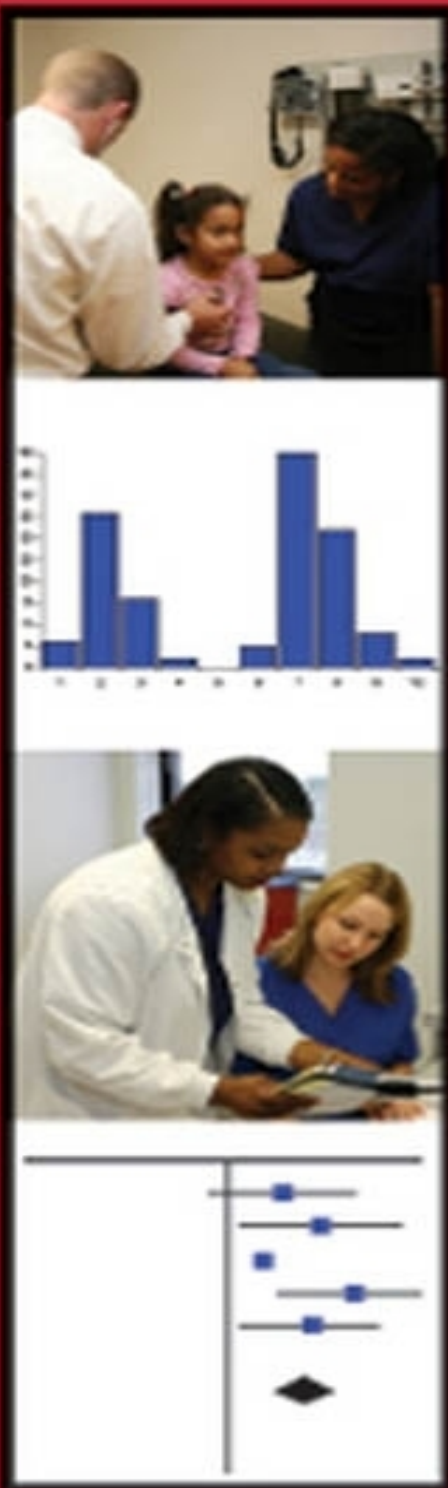


Statistics Toolkit

Rafael Perera
Carl Heneghan
Douglas Badenoch



Contents

Introduction

Data: describing and displaying

Summarizing your data

The arithmetic mean (numeric data)

The geometric mean

The weighted mean

The median and mode

Measures of dispersion: the range

The variance

The standard deviation

Percentiles

Standard error of the mean

Displaying data

Probability and confidence intervals

Probability distributions

Confidence intervals

Hypothesis testing

The steps in testing a hypothesis

Errors in hypothesis testing

Statistical power

Randomised controlled trials: mean, median, mode, RD, NNT, Mann-Whitney and log rank tests

The mean and the median

The risk difference (RD) and the number needed to treat (NNT)

The Mann-Whitney test

Log-rank test and Kaplan-Meier survival statistics

Systematic reviews 1: odds, odds ratio, heterogeneity, funnel plots

Odds ratios

Combining odds ratios

Heterogeneity tests

Visualising statistical significance

Subgroup analysis

Testing for publication bias

Other terminology used

Case-control studies: odds ratios, logistic regression

Odds ratios in case-control studies

Logistic regression and adjustment for crude odds ratios

Questionnaire studies I: weighted mean frequency, nonparametric tests

Weighted mean frequency

Nonparametric bootstrapping was done to calculate 95% CIs

The median time of onset and interquartile range

Questionnaire studies 2: inter-rater agreement

Likert scales

Assessing inter-rater agreement using kappa

Cohort studies: prevalence, incidence, risk, survival, mortality rate, association, prognostic factor

Using mortality or survival rates

Regression: are the outcomes associated with prognostic factors?

Interpreting multivariable analysis

Systematic reviews 2: Cohort study odds ratios and relative risk

Hazard ratio

Cox proportional hazards model

Diagnostic tests: sensitivity, specificity, likelihood ratios, ROC

curves, pre- and post-test odds, pre- and post-test probability

How well does the test detect the disease?

How likely is a given test result to be true?

Likelihood ratios (LR)

What about tests that have a range of possible results?

Measuring test performance at different cut-off points

Statistical significance

Pre-test odds and post-test odds: particularising the evidence

Using Bayes theorem and the Fagan nomogram

Scale validation: correlation

Visual analogue scale and CARIFS

Correlation coefficient (Pearson and Spearman)

Factor analysis and Cronbach's alpha

Statistical toolkit: Glossary

Software for data management and statistical analysis

References

Index

Statistics Toolkit

Rafael Perera

*Centre for Evidence-based Medicine
Department of Primary Health Care
University of Oxford
Old Road Campus
Headington
Oxford OX3 7LF*

Carl Heneghan

*Centre for Evidence-based Medicine
Department of Primary Health Care
University of Oxford
Old Road Campus
Headington
Oxford OX3 7LF*

Douglas Badenoch

*Minervation Ltd
Salter's Boat Yard
Folly Bridge
Abingdon Road
Oxford OX1 4LB*



BMJ | Books

© 2008 Rafael Perera, Carl Heneghan and Douglas
Badenoch

Published by Blackwell Publishing

BMJ Books is an imprint of the BMJ Publishing Group Limited,
used under licence

Blackwell Publishing, Inc., 350 Main Street, Malden,
Massachusetts 02148-5020, USA

Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4
2DQ, UK

Blackwell Publishing Asia Pty Ltd, 550 Swanston Street,
Carlton, Victoria 3053, Australia

The right of the Author to be identified as the Author of this
Work has been asserted in accordance with the 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.

First published 2008

1 2008

ISBN: 978-1-4051-6142-8

A catalogue record for this title is available from the British
Library and the Library of Congress.

Set in Helvetica Medium 7.75/9.75 by Sparks, Oxford -
www.sparks.co.uk Printed and bound in Singapore by
Markono Print Media Pte Ltd

Commissioning Editor: Mary Banks

Development Editors: Lauren Brindley and Victoria Pittman

Production Controller: Rachel Edwards

For further information on Blackwell Publishing, visit our website: <http://www.blackwellpublishing.com>

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

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.

This handbook was compiled by Rafael Perera, Carl Heneghan and Douglas Badenoch. We would like to thank all those people who have had input to our work over the years, particularly Paul Glasziou and Olive Goddard from the Centre of Evidence-Based Medicine. In addition, we thank the people we work with from the Department of Primary Health Care, University of Oxford, whose work we have used to illustrate the statistical principles in this book. We would also like to thank Lara and Katie for their drawings.

Introduction

This 'toolkit' is the second in our series and is aimed as a summary of the key concepts needed to get started with statistics in healthcare.



Often, people find statistical concepts hard to understand and apply. If this rings true with you, this book should allow you to start using such concepts with confidence for the first time. Once you have understood the principles in this book you should be at the point where you can understand and interpret statistics, and start to deploy them effectively in your own research projects.

The book is laid out in three main sections: the first deals with the basic nuts and bolts of describing, displaying and handling your data, considering which test to use and testing for statistical significance. The second section shows how statistics is used in a range of scientific papers. The final section contains the glossary, a key to the symbols used in statistics and a discussion of the software tools that can make your life using statistics easier.

Occasionally you will see the GO icon on the right. This means the difficult concept being discussed is beyond the scope of this textbook. If you need more information on this point you can either refer to the text cited or discuss the problem with a statistician.

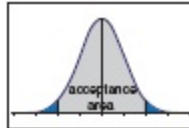


Essentials you
need to get
started

Types of data: categorical,
numerical, nominal, ordinal, etc.
Describing your data ([see p. 3](#))

The null (H_0)
and alternative
(H_1) hypothesis

For H_0 and H_1
([see p. 20](#))



What type of
test?

Choosing the right type of test to use
will prove the most difficult concept
to grasp. The book is set out with
numerous examples of which test to
use; for a quick reference refer to
the flow chart on [page 26](#)

Is it
significant?

Once you have chosen the test,
compute the value of the statistic
and then learn to compare the value
of this statistic with the critical value
([see p. 20](#))

Software tools

Start with Excel (learn how to
produce simple graphs) and then
move on to SPSS and when your
skills are improving consider STATA,
SAS or R ([see p. 103](#))

Data: describing and displaying

The type of data we collect determines the methods we use. When we conduct research, data usually comes in two forms:

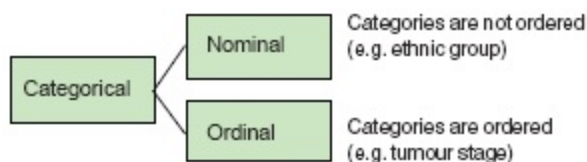
- Categorical data, which give us percentages or proportions (e.g. '60% of patients suffered a relapse').
- Numerical data, which give us averages or means (e.g. 'the average age of participants was 57 years').

So, the type of data we record influences what we can say, and how we work it out. This section looks at the different types of data collected and what they mean.

Any measurable factor, characteristic or attribute is a **variable**

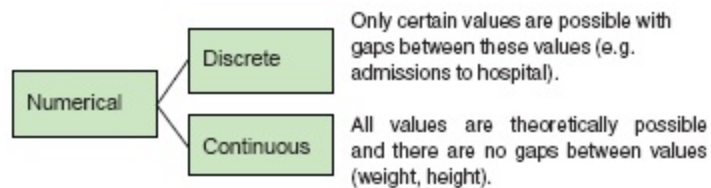
A variable from our data can be two types: categorical or numerical.

Categorical: the variables studied are grouped into categories based on qualitative traits of the data. Thus the data are labelled or sorted into categories.



A special kind of categorical variables are **binary** or **dichotomous** variables: a variable with only two possible values (zero and one) or categories (yes or no, present or absent, etc.; e.g. death, occurrence of myocardial infarction, whether or not symptoms have improved).

Numerical: the variables studied take some numerical value based on quantitative traits of the data. Thus the data are sets of numbers.



You can consider discrete as basically counts and continuous as measurements of your data.

Censored data - sometimes we come across data that can only be measured for certain values: for instance, troponin levels in myocardial infarction may only be detected for a certain level and below a fixed upper limit (0.2-180 µg/L)

Summarizing your data

It's impossible to look at all the raw data and instantly understand it. If you're going to interpret what your data are telling you, and communicate it to others, you will need to summarize your data in a meaningful way. Typical mathematical summaries include percentages, risks and the mean.

The benefit of mathematical summaries is that they can convey information with just a few numbers; these summaries are known as **descriptive statistics**.

Summaries that capture the average are known as measures of **central tendency**, whereas summaries that indicate the spread of the data usually around the average are known as measures of **dispersion**.

The arithmetic mean (numeric data)

The arithmetic mean is the sum of the data divided by the number of measurements. It is the most common measure of central tendency and represents the average value in a sample.

$$\bar{x} = \frac{\sum x_i}{n}$$

Consider the following test scores:

Test scores out of ten

6	4	2. Divided by the number of measurements
4	7	
5	2	(6+4+5+6+7+4+7+2+9+7) / 10 = 5.7
6	9	
7	7	1. The sum of the measurements 3. Gives you the mean

$\bar{x}$ = sample mean
 μ = population mean
 Σ = the sum of
 x = variable
 i = the total variables
 n = number of measurements

To calculate the mean, add up all the measurements in a group and then divide by the total number of measurements.

The geometric mean

If the data we have sampled are skewed to the right (see p. 7) then we transform the data using a natural logarithm (base $e = 2.72$) of each value in the sample. The arithmetic mean of these transformed values provides a more stable measure of location because the influence of extreme values is smaller. To obtain the average in the same units as the original data – called the geometric mean – we need to back transform the arithmetic mean of the transformed data:

$$\text{geometric mean original values} = e^{(\text{arithmetic mean } \ln(\text{original values}))}$$

The weighted mean

The weighted mean is used when certain values are more important than others: they supply more information. If all weights are equal then the weighted mean is the same as the arithmetic mean (see p. 54 for more).

We attach a weight (w_j) to each of our observations (x_j):

$$\frac{w_1x_1 + w_2x_2 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum w_jx_j}{\sum w_j}$$

The median and mode

The easiest way to find the median and the mode is to sort each score in order, from the smallest to the largest:

Test scores out of ten		
1) 2	6) 6	In a set of ten scores take the fifth and sixth values ↓ $(6+6) / 2 = 6$ ↑ The median is equal to the mean of the two middle values or to the middle value when the sample size is an odd number
2) 4	7) 7	
3) 4	8) 7	
4) 5	9) 7	
5) 6	10) 9	

The **median** is the value at the midpoint, such that half the values are smaller than the median and half are greater than the median. The **mode** is the value that appears most frequently in the group. For these test scores the mode is 7. If all values occur with the same frequency then there is no mode. If more than one value occurs with the highest frequency then each of these values is the mode. Data with two modes are known as **bimodal**.

Choosing which one to use: (arithmetic) mean, median or mode?

The following graph shows the mean, median and mode of the test scores. The x-axis shows the scores out of ten. The height of each bar (y-axis) shows the number of participants who achieved that score.