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.

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.

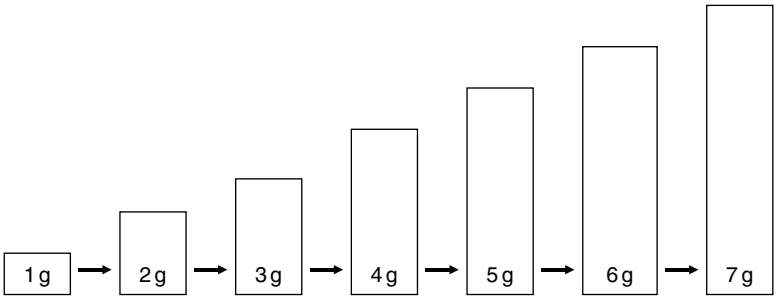

Figure 1.1 Interval scale data – a series of weights (1–7 g)

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

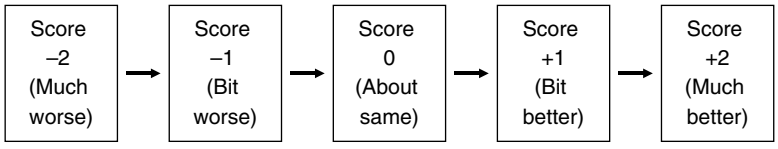

Figure 1.2 Ordinal scale data – scores for patient responses to treatment

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.