling hydrogen containing substances (now coded as one) as less active than those that were fluorinated. Hence, the choice of coding is not important.

15.3.9.3 Indicator variables with more than two possible values Where we want to include a nominal factor that is dichotomous (has two possible values), the choice of numbers we use to code them does not matter a great deal (although zero and one are commonly used). However, when the factor can take more than two values we need to be very careful.

As an example, assume that our phosphodiesterase inhibitors include substances with hydrogen, fluorine and chlorine substituents. The temptation might be to code these as (say) hydrogen = zero, fluorine = one and chlorine = two. However, that is not the answer. Such a coding scheme makes two assumptions, neither of which is justified:

- We know that fluorinated substances (code one) are more active than those containing hydrogen (code zero) and so the equation must associate a positive coefficient with the substituent. However, if chlorinated substances are to be given a code of two, then the positive coefficient will automatically mean that these are predicted as more active than either of the other groups; there is no particular reason why that should be the case.

- The coefficient models the extent to which the activity of fluorinated substances exceeds that of those containing hydrogen. If we use the zero, one, two coding system then this will mean that the size of the step up in activity as we go from hydrogen containing to fluorine containing substances must be the same as the step up in activity as we go from fluorinated to chlorinated substances; there is also no reason why this would necessarily be so.

The correct approach is to represent the factor using two variables. We make (another) arbitrary decision to use one variable as a yes/no indicator for one of the possible substituents and the other variable does the same job for a second substituent. Thus we might create the two following variables:

- A variable called 'Fluor': Indicates whether the compound contains fluorine. Zero codes that it does not and one codes that it does.

- A variable called 'Chlor': Indicates whether the compound contains chlorine. Zero = no; One = yes.

If the compound contains hydrogen, then it fits neither category and both variables are set to zero. We thus end up with the following possible situations: one or the other variable set to a value of one or both set to zero. There should, of course, be no cases where both variables are set to one.

This coding scheme will then allow us to produce a regression equation that includes both variables.

For example, we might obtain an equation such as:

$$\text{Activity} = 19.7 - (1.93 \times \text{pK}_\text{a}) + (2.21 \times \text{Fluor}) + (1.30 \times \text{Chlor})$$

Both of the variables describing substitution are associated with positive coefficients and so, if either of these elements is present, activity will be greater than we would see with hydrogen. As the fluorine coefficient is the greatest, we can rank the

hydrogen containing compounds as being the least active, chlorinated ones are intermediate and fluorinated substances are the most active. The important point is that the format of the equation is completely flexible; it could model any pattern of relative activities among the three substituents.

The general rule is that the number of variables required to represent a nominal factor is one less than the number of possible values that the factor can adopt. So if there were five different substituents, we would need four variables.

15.4 Chapter summary

Correlation and regression analyses describe the relationships between measured variables (generally on interval scales).

Correlation may be positive (values increase or decrease together) or negative (increases in one value associated with decreases in the other). The correlation coefficient describes what type of relationship exists (positive or negative) and the strength of the association. It may take any value between –1 and +1. Before calculating the correlation coefficient, a graphical check should be made for any clear evidence of a non-linear relationship. The statistical significance of correlation will depend upon the value of the correlation coefficient and the number of observations. A demonstration of correlation should not automatically be assumed to provide evidence of a cause and effect relationship.

In regression analysis, we identify the best fitting straight line through the observed data points. This is selected on the basis that it minimises the sum of the squared vertical deviations of the points from the line – the 'Least squares fit'. The goodness of fit is reported as an *R*-squared value which can vary between 0 (no fit) and 100% (perfect fit of line to points).

The equation of the regression line can then be used to predict the value of the dependent variable from that of the independent. This equation will take the form:

$$y = a + b.x$$

where '*a*' represents the intercept of the regression line with the vertical axis and '*b*' (the coefficient) is the gradient of the line.

When computing the regression equation, it is vital to identify correctly which is the dependent and which the independent variable. Once a regression equation has been developed, it can be used to predict the value of the dependent variable for a new case, using the value of the independent variable. When making such predictions, values of the independent variable should be within the range of observed values that were used to create the regression equation (Interpolation). Using values outside the observed values (Extrapolation) should be avoided unless there is good reason to believe that the linear relationship continues beyond the observed range. If it is necessary to estimate the value of the independent variable from the dependent

(Reverse prediction), the normal regression equation should be calculated initially and then algebraically re-arranged.

Regression can be extended to multiple regression, allowing several factors to participate in the prediction of the dependent variable. For every additional factor considered, an additional term is added to the regression equation shown above. It is necessary to establish not only that the equation is significant overall, but also that each individual contributing factor is significant. If some of the factors are found to be non-significant, these should be removed one at a time and significance re-tested at each stage, stopping once all remaining factors are individually significant. Readers should be wary of claims to have found significant regression equations, if the factors used in the equation are a small subset from a much larger initial collection; some (or all) of the factors claimed as statistically significant may be no more than false positives.

Nominal factors can be incorporated into multiple regression equations by the use of 'Indicator' (or 'Dummy') variables.

16
Analysis of covariance

This chapter will ...

- Describe how to use analysis of covariance (ANCOVA) in a scenario where an interval (measured) endpoint may be affected by two factors, one of which is also interval, but the other nominal.
- Discuss possible interaction between the two factors.
- Describe 'Common slopes' models.
- Explain how ANCOVA may achieve markedly greater statistical power than would be achieved if a *t*-test were applied to the same data.
- Show how ANCOVA can correct for baseline imbalances within the continuously measured factor.
- In the case of medical treatments, show how ANCOVA can help to identify significant prognostic factors that may affect the outcome of a treatment.

Chapters 7–14 described the use of *t*-tests and analyses of variance to deal with cases where a measured (interval) endpoint may be affected by one or more categorical (nominal) factors and then we met correlation and regression techniques

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

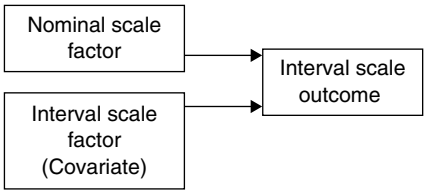


Figure 16.1 Structure of a study to be analysed by analysis of covariance (ANCOVA)

(Chapter 15) that also dealt with interval endpoints, but here the factor(s) were interval rather than nominal.

Not surprisingly, there are situations where an interval endpoint may be affected by mixed factors (nominal and interval) and this is where analyses of covariance (ANCOVA) are used.

Terminology used to describe experimental structures in not entirely consistent, but within this book the term 'Covariate' will be reserved for any factor which would be recorded on a measured (interval) scale. In other places you may see any factor (Whether interval or nominal) referred to as a covariate. A diagrammatic representation of an experimental structure where ANCOVA would be appropriate is shown in Figure 16.1.

16.1 A clinical trial where ANCOVA would be appropriate

The aim is to compare two treatments for chronic obstructive pulmonary disease (COPD). The treatments will be referred to as NewDrug and OldDrug. Suitable patients were randomly assigned to the two treatment groups. The clinical endpoint was percentage change in Forced Expiratory Volume in one second (FEV_1), comparing the pre-treatment value to the value after one week of treatment and calculating the percentage change. We suspected that the outcomes may vary according to patients' ages, so these were also recorded.

The outcome is measured (Percentage change in FEV_1). There is a nominal factor (New- or OldDrug) and an interval factor (Age), so ANCOVA is the natural choice of statistical analysis.

The results are shown in Table 16.1. The mean increase in FEV_1 among those treated with NewDrug (37.84%) is higher than for OldDrug (24.05%), but a formal test is needed to discern whether this is statistically significant.

The main point of interest is whether there is a difference in response according to the drug choice (i.e. the nominal factor). There may be some direct interest in the influence of age on response, but the principal reason for its inclusion is that it greatly improves the analysis of the effect of drug choice. This is a common pattern in ANCOVA – the nominal factor is the real point of interest and the covariate is included primarily to improve the analysis. However, this is not always the case;

Table 16.1 Patients' ages and the percentage changes in FEV_1 for the two treatment groups

OldDrug		NewDrug	
Age	FEV_1 change (%)	Age	FEV_1 change (%)
61	23	72	28
70	24	47	46
77	15	57	40
64	26	65	30
53	28	52	38
52	30	72	27
77	10	63	39
69	24	64	36
48	30	45	50
43	33	63	41
62	23	56	36
44	33	57	46
58	20	63	38
72	16	57	37
47	29	50	35
66	27	46	52
55	32	74	25
73	11	74	25
61	24	55	36
53	23	65	29
		58	42
60.25	24.05	56	38
		43	43
		50	48
		44	41
		57.92	37.84

ANCOVA could quite reasonably be used when the continuously varying covariate was the main point of interest or where the two factors were of equal interest.

16.2 General interpretation of ANCOVA results

Before undertaking the ANCOVA, it is useful to get a mental picture of the pattern of results using graphs such as those hypothetical cases in Figure 16.2. The endpoint is plotted against the covariate with separate lines based on the nominal variable. In this case, change in FEV_1 is plotted against age with separate lines for the two treatments.

- In Figure 16.2 (a), there is no effect of either factor. The outcome does not change with age – the lines have no positive or negative gradient. Choice of drug treatment also makes no difference – there is no vertical separation between the lines.

- Figure 16.2 (b) shows an effect of age – the outcome declines in older subjects, but there is still no difference between treatments.

- Figure 16.2 (c) does show an effect of drug choice – greater effect with NewDrug than OldDrug, but now there is no effect of age.

- Finally, in Figure 16.2 (d) we see effects of both age and drug choice. There is a definite gradient to the lines and vertical separation between them.

Section 14.3.2 introduced the concept of interaction. Wherever two (or more) factors are considered together, interaction is possible. In this case, it would take the form of a variable effect of drug choice depending upon patients' ages. In parts (c) and (d) of Figure 16.2, there is a drug effect, which is seen as greater values for the endpoint with NewDrug than OldDrug. The extent of this superiority is the same in

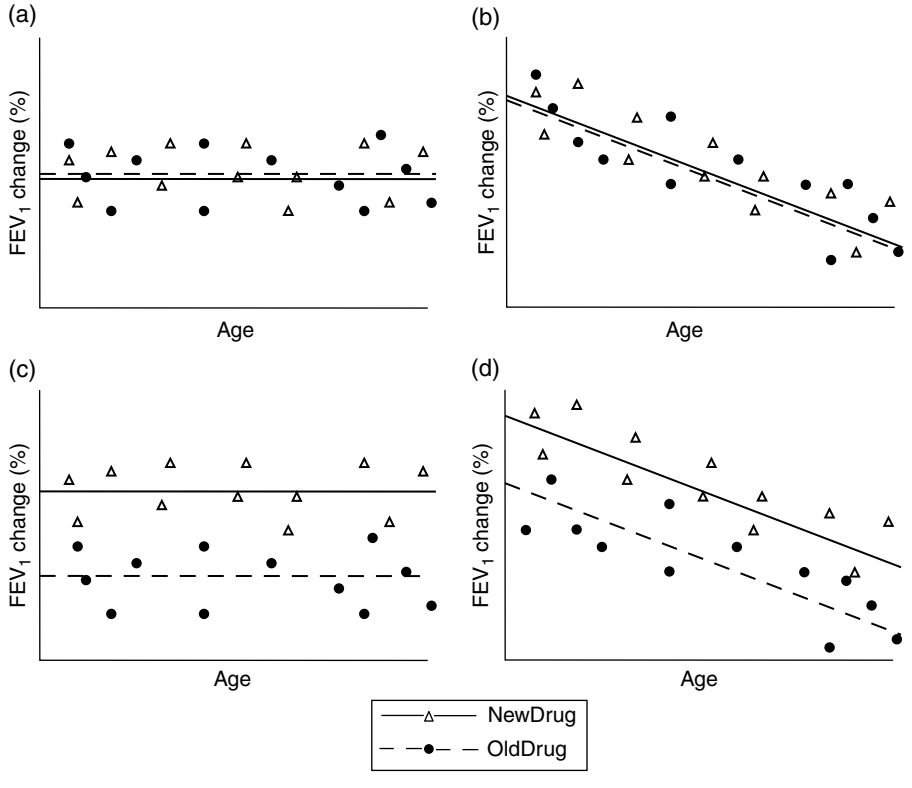


Figure 16.2 Four hypothetical outcomes for the COPD treatment trial (a, b, c, d)

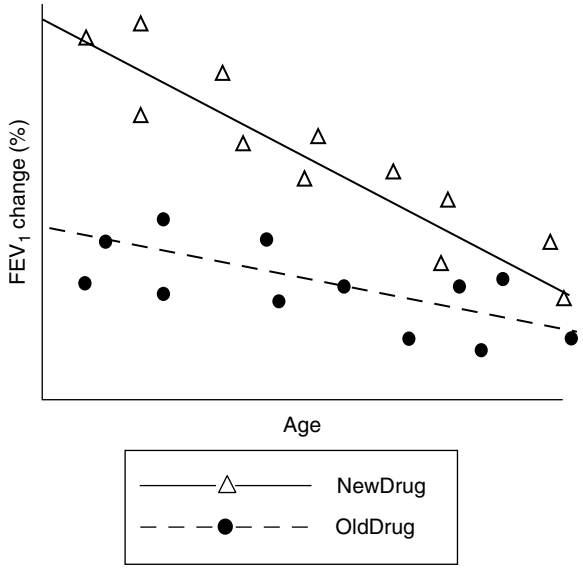


Figure 16.3 An example of interaction (non-parallelism). The difference between the two drugs is much greater at the young end of the age scale than at the older

younger and older subjects, which means that the in both cases the two lines are parallel. In these cases there is therefore no interaction.

Figure 16.3 shows a case where there is interaction. NewDrug produces a considerably greater effect than OldDrug among younger patients but among the oldest patients its superiority is much less marked. Interaction shows up as non-parallelism between the lines.

Graphs can give important insights prior to performing an analysis of covariance

It is always useful to construct graphs such as Figures 16.2 or 16.3 prior to carrying out an analysis of covariance. In particular, marked non-parallelism will show up in the analysis as significant evidence of interaction between the two factors.

16.3 Analysis of the COPD trial results

While the above graphs are important as a first stage in understanding the results, they are no substitute for the formal analysis. Figure 16.4 illustrates the results from Table 16.1. The general appearance is similar to Figure 16.2 (d), that is there probably are effects of both drug choice and age but no interaction.

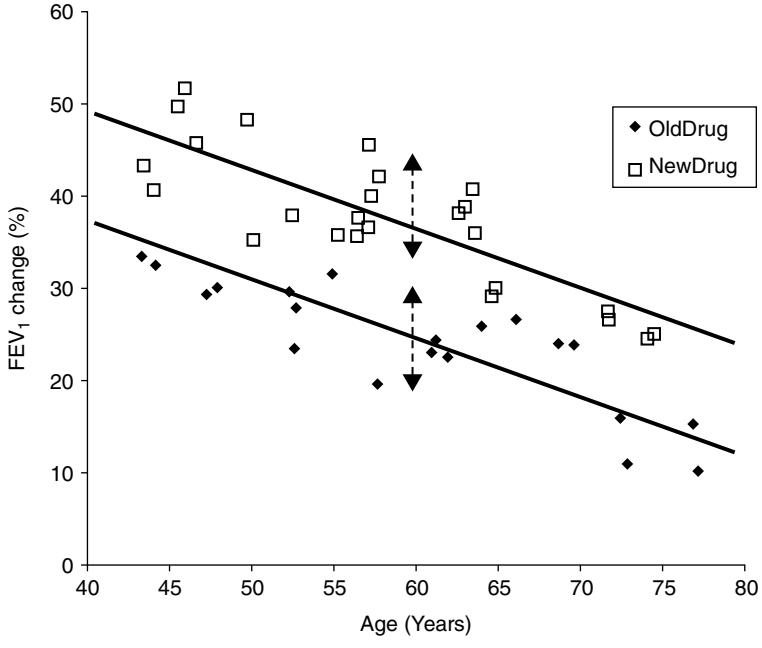


Figure 16.4 Results of the COPD trial. Percentage change in FEV_1 plotted against age with separate lines for each drug treatment group. Note that this shows a 'Common slopes model'. The vertical arrows indicate unexplained variation in response (see Section 16.4.1)

In most statistical packages an ANCOVA is carried using a technique called a General Linear Model. The associated website (www.ljmu.ac.uk/pbs/rowestats/) gives details of how to use SPSS or Minitab to perform ANCOVA. The first step is an analysis which includes the test for interaction. You will need to specify:

- Which variable forms the continuously measured outcome.
- Which variable forms the first factor and whether it is nominal or interval in nature.
- As above for the second factor.
- Specify that interaction is also to be tested for.

This first stage will model the data by establishing a best fitting line for the OldDrug data and then another independent line for the NewDrug data. These lines will almost certainly have different gradients. The main purpose of the first stage is to establish whether the difference between these gradients is large enough to be statistically convincing. If the lines differ markedly in gradient (e.g. as in Figure 16.3),

Table 16.2 Generic example of part of the output of an initial ANCOVA for the COPD trial data. An interaction term was included to allow for differing slopes in the lines fitting the data for the two treatments

Factor	Coefficient	*P* value
Treatment	19.295	0.014
Age	–0.536	0.000
Treatment*Age interaction	–0.117	0.356

then the test for interaction will be significant. Any minor difference will be dismissed as non-significant. For this initial analysis, the data in Table 16.1 will produce output similar to that in Table 16.2 (and probably much more besides!).

At this stage the only output that we should look at is the *P* value for interaction. Other results can be very misleading as they are likely to change dramatically in the next stage of the analysis. The relevant *P* value (0.356) shows no significant evidence of interaction – any slight deviation from perfect parallelism between the two fitted lines could simply be a result of random sampling error.

The initial analysis fitted lines to the two data sets independently and these were allowed to have different gradients. We have detected no convincing difference in the gradients, so we now take a compromise between the two gradients and fit lines through the two data sets, with both lines following this compromise gradient. This is referred to as a 'Common slopes model'. The lines shown in Figure 16.4 follow this common slopes model.

Common slopes model

In the absence of statistically significant evidence of interaction between the nominal and interval factor, it is common practice to model the data by fitting parallel lines through the two data sets. These have a common gradient which is a compromise between the gradients we would determine for the two data sets if they were fitted independently.

To execute a common slopes model, we proceed as above, but exclude interaction from the model. Output from the common slopes analysis will include the details in Table 16.3. This tells us two things:

- There is a statistically significant effect of drug treatment ($P < 0.001$) and that if two patients of the same age received the two treatments, the one receiving NewDrug would (on average) produce a treatment effect 12.4 percentage points greater than that seen in the patient taking OldDrug.

Table 16.3 Generic example of part of the results of a common slopes ANCOVA for the COPD trial data. Interaction term has now been excluded

Factor	Coefficient	*P* value
Treatment	12.405	0.000
Age	–0.594	0.000

- There is also a statistically significant effect of age ($P < 0.001$). The response to treatment (with either drug) tends to decline by 0.594 percentage points for every additional year in a patient's age.

16.4 Advantages of ANCOVA over a simple two-sample *t*-test

It would be perfectly possible to analyse the data from the COPD trial using a simple two-sample *t*-test; we would take account of the different drug treatments but ignore the data concerning age. Although the ANCOVA may be a more complex and less familiar technique than a *t*-test, it can bring significant advantages.

16.4.1 Greater statistical power and narrower confidence intervals

Throughout the preceding chapters there is a consistent theme; unexplained (random) variability in our data introduces uncertainty into our conclusions. The more unexplained variability there is, the less likely we are to achieve statistical significance. This unexplained variability showed up in *t*-tests and analyses of variance as the standard deviation of the data within our samples and in regression techniques as the amount of scatter in our graphs.

If we analysed the data from the COPD trial using a *t*-test, we would be viewing the data as shown in Figure 16.5. We would retain the information concerning which drug was used, but we are now blind to the differences in patients' ages. The vertical arrows indicate the unexplained variability in the two sets of data. All the patients in a particular group received the same treatment and so (ideally) they would all have responded to exactly the same extent. The whole of the within group variability is therefore unexplained.

Figure 16.4 showed the data as we would see it within an ANCOVA. We now have access to the age data and we can identify the fact that much of the variability within each treatment group is, in fact, explicable; it is due to differences among patients' ages. Even then, there is still some unexplained variability; ideally, all the data points would be located exactly on the two fitted lines. The vertical arrows in Figure 16.4 indicate the degree of scatter of the points away from the relevant lines – unexplained variability that remains even when the effects of different drugs and age have been accounted for. The important point is that the extent of this unexplained variability is now much less than that shown in Figure 16.5.

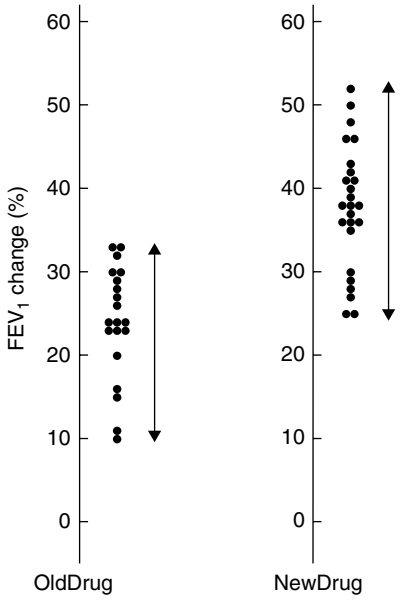


Figure 16.5 Unexplained (random) variability among the responses to treatment if we view the data as we would in a two-sample *t*-test. All within-group variation (vertical arrows) is unexplained

With the COPD data, the superiority of NewDrug over OldDrug is so great that we would achieve statistical significance (with $P < 0.001$), with either a *t*-test or an ANCOVA. However, in more marginal cases, an ANCOVA could well achieve statistical significance where a *t*-test would fail.

This ability of analyses of covariance to reduce unexplained variation also results in narrower confidence intervals for the 95% confidence intervals for the difference between treatments. Figure 16.6 shows the confidence intervals we would obtain if the COPD trial data is analysed by an ANCOVA or a *t*-test. (The difference in the point estimate will be explained in Section 16.4.2.) The important point at this stage is the narrower interval produced by the ANCOVA – by reducing the unexplained variability in the responses, we get a more precise estimate of the difference between the two treatments.

16.4.2 Correction for bias due to baseline imbalances

The patients were randomly allocated to the two treatment groups, so we would expect the two groups to be broadly similar in all regards except the drug received. However, it would be remarkable if characteristics such as age matched exactly. Table 16.1 shows that there was a small discrepancy in average ages (60.25 among

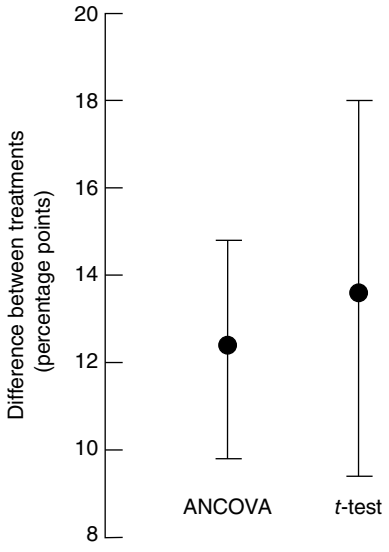


Figure 16.6 95% confidence intervals for the difference in treatment effect between OldDrug and NewDrug as estimated by an ANCOVA or by a two-sample *t*-test

those allocated to OldDrug but only 57.92 among the NewDrug group). Such mismatches are referred to as 'Baseline imbalances'.

The imbalance is small, but since we know that response declines with age, there must be some bias in the result. In this case, the bias will consist of an overestimation of the difference between the drugs; the group taking NewDrug are slightly younger, which will artificially boost the estimate of this drug's effectiveness.

The ANCOVA includes information about ages and so this analysis can recognise and correct for the slight bias. The *t*-test does not include the relevant information and no correction for age imbalance is possible. This is the reason for the discrepancy between the two point estimates for difference between treatments in Figure 16.6. The ANCOVA estimate is lower because it has been corrected; the *t*-test estimate retains the age-related bias.

It is usually possible to set options so that when an ANCOVA is executed, the output includes estimates for the effects of the two drugs that are corrected for any baseline imbalance. The correction for bias would usually be achieved by calculating the average effectiveness of the treatments among two groups of patients of exactly matched average ages. The common age chosen would be the mean age for all the patients pooled together. Table 16.4 shows a typical output. In this case effectiveness of the two drugs has been calculated in two groups both with average ages of 58.96.

In the current example, the baseline imbalance is small and the correction is almost certainly reasonable. However, this approach ceases to be acceptable if imbalances are excessive. Figure 16.7 shows a hypothetical case of severe baseline imbalance. Something has gone badly wrong with the allocation procedure and the

Table 16.4 Effect (percentage points increase in FEV_1) of two drugs when treating COPD. Means corrected for baseline imbalance in ages

Treatment	Point estimate	95% CI lower limit	95% CI upper limit
OldDrug	24.82	22.96	26.68
NewDrug	37.22	35.56	38.89

Calculated for Age = 58.96

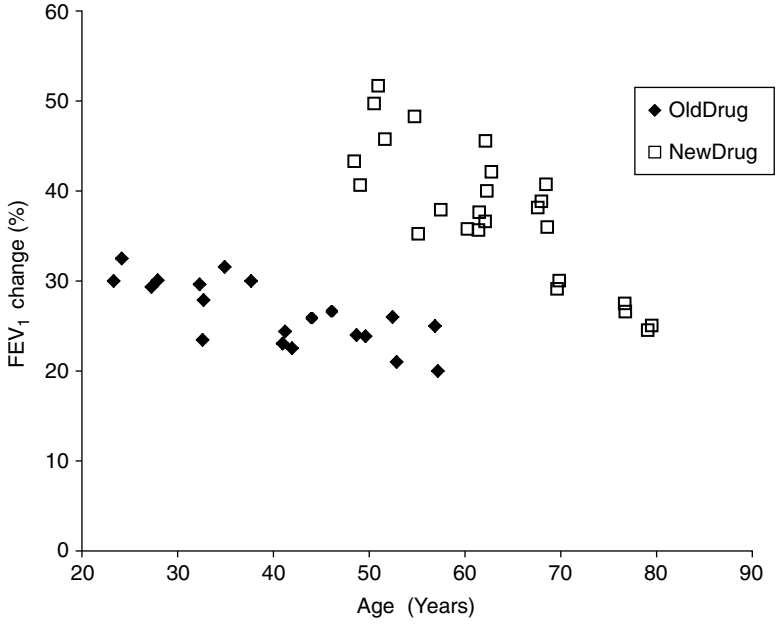


Figure 16.7 Severe baseline imbalance with one group much older than the other which cannot be corrected satisfactorily by ANCOVA

two subject groups are simply not comparable. Attempting to use ANCOVA to correct for age imbalance would effectively involve fitting lines to both data sets and then extrapolating that for OldDrug forwards to 80 years of age and a similar back extrapolation of that for the NewDrug group to 20 years.

Gross baseline imbalances

It is unreasonable to attempt to use ANCOVA to correct for baseline imbalance where there is a gross difference between groups to be compared.

16.4.3 Identification of significant prognostic factors and possible interactions

A final advantage of ANCOVA is that it allows the identification of factors that may influence the effectiveness of a treatment and also any interactions.

Whichever drug is used, we have found that its effect will be lower in older patients. This might mean that the balance between therapeutic benefit on the one hand and cost and possible side effects on the other may shift from a positive situation among younger patients to a detrimental one among the more elderly. We would say that we have identified a 'Prognostic factor' influencing the effectiveness of either drug.

With ANCOVAs we can also identify interactions. Figure 16.8 shows a moderate example of interaction. This is similar to the quantitative interaction we saw with analyses of variance (Section 14.3.8). We can still conclude that NewDrug is more effective than OldDrug for patients of any age, but we can no longer quote a single figure for the extent of superiority. The advantage is much greater among patients in their 40s than among those in their 70s.

Figure 16.9 shows a more extreme case; this now constitutes qualitative interaction. It is no longer possible to make any blanket claim of superiority of NewDrug over OldDrug. By age 75, the dugs' effects are indistinguishable.

With a *t*-test, there would be no opportunity to identify covariates or interactions.

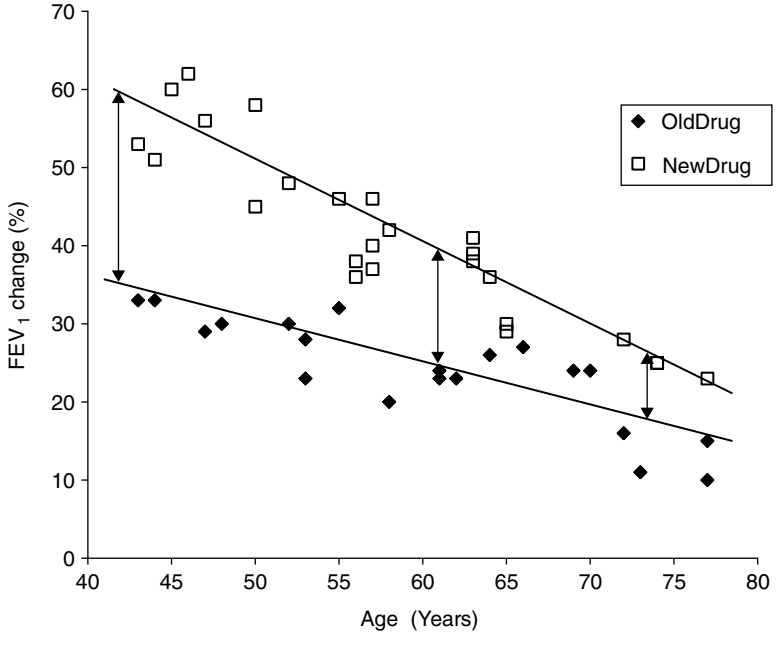


Figure 16.8 Quantitative interaction between treatment and age

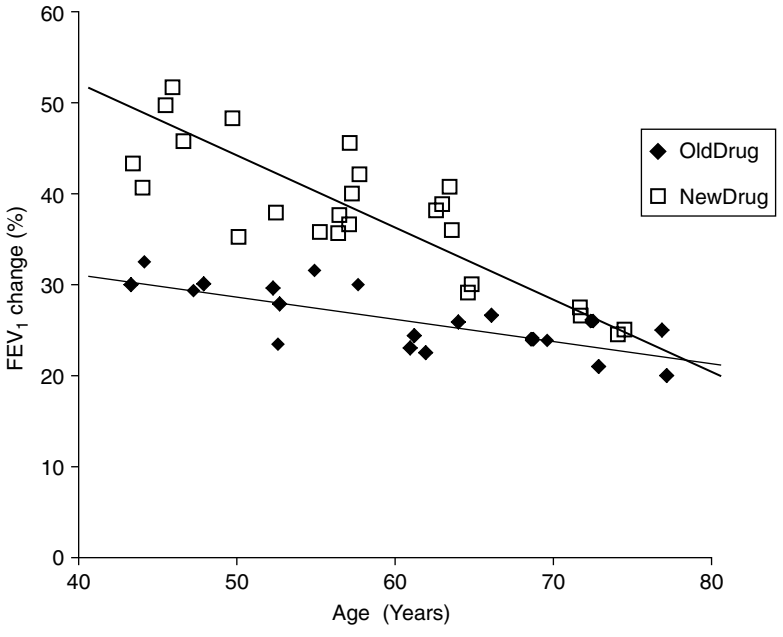


Figure 16.9 Qualitative interaction between treatment and age

16.5 Chapter summary

Analysis of Covariance (ANCOVA) is employed when an interval scale endpoint may be influenced by both a nominal and an interval scale factor. Quite commonly it is the nominal factor that is the real point of interest and the interval factor is included to improve the analysis, however this is not necessarily the case.

It is very useful to inspect the data graphically before performing an ANCOVA as this should clarify the pattern of outcomes, in particular any possible interaction.

The first stage of the analysis includes an interaction term to test whether there is clear evidence of non-parallelism. If the test for interaction is non-significant then we proceed to a 'Common slopes' analysis where we drop the interaction term and force the lines to adopt a common gradient (a compromise between the gradients that the two lines would adopt if independently fitted to the two data sets).

ANCOVA has several advantages over a simple *t*-test:

- Assuming that the interval scale factor is significant, its inclusion in the analysis will reduce the amount of unexplained variability in the outcome. Consequently, an ANCOVA has greater statistical power to detect any effect of the nominal factor than would be available if we used a *t*-test. The 95% confidence interval for the difference between the two treatments will also be narrower with an ANCOVA than with a *t*-test.

- Where there is a modest imbalance between the two study groups in terms of the interval scale factor, an ANCOVA can correct for the imbalance, however ANCOVAs should not be used to attempt to compensate for any gross imbalances.

- Finally, an ANCOVA may allow us to detect significant prognostic factors that may affect the effectiveness of the treatments that are being considered.

Part 3

Nominal-scale data

17

Describing categorised data and the goodness of fit chi-square test

This chapter will ...

- Show how we describe nominal scale data (data that consists of categorisations rather than measurements).
- Describe the production of a 95% confidence interval for a proportion.
- Emphasise the inefficient nature of nominal scale data.
- Describe the use of the goodness of fit chi-square test to determine whether the true proportion of a particular class of things/individuals might credibly be some pre-determined figure.

In this chapter we will start to look at data where no measurements are made on individuals. Instead, each individual is placed in a category and the numbers in each category are then counted. A classic example is where we look at a medical treatment and declare each patient's outcome as 'Successful' or 'Unsuccessful'. We then count the number of successes and failures. This type of data was introduced in Chapter 1 as 'Nominal' scale data.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

17.1 Descriptive statistics

17.1.1 Proportions in each category

To describe measurement data we needed indicators of both the overall magnitude of the values (Mean etc.) and their variability (SD). Describing nominal type data is simpler because all we can report is the proportion of individuals falling into each category. Many categorisations form just two groups, for example Succeed/Fail, Alive/Dead, Male/Female etc. This is referred to as a dichotomisation and we usually call the proportion falling into each category p and q. Which category is allocated to p and which to q is completely arbitrary. For example if 50 patients receive a treatment and 42 were considered to have had a successful outcome (leaving eight unsuccessful), then we might allocate p and q so that:

$$\text{Proportion successful} = (p) = 42 / 50 = 0.84 \text{ or } 84\%$$
$$\text{Proportion unsuccessful} = (q) = 8 / 50 = 0.16 \text{ or } 16\%$$

We have already met P values in the context of significance testing and it is singularly unfortunate that the same letter should be introduced for a second important function, but then that's statistics for you. (To achieve some clarity, I will use lower case 'p' for proportion and upper case 'P' in hypothesis testing.)

17.1.2 What determines the precision of sample estimates of a proportion?

We need to distinguish between sample and population proportions. The data we analyse will almost invariably be sample data collected in a survey or experiment which was intended to estimate the proportions within some wider population. The proportions derived from samples will rarely exactly match those in the underlying population.

As with all data, sampling error depends upon the number of observations; more data means greater precision. Figure 17.1 shows the pattern that emerged with repeated sampling from the same population of individuals, 50% of whom respond successfully to a treatment and 50% were treatment failures. Various different sample sizes were used. With samples as small as ten, there was wide scatter. Some samples contained only 2/10 successes while others contained 8/10, suggesting a success rate of anything between 20 and 80%. Samples this small are just about useless. Even samples of 25 or 50 were often pretty misleading and we really needed samples of a hundred or more before we got tolerably reliable estimates.

The other aspect of the data that affects the precision of an estimated proportion is the rarity of the various categories. Consider a survey of the proportion of patients who suffer an Adverse Drug Reaction (ADR) to a particular medicinal product. The following are two hypothetical outcomes. (In both cases the total number of patients observed is 1000.)

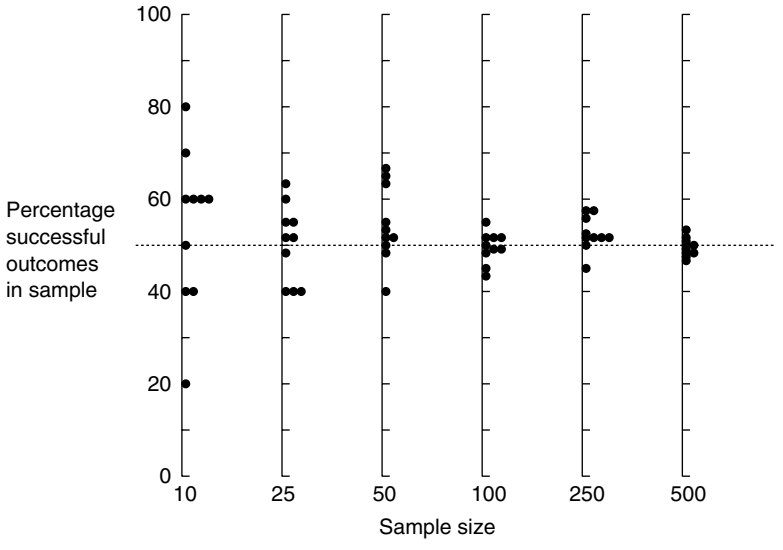


Figure 17.1 Ten sample estimates of the proportion of successful outcomes – Various sample sizes. (True proportion = 50%)

Outcome 1: Without ADR 949; With ADR 51

Outcome 2: Without ADR 996; With ADR 4

In the first scenario we observed 51 ADRs. If we were to repeat the exercise the numbers of ADRs we observed might vary somewhat but they would probably stay in the range of about 40–60; we have enough positive observations to allow a tolerably stable estimate of the rate of occurrence. However, in the second scenario we have only four ADRs and any repeat surveys might easily find (say) two or six positive cases and the estimated rate would now vary over a threefold range. We have the same total amount of data as in the first case, but the estimate is hopelessly imprecise.

Any situation in which there is a rare category will make it difficult to achieve a precise estimate of the proportions in the population.

Precision of an estimated proportion

The precision with which proportions within the population can be estimated from samples, depends on:

- The sample size and
- The presence of any rare categories.

17.1.3 95% C.I.s for a proportion

In an example quoted earlier (Section 17.1.1), we found that 42 out of a sample of 50 patients (84%) showed a successful response to treatment. But, what would happen if we were to adopt this treatment and record the outcomes for thousands of patients over the next few years? The proportion of successful outcomes will (hopefully) settle down to a figure in the region of 84%, but it would be most surprising if our original sample provided an exact match to the long-term outcome. To deal with this, we quote 95% confidence intervals for the proportion in the population based upon a sample proportion.

The essence of the calculation of a 95% confidence interval for a proportion is shown in Figure 17.2. Larger samples are associated with narrower intervals and the presence of rare categories causes greater uncertainty and hence wider intervals.

17.1.4 Using a stats package to produce a 95% C.I. for a dichotomized proportion

Statistical packages differ in the way in which they expect data to be supplied.

Some will only work from the raw data. For our success/fail data, you would provide the data as a column containing the 50 results coded suitably (possibly an 'S' or an 'F' for each success or failure).

Other packages will allow either the above format or 'Summarised data'. In the latter case you indicate the total number of individuals examined and the number who fell into one of the two available categories. (It's completely arbitrary which category you supply.)

In the generic output (Table 17.1) the data has been supplied in summarised form, indicating 50 cases examined and 42 classified as 'Successes'.

The 95% confidence interval is 70.9–92.8%. So, the most we can say is that if this therapy were to be implemented, we would expect the success rate to settle down eventually somewhere within the rather broad range of about 71–93%.

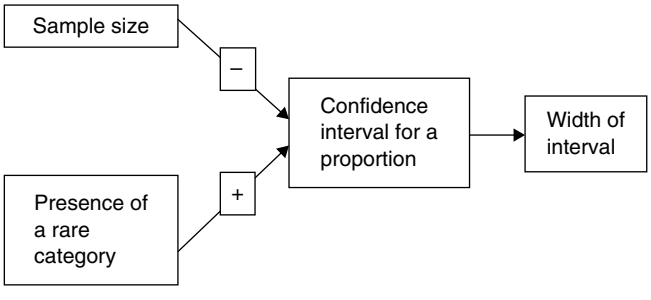


Figure 17.2 Aspects of the data that influence the width for a 95% C.I. for a proportion

Table 17.1 Generic output for calculation of a 95% C.I. for the proportion of successes

95% C.I. for proportions	
Number examined:	50
Number detected (Successful):	42
Point estimate (Successful):	84.0%
95% C.I. for proportion (Successful):	70.9–92.8%

SPSS does not provide any easy way to obtain the confidence interval, but Minitab does (see www.ljmu.ac.uk/pbs/rowestats).

17.1.5 Nominal data is not very efficient

With measurement data, we were often able to draw useful conclusions based on quite small numbers (10 or 15) of individuals. Here we have a somewhat larger sample (50) and yet the conclusion is horribly fuzzy, with a 95% C.I. covering a range of nearly 22 percentage points. Unfortunately it is a characteristic of this type of data that you need to gather an awful lot of it before you achieve any worthwhile degree of precision. Figure 17.1 showed that unless samples were quite large, they really weren't very reliable. This should translate into the widths of 95% C.I.s. Figure 17.3 shows that this is indeed the case. We have a series of samples of increasing size, each of which gave a point estimate of 50% for the success rate. Samples of 10 or 20 are almost useless, because the 95% C.I.s are ridiculously wide. If our target for precision was a C.I. covering no more than ten percentage points (Which doesn't seem overly ambitious), we would need a sample size of almost 500.

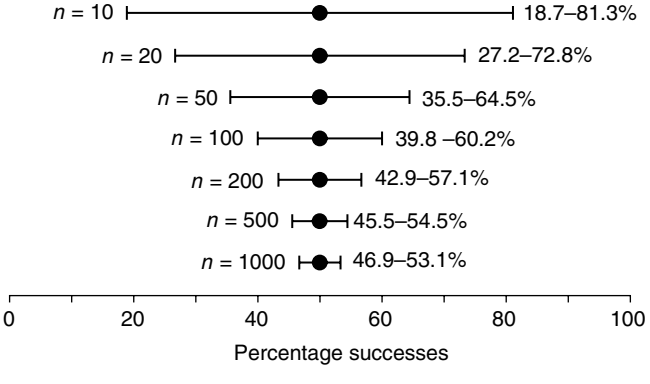


Figure 17.3 95% C.I.s for the proportion of successful outcomes with varying sample sizes. Point estimate equals 50% in all cases

Inefficiency of nominal scale data

You need to gather an awful lot of observations to achieve a useful degree of precision.

17.1.6 The alternative category

If you need a C.I. for the proportion in the alternative category (failures in the above case), it is fortunately just a case of subtracting the values we already obtained from 100%. Thus the confidence limits for the proportion of failures would be 7.2–29.1%.

17.1.7 95% C.I.s for proportions are generally asymmetrical

Figure 17.4 refers back to our trial where 42 out of 50 patients showed a successful outcome. Notice that the interval is asymmetrical.

The asymmetry arises because possible values are more tightly constrained on one side than the other. The upper limit of the interval could not logically be greater than 100%, so the upper limit can't be far above the point estimate of 84%. However, the lower limit could be anything down to 0%. Confidence intervals for proportions are always asymmetrical, unless the point estimate happens to be exactly 50% (as in Figure 17.3).

17.2 Testing whether the population proportion might credibly be some pre-determined figure

17.2.1 Using the 95% C.I. for the proportion

Where individuals are being categorised into two groups, we could test a hypothesis that the proportion in one of the categories was some stated figure by inspecting the 95% C.I. For example, in the case of the condition being treated in section 17.1.1, we

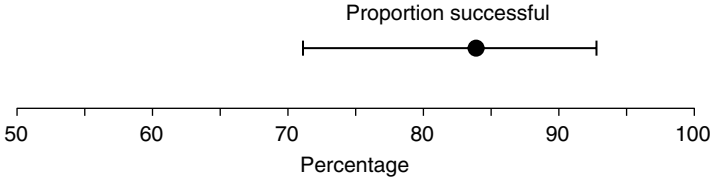


Figure 17.4 95% C.I. for the proportion of successful outcomes (42/50 in sample)

might know that about 50% of individuals will get better spontaneously. Is the success rate with our treatment any different from what would have happened anyway? We would set up a null hypothesis that the true proportion of successes with our treatment is 50%. If the figure of 50% fell within the C.I., the null hypothesis would remain credible. In fact it doesn't, so we have significant evidence that the true proportion of successes with our treatment is not 50%. (Fortunately it is above rather than below the assumed placebo rate of success!)

17.2.2 The goodness of fit chi-square test

The use of the 95% C.I. (as above) is a perfectly reputable approach, but we are going to introduce an alternative approach (the goodness of fit chi-square test). There are two reasons for this:

1. The chi-square test is one of the oldest and most venerable statistical tests and you will certainly come across cases where other people have used it. You may also want to use it yourself as you know your readers are likely to be familiar with it.

2. The chi-square test is flexible; it can easily be applied in situations where cases are categorised into more than two classes.

'Chi' is a letter of the Greek alphabet and chi-square is often written as χ^2. This test is not implemented in all statistical packages, but is so simple to calculate that we can realistically perform it manually.

Goodness of fit chi-square test

Compares the proportion within a sample with some hypothesized proportion for the population. Is the sample data consistent with the specified proportion?

17.2.3 Goodness of fit chi-square test for two categories – Returned pressurised canisters

The example we will consider in detail is a company that produces pressurised canisters of drug for asthma inhalers. Their factory has two machines that manufacture the canisters. One machine (Allegro model) is faster and produces 61% of the factory's output. The slower machine (Andante model) produces the other 39%. Occasional canisters are returned as faulty and, over the previous year, 126 of these have accumulated. From the batch numbers on the canisters we can determine that

52 were manufactured on the Allegro machine and 74 on the Andante. Since only a minority of canisters are made on the Andante machine, and yet it generates the majority of returns, we suspect that the output from this machine is less reliable. However a formal statistical test is required.

First we need a null hypothesis. This will state that the products of the two machines are equivalent and if we went on collecting faulty canisters for long enough the numbers of returns from each machine would simply be proportional to the numbers they produced. More formally it would be: 'Among a large population of returned canisters 61% would be from the Allegro machine and 39% from the Andante'.

To perform the test we first calculate so-called 'Expected frequencies'. These are an idealised set of frequencies calculated to match the null hypothesis exactly. So with our 126 returned canisters, the null hypothesis claims that 61% of these should be from the Allegro and 39% from the Andante machine. Hence, the 'Expected' frequencies are $126 \times 0.61 = 76.86$ (Allegro) and $126 \times 0.39 = 49.14$ (Andante). Notice that the term 'Expected' is being used in a technical sense. There is no expectation, in the normal sense, that we would actually observe these precise figures in the real world (especially as they involve decimal places!).

'Expected frequencies'

These are calculated so as to match exactly with the null hypothesis. There is no literal expectation that we would detect these numbers in a real world experiment.

We then calculate the test as in Table 17.2.

The first two rows show the real 'Observed' figures and the so-called 'Expected' frequencies that would have matched the null hypothesis exactly.

Table 17.2 Calculation of the goodness of fit chi-square test

	Allegro	Andante
Observed	52	74
Expected	76.86	49.14
Obs – Exp (Discrep)	52 – 76.86	74 – 49.14
	= –24.86	= +24.86
$(\text{Obs} - \text{Exp})^2$	618	618
$(\text{Obs} - \text{Exp})^2 / \text{Exp}$	618 / 76.86	618 / 49.14
	= 8.04	=12.58
$\chi^2 =$	8.04 + 12.58 = 20.62	

Note: Yates correction not applied at this stage. Ideally it would be (see Table 17.4)

The key line is the next one, which shows the discrepancies between what we observed and what the null hypothesis led us to expect. The negative figure (–24.86) indicates that we observed less sub-standard canisters from the Allegro machine than we would have expected on the basis of the null hypothesis. The positive figure is the excess of canisters from the Andante machine. It is the size of these numbers that determines whether or not the outcome will be significant. Small discrepancies would indicate that the null hypothesis was in reasonable agreement with what we actually observed and is therefore acceptable. However, large discrepancies suggest that the null hypothesis was completely wrong and should be dismissed. It is this attempt to match up the Observed and Expected frequencies that leads to the test being called a 'Goodness of fit' test.

The rest of the calculation is a classic statistical 'sausage machine'. In the next line the discrepancies are squared and then these are divided by the Expected frequency. In the final line, we sum the last two figures to produce the test statistic – χ^2. A large value for χ^2 would provide sufficient evidence that the null hypothesis is inconsistent with what was actually observed and should be dismissed. Exactly how big the χ^2 value has to be, for the result to be statistically significant, is shown in Table 17.3.

Notice that the required ('Critical') value of χ^2 depends upon the number of categories that have been used. In this case, there were only two categories (Allegro or Andante), so the critical χ^2 is 3.842. Our data set yielded a χ^2 of 20.62. As the value we achieved is way in excess of the critical value, our conclusion is very significant. There is clear evidence that the Andante machine is producing a disproportionately high number of the faulty canisters.

Thus far we know that the result is significant and so P must be less than 0.05, but we do not know its exact value. The easiest way to get an exact figure is to enter the formula below into a cell in Microsoft Excel.

$$=\text{CHIDIST}(20.62,1)$$

The parameter 20.62 is the χ^2 value and the value of 1 is the so-called 'degrees of freedom' of the study. This is simply the number of categories minus one; in this case two categories give one degree of freedom. The result is 0.0000056. We would never quote such a low value, but instead report it as '$P < 0.001$'.

Table 17.3 χ^2 Value required for a goodness of fit chi-square test to be statistically significant ($P < 0.05$)

Number of categories	Critical χ^2 value
2	3.842
3	5.991
4	7.815
5	9.488
6	11.070

17.2.4 The 'Continuity' problem

The mathematical basis of the test includes an assumption that χ^2 values are 'Continuous'. In other words, they could take any value. The reality however is that when we count canisters (or any other set of discrete items) the results are 'Discontinuous' – we may observe one, two or three canisters and so on, but not a fractional value. The subsequent chi-square values are therefore also discontinuous – some values of χ^2 could never arise because they do not match any outcome based upon whole numbers of canisters. This mismatch between the assumptions made by the test and the reality of the data introduces a bias that may inflate the χ^2 value and make the data look a little more significant than it really is.

The most commonly advocated solution to this problem is the Yates correction. However, the use of this correction is somewhat problematic as it is rather drastic and tends to overshoot, sometimes converting a liberal situation (too willing to declare significance) into a conservative one (too reluctant). We need a policy that avoids producing any markedly misleading results and is not so complex or obscure as to arouse suspicions that some sort of statistical fiddle is afoot. A simple and commonly used rule is that we should apply Yates correction only where there are just two categories. With more than two categories, the effect of discontinuity is so small, we are better off not trying to compensate for it.

As our problem with the canisters does have just two categories, the Yates correction should be added. The re-worked calculation, including the correction is shown in Table 17.4. The correction requires adjusting the discrepancies between the Observed and Expected frequencies by 0.5. The correction is applied so as to move the value towards zero. So a negative figure such as –24.86 is adjusted upwards to –24.36 whereas +24.86 moves down to +24.36.

The effect of the correction depends on the magnitude of the data, only being noticeable with very small figures. The numbers involved here are quite modest, but are still big enough to make the effect of the correction minimal. The value of χ^2 is reduced from 20.62 to 19.79 and the result remains clearly significant.

Table 17.4 Calculation of the goodness of fit chi-square test with Yates correction

	Allegro	Andante
Observed	52	74
Expected	76.86	49.14
Obs – Exp (Discrep)	52 – 76.86 = –24.86	74 – 49.14 = +24.86
Obs – Exp (Yates corr)	–24.36	+24.36
$(\text{Obs} - \text{Exp})^2$	593	593
$(\text{Obs} - \text{Exp})^2 / \text{Exp}$	7.72	12.07
$\chi^2 =$	7.72 + 12.07 = 19.79	

Yates correction

Apply the correction only when there are just two categories. Correct the discrepancies between observed and expected frequencies by 0.5, moving the value towards zero.

17.2.5 Cases with more than two categories – Patient preferences among three information leaflets

Dealing with cases where there are more than two categories is a simple extension of the method already shown. An example is given below.

A series of 90 patients are each given three different leaflets explaining the proper use of an inhaler. They are all asked to identify which of the three they considered the easiest to read and understand. The patients have thereby grouped themselves into three categories according to their first preference. We then want to test whether there is any significant evidence of differences among the acceptability of the leaflets. The numbers selecting each leaflet (A, B or C) are shown in Table 17.5.

There appears to be some preference for leaflet B, but we need a formal statistical test to see if the trend is significant. Our null hypothesis is that all leaflets are equally likely to be selected and so our 'Expected' outcome is that each leaflet will be selected by 90/3 = 30 patients. We then calculate χ^2 in the usual way (Table 17.6).

Notice that Yates correction has not been applied as there are more than two categories. According to Table 17.3 (3 categories), χ^2 would need to achieve a value of at least 5.991, for the result to be statistically significant. So, at this stage, we have not positively demonstrated any differences among the leaflets.

Care is needed in deciding what practical action should be taken on the basis of this result. Remember that a non-significant result does nor preclude a difference. Leaflet B has quite a strong lead over its competitors and our experiment may simply have inadequate power to detect a genuine superiority. This is another demonstration of how frustrating this type of data can be; 90 patients recruited and interviewed and we still aren't sure if it matters which leaflet we use!

Table 17.5 Number of patients selecting a leaflet as their first preference

Leaflet	Number of patients
A	23
B	39
C	28

Table 17.6 Calculation of the goodness of fit chi-square test for patients preferring one of three leaflets

	Leaflet A	Leaflet B	Leaflet C
Observed	23	39	28
Expected	30	30	30
Obs – Exp (Discrep)	–7	+9	–2
$(\text{Obs} - \text{Exp})^2$	49	81	4
$(\text{Obs} - \text{Exp})^2 / \text{Exp}$	1.63	2.70	0.13
χ^2		$1.63 + 2.70 + 0.13 = 4.46$	

17.3 Chapter summary

Categorisation (Nominal scale) data are dealt with using statistical methods entirely different from those previously encountered with measurement (Interval scale) data.

To describe such data, we report the proportion of individuals falling into each category. Sample data can be used to estimate the proportion of individuals falling into a given category in the general population, but this process is subject to random sampling error. The extent of sampling error depends upon the numbers observed – greater numbers giving greater precision – and the possible presence of rare categories – it is difficult to estimate the proportion falling into a rare category with any precision. This type of data is rather inefficient. Large numbers need to be observed before any real precision is achieved.

Where the data arises from a dichotomisation, most statistical packages provide a routine to calculate a 95% C.I. for the proportion of individuals in any particular category. Such C.I.s are asymmetrical (except where both categories accounts for exactly 50% of the sample).

The goodness of fit chi-square test can be used to determine whether the population proportion for any category might credibly be some pre-determined figure. The test can be applied to data arising from classification into any number of categories, but if only two categories are being considered, the Yates correction should be applied. The test is not well implemented by all statistical packages, but is simple enough to allow manual calculation.

18

Contingency chi-square, Fisher's and McNemar's tests

This chapter will ...

- Describe contingency tables.
- Demonstrate the use of the contingency chi-square test to detect changes in proportions.
- Advocate the use of simple 2 × 2 tables wherever possible.
- Show how to determine necessary sample size for an experiment that is to be analysed by a contingency chi-square test.
- Introduce Fisher's test as an alternative to the chi-square test for small numbers of observations.
- Describe McNemar's test for use where a study has a paired structure.

In the previous chapter we looked at the goodness of fit chi-square test; we will now move on to the contingency chi-square test. With the goodness of fit test, we hypothesised some proportions (61% of rejected canisters would be from the Allegro machine and each leaflet would be preferred by one-third of the patients). The test then compared the observed outcomes against these theoretical expectations.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

With the contingency test we do not start out with any particular theoretical expectation as to the proportions. Instead we compare one set of observed outcomes against another observed set.

Goodness of fit and Contingency Chi-Square tests

The goodness of fit test determines whether there is a convincing discrepancy between observed and theoretical proportions.

The contingency test determines whether there is a convincing difference between one set of observed proportions and another observed set.

18.1 Using the contingency chi-square test to compare observed proportions

18.1.1 Expulsion rates of IUDs – An example of a contingency table

In this chapter, we want to compare the proportion of individuals who fall into a particular category under two (or more) differing circumstances. For example we might want to compare two groups of women using alternative designs of intra-uterine device (IUD). We want to see if there is any difference in the proportion of women for whom the IUD is accidentally expelled from the womb during the first six months of use. We randomly allocate 4000 women to two equal sized groups. Women in one group are fitted with an existing (Control) design of IUD and those in the other group receive a new (Test) design. After six months, we follow up to determine the outcomes.

The results obtained can be expressed in a so-called contingency table as in Table 18.1.

The characteristic feature of a contingency table is that both the columns and rows are based on categorisations. Here, the columns are based on the category of IUD used and the rows are based on outcomes which are also categorised.

Contingency tables

A table where both the columns and rows are based upon categorisation.

Table 18.1 A contingency table showing the effect of IUD design upon the number of cases where the device was expelled

	Control design	Test design
Not expelled	1732 (86.6%)	1778 (88.9%)
Expelled	268 (13.4%)	222 (11.1%)

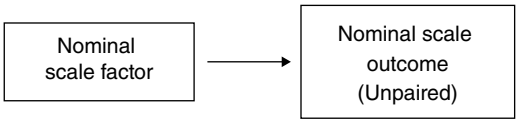


Figure 18.1 Summary of the circumstances in which a contingency chi-square test is appropriate. The comment that the outcome data is Unpaired will be clarified in Section 18.6

The table could be re-oriented so that the two different devices form the rows and the expelled/non-expelled categories form the columns. There is no absolute rule about how such tables should be laid out, but in the author's experience, they are more intuitive if the independent factor is used to form the columns and the dependent forms the rows. This is what has been done in Table 18.1 (expulsion may depend upon device design, but not vice versa). In some tables the two forms of categorisation, though potentially associated, may not have any dependent relationship and the table could equally well be presented in either orientation.

The percentage figures included in the table are column percentages. Thus, in the first column, 86.6% of women using the control IUD did not expel the device but 13.4% did. There is an apparently lower expulsion rate with the new test design (11.1%).

18.1.2 The contingency chi-square test

The data may suggest that the new device is superior to the old one, but these are only samples and the apparent difference could have arisen as a result of random sampling error. We therefore need to set up and test a null hypothesis. Our null hypothesis is that in larger populations of users, the rates of expulsion would be identical. We have two sets of observed proportions and we will compare them using a contingency chi-square test.

Figure 18.1 summarises the circumstances in which a contingency chi-square test is appropriate. The comment that the outcome data is unpaired will be clarified in Section 18.6.

18.1.3 What determines whether we obtain statistical significance?

For this test, two aspects of the data will determine whether the evidence is adjudged significant:

1. *How strong is the contrast between the two outcomes?* If one sample indicated an expulsion rate only marginally greater than that for the other, we would have to

accept that such a small difference could be due to sampling error alone. However, a large difference would be difficult to explain away on this basis – it is more likely to be a real difference.

2. *How precise are our estimates for the rates of expulsion of the two designs of IUD?* If our estimates of the two proportions of expulsions are imprecise then any apparent difference may be illusory. We know from Section 17.1.2 that the precision of these estimates depends upon the sample sizes and the presence/absence of rare categories. So if the estimates are precise (large sample sizes and absence of any rare categories) any apparent difference is likely to be real. But if sample sizes are small and/or there is a problem with a rare category then any difference may be nothing more than random sampling error.

The test will produce a test statistic (chi-square), which will be converted into a *P* value. As with other tests, the stronger the evidence, the greater the test statistic and the lower the *P* value. Figure 18.2 summarises the situation.

Likelihood of significance with a contingency chi-square test

Most likely: A large difference between the proportions in the two groups being compared, large sample sizes and all categories reasonably well represented in the samples.
Least likely: A small difference between the proportions in the two groups being compared, small sample sizes and rarity of one of the categories within the samples.

18.1.4 Using a statistical package to perform the contingency chi-square test

Packages differ in the way in which they expect the data to be presented.

Some will accept only the raw data. For our IUD trial, this would require a column with codes for the type of IUD each woman used (maybe 'C' or 'T' for Control or Test) and a second column with codes for the corresponding woman's outcome (e.g. 'S' or 'F' for Success/retained or Failed/expelled). There would then be 4000 rows of data – one for each subject.

Other packages will allow you to enter the data in summarised format. In that case the numbers of observed cases in Table 18.1 would simply occupy two rows of two columns. Notice however that only the actual counts should be used – not the percentages. If you are not clear just how important the last point is, just imagine entering the percentages. The analysis would have no indication whether the percentages

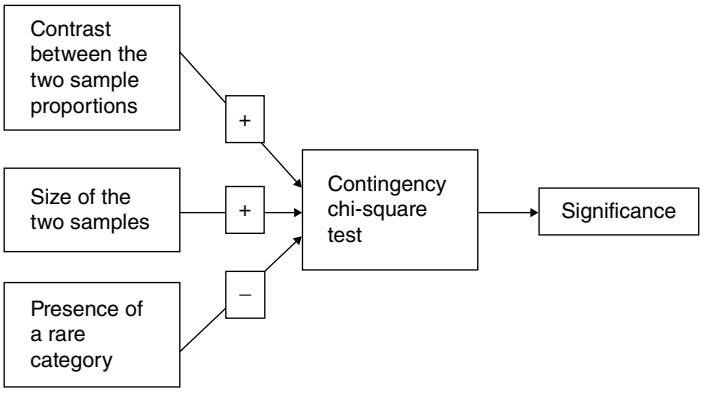


Figure 18.2 Aspects of the data that influence the outcome of a contingency chi-square test

were derived from ten women given each treatment (nowhere near significant) or ten million women (overwhelmingly significant).

Numbers *not* percentages

When entering data in summarised format, you must use the actual counts, not the percentages.

The previous chapter mentioned the 'Continuity' problem and introduced the Yates correction. The same issue arises with the contingency chi-square test. Opinions are divided on the application of the correction to this test. Some statistical packages offer both a corrected and an uncorrected result, others just the uncorrected. A commonly used stratagem is to quote the corrected result where the table contains only two columns and two rows, but for larger tables, the uncorrected result. Fortunately, the correction makes very little difference unless we have results hovering on the verge of significance and the numbers involved are small. For large studies (such as the present case), the correction has only a very slight effect. In this case we should use the corrected value.

Packages tend to produce extensive output, most of which we will refer to later. For now, the key output is the *P* value for the Yates corrected test (0.030 in Table 18.2). There is statistically significant evidence of a difference in expulsion rates between the two designs of IUD.

We would certainly report the *P* value. Additionally, the value of the test statistic (chi-square = 4.710) is commonly included, but it's not vital.

Table 18.2 Generic output from a contingency chi-square analysis of rates of expulsion of two designs of IUD

	Control	Test
Not expelled	1732 (86.6%) *[1755]*	1778 (88.9%) *[1755]*
Expelled	268 (13.4%) *[245]*	222 (11.1%) *[245]*

Expected frequencies indicated in square brackets

Statistical tests

	Chi-square	*P* value
Uncorrected chi-square	4.921	0.027
Yates corrected chi-square	4.710	0.030
Fisher's Exact test		0.030

18.2 Extent of change in proportion with an expulsion – Clinically significant?

Throughout this book, it has been emphasized that we should always try to go beyond simply reporting statistical significance and also consider the extent of any differences between treatments and so assess practical significance. The next chapter will describe the use of measures such as the Relative Risk to convey this information.

18.3 Larger tables – Attendance at diabetic clinics

In the example above there were only two groups to be compared (New and Old design) and there were only two possible outcomes (IUD expelled or not expelled). The results could therefore be expressed in a contingency table with only two columns and two rows (a '2 × 2 table') This is the simplest experimental design and has much to recommend it, as the results are so easy to interpret. However, other experiments may be more complex. For example, we might simultaneously compare more than two designs of IUD and/or there might be more than two outcomes (e.g. not expelled / expelled within one week / expelled after one week or more). The contingency chi-square test is easily extended to cover more complex designs.

As an example of a more complex case, we might want to compare various methods of encouraging diabetic patients to attend their next routine clinic appointment. During their May visit to the clinic, all patients will be issued with an appointment card that details when they should attend in June, but then various additional measures may be taken. Patients are randomly allocated to four groups:

Table 18.3 Attendance at a diabetes clinic following various additional reminders (with column percentages)

	None	Verbal emphasis	Letter	Phone call
Attended	49 (65.3%)	53 (71.6%)	61 (83.6%)	65 (86.7%)
Did not attend	26 (34.7%)	21 (28.4%)	12 (16.4%)	10 (13.3%)

- No additional methods
- Verbal emphasis at end of May visit of importance of attendance
- Letter in June
- Telephone call in June.

We then count those who do or do not attend their June appointment. The null hypothesis is that all four methods are equally effective. The results, expressed as a contingency table are shown in Table 18.3. This is described as a 2 × 4 table. Note that we quote it as (number of rows) × (number of columns).

Ideally the total number allocated to each approach would have been the same. However, this was not quite achieved. This is not a problem, because the chi-square test takes account of different 'column totals'. For optimum power, extreme variation in group sizes should be avoided.

The column percentages have been included and visual inspection of these suggests that the rate of attendance is higher where additional reminders have been used, but a formal statistical test is required to see whether the differences seen in these small samples would continue in the longer term.

18.3.1 The difficulty of interpreting the results from larger contingency tables

The result of a contingency chi-square test of the data in Table 18.3 is a *P* value of 0.006 (Highly significant). The IUD example (Table 18.1) produced results in a simple 2 × 2 table and the interpretation of the significant outcome was perfectly clear. However, in this case this study is more complex and a detailed interpretation of the outcome is problematic. We can reasonably conclude that at least one of the methods for reminding patients was more successful than one of the others, but it is difficult to be any more specific. There are two extreme views that we might take:

- *There are differences in effectiveness between all four methods of notification.* Doing nothing is worst, a verbal emphasis is a bit better, a letter is better still and a phone call best of all.

- *There are only really two degrees of effect.* Doing nothing and verbal emphasis are similarly ineffective and a letter or phone call produce about the same degree of improvement.

Various other intermediate views are also possible.

The output from the contingency chi-square procedure is generally of little help in choosing between interpretations. The moment we start doing experiments that go beyond the simple 2 × 2 contingency table design, our results become contentious and the greater the degree of complexity, the more opaque the outcome.

18.3.2 Sub-dividing large tables

Returning to our survey of possible reminders, let's assume that we would like to introduce a system of letters or phone calls, but are refused the necessary resources, so then it's verbal reminders or nothing. However, looking at Table 18.3, we suspect that despite the statistical significance of the table as a whole, it's doubtful whether there's really any difference between doing nothing and the verbal reminders.

One (contentious) way to test such a question is to sub-divide the table. We could set up a simple 2 × 2 table consisting only of the data for 'None' and 'Verbal emphasis'. A test of that reduced table gives $P = 0.409$, so there is no evidence that verbal reminders would do any good. However, be careful with this approach. If you start breaking down large contingency tables in smaller parts, any number of smaller tables could be created, leading to a high degree of multiple testing. This will increase the risk of a false positive finding creeping in. (Chapter 24 will deal with the problem of multiple testing in greater detail.) With the analysis suggested above, there would be little risk of generating a false positive as we would create only one sub-table. But you definitely should not start with a large, non-significant table and break it up into numerous sub-tables and test them all with a view to finding something significant somewhere.

Break a huge contingency table into every possible sub-table

You start with a massive contingency table, but it emerges as non-significant. Break it up into every conceivable smaller table and test each one. Each sub-table may offer only a 5% chance of a false positive, but with enough of them you're going to get lucky somewhere along the line.

18.3.3 Less experienced researchers gravitate to overly complex tables

In the author's experience, young, naïve experimentalists commonly feel the need to perform hugely complex surveys/experiments with half a dozen different groups being observed and outcomes categorized into a similar number of possibilities. The results then form a 6×6 contingency table or something equally ridiculous. However hard one may try, these over-ambitious pioneers rarely accept warnings that their experimental results will be almost incapable of interpretation. They carry on blithely and it all ends in tears. There seems to be a sense that their research project would look pretty feeble if three months' work could be summarized in a measly 2×2 table – a bigger table would look much more impressive. Supervisors have a major role in constraining such misplaced ambition.

Keep it simple, keep it clear

Where possible, stick to simple experimental designs, where the results can be expressed as a small (preferably 2×2) contingency table. The interpretation of the outcome will then be unambiguous.

18.4 Planning experimental size

18.4.1 Using a statistical package

For those commendably simple experiments that result in 2×2 contingency tables, some statistical packages include simple routines to calculate necessary sample size. For anything more complex, you are on your own – quite right too! The routine will require you to provide values for the following:

- The proportions anticipated in one of the groups to be considered.
- The difference between the two groups that should be detectable.
- Power required.

This is summarised in Figure 18.3. The importance of the first factor is that it will indicate any rare categories. If there is a rare category this will make the detection of any change more difficult and the necessary sample sizes will be greater. As usual, the detection of large differences is easier than for small changes and smaller samples will be adequate and finally if you require more power, you will need larger samples.

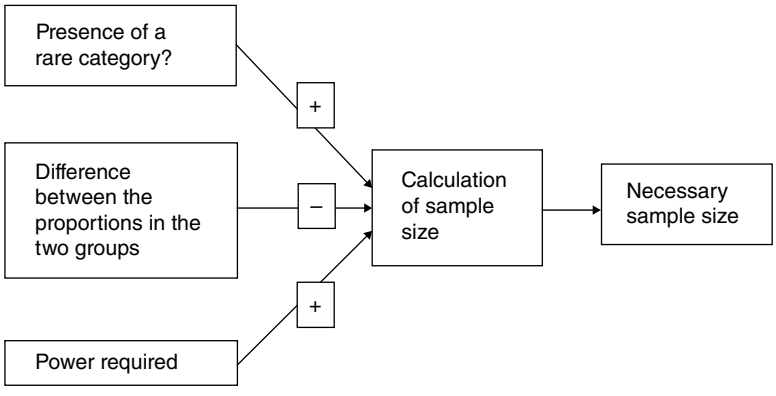


Figure 18.3 Calculation of necessary sample size for a contingency chi-square test

Table 18.4 Generic output from a calculation of necessary sample size for the IUD expulsion trial

Sample size for contingency chi-square test	
Assumed proportion for group 1:	0.15
Target proportion for group 2:	0.11
Target power:	0.90
Sample size (each group):	1484
Achieved power:	0.9001

18.4.2 Calculating the necessary sample size for our IUD expulsion trial

To plan an appropriate size for the trial of two IUDs, we need to establish values for the following:

- *The proportion of expulsions in one of the groups.* Literature reports suggest an expulsion rate of about 15% for the control device. This is generally entered into statistical packages in decimal format (0.15).

- *The change that we would wish to detect.* We will assume that a clinical expert has indicated that a change of four percentage points would be of practical significance. So, if the new device reduced rates to 11%, we would want to detect that change (entered as 0.11).

- *Required power.* We will start with 90% and if that requires unachievable numbers, we'll consider being a bit less ambitious (entered as 0.9).

Generic output is then as in Table 18.4.

The required sample size is 1484. Remember that this is for each treatment group, so we actually need a total of 2968 women. In the real world, we would need to start

out with rather more, to allow for a proportion that can't be followed up. Our original scheme was to use 2000 women in each group – so that was probably a bit generous. As usual, no exact number of subjects delivers precisely the requested power and numbers are set to slightly over-achieve.

18.5 Fisher's exact test

18.5.1 The problem of low expected frequencies

Apart from the controversy surrounding the use/non-use of Yates correction, chi-square testing also generates heated debate about possible inaccuracies caused by low frequencies. The calculation of the test involves some approximations that work satisfactorily for large counts but biases can creep in with lower counts.

We met 'Observed' and 'Expected' frequencies in regard to the goodness of fit chi-square test. The calculations for a contingency chi-square test also involve Expected frequencies and again they are a hypothetical set of values that would exactly fit the null hypothesis – the proportions within any one group exactly match those in all other groups. The first part of the computer output shown in Table 18.2 includes the Observed frequencies and also the calculated Expected frequencies. The expected frequencies maintain the total numbers in each treatment group but redistribute them so that the proportions of IUD expulsions are exactly the same for both designs.

It is important to be aware that it is low Expected frequencies that are the key issue – not the actual Observed frequencies. Low Expected frequencies arise when all the counts in any row or column are low. If a column contains a mixture of high and low counts the Expected frequencies will probably not be a problem. Consequently you should not automatically panic just because there are some low Observed counts; look for low values among the Expected counts.

Low expected frequencies

Low Expected frequencies are a direct cause for concern.

Low Observed frequencies are not a direct cause of concern – we only start to worry if these lead to low Expected frequencies.

18.5.2 Using Fisher's Exact test when there are low expected frequencies

There is general agreement that very low Expected frequencies can create problems, but exactly what constitutes 'low frequencies' is always the basis for a good argument whenever two or more statisticians are gathered together. At one (very cautious) extreme it may be claimed that any Expected frequency less than five is a problem.

Others would argue that the problem is only significant if either (a) several Expected frequencies are less than five or (b) any value is less than one.

Fortunately, Fisher's exact test offers an alternative approach that is free of any concerns about low frequencies. At one time, Fisher's test could be burdensome to calculate for some data sets, but modern computer systems can handle the test without difficulty. The safest approach is therefore to switch to Fisher's test if any expected frequency is less than five. That way nobody is likely to be seriously aggrieved.

If in doubt, use Fisher's exact test

In many cases, it is probably over-cautious, but a policy of switching to Fisher's exact test whenever any Expected frequency is less than five is unlikely to be open to serious criticism.

Figure 18.4 shows when Fisher's test is appropriate.

To exemplify Fisher's exact test we will use the results of an extremely small experiment on the effects of vitamin C supplementation on the incidence of the common cold. Eight subjects were randomised so that four received vitamin C and four got placebo tablets. These were taken from November to April the next year and the outcome was recorded as either did or did not suffer an attack of the common cold during that period. The results are shown in Table 18.5.

There is a lower rate of colds among the vitamin treated group, but common sense would suggest that such small numbers could not possibly be statistically significant. An attempt to analyse the results using a contingency chi-square test are shown in Table 18.6.

All the Expected frequencies are less than five and this is flagged as a warning at the bottom of the output. With all four cells having expected frequencies less than five, most analysts would agree that the chi-square test is seriously compromised. However, Table 18.6 also includes a *P* value for Fisher's test. This can be relied upon even in the presence of low frequencies and it shows (as expected) that the data is nowhere near significant ($P = 0.486$).

The method of calculation of Fisher's exact test is quite different from any other test we have seen. This book generally avoids setting out detailed calculations, but the interested reader can follow the method in the Appendix to this chapter.

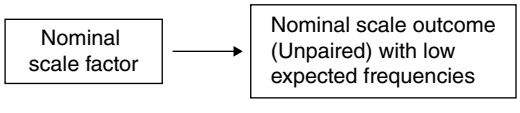


Figure 18.4 Circumstances in which Fisher's test is appropriate

Table 18.5 Occurrence of the common cold in subjects taking either placebo or vitamin C tablets

	Placebo	Vitamin C
No cold	1	3
With cold	3	1

Table 18.6 Generic output from a contingency chi-square analysis of incidence of common cold among users of placebo or vitamin C tablets

	Placebo	Vitamin C
No cold	1 (25.0%)	3 (75.0%)
	[2]	*[2]*
Cold	3 (75.0%)	1 (25.0%)
	[2]	*[2]*

Expected frequencies indicated in square brackets

Statistical tests

	Chi-square	*P* value
Uncorrected chi-square	2.000	0.157
Yates corrected chi-square	0.500	0.480
Fisher's Exact test		0.486

Warning: four cells have Expected frequencies less than 5

Programmes such as SPSS and Minitab can be used to obtain results for Fisher's test. See www.ljmu.ac.uk/pbs/rowestats for detailed instructions.

18.6 McNemar's test

18.6.1 Categorical outcomes with a paired structure

We have already met the two-sample and paired *t*-tests. The example scenario analysed by a two-sample test involved one group of subjects receiving rifampicin and another separate group receiving placebo (Chapter 7). In contrast, the paired test was applied to a single set of subjects who were studied both after placebo and active weight loss treatment. These were described as unpaired and paired studies respectively.

Studies with categorical endpoints can also be paired or unpaired. The IUD study was unpaired – one group of women used the control design of IUD and another separate group used the new design. McNemar's test is applied where the design is paired and the outcome is categorical.

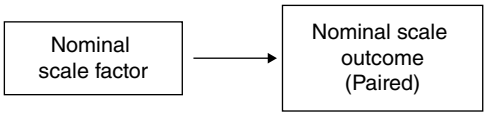


Figure 18.5 Use of McNemar's test for categorical outcomes with a Paired data structure

McNemar's test

When comparing categorical outcomes among two groups and the data is paired, use McNemar's test.

Figure 18.5 summarises the circumstances in which McNemar's test should be used in place of a contingency chi-square test.

18.6.2 Training in inhaler use – McNemar's test

Twenty-five patients with asthma are assessed for the quality of their inhaler use. There is a range of abilities, but the results are recorded as two broad categories 'Satisfactory' and 'Unsatisfactory'. All subjects then receive training and are given a leaflet to reinforce the training. After three months they are re-assessed and categorised as previously. The question we want to answer is whether there is significant evidence of a change in outcomes after training. Our null hypothesis would be that among a large number (population) of patients there would be no systematic trend – the number of patients changing from unsatisfactory to satisfactory would be exactly balanced by the number changing in the opposite direction.

The results are shown in Table 18.7. Notice that the results are presented in a special way that preserves the paired structure of the trial. There are nine patients whose technique was unsatisfactory prior to training and their technique remained unsatisfactory. There are also eight who were satisfactory both before after training. These results are described as 'Concordant' – the same classification on both occasions. The data in the other two cells are 'Discordant'. Seven patients were classed as unsatisfactory prior to training but (happily) were transformed to satisfactory afterwards. Sadly there was also one who deteriorated from satisfactory to unsatisfactory.

The concordant cases are considered uninformative as they give no indication of whether performance was better pre- or post-treatment (there were no changes in classification). The analysis therefore focuses on the discordant cases. In our sample there were a total of 1 + 7 = 8 discordant cases. These are our Observed frequencies. We then need to generate 'Expected' frequencies that exactly match the null hypothesis (change in one direction is exactly equal to that in the opposite direction). Our Expected frequencies are therefore that there would be 8 / 2 = 4 cases of conversion

Table 18.7 Effects of training on the quality of inhaler use among asthma patients

		Pre-training	
		Unsatisfactory	Satisfactory
Post-training	Unsatisfactory	9	1
	Satisfactory	7	8

Table 18.8 Calculation of the McNemar test for the effect of training on inhaler technique

	Unsatisfactory changing to Satisfactory	Satisfactory changing to Unsatisfactory
Observed	7	1
Expected	4	4
Obs. – Exp. (Discrep.)	7 – 4 = 3	1 – 4 = –3
Obs. – Exp. (Yates corr.)	2.5	–2.5
$(\text{Obs.} - \text{Exp.})^2$	6.25	6.25
$(\text{Obs.} - \text{Exp.})^2 / \text{Exp.}$	1.5625	1.5625
Chi-square	1.5625 + 1.5625 = 3.125	

from unsatisfactory to satisfactory and an equal number in the opposite direction. The calculation is then identical to that for the goodness of fit chi-square (see Section 17.1.5). The calculation is shown in Table 18.8. Notice that as there are only two categories (Satisfactory and Unsatisfactory), the Yates correction is applied.

With two categories, the critical value of chi-square is 3.842 (Table 17.3). In this case, the achieved value is 3.125 and there is not quite sufficient evidence to achieve formal statistical significance. The exact *P* value is 0.07.

The principal advantage of the paired *t*-test over the two-sample test was its greater statistical power. Similarly, the McNemar test is more powerful than the contingency chi-square test when data has a naturally paired structure.

Programmes such as SPSS and Minitab can be used to obtain results for McNemar's test. See www.ljmu.ac.uk/pbs/rowestats for detailed instructions.

18.7 Chapter summary

A contingency table presents data that is based entirely upon categorisation (both the columns and the rows are in this format).

The contingency chi-square test is used to determine whether the proportion of individuals falling into a particular category changes according to circumstances.

When performing this test the actual counts (not percentages) must be used. Where the data forms a simple 2 × 2 table, the Yates correction should be applied.

Results of the test are unambiguously interpretable when it is applied to the smallest (2 × 2) tables. The test can be applied to larger tables, but the interpretation will be less clear.

For experiments that can be described in a 2 × 2 contingency table, it is simple to calculate necessary sample sizes.

Low Expected frequencies can compromise the chi-square test. If even one cell has an Expected frequency of less than five, the safest approach is to switch to Fisher's Exact test.

If a study structure is paired (typically the same individuals studied under two sets of circumstances) and the endpoint is a categorisation, the results should be tested using McNemar's test.

18.8 Appendix

18.8.1 Calculation of Fisher's exact test

The Fisher's Exact test is fully calculated below for the example data set out in Table 18.5

We start from the null hypothesis that taking vitamin C has no effect on the likelihood of catching a cold. If the treatment is having absolutely no effect, we should be able to treat all subjects as equivalent. The test looks to see what would happen if we added patients to the results table in a random manner. In particular, it investigates how often we produce a table that is as strongly suggestive of a treatment effect as we actually saw in Table 18.5. If only a small proportion of our random arrangements of the patients (less than 5%) match the strength of evidence in the real table, we would conclude that *P* is less than 0.05 and the results are statistically significant.

We place a restriction that the random arrangements have to maintain the column and row totals seen in the original table. In that table, all the column and row totals were four. There are five possible tables that fulfil that requirement. These are shown in Table 18.9 parts (a) to (e).

However, if we are randomly feeding individual patients into these tables, some of the tables are more likely to arise than others.

If we start with the first part (a), we have four subjects receiving placebo – call these subjects A, B, C and D. There is only one way that we can achieve the first column of the first table and that is that all four subjects fall into the 'With cold' category. Similarly there is only one way to achieve the second column – all fall into the 'No cold' category. There is therefore only one way to arrange the eight subjects to achieve the first table.

For the second part (b), the first column could be achieved in four different ways. Subject A could fall into the 'No cold' category leaving the other three as 'With cold'

Table 18.9 (a) to (e) Possible arrangements of eight outcomes so as to retain the data structure seen in Table 18.5 (total of four individuals in all columns and rows)

	Placebo	Vitamin C	Number of ways to achieve this table	Probability of randomly achieving this table
(a)				
No cold	0	4	1	1/70 = 0.014
With cold	4	0		
(b)				
No cold	1	3	16	16/70 = 0.229
With cold	3	1		
(c)				
No cold	2	2	36	36/70 = 0.514
With cold	2	2		
(d)				
No cold	3	1	16	16/70 = 0.229
With cold	1	3		
(e)				
No cold	4	0	1	1/70 = 0.014
With cold	0	4		

and similarly we could have B or C or D as the single 'No cold' case. The same logic applies to the second column. As there are four ways to achieve both columns, there are $4 \times 4 = 16$ ways to achieve the table as a whole.

For the third part (c), there are six ways to achieve the first column. The two categorised as 'No cold' could be A+B, A+C, A+D, B+C, B+D or C+D and in each case the remaining two would constitute the 'With cold' category. There will also be six ways to achieve the second column and there will be $6 \times 6 = 36$ ways to arrive at the table as a whole.

The fourth and fifth parts, (d) and (e), follow the logic for the second and first and are achievable in 16 ways and one way, respectively.

In total there are $1 + 16 + 36 + 16 + 1 = 70$ ways to achieve a table that maintains the original data structure with four cases in all columns and rows. We can then calculate the relative likelihood of each of the tables. For example the first table will arise by only one of the 70 possible arrangements, so its probability is $1/70 = 0.014$. The probabilities of each of the other tables are indicated in Table 18.9.

The next step is to identify which tables present evidence of a difference in outcomes between the two treatments that is as strong as (or stronger than) the actual table observed.

Part (a) fulfils the above requirement; it would show even clearer evidence of a difference and part (b) is the one actually observed so it also matches the requirement. Part (c) shows no evidence of any difference, so it is excluded. Parts (d) and

(e) do fulfil the requirement; they provide evidence of a difference between treatments of the necessary strength. In these cases the evidence happens to point to a change in the opposite direction from that in the original table, but they are showing evidence of a difference.

So, the level of evidence that we actually achieved is matched (or exceeded) in strength by that in parts (a), (b), (d) and (e). The joint probability that one or other of these would arise if we randomly allocated subjects into the table is therefore $0.014 + 0.229 + 0.229 + 0.014 = 0.486$.

The *P* value is defined as the likelihood that we would achieve evidence as strong as (or stronger than) that actually seen, if the null hypothesis were true. The calculation that we carried out above has therefore arrived at precisely what the *P* value is supposed to measure. The *P* value for the results of our small trial is 0.486. The result is non-significant.

19

Relative risk, odds ratio and number needed to treat

This chapter will ...

- Describe the calculation and interpretation of the Relative Risk, Odds Ratio and Number Needed to Treat.
- Introduce 95% C.I.s for the Relative Risk and Odds Ratio.
- Discuss difficulties associated with attempts to calculate confidence intervals for the Number Needed to Treat

19.1 Measures of treatment effect – relative risk, odds ratio and number needed to treat

This book always emphasises that we should not only ask *whether* a treatment produces an effect, but we should also ask '*How great is the effect*?' That way we can judge whether any changes seen are practically significant. In this chapter we look at various measures of the size of change in a categorical (nominal) outcome. The methods we will discuss are mainly applicable to dichotomous outcomes (e.g. Success/Failure).

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

Table 19.1 The effect of IUD design upon the number of cases where the device was expelled

	Control design	Test design
Not expelled	1732	1778
Expelled	268	222
Total	2000	2000

19.1.1 Relative Risk for the expulsion of intrauterine devices comparing test to control design

In the previous chapter, Table 18.1 presented results comparing the rates of expulsion for test and control designs of intrauterine device (IUD). Table 19.1 presents the same data, but additionally includes the totals for both treatment groups as these are required for the calculations in this chapter.

Risk is defined as the ratio of the number of cases where the relevant event occurred to the total number of subjects observed. For the purposes of the present study we will consider expulsion of the IUD to be the event.

Definition of Risk:

$$\text{Risk} = \frac{\text{Number where event occurred}}{\text{Total subjects observed}}$$

The risks for the two groups are therefore:

$\text{Risk}_{\text{Control}} = 268/2000 = 0.134$ $\text{Risk}_{\text{Test}} = 222/2000 = 0.111$

The risks can be reported as decimals or as percentages, so either 0.134 and 0.111 or 13.4% and 11.1% are acceptable.

The Relative Risk (RR) is then the ratio of the two risks. Normal practice is to place the risk for the new or active treatment on top of the fraction and use the old or control figure as the divisor.

Calculation of Relative Risk

$$\text{RR} = \frac{\text{Risk in group receiving new or active treatment}}{\text{Risk in group receiving old or control treatment}}$$

In this case RR = $\text{Risk}_{\text{Test}}$ / $\text{Risk}_{\text{Control}}$
= 0.111 / 0.134
= 0.828

When reporting a relative risk, it is important to avoid ambiguity.

- Define what constitutes the 'Event'. We have taken the event to be expulsion, rather than retention of the IUD.

- Define which way the relative risk has been calculated. We have followed normal practice and expressed the risk for the test design relative to that for the control.

A good description is therefore 'The Relative Risk of expulsion for the test design compared to the control was 0.828.'

The null hypothesis is that the risk of expulsion is the same in both groups and so the null hypothesis value for the RR is 1.0. Values below 1.0 indicate a lower risk with test design. Any value greater than 1.0 would suggest an increased risk. The figure of 0.828 indicates a modest reduction in risk of expulsion with the new design.

Relative Risk is sometimes called the Risk Ratio. They are exactly the same thing and it is fortunate that both give the initials RR.

19.1.1.1 The 'risk' of being cured In common parlance the term 'Risk' is decidedly negative – the likelihood that something unpleasant will transpire. Unfortunately, statisticians tend to use the term in a technical sense – the likelihood of any event, be it good or bad. Hence if a drug treatment may produce a cure, the relevant event might quite reasonably be taken to be a cure and this may generate a reference to the risk of a cure (Number of patients experiencing a cure/total number treated).

This also means that Relative Risk can be counterintuitive. If a new treatment increases the likelihood of a cure, then the Relative Risk for the new treatment will be greater than 1.0, but this is actually good news – greater 'risk' of a cure.

19.1.1.2 Relative Risks greater than 1.0 In the example quoted, the experimental data (New design of IUD) was associated with less events than the control and so the RR was below 1.0. In other circumstances, the existence of a treatment effect may be revealed by an RR greater than 1.0. For example, in a comparison of rates of dry cough in patients treated with active or placebo ACE inhibitors, the RR for a dry cough among active drug users relative to those receiving placebo might be (say) 3.0.

19.1.2 Odds Ratio for the expulsion of IUDs

The odds are defined as the ratio of the number of cases where the relevant event occurred to the number where the event did not occur.

Definition of Odds:

$$\text{Odds} = \frac{\text{Number where event did occur}}{\text{Number where event did not occur}}$$

The odds for our two groups of IUD users are therefore:

$$\text{Odds}_{\text{Control}} = 268/1732 = 0.1547 \quad \text{Odds}_{\text{Test}} = 222/1778 = 0.1249$$

As previously, the Odds Ratio (OR) is calculated for the new or active treatment relative to the old or control case. For our study:

$$\text{OR} = 0.1249/0.1547 = 0.807$$

As with the RR, the null hypothesis value for the OR is 1.0 and the current result indicates reduced odds of expulsion with the new design.

19.1.3 Number Needed to Treat with new design of IUD

19.1.3.1 Purpose of the Number Needed to Treat The Number Needed to Treat (NNT) tells us how many patients we would need to transfer to a new treatment, in order to achieve one additional positive outcome. It has obvious value in pharmacoeconomics.

19.1.3.2 Calculation of the NNT The calculation is in two stages. First calculate the Absolute Risk Difference (ARD). For this we subtract one risk from the other:

$$\begin{aligned}\text{ARD} &= \text{Risk}_{\text{Test}} - \text{Risk}_{\text{Control}} \\ &= 0.134 - 0.111 \\ &= 0.023\end{aligned}$$

This means that by swapping to the new design a woman reduces her risk of IUD expulsion by 2.3 percentage points.

The second stage is to calculate how many women would have to make that swap from control to new design if we are to reduce the number of expulsions by one:

$$\begin{aligned}\text{NNT} &= 1/\text{ARD} \\ &= 1/0.023 \\ &= 43.48\end{aligned}$$

Conventionally we always round the result upwards and NNT becomes 44. If 44 women were to be swapped from the old to the new design, we would expect to see one less case of IUD expulsion.

General definition of the Number Needed to Treat

The number of individuals who would need to be transferred from one treatment to another in order to prevent one harmful event or produce one additional beneficial outcome.

19.2 Similarity between relative risk and odds ratio

The calculations for risk and odds are:

$$\text{Risk} = \frac{\text{Number with the event}}{\text{Number of subjects observed}} \qquad \text{Odds} = \frac{\text{Number with the event}}{\text{Number without the event}}$$

The numerator is the same in both cases. The denominators are defined differently, but if the event is rare, the number of subjects observed and the number without the event will be very similar. Thus, so long as the event is rare, the risk and the odds take very similar values. This also means that the Relative Risk and Odds Ratio will take similar values. IUD expulsion was a reasonably rare event and consequently the RR (0.828) and the OR (0.808) do not differ greatly.

As a contrast, consider a fairly common event such as diarrhoea among children in a third-world nation. We provide some mothers with water sterilising tablets while others act as controls (no tablets). We follow up for 12 months recording any cases of diarrhoea. The results are shown in Table 19.2.

We then calculate the RR and OR as:

$\text{Risk}_{\text{Control}} = 798/1110 = 0.719 \quad \text{Risk}_{\text{Treated}} = 345/1175 = 0.294$

Relative Risk = 0.294/0.719 = 0.409

Table 19.2 The effect of providing water sterilising tablets on rates of childhood diarrhoea

	No tablets	Tablets supplied
No diarrhoea	312	830
Diarrhoea	798	345
Total	1110	1175

$\text{Odds}_{\text{Control}} = 798/312 = 2.558 \quad \text{Odds}_{\text{Treated}} = 345/830 = 0.416$

Odds Ratio = 0.416/2.558 = 0.163

With this common event, the RR and OR are markedly different (0.409 versus 0.163).

For both the RR and the OR, we compare the result against a null hypothesis value of 1.0. The danger is that the OR (0.163) tends to make the results look more dramatically different from unity than the RR (0.409), even though it is exactly the same data that has been analysed in both cases.

Similarity between the value of the Relative Risk (RR) and the Odds Ratio (OR)

Where an event is fairly rare, the RR and OR take similar values, but with commoner events, the OR deviates further from the null hypothesis value of 1.0 than is the case for the RR.

Beware of drug companies bearing Odds Ratios

You want to make the effectiveness of your new treatment look as dramatic as possible. A set of clinical trial results could be represented either as a Relative Risk of 0.409 or as an Odds Ratio of 0.163. You know that in either case, the reader will asses the effectiveness of the treatment by comparing the figure you quote against a null hypothesis value of 1.0. Which are you going to quote, 0.409 or 0.163? It's a no brainer.

19.3 Interpreting the various measures

19.3.1 Interpreting the Relative Risk

Relative Risk is easily understood. For example, in the childhood diarrhoea example, the likelihood that a child will suffer an attack can be reduced to about 41% of its control level by providing sterilising tablets.

19.3.2 Interpreting the Odds Ratio

IUD expulsion is a relatively rare event and so, as explained in Section 19.2, the Odds Ratio will be so similar to the Relative Risk that it can safely be interpreted in the same way – the OR can effectively be read as if it were the RR. However,

with a commoner event such as childhood diarrhoea, the meaning of the OR is far from intuitive.

For all the difficulties in interpretation that may plague the Odds Ratio, there are circumstances where it is unavoidably the measure of effect size with which we have to make do.

- For case-control epidemiological studies, it is only possible to calculate the Odds Ratio.

- The next chapter describes logistic regression where the effect measure we obtain is the Odds Ratio.

19.3.3 Interpreting the Number Needed to Treat

The main value of the Number Needed to Treat lies in health economics. If, for example, we know that it would cost £35 to provide a new treatment to an individual patient and the NNT is 25 and the event in question is death, then in order to prevent one death we will have to treat 25 patients at a cost of 25 × £35 = £875. Most people would probably agree that this was money well spent. The arguments begin when the cost per patient is high and/or the NNT is large.

The NNT can provide a valuable reality check when an event is rare. Consider, for example, an existing drug that causes a particular side-effect in one in 1000 patients and a new drug that causes the problem in only one in 2000 patients. The RR for the side effect will be 0.5 for the new drug relative to the older one. Considered in isolation, that figure of 0.5 sounds like an impressive effect size – surely we should switch to the new drug (even if it is more expensive). However, the NNT warns us that we would need to switch 2000 patients from the old to the new drug to prevent one case of the side effect. Unless the side effect is very serious and/or the new drug doesn't cost much more than the old one, the wisdom of a drug switch is in fact questionable.

19.4 95% confidence intervals for measures of effect size

Any figures we arrive at for an RR or OR and so on will always be based upon sample data and the true long-term effect may well be somewhat less than or greater than the sample estimate. As usual, we allow for this uncertainty by quoting a 95% confidence interval for the measure.

The precision of a measure such as the RR or OR will depend upon the precision with which we determined the event rates in the two groups being compared and as explained in Section 17.1.2, this will depend upon the sample sizes and the existence of any rare categories. If the sample sizes are small and/or one of the categories is rare, the confidence interval will be wide.

19.4.1 Confidence intervals for the RR, OR and NNT for clearly significant results (e.g. childhood diarrhoea trial)

Some of the better known statistical packages cannot easily be used to obtain confidence intervals for endpoints such as the Relative Risk, but the free analytical package 'Epi Info' distributed by the United States Government's Centers for Disease Control and Prevention does a good job in this respect. An XL spreadsheet that calculates the confidence intervals is available from the website associated with this book (www.ljmu.ac.uk/pbs/rowestats/).

Whatever package is used, the IUD expulsion data (Table 19.1) would lead to output similar to Table 19.3. It looks strange that the confidence limits for the NNT are both 3; the explanation is that the exact limits are 2.16 and 2.58, which both round up to 3.

19.4.2 Statistical significance and 95% confidence intervals

Figure 19.1 shows the 95% confidence intervals for the Relative Risk, Odds ratio and Absolute Risk Difference for the childhood diarrhoea trial (Table 19.2). The null hypothesis value for the RR and OR are both 1.0. As the Absolute Risk Difference is calculated by subtracting one risk from the other, its null hypothesis value is zero. These are indicated in Figure 19.1. All three intervals exclude their relevant null hypothesis values and therefore show the data to be statistically significant.

The childhood diarrhoea data could also be tested by a contingency chi-square test as described in Section 18.1.2. This test would give $P < 0.001$ (strongly significant). So, in this case, all four possible methods to test the data (chi-square, and the three 95% confidence intervals) agree with one another – the results are statistically significant. The mathematical calculations involved in producing the confidence intervals for the RR, OR and ARD dictate that they will always agree. They will either all be significant or all non-significant; they will never produce discordant conclusions. The chi-square test will generally agree with the conclusions from the confidence intervals, but it does use a quite different method of calculation and in marginal cases, there could be a disagreement.

Table 19.3 Measures of treatment effect with 95% confidence intervals for the childhood diarrhoea trial with water sterilising tablets

	Point estimate	Lower limit	Upper limit
Relative Risk	0.409	0.371	0.450
Odds Ratio	0.163	0.136	0.195
Absolute Risk Difference	0.425	0.387	0.463
Number Needed to Treat	3	3	3

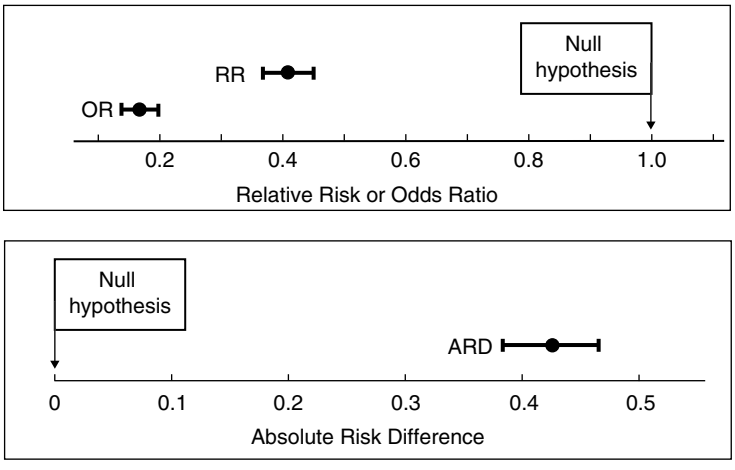


Figure 19.1 95% confidence intervals for the Relative Risk (RR), Odds Ratio (OR) and Absolute Risk Difference (ARD) for childhood diarrhoea (comparing families with or without access to water sterilising tablets) along with their null hypothesis value

Significance testing using the 95% confidence interval for the RR, OR or Absolute Risk Difference

Statistical significance can be assessed by checking whether the 95% confidence intervals for the RR, OR or ARD exclude their respective null hypothesis values. The results of any one of these will always concur with any other.

Results can also be tested using the contingency chi-square test. The result will generally agree with that from the confidence intervals, but in marginal cases a disagreement is possible.

19.4.3 Confidence intervals with marginally significant or non-significant data (e.g. IUD expulsion trial)

The results of the IUD expulsion trial were analysed by contingency chi-square test in Section 18.1.2 and were found to be statistically significant, but this was a much less clear cut outcome than we saw with the childhood diarrhoea trial ($P = 0.030$ versus $P < 0.001$). Table 19.4 shows the effect measures and their 95% confidence intervals for the IUD trial.

The output for the RR and the OR are perfectly straightforward and very much as we might expect; they confirm statistical significance, but the result is marginal (upper limits of the intervals are close to the null hypothesis value of 1.0).

Table 19.4 Measures of treatment effect with 95% confidence intervals for the IUD expulsion trials (new versus old design)

	Point Estimate	Lower limit	Upper limit
Relative Risk	0.828	0.701	0.979
Odds Ratio	0.807	0.667	0.976
Absolute Risk Difference	0.0230	0.0027	0.0433
Number Needed to Treat	44	24	372

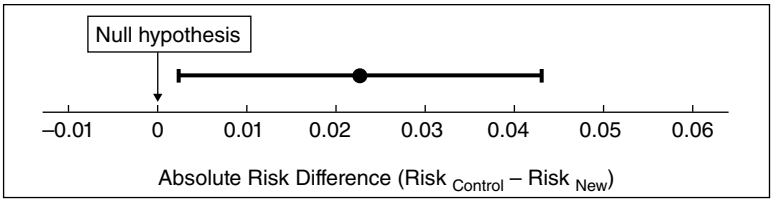


Figure 19.2 The 95% confidence interval for the Absolute Risk Difference for expulsion between original and new designs of intrauterine device

However, the results for the NNT hold a surprise. The relatively high point estimate (44) is reasonable, given that the new design only marginally reduces the risk of expulsion and the lower limit for the confidence interval (24) is also unsurprising, but the upper limit of 372 is nine times greater than the point estimate – not something that would easily be anticipated.

Figure 19.2 shows the confidence interval for the ARD, The lower limit is close to zero. If the results had been on the very limit of statistical significance that limit would move to zero, at which point the corresponding NNT would be 1/0 = Infinity patients. Consequently, as the lower limit for the ARD approaches zero, the upper limit for the NNT will escalate dramatically. Wherever results are marginally significant, one of the limits for the confidence interval for the NNT will not only be extreme; it will also be highly unstable. When we calculate the NNT for a marginally significant change, we are taking the reciprocal of a number close to zero and even small changes in the data would produce disproportionate changes in the upper confidence limit for the NNT. Consider an example of this instability: The total sample size for the each design of IUD was 2000. If we take the results for the new design and change just five subjects' results, so the number with an expulsion increases from 222 to 227, the upper limit of the confidence of the NNT increases from 372 to 10 010. The change we made in the data was quite small (just five out of 2000 subjects), but the effect on the upper confidence limit for the NNT was dramatic. Such instability is statistically very undesirable.

In cases where the data is non-significant, the problems are even more fundamental. The confidence interval for the Absolute Risk Difference will include zero and thus

one limit must be positive and the other negative. If we take the reciprocal of the negative value it will generate an NNT that is also negative. How we should go about treating a negative number of patients is problematic. There are theoretical approaches to making sense of these negative numbers, but the logic is highly convoluted. As a parting shot, one might question the whole point of calculating a confidence interval for the NNT for any treatment that is not even statistically significant.

The childhood diarrhoea trial gave a confidence interval for the NNT that raised fewer problems (Table 19.3). This was because the confidence limits for the ARD (Figure 19.1) went nowhere near zero. It is only when one of these limits gets close to zero that the corresponding NNT starts to escalate in an unstable manner.

In summary, attempting to calculate confidence limits for the NNT where results are only marginally significant, produces values that are unstable and can be very unexpected. With non-significant results, confidence intervals for the NNT are almost meaningless, with one limit indicating negative numbers of subjects. These problems give confidence intervals for the NNT a generally murky reputation. If they are to be calculated, they should be restricted to cases where the outcome is strongly significant.

Calculating a 95% confidence interval for the RR, OR and Number Needed to Treat

There are no problems calculating confidence intervals for the RR or OR and they are widely used and extremely useful.

The calculation of a confidence interval for the NNT is beset with problems and many prefer to avoid it entirely. If it is undertaken, it needs to be done cautiously and preferably only with data that is strongly statistically significant.

19.5 Chapter summary

The Relative Risk (RR), Odds Ratio (OR) and Number Needed to Treat (NNT) are commonly used to describe the extent of change seen in the proportion of subjects experiencing a particular event.

The RR is easily understood as the proportional change in the likelihood of the relevant event.

For events that occur only rarely, the OR will approximate the RR, but for more common events, the OR will diverge more strongly from the null hypothesis value (1.0) than the RR.

The Absolute Risk Difference is calculated by subtracting one risk from the other. The null hypothesis value for the ARD is zero. The NNT is the reciprocal of the ARD

and it describes the number of subjects who would have to be transferred from one treatment to another in order to achieve one additional advantageous outcome. It is of value in pharmacoeconomics.

It is common practice to report 95% confidence intervals for the RR and OR and occasionally the ARD. Statistical significance can be established if any one of these intervals excludes the relevant null hypothesis value. All three tests are mathematically equivalent and will always lead to the same conclusion. If the contingency chi-square test is applied to the same data, its conclusion will generally agree with those from the confidence intervals, but in marginal cases there may be disagreement.

It is less common to see authors quoting 95% confidence intervals for the NNT as it has some undesirable properties. If such intervals are to be quoted they should preferably be restricted to cases that are strongly statistically significant.

20
Logistic regression

This chapter will ...

- Describe how logistic regression can be used to analyse studies where the recorded outcome is dichotomous (e.g. Yes/No, Agree/Disagree or Success/Failure etc.).
- Describe how we can model the likelihood of a particular outcome by taking account of either nominal or interval factors (or a mixture of both).
- Explain the concept of 'Confounding' where a factor falsely appears to influence an outcome.
- Describe how logistic regression can be used to distinguish factors that genuinely affect an outcome from those that are merely confounded.

20.1 Modelling a binary outcome

In the previous two chapters we considered whether a nominal outcome (e.g. an IUD is or is not expelled) may be influenced by a nominal factor (e.g. using a control or test model of IUD). However, nominal outcomes may also be affected by measured (Interval) factors such as subjects' ages or weights and so on. We can model such cases using logistic regression. The method is similar to linear regression

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

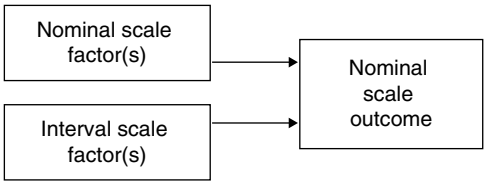


Figure 20.1 Diagrammatic representation of circumstances in which we would use logistic regression

(Chapter 15); it generates a regression equation, but the outcome is now categorised rather than measured.

We will see later in this chapter that logistic regression is a very versatile method that can consider the influence of any combination of factors on a dichotomous endpoint. There may be one or more continuously varying (interval) factors, one or more categorical (nominal) factors or a mix of both types of factor. This is shown in Figure 20.1. The only scenario where we almost certainly would not use logistic regression is where there is a single nominal factor under investigation. Such a situation could in principle be analysed by logistic regression, but normal practice would be to use a chi-square test.

20.1.1 The Logistic Model. The influence of a continuously varying factor (Age) on the probability that pharmacists will be willing to provide training

The example we will start with concerns the influence of age on pharmacists' willingness to train patients in the use of a phone 'App'.

Pharmacists' roles include promoting public health campaigns and we are investigating the willingness of pharmacists to become involved in encouraging and training patients with weight problems to use a phone 'App' to track calorie intake. A questionnaire has been used to gather demographic data for some pharmacists; the data includes their age and gender. It also asks about various opinions including the question 'Would you personally be happy to provide training in the use of the App? (Yes/No)'.

A spread sheet (see www.ljmu.ac.uk/pbs/rowestats) gives details of respondents' ages, gender and willingness to provide training. The first research question is whether Age (a continuously varying factor) influences willingness to train (the categorised outcome).

The status of each pharmacist is recorded as a categorisation (will/will not provide training) and as such cannot be the subject of a regression model. However, the probability of a particular pharmacist being a willing trainer is a continuous measure and can potentially be predicted by a regression approach.

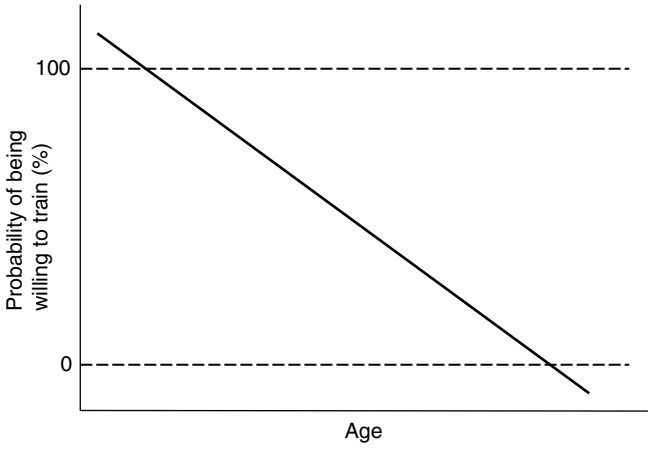


Figure 20.2 Attempting to use linear regression to predict probability of being willing to provide phone App training on the basis of the respondent's age

We might attempt to predict the probability of being willing to train by using linear regression as described in Chapter 15. In the example shown (Figure 20.2), it has been assumed that the greater the pharmacist's age, the lower the likelihood that he/she will be willing to train. However, the regression line would extend indefinitely at both ends and include probabilities of more than 100% and less than 0 which are logically impossible; probabilities can only range from 0 (no chance) to 100% (a certainty).

Figure 20.3 shows a more realistic relationship where the probability of being willing to train still decreases with age, but the graph is now sigmoidal, asymptotically approaching 0% at greater age and similarly approaching 100% among the very young.

Fortunately, the sort of relationship illustrated in Figure 20.3 can be linearised by a suitable mathematical transformation. A commonly used transformation is the 'Logit' (Log odds) of the probability. The odds of an event occurring has already been defined (Section 19.1.2) as the ratio between the probability that the event will occur and the probability that it will not. Hence, if P is the probability that the event will occur:

$$\text{Odds} = P/(1-P)$$

If the probability is 60% (0.6), then the odds are 0.6 / (1 – 0.6) = 0.6 / 0.4 = 1.5.

The logit of the probability of such an event is then defined as the natural log of these odds and so:

$$\text{Logit}(P) = \text{Ln}[P/(1-P)]$$

In the above case where the probability was 0.6, the Logit would be:

$$\text{Logit} = \text{Ln}(\text{Odds}) = \text{Ln}(1.5) = 0.405.$$

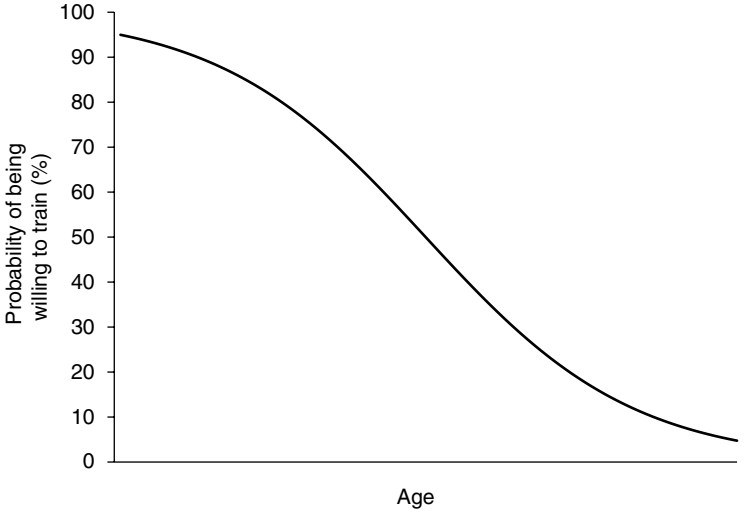


Figure 20.3 A more realistic, sigmoidal relationship between willingness to provide training and age

Table 20.1 Relationship between the probability of being willing to provide training and the Logit of the probability

Scenario	Odds = P/(1 – P)	Logit = ln(Odds)
Certainly not willing Probability = 0	0/(1 – 0) = 0	ln(0) = – ∞
Neutral point Probability = 0.5	0.5/(1 – 0.5) = 1	ln(1) = 0
Certainly willing Probability = 1	1/(1 – 1) = ∞	ln(∞) = + ∞

Table 20.1 shows that, in the most extreme case of an individual who will definitely not be willing to train, the probability and its odds are zero and the log of the odds is therefore minus infinity. At the other extreme, someone certain to be willing has a probability of 1.0 and then the odds and the Logit are both plus infinity. So, while probabilities are restrained between values of zero and one, the Logit of the probability can meaningfully take any value from minus to plus infinity. In the exactly balanced case (0.5 probability), the odds take a value of one and the Logit is zero.

The relationship between the logit of the probability and age is shown in Figure 20.4. This line still theoretically extends from minus to plus infinity, but such values make perfectly good sense for the logit of the probability even if the same values are nonsensical in the context of the probability itself.

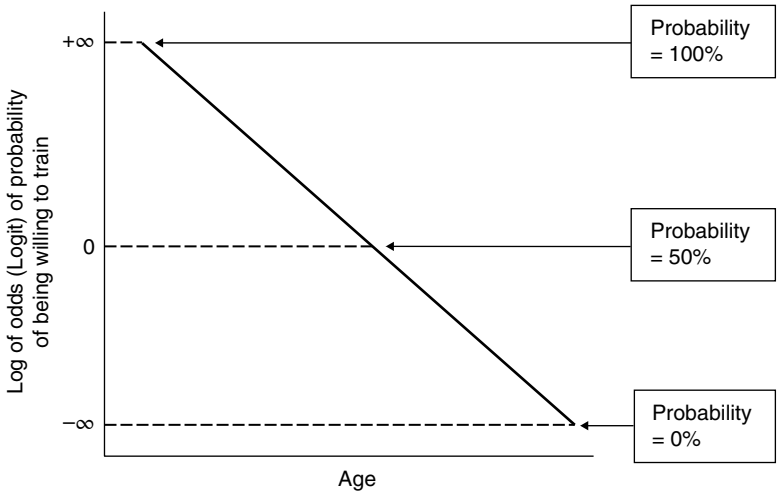


Figure 20.4 The logit of the probability of being willing to train can take values between plus infinity (Corresponds to 100% probability) and minus infinity (0% probability). A logit of zero corresponds to 50% probability

20.1.2 Fitting a logistic model to observed data

A regression equation is set up in the form:

$$\text{Logit}\,(P) = \text{Constant} + \text{Coefficient} \times \text{Predictor}$$

Because of the use of the logit transform, such equations are given the special name of 'Logistic regressions'.

Fitting a model such as that in Figure 20.4 to the observed data is complicated by the fact that each pharmacist is recorded simply as either willing or not willing to train. These can be recorded as values of one and zero, but we end up with data consisting of just these two values and nothing in between.

With linear regression we used the 'Least squares' criterion to establish the best fit of the regression line to the observed data. However, because the current data consists simply of ones and zeroes we have to use a different method which is referred to as 'Maximum likelihood fitting'. The principle is that the constant and coefficient in the logistic equation are adjusted so that model predictions are as closely aligned as possible to the observed outcomes. A model will have a low likelihood if it calculates high probabilities of willingness to train for pharmacists who actually are not so prepared and/or assigns low probabilities to those who are willing. The coefficients are adjusted until the overall likelihood for the model is as high as possible.

Unlike least squares fitting, maximum likelihood fitting cannot be achieved in a single step; instead an iterative approach is used. An initial attempt at a fit is

established and then the constant and coefficient are adjusted to improve the match between predicted and observed outcomes. These adjustments are repeated over and over, improving the match each time. Eventually a point is reached where further iterations produce minimal improvement. The computer algorithm will include a threshold below which any improvement is considered trivial and the process will terminate.

20.1.3 Modelling the willingness to train by taking account of age

Age will be used as a single predictor of the probability that a particular pharmacist will be willing to train.

20.1.3.1 Produce a logistic regression equation The first step is to use a computer program to generate the logistic regression equation using the observed data. The resultant equation is:

$$\text{Logit (Probability of willingness to train)} = 4.605 - 0.0921 \times \text{Age}$$

The negative coefficient for Age tells us that the model will allocate lower probabilities of being willing to train to pharmacists of greater age.

20.1.3.2 Consider the statistical significance of the regression Chapter 15 explained that we need to test the statistical significance of any linear regression relationship. The same principle applies in the context of logistic regression. Is the evidence convincing or might the apparent trend towards lower willingness to train among older pharmacists be a chance correlation within this particular sample that is not genuinely present within the generality of the profession?

Significance depends upon the strength of the correlation. The strongest possible correlation would arise if there was some critical age below which everybody was prepared to train and above which nobody was. At the other extreme, there might be the same mix of willing and unwilling among all ages. In addition to this strength of correlation, the sample size is also important in determining significance – large samples being more likely to achieve formal significance than small ones (see Figure 20.5).

The *P* value for the equation given in Section 20.1.3.1 is <0.001; the apparent relationship between age and willingness to train is clearly statistically significant.

20.1.3.3 Calculating the probability that a given pharmacist will be willing to train Using the logistic regression equation, it is possible to take each pharmacist and calculate the logit of the probability and hence the actual probability that he/she will be willing to train. This is illustrated below for a pharmacist aged 65 years:

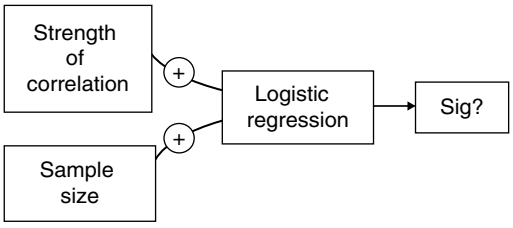


Figure 20.5 Determination of statistical significance (Sig) for a logistic regression

$$\begin{aligned}\text{Logit}(P) &= 4.605 - 0.0921 \times \text{Age} \\ &= 4.605 - 0.0921 \times 65 \\ &= 4.605 - 5.987 \\ &= -1.382\end{aligned}$$

The Logit of the probability can then converted to the actual probability (P) using the equation below. (If you are unfamiliar with the function 'Exp', see the nearby key box.)

$$\text{P} = \text{Exp}\,(\text{Logit})/[\text{Exp}(\text{Logit}) + 1]$$

So, for our 65 year old:

$$\begin{aligned}\text{Probability} &= \text{Exp}\,(\text{Logit})\,/[\text{Exp}(\text{Logit}) + 1] \\ &= 0.251/1.251 \\ &= 0.201\end{aligned}$$

Exponent (Exp)

Exponentiation uses a mathematical constant (Euler's e) which takes the value 2.718. The exponent then consists of e raised to the relevant power, thus the exponent of n is e^n. This is often written as 'Exp(n)'.

Most scientific calculators will easily calculate exponents.

The interpretation is that, based on his/her age, there is a 20% probability that this pharmacist will be prepared to provide training. As the figure is less than 50%, our categorical prediction would be that this individual will not be a willing trainer.

Table 20.1 showed that a logit of zero corresponds to the balance point where probability equals 50%, so any pharmacist for whom the logit is greater than zero will be predicted as willing to train and those with values below zero will be predicted as unwilling. For the current data set, age 50 years has special significance; at

that point the logit is exactly zero and consequently the probability of being willing to train is at the balance point of 50%. This means that pharmacists aged under 50 are more likely than not to be willing to train and those over 50 will tend to be unwilling. Such predictions will never be entirely reliable and it is especially foolhardy to rely on any prediction where the calculated probability is close to 50%; if the probability is close to 0 or 100%, then the prediction should be much more reliable. For those aged around 50, we cannot reliably predict whether they will agree to train, but for youngsters and veterans, we can make fairly confident predictions.

20.1.3.4 Assess the effectiveness of the model The overall effectiveness of the predictions can be illustrated in a couple of different ways. The first way is to classify each case as a true positive or negative (Prediction matches observed situation) or as false positives or negatives (Predicted as willing but actually unwilling or vice versa, respectively). The results are shown in Table 20.2.

There are 116 + 46 = 162 true predictions from a total of 236 cases; a success rate of 68.6%. So, while there is statistically significant evidence that age is related to willingness to train, it is clearly not a terribly effective predictor.

Another way to illustrate the outcome is a graph such as Figure 20.6. Probability of being willing to train is shown on the vertical axis and ages are shown along the horizontal. The sigmoidal graph shows the logistic relationship between probability and age. The key age of 50 years at which there is 50% probability is shown as a vertical dotted line. Each pharmacist is superimposed on the graph by plotting his/her age on the horizontal axis with the vertical position fixed at either one or zero according to whether they are or are not observed to be willing trainers.

Individuals represented by the dots to the left of the dotted line are all predicted as willing to train; those to the right are predicted as not willing. Among those to the left of the line, the dots at the top of the figure are true positives and those at the bottom are false positives. To the right of the line, we have true negatives at the bottom and false negatives at the top.

At the extremes of the age range, the predictions are reasonably reliable – those below age 40 are very likely to be willing and those over 65 are unlikely. However, there is a wide central range (40–65 years) where the calculated likelihood of willingness is not far from 50% and the predictions are unreliable.

Table 20.2 Outcomes of predicting willingness to train using Age as the single predictor

	Observed to be willing	Observed to be not willing
Predicted as willing	True positives = 116	False positives = 45
Predicted as not willing	False positives = 29	True negatives = 46

[Note that two individuals were calculated to have exactly 50% probability of being willing to train. These have both been counted as positive predictions; this approach places them among the true positives. Some computer programs consider them as negative predictions and they then become false negatives.]

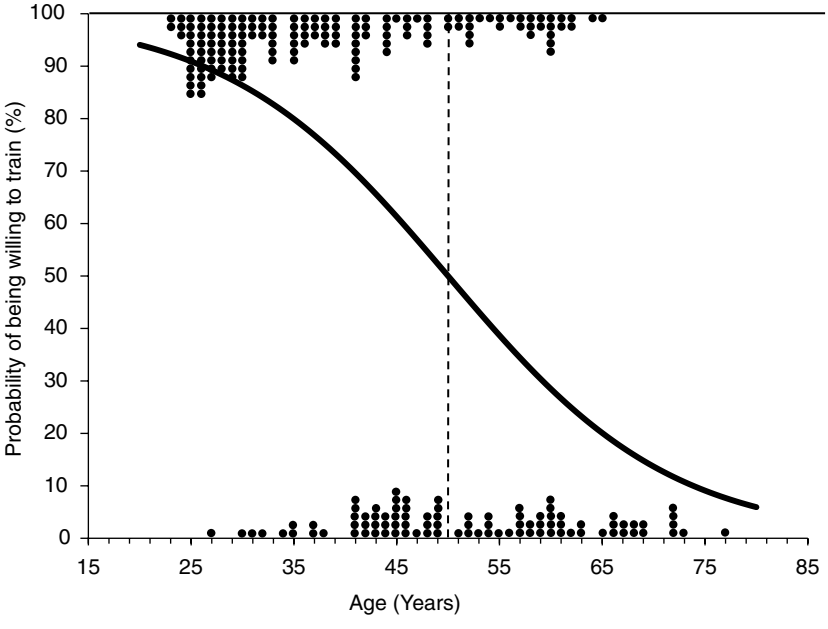


Figure 20.6 True and false classifications of willingness to train, achieved by logistic regression based on age

20.1.3.5 Odds ratio – Describes the magnitude of the effect of Age on willingness to train As always, we should not just ask 'Is the factor statistically significant?'; we should also ask about how great an effect it has (remember Chapter 10). It is the value of the coefficient in the equation that we look at. This determines whether the outcome (willingness to train) is strongly or weakly influenced by Age.

The regression equation takes the form:

$$\text{Logit}(P) = 4.605 - 0.0921 \times \text{Age}$$

It is the figure of –0.0921 that links age to the probability of being willing to train. However, it does not directly link the two; it links age to the logit of the probability. To give a direct link we take the coefficient and use it to calculate the 'Odds Ratio' (OR). This is the proportional change in the odds associated with a one unit change in the predictor. In our study it would be the relative change in the odds of being willing to train associated with a one year increase in age. Thus if two pharmacists are aged (say) 30 and 31 years and the odds of them being willing trainers are represented as Odds(30) and Odds(31), the odds ratio would be:

$$\text{OR} = \text{Odds}(31) / \text{Odds}(30)$$

The OR can be calculated as the exponent of the coefficient. The appendix to this chapter provides a demonstration of this relationship. In our equation the coefficient for Age was –0.0921, so the Odds Ratio for Age is:

$$\text{OR} = \text{Exp}(-0.0921) = 0.912$$

Thus if one pharmacist is one year older than another, the odds that the older pharmacist will be willing to train will be 0.912 times those for the younger individual.

The practical conclusion would be that two individuals separated by an age gap of one year won't differ greatly in the odds that they will be willing to train, but over the course of time, the odds will fall to 91.2% of its previous value during each additional year. So, an age difference of (say) 20 years would be associated with a difference in odds of $0.912^{20} = 0.158$. The odds of an older pharmacist being willing to train are only 0.158 those for an individual who is 20 years younger. A considerable effect accumulates over time.

Odds ratio from logistic regression

The exponential of the coefficient in the logistic regression equation is the 'Odds ratio'. It is the key measure of the strength of the relationship between the predictor and the dichotomous outcome. It records the extent to which the odds on a particular outcome change in response to a one unit change in the predictor.

20.2 Additional predictors and the problem of confounding

20.2.1 Consider the effect of gender alone

The Excel spreadsheet contains details of the gender of each pharmacist – data that we have not so far considered. Having carried out a single factor analysis looking at the effect of age, we might continue in the same vein and look at the effect of gender in another single factor analysis. Gender is a categorical factor and we could use the familiar chi-square test (as in Chapter 18). The findings would be that among males 53.4% are prepared to train compared to 69.2% among females. Females seem to be more willing to train and the *P* value from the chi-square test (0.013) suggests that this effect is statistically significant.

20.2.2 The problem of confounding

So now the simplistic conclusion might be that willingness to train is affected by both age and gender. However, as soon as we start to consider two or more potential factors we must bear in mind the risk of 'confounding'. This concept is

probably best clarified by working through the specific problem that arises in our example.

Pharmacy training in the United Kingdom has seen a considerable shift in gender balance – a predominance of male trainees in the 1960s has gradually been replaced by a current majority of females. This means that in the UK workforce, older pharmacists are predominantly male but younger ones include more females. This shows in our sample; the mean ages are 39.2 for females and 48.0 for males. We could therefore construct three distinct models, each of which would be consistent with the associations of age and gender with willingness to train:

- Age has a true, direct effect on willingness but the apparent effect of gender is false. Men may well be less willing to train, but this arises because they are generally older; it is not a direct effect of their gender. In this model, gender would be 'Confounded'.

- The real factor is gender and age falsely appears to influence the outcome; older pharmacists would be less willing to train because so many of them are male, not because of their age. Here age is confounded.

- Both factors do genuinely, independently influence the outcome; there would be no confounding.

20.2.3 Determining which of the apparent factors really affect willingness to train

How can any statistical technique unravel this confusion? The key is that while younger pharmacists may be predominantly female and older ones largely male, there are significant numbers of individuals who do not follow this trend; we have useful numbers of younger males and older females. It is these that help us to see what is happening. If age is the real factor and gender is just confounded, then our younger males should be influenced by their age and therefore be more willing to train than our older females. On the other hand, if gender is what really matters, younger males will follow the influence of their gender and be less willing than the older females.

Table 20.3 illustrates the approach described above. The pharmacists have been divided into four groups based on age and gender. Age was dichotomised at 42 years as this was the median. The proportion willing to train is shown for each sub-group. As explained above, the key groups are the younger males and older females and these have been emphasised in the table. Both key groups show levels of willingness to train that are very similar to the other subgroup of similar age, but quite different from the other group of the same gender. This suggests that age is genuinely influencing the outcome, but gender has no real effect; it is merely confounded.

Table 20.3 Proportions willing to train among pharmacists grouped by age and gender

	Younger (Age 42 or less)	Older (Age over 42)
Male	*33/42 = 78.6%*	29/74 = 39.2%
Female	67/76 = 88.2%	*16/44 = 36.4%*

Younger males and older females are italisised as they are the key groups.

If correlation between the two possible factors was excessively strong (e.g. virtually all the younger pharmacists were female and all older ones male), there would be too few of the younger males and older females to allow us to carry out the sort of analysis suggested above; it would be impossible to identify the genuine factor.

20.2.4 Using logistic regression to detect confounding

As a more formal approach to detecting confounding, we will use logistic regression. Logistic regression equations can be modified in a similar manner to that seen in Chapter 15, to make them multiple regressions. We simply add an extra term for each additional predictor. The equation then takes the form:

$$\text{Logit}\,(P) = \text{Const} + \text{coeff}_1 \times \text{predictor}_1 + \text{coeff}_2 \times \text{predictor}_2 \ldots \text{etc}$$

We will now replace the two separate single-factor analyses with just one analysis that incorporates both age and gender as factors – a multi-factorial logistic regression. To enter gender as a factor, we will use codes of F and M and the computer program will arbitrarily convert these to indicator variables of one and zero in a similar manner to that discussed in Section 15.3.9. You need to check the program's output to ascertain how this coding was done.

When the data is modelled by multiple logistic regression, the *P* value for age is <0.001 and for gender *P* is 0.975 (See Table 20.4). Age is confirmed as directly

Table 20.4 Generic output for multiple logistic regression investigating Age and Gender as possible factors influencing the probability of being willing to provide training

Coding:		Gender: 0 = Female, 1 = Male	
Logistic regression of Training on Age, Gender			
	Coefficient	Exp(B)/Odds ratio	*P*
Constant (A)	4.606		0.000
Age (B1)	–0.09203	0.912	0.000
Gender(B2)	–0.01010	0.990	0.975

associated with willingness to train, but Gender falls far short of significance (in agreement with our less formal conclusion in Section 20.2.3). The important thing to note is the stark contrast between the outcomes of the single and multi-factorial analyses. The two separate, single factor analyses suggested that both Age and Gender were significant factors. Performing one multifactorial analysis identifies Age as a genuinely relevant factor but dismisses Gender as confounded.

As a matter of practicality, it may be convenient to carry out initial, separate, single factorial tests on each of several pieces of data that might influence the relevant outcome. However, if we do, we certainly should not immediately claim that a multiplicity of factors have all been shown to influence the outcome. We should always gather the possibly relevant factors together into one overall multiple logistic regression model and see which factors survive as statistically significant.

Use logistic regression to detect confounding

Logistic regression can cope with any mix of continuously varying and/or categorical factors (e.g. age and gender) and determine whether apparently significant factors genuinely influence the outcome or are merely confounded.

20.2.5 Removing non-significant factors from a multiple logistic regression

If you create an initial multiple logistic regression model and two or more factors are indicated as non-significant, it is important to follow the approach discussed in Section 15.3.3.2. You should remove just one non-significant factor at a time and re-test the reduced model each time. You may find that the removal of one non-significant factor reveals that one of the other factors which previously appeared to be non-significant actually is significant. Peel away non-significant factors one at a time until all the remaining factors are significant.

20.3 Analysis by computer package

Whatever package you use, you will need to identify a variable that encodes the outcome. This variable must have just two values. It is best to use 0 and 1, allocated so that the event we want to predict is encoded as 1 and its absence is encoded as 0. In the current case, willingness to train would be encoded as 1 and unwillingness as 0. Many programs allow the use of (say) Y and N to encode willingness to train, but the program will convert these to values of 0 and 1 and it may allocate 1 as the code for N. You would then end up with a logistic regression equation that predicts the likelihood

of being unwilling to train which is counterintuitive. It is therefore better to impose 0 and 1 codes in a manner that is appropriate for your purposes.

You will also need variables to encode the factors that may influence the outcome. It is necessary to distinguish between those factors that are measured, interval type variables and those that are nominal categorisations; different programs use different methods to achieve this. Any factor that is categorical (such as Gender) will be converted to a dummy numerical variable. This is done automatically by the computer program; if SPSS is used for the analysis, males are coded as 0 and females as 1, while Minitab uses the opposite pattern. It is important to check the program's output to see what coding scheme has been generated.

Typical computer output from a multiple logistic regression with both age and gender as factors is shown in Table 20.4. The first part of the output described the allocation of dummy values for Gender; in this case 0 = Female and 1 = Male.

The next key pieces of output to look for are the *P* values which identify that only Age is truly significant. The following two paragraphs explain how to find values for the regression coefficients and Odds ratios in Table 20.4. However, as gender is not a significant factor, the definitive values for the regression coefficient and odds ratio for Age would be taken from a reduced analysis, considering Age as a single factor. In practice, the numerical values are only very marginally affected by removing Gender from the model.

Regression coefficients and odds ratios are included in Table 20.4. If you struggle to find the coefficients in the output from your statistics package, some programmes refer to them as 'B' values. The Odds Ratio for age is a key piece of output that you should report. The latter will be easily identifiable if it is labelled as 'OR' or 'Odds Ratio', but if you can't find it, it may be labelled as something like 'Exp(B)'; as explained in Section 20.1.3.5 the Odds Ratio is calculated as the exponent of the coefficient in the regression equation, hence the obscure label.

Gender was encoded as 0 for female and 1 for male; the negative coefficient for gender therefore indicates a trend for males to be less willing to train. However, that trend was not statistically significant.

20.4 Extending logistic regression beyond dichotomous outcomes

So far we have restricted the scenarios that are subjected to analysis by logistic regression to those where there is a dichotomous outcome. In fact, logistic regression can be extended to cover any scenario where the outcome is recorded as a series of three or more categories. The obvious extension is to cases where the outcome is still something that would be recorded as nominal categories, but with more than two possible outcomes. Moving even further, the various outcome categories could form an ordinal scale of measurement.

All of the above can be modelled by logistic regression. Where there is a simple dichotomous outcome (as in the case we looked at) the relevant technique may be

referred to as 'Binary logistic regression'. For nominal outcomes with more than two possibilities we use 'Nominal logistic regression' and for ordinal outcomes (not surprisingly) we use 'Ordinal logistic regression'.

These techniques are more complex and subject to significant potential pitfalls and realistically go beyond the scope of a general book such as this. If you want to use the more advanced forms of logistic regression, you would probably be wise to let a statistician guide you. However, don't be put off – they are manageable.

The only situation with a categorical outcome where we would probably not use logistic regression is where we have a single categorical factor; the more familiar chi-square test can be used.

20.5 Chapter summary

Logistic regression is a highly versatile method that allows us to determine the influences of nominal and/or interval factors on a dichotomous nominal outcome. It produces a regression equation that will allow us to predict the likelihood that a given individual will fall into a particular category by taking account of one or more characteristics.

We predict that a given individual will (or will not) fall into a particular category, based upon whether the probability is greater than or less than 0.5. However, where the probability is close to 0.5, such predictions are highly error prone. If the probability is close to zero or one, the prediction is more secure.

If several apparently significant factors have been identified in a series of single factor analyses, these should all be incorporated into a unified multiple logistic regression to check for possible confounding.

A key part of the output from logistic regression is the Odds Ratio associated with a particular factor. This describes the effect of an increase in the value of the factor by 1.0 – for example an increase in age by one year or an increase in the value of an indicator variable from zero to one (e.g. female to male). If the OR is (say) 1.5, then a one point increase in the value of the factor will cause the odds to increase to 1.5 times its original value.

20.6 Appendix

Demonstration of the equality of the Odds Ratio and the exponent of the regression coefficient

We can express the logistic equation as :

$$\text{Ln(Odds)} = \text{A} + \text{B} \times \text{Predictor}$$

In the above, A is a constant and B is the coefficient governing the effect of the predictor.

If we take the exponential of both sides:

$$\text{Odds} = e^{(\text{A} + \text{B} \times \text{Predictor})}$$

This can then be re-expressed as:

$$\text{Odds} = e^{(\text{A})} \times e^{(\text{B} \times \text{Predictor})}$$

The term $e^{(A)}$ is insensitive to any change in the predictor and can be considered as a constant. A one unit change in the value of the predictor will therefore change the odds by a factor of e^{B}.

Hence the Odds Ratio equals e^{B} or Exp(B).

Part 4

Ordinal-scale data

21

Ordinal and non-normally distributed data: Transformations and non-parametric tests

This chapter will ...

- Describe the requirement for normally distributed data when using parametric tests (*t*-tests, ANOVAs etc.).
- Show how such tests can be used and interpreted after non-normal data has been transformed to normality.
- Introduce non-parametric methods where data is converted to rankings, so they become 'distribution free tests'.
- Explain why ordinal data is generally subjected to non-parametric tests.
- Discuss the appropriate wording of conclusions where a non-parametric test has proved significant.
- Describe how to decide whether data should be tested: (i) directly, (ii) after transformation to normality or (iii) non-parametrically.
- Describe four widely applicable non-parametric methods (Mann–Whitney, Wilcoxon paired samples, Kruskal–Wallis and Spearman correlation).

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

In Chapter 6 we saw how the calculation of the 95% C.I. for the mean can lead to nonsensical results if the data deviated severely from a normal distribution. This requirement for a normal distribution also applies to the *t*-tests, analyses of variance and so on that we met in Chapters 7 to 15. These procedures are termed 'parametric methods' and are quite robust, so moderate non-normality does little damage, but in more extreme cases, some pretty dumb conclusions can emerge. This chapter looks at steps that can be taken to allow the analysis of seriously non-normal data and also of ordinal scale data.

21.1 Transforming data to a normal distribution

21.1.1 Production of a toxic metabolite in smokers and non-smokers

In a small minority of users, an analgesic produces a serious side-effect – inflammation of the liver. It is suspected that this may be due to a very minor, but toxic metabolite. It has been noted that the reaction is about twice as frequent among smokers compared to non-smokers. A theory is advanced that because smoking induces greater levels of certain metabolic enzymes in the liver, the increased susceptibility among smokers might be due to the production of greater quantities of the toxic metabolite. If this theory were correct then it ought to be possible to detect increased production of the rogue metabolite in smokers. Table 21.1 shows the amount of the relevant metabolite recovered in the urine of 20 smokers and non-smokers, following the ingestion of an oral dose (50 mg) of the drug. For now, focus on the first two columns and ignore the log data.

These quantities of metabolite are shown in Figure 21.1. There is a fairly strong visual impression that metabolite production is indeed higher among smokers, but unfortunately there is also a distinct impression that the data may not be normally distributed – there seems to be a scattering of high values above the main clusters of points.

Figure 21.2 (a) provides an even stronger impression of positively skewed data with the smokers' data. Although not shown here, the non-smokers' data is similarly skewed.

If we were just to ignore this non-normality and perform a two-sample *t*-test on the raw data, the result would be a *P* value of 0.115 indicating a lack of statistical significance. However, we would be very unwise simply to accept this negative result, given that the test used is not appropriate for highly skewed data.

There are two possible solutions and we are going to look at both. The first is to use the same trick we saw in Chapter 6 – transformation.

21.1.2 Carrying out a log transformation

We can try to find a mathematical transformation of the data that shows a better approximation to a normal distribution. With positive skew, a log transform may be useful. The results of the log transform are presented in Table 21.1 and Figure 21.2 (b).

Table 21.1 Production of a toxic metabolite (μg) from an analgesic drug in smokers and non-smokers – actual masses and base 10 logs

Smokers	Non-smokers	Log for smokers	Log for non-smokers
7.75	2.50	0.889	0.398
7.80	7.45	0.892	0.872
23.40	4.95	1.369	0.695
27.50	3.10	1.439	0.491
5.65	17.10	0.752	1.233
12.05	4.20	1.081	0.623
11.95	6.05	1.077	0.782
21.15	25.20	1.325	1.401
15.40	20.20	1.188	1.305
14.05	6.10	1.148	0.785
10.15	9.55	1.006	0.980
12.40	7.15	1.093	0.854
6.20	3.90	0.792	0.591
16.30	13.00	1.212	1.114
7.15	12.05	0.854	1.081
5.95	4.25	0.775	0.628
9.20	15.20	0.964	1.182
5.15	4.45	0.712	0.648
9.70	5.90	0.987	0.771
12.90	6.10	1.111	0.785

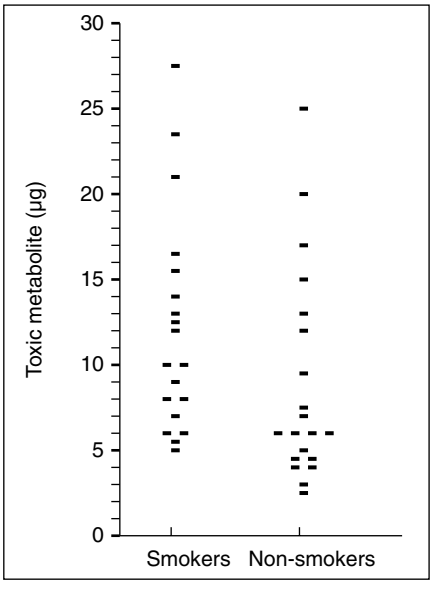


Figure 21.1 Production of a toxic metabolite from an analgesic drug in smokers and non-smokers

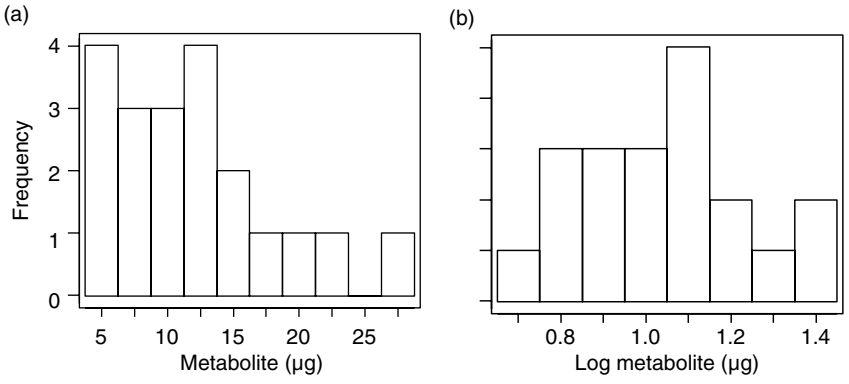


Figure 21.2 Histograms of: (a) metabolite production and (b) log metabolite production from an analgesic drug in smokers

Table 21.2 Generic output for a two-sample *t*-test comparing the logs of the amounts of toxic metabolite produced by smokers and non-smokers (last two columns of Table 21.1)

Two-sample *t*-test	
Mean (LogSmoke):	1.033
Mean (LogNonsmoke):	0.861
Difference (LogSmoke – LogNonsmoke):	0.172
95% C.I. Difference:	0.014–0.331
P:	0.034

The latter shows that the distribution for the smokers' data is now much more symmetrical. The effect on the non-smokers' data is not shown but is also satisfactory.

We would then perform a standard two sample *t*-test, but apply it to the last two columns in Table 21.1. Generic output is shown in Table 21.2.

Wonder of wonders! Data that was non-significant is now revealed as significant ($P = 0.034$).

It is usually at about this point that the cynical cry 'Cheat!'. How dare we use this statistical fiddle to convert non-significant results into significant ones? Essentially, we need have no qualms about this approach. It is entirely respectable and is definitely superior to the analysis of the original data, because the transformed data is much closer to a normal distribution. The only caveat would be that if we are going to do this kind of thing, we should ideally declare our intentions in advance. It is not good practice to gather data and then thrash around, trying every possible statistical approach until we find one that produces the desired result (Generally a significant one!).

Carrying out tests on transformed data

It is a normal and legitimate practice to use transformations to convert data to a better approximation of a normal distribution and then carry out tests on the transformed data.

The contrast between the successful outcome when testing normally distributed transformed data and the failure with highly skewed raw data is an example of the large loss of power that often accompanies the application of procedures such as *t*-tests to inappropriate data.

Major loss of power with inappropriate data

It is quite common to suffer a large loss of power if highly skewed or otherwise non-normal data is analysed by methods that assume normality.

If we want to comment upon the size of the difference between smokers and non-smokers, we need to be very careful when interpreting these results. The limits of the 95% C.I. for the difference between the two groups are reported as 0.014 to 0.331. However, these were calculated from log transformed data and need to be converted back to their antilogs. This then gives a C.I. of 1.03 to 2.14. However, these differences were arrived at by subtracting one log value from another and then taking the anti-log. When you carry out that sequence of calculations, you are actually performing a division. Consequently the numbers we end up with are not the difference in numbers of micrograms of toxin produced; they are the ratio between toxin production in the two groups. The correct interpretation is that we can be 95% confident that the smokers produce between 1.03 and 2.14 times more toxin than the non-smokers. In such a case, the null hypothesis should be expressed as 'The ratio of toxin production in smokers to that in non-smokers is 1.0'. Statistical significance is then based on the fact that the 95% C.I. does not include the figure of 1.0.

The point estimate for the effect size is subject to the same logic. The estimated difference is given as 0.172 and the antilog of that is 1.49. So, it is estimated that smokers produce about 50% more toxin.

Effect size when a log transform has been used in a *t*-test

We obtain a 95% C.I. for the ratio between the two means, rather than their absolute difference.

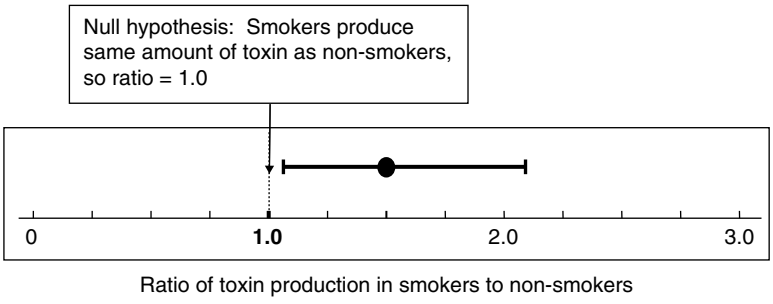


Figure 21.3 95% C.I. for the ratio between toxin production in smokers and non-smokers

The point estimate and 95% C.I. for the effect of smoking are shown in Figure 21.3.

21.2 The Mann–Whitney test – a non-parametric method

21.2.1 A substitute for the two-sample *t*-test

Instead of transforming the data to normality, we could employ one of a range of procedures referred to as 'Non-parametric tests'. These partially duplicate the functionality of tests we have already met, but use a method of calculation that does not depend upon a normal distribution. The non-parametric test that is generally substituted for the two sample *t*-test goes by a variety of names, but most commonly the 'Mann–Whitney' test.

21.2.2 Converting data to rankings

The defining characteristic of all non-parametric tests is that we take the data and convert it into ranks. To convert the toxin data we start with the actual quantities of toxin produced (forget about the log transformed data for this test), and search through it looking for the lowest single value. This turns out to be the value of 2.5 μg for the first of the non-smokers, so this is awarded the rank value of 1. The next lowest values (3.10, 3.90, 4.20, 4.25, 4.45 and 4.95 μg) are also among non-smokers and get ranks of 2, 3, 4, 5, 6 and 7. The next value is then for one of the smokers (5.15 μg) and it gets rank 8. This process continues simply enough for the lowest 12 values, but then we find that the next value (6.10 μg) occurs twice (among the non-smokers). These are referred to as tied values. They should get ranks 13 and 14, but there is no logical reason why one should be ranked higher than the other, so they each receive a rank of 13.5. We eventually reach the highest value (27.5 μg) which is for one of the smokers and it gets rank 40.

The original data and their rankings are shown in Table 21.3.

Table 21.3 Conversion of quantities of toxin into rank values

Smokers		Non smokers	
Toxin (μg)	Rank	Toxin (μg)	Rank
7.75	19	2.50	1
7.80	20	7.45	18
23.40	38	4.95	7
27.50	40	3.10	2
5.65	9	17.10	35
12.05	26.5	4.20	4
11.95	25	6.05	12
21.15	37	25.20	39
15.40	33	20.20	36
14.05	31	6.10	13.5
10.15	24	9.55	22
12.40	28	7.15	16.5
6.20	15	3.90	3
16.30	34	13.00	30
7.15	16.5	12.05	26.5
5.95	11	4.25	5
9.20	21	15.20	32
5.15	8	4.45	6
9.70	23	5.90	10
12.90	29	6.10	13.5
Rank total	488		332

Non-parametric tests are based upon ranks

In non-parametric tests the data is transformed into rank values and then all further calculations are based solely upon these rankings.

21.2.3 Using rankings for further calculation

All further calculations are then carried out on these rank values rather than on the original data. The next stage is to add up all the rank values for the two groups to produce 'Rank Totals'. On a null hypothesis that there is no systematic difference between smokers and non-smokers, we would expect high and low rank values to be scattered randomly between the two groups and the rank totals to be fairly similar. However, if (as we suspect) the higher values are generally among the smokers, then this is where we will also find the highest rank values and the higher rank total. The rank totals shown above (488 and 332) hint fairly strongly at a difference, but is the difference big enough to be convincing? The rest of the Mann–Whitney test answers precisely that question.

21.2.4 Conducting a Mann–Whitney test

Depending upon the particular statistical package used, the data may be entered either in two columns or all in a single column with a separate column containing codes indicating which groups the values belong to. What you enter are the actual quantities of toxin produced; you don't have to work out the rankings – that should all be done by the statistical package.

Output (Table 21.4) varies from package to package, but generally includes a median value for each group. Note that while these are useful descriptively, they do not enter into the calculation of the test. There may be two *P* values. This arises because there were some tied values among the data. The existence of ties somewhat undermines the method of calculation used within the Mann–Whitney test and it is possible to apply a correction to compensate for this. The general preference seems to be for the latter value (labeled as 'Adjusted for ties'), but unless there are a huge number of ties the difference is not usually great. In this case there are so few ties (three pairs) that the two *P* values are apparently identical and the result is significant either way.

21.2.5 Interpreting a significant outcome

When a *t*-test produces a significant outcome, its interpretation is quite straightforward: there is evidence that the two population mean values differ. Unfortunately, with non-parametric tests such as Mann–Whitney, it's not so simple. The conclusion may be worded simply in terms of evidence of an increase/decrease in values, or we may want to be more specific and talk about changes in the median or mean.

Strictly speaking, because no mean or median is even calculated as part of the test, the null hypothesis should be that:

'*Smokers and non-smokers produce the same amounts of toxic metabolite.*'

What this is understood to mean is that if we randomly selected one smoker and one non-smoker, the chances that the smoker would produce more toxin than the non-smoker is exactly equal to the chances of the opposite pattern.

Table 21.4 Generic output from a Mann–Whitney test of the amounts of toxin produced by smokers and non-smokers

Mann-Whitney test	
Median (Smoke):	11.05
Median (NoSmoke):	6.10
P:	0.036
P (Adjusted for ties):	0.036

That way we make no reference to the mean, median or any other statistical parameter.

As the null hypothesis has been dismissed, we have evidence that there must be a difference in the amount of toxin produced. Based on Figure 21.1, we can conclude that:

'*Smokers tend to produce more toxin than non-smokers.*'

That would be a minimum conclusion that nobody is likely to challenge.

Can we then go on and say that this must also imply an increase in the median or mean amount produced?

- **The median:** In the majority of cases a demonstration of generally increased (or decreased) values can be taken to imply a corresponding change in the median. However, just be aware that there are some bizarre distributions (usually involving extreme skewness) where the median may not change in the way you would anticipate (see Appendix to this chapter).

- **The mean:** The process of ranking destroys all information regarding the absolute magnitude of the individual results and consequently it would be very risky to try to claim that a Mann–Whitney test had demonstrated a change in the mean. To justify such a conclusion, you would have to make such extensive assumptions about the distribution of the data, that you could probably use a parametric test anyway!

Interpreting a significant Mann–Whitney test

'**Values are generally higher in this group than in that.':** This makes no assumptions about how the data is distributed. Minimum claim – Little risk.
'**The median is greater in this group than in that.':** Only assumption is that the data is not distributed in a totally bizarre manner. Generally OK, but check with an expert if the data sets have extreme distributions.
'**The mean is greater in this group than in that.':** Rarely justifiable.

21.2.6 Choosing: parametric or non-parametric?

When non-parametric methods are applied to data that is normally distributed, they are slightly less powerful than their parametric equivalents although the difference is not great. For the tests covered in this chapter, the non-parametric test has about 95% of the power of its parametric equivalent. In other words, if our data is normally distributed then a sample of 19 tested by a parametric method would provide about

the same power as a sample of 20 tested by the non-parametric equivalent. Since the power of the two types of test is so similar, it is not surprising that the *P* values generated (0.034 for the *t*-test (when applied to the transformed data) and 0.036 for the Mann–Whitney test) are barely different.

However, if the data is severely non-normal we can loose a huge amount of power by using a parametric test. We saw this loss of power when we obtained a non-significant result by applying a *t*-test to the untransformed (and highly skewed) toxin data.

One disadvantage of non-parametric tests is that while they detect statistical significance quite satisfactorily, they do not provide any meaningful 95% confidence interval for the size of the difference in outcome and are therefore little use in determining practical significance.

Strengths of parametric and non-parametric tests

Parametric: Slightly more powerful than non-parametric where data is reasonably normally distributed and they produce a 95% C.I. for the size of the experimental effect.
Non-parametric: May be much more powerful than parametric tests if data is seriously non-normal.

A good general rule is to use a parametric test whenever possible even if that necessitates a transformation of the data. However, some data sets are such a mess that no amount of jiggery-pokery will render them normal and in these cases a non-parametric test is our ultimate fall-back.

Dealing with non-normally distributed data

First choice: Convert to normal distribution by transformation and use parametric test.
Second choice: Resort to a non-parametric test.

21.2.7 Distribution free tests

When we convert measurement data to rankings we destroy all information about the distribution of the data. For instance, when we ranked the toxin measurements, all that remained was a series of ranks from 1 to 40 and we would have got exactly the same set of values regardless of whether the original distribution was normal, skewed, bimodal or any other shape. Because ranked data is blind to the initial distribution, the non-parametric tests are sometimes called 'Distribution free tests'.

21.3 Dealing with ordinal data

Back in Chapter 1, data were described as 'Interval' (measurements on a regular scale), 'Ordinal' (measurements on a scale of undefined steps) and 'Nominal' (classifications). We have dealt extensively with two of these, but ordinal data have been largely ignored.

21.3.1 Why ordinal data are generally analysed by non-parametric methods

Ordinal data typically includes things like patients' subjective descriptions of their condition. A score may be allocated, ranging from 1 to 4, where:

1 = No/almost no pain
2 = Slight pain
3 = Significant pain
4 = Severe pain

Ordinal data tends not to form normal distributions. For a start, it is often recorded on scales with a very limited number of possible values. Scales of four, five or six points are frequently seen. In such cases, it is impossible for the data to form the sort of smooth, bell-shaped distribution that constitutes a true normal distribution. But then the problem is further exacerbated. Although there is no necessary reason for it, anybody who has worked with real-world, ordinal data knows that it is frequently hideously non-normal. Offered a scale of possible scores, people will quite frequently do bizarre things like only using the extreme upper and lower values but not the middle ones, or else they will produce a completely flat distribution, with no peak frequencies anywhere. No amount of mathematical transformation is going to convert that sort of mess into anything remotely resembling a normal distribution.

It's normal to be non-normal

It is theoretically possible for ordinal scale data to approximate a normal distribution, but marked non-normality is all too common.

There is no absolute case that parametric methods cannot be used with ordinal scale data. If the scoring system allows a reasonably wide range of possible values and if these happen to approximate a normal distribution, parametric methods could be used. However the reality of working with ordinal data is:

- Frequently horribly non-normal distributions.
- A small potential gain in power if a parametric method is deployed with data that is reasonably normal.

- A very large potential loss of power if a parametric method is used with data that is badly non-normal.

A common view is that when planning any experiment where data will be collected on an ordinal scale, we may as well reconcile ourselves to the use of non-parametric methods from the outset.

Dealing with ordinal scale data

Unless there is specific evidence that the data is likely to behave itself unusually well, just accept that non-parametric methods will have to be used. Power loss will, at worst, be very slight.

21.3.2 An example of dealing with ordinal scale data: Applying the Mann–Whitney test to the effectiveness of an analgesic

Two teams of patients rate the effectiveness of either an active herbal analgesic or a placebo for the treatment of mild pain. The design is unpaired – one team are allocated to active and the other to placebo tablets. The scale used to report effectiveness is:

4 = Completely/almost completely effective
3 = Strongly effective
2 = Moderately effective
1 = Slightly effective
0 = No/Almost no effect

The results are shown in Table 21.5 and graphically in Figure 21.4.

Two things are apparent:

1. There is a strong impression that the active preparation is receiving higher scores than the placebo, but formal testing is still required.

2. Both distributions are strongly skewed, but as one shows positive and the other negative skew, there is no way we are going to be able to transform them both to normality.

21.3.3 Analysing the results

With an unpaired design and measurements on an interval scale, we would normally have used a two-sample *t*-test to check for any difference. But, this data is ordinal and not remotely normally distributed so we will have to move to its non-parametric equivalent – the Mann–Whitney test.

Table 21.5 Scores for effectiveness of placebo/active analgesic. Greater scores mean greater effect

Placebo	Active
3, 0, 2, 4, 0	1, 3, 3, 3, 4,
0, 2, 0, 0, 1,	4, 3, 0, 4, 3,
1, 0, 0, 0, 3,	4, 4, 3, 1, 2,
2, 1, 2, 1, 3	4, 4, 3, 2, 2

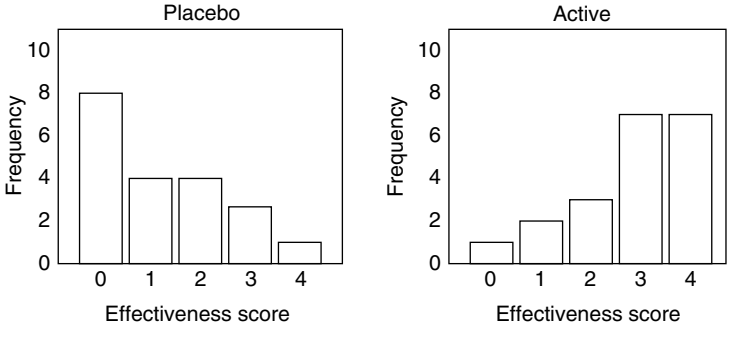


Figure 21.4 Effectiveness of placebo or active analgesic

The result of a Mann–Whitney test on this data is an uncorrected *P* value of 0.0008. There are many tied values in this set of results and adjusting for ties does now slightly reduce *P* to 0.0006. Either way the result is strongly significant.

As explained previously, the significant result can certainly be interpreted as evidence that the active preparation generally had a greater effect than the placebo and a more detailed claim that the active preparation has a greater median effect is pretty uncontroversial even with these distinctly skew distributions.

21.4 Other non-parametric methods

There are huge numbers of other non-parametric tests available. Three are especially useful as substitutes for parametric tests that have already been covered in this book.

21.4.1 Wilcoxon paired samples test – a substitute for the paired *t*-test

In the paired *t*-test we calculate the change in a measured endpoint for each individual and the test expects these differences to form a normal distribution. If this condition is not met, the Wilcoxon paired samples test can be used instead.

As an example, we will consider (Table 21.6) some changes that occurred in haemoglobin levels in a group of vegans when given vitamin B12 supplementation. The study was paired, each individual providing a pre-treatment blood sample and then a further sample after four weeks of supplementation.

Figure 21.5 shows this to be a very difficult data set, with a strong suspicion of a bimodal distribution. Such a distribution is biologically quite credible – there may well be a majority whose diet already contains adequate B12 in whom supplementation will be associated with small random changes clustered around zero, but a minority (approximately 30%) who are deficient and where we see decisive increases in haemoglobin levels.

Figure 21.5 shows far more positive than negative changes, which strongly suggests that there is a treatment effect, but a formal test is required. No transformation

Table 21.6 Haemoglobin levels ($g.L^{-1}$) in a group of vegans before and after vitamin B12 supplementation

Pre-treatment	Post-treatment	Change
142	146	+4
140	160	+20
135	143	+8
153	153	0
136	155	+19
142	141	–1
146	151	+5
117	133	+16
139	139	0
156	153	–3
154	155	+1
152	150	–2
154	156	+2
133	155	+22
146	151	+5
153	153	0
126	140	+14
115	114	–1
159	164	+5
146	152	+6
136	142	+6
158	161	+3
129	144	+15
137	152	+15
150	153	+3
136	141	+5
137	149	+12
136	151	+15

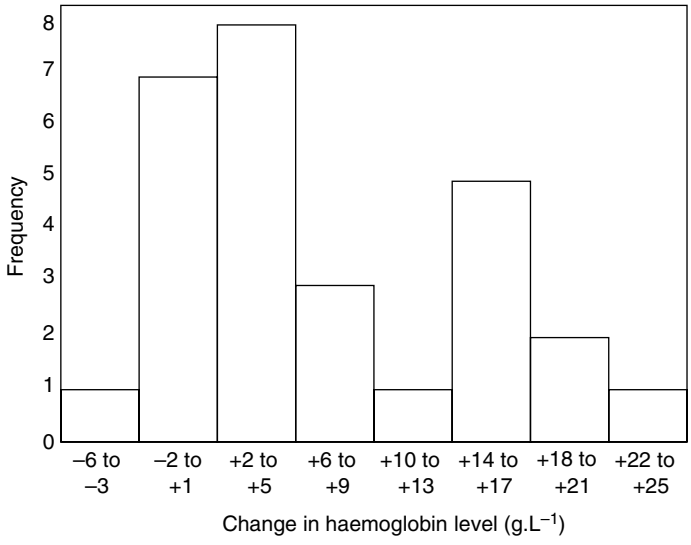


Figure 21.5 Histogram of changes in haemoglobin levels after vitamin B12 treatment

is going to convert this to a normal distribution. With normally distributed changes, we would have used a paired *t*-test, but with this data we will change to the Wilcoxon paired sample test.

The test is directly available in some statistical packages (e.g. SPSS) but not in others such as Minitab. Where it is available, the pre- and post-treatment values are entered into two columns and the test can be performed directly. With the likes of Minitab, the test can be achieved, but it's messy. You will first have to calculate the change that occurs in each individual and enter these into a column. Then the 'One-sample Wilcoxon' procedure is used to compare these values against a null hypothesis of no systematic change.

The output obtained varies so much between packages that there is no such thing as generic output. However, among whatever you do get, there should be a *P* value of <0.001. There is significant evidence that B12 has an effect on haemoglobin values.

21.4.2 Kruskal–Wallis test – a substitute for the one-way ANOVA

This acts as an equivalent of the one-way ANOVA that we met in Chapter 14. The requirements underlying the ANOVA are that all the populations from which samples are drawn should be normally distributed and have equal SDs. The Kruskal–Wallis test allows us to circumvent these requirements.

To illustrate the test, consider a trial of analgesia in palliative care. Sixty patients are randomized into three groups. One group receives oxycodone, one morphine

and the final group diamorphine. Over the next five days, the dosage is adjusted to what is considered the optimum for each patient. The patients then score their pain on a 'Visual analogue' scale. This consists of a line on a piece of paper. One end of the line is labelled 'No pain' and the other 'Severe pain'. The patient then makes a pencil mark on the line at a point indicating their impression of their pain. The scale is 10 cm long and the distance is measured to the nearest millimetre, giving potential scores from zero to 100.

The results are shown in Table 21.7.

For a visual presentation of the data, this is ordinal data, so a box and whisker plot is probably as good as anything (see Figure 21.6).

The data is ordinal and with only 20 observations for each drug, a histogram would provide little guidance as to whether the data is normally distributed. Under these circumstances, it would be risky to assume a normal distribution. The non-parametric Kruskal–Wallis test is preferable to a one-way analysis of variance with ordinal data.

To perform this test, most statistical packages require all the data to be entered into a single column with a further column contains codes indicating which group a result belongs to (as described for the one-way ANOVA). Generic output is shown in Table 21.8.

Table 21.7 Pain scores with three different analgesics

Oxycodone	Morphine	Diamorphine
64	62	23
71	50	42
21	76	30
75	51	70
12	45	71
11	35	65
15	54	35
20	71	20
69	31	25
38	49	45
9	19	9
10	10	15
74	55	47
79	80	6
17	15	15
32	62	14
15	31	15
51	80	17
72	41	20
64	19	25

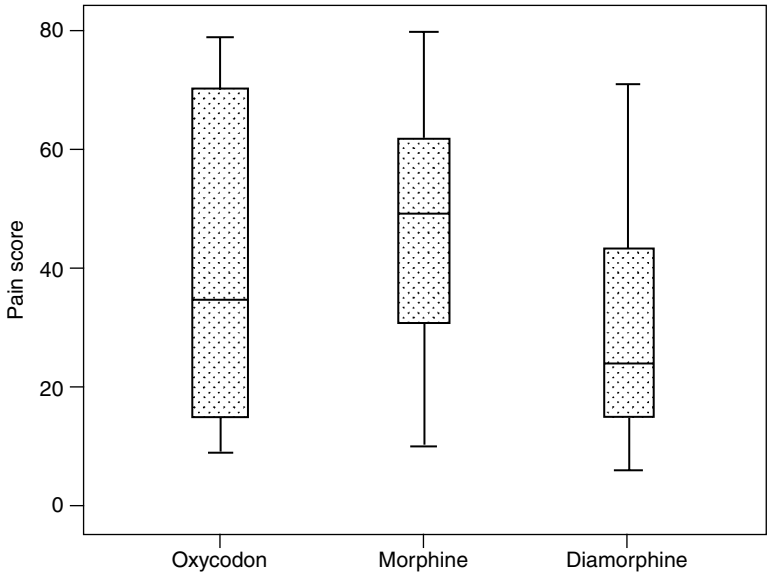


Figure 21.6 Box plots of pain scores with three different analgesics

Table 21.8 Generic output from a Kruskal–Wallis test of pain scores with different analgesics

Kruskal–Wallis test	
Analgesic	Median Pain_score
Oxycodone	35.0
Morphine	49.5
Diamorphine	24.0
$P = 0.097$	
P (Adjusted for ties) = 0.097	

Figure 21.6 did not suggest any clear difference between the various analgesics and the results are not statistically significant.

With the one-way ANOVA, most statistical packages implement a series of follow-up tests to determine exactly where any differences lie. Similar procedures exist to allow follow up after a significant Kruskal–Wallis test, but unfortunately they are not widely implemented in statistical packages. There would be no point in doing so in the present case, but if another data set proves significant and you want to perform follow up tests, you will either have to resort to a very powerful (and probably not very friendly) statistical package, or do the calculation manually. The latter is tedious, but recipes are available. A clear account is available in Zar, J.H. (1999) *Biostatistical analysis*, Prentice Hall, NJ; pp 223–226.

21.4.3 Spearman correlation – a substitute for Pearson correlation

21.4.3.1 Performing Spearman correlation Technically, Pearson correlation (Chapter 15), does have an assumption that the two sets of data are normally distributed, but it's pretty rare to see anybody bothering to check whether they actually are and there's precious little evidence of anything going radically wrong if they're not. However, non-parametric Spearman correlation is quite frequently resorted to with ordinal data.

To illustrate its use we will look at an investigation of a new explanatory leaflet. A previous version of the leaflet proved difficult to understand for patients with restricted educational achievement. The new version was drawn up by a specialist consultant (at great expense) and is supposed to be generally clearer and, crucially, equally accessible to all users. To test it, patients with a wide spectrum of educational achievement are asked to read the new leaflet and then answer five knowledge testing questions. Each patient is then awarded a knowledge score of between 0 and 5. The patients are also allocated an education score of between 1 and 6: 1 for those with no formal educational qualifications; 6 for graduates with additional professional qualifications. The education and knowledge scores, along with their rankings, are shown in Table 21.9.

Figure 21.7 suggests that we are still not doing a very good job – there are an awful lot of low scores. We also appear to have failed to achieve equal accessibility – there

Table 21.9 Educational levels and knowledge scores after reading revised information leaflet

Education level	Knowledge score	Ranked education level	Ranked knowledge score
1	0	2.0	1.0
1	3	2.0	8.0
1	1	2.0	3.0
2	3	4.5	8.0
2	1	4.5	3.0
3	1	7.5	3.0
3	3	7.5	8.0
3	4	7.5	12.0
3	3	7.5	8.0
4	4	11.0	12.0
4	5	11.0	16.0
4	3	11.0	8.0
5	5	14.5	16.0
5	5	14.5	16.0
5	2	14.5	5.0
5	4	14.5	12.0
6	5	17.5	16.0
6	5	17.5	16.0

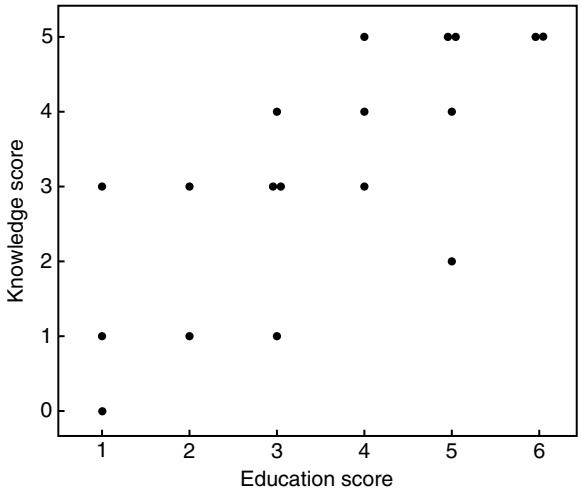


Figure 21.7 Relationship between educational level and score for knowledge after reading information leaflet

is a strong suspicion that those with lower educational levels are also achieving poorer knowledge scores.

To test for a relationship between educational level and amount gleaned from the leaflet, we need some from of correlation analysis. Both characteristics are assessed on ordinal scales with a very limited range of possible values, so we would probably use non-parametric Spearman Correlation.

The procedure is directly implemented in SPSS, but not in Minitab. In the latter case, it can still be performed but it's not elegant. With packages where the test is directly available, the data is simply entered into two columns and the appropriate test selected.

With other packages, we proceed in two stages as below:

Stage 1: Convert the data to ranks. Your statistical package may contain a routine to automate the process of calculating rankings from your data, otherwise it will have to be done manually. (Minitab can be used to do this – In the menu structure, go to 'Data' and then 'Rank'.). To perform ranking manually see Section 21.2.2. Table 21.9 includes the rankings.

Stage 2: Carry out correlation analysis of the rank values. Correlation of the rank values is carried out in the usual way, indicating the two columns containing the rankings. As rankings are being used, the overall procedure will constitute Spearman correlation.

Results will be as in Table 21.10.

If you are using a package that offers Spearman correlation, the fact that it is Spearman correlation will probably be indicated. Unfortunately, with a package like Minitab where we needed a work-around, the program will not be aware that the

Table 21.10 Generic output for Spearman correlation analysis of education levels and knowledge scores

Spearman correlation
Correlation of Education and Knowledge = +0.748
P = <0.001

data you fed it was actually a series of rankings, so the output may say 'Pearson correlation' when this is now in fact Spearman correlation.

The Spearman correlation coefficient works in essentially the same way as Pearson's. Its value can be anything between –1 and +1. Zero represents no correlation and the extreme values either perfect negative or perfect positive correlation. In this case, a Spearman correlation coefficient of +0.748 indicates quite a strong relationship. The *P* value is strongly significant, so it looks as if we were wasting our money when we hired the consultant. (How unusual is that?)

21.4.3.2 Spearman correlation and non-linear relationships – requirement for 'Monotonicity' When we were discussing Pearson correlation it was emphasised that problems would arise if this method was applied to data displaying a non-linear relationship (Section 15.1.6). With Spearman correlation, the problem is less severe. Figure 21.8 (a) shows a clearly non-linear relationship where the use of Pearson correlation would be questionable. However Spearman correlation would convert the data to rankings and then test for a linear relationship; part (b) of Figure 21.8 shows that a linear relationship is exactly what we have now got. The rankings for

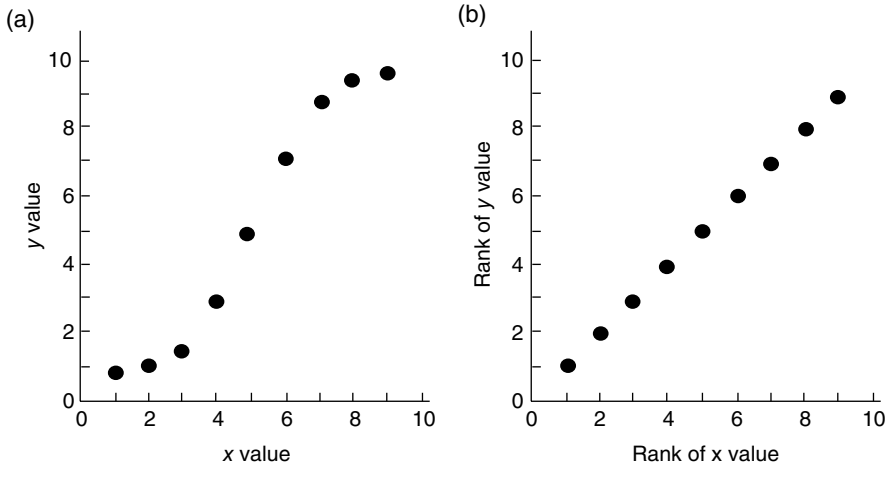


Figure 21.8 Non-linear relationships become linear when data is converted to ranks, with the sole proviso that the relationship is monotonic

the x values are simply 1, 2, 3 … 9 and the y rankings follow the same pattern giving a perfectly linear graph. With Spearman correlation, the initial non-linearity is therefore no problem.

The only requirement for Spearman correlation is that any relationship is 'Monotonic'. This term means that, throughout the range of x values studied, y values either always increase or always decrease. It does not matter if the relationship changes gradient; the only restriction is that the graph of y against x should not reach a maximum and then decline or fall to a minimum and then increase.

Non-parametric equivalents of other procedures

Parametric	**Non-Parametric**
Two-sample *t*-test:	Mann–Whitney test
Paired *t*-test:	Wilcoxon paired samples test
One-way ANOVA:	Kruskal–Wallis test
Pearson correlation:	Spearman correlation

21.5 Chapter summary

t-tests, analyses of variance and so on are referred to as 'Parametric methods' and they all make an assumption that data will be normally distributed. These methods can be very misleading if applied to severely non-normal data.

Data may be converted to a normal distribution using log transforms and so on as described in Chapter 6.

Strongly positively skewed data is often converted to a normal distribution by log transformation. When this is done to allow analysis by a two-sample *t*-test, you should be aware that the 95% C.I. for the size of the treatment effect will estimate the ratio between the values of the endpoint under the two conditions instead of the absolute difference in the value.

An alternative approach is to use a non-parametric test which does the same job as one of the parametric tests. These tests convert the data to rankings and the distribution of the data (largely) ceases to be an issue. They are sometimes called 'Distribution free tests'.

A significant result from a non-parametric equivalent of a *t*-test or an analysis of variance can always be interpreted as evidence of a change in values. Beyond that, so long as the data is not very awkwardly distributed (e.g. extremely skew), a significant result can probably be interpreted as evidence of a change in the median.

Where data is reasonably normally distributed, non-parametric methods are a little less powerful than their parametric equivalents, but where the data is severely non-normal, the non-parametric test may be much more powerful. With non-normally distributed interval scale data, the best solution is transformation to normality, but failing that, non-parametric methods can be used with only modest loss of power.

Ordinal data can potentially approximate to a normal distribution, but tends to be severely non-normal. The use of non-parametric methods is not obligatory with such data, but is common practice.

Four non-parametric tests are briefly introduced:

- Mann–Whitney test equivalent to the two-sample *t*-test
- Wilcoxon paired samples test equivalent to the paired *t*-test
- Kruskal–Wallis test equivalent to the one-way ANOVA
- Spearman correlation equivalent to Pearson correlation

21.6 Appendix

Do not assume that a significant result from a non-parametric test can necessarily be interpreted as evidence of a difference in median values.

Despite what some books claim, it is NOT safe to assume that a significant Mann–Whitney test can always be interpreted as evidence of a difference in the medians for the two groups being compared. Table 21.11 shows self-assessed severity scores for gastro-intestinal discomfort in groups of 50 patients taking two different analgesic drugs. The scoring is:

0 = None/virtually none
1 = Mild
2 = Moderate
3 = Severe

Table 21.11 Gastro-intestinal discomfort scores in patients taking two different analgesic drugs

	Number reporting this score	
Discomfort score	(Drug 1)	(Drug 2)
0	39	28
1	6	12
2	4	6
3	1	4

The data is ordinal and extremely positively skewed, with a majority of zero scores and a tail of other scores on one side only. Group 2 appears to have somewhat higher scores (22 positive scores compared to only 11 in group 1). For a formal comparison, we would use the non-parametric Mann–Whitney test. That yields a *P* value (adjusted for ties) is 0.021, so there is significant evidence of higher scores in group 2.

More than half of each group has a score of zero, so the median score must perforce be zero in both cases. So, while it is OK to conclude that 'Scores are higher in group 2 than group 1', you certainly could not claim that 'Group 2 has a higher median score' – the median is zero for both groups.

Part 5

Other topics

22

Measures of agreement

This chapter will ...

- Describe how we determine the degree of agreement where we have two or more sets of assessments or measurements and there is reason to expect agreement.
- Describe the use of Chronbach's Alpha as a measure of correlation among groups of questions that are all intended to reflect a single common feature.
- Council against the automatic use of Chronbach's Alpha in cases where the results of several questions are to be aggregated into a single score.
- Introduce the use of Cohen's Kappa in cases where a question produces results in a categorical (nominal) format and where we have two sets of assessments of the same things or people.
- Describe the extension of Cohen's Kappa to cases where assessments produce results on an ordinal scale by adding weighting factors to reflect the severity of any disagreements. This is then called Weighted Kappa.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

- Explore the problem of describing agreement among continuous measured endpoints, for example in comparisons of analytical methodologies. It will:
  - Explain why it is inappropriate to use the Pearson correlation coefficient in this context.
  - Describe the use of the Intraclass Correlation Coefficient (ICC) to describe such agreement.
  - Show the use of Bland–Altman plots to diagnose the form of disagreement in cases where the ICC takes a low value.

There are a number of scenarios where we have two or more sets of results and, in some sense, we want to determine their level of agreement. Precisely what is meant by 'Agreement' depends on the context and the type of data collected.

The first major divide is between those cases where we have answers to several questions (Section 22.1) or repeated answers to a single question (Section 22.2). In the latter case, we also need to consider the type of data that the question generates – nominal, ordinal or interval (See Sections 22.2.1 to 22.2.3).

22.1 Answers to several questions

22.1.1 Are all our questions measuring the same thing? Cronbach's alpha

The common scenario is that we have asked several questions all of which should reflect some single common feature. The intention would probably be to aggregate the various answers together into a single indicator.

A classic example is the Hamilton Depression Rating Scale. This uses a number of questions (17 or more) concerning anxiety, guilt, insomnia, suicidal thoughts and so on, all of which are associated with depression. Each question is assessed on a scale of zero to two or four points; these are then totalled to produce an overall depression rating score.

Another source of related sets of questions is where we deliberately include two or more questions in a questionnaire that may be worded differently, but which have essentially the same meaning. For example we might ask respondents to indicate their level of agreement with the statements 'After taking the tablets I feel that I am going to be sick' and 'I feel nauseous after I have taken the tablets'. We would normally place such questions well apart in the questionnaire, not immediately next to each other.

We may do this simply to determine whether respondents are being reasonably careful about the responses they supply or in the hope that using an average score for several similar questions will reduce problems caused by occasional aberrant responses.

Investigators sometimes use contradictory statements, to assess whether respondents are reading the questions carefully. For example in addition to the two statements about nausea already suggested above, we might add 'I have no problems with sickness while taking the tablets'. In such cases we have to reverse the scoring. Thus, if strong agreement with the first two statements was scored as five and strong disagreement as one, then for the new question, it would be strong disagreement that scored five. In that way, we can meaningfully add all three results to produce an aggregate score out of 15, where higher scores are associated with greater nausea.

When we come to assess the level of agreement among the questions, what we want to know is whether they are all measuring the same thing. If they are, then the answers should be correlated. In the case of the depression scale, each individual would ideally produce a consistent pattern of responses – all the responses indicating a low level of depression or all showing a high level (or perhaps all intermediate). What we do not want is respondents generating randomly mixed patterns. In the second example (nausea following medication), subjects should either indicate nausea in all responses or consistently deny it.

22.1.2 Indicating agreement with Cronbach's Alpha

Cronbach's Alpha is commonly used to measure the degree of correlation among sets of answers to related questions.

Chronbach's Alpha

We have a number of questions each of which produces an answer on an ordinal scale. Respondents will answer all the questions. Are those questions all measuring the same thing?

Unfortunately, the very fact that this measure is available seems to convince some investigators that it should be applied blindly in every case where several sets of results are to be amalgamated into a single score. The author has engaged in several debates with researchers who seemed to be heading towards a quite inappropriate use of the statistic. Consider as an example a comparison of subjects' opinions of several study centres being used in a clinical trial. We could ask questions about the clarity of the instructions for finding each centre, accessibility by public transport, convenience of car parking, problems accessing the building, friendliness of the staff and so on. We could then bundle the scores for all these questions into an overall

assessment of satisfaction with each centre. In this case, there is little reason to expect correlation among the responses. A group of respondents might all drive particularly large cars that cannot easily be manoeuvred in the car park and will all give low scores for that aspect, but that is unlikely to be linked to any systematic trend among their answers to the other questions. It remains perfectly reasonable to bundle all these responses into a single rating; each of the questions probes a relevant facet of satisfaction. However, they are independent facets unlikely to be correlated with one another and there is no reason to expect a high value for Cronbach's Alpha.

The author's experience is that those who persisted in the inappropriate use of this statistic found that (unsurprisingly) it yielded a very low Alpha value and they were then left with a false impression that there must be some flaw in their proposed approach.

22.1.3 Calculating and interpreting Cronbach's Alpha

Packages such as SPSS and Minitab include the ability to calculate Chronbach's Alpha.

It is a measure of correlation and should normally be restricted to values between zero and one. Zero indicates a completely random pattern, with no indication that the questions are measuring any common feature. A value of one will only arise if each respondent scores every question as two or all as three and so on, in which case they certainly do appear to be measuring the same thing.

Negative values of Alpha are theoretically possible, but probably indicate an error in data entry. For example, you might want to check for any negatively worded questions where the scores should have been reversed. Has this been done?

It is commonly claimed that Alpha = 0.7 represents an adequate level of correlation. There is no obvious rationale for this exact figure, but it is certainly true that when Alpha falls much below 0.7, it becomes difficult to see any convincing evidence that the questions are measuring the same thing.

Very high values of Alpha can also be a source of concern. If it gets above 0.95, you need to ask whether some of the questions are redundant; you might well be able to get away with a shorter questionnaire with little loss of real information.

22.1.4 An example using Chronbach's Alpha

Let us continue with the example of questions about nausea after medication. The three statements are:

- After taking the tablets I feel that I am going to be sick.
- I feel nauseous after I have taken the tablets.
- I have no problems with sickness while taking the tablets.

For the first two questions, responses are on a five point scale, where scores of one to five represent (respectively) 'Strongly disagree', 'Disagree', 'Neutral', 'Agree' and 'Strongly agree'. For the third statement, the scoring is reversed – five down to one.

For ten respondents the scores are shown in Table 22.1.

Simple inspection suggests a good level of correlation. With a couple of exceptions, the data for any individual patient indicates a consistent opinion concerning level of nausea. This is reflected in the Alpha value which is +0.853 – a comfortably high value (see Table 22.2).

Most packages deliver more than just the simple Alpha value. A useful additional piece of information is the Alpha values with one omitted item. In this case, there are three of these. First we re-calculate Alpha with question one omitted, so we use just questions two and three. Then we omit number two and so on. These are included in Table 22.2. It is noticeable that Alpha increases markedly if the third question is omitted, suggesting that it has been spoiling the correlation. Respondents number two and ten look like they may well have failed to notice the negative wording of question three and so those responses are out of line with their other two. We might consider dropping this question.

Table 22.1 Scores indicating post-medication nausea using three differently worded questions

Respondent number	Question 1	Question 2	Question 3
1	1	2	1
2	4	4	1
3	5	5	4
4	2	2	2
5	3	3	2
6	1	1	1
7	4	5	5
8	1	1	1
9	3	4	3
10	2	2	5

A score of 1 always indicates a lack of nausea and 5 its presence

Table 22.2 Generic output for Chronbach's Alpha

Chronbach's Alpha	
Alpha = 0.853	
Omitted items	
Item omitted	Alpha
Question 1	0.71
Question 2	0.678
Question 3	0.973

22.2 Several answers to one question – do they agree?

22.2.1 Nominal outcomes – Cohen's Kappa

The usual scenario is that there is a judgement to be made and the decision is subject to a degree of individual opinion. The first example is a simple dichotomisation. We will consider a study of patient understanding of medicine use. Understanding was assessed as 'Satisfactory' or 'Not satisfactory'. Two researchers are involved in making the assessments and we want to know if they agree. Table 22.3 shows the results when 187 patients are assessed by both researchers. There are 137 cases where the two assessors agree that the individual patient did understand the instructions and 34 where they agreed that there was a lack of understanding. However, there are an additional six and ten cases where the assessors disagree.

A simple way to express the assessors' degree of agreement is simply to calculate that there are 137 + 34 = 171 agreed cases out of 187 assessments, which is 91.44% agreement.

The result above sounds very impressive but to put it in context, we need to calculate the level of 'agreement' we would achieve if both assessors assigned their judgements randomly. Cohen's Kappa measures how much better we have done than we would achieve by random guessing. First we calculate the level of agreement that would arise from random assignment of judgements as below:

Assessor 1 used the grade 'Satisfactory' for 137 + 10 = 147 out of 187 cases, which (as a fraction) is 0.786. This assessor must have used 'Not satisfactory' on 1 – 0.786 = 0.214 of occasions.

For assessor 2, the corresponding figures for 'Satisfactory' are 137 + 6 = 143 which is 143/187 = 0.765 and 0.235 for 'Not satisfactory'.

Then consider what would happen if the two judges assessed each patient randomly, with assessor 1 always allowing a 0.786 likelihood of giving 'Satisfactory' and assessor 2 giving 'Satisfactory' with a 0.765 likelihood.

For any particular case, the likelihood of both assessors giving a grade of 'Satisfactory' is $0.786 \times 0.765 = 0.601$. Similarly, the likelihood that both assign 'Not satisfactory' is $0.214 \times 0.235 = 0.050$.

For 187 patients, there would be $187 \times 0.601 = 112.41$ cases where both assessors declared them 'Satisfactory' (an agreement) and $187 \times 0.050 = 9.41$ cases where they both gave 'Not satisfactory' (also an agreement). The total number of agreements is then 112.41 + 9.41 = 121.82. That would be 121.82 / 187 = 65.14% agreement.

Table 22.3 Two assessors' opinions on 187 patients' understanding of how they should use their medicine

		Assessor number 1	
		Satisfactory	Not satisfactory
Assessor number 2	Satisfactory	137	6
	Not satisfactory	10	34

The 91.44% agreement we actually achieved was good, but it starts to look rather less dramatic when we realise that random assignment of judgements would achieve 65.14% agreement.

We need to consider the extent to which our actual level of agreement exceeds that which we would achieve by random assignment of judgements. In this case, the level of superiority is 91.44 – 65.14 = 26.30%.

Finally, Cohen's Kappa compares this to the maximum possible improvement over random guessing, that is if we were to achieve 100% agreement. In the current case, the maximum possible improvement over random assignment would be 100 – 65.14 = 34.86%. So …

$$\begin{aligned}\text{Kappa} &= \text{Actual improvement}/\text{Maximum potential improvement}\\ &= 26.30\% \,/\, 34.86\%\\ &= 0.754\end{aligned}$$

(The exact figure for Kappa is 0.755; the above calculation gives a small rounding error.)

The two assessors have achieved about three-quarters of the possible improvement over random guessing.

The Kappa value of 0.754 probably gives a more realistic sense of the level of agreement than the bald figure of 91.4% agreement. Simple calculations of proportions of agreement are especially misleading where either of the two judgements is very frequently used (i.e. either most cases are judged Satisfactory or most as Not satisfactory). For example, if both assessors randomly assigned 95% of cases as Satisfactory, this would result in 90.5% (so called) agreement.

22.2.1.1 Interpreting Kappa Normally Kappa will take a value of between zero and one. Zero indicating that one (or possibly both) of the assessors has provided effectively random assignment of classifications while Kappa = 1.0 shows exact agreement. Theoretically, negative values are possible, but this would require systematic disagreement; for any case where one assessor gave a rating of Satisfactory, the other would have to have a specific tendency to rate it as Not satisfactory and vice versa.

How high does Kappa need to be if we are to claim satisfactory agreement between assessments? This obviously depends to an extent, on the context. In some areas of

research, categories may be clearly demarcated and agreement should be almost complete (Kappa close to 1.0), but in others there may be a strong element of subjectivity and Kappa could never realistically achieve such a high value.

A commonly quoted interpretation of Kappa was provided by Landis and Koch (1977; *Biometrics* **33**, 159–174). They banded levels of agreement as: 0–0.20 Slight; 0.21–0.40 Fair; 0.41–0.60 Moderate; 0.61–0.80 Substantial; 0.81–1.0 Almost perfect. However, there is no real rationale for these bands and they take no account of the specific context.

22.2.1.2 Random versus biased disagreement If a low value of Kappa is obtained, we know immediately that there is disagreement, but it is worth digging a little deeper to determine the exact form of disagreement.

In Table 22.3 there were 16 cases where the assessors disagree and that number splits fairly evenly between ten patients that assessor one considered satisfactory, but assessor two reckoned as unsatisfactory and six patients producing the opposite pattern. This is explicable as random disagreement.

In Table 22.4 there are 23 disagreements, but now these are very unbalanced. In 22 cases, assessor one grades the patient as unsatisfactory but assessor two considers them satisfactory. In contrast, there is just one case where we see the opposite pattern. Here there is bias, assessor one is demanding a higher standard than assessor two.

To test whether any imbalance achieves statistical significance, we would apply McNemar's test (Section 18.6). With the data in Table 22.4, the imbalance is highly significant ($P < 0.001$); one assessor is biased relative to the other.

22.2.1.3 Using Kappa with two assessments by a single investigator: Test – re-test stability In the example above, there are two separate assessors. An alternative scenario would have an assessor make judgements on all 187 cases and then go back over the same cases a week later and make fresh judgements. This would be termed a 'Test – re-test' structure. Cohen's Kappa could be calculated as above to express the test – re-test stability of the assessor's judgement.

22.2.1.4 More than two categories Kappa can also be calculated where cases are being classified into more than two categories. The basic approach is as before – compare agreement actually achieved with what we would expect if classification

Table 22.4 Bias – Assessor 1 is applying higher standards than Assessor 2

		Assessor number 1	
		Satisfactory	Not satisfactory
Assessor number 2	Satisfactory	122	22
	Not satisfactory	1	42

was random. An example might be diagnoses of respiratory disease into categories such as chronic obstructive pulmonary disease, asthma, bronchitis and so on. Note that these are unordered categories; they do not form an ordinal scale. For ordinal categories, see Section 22.2.2.

22.2.1.5 Programs to calculate Kappa Several statistics packages (e.g. SPSS and Minitab) have routines to calculate Cohen's Kappa for cases where cases fall into two or more categories. A simple XL spreadsheet is also available on the website associated with this book (www.ljmu.ac.uk/pbs/rowestats/).

22.2.2 Ordinal outcomes – Weighted Kappa

We will consider an example where two experts review 157 clinical cases where there is a suspected Adverse Drug Reaction (ADR). The Naranjo algorithm (1981; *Clin. Pharmacol. Ther.* **30**, 239–245) uses a series of aspects of the case to judge how likely it is that an adverse event was actually due to the drug rather than being the result of other factors. The likelihood of the case being a true ADR is graded on a four point ordinal scale as Doubtful, Possible, Probable or Definite. What we want to asses is the extent to which the two investigators agree on the grading. The results are shown in Table 22.5.

There are 25 cases where the two researchers are in exact agreement that it was Doubtful that the event was drug related. Then we have 27, 57 and 32 other cases of exact agreement on Possible, Probable or Definite relationships between event and drug use. These cases of exact agreement are said to lie on the 'Diagonal' of the matrix. Other smaller numbers of cases are 'Off diagonal' and show disagreement.

With ordinal outcomes we have various levels of severity of discrepancy in classification. We might not be too surprised if one of our experts classified a case as Definite and the other as Probable, but if their respective conclusions were Definite and Doubtful there would be a more serious issue. Fortunately most of the

Table 22.5 Grading by two experts of the likelihood that a series of adverse clinical events were causally related to drug use

		Expert One			
		Doubtful	Possible	Probable	Definite
Expert Two	Doubtful	*25*	1	0	0
	Possible	2	*27*	3	1
	Probable	1	3	*57*	4
	Definite	0	0	1	*32*

Values on the diagonal (Exact agreement) are given in italics

disagreements in Table 22.5 are only one square off the diagonal, but there are two cases where there is more serious disagreement. In the first, Expert One graded a case as Definite but Expert Two declared it as only Possible and in the other we have grades of Doubtful and Probable.

Thus with ordinal classifications we should not simply consider cases as agreed or disagreed, but should use a weighting system to reflect the degree of any disagreement. A 'Weighted Kappa' value operates in a similar manner to the previous section, but cases of disagreement are weighted according to the extent of the discrepancy.

Weighted Kappa

Weighted Kappa measures agreement between assessors where results form an ordinal scale. Disagreements are weighted according to their severity. Discrepancies of a single point reduce Kappa to a smaller extent than more severe disagreements.

To calculate weighted Kappa, we have to devise a weighting scheme. There are two commonly used schemes [Table 22.6 (a) and (b)]. Scheme (a) is referred to as 'Linear' weighting; the weights are simply equal to the number of grades by which the two assessors disagree in a particular case. Thus weights of zero are used when there is no discrepancy and the maximum of three would arise where there is complete

Table 22.6 Two commonly used weighting schemes used to calculate weighted Kappa where classification is on a four point ordinal scale

(a) Linear weighting scheme

		Assessor One			
		Doubtful	Possible	Probable	Definite
Assessor Two	Doubtful	0	1	2	3
	Possible	1	0	1	2
	Probable	2	1	0	1
	Definite	3	2	1	0

(b) Quadratic weighting scheme

		Assessor One			
		Doubtful	Possible	Probable	Definite
Assessor Two	Doubtful	0	1	4	9
	Possible	1	0	1	4
	Probable	4	1	0	1
	Definite	9	4	1	0

disagreement; one assessor concludes Doubtful and the other Definite. The common alternative, 'Quadratic' weighting uses the squares of the discrepancies.

22.2.2.1 Computing Weighted Kappa Not many statistical packages offer the calculation of weighted Kappa. An XL spreadsheet is available on www.ljmu.ac.uk/pbs/rowestats which will calculate Kappa for ordinal scales with a range of up to ten points. It can be used with any weighting scheme that you want but, unless there is a strong case for using something more complex, it is probably best to stay with one of the two schemes in Table 22.6. The XL sheet uses general phrasing so it does not refer to the specific grades used in Table 22.5; instead it simply refers to various 'Grades' on an ordinal scale. Where there are fewer than ten grade points on the scale, the additional spaces in the tables are left empty.

Using the linear weighting scheme, weighted Kappa for the data in Table 22.5 is 0.896. The simple unweighted Kappa would be slightly lower at 0.858. The more generous result with weighted Kappa reflects the fact that where there are discrepancies, they are generally of low severity, there being several minor disagreements but no cases of the most extreme disagreement (Doubtful versus Definite).

22.2.2.2 Random and biased disagreement Section 22.2.1.1 pointed out that a low value of Kappa when analysing nominal categorisations could be due to random or biased disagreement. The same applies with ordinal data such as these gradings of ADRs. In Table 22.5 there are 16 cases where the experts did not achieve exact agreement. These appear to be scattered fairly evenly above and below the diagonal and the disagreements are almost certainly simply random in nature. However, we could have a situation where the disagreements were (say) predominantly below the diagonal. That would mean we had a surfeit of cases where expert number two was assigning higher grades than the other expert. This would then be bias. If there is an appearance of bias, it could be tested for formally using the Wilcoxon paired samples test (Section 21.4.1) as we have paired results for each of the possible ADRs.

22.2.3 Continuous measured (Interval) outcomes

The problem here is to compare two sets of measurements on the same samples/patients and so on where the outcome is on an interval scale.

We will consider a method comparison problem. At question is the level of agreement between measurements of serum fluconazole made by High Performance Liquid Chromatography (HPLC) and by Gas Chromatography/Mass Spectroscopy (GC/MS). Twenty serum samples have been analysed by both methods (Table 22.7). There is just one set of measurements by HPLC, but six hypothetical sets by GC/MS; these illustrate a number of features that may emerge. In the final data set, only six

Table 22.7 Measurements of serum fluconazole (mg.L^{-1}) by HPLC and GC/MS for 20 serum samples

Sample number	HPLC	GC/MS 1	GC/MS 2	GC/MS 3	GC/MS 4	GC/MS 5	GC/MS 6
1	24.3	23.8	27.7	25.8	29.7	21.5	
2	14.4	12.0	11.0	14.0	13.0	13.7	12.0
3	21.7	21.5	26.8	23.5	28.8	20.0	
4	25.3	23.8	20.6	25.8	22.6	21.5	
5	16.5	16.8	22.8	18.8	24.8	16.9	
6	18.5	19.8	18.7	21.8	20.7	18.9	
7	13.1	13.9	9.8	15.9	11.8	14.9	13.9
8	22.2	21.8	16.9	23.8	18.9	20.2	
9	18.5	18.9	25.8	20.9	27.8	18.3	
10	24.7	23.9	31.6	25.9	33.6	21.6	
11	10.9	12.5	8.4	14.5	10.4	14.0	12.5
12	10.1	9.2	14.1	11.2	16.1	11.8	9.2
13	24.4	23.2	27.0	25.2	29.0	21.1	
14	11.6	11.9	10.1	13.9	12.1	13.6	11.9
15	14.3	13.7	10.8	15.7	12.8	14.8	13.7
16	8.0	6.8	9.3	8.8	11.3	10.2	
17	26.0	26.2	26.8	28.2	28.8	23.1	
18	18.2	18.1	17.7	20.1	19.7	17.7	
19	9.0	10.4	16.4	12.4	18.4	12.6	
20	6.0	6.0	2.1	8.0	4.1	9.7	

samples have been re-analysed by GC/MS. We will compare each of these six sets of GC/MS data against the one set of HPLC measurements.

Figure 22.1 shows plots of fluconazole concentration for each of the sets of data for GC/MS against the results for HPLC. Exact agreement is indicated by the dotted 45° line.

- Figure 22.1 (a) shows good agreement; all points are close to the 45° line.
- Part (b) shows a problem of random disagreement. The points are widely scattered away from the line, but there is no systematic displacement of the points either above or below the line.
- Part (c) shows bias. For any sample, the result obtained by GC/MS is likely to be greater that that by HPLC.
- Part (d) shows both bias and random disagreement. There is severe scatter and the GC/MS results are systematically greater than those from HPLC.

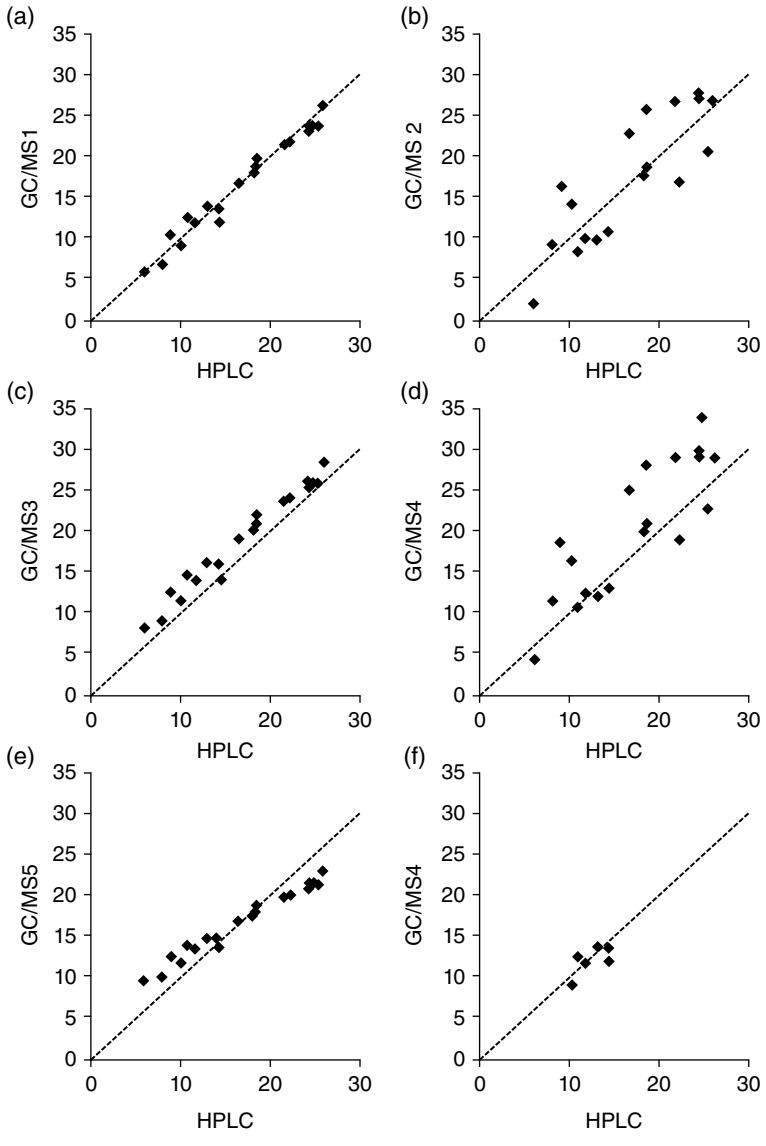


Figure 22.1 Measurements of fluconazole (mg.L^{-1}) in serum samples by GC/MS and HPLC. The dotted line indicates exact agreement between the two methods

- Part (e) shows correlated disagreements. Overall, cases where GC/MS produced higher values than HPLC are balanced by the opposite pattern. However, the discrepancies show a systematic tendency to be higher with GC/MS for low values but lower for high values.

- Part (f) shows a subset of the data from part (a) of this figure; it includes those cases where HPLC yielded a result within the range of 10–15 $mg.L^{-1}$.

Disagreement can take several forms

Disagreement between measurements on an interval scale may take the form of:

- Random disagreement
- Bias
- Correlated disagreement

A satisfactory method for indicating the level of agreement should achieve two things:

- Distinguish between cases such as (a) where there is reasonable agreement and any such as (b) to (e) where there is some form of disagreement.
- Where there is disagreement, help us to diagnose the nature of the problem (Random disagreement, bias or correlated disagreement).

22.2.3.1 The correlation coefficient and why it should not be used Since both analytical methods yield interval scale data, the Pearson correlation coefficient (Section 15.1) would seem to be an obvious way to assess their relationship. However, this only measures how closely the points adhere to a straight line – any straight line, not necessarily the ideal 45° line. The data shown in Figures 22.1 (c) and (e) are almost perfectly linear and would both give a correlation coefficient of +0.99 – near perfect correlation in spite of significant disagreement due to bias in (a) or correlated disagreement (e).

The Pearson correlation coefficient is not appropriate

The Pearson correlation should not be used to compare analytical methods. It only detects random error; it is insensitive to bias and correlated disagreement.

22.2.3.2 Intraclass Correlation Coefficient The Intraclass Correlation Coefficient (ICC) operates in a similar manner to all correlation measures; it indicates degree of relatedness and takes values between minus and plus one. Zero represents no

Table 22.8 Intraclass Correlation Coefficients between each of the GC/MS analyses and the HPLC results

Data set	Intraclass Correlation Coefficient with HPLC results
GC/MS1	0.99
GC/MS2	0.83
GC/MS3	0.95
GC/MS4	0.78
GC/MS5	0.91
GC/MS6	0.71

relationship; and plus one, perfect agreement. Negative values are theoretically possible, but would only arise where one method of measurement tended to produce high values when the other produced low ones and vice versa.

What distinguishes the Pearson correlation from the ICC is that the former measures adherence of the points to any straight line, but the latter determines adherence specifically to the 45° line of agreement. Thus the data in Figure 22.2 (a), where all the points are close to the ideal line, produce an ICC of 0.99. In all of data sets GC/MS2–6 the points fall short of this ideal and the ICC values are lower. Table 22.8 sets out the ICC values for all the data sets in Table 22.7. The data set GC/MS3 [Part (c) of Figure 22.2] is especially important as these results show excellent linearity and hence a near perfect Pearson correlation coefficient ($r = 0.99$), but the points are not on the ideal line and the ICC is lower at 0.95.

Intraclass Correlation Coefficient (ICC)

Unlike the Pearson correlation coefficient, the ICC can only take a value of +1.0 if there is absolute agreement between the two measurement methods. All the points on a graph of one method versus the other must fall on a straight line that follows the ideal 45° gradient.

ICC values can be calculated by some statistical packages such as SPSS. Researchers do need to be aware that the ICC can be calculated in several different ways. To be certain of getting the appropriate form, it may be necessary to ensure that the following options are set correctly:

- The statistical model should assume that the samples/items analysed are a random selection from a potentially much larger population of such items, that is a 'Random factor'. However, the measurement methods used were specifically pre-selected, that is a 'Fixed factor' (see Section 14.4 for a discussion of random and fixed factors). In SPSS the relevant model is referred to as 'Two-Way Mixed'.

- We want the ICC to reflect absolute agreement, not just correlation. For example, the biased nature of data set (c) in Figure 22.2 will be reflected if we opt for Absolute agreement, but not if we select correlation. In SPSS, use 'Absolute Agreement'.

- We need to specify whether our final intention is to use the results from just one of the available methods or to use both of them and take their average as our final answer. For our example, the eventual intention would almost certainly be to use just one analytical method. It is unlikely that we would apply both methods and take their average. The ICC values in Table 22.8 are based on this assumption. In SPSS, you would note the output line labelled as 'Single Measures'.

The ICC is far preferable to the Pearson correlation coefficient and does have a useful ability to identify the presence of any form of disagreement between sets of results. However it is far from perfect – see three points below.

Problem 1 ICC does not identify the specific form of disagreement. The data sets shown in parts (b), (c) and (e) of figure 22.2 all show quite different forms of disagreement. Unfortunately, the ICC is essentially a composite measure whose value is reduced if any (or all) of these problems are present. Consequently a low value of the ICC only flags the fact that there is a problem and does not give any guidance as to the form of the problem.

Problem 2 The value of the ICC is dependent on the range of values among the items being assessed. As explained in Section 15.1.9, correlation coefficients are dependent upon the dynamic range in the data being analysed. The data shown in part (f) of Figure 22.2 considers a limited range of values. It is simply a subset of that in part (a) and ought to be assessed as being just as 'good'. However, because of its limited data range, it actually produces a much lower ICC (0.71 versus 0.99). We could manipulate the ICC to take almost any value we wanted by measuring samples with a wider or narrower range of values. Worryingly, we could generate an unrealistically high value for the ICC by assessing it with a set of samples that contained a much wider range of drug concentrations than would actually be seen in the real samples that would eventually be subjected to analysis.

Problem 3 Over-generous interpretation It is common to see ICC values divided into bands such as: Greater than 0.8 = Excellent; 0.61-0.8 = Good; 0.41-0.6 = Moderate etc. For very subjective measures, these bandings might be perfectly justified and indeed they are often seen in areas such as psychometric assessment where high levels of agreement are difficult to achieve. However, the use of these bandings with instrumental analytical methods would be quite inappropriate. With instrumental methods a high degree of agreement ought to be achievable. Should data such as that shown in Figure 22.7 (b) really be said to show 'Excellent' agreement on the grounds that ICC = 0.83? The target we should be looking for depends hugely on the context. For instrumental analytical methods we should only be satisfied with very high values of the ICC.

22.2.3.3 Bland–Altman plots and accompanying statistics If the ICC value suggests a problem is present then we can diagnose the type of problem by using the graphical method recommended by Altman and Bland (1983; *The Statistician* **32**, 307–317) along with two or three statistical descriptors.

Bland–Altman graphs show the discrepancy between the two methods plotted against the average for the two methods. Notice that it is important that the horizontal axis represents the average for the two methods and not simply one set of measurements or the other.

Table 22.9 shows the necessary calculations for the comparison of the HPLC results versus the first the first set of GC/MS data for fluconazole. Figure 22.2 then illustrates the assessments of agreement between all of the GC/MS measurements and the HPLC values.

- Part (a) in Figure 22.2 shows a more or less ideal outcome. There is relatively little random disagreement; none of the points are far from the zero line. There is also no evidence of systematic bias; there is a fairly even balance of positive and negative differences between the methods. Finally there is no sign of a gradient among the data points which might have suggested correlated disagreements.

Table 22.9 Calculations forming the basis for a Bland–Altman plot to visualise agreement between HPLC and the first set of GC/MS measurements of serum fluconazole concentrations ($mg.L^{-1}$).

Sample number	HPLC	GC/MS1	Average of HPLC and GC/MS1	Difference (GC/MS1 – HPLC)
1	24.3	23.8	24.05	–0.5
2	14.4	12.0	13.20	–2.4
3	21.7	21.5	21.60	–0.2
4	25.3	23.8	24.55	–1.5
5	16.5	16.8	16.65	0.3
6	18.5	19.8	19.15	1.3
7	13.1	13.9	13.50	0.8
8	22.2	21.8	22.00	–0.4
9	18.5	18.9	18.70	0.4
10	24.7	23.9	24.30	–0.8
11	10.9	12.5	11.70	1.6
12	10.1	9.2	9.65	–0.9
13	24.4	23.2	23.80	–1.2
14	11.6	11.9	11.75	0.3
15	14.3	13.7	14.00	–0.6
16	8.0	6.8	7.40	–1.2
17	26.0	26.2	26.10	0.2
18	18.2	18.1	18.15	–0.1
19	9.0	10.4	9.70	1.4
20	6.0	6.0	6.00	0.0

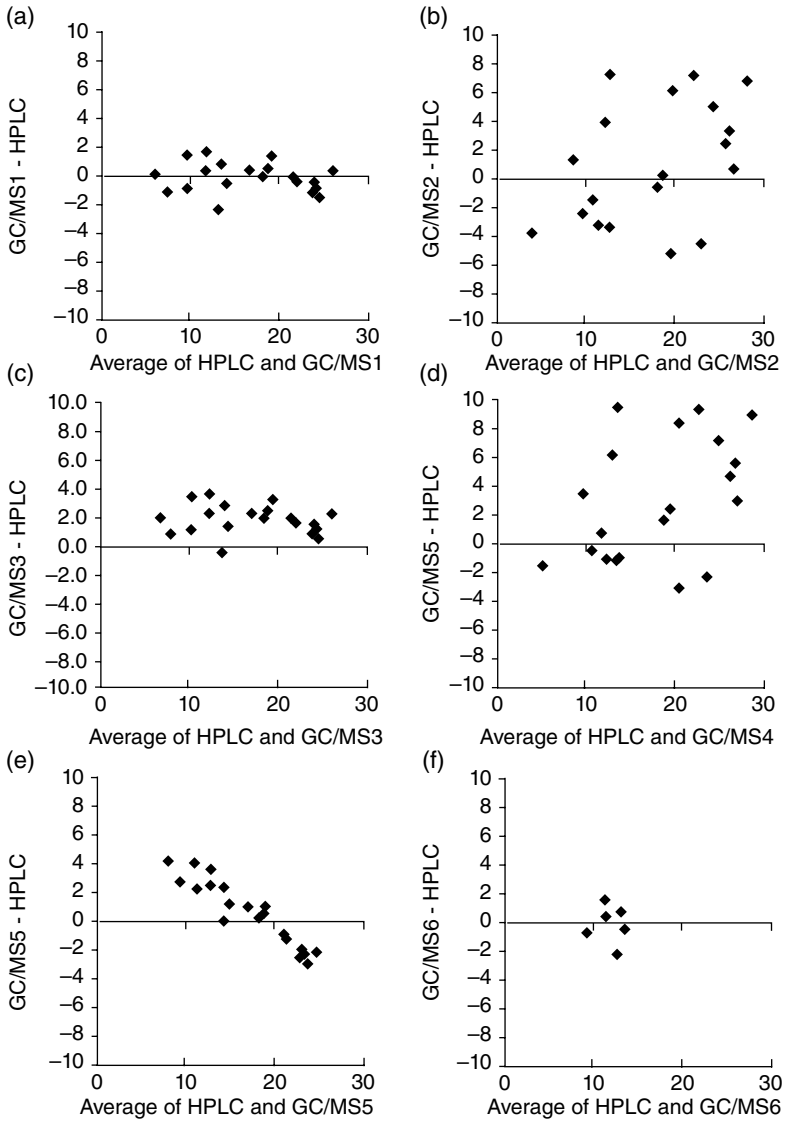


Figure 22.2 Bland–Altman plots – Discrepancies between GC/MS and HPLC measurements plotted against the averages for the two methods. Serum fluconazole concentrations ($mg.L^{-1}$)

- Part (b) shows extensive random disagreement, but no bias or correlated disagreement.

- Part (c) shows bias – the GC/MS values are consistently higher than those from HPLC.

- Part (d) shows both bias and severe random differences.

- Part (e) shows correlated discrepancies – GC/MS values exceed those from HPLC at low concentrations, but at higher concentrations the pattern reverses.

In reality, the Bland–Altman plots pretty well tell the whole story, but folk do like to quote some statistics. Bland and Altman (1990; *Comput. Biol. Med.* **20**, 337–340) recommend calculation of both the mean and standard deviations among the discrepancies between the methods. In the absence of bias, the mean should be close to zero; any significant positive or negative mean value indicates bias. The standard deviation should be small. Graphs such as parts (a) or (c) will produce a low SD, but the discrepancies seen in (b) are much more variable – higher SD.

The only problem that the mean and SD will not identify is correlated discrepancies as in (e). In fact (e) would produce a mean discrepancy close to zero and a high SD, which is the same pattern that we would see with (b). To identify correlated disagreements, it also useful to calculate the correlation coefficient between the discrepancies and the average for the two methods.

Table 22.10 shows the three diagnostic statistics suggested above for all the method comparisons.

For two of the data sets (GC/MS3 and 4) the 95% confidence interval for the mean discrepancy with the HPLC results excludes zero, indicating that there is systematic bias. Two of the standard deviations have been highlighted as high values. There is no exact criterion for what constitutes a 'High' S.D.; the highlighted cases

Table 22.10 Mean and Standard Deviation for discrepancies between HPLC and various GC/MS data sets and Correlation Coefficients with the averages for the two methods. Results that are a cause for concern are given in italics

Data set compared to HPLC	Discrepancy mean (95% C.I.)	Discrepancy standard deviation	Correlation coefficient (*P* value)	Diagnosis
GC/MS1	–0.175 (–0.652 to 0.302)	1.020	–0.175 (*P* = 0.460)	No problem
GC/MS2	0.835 (–1.182 to 2.852)	*4.310*	0.428 (*P* = 0.060)	Random discrepancies
GC/MS3	*1.825* *(1.348 to 2.302)*	1.020	–0.175 (*P* = 0.460)	Bias
GC/MS4	*2.835* *(0.818 to 4.52)*	*4.310*	0.428 (P = 0.060)	Bias and random discrepancies
GC/MS5	–0.080 (–1.209 to 1.049)	2.412	*–0.943* *(P<0.001)*	Correlated discrepancies
GC/MS6	–0.200 (–1.683 to 1.283)	1.413	–0.085 (*P* = 0.873)	Too little data for diagnosis

are simply rather high values in the context of the concentrations being measured. Finally one case (GC/MS5) shows significant correlation between the discrepancies and the values being measured.

Use a Bland–Altman plot accompanied by appropriate descriptive statistics

The best way to diagnose the form of any disagreement between analytical methods is a Bland–Altman plot accompanied by the mean and standard deviation of the discrepancies between the two methods and their correlation coefficient with the average of the two methods.

22.3 Chapter summary

This chapter looked generally at measures of how well sets of answers to questions agree with each other.

We may have several questions and (probably) intend to integrate the answers into a single measure concerning the subject or their opinions. The concern is to establish that all the questions are assessing the same thing. If this is the case, then the answers should all be correlated with one another. The usual measure of this correlation is Cronbach's Alpha. Alpha is a form of correlation and will normally take a value between 0 (no agreement) and +1.0 (Exact agreement). The case is made that Chronbach's Alpha should not automatically be applied whenever we intend to combine several questions into a single indicator; there are cases where a group of questions may assess various factors that contribute to the thing we want to measure, without necessarily being correlated.

We next considered cases where we have a single question that leads to a nominal (categorical) response and two observers are assessing a number of subjects or samples. We want to establish that when the two assessors consider the same subject/sample they generally produce the same answer. For this we use Cohen's Kappa. This also takes values between 0 and +1.0. Where Kappa takes a low value, it makes sense to inspect the results to see whether the problem is one of random or biased disagreement.

In a case similar to that above, but where the outcomes form an ordinal scale of measurement, Kappa can be extended to Weighted Kappa. We grade disagreements; those disagreements that amount to a single point on the ordinal scale are weighted less highly than any cases where the assessments differ by several points on the scale.

Finally, we considered cases where there are two methods to determine an endpoint on a continuous measured (Interval) scale. There are three forms of potential disagreement:

- Random
- Biased
- Correlated

Simple Pearson correlation was dismissed as it only detects the first of these error types, but is insensitive to the other two.

The Intraclass Correlation Coefficient (ICC) is superior in that it will detect any of the three forms of disagreement. The ICC will only take a value near to +1.0 if the data avoids all three potential pitfalls.

The ICC is a composite indicator; a low value only tells you that there is some sort of problem, but does not indicate which of the three types has arisen. If a low ICC value is encountered, the nature of the problem can be diagnosed using a Bland–Altman plot. This is a graph of the discrepancy between the two methods plotted versus the average of the results from the two methods.

23
Survival analysis

This chapter will ...

- Describe the special features of survival data that necessitate distinctive analytical methodologies including the problem of 'censoring'.
- Demonstrate the use of the Kaplan–Meier method to calculate survival, using all available information including that from censored cases.
- Show the use of median survival time and the hazard rate to describe survival.
- Describe the use of survival curves to illustrate the results of Kaplan–Meier analysis.
- Use the log rank test to compare survival in two groups that are differentiated by a single categorical factor. Introduce alternative tests such as the generalised Wilcoxon test that are appropriate where it is believed that a treatment may delay but not ultimately prevent an event such as death.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

- Show how Cox regression can be used as a highly flexible method to analyse the effects of one or more categorical factors, one or more continuous measured factors or a mix of factors.
- Describe the use of the hazard ratio as a measure of how strongly a factor influences survival.

The analyses to be described in this chapter are concerned with the time that passes before a particular event occurs. The more general term 'Time to event' analysis is sometimes used, but in early studies the relevant event was death, hence the usual term 'Survival analysis'.

23.1 What special problems arise with survival data?

23.1.1 Often highly skewed

Survival data is frequently highly positively skewed. Figure 23.1 shows some survival data for cancer patients. Unfortunately some of the patients' disease was so advanced that little could be done and they died within a relatively short period. In others the disease was caught early and effectively eradicated, allowing these patients to survive for many years, achieving a normal life span. The skewed nature of the data is problematic but might be overcome by data transformation.

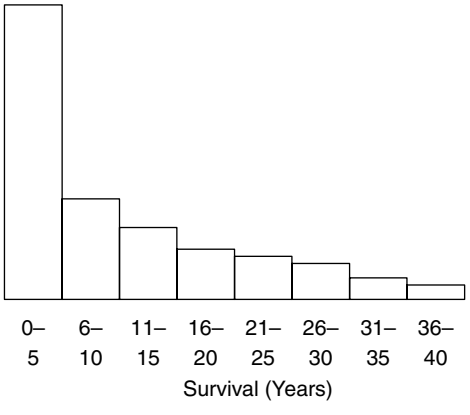


Figure 23.1 Survival data is often strongly positively skewed

23.1.2 Often incomplete – 'Censored'

A more fundamental problem is that survival data is often incomplete. As we saw in Figure 23.1, some subjects may have very long survival times. In most trials and experiments, it is unrealistic to wait for every last individual to die. We need to make reasonable progress and at some time we have to draw a line and analyse the data as it stands, even if currently incomplete. Such data is referred to as 'Censored'. Because we know when these patients' survival period started but not when it would eventually end, you may meet the more precise term 'Right censored'.

It is this problem of censoring that is fundamental; it cannot be overcome by data transformation. This is why we need special techniques to analyse survival data.

23.1.3 Discard the incomplete data?

One obvious thought is that censored cases effectively represent missing data, so should we not simply discard these cases? However, this would bias the outcome. With a situation such as that shown in Figure 23.1, we would easily be able to obtain full data for those who died quickly; it would be the long term survivors that were discarded. Consequently our estimate of survival would be biased downwards.

Cannot simply discard incomplete cases

Discarding the cases where the event was never observed would give a biased (low) estimate of survival.

23.2 Kaplan–Meier survival estimation

Consider a patient who is still alive after five years of follow-up and now has to be considered as a censored case. It is certainly true that we do not know what their ultimate survival will be, but we do have some useful information; this individual survived for *at least* five years. Kaplan–Meier estimation allows us to make use of that partial information.

23.2.1 A small example of Kaplan–Meier estimation

23.2.1.1 Calculating survival Table 23.1 shows the outcome of a small observational study. For a period of 100 days we record every newly diagnosed patient. The day of their diagnosis is recorded as 'Start day'. The outcome is recorded as 'Died' or

'Censored', the latter meaning that they were still alive on day 100 when the study was terminated. 'End day' records when we ceased to observe them, either because they had died or we had reached the end of the study. (All censored cases end on day 100.) 'Observation time' is the time from diagnosis to death or to the end of the study. So, subject number one was diagnosed on day one of the study and died on day 78 giving 77 days of observation whereas subject three was diagnosed on day 40 and was still alive at the end of the study period and is recorded as censored on day 100 (60 days of observation).

Figure 23.2 shows the data diagrammatically. Creating a figure like this is not formally part of the calculation of survival but helps greatly in picturing what is going on. There are a number of things to note:

- The time scale does not represent the day numbers within the study. Dates are reckoned from the start of each patient's period of observation and so they all start at time zero.

Table 23.1 Survival outcomes for a very small study

Subject number	Start day	End day	Outcome	Observation time (days)
1	1	78	Died	77
2	22	76	Died	54
3	40	100	Censored	60
4	54	85	Died	31
5	59	70	Died	11
6	71	100	Censored	29
7	78	98	Died	20
8	90	100	Censored	10

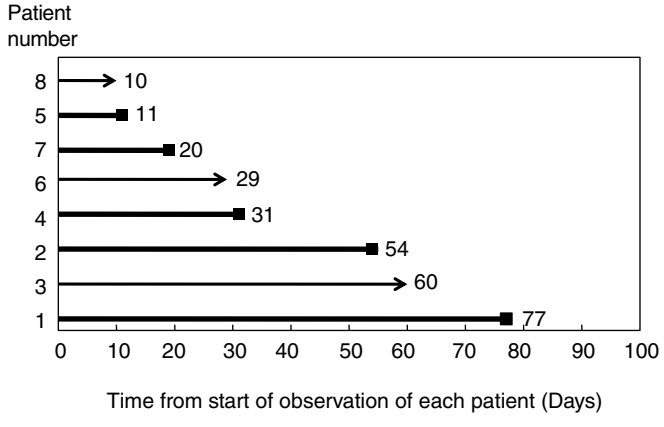


Figure 23.2 Diagrammatic representation of patients' survival data

- The patients have been ranked in terms of their period of observation. Patient eight was observed for the shortest period (ten days) and patient one for the longest (77 days).

- Some lines end with an arrow head – A patient that was censored and we know they lived on for some further (undetermined) period. Others end in a square indicating that they died.

The easiest way to understand the calculations that follow is to imagine that all the patients began their period of observation on the same day (as shown in Figure 23.2). The fact that the observation periods actually started at different times is immaterial. The calculation is shown in Table 23.2.

The first column of the table identifies those days when something of relevance occurred. Day zero sees the initiation of the study and then on each of the other days there is either a death or a patient becomes censored. (Remember that the day numbers are calculated as in Figure 23.2 – all patients considered to have come under observation on day zero.)

In the first line we consider day zero. All eight patients were under observation at the start of this day and there were no deaths or censorings during this day. Zero out of eight patients being observed died, so the proportion of deaths was zero. In the next column we calculate that the survival rate for the day was 1.00. The final column calculates the cumulative survival at the end of the day. To calculate this we take the proportion of the initial cohort that survived up to the beginning of the day and multiply that by the proportion that made it through the current day. For day zero 100% were alive at the beginning of the day and 100% of them survived the day, so we still have 100% of the initial group alive.

In the next line (considering day ten) we still have eight under observation at the start of the day and while somebody is censored on that day, there are no deaths, so as in the previous line there is 100% survival for the day and we still have 100% of the initial cohort alive.

Table 23.2 Kaplan–Meier calculation of survival

Day no.	No. at start of day	Deaths during day	Censored during day	Proportion died this day	Proportion survived this day	Cumulative survival
0	8	0	0	0/8 = 0.00	1.00 – 0.00 = 1.00	1.00 × 1.00= 1.00
10	8	0	1	0/8 = 0.00	1.00 – 0.00 = 1.00	1.00 × 1.00= 1.00
11	7	1	0	1/7 = 0.14	1.00 – 0.14 = 0.86	1.00 × 0.86= 0.86
20	6	1	0	1/6 = 0.17	1.00 – 0.17 = 0.83	0.86 × 0.83= 0.71
29	5	0	1	0/5 = 0.00	1.00 – 0.00 = 1.00	0.71 × 1.00= 0.71
31	4	1	0	1/4 = 0.25	1.00 – 0.25 = 0.75	0.71 × 0.75= 0.53
54	3	1	0	1/3 = 0.33	1.00 – 0.33 = 0.67	0.53 × 0.67= 0.36
60	2	0	1	0/2 = 0.00	1.00 – 0.00 = 1.00	0.36 × 1.00= 0.36
70	1	1	0	1/1 = 1.00	1.00 – 1.00 = 0.00	0.36 × 0.00= 0.00

In the third line (day 11) we now start with only seven under observation having lost one by censoring. We now have a death, so for this day, the proportional death rate is one out of seven = 0.14; survival is therefore 0.86 and cumulative survival is now calculated as 1.00 entering the day but only 0.86 alive at the end.

On line four (day 20) we have one death among the six now under observation. The proportional death rate = 0.17 and proportional survival = 0.83. We then calculate that 0.86 of the cohort had survived up to the beginning of day 20 and 0.83 of these made it through that day, giving a cumulative survival of $0.86 \times 0.83 = 0.71$ at the end of that day.

The process continues in this way through to the last day when the one remaining patient dies giving us a survival proportion for the day of zero and so cumulative survival also falls to zero.

Notice how each line of calculation is dependent upon a figure carried down from the line above. Cumulative survival is always calculated as the proportion making it to the end of the previous day multiplied by the proportion that survived the current day.

23.2.1.2 Ability of Kaplan–Meier analyses to use censored data If you look at the third row of the calculations in Table 23.2, you can see that we calculated the proportional death rate for day 11 as one out of the seven under observation at that time (14%). Those seven under observation included five who were eventually observed to die and two who went on to be censored. Had we not included the two censored individuals in our data set, the proportional death rate would have been one out of five (20%). The Kaplan–Meier method therefore produces lower estimated death rates than we would obtain by simply ignoring censored cases. Consequently Kaplan–Meier increases estimated survival rates.

In Section 23.1.3 it was pointed out that simply ignoring censored cases would produce biased underestimates of survival. Kaplan–Meier is doing exactly what we want; it uses the information available from the censored cases to correct for that potential bias and generate higher (and more appropriate) survival estimates.

23.2.1.3 Graphical representation of survival The data generated by the Kaplan–Meier calculation is normally presented as a survival curve (proportion surviving plotted versus time) as in Figure 23.3. Small vertical tick marks are usually included to indicate the time points where a subject was censored.

Kaplan–Meier analysis uses all the available information

Kaplan–Meier analysis uses the partial information available from censored cases. The subject is known to have survived for *at least* a certain period. This overcomes the downward bias that would arise if we simply discarded the censored cases.

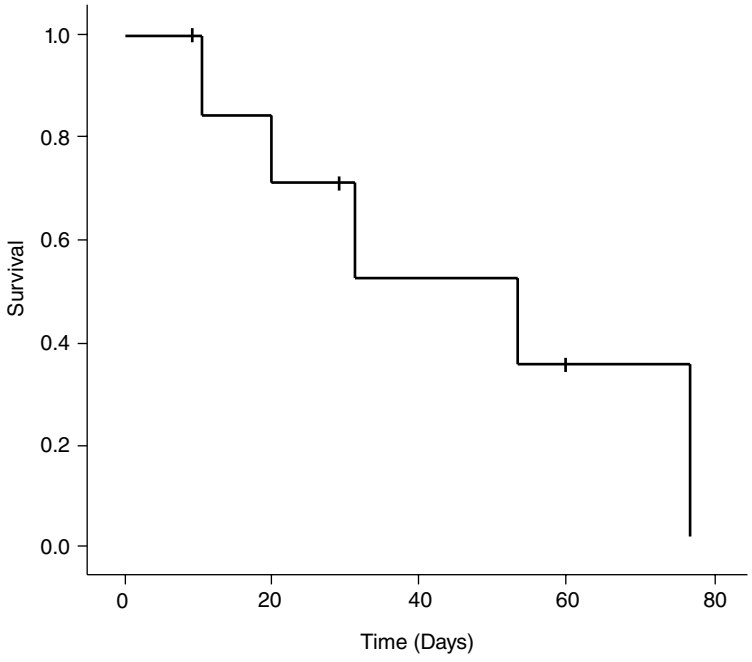


Figure 23.3 Kaplan–Meier survival graph. Small vertical marks indicate times where a censoring occurred

Survival graphs contain rather strange looking sudden downward steps with flat portions in between. These are artefacts generated by the use of sample data to estimate what is going on in the general population. Survival among a large population would be represented by a smoothly declining curve. In our limited sample there are periods when no deaths occur – the flat parts of the graph. When a death does occur it generates a sudden drop in estimated survival. These drops become particularly large in the later parts of the graph when sample size has declined and even one death may represent a significant proportion of the sample under observation. Figure 23.4 superimposes a plausible population survival curve on the sample estimate. With very large samples the drops associated with each death become much smaller and we get a survival graph that comes close to being a smooth curve.

Where death is the relevant event, a graph such as Figure 23.3 showing the proportion who have so far avoided this event is intuitively appropriate. However, there are circumstances where it would be more natural to think in terms of the proportion of subjects who have experienced the event. An example would be a study of time to become pregnant during fertility treatment. 'Survival' in such a situation would be the avoidance of pregnancy – hardly a natural way to look at the data! In this case we can plot one minus survival which represents those who have become pregnant. Figure 23.5 shows a rising curve as more women become pregnant. If the

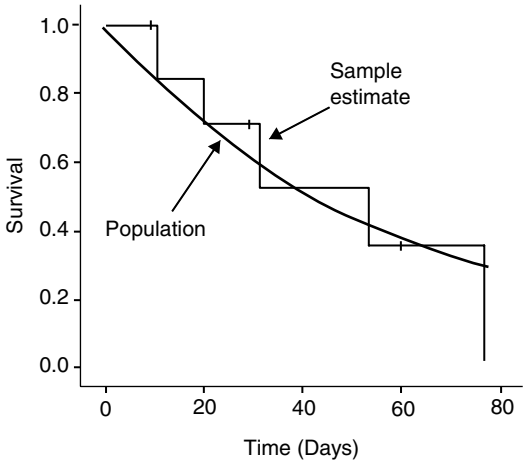


Figure 23.4 The sample estimate of survival contains artificial sudden drops whereas the true population survival pattern would be a smooth curve

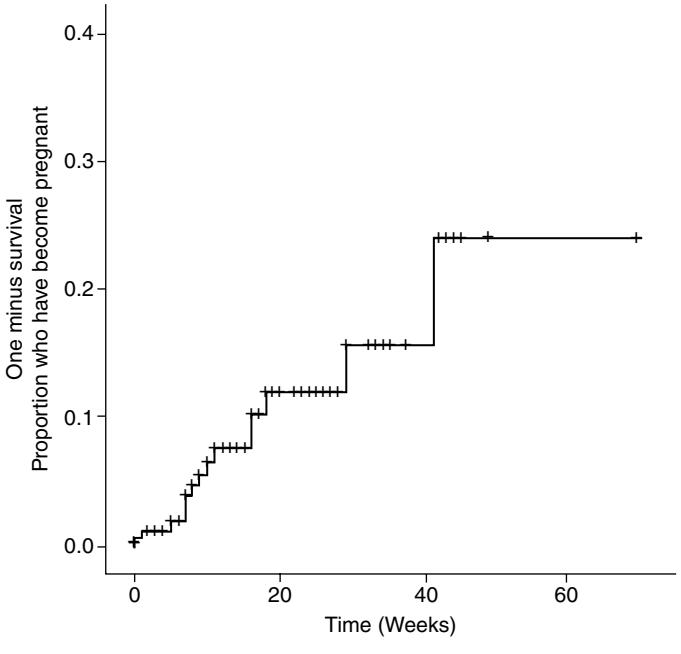


Figure 23.5 Plotting one minus survival to illustrate the proportion of women who have experienced the relevant event – become pregnant during fertility treatment

study had concerned a trial of a contraceptive method, it would make more sense to stay with the traditional plot of survival versus time and emphasise the avoidance of pregnancy.

23.2.1.4 Using a computer to carry out a Kaplan–Meier analysis Some programmes such as SPSS offer good suites of survival analysis routines. To perform a Kaplan–Meier analysis you will need a minimum of two variables that encode:

- The amount of time for which the subject was observed – an interval variable.
- A categorical indicator of whether that subject underwent the relevant event or was censored.

The programme will require you to indicate which value of the second variable codes for the event having occurred. Detailed instructions for using SPSS are included on the website associated with the book (www.ljmu.ac.uk/pbs/rowestats/).

23.3 Declining sample sizes in survival studies

Survival studies are unusual in that the sample size is not fixed. Every time a patient dies the sample size declines. Censorings have the same effect. The precision of any statistical estimate always depends upon the sample size and this applies to survival studies. In the early period of such studies, the sample sizes are greatest and the survival curve is most reliable, but later on as sample size declines, the curve becomes more error prone.

It is dangerously easy to be misled by a visual impression of a change in the later parts of survival curves. Figure 23.6 shows an example where survival among actively treated patients looks superior to that for the controls. However, the most eye-catching differences occur after 40 weeks when there are only nine actively treated and six control patients under observation. The late parts of the graphs are consequently highly error prone and the visual impression of a difference should not be relied upon. In a later section we will give full consideration to formal statistical testing for differences in survival, but it can be said now that this apparent difference is not statistically significant ($P = 0.118$).

It may be useful to indicate sample size at a few strategic time points, so the reader can see if and when the sample size falls to a very low level. This has been done at 20 and 40 days in Figure 23.6.

23.4 Precision of sampling estimates of survival

There are three aspects of any study that determine the precision with which we can estimate survival.

- Initial sample size: In all statistical methods, greater sample sizes always give more precise estimates.

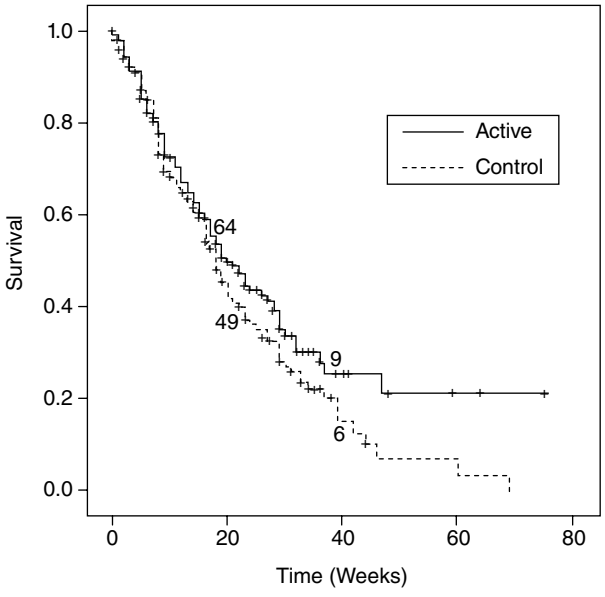


Figure 23.6 Misleading impression of a difference in survival. In the late period, the sample sizes have become meaninglessly small

- Period of time covered by the study: If we want to establish the risk of cardiovascular events in patients with hypertension and follow a group for one month, we are unlikely to learn very much, but follow them for five years and a clearer picture should emerge.

- The event rate: If you want to estimate the risk rate for being struck by lightning it is going to be quite a challenge! Even if you follow 10 000 people for five years you are unlikely to accumulate enough events to achieve any useful precision.

In fact, we can summarise these three into one single measure. The precision of any survival estimate simply depends upon the number of events observed, which in turn, depends upon the three points highlighted above.

Precision of survival estimates depends upon the number of events observed

The greater the number of subjects in whom we observe the relevant event, the greater will be the precision of our survival estimate.

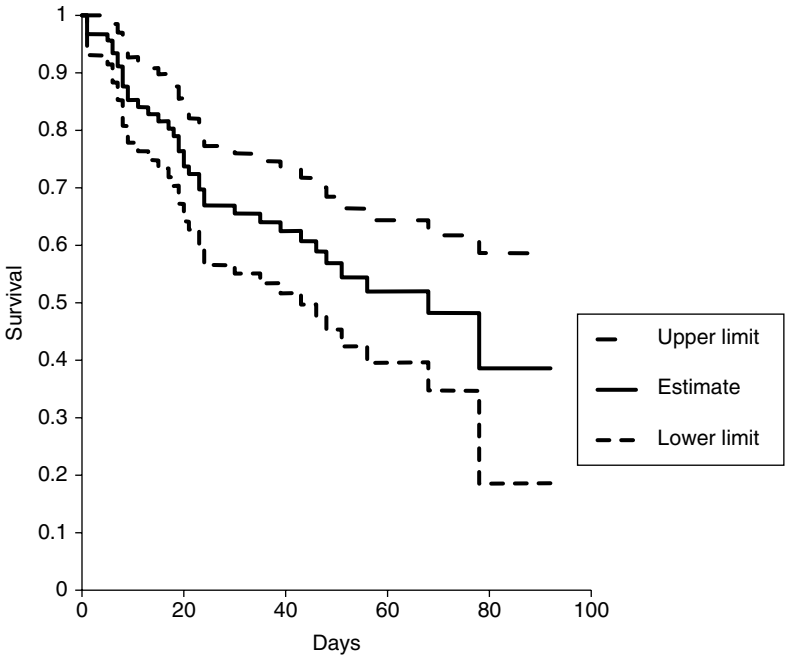


Figure 23.7 95% confidence limits for a survival graph

We can use a 95% confidence interval to indicate the level of uncertainty caused by imprecision in sampling. Figure 23.7 shows an example where the central graph is the point estimate of survival based on our sample and the outer lines indicate how much greater or less survival might be in the general population, taking account of random sampling error.

Section 23.3 pointed out that survival curves become less precise in their later stages due to declining sample size. This is reflected in the width of the confidence interval. At early times In Figure 23.7, the interval covers less than ten percentage points, but by the end it is about 40 percentage points wide.

23.5 Indicators of survival

We will look at the use of median survival time and the hazard rate as indicators of survival.

23.5.1 Median survival time

In other areas of statistics, our usual indicator of magnitude is the mean, however in survival studies it functions poorly. For one thing, survival data is commonly very positively skewed and the influence of a few individuals with long survival can

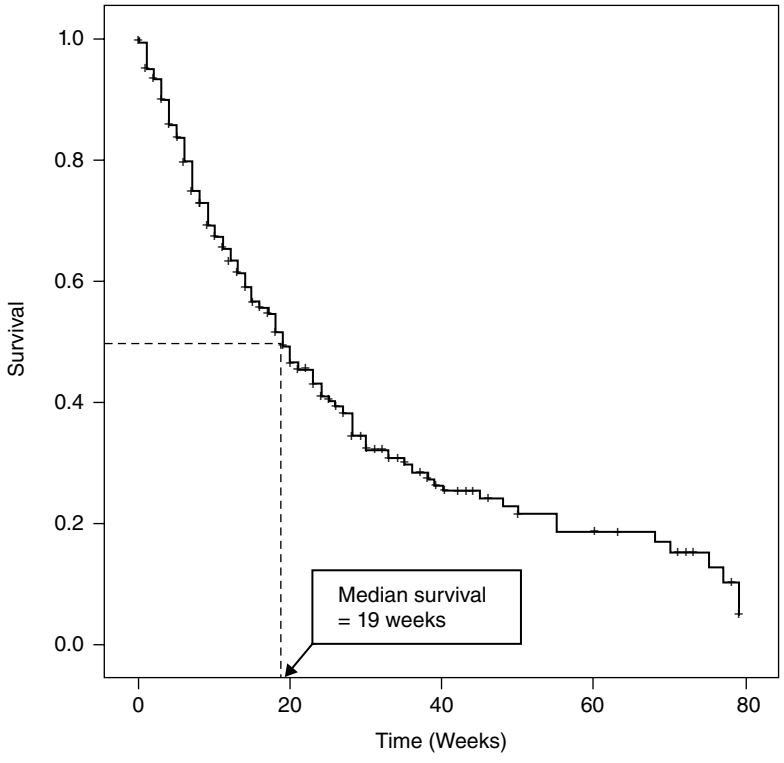


Figure 23.8 Estimation of median survival time

overwhelm larger numbers of more typical individuals with shorter survival. Consequently we generally use the median. This is estimated by reading off the time required for survival to fall to 50% as shown in Figure 23.8; here median survival is 19 weeks.

The figure of 19 weeks is a sample estimate subject to sampling error, so it is useful to report it along with 95% a confidence interval. In this case the interval would be 15.8–22.2 weeks.

Median survival time

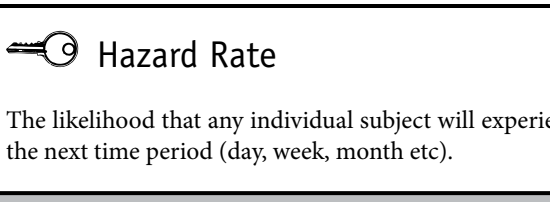
The time at which survival fell to 50%.

One potential problem arises if calculated survival never falls to 50% in which case there is no direct way to calculate median survival.

23.5.2 Hazard rate

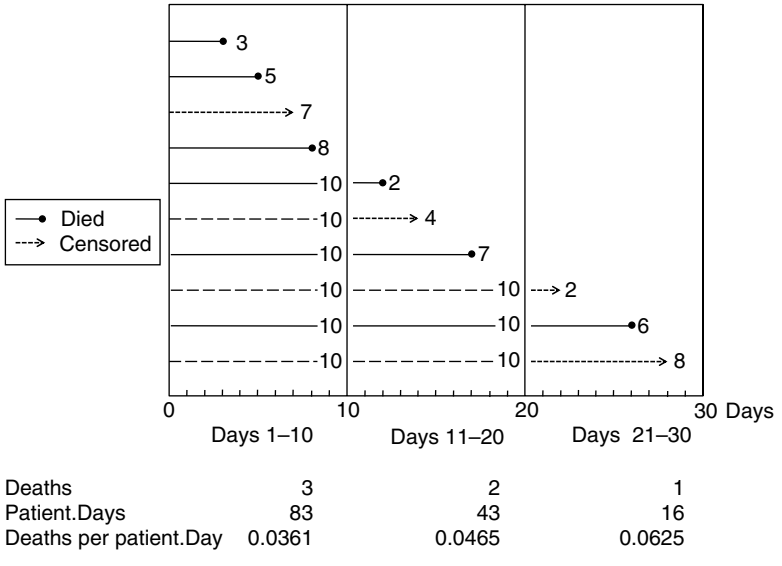
23.5.2.1 Method of calculation Hazard rate is the risk that a particular individual will experience the relevant event during the next period of time. Figure 23.9 shows a very small study (ten subjects) lasting 30 days. Some subjects died and others were censored.

The hazard rate has been calculated separately for each ten day period within the study. During the first ten days there were three deaths. We next calculate the period of observation in terms of patient.days. The lower six patients represented in the figure, all survived right through the first ten days and so they jointly contributed 60 patient.days of observation. The upper four patients contributed three, five, seven and eight days, making a further 23 days. The grand total is therefore 83 patient.days. The hazard rate is then calculated as three deaths divided by 83 patient.days of observation which equals 0.0361 deaths per patient.day. In other words starting from some point in time within that ten day period, any individual patient faced a 0.0361 (3.61%) risk of dying during the next 24 hours. The hazard rates for the next two periods are calculated in Figure 23.9 as 0.0465 and 0.0625 deaths per patient.day. All of these estimates are likely to be rather imprecise as the sample sizes are small and the figure for the final period is especially suspect as it uses so little data.



Hazard Rate

The likelihood that any individual subject will experience the relevant event during the next time period (day, week, month etc).



	Days 1–10	Days 11–20	Days 21–30
Deaths	3	2	1
Patient.Days	83	43	16
Deaths per patient.Day	0.0361	0.0465	0.0625

Figure 23.9 Calculation of hazard rate

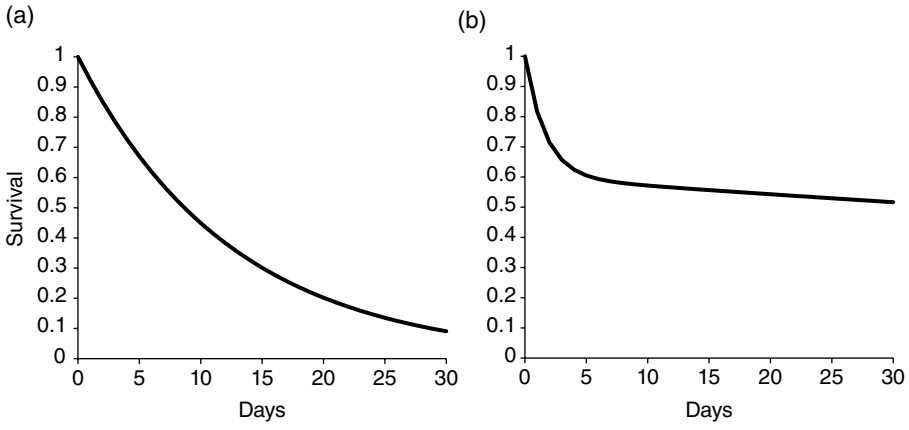


Figure 23.10 Common patterns of hazard. (a) Constant hazard. (b) High initial hazard

23.5.2.2 Patterns of hazard In many situations the hazard rate remains fairly constant over an extended period producing a pattern as in Figure 23.10 (a). A common alternative is shown in part (b) of the figure. Here the initial risk is high but then it falls. An example would be survival of patients following major surgery. The risk of dying during the first 24 hours post surgery is likely to be a lot greater than the risk during the 30th day.

23.6 Testing for differences in survival

Survival analyses really come alive when we start comparing survival in different groups.

23.6.1 The log rank test

23.6.1.1 Performing the log rank test The log rank test allows us to test for the possible effect of a categorical factor. For example we may compare survival among actively treated patients versus that in controls or compare males versus females.

We will look at a simple clinical trial comparing survival among small cell lung cancer sufferers who receive active or placebo treatment. The results are available from a spreadsheet (Figure_23_11_Data.XLS on www.ljmu.ac.uk/pbs/rowestats). Figure 23.11 shows the results and these certainly suggest longer survival among the actively treated patients. We would normally quote separate values for median survival in the two groups, which are 19 weeks for the control group and 29 weeks for those actively treated.

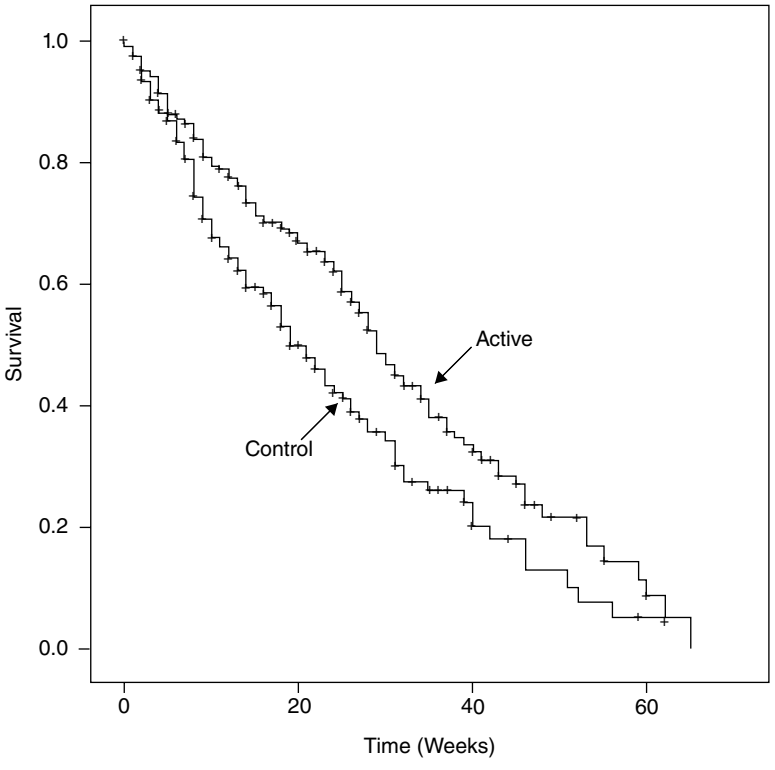


Figure 23.11 A comparative survival study to be tested by the log rank test – Survival among actively treated versus control patients

Survival graphs are notoriously difficult to judge visually and although the difference may look impressive, a formal test is required. The log rank test takes account of two aspects of the data.

- First, we need to consider how different the two outcomes are; how great is the difference between the two groups in terms of the hazard (rates of decline in survival)?

- Second, we take account of how precisely any difference has been estimated. This depends on the precision of the estimates for the two individual groups and as explained in Section 23.4, this depends on the number of events (deaths) observed. Finally, this in turn depends on sample sizes, the period for which the study was continued and the frequency of the relevant event (hazard rate).

This is summarised in Figure 23.12.

The test initially generates a chi square value which is converted to a *P* value. The classical version of the test can be somewhat biased if the number of events observed

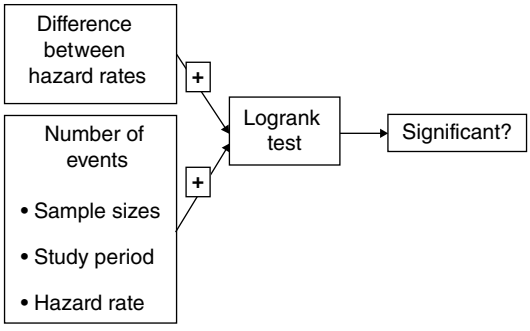


Figure 23.12 Aspects of the data that influence whether a log rank test will prove statistically significant

is small, so statistical packages (e.g. SPSS) may use a variant usually referred to as the 'Mantel–Cox' version. Table 23.3 shows generic output from the log rank test of this data. The difference in survival is statistically significant.

23.6.1.2 Alternative tests that give extra weight to the early period Figure 23.13 shows the survival of two groups of patients in a clinical trial. The actively treated patients seem to have benefitted in the sense that their deaths have been delayed, but at about a year into treatment their hazard rate increased and their survival ultimately became indistinguishable from the controls.

It is clearly to the benefit of the actively treated patients that their deaths have been delayed even if they were not ultimately prevented and we would want this difference to be detected by our statistical analysis. The problem is that there is a substantial part of the graph where there is no difference between the groups; it would be desirable to focus attention on the initial period. This is referred to as increasing the weighting on the early data. There are a number of alternative tests that do precisely that. A commonly used version is the generalised Wilcoxon test.

With the data represented in Figure 23.13, the standard log rank test would yield a non-significant P value of 0.190. However, the generalised Wilcoxon test gives $P = 0.010$. By focusing on the early period, which is most relevant, we achieve statistical significance. These alternative methods are effectively testing for the ability to delay death rather than necessarily effecting a cure.

Log rank and generalised Wilcoxon tests

The most commonly used test to detect the possible effect of a single categorical factor on survival is the log rank test. The generalised Wilcoxon test does a similar job, but is useful where death (or some other event) is delayed rather than prevented.

Table 23.3 Generic output from log rank test comparing survival in actively treated versus control patients with lung cancer

	Log rank test	
	Chi-square	*P* value
Log rank (Mantel–Cox)	8.474	0.004

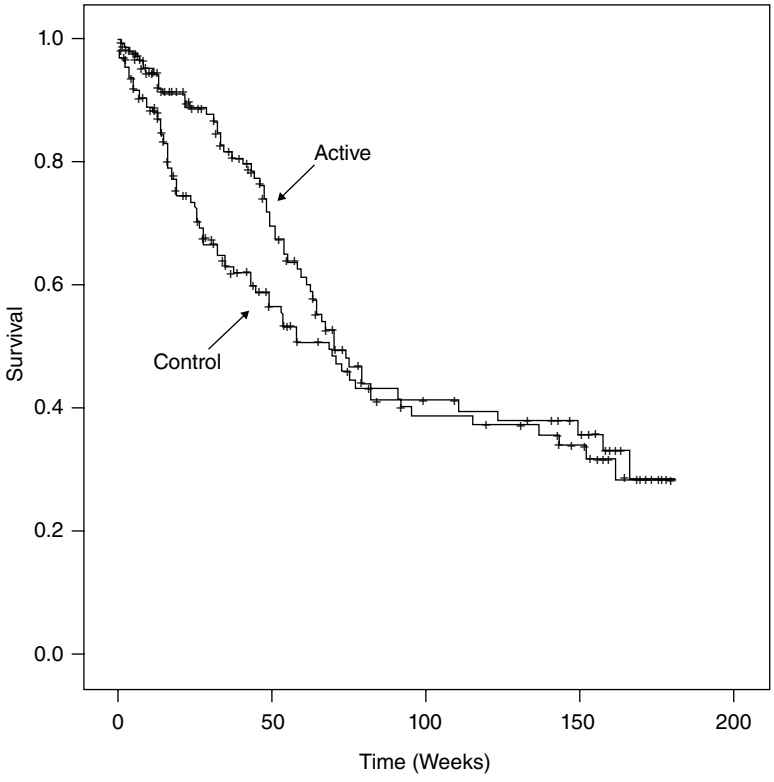


Figure 23.13 A case where active treatment causes a delay in death rather than preventing it entirely. The generalised Wilcoxon test is indicated

23.6.1.3 Using a computer to perform a log rank or generalised Wilcoxon test These tests are usually called up as optional additions to a Kaplan–Meier analysis. The generalised Wilcoxon test may be referred to by other names; for example SPSS refers to it as the Breslow test. In addition to the two variables required for a Kaplan–Meier analysis you will need a further categorical indicator of the group to which each subject belongs.

23.6.2 Cox regression

23.6.2.1 The theory behind Cox regression Cox regression offers a completely general method for the analysis of survival data; it can consider any mix of factors – one or more categorical (nominal) factors, one or more continuous measured (Interval) factors or a mix of factor types. We will start by considering a case with a single interval factor and move onto a more complex case with mixed factor types.

Cox regression aims to model the hazard rate based on the relevant factor(s). Simple regression generates equations in the form:

$$Y = \text{Constant} + \text{B} \times X$$

We could attempt to model hazard, using age as a predictor. On the above basis we would get an equation in the form:

$$\text{Hazard} = \text{Constant} + \text{B} \times \text{Age}$$

However, this is an unrealistic model. For one thing, it could predict a negative hazard if Age took a sufficiently extreme value. To produce a more realistic model, Cox regression simply exponentiates the right hand term of the equation (see explanation of '*Exponentiation*' in nearby *Key point box*). A Cox regression equation takes the form:

$$\text{Hazard} = \text{Exp}\left(\text{Constant} + \text{B} \times \text{Age}\right)$$

Exponentiation

Exponentiation uses a mathematical constant called Euler's *e* which takes the value 2.718. Exponentiation then consists of raising *e* to the power that we want to exponentiate, thus the exponent of *n* is e^n. This is often written as 'Exp(*n*)'.

A Cox regression equation might predict hazard (likelihood of event in the next year) as say:

$$\text{Hazard} = \text{Exp}\left(-4.5 + 0.03 \times \text{Age}\right)$$

If a subject was aged 60, their hazard would be estimated as:

$$\begin{aligned}\text{Hazard} &= \text{Exp}\left(-4.5 + 0.03 \times \text{Age}\right)\\ &= \text{Exp}\left(-4.5 + 0.03 \times 60\right)\\ &= \text{Exp}\left(-4.5 + 1.8\right)\\ &= \text{Exp}\left(-2.7\right)\\ &= 2.718^{-2.7}\\ &= 0.0672\end{aligned}$$

We estimate that the subject has a 6.72% likelihood of experiencing the relevant event during the coming year.

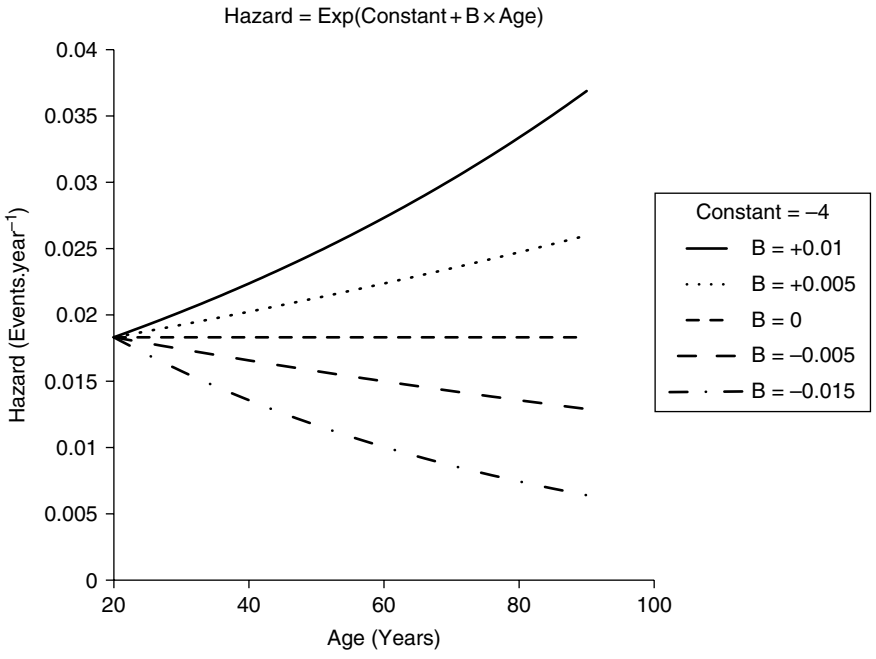


Figure 23.14 The effect of changing the value of the B coefficient in a Cox regression equation predicting hazard from age

Based on this form of equation, Figure 23.14 shows the influence of age on hazard for various positive and negative values of the coefficient B. In all cases the constant is maintained at a value of –4. Positive coefficients model an effect whereby greater age is associated with greater hazard. In contrast, negative coefficients suggest hazard declines with age. The null hypothesis is that the coefficient B = 0; hazard is not affected by age.

The effect of the constant is to control the vertical position of the curves. With greater values of the constant, all the graphs would simply move up the vertical axis.

One thing to be wary of with Cox regression is that it is very easy to think that a positive coefficient means that the factor is positively associated with survival. In fact the opposite is the case. Cox regression models hazard. Thus, if (for example) age carries a positive coefficient, then hazard will increase with age, but survival will be reduced among older subjects.

It can be seen in Figure 23.14 that with negative coefficients, the graphs approach zero asymptotically. Even if we considered some extreme value for age, the calculated hazard would remain positive. Thus, the Cox model avoids any possible unrealistic predictions of a negative hazard (unlike the simpler equation suggested above).

In Cox regression, the values of the constant and the coefficient are adjusted so that the resultant equation is maximally consistent with observed data. In most cases, it is the value of the B coefficient(s) that we are interested in rather than the constant. The constant is essential to the modelling process, but rarely of direct interest.

23.6.2.2 An example with a single interval factor. Does weight influence the development of Type II diabetes? A spreadsheet (Table_23_4.XLS) is available from www.ljmu.ac.uk/pbs/rowestats. It describes subjects aged between 58 and 63 none of whom had any type of diabetes at the time of recruitment into the observational study. Details are given for their weights (in kg) when recruited into the study, the time (months) for which they were observed and whether they developed Type II diabetes while under observation.

Table 23.4 gives generic output for a Cox regression analysis of the data considering weight as a continuously varying factor that might influence the hazard of developing diabetes. The *P* value (0.017) shows that weight is a statistically significant factor. The coefficient (B) is positive indicating that greater weight is associated with a greater hazard for the development diabetes. However the value of +0.024 does not directly reflect the strength of the relationship because the regression equation includes exponentiation. The real measure of effect size is therefore Exp(B). This is shown in the table as 1.025 and is referred to as the Hazard Ratio (HR). Its interpretation is that if two individuals have values of the relevant factor that differ by 1.0, then the ratio of their hazards will be 1.025. In the current case if one individual was 1 kg heavier than another then the risk that the heavier individual would develop diabetes during the next month would be 1.025 times that for the lighter person. With greater weight differences, the difference in hazard increases exponentially; a weight difference of 20 kg would be associated with a $1.025^{20} = 1.64$-fold increase in risk.

23.6.2.3 An example with both continuous measured and categorical factors. Treatment of severe congestive heart failure survival The example considers multiple factors that may affect survival among patients being treated for severe congestive heart failure. The main interest is the effect of ACE inhibitor treatment, but we want to control for the possible effects of various other factors (similar to the concept underlying analysis of covariance – Chapter 16) The factors are:

- BMI: Interval factor ($kg.m^{-2}$).
- Age: Interval factor (years).

Table 23.4 Generic output for Cox regression analysis of effect of weight (kg) on development of Type II diabetes

Cox regression			
	B	Hazard Ratio Exp(B)	*P* value
Weight	0.024	1.025	0.017

- Gender: Nominal factor recorded as 'M' or 'F'.
- Treatment: Drug treatment. Nominal factor recorded as 'Active' or 'Placebo'.

The relevant event is death. Observation time is recorded in months. The data is available in a spreadsheet (Table_22_5.XLS on www.ljmu.ac.uk/pbs/rowestats).

Most statistical packages will allow the inclusion of factors in a non-numerical format (such as gender above). The program will convert these to numerical indicator variables.

When there is more than one factor to consider, the Cox regression equation simply adds extra terms to the format seen previously:

$$\text{Hazard} = \text{Exp}\left(\text{Constant} + B_1 \text{x Factor}_1 + B_2 \text{x Factor}_2 \ldots.\right)$$

Table 23.5 provides generic output for the Cox regression analysis.

The main interest is drug treatment and its *P* value is <0.001 – highly significant.

To interpret this finding we need to look at how the program has encoded the two treatments. The first part of the output tells us that Active was encoded as 1.0 and Placebo as 0.0. The negative B coefficient tells us that the hazard is lower for the actively treated group since it was coded with the higher value.

The Hazard Ratio – Exp(B) – for treatment is given as 0.434. Remember that the HR tells us the change in hazard associated with a one unit change in the value of the factor. The difference between the numerical codes for the two treatment groups is exactly 1.0, so the hazard for an actively treated patient is only 0.434 times that for somebody placebo-treated.

BMI is also highly significant with a positive B coefficient, so heavier patients are at greater hazard. It seems to have a massive hazard ratio of 2.397. However the way BMI has been recorded gives a standard deviation of only 0.286, so an increase in

Table 23.5 Generic output from Cox regression analysis of possible effects of BMI, age, gender and ACE inhibitor treatment on survival among patients with severe congestive heart failure

Cox regression

Coding of categorical variables

Gender: F = 1; M = 0

Treatment: Active = 1; Placebo = 0

	B	Hazard Ratio Exp(B)	*P* value
BMI	0.874	2.397	0.001
Age	0.032	1.033	0.000
Gender	0.004	1.004	0.976
Treatment	–0.835	0.434	0.000

BMI of 1.0 would represent an enormous increase in obesity; an over twofold increase in hazard is not so surprising.

Age is also statistically significant with a positive B coefficient but a seemingly rather small hazard ratio of only 1.033. However, this represents the increased hazard associated with just a one year age difference. A 25 year age difference would bring a 1.033^{25} = approximately 2.25-fold increase in hazard.

Finally gender is non-significant. Had it been significant, it is females who received the coding of 1.0 and there is a positive B coefficient, so females would have been at greater risk than males.

Cox regression is completely general

Cox regression can be used to model any experimental structure. There may be:

- One or more categorical factors.
- One or more continuously varying measured factors.
- A mix of both types of factor.

23.6.2.4 Does Cox regression make log rank testing redundant? Cox regression is quite capable of handling the sort of studies that would be suitable for analysis by log rank testing (a single categorical factor) and it has the advantage of producing an indicator of the size of any effect (the hazard ratio). Log rank testing could therefore be dispensed with. However, it has been around for a long time and is familiar to many workers, so it is unlikely to disappear any time soon.

23.6.2.5 Using a computer to carry out a Cox regression analysis You will need the following variables:

- An interval variable to represent the time of observation of each subject.
- A categorical variable to represent the outcome (Event occurred or censored).
- An interval variable for each continuously measured factor.
- A nominal variable to represent each categorical factor.

The website associated with this book (www.ljmu.ac.uk/pbs/rowestats/) includes detailed instructions for using SPSS to perform a Log Rank test or Cox regression.

23.7 Chapter summary

'Survival' or 'Time to event' analyses are essential in those cases where we are recording how long it takes until an event occurs and we have some individuals in whom the event is never recorded (censored cases).

Kaplan–Meier analyses are able to use all available information. Even the censored cases provide the information that an individual survived for at least a certain period of time without the event occurring and this is incorporated into a Kaplan–Meier analysis.

The usual measures of how well subjects survive, on average, are the median survival time and the hazard rate.

The precision of sample estimates of survival depends solely on the number of events observed. This in turn depends upon the initial sample size, the length of the observation period and the hazard rate.

Where we want to determine whether a single categorical factor influences survival the most commonly used test is the log rank test. Alternative tests are available that place additional emphasis on the early part of the period studied, for example the generalised Wilcoxon test. These are useful where it is suspected that a treatment may delay but not prevent an event.

Cox regression allows us to model the effect of any combination of categorical and/or continuous measured factors on the hazard rate for an event. It generates a coefficient (usually called a B coefficient) that describes the effect of each factor. A positive coefficient implies a positive relationship – as the value of the factor increases, the hazard also increases. But, remember that this in turn means reduced survival. Negative coefficients mean that increasing the factor reduces the hazard and increases survival.

The exponential of a B coefficient [Exp(B)] is referred to as the Hazard Ratio (HR). This is a measure of the relative change in hazard if the value of the factor increases by 1.0. It is the main measure of the extent of the effect of the factor on hazard and hence on survival.

24
Multiple testing

This chapter will ...

- Explain what multiple testing is and why it is a problem.
- Describe the circumstances where it is most likely to arise.
- Review methods that can be used to prevent false conclusions being relied upon.

24.1 What is it and why is it a problem?

With every statistical test we perform, we face the fact that if no real effect were present, there would still be a 5% risk of a false positive. We accept this small risk along with every *t*-test or correlation analysis and so on. The real problem comes when we indulge in a plethora of tests and the generic term 'Multiple testing' covers this scenario. It is one thing learning to live with the standard 5% risk, but multiple tests will jointly entail a much greater hazard that somewhere, a false positive will creep in.

Taken to extremes, multiple testing is virtually guaranteed to find 'statistical significance' even in the absence of any real effects. For example we could measure ten different endpoints in a series of people (Height, blood pressure, hand-grip strength etc.) and those endpoints might be entirely independent of one another.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

We could then test for significant correlation between all possible pairs of endpoints. This would require 45 analyses, each carrying the standard 5% risk. The chances of hitting at least one meaningless (but apparently significant) correlation would then be about 90%.

In most cases multiple testing arises from naivety, but in others it's the product of an overzealous searching for statistical significance and no doubt there is a hardcore of outright villainy – the sure and certain knowledge that if you do enough tests, you're bound to get lucky eventually.

Multiple testing

A single statistical test carries a 5% risk of producing a false positive conclusion – generally considered an acceptable level of risk.

Repeated tests rack up a much greater (and ultimately unacceptable) risk.

In this chapter, we will first review the situations where multiple testing can arise and then look at various stratagems that should help avoid excessive risks of false positives.

24.2 Where does multiple testing arise?

The commoner sources of this problem are:

- Comparing several treatments against each other.
- Recording numerous endpoints and testing each one for changes.
- Measuring the same end point on several occasions and testing at each time point.
- Breaking the data into numerous sub-sets and testing within each of these.

We will look briefly at each of these.

24.2.1 Comparing several treatments

Within a single piece of work, we might assess several possible treatments and measure the same endpoint for each treatment. If the endpoint is a measured variable, a *t*-test for each possible pair of treatments might look tempting or with a categorical endpoint, repeated chi-square tests. The potential for multiple testing rises rapidly if we insist on making every possible comparison among 3, 4 or 5 and so on different treatments.

24.2.2 Comparing several endpoints

Even when comparing just two treatments, there may be a whole raft of endpoints. When comparing two chemical products, we might measure levels of large numbers of different impurities. Similarly in a biological setting, we might want to see whether a dietary change alters blood lipids and end up measuring triglycerides, total cholesterol and high, low and very low density lipoprotein cholesterol. If we performed separate *t*-tests on each impurity or each type of lipid, we would run a considerable risk of detecting false differences.

24.2.3 Testing at several time points

Particularly in biological and medical experiments, we frequently determine the relevant endpoint at several time points. So, we might take a group who begin yoga and a control group who do not, and measure subjects' blood pressure every month for the following year. If we then carried out *t*-tests (control versus yoga) at each of the 12 monthly observation points, we would again run an elevated risk of false positives.

24.2.4 Testing within numerous subgroups

This is one of the worst danger areas for multiple testing. Problems arise if we experiment on a mixed bag of subjects and then start breaking the subjects down into subgroups based on their gender, age, ethnicity, disease status or whatever. When testing a medical treatment, we might start with a single statistical analysis based on the whole data set, which is perfectly respectable. If that fails to achieve significance, we might then think 'Well, maybe this stuff only works in people of a certain age', so we divide the data into six subsets each containing subjects in the same decade of life. We then compare controls versus actively treated within each age group separately. We now have a significant degree of multiplicity and are likely to 'discover' that (say) subjects aged 60–69 are responsive to our treatment, even if nobody else is. If that doesn't work then we could subdivide by gender (two subgroups) or the severity of disease (mild, moderate, severe – three subgroups). The real killer is to start combining several criteria. In the case above we could potentially generate $6 \times 2 \times 3 = 36$ subgroups ranging from males, aged 20–29 with mild disease to females, aged 70–79 with severe disease. With 36 subgroups, it would be very surprising if none of them generated an apparently significant effect.

If an honest author describes the analysis of half a dozen different endpoints or 20 subgroup analyses, one of which is significant, then the multiplicity is overt and an educated reader can be suitably cautious. However, the unscrupulous may conduct endless analyses and only report the one that hit the jackpot – covert multiplicity. The latter is out and out dishonesty – you might as well just make the results up!

Against covert multiple testing, the reader will remain pretty much defenseless until journals upgrade their procedures (see Section 24.4).

Break the data down into large numbers of subgroups and publish the one that produces a 'significant' result

Wicked! The effectiveness of this one depends only on how unscrupulous you are prepared to be. If you are prepared simply to publish the subgroup that best suits your purposes and keep quiet about the fact that it was part of a larger experiment, there is no limit to what you can prove. Those with some residual scruples may prefer to own up to the full data set, but invent some spurious reason why there is justification for focusing on the particular subgroup where a significant result was obtained.

This trick works best where there are plenty of apparently logical reasons for subdividing the data and human data is ideal for this. Subdivide by gender, age, ethnicity, social class and disease status (or any combination of these) and when you find that your new treatment reduces blood pressure to a statistically significant extent among caucasian, protestant, males aged 35–55 who also suffer from migraine, then this is obviously the group upon whom we should especially focus.

24.3 Methods to avoid false positives

24.3.1 Use a single ('omnibus') test to avoid a series of pair wise comparisons

24.3.1.1 Multiple treatments Where we've compared several different treatments, recording the same endpoint for each one, an omnibus test should generally be the first step in analysis. If an experiment addresses the question of whether a measured variable changes under a range of different conditions, we have already seen (Chapter 14) that a single analysis of variance will avoid the problems that would arise with repeated *t*-tests. If the omnibus test proves significant, then we can start digging into the detail to find out which treatment differs from which other. We have seen that a Dunnett's or Tukey's test can legitimately be used to make detailed comparisons.

24.3.1.2 Multiple endpoints For this scenario, there is a whole other area of analysis referred to as 'Multivariate' statistics. These methods also allow omnibus testing – all endpoints are considered simultaneously. Multivariate stats are among the most complex to perform and interpret – there are generally six different ways to do any one job! However, don't be put off. Multivariate stats can uncover subtle effects that

you would never find by looking at each endpoint in isolation. What you will need is seriously competent statistical advice.

24.3.1.3 Multiple time points With observations at several time points, all the data can be considered simultaneously using a technique called Repeated Measures Analysis of Variance. This can be carried out using packages such as SPSS or Minitab. Its execution and interpretation is not unduly complex, but is probably best undertaken with expert statistical support.

24.3.2 The Bonferroni correction

24.3.2.1 Bonferroni keeps the rate of type I errors down to 5% … In many instances, omnibus tests are not possible. If there are several comparisons to be made, each involving different groups of individuals and/or different endpoints, there may be no choice but to use a series of discrete statistical tests. In these circumstances, the Bonferroni correction may be applied to maintain the overall risk of false positives at the standard level of 5%.

What the Bonferroni correction does is to raise the standard of proof for all the individual tests. Each test is then less likely to produce a false positive and the complete series of analyses will jointly generate a 5% risk.

The usual way to implement the correction is to adjust the critical P value below which we claim statistical significance. This is calculated as:

$$\text{Critical } P = 0.05/n$$

where n is the number of tests to be carried out.

With five tests, we would only declare the result of any individual test to be significant if P was less than 0.05/5 = 0.01. Under these arrangements, each test will generate only a 1% chance of a false positive and the series of five will entail a 5% risk.

24.3.2.2 … but it reduces the power of the tests to which it is applied The Bonferroni correction does a good job of curing the problem it's advertised as fixing – keeping the risk of type I errors down to 5%, but unfortunately it also introduces a new problem. The raised standard of proof required for each individual test, increases the chances that a real effect will not be declared significant (more type II errors and therefore reduced power). With the example above, one of the five tests might be for an effect that is genuinely present and it might generate a P value of (say) 0.02. If we had tested for this effect in isolation, we would have used the normal criterion ($P < 0.05$) and we would have obtained a perfectly legitimate significant result. However, once the result is lumped in with four others, it will fail to meet the more demanding Bonferroni corrected target. The greater the number of possibilities considered, the greater the risk that a real effect will be lost by dilution among a load of irrelevant factors.

Bonferroni correction – There is no such thing as a free lunch

Keeps the overall risk of any false positives down to 5% by adjusting the level at which *P* is considered significant.

It also has the unfortunate effect of reducing the power of all tests to which it is applied.

24.3.3 Distinction between primary and secondary (exploratory) analyses

24.3.3.1 Primary question, primary endpoint and primary analysis Another way to avoid the hazards of multiple testing is to highlight one particular route through the experimentation and data analysis. Whatever conclusion arises from this route will then be claimed as definitive.

The work may have posed several questions, collected several endpoints and each endpoint may have been subject to more than one statistical analysis. To fully specify a definitive route, we need to specify:

- What is the primary question being asked?
- What will be the primary endpoint that answers that question?
- What will be the primary statistical analysis of that endpoint?

There is then one definitive answer based upon one definitive route and we are guilty of no multiplicity. We can then look at as many additional questions and indulge in as many secondary (or 'Exploratory') analyses as we wish. Any conclusions arising from exploratory analyses are subject to all the hazards associated with multiple testing and must be considered suitably suspect. Indeed a good way to view such 'Conclusions' is that they are not really answers to questions at all, rather they provide guidance as to what might be useful further research. You may see a distinction drawn between 'Hypothesis testing' – the role of the primary analysis and 'Hypothesis generation' – the role of secondary analyses.

For example, a primary analysis shows that there is no significant evidence that a medical treatment changes outcomes, but a secondary subgroup analysis suggests that it does produce an effect in younger subjects. In that case, the definitive conclusion would be that the current experiment has not shown an effect. However, the secondary conclusion might provide adequate motivation to repeat the work using only young subjects and for the new experiment, the primary question would become 'Does this treatment work in young people?'.

To have any legitimacy, the identification of a primary analysis must be finalised before the experimental data has been seen.

Primary question – Primary endpoint – Primary analysis

Specify a primary question, primary endpoint and primary analysis of that endpoint. The answer obtained from these can legitimately be claimed as definitive with no contamination by multiplicity.

However, you may also want to take the opportunity to look at several subsidiary questions, which will involve the measurement of additional endpoints. It may then be interesting to carry out all manner of analyses of the various endpoints. Any conclusions thus obtained must be recognised as secondary and tentative. If one of these other conclusions is of particular theoretical or practical importance, it can become the primary question of further research.

The dangers of relying upon the findings of exploratory analyses were nicely summed up by Stephen Senn:

> '*Enjoy the result you have found by exploratory data analysis, for you will not find it again.*'

24.3.3.2 Fishing trips We could push the previous idea one stage further and publish indiscriminately (all the data and all the analyses) without any correction for multiple testing, but accept that all conclusions are tentative and should only be used as the basis for hypothesis generation not hypothesis testing.

The collection of masses of data and endless statistical analyses is often referred to derogatorily as a 'Fishing trip'. If the authors of this type of work try to pretend that they have produced any reliable conclusions, then they deserve to be pilloried, but so long as they identify their purpose as looking for possible effects that could then be followed up definitively, it is perfectly respectable.

24.3.4 Look for patterns of significant results

One thing to check, especially when assessing published data that includes multiple testing, is whether any apparently significant results form meaningful patterns.

Say an author has reported a programme of work that resulted in 30 statistical analyses and has taken no steps to account for multiple testing. We would expect 30 analyses to produce one or two apparently significant results even if there are no real effects present.

If the result is 28 non-significant tests and two marginally 'significant' ones, then it's a nap that these are false positives, meaning absolutely nothing. On the other hand, if there are 25 significant results, it would be churlish to reject all the author's conclusions just because there was multiple testing. It is highly unlikely that such a consistent pattern of significant results would arise by chance and it would have to be accepted that the subject area under investigation must contain real effects. There would still be a danger that an odd false positive might have crept in, and if any particular conclusion was especially critical, it would be useful to check whether it would have survived a Bonferroni correction. If it wouldn't, then it should not be overly relied upon.

Beware of odd isolated 'Significant' results

If multiple tests yield a mass of non-significant results and one or two marginally significant ones, be very suspicious of any claim that the 'Significant' ones have any real meaning. If the outcome is a high proportion of significant results, this provides strong evidence, despite the multiple testing.

The other thing to look out for is whether any allegedly significant results make sense in terms of the science underlying the area of investigation. For example, a handful of statistically significant results that form no obvious pattern and which would be difficult to account for on the basis of any known chemical or biological mechanism would not be very convincing. But, if we have a similar number of significant results that all suggest the same sort of effect and if that effect could be readily explained by known mechanisms, these might be taken a lot more seriously.

24.4 The role of scientific journals

If you look through the pirate boxes scattered around this book, you may notice that, in a high proportion of cases, the fiddle is worked by seeing the data first and then chopping and changing the analysis until the desired result is obtained. In extreme cases, the process has been memorably referred to as 'The data was tortured until it confessed'.

If those who conduct clinical trials want to be taken remotely seriously by regulatory authorities, they know they will have to record all the main aspects of their proposed trial in advance.

- What question is the trial going to answer?
- What endpoint will be used?
- What statistical analysis will be performed?

As a result, most of the pirate boxes have long ceased to be a hazard in clinical trials. Certainly, the real shockers – changing to a one-sided test or selectively reporting the analysis of one subgroup – are dead and buried.

Sadly, academic science and the journals in which it is published fall far short of that standard. How many journals require prior notification of the experiment you are about to perform, analyse and (hopefully) publish? Next time you read an interesting paper, ask yourself a few questions:

- They say the purpose of the experiment was to see if the new process produced harder tablets, but was it really? Maybe they tested six different endpoints and only hardness came up as statistically significant.

- It is claimed that 'We planned to test for an increase in fertility and therefore performed a one-sided test'. Did they? Maybe they were looking for a contraceptive effect but found that there were actually more babies and the initial two-sided test produced a P value of 0.07. Was the pirate box at the end of Chapter 11 applied?

- The purpose of the experiment is stated as 'To test whether our new dispensing system increased patient compliance among Bangladeshis aged 27 to 44 with a severe form of the disease'. Really? Could this just possibly be a covert subgroup analysis?

While journals continue to function in their current manner, you may have doubts such as those above, but you'll never know. If a minority of journals took the simple precaution of requiring authors to provide prior notification of what they intended to do, the papers they published would be in a different league from the general CV fodder.

24.5 Chapter summary

Unless special precautions are used, multiple testing will increase the risk of generating false positive findings beyond the level of 5% that is normally tolerated.

- Where several treatments are compared against each other, the first step in analysis should be an omnibus test (such as an ANOVA) and if this proves significant, more detailed analyses can then be undertaken.

- If several endpoints have been determined, a multivariate statistical test may be employed and this can then be followed up by tests concerning the individual endpoints.

- Where an endpoint has been determined on several occasions, a repeated measure ANOVA can be used.

- Data should not be broken down into multiple subgroups unless either a Bonferroni correction or a distinction between primary and secondary analyses is used to provide additional protection.

The Bonferroni correction raises the standard of proof required for each individual test. This has the effect of maintaining the overall risk of any false positives at 5%, but reduces statistical power.

It is legitimate to declare in advance that the primary purpose of the work is to answer one identified question, and that one specified endpoint and statistical analysis will be used to answer that question. Limitless additional analyses can then be undertaken, so long as any conclusions are recognised as subject to the hazards of multiple testing and cannot be considered definitive.

When authors have used multiple testing, a small number of apparently significant results amid a sea of non-significance should be viewed with suspicion. A high proportion of significant results is however more convincing, especially if the significant results form a consistent and logical pattern.

Scientific journals could improve the quality (and possibly the honesty) of data analysis if they insisted on the prior declaration of authors' intentions. Most of the pirate boxes in this book could be abolished overnight.

25
Questionnaires

This chapter will ...

- Explore the range of data generated by questionnaires.
- Emphasise the need to achieve both an adequate number of completed questionnaires and an adequate rate of completion.
- Warn of the dangers of bias if there is an excessively low return rate for a questionnaire.
- Describe the two stages of analysis – Frequency analysis and hypothesis testing.
- Recall the danger of confounding and how we use logistic regression to detect it.
- Emphasise the potential hazard of multiple testing with questionnaire data.

Although questionnaires can generally be analysed using techniques that have already been described, they are associated with some distinctive problems and hazards. Three common sources of difficulty are:

- Low return rates
- Confounding
- Multiple testing.

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

25.1 Types of questions

When we come to analyse the results, we will probably be looking for various possible cause and effect relationships. So let's start by separating those bits of information that are likely to be treated as causes (from here on, I refer to these as 'Demographics') from those that are more likely to be seen as effects (referred to as 'Outcomes').

25.1.1 Causal factors – 'Demographics'

Typically it is things like age, gender, post code, employment, disease status and so on that are likely to be investigated as possible causal factors influencing outcomes recorded in other questions. It is these that I refer to as 'Demographics'.

25.1.2 Outcome data

Outcomes are aspects that may vary according to the responses to the demographic questions. These can be fairly clearly divided into three types:

- Factual
- Opinion seeking
- Knowledge testing.

25.1.2.1 Factual questions Examples include things such as:

- The pain lasted for: 0–19 min/20 min to 1 h/1–3 h/3–6 h/more than 6 h.
- Were you able to see a hospital specialist within one calendar month of first seeing your GP? Yes/No.

25.1.2.2 Opinion seeking questions These assess more subjective impressions, for example:

- Did you think the instructions were clear? Yes/No.
- In your opinion the staff were: Very friendly/Friendly/Neutral/Unfriendly/Very unfriendly.

Questionnaires containing opinion seeking questions are one of the commonest sources of ordered categorical data. The last example given above would probably be recorded as a score, ranging from (say) 1 (Very unfriendly) to 5 (Very friendly).

Some investigators like to have an even number of options to avoid a neutral response. If you wanted to force respondents to come off the fence, you could offer the range suggested above, but omitting the option 'Neutral'. Then they will have to jump one way or the other – positive or negative.

25.1.3 Knowledge testing questions

Here we are looking to determine how well informed the subject is. We might target either the public (Is an educational campaign needed?) or health professionals (How effective was the training?). For example:

- What is the normal oral dose of paracetamol for an adult? Choose just one of the following: 10 mg/100 mg/200 mg/1 g/5 g.
- Which one of the following should not be taken with combined oral contraceptives? Paracetamol/Rifampicin/Vitamin D/Lithium/SSRIs.

There may be odd occasions when we only need to find out whether the subject knows a single piece of information and a single question will suffice, but generally we have clusters of such questions to allow a broader test of knowledge.

25.1.4 Closed versus open questions

The form in which questions are posed may be closed or open.

Closed questions Here the subject is asked to choose from a limited list of possibilities, for example:

> 'What form of Hormone Replacement Therapy are you presently receiving? Please indicate just one: None/Tablets/Patches/Gel/Implant/Nasal spray/Vaginal ring.'

You'll receive a maximum of seven different responses and you know exactly what they will be. Two things need to be ensured:

1. All possibilities are covered.
2. It is clear whether the respondent should select just one option or is free to tick several if appropriate.

Open questions These allow free text response, for example:

> 'Why did you come to the Accident and Emergency department?'

Results from closed questions are relatively easy to analyse because the responses are predictable and you can simply count how many choose each option. In contrast, analysing responses to open questions is trickier. You will need to read through all the responses looking for common themes, that is clusters of answers that say essentially the same thing even if the exact wording is different. Where a theme is mentioned reasonably frequently, it may be worth creating a new variable that records whether or not a particular respondent mentioned the relevant point. This is then recorded as Yes/No. You can then look for relationships with demographics in the usual way, for example: Are males or females more likely to mention the relevant aspect? Where only one or two respondents mention a particular idea it will not be worth recording that information in any statistical package as statistical analyses will be inappropriate, but nonetheless it may occasionally be worth reporting something such as 'Two subjects made the interesting observation that ….' Such analyses are time-consuming and it is perilously easily to slip into a degree of subjectivity.

Despite the difficulties, open questions can be valuable:

- They allow respondents to mention aspects that you failed to list in a closed question.
- Open questions can avoid excessive prompting.

Closed and open questions

A mix of closed and open questions will assure access to some data that can be easily and objectively analysed, but will also offer the respondent the opportunity to record facts/opinions that you may have accidentally excluded.

25.2 Sample sizes and low return rates

25.2.1 Calculating an appropriate sample size

As with any form of sample based research, it's important to know in advance how much data (completed questionnaires) will be required. We can calculate a necessary sample size using the approaches described earlier in this book. We just need to identify: (i) the primary hypothesis that the questionnaire is intended to test, (ii) the key question(s) that will determine the outcome and (iii) what statistical analysis will be applied. We can then perform a sample size calculation in the normal way.

25.2.2 Problem number one: Low response rates and self-selected minorities

However, there is still one very large hazard waiting to get you. The danger is that you calculate 100 completed questionnaires are needed, post out 500 and get 110 returned. You've got back more than enough, so what's the problem?

The 110 out of 500 (22%) who did return the questionnaire are a self-selected minority. We already know that they differ in one way – they did return the questionnaire, while the other 78% didn't. But is this decision to complete the form just random or is it linked to other personal characteristics? It would be a brave (foolhardy?) researcher who was prepared to assume that returning or not returning the questionnaire was a purely random choice. The reality is that there are all sorts of ways in which the returners may differ from the non-returners. With a return rate as low as 22%, we could be dealing with a highly unrepresentative minority and any conclusions will be biased towards whatever type of person happens to be most likely to return our questionnaire.

One obvious likelihood is that those with a strong interest in, or opinion concerning, the relevant subject will be more likely to respond.

You need adequate numbers and an adequate return rate

It s not enough simply to get an adequate ***number*** of completed questionnaires, it is also vital to obtain a satisfactory return ***rate***.

It's impossible to lay down the law about what constitutes an adequate return rate. You would have to consider the potential for bias associated with the particular piece of research. However, once return rates get below about 40%, the onus really is on the researcher to convince his/her readers that the sample isn't at risk of bias.

Section 25.3.1.2 will suggest that part of our analysis should routinely be a simple frequency analysis of the demographic data. This may reveal that our sample is biased in terms of gender, age groups or whatever. However, with very low return rates, it is impossible to exclude the possibility that our sample may be biased in some way that will not be detected by the demographics we are able to check.

25.3 Analysing the results

Generally the results will be analysed in two stages – frequency analyses and hypothesis testing as described below.

25.3.1 Frequency analyses

This initial stage is a simple reporting of the results for each question in isolation. Responses in categorical (nominal) or ordinal form will be reported as frequencies for each category and may be represented by bar charts and so on. Responses in the form of continuous measurements such as age could be reported as a mean and standard deviation but will more likely be broken up into bands and the number in each band reported.

These frequencies serve two possible purposes – they are likely to be of intrinsic interest and they act as checks on your demographics.

25.3.1.1 Intrinsic interest Less experienced researchers are surprisingly resistant to the idea that the primary conclusion of a survey can be based upon a simple frequency analysis. Determining the proportion of patients who would prefer to have their drug treatment monitored in a local health centre rather than a hospital or the proportion of people who have various levels of satisfaction with a new service (etc.) can be perfectly respectable outcomes. The problem is that there are no fancy statistical tests and there can be a sense that something more 'sophisticated' would be of greater merit.

It doesn't have to be complicated

A simple frequency analysis may be entirely appropriate as the primary outcome of a survey.

25.3.1.2 Checking demographics The other key use of frequency analyses is to check that the demographics of a sample are reasonable. If your survey data is supposed to represent all patients being treated at a local health centre, but your frequency analyses reveal that 90% of the responses are from women or there is a similarly disproportionate response from younger people or a particular ethnic group, there are going to be serious questions as to whether your data is credible as a supposedly random sample from the population of service users. If the health centre serves a large new estate full of young families, then you might conclude that a high proportion of younger patients was perfectly reasonable. However, in the absence of such explanations, you may face accusations that you have been selective in questioning certain groups and shunning others.

25.3.2 Hypothesis testing

This is where we start checking whether the answers to one question are linked to those for another. Examples include:

- Do opinions vary between those who are young/old, male/female and so on?
- Are knowledge levels greater among those who have received a leaflet than among those who have not?

We will be asking whether any of the outcomes vary according to the demographic or treatment group to which the respondent belongs. Apart from the problem of

confounding (see next section), there are no special considerations so far as performing statistical tests is concerned. We just apply the same tests outlined in Chapters 7 to 23. With questionnaires, many outcomes are recorded as simple categorisations – nominal scale data. Consequently questionnaires tend to generate a lot of contingency chi-square tests. Opinions are likely to be recorded on ordinal scales, so questionnaires are an area where non-parametric methods (Chapter 21) are likely to be useful. Interval scale (measured) endpoints aren't all that common and parametric tests (*t*-tests, etc.) feature less prominently in questionnaire analysis than in other areas. Scores arising from blocks of knowledge testing questions might in principal, be analysed as interval scale data, but they have a nasty habit of forming non-normal distributions, again necessitating a shift to non-parametric alternative tests.

25.4 Problem number two: Confounded questionnaire data

The problem of confounding was dealt with in detail in Section 20.2.2. A quick summary is that if two demographic features have a degree of correlation, this can lead to false impressions of relationships between one of the factors and one (or more) of the outcomes. The example considered in Chapter 20 was one where male pharmacists tended to be older than female ones and consequently if age had the genuine effect of causing older pharmacists to hold a particular opinion, then that opinion would also automatically tend to be commoner among male pharmacists. The important point is that male pharmacists might hold the opinion, but that may arise simply because they are generally older, not because they are male.

Most questionnaires contain quite large numbers of questions. If we record each respondent's age, gender, ethnicity, educational level, profession, area of residence and so on, it would be very surprising if these were all uncorrelated. If one of these factors is genuinely associated with a particular outcome, then any other factor that correlates with the original factor is likely to be falsely detected as being related to the outcome.

In the author's experience, questionnaire-based, undergraduate projects frequently identify numerous factors as being statistically significantly associated with a particular outcome, but the possibility that some of these 'relationships' are mere confounding is not always addressed. Section 20.2 considers in detail how logistic regression can be used to build a model that simultaneously considers all the apparently significant factors and will identify any that are actually confounded.

25.5 Problem number three: Multiple testing with questionnaire data

The general problem of multiple testing was covered in the previous chapter. Questionnaires are unfortunately the multiple-tester's happy hunting ground. A questionnaire containing four demographic questions plus three factual, opinion or knowledge testing questions would be a pretty barebones effort (most are far longer).

However, with even this minimal example, if we were to take each demographic factor and look to see whether it influenced each of the outcome questions, we would be examining 12 combinations of questions. The chances of obtaining at least one false positive result would exceed 50%. The reality is that most questionnaires present considerably greater temptation than this.

There are a couple of ways to reduce the problem.

25.5.1 For opinion seeking or knowledge testing questions consider basing your primary analysis on an overall knowledge score

If a group of questions all test the same topic, consider combining the results from all the questions to produce an overall score. You can then test whether key demographic factors affect this single result rather than investigating each test question separately. A good example is clusters of knowledge testing questions where we can combine these into a single knowledge score.

Combine a cluster of knowledge testing questions into a single score

Rather than commit multiple testing by investigating each knowledge testing question separately, carry out a single test on the overall test score.

Apart from knowledge testing, there are other areas where the results from groups of questions may usefully be aggregated into a single score. Level of satisfaction with a mode of treatment is an obvious case. When combining the results of several questions, it may be appropriate to use Chronbach's Alpha to determine the level of agreement among those questions, but read Section 22.1.2 carefully; the appropriateness of using this statistic needs to be decided on a case by case basis.

25.5.2 Identify your primary question in advance

The best defence against multiplicity is to identify, in advance, which potential relationships will be tested for, as primary analyses (as described in the previous chapter). A typical example would be where we have produced a knowledge score and as a primary analysis, found that two demographic groups scored differently. This would produce a reliable answer to our main question – Does group A know more than group B?

As a follow up to this, you might start to ask more detailed questions. For example, did the differences arise because one group was better at answering all the questions

or was there just one killer question that one group could answer but the other couldn't? You could tackle this in a secondary analysis where you would be free to look at individual questions and do as much multiple testing as you wish. The caveat is of course that any conclusions must also be recognised as secondary and not to be unduly relied upon (unless confirmed in subsequent work).

25.6 Chapter summary

The data collected via questionnaires largely falls into two categories – 'Demographics' and 'Outcomes'. Demographic data describes the characteristics of respondents that will commonly be tested for possible causal relationships with the various outcomes.

Outcome data is commonly classified as 'Factual', 'Opinion seeking' or 'Knowledge testing'.

'Closed' questions have the advantage of bringing only predictable answers that are easily analysed, while 'Open' questions allow respondents to report facts/opinions that might have been excluded by a closed question and they also avoid excessive prompting.

An appropriate sample size – Number of completed questionnaires – should be calculated (as described in earlier chapters).

The first stage in analysing the results is generally a simple frequency analysis – a count of how many respondents produced each possible response, on a question by question basis. This information is likely to be of value in its own right and, when applied to your demographic data, may also alert you to any biases in your sample.

Analysis will then probably progress to hypothesis testing – looking to see whether the answer provided to one question influences the pattern of answers to another question. As so much questionnaire data is nominal or ordinal (e.g. Likert scales), contingency chi-square and non-parametric tests tend to dominate.

There are three potential errors that are particularly easy to commit during questionnaire based research.

Low return rate: If only a small proportion return the questionnaire you are in danger of getting a biased return. It is not satisfactory to attempt to compensate for a low return rate, by simply increasing the numbers sent out. You need an adequate number of returns and an adequate return rate.

Confounding: When interpreting your conclusions, it is vital to bear in mind the risk of confounding. If more than one factor appears to be statistically significantly associated with a particular outcome, logistic regression should be used to exclude confounding.

Multiple testing: Endless pairs of questions could be tested for correlation and questionnaires are a heaven-sent opportunity for multiple testing. Because of this, it is particularly important that the primary analysis is planned in advance. Any amount of secondary analysis can ultimately be performed, so long as it is identified as such. It may also be possible to reduce the number of tests carried out by aggregating the results of several questions into a single combined score.

Index

Essential Statistics for the Pharmaceutical Sciences, Second Edition. Philip Rowe.

© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.