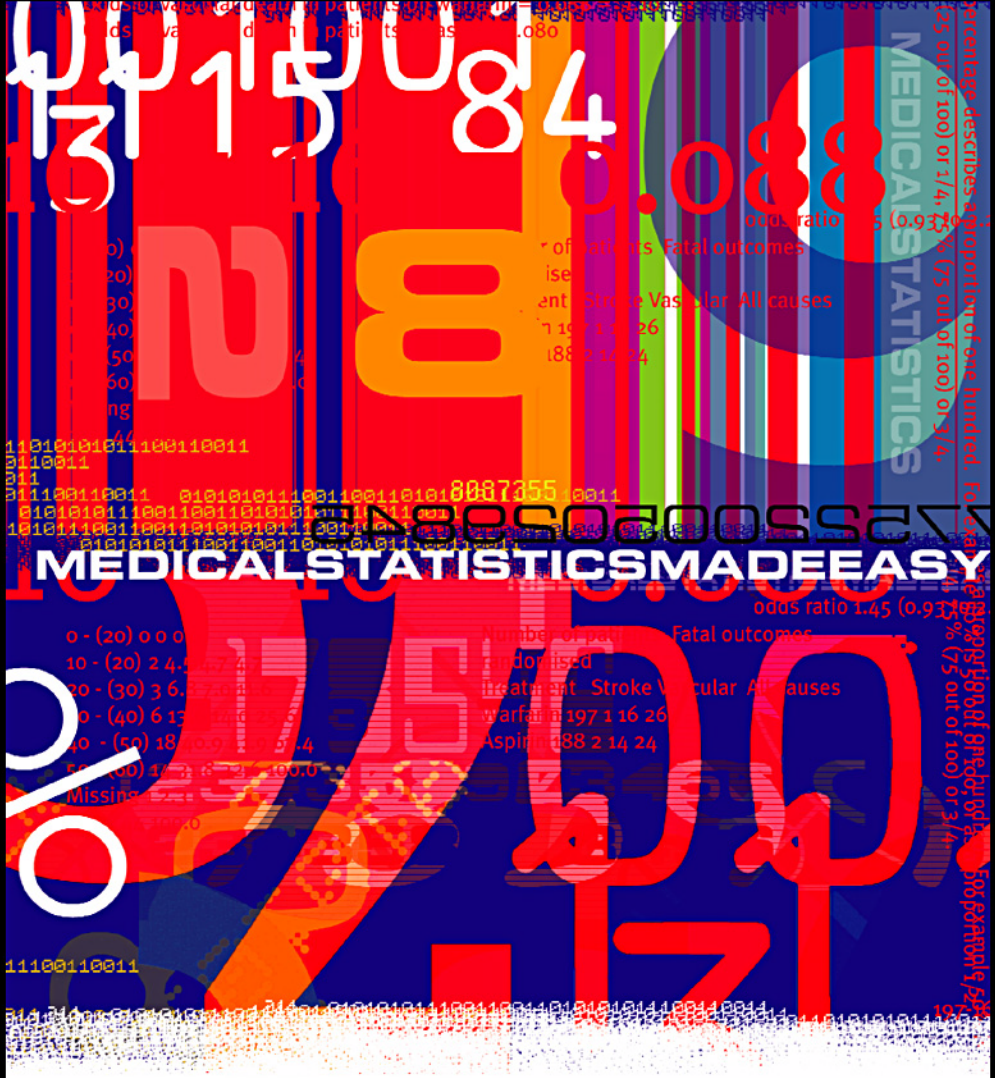

MEDICAL STATISTICS MADE EASY

M. Harris and G. Taylor

MD Martin Dunitz
Taylor & Francis Group

Also available as a printed book
see title verso for ISBN details

MEDICAL STATISTICS MADE EASY

MEDICAL STATISTICS MADE EASY

Michael Harris
General Practitioner and Lecturer in General Practice, Bath, UK

and

Gordon Taylor
Senior Research Fellow in Medical Statistics, University of Bath, Bath, UK

Martin Dunitz
Taylor & Francis Group
LONDON AND NEW YORK

© 2003 Martin Dunitz, an imprint of the Taylor & Francis Group

First published in the United Kingdom 2003
by Martin Dunitz, an imprint of the Taylor and Francis Group,
11 New Fetter Lane, London EC4P 4EE
Tel: +44 (0) 20 7583 9855; Fax: +44 (0) 20 7842 2298;
Email: info@dunitz.co.uk; website: http://www.dunitz.co.uk

This edition published in the Taylor & Francis e-Library, 2004.

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, without the prior permission of the publisher or in accordance with the provisions of the Copyright, Designs and Patents Act 1988 or under the terms of any licence permitting limited copying issued by the Copyright Licensing Agency, 90 Tottenham Court Road, London W1P 0LP.

Although every effort has been made to ensure that all owners of copyright material have been acknowledged in this publication, we would be glad to acknowledge in subsequent reprints or editions any omissions brought to our attention.

Although every effort has been made to ensure that drug doses and other information are presented accurately in this publication, the ultimate responsibility rests with the prescribing physician. Neither the publishers nor the authors can be held responsible for errors or for any consequences arising from the use of information contained herein. For detailed prescribing information or instructions on the use of any product or procedure discussed herein, please consult the prescribing information or instructional material issued by the manufacturer.

ISBN 0-203-50277-9 Master e-book ISBN

ISBN 0-203-59739-7 (Adobe eReader Format)
ISBN 1 85996 219 x (Print Edition)

Published in the United States of America,
its dependent territories and Canada
by Springer-Verlag New York Inc.,
175 Fifth Avenue,
New York, NY 10010-7858,
in association with Martin Dunitz,
an imprint of the Taylor and Francis Group.

Distributed in the rest of the world by
Thomson Publishing Services
Cheriton House
North Way
Andover, Hampshire SP10 5BE, UK
Tel.: +44 (0)1264 332424; E-mail:
salesorder.tandf@thomsonpublishingservices.co.uk

Production Editor: Andrea Bosher

CONTENTS

Abbreviations vii
Preface ix
About the authors x
Foreword xi
How to use this book 1
How this book is designed 4

Statistics which describe data

Percentages 7
Mean 9
Median 12
Mode 14
Standard deviation 16

Statistics which test confidence

Confidence intervals 20
P values 24

Statistics which test differences

t tests and other parametric tests 28
Mann-Whitney and other non-parametric tests 31
Chi-squared 34

Statistics which compare risk

Risk ratio 37
Odds ratio 40
Risk reduction and numbers needed to treat 43

Statistics which analyze relationships

Correlation 48
Regression 53

Statistics which analyze survival

Survival analysis: life tables and Kaplan–Meier plots 57
The Cox regression model 60

Statistics which analyze clinical investigations and screening

Sensitivity, specificity and predictive value 62
Level of agreement and Kappa 67

Other concepts 69

Statistics at work 73

Standard deviation, relative risk, confidence intervals, chi-squared and *P* values 74
Odds ratios and confidence intervals 78
Correlation and regression 82
Survival analysis and risk reduction 86
Sensitivity, specificity and predictive values 90

Glossary 94

Index 113

ABBREVIATIONS

ARR	absolute risk reduction
BMI	body mass index
BP	blood pressure
CI	confidence interval
df	degrees of freedom
HR	hazard ratio
IQR	inter-quartile range
LR	likelihood ratio
NNH	number needed to harm
NNT	number needed to treat
NPV	negative predictive value
P	probability
PPV	positive predictive value
RRR	relative risk reduction
SD	standard deviation

PREFACE

This book is designed for healthcare students and professionals who need a basic knowledge of when common statistical terms are used and what they mean.

Whether you love or hate statistics, you need to have some understanding of the subject if you want to critically appraise a paper. To do this, you *do not* need to know how to do a statistical analysis. What you *do* need is to know whether the right test has been used and how to interpret the resulting figures.

This book does not assume that you have any prior statistical knowledge. However basic your mathematical or statistical knowledge, you will find that everything is clearly explained.

A few readers will find some of the sections ridiculously simplistic, others will find some bafflingly difficult. The "smiley faces" grading will help you pick out concepts that suit your level of understanding.

The "star" system is designed to help you pick out the most important concepts if you are short of time.

This book is also produced for those who may be asked about statistics in an exam. Pick out the "exam tips" sections if you are in a hurry.

You can test your understanding of what you have learnt by working through extracts from original papers in the "Statistics at work" section.

ABOUT THE AUTHORS

Dr Michael Harris MB BS FRCGP MMEd is a GP and lecturer in General Practice in Bath. He teaches nurses, medical students and GP registrars. Until recently he was an examiner for the MRCGP.

Dr Gordon Taylor PhD MSc BSc (Hons) is a senior research fellow in medical statistics at the University of Bath. His main role is in the teaching, support and supervision of health care professionals involved in non-commercial research.

FOREWORD

A love of statistics is, oddly, not what attracts most young people to a career in medicine and I suspect that many clinicians, like me, have at best a sketchy and incomplete understanding of this difficult subject.

Delivering modern, high quality care to patients now relies increasingly on routine reference to scientific papers and journals, rather than traditional textbook learning. Acquiring the skills to appraise medical research papers is a daunting task. Realizing this, Michael Harris and Gordon Taylor have expertly constructed a practical guide for the busy clinician. One a practising NHS doctor, the other a medical statistician with tremendous experience in clinical research, they have produced a unique handbook. It is short, readable and useful, without becoming overly bogged down in the mathematical detail that frankly puts so many of us off the subject.

I commend this book to all healthcare professionals, general practitioners and hospital specialists. It covers all the ground necessary to critically evaluate the statistical elements of medical research papers, in a friendly and approachable way. The scoring of each brief chapter in terms of usefulness and ease of comprehension will efficiently guide the busy practitioner through his or her reading. In particular it is almost unique in covering this part of the syllabus for royal college and other postgraduate examinations. Certainly a candidate familiar with the contents of this short book and taking note of its

numerous helpful examination tips should have few difficulties when answering the questions on statistics in both the MCQ and Written modules of the current MRCGP exam.

Bill Irish
January 2003.

(BSc MB BChir DCH DRCOG MMEd FRCGP, General Practice Trainer and Examiner for the MRCGP(UK)).

HOW TO USE THIS BOOK

You can use this book in a number of ways.

If you want a statistics course

- Work through from start to finish for a complete course in commonly used medical statistics.

If you are in a hurry

- Choose the sections with the most stars to learn about the commonest statistical methods and terms.

- You may wish to start with these 5-star sections: percentages (page 7), mean (page 9), standard deviation (page 16), confidence intervals (page 20) and *P* values (page 24).

If you are daunted by statistics

- If you are bewildered every time someone tries to explain a statistical method, then pick out the sections with the most smiley faces to find the easiest and most basic concepts.

- You may want to start with percentages (page 7), mean (page 9), median (page 12) and mode (page 14), then move on to risk ratio (page 37), incidence and prevalence (page 70).

If you are taking an exam

- The "Exam Tips" give you pointers to the topics which examiners like to ask about.
- You will find these in the following sections: mean (page 9), standard deviation (page 16), confidence intervals (page 20), *P* values (page 24), risk reduction and NNT (page 43), sensitivity, specificity and predictive value (page 62), incidence and prevalence (page 70).

Test your understanding

- See how statistical methods are used in five extracts from real-life papers in the "Statistics at work" section (page 73).
- Work out which statistical methods have been used, why, and what the results mean. Then check your understanding in our commentary

Glossary

- Use the glossary (page 94) as a quick reference for statistical words or phrases that you do not know.

Study advice

- Go through difficult sections when you are fresh and try not to cover too much at once.
- You may need to read some sections a couple of times before the meaning sinks in. You will find that the examples help you to understand the principles.

- We have tried to cut down the jargon as much as possible. If there is a word that you do not understand, check it out in the glossary.

HOW THIS BOOK IS DESIGNED

Every section uses the same series of headings to help you understand the concepts.

"How important is it?"

We noted how often statistical terms were used in 250 randomly selected papers in mainstream medical journals. All the papers selected were published during the last year in the *British Medical Journal*, the *Lancet*, the *New England Journal of Medicine* and the *Journal of the American Medical Association*.

We grouped the terms into concepts and graded them by how often they were used. This helped us to develop a star system for importance. We also took into account usefulness to readers. For example, "numbers needed to treat" are not often quoted but are fairly easy to calculate and useful in making treatment decisions.

★★★★★ Concepts which are used in the majority of medical papers.

★★★★ Important concepts which are used in at least a third of papers.

★★★ Less frequently used, but still of value in decision-making.

★★ Found in at least 1 in 10 papers.

★ Rarely used in medical journals.

How easy is it to understand?

We have found that the ability of health care professionals to understand statistical concepts varies more widely than their ability to understand anything else related to medicine. This ranges from those that have no difficulty learning how to understand regression to those that struggle with percentages.

One of the authors (not the statistician!) fell into the latter category. He graded each section by how easy it is to understand the concept.

☺☺☺☺☺ Even the most statistic-phobic will have little difficulty in understanding these sections.

☺☺☺☺ With a little concentration, most readers should be able to follow these concepts.

☺☺☺ Some readers will have difficulty following these. You may need to go over these sections a few times to be able to take them in.

☺☺ Quite difficult to understand. Only tackle these sections when you are fresh.

☺ Statistical concepts that are very difficult to grasp.

When is it used?

One thing you need to do if critically appraising a paper is check that the right statistical technique has been used. This part explains which statistical method should be used for what scenario.

What does it mean?

This explains the bottom line – what the results mean and what to look out for to help you interpret them.

Examples

Sometimes the best way to understand a statistical technique is to work through an example. Simple, fictitious examples are given to illustrate the principles and how to interpret them.

Watch out for . . .

This includes more detailed explanation, tips and common pitfalls.

Exam tips

Some topics are particularly popular with examiners because they test understanding and involve simple calculations. We have given tips on how to approach these concepts.

PERCENTAGES

How important are they?

★★★★★ An understanding of percentages is probably the first and most important concept to understand in statistics!

How easy are they to understand?

☺☺☺☺☺ Percentages are easy to understand.

When are they used?

Percentages are mainly used in the tabulation of data in order to give the reader a scale on which to assess or compare the data.

What do they mean?

"Per cent" means per hundred, so a percentage describes a proportion of 100. For example 50% is 50 out of 100, or as a fraction ½. Other common percentages are 25% (25 out of 100 or ¼), 75% (75 out of 100 or ¾).

To calculate a percentage, divide the number of items or patients in the category by the total number in the group and multiply by 100.

EXAMPLE

Data were collected on 80 patients referred for heart transplantation. The researcher wanted to compare their ages. The data for age were put in "decade bands" and are shown in *Table 1*.

Table 1. Ages of 80 patients referred for heart transplantation

Years[a]	Frequency[b]	Percentage[c]
0–9	2	2.5
10–19	5	6.25
20–29	6	7.5
30–39	14	17.5
40–49	21	26.25
50–59	20	25
≥ 60	12	15
Total	80	100

[a] Years = decade bands;

[b] Frequency = number of patients referred;

[c] Percentage = percentage of patients in each decade band. For example, in the 30–39 age band there were 14 patients and we know the ages of 80 patients, so $\frac{14}{80} \times 100 = 17.5\%$.

Watch out for . . .

Authors can use percentages to hide the true size of the data. To say that 50% of a sample has a certain condition when there are only four people in the sample is clearly not providing the same level of information as 50% of a sample based on 400 people. So, percentages should be used as an additional help for the reader rather than replacing the actual data.

MEAN

Otherwise known as an arithmetic mean, or average.

How important is it?

★★★★★ A mean appeared in ⅔ papers surveyed, so it is important to have an understanding of how it is calculated.

How easy is it to understand?

☺☺☺☺☺ One of the simplest statistical concepts to grasp. However, in most groups that we have taught there has been at least one person who admits not knowing how to calculate the mean, so we do not apologize for including it here.

When is it used?

It is used when the spread of the data is fairly similar on each side of the mid point, for example when the data are "normally distributed".

The "normal distribution" is referred to a lot in statistics. It's the symmetrical, bell-shaped distribution of data shown in *Fig. 1*.

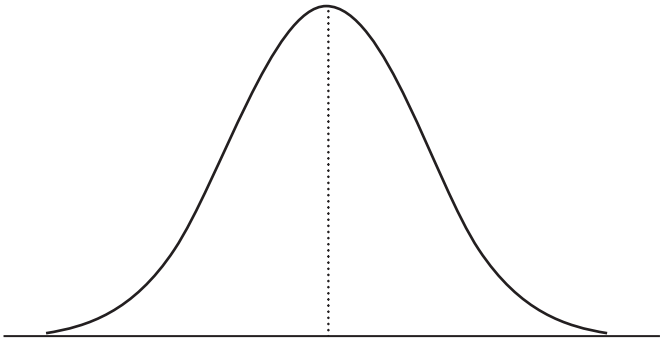

Fig. 1. The normal distribution. The dotted line shows the mean of the data.

What does it mean?

The mean is the sum of all the values, divided by the number of values.

EXAMPLE

Five women in a study on lipid-lowering agents are aged 52, 55, 56, 58 and 59 years.

Add these ages together:

$52 + 55 + 56 + 58 + 59 = 280$

Now divide by the number of women:

$$\frac{280}{5} = 56$$

So the mean age is 56 years.

Watch out for...

If a value (or a number of values) is a lot smaller or larger than the others, "skewing" the data, the mean will then not give a good picture of the typical value.

For example, if there is a sixth patient aged 92 in the study then the mean age would be 62, even though only one woman is over 60 years old. In this case, the "median" may be a more suitable mid-point to use (see page 12).

A common multiple choice question is to ask the difference between mean, median (see page 12) and mode (see page 14) – make sure that you do not get confused between them.

MEDIAN

Sometimes known as the mid-point.

How important is it?

★★★★ It is given in over a third of mainstream papers.

How easy is it to understand?

☺☺☺☺☺ Even easier than the mean!

When is it used?

It is used to represent the average when the data are not symmetrical, for instance the "skewed" distribution in *Fig. 2*.

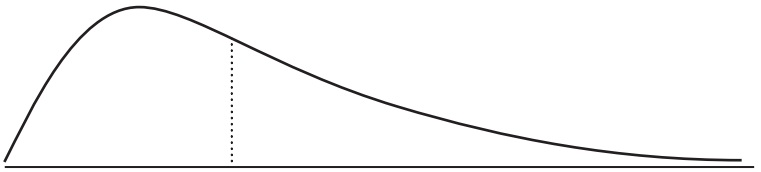

Fig. 2. A skewed distribution. The dotted line shows the median. Compare the shape of the graph with the normal distribution shown in ***Fig. 1***.

What does it mean?

It is the point which has half the values above, and half below.

EXAMPLE

Using the first example from page 10 of five patients aged 52, 55, 56, 58 and 59, the median age is 56, the same as the mean – half the women are older, half are younger.

However, in the second example with six patients aged 52, 55, 56, 58, 59 and 92 years, there are two "middle" ages, 56 and 58. The median is halfway between these, i.e. 57 years. This gives a better idea of the mid-point of this skewed data than the mean of 62.

Watch out for...

The median may be given with its inter-quartile range (IQR). The 1st quartile point has the ¼ of the data below it, the 3rd quartile has the ¾ of the sample below it, so the IQR contains the middle ½ of the sample. This can be shown in a "box and whisker" plot.

EXAMPLE

A dietician measured the energy intake over 24 hours of 50 patients on a variety of wards. One ward had two patients that were "nil by mouth". The median was 12.2 megajoules, IQR 9.9 to 13.6. The lowest intake was 0, the highest was 16.7. This distribution is represented by the box and whisker plot in *Fig. 3*.

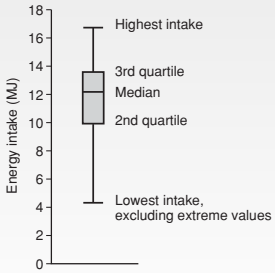

Fig. 3. Box and whisker plot of energy intake of 50 patients over 24 hours. The ends of the whiskers represent the maximum and minimum values, excluding extreme results like those of the two "nil by mouth" patients.

MODE

How important is it?

★ Rarely quoted in papers and of limited value.

How easy is it to understand?

☺☺☺☺☺ An easy concept.

When is it used?

It is used when we need a label for the most frequently occurring event.

What does it mean?

The mode is the most common of a set of events.

EXAMPLE

An eye clinic sister noted the eye-colour of 100 consecutive patients. The results are shown in *Fig. 4*.

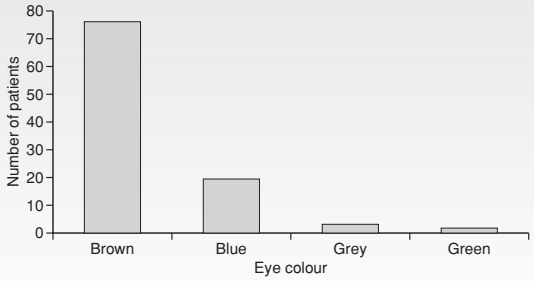

Fig. 4. Graph of eye colour of patients attending an eye clinic.

In this case the mode is brown, the commonest eye colour.

You may see reference to a "bi-modal distribution". Generally when this is mentioned in papers it is as a concept rather than from calculating the actual values, e.g. "The data appear to follow a bi-modal distribution". See *Fig. 5* for an example of where there are two "peaks" to the data, i.e. a bi-modal distribution.

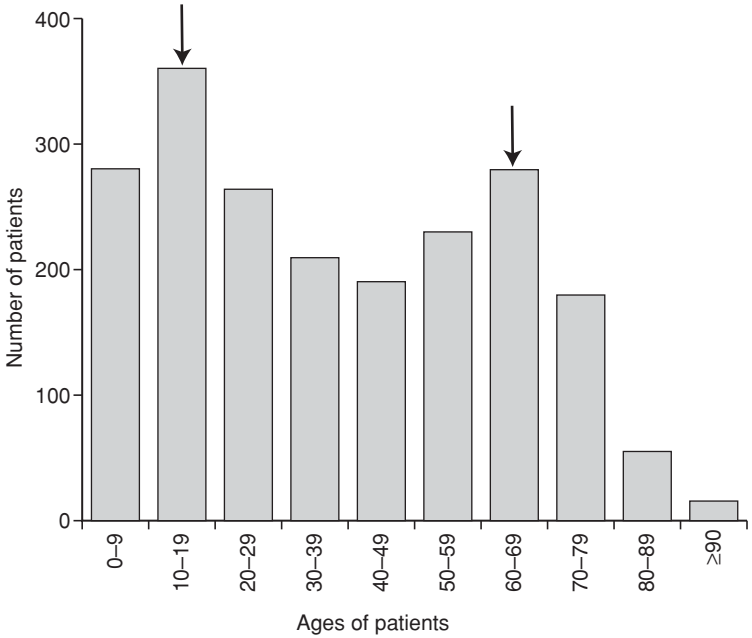

Fig. 5. Graph of ages of patients with asthma in a practice.

The arrows point to the modes at ages 10–19 and 60–69.

Bi-modal data may suggest that two populations are present that are mixed together, so an average is not a suitable measure for the distribution.

STANDARD DEVIATION

How important is it?

★★★★★ Quoted in half of papers, it is used as the basis of a number of statistical calculations.

How easy is it to understand?

☺☺☺ It is not an intuitive concept.

When is it used?

Standard deviation (SD) is used for data which are "normally distributed" (see page 9), to provide information on how much the data vary around their mean.

What does it mean?

SD indicates how much a set of values is spread around the average.

A range of one SD above and below the mean (abbreviated to ± 1 SD) includes 68.2% of the values.

±2 SD includes 95.4% of the data.

±3 SD includes 99.7%.

EXAMPLE

Let us say that a group of patients enrolling for a trial had a normal distribution for weight. The mean weight of the patients was 80 kg. For this group, the SD was calculated to be 5 kg.

1 SD below the average is 80 – 5 = 75 kg.

1 SD above the average is 80 + 5 = 85 kg.

±1 SD will include 68.2% of the subjects, so 68.2% of patients will weigh between 75 and 85 kg.

95.4% will weigh between 70 and 90 kg (±2 SD).

99.7% of patients will weigh between 65 and 95 kg (±3 SD).

See how this relates to the graph of the data in *Fig. 6*.

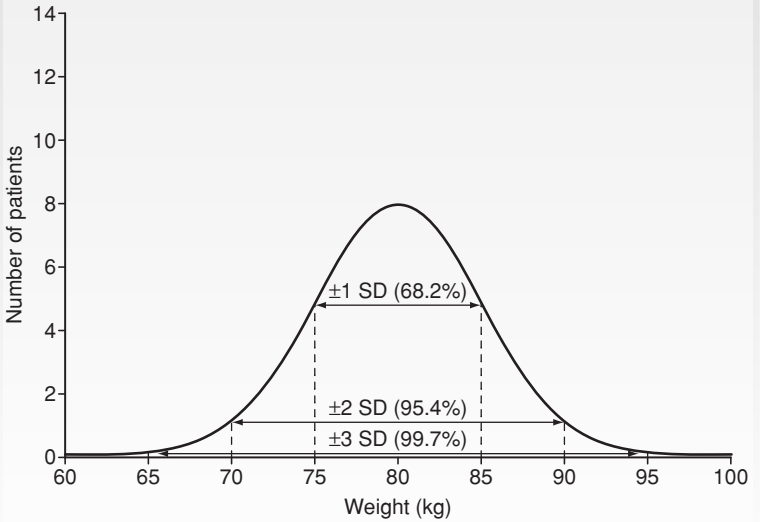

Fig. 6. Graph showing normal distribution of weights of patients enrolling in a trial with mean 80 kg, SD 5 kg.

If we have two sets of data with the same mean but different SDs, then the data set with the larger SD has a wider spread than the data set with the smaller SD.

For example, if another group of patients enrolling for the trial has the same mean weight of 80 kg but an SD of only 3, ±1 SD will include 68.2% of the subjects, so 68.2% of patients will weigh between 77 and 83 kg (Fig. 7). Compare this with the example above.

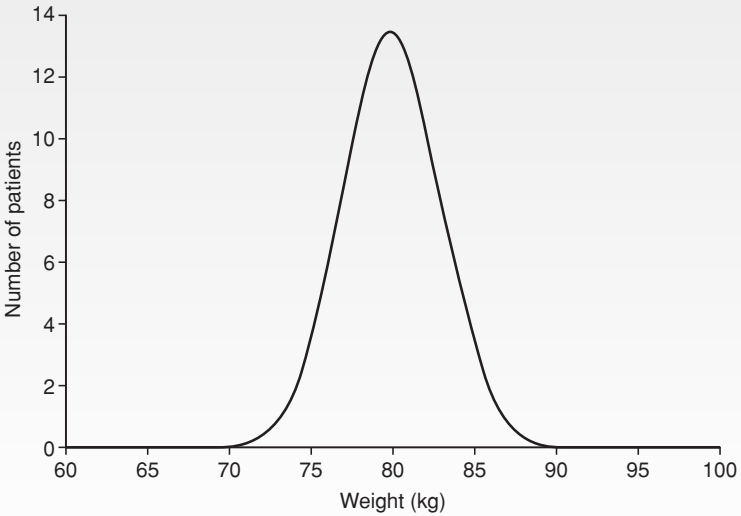

Fig. 7. Graph showing normal distribution of weights of patients enrolling in a trial with mean 80 kg, SD 3 kg.

Watch out for...

SD should only be used when the data have a normal distribution. However, means and SDs are often wrongly used for data which are not normally distributed.

A simple check for a normal distribution is to see if 2 SDs away from the mean are still within the possible

range for the variable. For example, if we have some length of hospital stay data with a mean stay of 10 days and a SD of 8 days then:

$$\text{mean} - 2 \times \text{SD} = 10 - 2 \times 8 = 10 - 16 = -6 \text{ days.}$$

This is clearly an impossible value for length of stay, so the data cannot be normally distributed. The mean and SDs are therefore not appropriate measures to use.

Good news – it is not necessary to know how to calculate the SD.

It *is* worth learning the figures above off by heart, so a reminder –

±1 SD includes 68.2% of the data

±2 SD includes 95.4%,

±3 SD includes 99.7%.

Keeping the "normal distribution" curve in *Fig. 6* in mind may help.

Examiners may ask what percentages of subjects are included in 1, 2 or 3 SDs from the mean. Again, try to memorize those percentages.

CONFIDENCE INTERVALS

How important are they?

★★★★★ Important – given in ¾ of papers.

How easy are they to understand?

☺☺ A difficult concept, but one where a small amount of understanding will get you by without having to worry about the details.

When is it used?

Confidence intervals (CI) are typically used when, instead of simply wanting the mean value of a sample, we want a range that is likely to contain the true population value.

This "true value" is another tough concept – it is the mean value that we would get if we had data for the whole population.

What does it mean?

Statisticians can calculate a range (interval) in which we can be fairly sure (confident) that the "true value" lies.

For example, we may be interested in blood pressure (BP) reduction with antihypertensive treatment. From a sample of treated patients we can work out the mean change in BP.

However, this will only be the mean for our particular sample. If we took another group of patients we would not expect to get exactly the same value, because chance can also affect the change in BP.

The CI gives the range in which the true value (i.e. the mean change in BP if we treated an infinite number of patients) is likely to be.

EXAMPLES

The average systolic BP before treatment in study A, of a group of 100 hypertensive patients, was 170 mmHg. After treatment with the new drug the mean BP dropped by 20 mmHg.

If the 95% CI is 15–25, this means we can be 95% confident that the true effect of treatment is to lower the BP by 15–25 mmHg.

In study B 50 patients were treated with the same drug, also reducing their mean BP by 20 mmHg, but with a wider 95% CI of −5 to +45. This CI includes zero (no change). This means there is more than a 5% chance that there was no true change in BP, and that the drug was actually ineffective.

Watch out for...

The size of a CI is related to the sample size of the study. Larger studies usually have a narrower CI.

Where a few interventions, outcomes or studies are given it is difficult to visualize a long list of means and CIs. Some papers will show a chart to make it easier.

For example, "meta-analysis" is a technique for bringing together results from a number of similar studies to give one overall estimate of effect. Many meta-analyses compare the treatment effects from

those studies by showing the mean changes and 95% CIs in a chart. An example is given in *Fig. 8*.

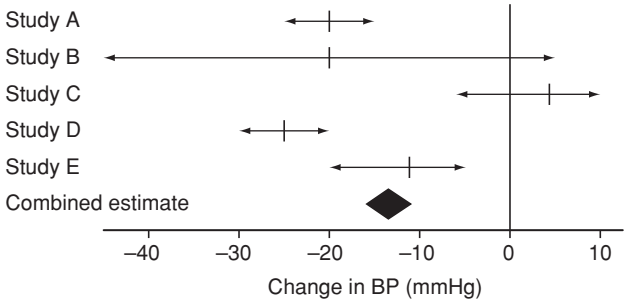

Fig. 8. Plot of 5 studies of a new antihypertensive drug. See how the results of studies A and B above are shown by the top two lines, i.e. 20 mmHg, 95% CI 15–25 for study A and 20 mmHg, 95% CI −5 to +45 for study B.

The vertical axis does not have a scale. It is simply used to show the zero point on each CI line.

The statistician has combined the results of all five studies and calculated that the overall mean reduction in BP is 14 mmHg, CI 12–16. This is shown by the "combined estimate" diamond. See how combining a number of studies reduces the CI, giving a more accurate estimate of the true treatment effect.

The chart shown in *Fig. 8* is called a "Forest plot" or, more colloquially, a "blobbogram".

Standard deviation and confidence intervals – what is the difference? Standard deviation tells us about the variability (spread) in a sample.

The CI tells us the range in which the true value (the mean if the sample were infinitely large) is likely to be.

An exam question may give a chart similar to that in *Fig. 8* and ask you to summarize the findings. Consider:

- Which study showed the greatest change?
- Did all the studies show change in favour of the intervention?
- Were the changes statistically significant?

In the example above, study D showed the greatest change, with a mean BP drop of 25 mmHg.

Study C resulted in a mean increase in BP, though with a wide CI. The wide CI could be due to a low number of patients in the study.

The combined estimate of a mean BP reduction of 14 mmHg, 95% CI 12–16, is statistically significant.

P VALUES

How important is it?

★★★★★ A really important concept, *P* values are given in more than four out of five papers.

How easy is it to understand?

☺☺☺ Not easy, but worth persevering as it is used so frequently.

It is not important to know how the *P* value is derived – just to be able to interpret the result.

When is it used?

The *P* (probability) value is used when we wish to see how likely it is that a hypothesis is true. The hypothesis is usually that there is *no* difference between two treatments, known as the "null hypothesis".

What does it mean?

The *P* value gives the probability of any observed difference having happened by chance.

$P = 0.5$ means that the probability of the difference having happened by chance is 0.5 in 1, or 50:50.

$P = 0.05$ means that the probability of the difference having happened by chance is 0.05 in 1, i.e. 1 in 20.

It is the figure frequently quoted as being "statistically significant", i.e. unlikely to have happened by chance and therefore important. However, this is an arbitrary figure.

If we look at 20 studies, even if none of the treatments work, one of the studies is likely to have a *P* value of 0.05 and so appear significant!

The lower the *P* value, the less likely it is that the difference happened by chance and so the higher the significance of the finding.

$P = 0.01$ is often considered to be "highly significant". It means that the difference will only have happened by chance 1 in 100 times. This is unlikely, but still possible.

$P = 0.001$ means the difference will have happened by chance 1 in 1000 times, even less likely, but still just possible. It is usually considered to be "very highly significant".

EXAMPLES

Out of 50 new babies on average 25 will be girls, sometimes more, sometimes less.

Say there is a new fertility treatment and we want to know whether it affects the chance of having a boy or a girl. Therefore we set up a null hypothesis – that the treatment *does not* alter the chance of having a girl. Out of the first 50 babies resulting from the treatment, 15 are girls. We then need to know the probability that this just happened by chance, i.e. did this happen by chance or has the treatment had an effect on the sex of the babies?

The *P* value gives the probability that the null hypothesis is true.

The *P* value in this example is 0.007. Do not worry about how it was calculated, concentrate on what it means. It means the result would only have happened by chance in 0.007 in 1 (or 1 in 140) times if the treatment did not actually affect the sex of the baby. This is highly unlikely, so we can reject our hypothesis and conclude that the treatment probably *does* alter the chance of having a girl.

Try another example: Patients with minor illnesses were randomized to see either Dr Smith or Dr Jones. Dr Smith ended up seeing 176 patients in the study whereas Dr Jones saw 200 patients (Table 2).

Table 2. Number of patients with minor illnesses seen by two GPs

	Dr Jones (*n*=200)[a]	**Dr Smith (*n*=176)**	***P* value**	**i.e. could have happened by chance**
Patients satisfied with consultation (%)	186 (93)	168 (95)	0.4	Four times in 10 – possible
Mean (SD) consultation length (minutes)	16 (3.1)	6 (2.8)	<0.001	< One time in 1000 – very unlikely
Patients getting a prescription (%)	58 (29)	76 (43)	0.3	Three times in 10 – possible
Mean (SD) number of days off work	3.5 (1.3)	3.6 (1.3)	0.8	Eight times in 10 – probable
Patients needing a follow-up appointment (%)	46 (23)	72 (41)	0.05	Only one time in 20 – fairly unlikely

[a] *n*=200 means that the total number of patients seen by Dr Jones was 200.

Watch out for...

The "null hypothesis" is a concept that underlies this and other statistical tests.

The test method assumes (hypothesizes) that there is *no* (null) difference between the groups. The result of the test either supports or rejects that hypothesis.

The null hypothesis is generally the opposite of what we are actually interested in finding out. If we are interested if there is a difference between two treatments then the null hypothesis would be that there is no difference and we would try to disprove this.

Try not to confuse statistical significance with clinical relevance. If a study is too small, the results are unlikely to be statistically significant even if the intervention actually works. Conversely a large study may find a statistically significant difference that is too small to have any clinical relevance.

You may be given a set of P values and asked to interpret them. Remember that $P = 0.05$ is usually classed as "significant", $P = 0.01$ as "highly significant" and $P = 0.001$ as "very highly significant".

In the example above, only two of the sets of data showed a significant difference between the two GPs. Dr Smith's consultations were very highly significantly shorter than those of Dr Jones. Dr Smith's follow-up rate was significantly higher than that of Dr Jones.

t TESTS AND OTHER PARAMETRIC TESTS

How important are they?

★★★★ Used in one in three papers, they are an important aspect of medical statistics.

How easy are they to understand?

☺ The details of the tests themselves are difficult to understand.

Thankfully you do not need to know them. Just look for the *P* value (see page 24) to see how significant the result is. Remember, the smaller the *P* value, the smaller the chance that the "null hypothesis" is true.

When are they used?

Parametric statistics are used to compare samples of "normally distributed" data (see page 9). If the data do *not* follow a normal distribution, these tests should not be used.

What do they mean?

A parametric test is any test which requires the data to follow a specific distribution, usually a normal distribution. Common parametric tests you will come across are the *t* test and the χ^2 test .

Analysis of variance (ANOVA) This is a group of statistical techniques used to compare the means of two or more samples to see whether they come from the same population – the "null hypothesis". These techniques can also allow for independent variables which may have an effect on the outcome.

Again, check out the *P* value.

t *test (also known as Student's* t*)* *t* tests are typically used to compare just two samples. They test the probability that the samples come from a population with the same mean value.

χ^2 test A frequently used parametric test is the χ^2 test. It is covered separately (page 34).

EXAMPLE

Two hundred adults seeing an asthma nurse specialist were randomly assigned to either a new type of bronchodilator or placebo.

After 3 months the peak flow rates in the treatment group had increased by a mean of 96 l/min (SD 58), and in the placebo group by 70 l/min (SD 52). The null hypothesis is that there is no difference between the bronchodilator and the placebo.

The *t* statistic is 11.14, resulting in a *P* value of 0.001. It is therefore very unlikely (1 in 1000 chance) that the null hypothesis is correct so we reject the hypothesis and conclude that the new bronchodilator is significantly better than the placebo.

Watch out for...

Parametric tests should only be used when the data follow a "normal" distribution. You may find reference to the "Kolmogorov Smirnov" test. This

tests the hypothesis that the collected data are from a normal distribution and therefore assesses whether parametric statistics can be used.

Sometimes authors will say that they have "transformed" data and then analyzed it with a parametric test. This is quite legitimate – it is not cheating! For example, a skewed distribution might become normally distributed if the logarithm of the values is used.

MANN–WHITNEY AND OTHER NON-PARAMETRIC TESTS

How important are they?

★★ Used in one in six papers.

How easy are they to understand?

☺ Non-parametric testing is difficult to understand.

However, you do not need to know the details of the tests. Look out for the *P* value (see page 24) to see how significant the results are. Remember, the smaller the *P* value, the smaller the chance that the "null hypothesis" is true.

When are they used?

Non-parametric statistics are used when the data are *not* normally distributed and so are not appropriate for "parametric" tests.

What do they mean?

Rather than comparing the values of the raw data, statisticians "rank" the data and compare the ranks.

EXAMPLE

Mann–Whitney U test A GP introduced a nurse triage system into her practice. She was interested in finding out whether the age of the patients attending for triage appointments was different to that of patients who made emergency appointments with the GP.

Six hundred and forty-six patients saw the triage nurse and 532 patients saw the GP. The median age of the triaged patients was 50 years (1st quartile 40 years, 3rd quartile 54), for the GP it was 46 (22, 58). Note how the quartiles show an uneven distribution around the median, so the data cannot be normally distributed and a non-parametric test is appropriate. The graph in Fig. 9 shows the ages of the patients seen by the nurse and confirms a skewed, rather than normal, distribution.

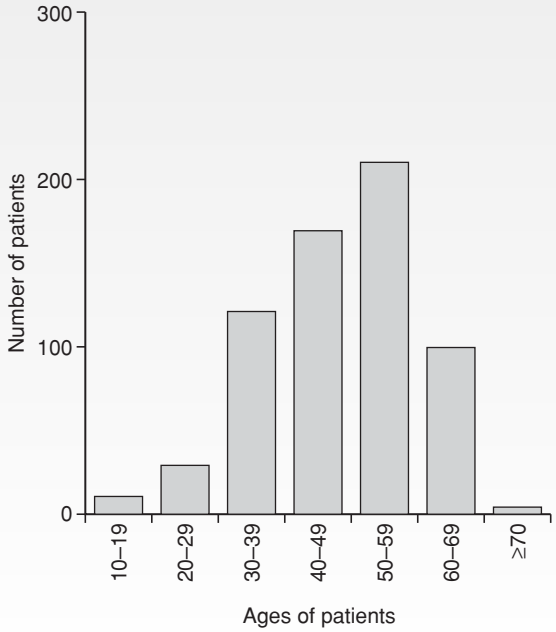

Fig. 9. Graph of ages of patients seen by triage nurse.

The statistician used a "Mann–Whitney U test" to test the hypothesis that there is no difference between the ages of the two groups. This gave a U value of 133 200 with a *P* value of < 0.001. Ignore the actual U value but concentrate on the *P* value, which in this case suggests that the triage nurse's patients were very highly significantly older than those who saw the GP.

Watch out for...

The "Wilcoxon signed rank test", "Kruskal Wallis" and "Friedman" tests are other non-parametric tests. Do not be put off by the names – go straight to the *P* value.

CHI-SQUARED

Usually written as χ^2 (for the test) or X^2 (for it's value); Chi is pronounced as in sky without the s.

How important is it?

★★★★ A frequently used test of significance, given in a third of papers.

How easy is it to understand?

☺☺ Do not try to understand the X^2 value, just look at whether or not the result is significant.

When is it used?

It is a measure of the difference between actual and expected frequencies.

What does it mean?

The "expected frequency" is that there is no difference between the sets of results (the null hypothesis). In that case, the X^2 value would be zero.

The larger the actual difference between the sets of results, the greater the X^2 value. However, it is difficult to interpret the X^2 value by itself as it depends on the number of factors studied.

Statisticians make it easier for you by giving the *P* value (see p 24), giving you the likelihood there is no real difference between the groups.

So, do not worry about the actual value of X^2 but look at its *P* value.

EXAMPLES

A group of patients with bronchopneumonia were treated with either amoxicillin or erythromycin. The results are shown in *Table 3*.

Table 3. Comparison of effect of treatment of bronchopneumonia with amoxicillin or erythromycin

	Type of antibiotic given					
	Amoxicillin		Erythromycin		Total	
Improvement at 5 days	144	(60%)	160	(67%)	304	(63%)
No improvement at 5 days	96	(40%)	80	(33%)	176	(37%)
Total	240	(100%)	240	(100%)	480	(100%)
	$X^2 = 2.3$; $P = 0.13$					

A table like this is known as a "contingency table" or "two-way table".

First, look at the table to get an idea of the differences between the effects of the two treatments.

Remember, do not worry about the X^2 value itself, but see whether it is significant. In this case *P* is 0.13, so the difference in treatments is not statistically significant.

Watch out for...

Some papers will also give the "degrees of freedom" (df), for example $X^2 = 2.3$; df 1; $P = 0.13$. See page 99 for an explanation. This is used with the X^2 value to work out the *P* value.

Other tests you may find. Instead of the χ^2 test, "Fisher's exact test" is sometimes used to analyze contingency tables. Fisher's test is the best choice as it always gives the exact *P* value, particularly where the numbers are small.

The χ^2 test is simpler for statisticians to calculate but gives only an approximate *P* value and is inappropriate for small samples. Statisticians may apply "Yates' continuity correction" or other adjustments to the χ^2 test to improve the accuracy of the *P* value.

The "Mantel Haenszel test" is an extension of the χ^2 test that is used to compare several two-way tables.

RISK RATIO

Often referred to as relative risk.

How important is it?

★★★ Used in one in six papers.

How easy is it to understand?

☺☺☺☺ Risk is a relatively intuitive concept that we encounter every day, but interpretation of risk (especially low risk) is often inconsistent. The risk of death while travelling to the shops to buy a lotto ticket can be higher than the risk of winning the jackpot!

When is it used?

Relative risk is used in "cohort studies", prospective studies that follow a group (cohort) over a period of time and investigate the effect of a treatment or risk factor.

What does it mean?

First, risk itself. *Risk* is the probability that an event will happen. It is calculated by dividing the number of events by the number of people at risk.

One boy is born for every two births, so the probability (risk) of giving birth to a boy is

½ = 0.5

If one in every 100 patients suffers a side-effect from a treatment, the risk is

$\frac{1}{100} = 0.01$

Compare this with odds (page 40).

Now, risk *ratios*. These are calculated by dividing the risk in the treated or exposed group by the risk in the control or unexposed group.

A risk ratio of one indicates no difference in risk between the groups.

If the risk ratio of an event is >1, the rate of that event is increased compared to controls.

If <1, the rate of that event is reduced.

Risk ratios are frequently given with their 95% CIs – if the CI for a risk ratio *does not* include one (no difference in risk), it is statistically significant.

EXAMPLES

A cohort of 1000 regular football players and 1000 non-footballers were followed to see if playing football was significant in the injuries that they received.

After 1 year of follow-up there had been 12 broken legs in the football players and only four in the non-footballers.

The *risk* of a footballer breaking a leg was therefore 12/1000 or 0.012. The risk of a non-footballer breaking a leg was 4/1000 or 0.004.

The risk *ratio* of breaking a leg was therefore 0.012/0.004 which equals three. The 95% CI was calculated to be 0.97 to 9.41. As the CI includes the value 1 we cannot exclude the possibility that there was no difference in the risk of footballers and non-footballers breaking a leg. However, given these results further investigation would clearly be warranted.

ODDS RATIO

How important is it?

★★★ Used in a ¼ of papers.

How easy is it to understand?

☺☺ Odds are difficult to understand. Just aim to understand what the ratio means.

When is it used?

Used by epidemiologists in studies looking for factors which do harm, it is a way of comparing patients who already have a certain condition (cases) with patients who do not (controls) – a "case–control study".

What does it mean?

First, *odds*. Odds are calculated by dividing the number of times an event happens by the number of times it does not happen.

One boy is born for every two births, so the odds of giving birth to a boy are 1:1 (or 50:50) = ⅟₁ = 1

If one in every 100 patients suffers a side-effect from a treatment, the odds are

1:99 = ⅟₉₉ = 0.0101

Compare this with risk (page 37).

Next, odds *ratios*. They are calculated by dividing the odds of having been exposed to a risk factor by the odds in the control group.

An odds ratio of 1 indicates no difference in risk between the groups, i.e. the odds in each group are the same.

If the odds ratio of an event is >1, the rate of that event is increased in patients who have been exposed to the risk factor.

If <1, the rate of that event is reduced.

Odds ratios are frequently given with their 95% CI – if the CI for an odds ratio *does not* include 1 (no difference in odds), it is statistically significant.

EXAMPLES

A group of 100 patients with knee injuries, "cases", was matched for age and sex to 100 patients who did not have injured knees, "controls".

In the cases, 40 skied and 60 did not, giving the *odds* of being a skier for this group of 40:60 or 0.66.

In the controls, 20 patients skied and 80 did not, giving the odds of being a skier for the control group of 20:80 or 0.25.

We can therefore calculate the odds *ratio* as 0.66/0.25 = 2.64. The 95% CI is 1.41 to 5.02.

If you cannot follow the maths, do not worry! The odds ratio of 2.64 means that the number of skiers in the cases is higher than the number of skiers in the controls, and as the CI does not include 1 (no difference in risk) this is statistically significant. Therefore, we can conclude that skiers are more likely to get a knee injury than non-skiers.

Watch out for...

Authors may give the percentage *change* in the odds ratio rather than the odds ratio itself. In the example above, the odds ratio of 2.64 means the same as a 164% increase in the odds of injured knees amongst skiers.

Odds ratios are often interpreted by the reader in the same way as risk ratios. This is reasonable when the odds are low, but for common events the odds and the risks (and therefore their ratios) will give very different values. For example, the odds of giving birth to a boy are 1, whereas the risk is 0.5. However, in the side-effect example given above the odds are 0.0101, a similar value to the risk of 0.01. For this reason, if you are looking at a case–control study, check that the authors have used odds ratios rather than risk ratios.

RISK REDUCTION AND NUMBERS NEEDED TO TREAT

How important are they?

★★★ Although only quoted in 5% of papers, they are helpful in trying to work out how worthwhile a treatment is in clinical practice.

How easy are they to understand?

☺☺☺ "Relative risk reduction" (RRR) and "absolute risk reduction" (ARR) need some concentration. "Numbers needed to treat" (NNT) are pretty intuitive, useful and not too difficult to work out for yourself.

When are they used?

They are used when an author wants to know how often a treatment works, rather than just whether it works.

What do they mean?

ARR is the difference between the event rate in the intervention group and that in the control group. It is also the reciprocal of the NNT and is usually given as a percentage, i.e. $\text{ARR} = \frac{100}{\text{NNT}}$

NNT is the number of patients who need to be treated for one to get benefit.

RRR is the proportion by which the intervention reduces the event rate.

EXAMPLES

One hundred women with vaginal candida were given an oral antifungal, 100 were given placebo. They were reviewed 3 days later. The results are given in *Table 4*.

Table 4. Results of placebo-controlled trial of oral antifungal agent

Given antifungal		Given placebo	
Improved	No improvement	Improved	No improvement
80	20	60	40

ARR = improvement rate in the intervention group – improvement rate in the control group = 80% – 60% = 20%

$$\text{NNT} = \frac{100}{\text{ARR}} = \frac{100}{20} = 5$$

So five women have to be treated for one to get benefit.

The incidence of candidiasis was reduced from 40% with placebo to 20% with treatment , i.e. by half.

Thus, the RRR is 50%.

In another trial young men were treated with an expensive lipid-lowering agent. Five years later the death rate from ischaemic heart disease (IHD) is recorded. See *Table 5* for the results.

Table 5. Results of placebo-controlled trial of Cleverstatin

Given Cleverstatin		Given placebo	
Survived	Died	Survived	Died
998 (99.8%)	2 (0.2%)	996 (99.6%)	4 (0.4%)

ARR = improvement rate in the intervention group – improvement rate in the control group = 99.8% – 99.6% = 0.2%

$$\text{NNT} = \frac{100}{\text{ARR}} = \frac{100}{0.2} = 500$$

So 500 men have to be treated for 5 years for one to survive who would otherwise have died.

The incidence of death from IHD is reduced from 0.4% with placebo to 0.2% with treatment – i.e. by half.

Thus, the RRR is 50%.

The RRR and NNT from the same study can have opposing effects on prescribing habits. The RRR of 50% in this example sounds fantastic. However, thinking of it in terms of an NNT of 500 might sound less attractive: for every life saved, 499 patients had unnecessary treatment for 5 years.

Watch out for...

Usually the necessary percentages are given in the abstract of the paper. Calculating the ARR is easy: subtract the percentage that improved without treatment from the percentage that improved with treatment.

Again, dividing that figure into 100 gives the NNT.

With an NNT you need to know:

(a) What treatment?

- What are the side-effects?
- What is the cost?

(b) For how long?

(c) To achieve what?

- How serious is the event you are trying to avoid?
- How easy is it to treat if it happens?

For treatments, the lower the NNT the better – but look at the context.

(a) NNT of 10 for treating a sore throat with expensive blundamycin

- not attractive

(b) NNT of 10 for prevention of death from leukaemia with a non-toxic chemotherapy agent

- worthwhile

Expect NNTs for prophylaxis to be much larger. For example, an immunization may have an NNT in the thousands but still be well worthwhile.

Numbers needed to harm (NNH) may also be important.

$$\text{NNH} = \frac{100}{(\%\text{ on treatment that had SEs}) - (\%\text{ not on treatment that had SEs})}$$

In the example above, 6% of those on cleverstatin had peptic ulceration as opposed to 1% of those on placebo.

$$\text{NNH} = \frac{100}{6-1} = \frac{100}{5} = 20$$

i.e. for every 20 patients treated, one peptic ulcer was caused.

You may see ARR and RRR given as a proportion instead of a percentage. So, an ARR of 20% is the same as an ARR of 0.2

Be prepared to calculate RRR, ARR and NNT from a set of results. You may find that it helps to draw a simple table like *Table 5* and work from there.

CORRELATION

How important is it?

★★ Only used in 10% of medical papers.

How easy is it to understand?

☺☺☺

When is it used?

Where there is a linear relationship between two variables there is said to be a correlation between them. Examples are height and weight in children, or socio-economic class and mortality.

The *strength* of that relationship is given by the "correlation coefficient".

What does it mean?

The correlation coefficient is usually denoted by the letter "*r*" for example $r = 0.8$

A *positive* correlation coefficient means that as one variable is increasing the value for the other variable is also increasing – the line on the graph slopes up from left to right. Height and weight have a positive correlation: children get heavier as they grow taller.

A *negative* correlation coefficient means that as the value of one variable goes up the value for the other variable goes down – the graph slopes down from left

to right. Higher socio-economic class is associated with a lower mortality, giving a negative correlation between the two variables.

If there is a perfect relationship between the two variables then $r = 1$ (if a positive correlation) or $r = -1$ (if a negative correlation).

If there is no correlation at all (the points on the graph are completely randomly scattered) then $r = 0$.

The following is a good rule of thumb when considering the size of a correlation:

$r = 0–0.2$: very low and probably meaningless.

$r = 0.2–0.4$: a low correlation that might warrant further investigation.

$r = 0.4–0.6$: a reasonable correlation.

$r = 0.6–0.8$: a high correlation.

$r = 0.8–1.0$: a very high correlation. Possibly too high! Check for errors or other reasons for such a high correlation.

This guide also applies to negative correlations.

Examples

A nurse wanted to be able to predict the laboratory HbA_{1c} result (a measure of blood glucose control) from the fasting blood glucoses which she measured in her clinic. On 12 consecutive diabetic patients she noted the fasting glucose and simultaneously drew blood for HbA_{1c}. She compared the pairs of measurements and drew the graph in *Fig. 10*.

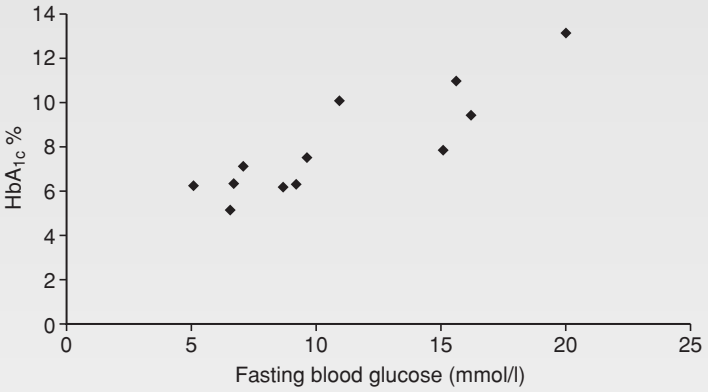

Fig. 10. Plot of fasting glucose and HbA_{1c} in 12 patients with diabetes. For these results $r = 0.88$, showing a very high correlation.

A graph like this is known as a "scatter plot".

An occupational therapist developed a scale for measuring physical activity and wondered how much it correlated to Body Mass Index (BMI) in 12 of her adult patients. *Fig. 11* shows how they related.

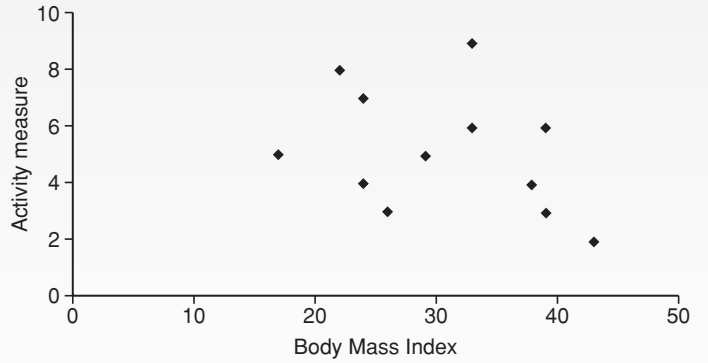

Fig. 11. BMI and activity measure in 12 adult patients.

In this example, $r = -0.34$, indicating a low correlation. The fact that the r value is negative shows that the correlation is negative, indicating that patients with a higher level of physical activity tended to have a lower BMI.

Watch out for...

Correlation tells us how strong the association between the variables is, but does not tell us about cause and effect in that relationship.

The "Pearson correlation coefficient", Pearson's *r*, is used if the values are sampled from "normal" populations (page 9). Otherwise the "Spearman correlation coefficient" is used. However, the interpretation of the two is the same.

Where the author shows the graph, you can get a good idea from the scatter as to how strong the relationship is without needing to know the *r* value.

Authors often give *P* values with correlations, however take care when interpreting them. Although a correlation needs to be significant, we need also to consider the size of the correlation. If a study is sufficiently large, even a small clinically unimportant correlation will be highly significant.

R^2 is sometimes given. As it is the square of the *r* value, and squares are always positive, you cannot use it to tell whether the graph slopes up or down.

What it *does* tell you is how much of the variation in one value is caused by the other.

In *Fig. 10*, $r = -0.88$. $R^2 = -0.88 \times -0.88 = 0.77$. This means that 77% of the variation in HbA_{1c} is related to the variation in fasting glucose.

Again, the closer the R^2 value is to 1, the higher the correlation.

It is very easy for authors to compare a large number of variables using correlation and only present the ones that happen to be significant. So, check to make sure there is a plausible explanation for any significant correlations.

Also bear in mind that a correlation only tells us about linear (straight line) relationships between variables. Two variables may be strongly related but not in a straight line, giving a low correlation coefficient.

REGRESSION

How important is it?

★★★ Regression analysis is used in ¼ of papers.

How easy is it to understand?

☺☺ The idea of trying to fit a line through a set of points to make the line as representative as possible is relatively straightforward. However, the mathematics involved in fitting regression models are more difficult to understand.

When is it used?

Regression analysis is used to find how one set of data relates to another.

This can be particularly helpful where we want to use one measure as a proxy for another – for example, a near-patient test as a proxy for a lab test.

What does it mean?

A regression line is the "best fit" line through the data points on a graph.

The regression coefficient gives the "slope" of the graph, in that it gives the change in value of one outcome, per unit change in the other.

EXAMPLE

Consider the graph shown in *Fig. 10* (page 50). A statistician calculated the line that gave the "best fit" through the scatter of points, shown in *Fig. 12*.

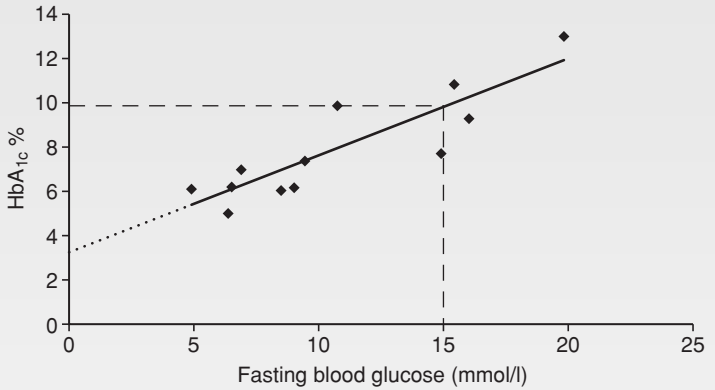

Fig. 12. Plot with linear regression line of fasting glucose and HbA_{1c} in 12 patients with diabetes.

The line is called a "regression line".

To predict the HbA_{1c} for a given blood glucose the nurse could simply plot it on the graph, as here where a fasting glucose of 15 predicts an HbA_{1c} of 9.95.

This can also be done mathematically. The slope and position of the regression line can be represented by the "regression equation":

$HbA_{1c} = 3.2 + (0.45 \times \text{blood glucose})$.

The 0.45 figure gives the *slope* of the graph and is called the "regression coefficient".

The "regression constant" that gives the *position* of the line on the graph is 3.2: it is the point where the line crosses the vertical axis.

Try this with a glucose of 15:

$HbA_{1c} = 3.2 + (0.45 \times 15) = 3.2 + 6.75 = 9.95$

This regression equation can be applied to any regression line. It is represented by:

$y = a + bx$

To predict the value y (value on the vertical axis of the graph) from the value x (on the horizontal axis), b is the regression coefficient and a is the constant.

Other values sometimes given with regression

You may see other values quoted. The regression coefficient and constant can be given with their "standard errors". These indicate the accuracy that can be given to the calculations. Do not worry about the actual value of these but look at their *P* values. The lower the *P* value, the greater the significance.

The R^2 value may also be given. This represents the amount of the variation in the data that is explained by the regression. In our example the R^2 value is 0.77. This is stating that 77% of the variation in the HbA_{1c} result is accounted for by variation in the blood glucose.

Other types of regression

The example above is a "linear regression", as the line that best fits the points is straight.

Other forms of regression include:

Logistic regression This is used where each case in the sample can only belong to one of two groups (e.g. having disease or not) with the outcome as the probability that a case belongs to one group rather than the other.

Poisson Regression Poisson regression is mainly used to study waiting times or time between rare events.

Cox proportional hazards regression model The Cox regression model (page 60) is used in survival analysis where the outcome is time until a certain event.

Watch out for...

Regression should not be used to make predictions outside of the range of the original data. In the example above, we can only make predictions from blood glucoses which are between 5 and 20.

Regression or correlation?

Regression and correlation are easily confused.

Correlation measures the *strength* of the association between variables.

Regression *quantifies* the association. It should only be used if one of the variables is thought to precede or cause the other.

SURVIVAL ANALYSIS: LIFE TABLES AND KAPLAN–MEIER PLOTS

How important are they?

★★★ Survival analysis techniques are used in 20% of papers.

How easy are they to understand?

☺☺☺ Life tables are difficult to interpret. Luckily, most papers make it easy for you by showing the resulting plots – these graphs give a good visual feel of what has happened to a population over time.

When are they used?

Survival analysis techniques are concerned with representing the time until a single event occurs. That event is often death, but it could be any other single event, for example time until discharge from hospital.

Survival analysis techniques are able to deal with situations in which the end event has not happened in every patient or when information on a case is only known for a limited duration – known as "censored" observations.

What do they mean?

Life table. A life table is a table of the proportion of patients surviving over time.

Life table methods look at the data at a number of fixed time points and calculate the survival rate at those times. The most commonly used method is Kaplan–Meier.

Kaplan–Meier

The Kaplan–Meier approach recalculates the survival rate when an end event (e.g. death) occurs in the data set, i.e. when a change happens rather than at fixed intervals.

This is usually represented as a "survival plot". *Fig. 13* shows a fictitious example.

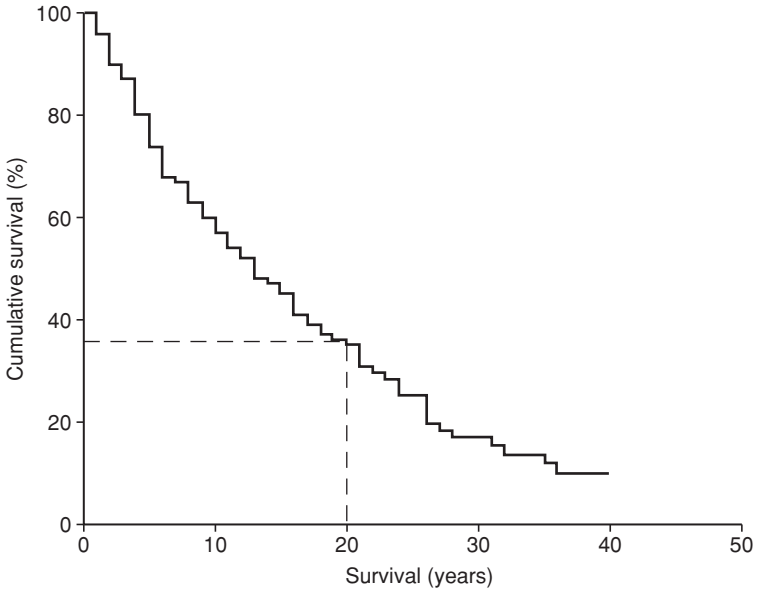

Fig. 13. Kaplan–Meier survival plot of a cohort of patients with rheumatoid arthritis.

The dashed line shows that at 20 years, 36% of this group of patients were still alive.

Watch out for...

Life tables and Kaplan–Meier survival estimates are also used to compare survival between groups. The plots make any difference between survival in two groups beautifully clear. *Fig. 14* shows the same group of patients as above, but compares survival for men and women.

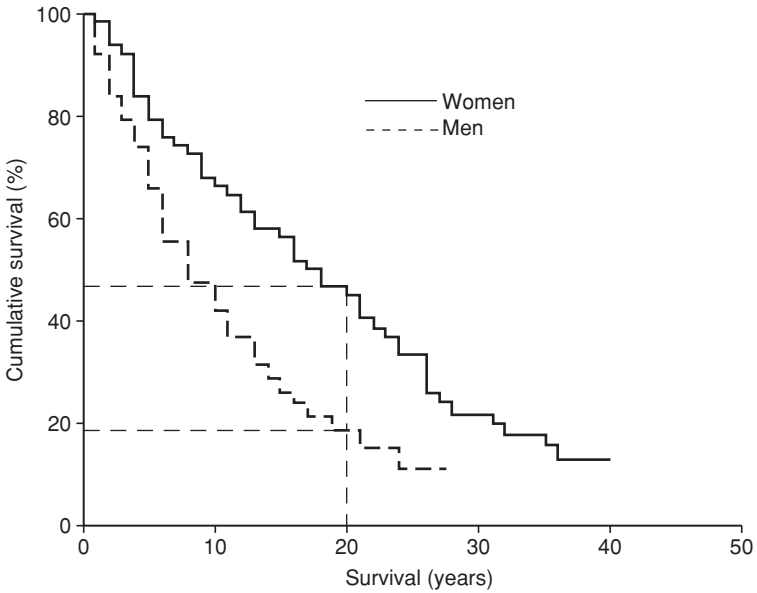

Fig. 14. Kaplan–Meier survival plot comparing men and women with rheumatoid arthritis.

In this example 46% of women were still alive at 20 years but only 18% of men.

The test to compare the survival between these two groups is called the "log rank test". Its *P* value will tell you how significant the result of the test is.

THE COX REGRESSION MODEL

Also known as the proportional hazards survival model.

How important is it?

★★ It appeared in one in ten papers.

How easy is it to understand?

☺☺ Just aim to understand the end result – the "hazard ratio" (HR).

When is it used?

The Cox regression model is used to investigate the relationship between an event (usually death) and possible explanatory variables, for instance smoking status or weight.

What does it mean?

The Cox regression model provides us with estimates of the effect that different factors have on the time until the end event.

As well as considering the significance of the effect of different factors (e.g. how much shorter male life expectancy is compared to that of women), the model can give us an estimate of life expectancy for an individual.

The "HR" is the ratio of the hazard (chance of something harmful happening) of an event in one group of observations divided by the hazard of an event in another group. A HR of 1 means the risk is 1 × that of the second group, i.e. the same. A HR of 2 implies twice the risk.

EXAMPLE

The Cox regression model shows us the effect of being in one group compared with another.

Using the rheumatoid arthritis cohort on page 58, we can calculate the effect that gender has on survival. *Table 6* gives the results of a Cox model estimate of the effect.

Table 6. Cox model estimate of the effect of sex on survival in a cohort of patients with rheumatoid arthritis

Parameter	HR[a] (95% CI)[b]	df [c]	*P* value[d]
Sex (Male)	1.91 (1.21 to 3.01)	1	<0.05

[a] The HR of 1.91 means that the risk of death in any particular time period for men was 1.91 times that for women.

[b] This CI means we can be 95% confident that the true HR is between 1.21 and 3.01.

[c] Degrees of freedom – see glossary, page 99.

[d] The *P* value of <0.05 suggests that the result is significant.

SENSITIVITY, SPECIFICITY AND PREDICTIVE VALUE

How important are they?

★★★ Even though they are only used in less than 10% of papers, a working knowledge is important in interpreting papers that study screening.

How easy are they to understand?

☺☺☺ The tables themselves are fairly easy to understand. However, there is a bewildering array of information that can be derived from them.

To avoid making your head spin, do not read this section until you are feeling fresh. You may need to go over it a few days running until it is clear in your mind.

When are they used?

They are used to analyze the value of screening or tests.

What do they mean?

Think of any screening test for a disease. For each patient:

- the disease itself may be present or absent;
- the test result may be positive or negative.

We need to know how useful the test is.

The results can be put in the "two-way table" shown in *Table 7*. Try working through it.

Table 7. Two-way table

		Disease:	
		Present	Absent
Test result:	Positive	A	B (False positive)
	Negative	C (False negative)	D

Sensitivity. If a patient has the disease, we need to know how often the test will be positive, i.e. "positive in disease".

This is calculated from: $\frac{A}{A+C}$

This is the rate of pick-up of the disease in a test, and is called the *Sensitivity.*

Specificity. If the patient is in fact healthy, we want to know how often the test will be negative, i.e. "negative in health".

This is given by: $\frac{D}{D+B}$

This is the rate at which a test can exclude the possibility of the disease , and is known as the S*pecificity.*

Positive Predictive Value. If the test result is positive, what is the likelihood that the patient will have the condition?

Look at: $\frac{A}{A+B}$

This is known as the *Positive Predictive Value* (PPV).

Negative Predictive Value. If the test result is negative, what is the likelihood that the patient will be healthy?

Here we use: $\frac{D}{D+C}$

This is known as the *Negative Predictive Value* (NPV).

In a perfect test, the sensitivity, specificity, PPV and NPV would each have a value of 1. The lower the value (the nearer to zero), the less useful the test is in that respect.

EXAMPLES

Confused? Try working through an example.

Imagine a blood test for gastric cancer, tried out on 100 patients admitted with haematemesis. The actual presence or absence of gastric cancers was diagnosed from endoscopic findings and biopsy. The results are shown in *Table 8*.

Table 8. Two-way table for blood test for gastric cancer

		Gastric cancer:	
		Present	Absent
Blood result:	Positive	20	30
	Negative	5	45

$$\underline{\text{Sensitivity}} = \frac{20}{20+5} = \frac{20}{25} = 0.8$$

If the gastric cancer is present, there is an 80% (0.8) chance of the test picking it up.

$$\underline{\text{Specificity}} = \frac{45}{30+45} = \frac{45}{75} = 0.6$$

If there is no gastric cancer there is a 60% (0.6) chance of the test being negative – but 40% will have a false positive result.

$$\underline{\text{PPV}} = \frac{20}{20 + 30} = \frac{20}{50} = 0.4$$

There is a 40% (0.4) chance, if the test is positive, that the patient actually has gastric cancer.

$$\underline{\text{NPV}} = \frac{45}{45 + 5} = \frac{45}{50} = 0.9$$

There is a 90% (0.9) chance, if the test is negative, that the patient does <u>not</u> have gastric cancer. However, there is still a 10% chance of a false negative, i.e. that the patient does have gastric cancer.

Watch out for...

One more test to know about.

The "Likelihood Ratio" (LR) is the likelihood that the test result would be expected in a patient with the condition compared to the likelihood that that same result would be expected in a patient without the condition.

To calculate the LR, divide the sensitivity by (1 – specificity).

Head spinning again? Try using the example above to calculate the LR for a positive result.

$$\text{LR} = \frac{\text{sensitivity}}{(1 - \text{specificity})} = \frac{0.8}{1 - 0.6} = \frac{0.8}{0.4} = 2$$

In this example, LR for a positive result = 2. This means that if the test is positive in a patient, that patient is twice as likely to have gastric cancer than not have it.

Tip: Invent an imaginary screening or diagnostic test of your own, fill the boxes in and work out the various values. Then change the results to make the

test a lot less or more effective and see how it affects the values.

One thing you may notice is that in a rare condition, even a diagnostic test with a very high sensitivity may result in a low PPV.

If you are still feeling confused, you are in good company. Many colleagues far brighter than us admit that they get confused over sensitivity, PPV etc.

Try copying the following summary into your diary and refer to it when reading a paper:

- **Sensitivity:** how often the test is positive if the patient has the disease.
- **Specificity:** if the patient is healthy, how often the test will be negative.
- **PPV:** If the test is positive, the likelihood that the patient has the condition.
- **NPV:** If the test is negative, the likelihood that the patient will be healthy.
- **LR:** If the test is positive, how much more likely the patient is to have the disease than not have it.

Examiners love to give a set of figures which you can turn into a two-way table and ask you to calculate sensitivity, PPV etc. from them. Keep practising until you are comfortable at working with these values.

LEVEL OF AGREEMENT AND KAPPA

Kappa is often seen written as κ.

How important is it?

★ Not often used.

How easy is it to understand?

☺☺☺ When is it used?

It is a comparison of how well people or tests agree and is used when data can be put into ordered categories.

Typically it is used to look at how accurately a test can be repeated.

What does it mean?

The kappa value can vary from zero to 1.

A κ of zero means that there is no significant agreement – no more than would have been expected by chance.

A κ of 0.5 or more is considered a good agreement, a value of 0.7 shows very good agreement.

A κ of 1 means that there is perfect agreement.

EXAMPLE

If the same cervical smear slides are examined by the cytology departments of two hospitals and $\kappa = 0.3$, it suggests that there is little agreement between the two laboratories.

Watch out for...

Kappa can be used to analyze cervical smear results because they can be ordered into separate categories, e.g. CIN 1, CIN 2, CIN 3 – so-called "ordinal data". When the variable that is being considered is continuous, for example blood glucose readings, the "intra-class correlation coefficient" should be used (see glossary).

OTHER CONCEPTS

Multiple testing adjustment

Importance: ★
Ease of understanding: ☺☺

One fundamental principle of statistics is that we accept there is a chance we will come to the wrong conclusion. If we reject a null hypothesis with a *P* value of 0.05, then there is still the 5% possibility that we should not have rejected the hypothesis and therefore a 5% chance that we have come to the wrong conclusion.

If we do lots of testing then this chance of making a mistake will be present each time we do a test and therefore the more tests we do the greater the chances of drawing the wrong conclusion.

Multiple testing adjustment techniques therefore adjust the *P* value to keep the overall chance of coming to the wrong conclusion at a certain level (usually 5%).

The most commonly used method is the "Bonferroni" adjustment.

1- and 2-tailed tests

Importance: ★
Ease of understanding: ☺

When trying to reject a "null hypothesis" (page 27), we are generally interested in two possibilities: either

we can reject it because the new treatment is better than the current one, or because it is worse. By allowing the null hypothesis to be rejected from either direction we are performing a "two-tailed test" – we are rejecting it when the result is in either "tail" of the test distribution.

Occasionally there are situations where we are only interested in rejecting a hypothesis if the new treatment is worse that the current one but not if it is better. This would be better analyzed with a one-tailed test. However, be very sceptical of one-tailed tests. A *P* value that is not quite significant on a two-tailed test may become significant if a one-tailed test is used. Authors have been known to use this to their advantage!

Incidence

Importance: ★★★★
Ease of understanding: ☺☺☺☺

The number of **new** cases of a condition **over a given time** as a percentage of the population.

Example: Each year 15 people in a practice of 1000 patients develop Brett's palsy.

$$\frac{15}{1000} \times 100 = \text{yearly incidence of } 1.5\%$$

Prevalence (= Point Prevalence Rate)

Importance ★★★★
Ease of understanding: ☺☺☺☺

The **existing** number of cases of a condition at **a single point in time** as a percentage of the population.

Example: At the time of a study 90 people in a practice of 1000 patients were suffering from Brett's palsy (15 diagnosed in the last year plus 75 diagnosed in previous years).

$$\frac{90}{1000} \times 100 = \text{a prevalence of } 9\%$$

With chronic diseases like the fictitious Brett's palsy, the incidence will be lower than the prevalence – each year's new diagnoses swell the number of existing cases.

With short-term illnesses the opposite may be true. 75% of a population may have a cold each year (incidence), but at any moment only 2% are actually suffering (prevalence).

Check that you can explain the difference between incidence and prevalence.

Power

Importance: ★★
Ease of understanding: ☺☺☺

The power of a study is the probability that it would detect a statistically significant difference.

If the difference expected is 100% cure compared with 0% cure with previous treatments, a very small study would have sufficient power to detect that.

However if the expected difference is much smaller, e.g. 1%, then a small study would be unlikely to have enough power to produce a result with statistical significance.

Bayesian statistics

Importance: ★
Ease of understanding: ☺

Bayesian analysis is not often used. It is a totally different statistical approach to the classical, "frequentist" statistics explained in this book.

In Bayesian statistics, rather than considering the sample of data on its own, a "prior distribution" is set up using information that is already available. For instance, the researcher may give a numerical value and weighting to previous opinion and experience as well as previous research findings.

One consideration is that different researchers may put different weighting on the same previous findings.

The new sample data are then used to adjust this prior information to form a "posterior distribution". Thus these resulting figures have taken *both* the disparate old data *and* the new data into account.

Bayesian methodology is intuitively appealing because it reflects how we think. When we read about a new study we do not consider its results in isolation, we factor it in to our pre-existing opinions, knowledge and experience of dealing with patients.

It is only recently that computer power has been sufficient to calculate the complex models that are needed for Bayesian analysis.

STATISTICS AT WORK

In this section we have given five real-life examples of how researchers use statistical techniques to describe and analyze their work.

The extracts have been taken from research papers published in the *British Medical Journal* (reproduced with permission of The BMJ Publishing Group). If you want to see the original papers, you can download them from the *BMJ* website http://www.bmj.com/

If you wish, you can use this part to test what you have learnt:

- First, go through the abstracts and results and note down what statistical techniques have been used.
- Then try to work out why the authors have used those techniques.
- Next, try to interpret the results.
- Finally, check out your understanding by comparing it with our commentary.

The following extract is reproduced with permission from The BMJ Publishing Group

Standard deviation, relative risk, confidence interval, chi-squared, and *P* values

Diggle, L. and Deeks, J. (2000). Effect of needle length on incidence of local reactions to routine immunisation in infants aged 4 months: randomised controlled trial. *BMJ* **321:** 931–933.

Abstract

Objective: To compare rates of local reactions associated with two needle sizes used to administer routine immunisations to infants.

Design: Randomised controlled trial.

Setting: Routine immunisation clinics in eight general practices in Buckinghamshire.

Participants: Healthy infants attending for third primary immunisation due at 16 weeks of age: 119 infants were recruited, and 110 diary cards were analysed.

Interventions: Immunisation with 25 gauge, 16 mm, orange hub needle or 23 gauge, 25 mm, blue hub needle.

Main outcome measures: Parental recordings of redness, swelling, and tenderness for three days after immunisation.

Results: Shown in *Table 9*.

Table 9. Mean weights 4 month old infants and rate of local reactions to immunisation over 3 days by needle used for vaccination. Values are numbers of infants unless stated otherwise

	Size of needle			
	23 G, 25 mm (*n*=53)	25 G, 16 mm (*n*=57)	Relative risk (95% CI); *P* value[1]	
Mean (SD) weight (kg)	6.7 (0.9)	6.8 (0.9)		
Any local reaction (%)	33 (62)	48 (84)	0.74	(0.58–0.94);
P = 0.009				

[1] Analysed using χ^2 test.

What statistical methods were used and why?

At the time of the study, many clinicians were using the shorter (16 mm) needle as they assumed that it would cause less local reactions than the longer one. The researchers wished to know if this was indeed true, or whether use of longer needles might actually reduce the number of reactions.

Before analyzing the outcomes, the authors needed to know how comparable the two groups of babies were before they had their injections. They gave a number of comparisons of these "baseline characteristics", but we have only quoted the mean weights.

The authors assumed that the weights of the babies were *normally distributed*, so compared the two groups by giving their *mean* weights.

They wanted to indicate how much the babies' weights were spread around the mean, so gave the *standard deviations* of the weights.

As there were different numbers of babies in each group, *percentages* were given as a scale on which to compare of the number of babies getting reactions (the *incidence* of reactions).

This was a *prospective* study, following two *cohorts* of babies over 3 days, so *relative risk* was used to investigate the effect of the two different needles.

The authors wanted to give the range that was likely to contain the true relative risk, so gave it with its *95% confidence interval*.

The *null hypothesis* was that there would be no difference between the number of reactions in the

two groups. They used *chi-squared* (χ^2) as a measure of the difference between actual and expected number of reactions.

From the X^2 figure (not given by the authors) they calculated a *P value* giving the probability of the difference in risk between the two groups happened by chance.

What do the results mean?

The SD of the mean weight in each group was 0.9. In the 25 mm needle group:

1 SD below the average is 6.7 – 0.9 = 5.8 kg.

1 SD above the average is 6.7 + 0.9 = 7.6 kg.

±1 SD will include 68.2% of the subjects, so 68.2% of babies will have been between 5.8 and 7.6 kg.

95.4% weighed between 4.9 and 8.5 kg (±2 SD).

99.7% of the babies would have been between 4.0 and 9.4 kg (±3 SD).

33 babies out of 53 in the 25 mm needle group had a reaction. As a percentage, this is:

$$\frac{33}{53} \times 100 = 62\%$$

This is the same as the risk, or probability, that a reaction would happen in this group.

The risk ratio was calculated by dividing the risk in the 25 mm needle group with that in the 16 mm group:

$$\frac{62}{84} = 0.74$$

The risk ratio of <1 shows that the rate of reactions in the 25 mm group was lower than that in the other babies.

The CI of 0.58–0.94 for the risk ratio does not include 1 (no difference in risk), so it is statistically significant.

The X^2 value for the difference between actual and expected number of reactions was not given. However, the *P* value of 0.009 means that the probability of the difference having happened by chance is 9 in 1000. This is less than 1 in 100 and is therefore considered to be highly significant.

From the results of the research, the clinician can calculate ARR, NNT and RRR from changing to a 23 G, 25 mm needle.

ARR = [risk in the 25 mm group] – [risk in the 16 mm group] = 84 – 62 = 22%

$$\text{NNT} = \frac{100}{\text{ARR}} = \frac{100}{22} = 4.5$$

So, for every five infants vaccinated, use of the longer needle instead of the shorter needle would prevent one infant from experiencing a local reaction.

The risk of reactions was reduced from 84% to 62% by using the longer needles, so the RRR is given by:

$$\text{RRR} = \frac{84 - 62}{84} = \frac{22}{84} = 0.26 = 26\%.$$

The researchers concluded that use of 25 mm needles significantly reduced rates of local reaction to routine infant immunization, with an NNT of 5.

The following extract is reproduced with permission from The BMJ Publishing Group

Odds ratios and confidence intervals

Taylor, F., Cohen, H. and Ebrahim, S. (2001). Systematic review of long term anticoagulation or antiplatelet treatment in patients with non-rheumatic atrial fibrillation. *BMJ* **322:** 321–326.

Abstract

Objective: To examine the benefits and risks of long term anticoagulation (warfarin) compared with antiplatelet treatment (aspirin/indoprofen) in patients with non-rheumatic atrial fibrillation.

Methods: Meta-analysis of randomised controlled trials. Odds ratios (95% confidence intervals) calculated to estimate treatment effects.

Outcome measures: Fatal and non-fatal cardiovascular events, reductions of which were classified as benefits. Fatal and major non-fatal bleeding events classified as risks.

Results: There were five randomised controlled trials. There were no significant differences in mortality between the two treatment options: odds ratio 0.74 (95% CI 0.39 to 1.40) for stroke deaths; 0.86 (0.63 to 1.17) for vascular deaths.

There were more major bleeding events among patients on anticoagulation than on antiplatelet treatment: odds ratio 1.45 (0.93 to 2.27).

Table 10. Extract from table showing clinical events in trials of oral anticoagulation versus antiplatelet treatment in patients with atrial fibrillation: SPAF II 75+ trial

Treatment	Number of patients randomised	Fatal outcomes		
		Stroke	Vascular	All causes
Warfarin	197	1	16	26
Aspirin	188	2	14	24

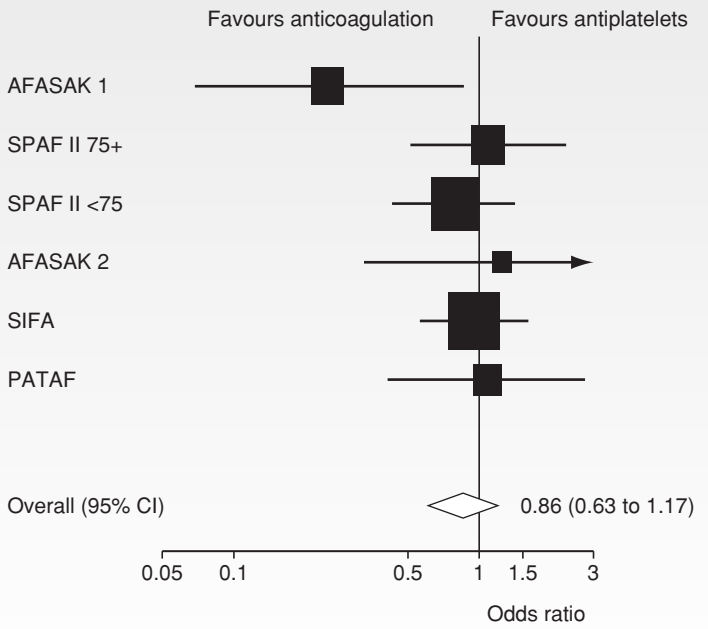

Fig. 15. Comparison of fatal vascular outcomes in trials of anticoagulation vs. antiplatelet treatment.

What statistical methods were used and why?

The researchers chose *odds ratios* as a method of comparing the outcomes of groups of patients with atrial fibrillation treated in two different ways.

The authors wanted to give the ranges that were likely to contain the true odds ratios, so gave them with their 95% *CI*.

What do the results mean?

In the *BMJ* paper, the authors gave the results and OR for eight outcomes of each of the trials. We have extracted the results for deaths for just one of the trials (*Table 10*).

Consider the fatal vascular outcomes:

Odds for vascular death in patients on warfarin =

$$\frac{\text{number times an event happens}}{\text{number of times it does not happen}} = \frac{16}{197-16} = 0.088$$

Odds for vascular death in patients on aspirin =

$$\frac{14}{188-14} = \frac{14}{174} = 0.080$$

OR =

$$\frac{\text{odds of vascular death in patients on warfarin}}{\text{odds of vascular death in patients on aspirin}} = \frac{0.088}{0.080} = 1.10$$

The OR of >1 indicates that the rate of vascular death is increased in patients on warfarin compared to patients on aspirin.

See how this is shown in the second line of *Figure 15*. The line indicates the 95% CI (0.52–2.32), the box shows the OR of 1.1. The relative sizes of the boxes indicate the number of patients in each trial.

As the CI for the OR includes 1 (no difference in odds), the difference between the results in this trial is not statistically significant.

The OR for combined fatal and non-fatal events was 0.79 (95% CI 0.61–1.02). Here, the OR of <1 indicates that the overall rate of events is reduced in patients on warfarin compared to patients on aspirin.

Again however, as the CI for the odds ratio includes 1 (no difference in odds), the difference between the results is not statistically significant.

Note also the OR of 1.45 (95% CI 0.93–2.27) for major bleeding events given in the abstract: major bleeds are 1.45 times more common among patients on anticoagulation compared to antiplatelet treatment, though again the 95% CI suggests that the results are not statistically significant.

The authors felt that there was uncertainty about the value of long term anticoagulation compared with antiplatelet treatment. They commented that the risks of bleeding made it an even less convincing treatment option.

The following extract is reproduced with permission from The BMJ Publishing Group

Correlation and regression

Priest, P., Yudkin, P., McNulty, C. and Mant, D. (2001). Antibacterial prescribing and antibacterial resistance in English general practice: cross sectional study. *BMJ* **323:** 1037–1041.

Abstract

Objective: To quantify the relation between community based antibacterial prescribing and antibacterial resistance in community acquired disease.

Design: Cross sectional study of antibacterial prescribing and antibacterial resistance of routine isolates within individual practices and primary care groups.

Setting: 405 general practices in south west and north west England.

Main outcome measures: Correlation between antibacterial prescribing and resistance for urinary coliforms.

Results: Antibacterial resistance in urinary coliform isolates is common but the correlation with prescribing rates was relatively low for individual practices (ampicillin and amoxicillin r_s=0.20, P=0.001). The regression coefficient was also low; a practice prescribing 20% less ampicillin and amoxicillin than average would have about 1% fewer resistant isolates.

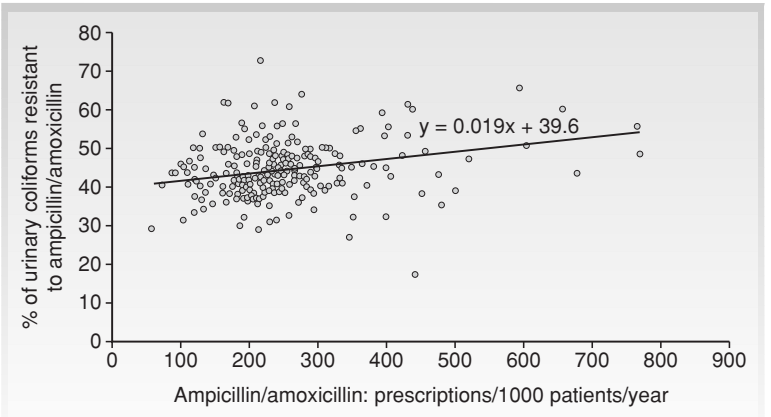

Fig. 16. Graph of regression of antibacterial resistance of urinary coliforms on prescribing at practice level

What statistical methods were used and why?

The researchers wanted to know whether there was a relationship between antibacterial prescribing and antibiotic resistance, so needed to measure the *correlation* between the two.

As the data were *skewed*, i.e. not *normally distributed*, they needed to use the *Spearman correlation coefficient* as a measure of the correlation.

They wished to predict how much a 20% reduction in antibiotic prescribing would be likely to reduce antibacterial resistance. To do this they needed to calculate the *regression coefficient*.

What do the results mean?

The Spearman correlation coefficient for penicillin prescribing and resistance for urinary coliforms is given as $r_s = 0.20$.

The *r* indicates that it is the correlation coefficient, and the small *s* behind it shows that it is a Spearman correlation coefficient.

The *r* value of 0.20 indicates a low correlation.

$P = 0.001$ indicates that an *r* value of 0.20 would have happened by chance 1 in 1000 times if there were really no correlation.

Fig. 16 shows the authors' *regression graph* comparing antibiotic prescribing with antibiotic resistance. Each dot represents one practice's results. The line through the dots is the *regression line*.

The authors show the *regression equation* in the graph: $y = 0.019x + 39.6$.

In this equation:

x (the horizontal axis of the graph) indicates the number of penicillin prescriptions per 1000 patients per year.

y (the vertical axis of the graph) indicates the percentage of urinary coliforms resistant to penicillin.

The *regression constant* of 39.6 is the point at which the regression line would hit the vertical axis of the graph (when number of prescriptions = 0).

The *regression coefficient* gives the slope of the line – the percentage of resistant bacteria reduces by 0.019 for every one less penicillin prescription per 1000 patients per year.

Using this value, the authors calculated that a practice prescribing 20% less penicillin than average would only have about 1% fewer resistant bacteria.

The authors therefore suggested that trying to reduce the overall level of antibiotic prescribing in UK general practice may not be the most effective strategy for reducing antibiotic resistance in the community.

The following extract is reproduced with permission from The BMJ Publishing Group

Survival analysis and risk reduction

Bosch, J., Yusuf, S., Pogue, J., Sleight, P., Lonn, E. et al. (2002). Use of ramipril in preventing stroke: double blind randomised trial. *BMJ* **324:** 699–702.

Abstract

Objective: To determine the effect of the angiotensin converting enzyme inhibitor ramipril on the secondary prevention of stroke.

Design: Randomised controlled trial.

Setting: 267 hospitals in 19 countries.

Participants: 9297 patients with vascular disease or diabetes plus an additional risk factor, followed for 4.5 years as part of the HOPE study.

Outcome measures: Stroke (confirmed by computed tomography or magnetic resonance imaging when available), transient ischaemic attack, and cognitive function. Blood pressure was recorded at entry to the study, after 2 years, and at the end of the study.

Results: Reduction in blood pressure was modest (3.8 mmHg systolic and 2.8 mmHg diastolic). The relative risk of any stroke was reduced by 31% (156/4635 v 226/4652) in the ramipril group compared with the placebo group, and the relative risk of fatal stroke was reduced by 61% (17/4645 v 44/4652).

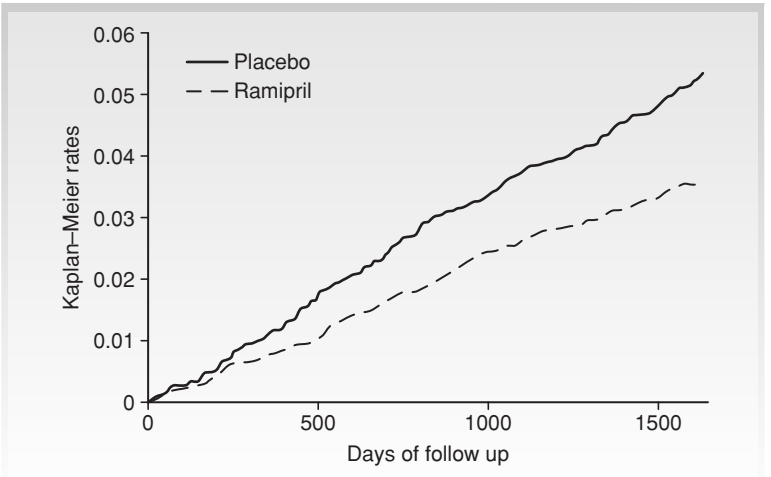

Fig. 17. Kaplan–Meier estimates of the development of stroke by treatment group.

What statistical methods were used and why?

The authors wanted to find out whether ramipril prevented stroke in high-risk patients.

The *null hypothesis* was that there was no difference between the two groups of patients.

The *Kaplan–Meier plot (Fig. 17)* was used to give a visual representation of the time until having a stroke in each group. The survival between these two groups was compared using the *log rank test*.

Because the outcome was time until a stroke, the *Cox regression model* was used to estimate the *RRR*. The significance of the RRR was indicated by its *95% CI*. The likelihood there was no real difference between the groups was given by the *P value*.

What do the results mean?

The authors found that 156 out of 4645 patients (3.36%) in the ramipril group had a stroke. In the placebo group, 226 of 4652 patients (4.86%) had a stroke.

The risk ratio was therefore

$$\frac{3.36}{4.86} = 0.69$$

For this risk ratio, the 95% CI was 0.56–0.84. Its *P* value of 0.0002 means that such a result would have happened by chance only two times in 10 000, so the result is very highly significant.

The RRR is given by

$$\frac{4.86 - 3.36}{4.86} = \frac{1.50}{4.86} = 0.31 = 31\%$$

From the results we can also calculate the ARR and NNT. The risk has reduced from 4.86 to 3.36, so

$$\text{ARR} = 4.86 - 3.36 = 1.5\%$$

$$\text{NNT} = \frac{100}{\text{ARR}} = \frac{100}{1.5} = 67$$

Seventeen (0.37%) patients in the ramipril group had a fatal stroke compared with 44 (0.95%) in the placebo group (relative risk 0.39, 95% CI 0.22–0.67), giving a RRR of

$$\frac{0.95 - 0.37}{0.95} = \frac{0.58}{0.95} = 0.61 = 61\%$$

The Kaplan–Meier survival plot shows how the trend towards reduction in stroke risk with ramipril continued for the length of the study.

This is an example where the RRR and NNT from the same study can have opposing effects on prescribing habits. The authors' conclusion that ramipril reduced the incidence of stroke in patients at high risk by 31%, despite a modest reduction in blood pressure, sounded attractive. However, some clinicians commented that the need to treat 67 patients for 4½ years to prevent one stroke was less convincing: for every stroke prevented, 66 patients had unnecessary treatment for 4½ years.

The following extract is reproduced with permission from The BMJ Publishing Group

Sensitivity, specificity and predictive values

Herren, K., Mackway-Jones, K., Richards, C., Seneviratne, C., France, M. and Cotter, L. (2001). Is it possible to exclude a diagnosis of myocardial damage within six hours of admission to an emergency department? Diagnostic cohort study. *BMJ* **323:** 372–374.

Abstract

Objective: To assess the clinical efficacy and accuracy of an emergency department based six hour rule-out protocol for myocardial damage.

Design: Diagnostic cohort study.

Setting: Emergency department of an inner city university hospital.

Participants: 383 consecutive patients aged over 25 years with chest pain of less than 12 hours' duration who were at low to moderate risk of acute myocardial infarction.

Intervention: Serial measurements of creatine kinase MB mass and continuous ST segment monitoring for six hours with 12 leads.

Main outcome measure: Performance of the diagnostic test against a gold standard consisting of either a 48 hour measurement of troponin T concentration or screening for myocardial infarction according to the World Health Organization's criteria.

Results: Outcome of the gold standard test was available for 292 patients. On the diagnostic test for the protocol, 53 patients had positive results and 239 patients had negative results.

There were 18 false positive results and one false negative result. Sensitivity was 97.2%, specificity 93.0%, the negative predictive value 99.6%, and the positive predictive value 66.0%. The likelihood ratio was 13.8.

Table 11. Two-way table of number of patients categorised as positive or negative by rule-out protocol and by gold standard

		Outcome of gold standard test		Total
		Positive	Negative	
Result in chest pain	Positive	35	18	53
assessment unit	Negative	1	238	239
Total		36	256	292

What statistical methods were used and why?

The researchers wanted to know whether they could use a diagnostic test to rule-out a myocardial infarction (MI) in patients with cardiac type chest pain. Their test is summarized in the "intervention" section of the abstract.

They decided whether patients had actually had an MI by using a "gold standard" test, given in the "main outcome measure" section.

They used a *two-way table* to show the results, shown in *Table 11*.

Sensitivity, *specificity*, *predictive values* and *likelihood ratios* were used to give the value of their diagnostic test.

What do the results mean?

In a perfect test, the sensitivity, specificity and predictive values would each have a value of 1. The lower the value (the nearer to zero), the less useful the test is in that respect.

Sensitivity shows the rate of pick-up of the disease – if a patient has had an MI, how often the test will be positive.

This is calculated from: $\frac{35}{36} = 0.972$, or 97.2%

Specificity is the rate at which the test can exclude an MI – if the patient is in fact healthy, how often the test will be negative.

This is given by: $\frac{238}{256} = 0.93$, or 93%

PPV is the likelihood that the patient has had an MI if the test result is positive.

$$\frac{35}{53} = 0.66, \text{ or } 66\%$$

So, if the test is positive, 34 in 100 patients will not actually have had an MI.

NPV shows the likelihood that the patient will be healthy if the test result is negative.

$$\frac{238}{239} = 0.996, \text{ or } 99.6\%$$

So, if the test is negative, only 4 in 1000 patients will actually have had an MI.

LR is the likelihood that the test result would be expected in a patient with an MI, compared to the likelihood that that same result would be expected in a healthy patient.

$$\text{LR} = \frac{\text{sensitivity}}{(1 - \text{specificity})} = \frac{0.972}{1 - 0.93} = \frac{0.972}{0.07} = 13.8$$

So, if the test is positive in a patient, that patient is almost 14 times as likely to have had an MI than not have had one.

The authors concluded that the chest pain rule-out protocol for MI had a high enough NPV to allow patients with a negative result to be sent home safely.

GLOSSARY

Cross-references to other parts of the glossary are given in *italics*.

Absolute risk reduction, ARR

The proportion by which an intervention reduces the *risk* of an event. Compare this with r*elative risk reduction*. See page 43.

Alpha, α

The alpha value is equivalent to the *P value* and should be interpreted in the same way.

ANCOVA (ANalysis of COVAriance)

Analysis of covariance is an extension of *analysis of variance* (ANOVA) to allow for the inclusion of *continuous variables* in the model. See page 29.

ANOVA (ANalysis Of VAriance)

This is a group of statistical techniques used to compare the *means* of two of more samples to see whether they come from the same population. See page 29.

Association

A word used to describe a relationship between two *variables*.

Beta, β

The beta value is the probability of accepting a *hypothesis* that is actually false. 1 – β is known as the *power* of the study. See page 107.

Bayesian statistics

An alternative way of analyzing data, it creates and combines numerical values for prior belief, existing data and new data. See page 72.

Binary variable

See *categorical variable* below.

Bi-modal distribution

Where there are 2 *modes* in a set of data, it is said to be bi-modal. See page 15.

Binomial distribution

When data can only take one of two values (for instance male or female), it is said to follow a binomial *distribution*.

Bonferroni

A method that allows for the problems associated with making multiple comparisons. See page 69.

Box and whisker plot

A graph showing the *median*, *range* and *inter-quartile range* of a set of values. See page 13.

Case–control study

A retrospective study which investigates the relationship between a risk factor and one or more outcomes. This is done by selecting patients who already have the disease or outcome *(cases),* matching them to patients who do not (controls) and then comparing the effect of the risk factor on the two groups. Compare this with *Cohort study*. See page 40.

Cases

This usually refers to patients but could refer to hospitals, wards, counties, blood samples etc.

Categorical variable

A *variable* whose values represent different categories of the same feature. Examples include different blood groups, different eye colours, and different ethnic groups.

When the variable has only two categories, it is termed "binary" (e.g. gender). Where there is some inherent ordering (e.g. mild, moderate, severe), this is called an "*ordinal*" variable.

Causation

The direct relationship of the cause to the effect that it produces, usually established in experimental studies.

Censored

A censored observation is one where we do not have information for all of the observation period. This is usually seen in *survival analysis* where patients are

followed for some time and then move away or withdraw consent for inclusion in the study. We cannot include them in the analysis after this point as we do not know what has happened to them. See page 57.

Central tendency

The "central" scores in a set of figures. *Mean*, *median* and *mode* are measures of central tendency.

Chi-squared test, χ^2

The chi-squared test is a test of association between two *categorical variables*. See page 34.

Cohort Study

A prospective, observational study that follows a group (cohort) over a period of time and investigates the effect of a treatment or risk factor. Compare this with *case–control study*. See page 37.

Confidence interval, CI

A range of values within which we are fairly confident the true *population* value lies. For example, a 95% CI means that we can be 95% confident that the population value lies within those limits. See page 20.

Confounding

A confounding factor is the effect of a *covariate* or factor that cannot be separated out. For example, if women with a certain condition received a new treatment and men received placebo, it would not be possible to separate the treatment effect from the

effect due to gender. Therefore gender would be a confounding factor.

Continuous variable

A *variable* which can take any value within a given range, for instance BP. Compare this with *discrete variable*.

Correlation

When there is a linear relationship between two *variables* there is said to be a correlation between them. Examples are height and weight in children, or socio-economic class and mortality.

Measured on a scale from −1 (perfect negative correlation), through 0 (no relationship between the variables at all), to +1 (perfect positive correlation). See page 48.

Correlation coefficient

A measure of the strength of the linear relationship between two *variables*. See page 48.

Covariate

A covariate is a *continuous variable* that is not of primary interest but is measured because it may affect the outcome and may therefore need to be included in the analysis.

Cox regression model

A method which explores the effects of different *variables* on survival. See page 60.

Database

A collection of records that is organized for ease and speed of retrieval.

Degrees of freedom, df

The number of degrees of freedom, often abbreviated to df, is the number of independent pieces of information available for the statistician to make the calculations.

Descriptive statistics

Descriptive statistics are those which describe the *data* in a *sample*. They include *means*, *medians*, *standard deviations*, *quartiles* and *histograms*. They are designed to give the reader an understanding of the data. Compare this with *inferential statistics*.

Discrete variable

A *variable* where the data can only be certain values, usually whole numbers, for example the number of children in families. Compare this with *continuous variable*.

Distribution

A distinct pattern of data may be considered as following a distribution. Many patterns of data have been described, the most useful of which is the *normal* distribution. See page 9.

Fisher's exact test

Fisher's exact test is an accurate test for association between *categorical variables*. See page 35.

Fields

See *variables* below.

Hazard ratio, HR

The HR is the ratio of the hazard (chance of something harmful happening) of an event in one group of observations divided by the hazard of an event in a different group. A HR of 1 implies no difference in risk between the two groups, a HR of 2 implies double the risk. The HR should be stated with its *confidence intervals*. See page 61.

Histogram

A graph of *continuous* data with the data categorized into a number of classes. See example on page 15.

Hypothesis

A statement which can be tested that predicts the relationship between *variables*.

Incidence

The rate or proportion of a group developing a condition within a given period.

Inferential statistics

All statistical methods which test something are inferential. They estimate whether the results suggest that there is a real difference in the *populations*. Compare this with *descriptive statistics*.

Intention to treat

An intention to treat analysis is one in which patients are included in the analysis according to the group into which they were randomized, even if they were withdrawn from the study and did not actually receive the treatment.

Interaction

An interaction is when two or more *variables* are related to one another and therefore not acting independently.

Inter-quartile range, IQR

A measure of spread given by the difference between the first *quartile* (the value below which 25% of the cases lie) and the third quartile (the value below which 75% of the cases lie). See page 13.

Intra-class correlation coefficient

This correlation measures the *level of agreement* between two *continuous variables*. It is commonly used to look at how accurately a test can be repeated, for instance by different people. See page 68.

Kaplan–Meier survival plot

Kaplan–Meier plots are a method of graphically displaying the survival of a sample *cohort* of which the survival estimates are re-calculated whenever there is a death. See page 58.

Kappa, κ

This is a measure of the *level of agreement* between two *categorical* measures. It is often used to look at how accurately a test can be repeated, for instance by different people. See page 67.

Kolmogorov Smirnov test

Kolmogorov Smirnov tests the hypothesis that the collected data are from a *normal distribution*. It is therefore used to assess whether *parametric* statistics can be used. See page 29.

Kruskal Wallis test

This is a *non-parametric* test which compares two or more independent groups. See page 33.

Level of agreement

A comparison of how well people or tests agree. See page 67.

Life table

A table of the proportion of patients surviving over time. It is used in survival analysis. See page 57.

Likelihood ratio, LR

This is the likelihood that a test result would be expected in patients with a condition, divided by the likelihood that that same result would be expected in patients without that condition. See page 65.

Log rank test

A *non-parametric* test used for the comparison of survival estimates using *Kaplan–Meier* or *life table* estimates. See page 59.

Logistic regression

Logistic regression is a variation of *linear regression* that is used when there are only two possible outcomes. See page 55.

Mann-Whitney U test

A *non-parametric* test to see whether there is a significant difference between two sets of data that have come from two different sets of *subjects*. See page 31.

Mantel Haenszel test

An extension of the *chi-squared* test to compare several two-way tables. This technique can be applied in *meta-analysis*. See page 36.

Mean

The sum of the observed values divided by the number of observations. Compare this with *median* and *mode*. See page 9.

Median

The middle observation when the observed values are ranked from smallest to largest. Compare this with *mean* and *mode*. See page 12.

Meta-analysis

Meta-analysis is a method of combining results from a number of independent studies to give one overall estimate of effect. See page 21 for an example.

Mode

The most commonly occurring observed value. Compare this with *mean* and *median*. See page 14.

Negative predictive value (NPV)

If a diagnostic test is negative, the NPV is the chance that a patient does not have the condition. Compare this with *positive predictive value*. See page 64.

Nominal data

Data that can be placed in named categories that have no particular order, for example eye colour.

Non-parametric test

A test which is not dependent on the *distribution* (shape) of the data. See page 31.

Normal distribution

This refers to a *distribution* of data that is symmetrical. In a graph it forms a characteristic bell shape. See page 9.

Null hypothesis

A hypothesis that there is no difference between the groups being tested. The result of the test either supports or rejects that hypothesis.

Paradoxically, the null hypothesis is usually the opposite of what we are actually interested in finding out. If we want to know whether there is a difference between two treatments, then the null hypothesis would be that there is no difference. The statistical test would be used to try to disprove this. See page 27.

Number needed to harm, NNH

NNH is the number of patients that need to be treated for one to be harmed by the treatment. See page 46.

Number needed to treat, NNT

NNT is the number of patients that need to be treated for one to get benefit. See page 43.

Odds

The ratio of the number of times an event happens to the number of time it does not happen in a group of patients. Odds and *risk* give similar values when considering rare events (e.g. winning the lottery), but may be substantially different for common events (e.g. not winning the lottery!). See page 40.

Odds ratio (OR)

The *odds* of an event happening in one group, divided by the odds of it happening in another group. See page 40.

One-tailed test

A test where the *null hypothesis* can only be rejected in one direction, for example if new treatment is

worse than current treatment but not if it is better. It should only rarely be used. Compare this with a *two-tailed test*. See page 69.

Ordinal data

Data that can be allocated to categories that can be "ordered", e.g. from least to strongest. An example is the staging of malignancy.

P value

Usually used to test a *null hypothesis*, the *P* value gives the probability of any observed differences having happened by chance. See page 24.

Parametric test

Any test that has an assumption that the data needs to follow a certain distribution can be considered to be a parametric test. The most common distribution that the data need to follow is the *normal distribution*. Examples are the *t test* and *ANOVA*. See page 28.

Pearson correlation coefficient

A method of calculating a correlation coefficient if the values are sampled from a *normal* population. See page 51.

Percentage

The number of items in a category, divided by the total number in the group, then multiplied by 100. See page 7.

Poisson distribution

This *distribution* represents the number of events happening in a fixed time interval, for instance the number of deaths in a year.

Poisson regression

A variation of *regression* calculations which allows for the frequency of rare events. See page 56.

Population

The complete set of subjects from which a sample is drawn.

Positive predictive value, PPV

If a diagnostic test is positive, the PPV is the chance that a patient has the condition. Compare this with *negative predictive value*. See page 63.

Power

The power of a study is the probability that it will detect a statistically *significant* difference. See page 71.

Prevalence

The proportion of a group with a condition at a single point in time. See page 70.

Quartiles

Quartiles may be given with the *median*. The 1st quartile point has ¼ of the data below it, the 3rd quartile has ¾ of the data below it. See page 13.

r

Where there is a linear relationship between two variables there is said to be a *correlation* between them. The *correlation coefficient* gives the strength of that relationship. See page 48.

R^2

An estimate of the amount of the variation in the data that is being explained by a *regression* model. See page 55.

Range

The difference between the maximum and minimum score in a set of figures.

Rank

A numerical value given to an observation showing its relative order in a set of data.

Rate

The number of times that an event happens in a fixed period of time.

Regression

Regression analysis is a technique for finding the relationship between two *variables*, one of which is dependent on the other. See page 53.

Relative risk

Risk ratio is often referred to as *relative risk*. However, *odds ratios* are also a measure of relative risk.

Relative risk reduction, RRR

The proportion by which an intervention reduces the *risk* of an event. Compare this with a*bsolute risk reduction*. See page 43.

Risk

The probability of occurrence of an event. Calculated by dividing the number of events by the number of people at risk. See page 37.

Risk ratio, RR

The *risk* of an event happening in one group, divided by the risk of it happening in another group. See page 37.

Sample

A small group drawn from a larger population.

Sensitivity

This is the rate of pick-up of a condition in a test. In other words, the proportion of patients with the condition having a positive test result. See page 63. Compare this with *specificity.*

Significance

The probability of getting the results if the *null hypothesis* is true. See page 25.

Spearman correlation coefficient

An estimate of *correlation* used for *non-parametric variables*. See page 51.

Specificity

The rate of elimination of the possibility of disease by a test. In other words, the proportion of patients without the condition that has a negative test result. See page 63.

Skewed data

A lack of symmetry in the *distribution* of data. See page 12.

Standard deviation, SD

A measure of the spread of scores away from the *mean*. See page 16.

Standard error of the mean

A measure of how close the *sample mean* is likely to be to the *population* mean.

Stratified

A stratified *sample* is one that has been split into a number of subgroups.

Student's *t* test

See *t test*.

Subjects

The *sample* in a study.

t test (also known as Student's *t* test)

The *t* test is a *parametric* test used to compare the *means* of two groups. See page 28.

Transformation

A transformation is where a mathematical formula is used to change the data. This will often be done to try to make the data follow a *normal distribution* so that a *parametric* test can be used. See page 30.

Two-tailed test

A test where the *null hypothesis* can be rejected whether the new treatment is better, or worse, than the current treatment. Compare this with a *one-tailed test*. See page 69.

Type I and II errors

Any statistical test can fail in two ways. A *hypothesis* that is correct can be rejected (type I error), or a hypothesis that is incorrect can be accepted (type II error). The chance of making a type I error is the same as the *P value*.

Variable

Any characteristic that differs from subject to subject or time to time. In data analysis, variables may be called *fields* and refer to all the things recorded on the *cases*.

Variance

A measure of the spread of scores away from the mean. It is the square of the *standard deviation*.

Wilcoxon signed rank test

A *non-parametric* test for comparing the difference between paired groups, for instance before and after treatment. See page 33.

χ^2 test

The chi-squared test is a test of association between two *categorical variables*. See page 34.

Yates' continuity correction

This is an adjustment to the *chi-squared* test to improve the accuracy of the *P value*. See page 36.

INDEX

(Page numbers in *italics* refer to the glossary)

2-way table 35, 63, 92
Absolute risk reduction, ARR 43, 89, *94*
Alpha value, α *94*
ANCOVA *94*
ANOVA 29, *94*
Arithmetic mean 9
Association *94*
Average 9

Beta, β *95*
Bayesian statistics 72, *95*
Best fit 54
Binary variable *95*
Bi-modal distribution 15, *95*
Binomial distribution *95*
Blobbogram 22
Bonferroni 69, *95*
Box and whisker plot 13

Case–control study 40, *96*
Cases 40, *96*
Categorical variable *96*
Causation *96*
Censored 57, *96*
Central tendency *97*
Chi-squared test, χ^2 34, 76, *97*
Cohort study 37, 75, *97*
Confidence Interval, CI 20, 75, 80, *97*
Confounding *97*
Contingency table 35
Continuous variable 68, *98*
Correlation 48, 56, 84, *98*
Correlation coefficient 48, *98*
Covariate *98*
Cox regression model 60, 88, *98*

Database *99*
Degrees of freedom, df *99*
Descriptive statistics *99*
Discrete variable *99*
Distribution *99*

Exam tips 2, 11, 19, 23, 27, 47, 66, 71

False negative 63
False positive 63
Fisher's exact test 35, *99*
Fields *111*
Forest plot 22
Frequentist statistics 72
Friedman test 33

Hazard ratio 61, *100*
Histogram *100*
Hypothesis *100*

Incidence 70, 75, *100*
Inferential statistics *100*
Intention to treat *101*
Interaction *101*
Inter-quartile range, IQR 13, *101*
Intra-class correlation coefficient 68, *101*

Kaplan-Meier plots 58, 88, *101*
Kappa, κ 67, *102*
Kolmogorov Smirnov 29, *102*
Kruskal Wallis test 33, *102*

Level of agreement 67, *102*
Life table 57, *102*
Likelihood ratio, LR 65, 93, *102*
Linear regression 55
Log rank test 59, 88, *103*
Logistic regression 55, *103*

Mann–Whitney U test 31, *103*
Mantel Haenszel test 36, *103*
Mean 9, 75, *103*
Median 12, *103*
Meta-analysis 21, *104*
Mode 14, *104*

Negative predictive value, NPV 64, 93, *104*
Nominal data *104*
Non-parametric test 31, *104*
Normal distribution 9, 28, 75, *104*
Null hypothesis 27, 75, 88, *104*
Number needed to harm, NNH 46, *105*
Number needed to treat, NNT 43, 89, *105*

Odds 40, *105*
Odds ratio, OR 40, 80, *105*
One-tailed test 69, *105*
Ordinal data 68, *106*

P value, *P* 24, 76, 88, *106*
Parametric test 28, *106*
Pearson correlation coefficient 51, *106*
Percentage 7, 75, *106*
Poisson distribution *107*
Poisson regression 56, *107*
Population *107*
Positive predictive value, PPV 63, 93, *107*
Posterior distribution 72
Power 71, *107*
Prevalence 70, *107*
Prior distribution 72
Proportional hazards survival model 60

Quartiles 13, *107*

r 48, 84, *108*
R^2 51, 55, *108*
Range *108*
Rank 31, *108*
Rate *108*
Regression 53, 84, *108*
Regression coefficient 53, 85
Regression constant 54, 85
Relative risk, RR 37, 75, 88, *108*
Relative risk reduction, RRR 43, 88, *109*
Risk 37, *109*
Risk ratio, RR 37, *109*

Sample *109*
Scatter plot 50
Sensitivity 63, 92, *109*
Significance 25, *109*
Spearman correlation coefficient 51, 84, *109*
Specificity 63, 92, *110*
Skewed data 10, 12, 32, 84, *110*
Standard deviation 16, 22, 75, *110*
Standard error of the mean *110*
Stratified *110*
Student's *t* test 29, *110*
Study advice 2
Subjects *110*
Survival analysis 57
Survival plot 58

t test 28, *111*
Transformation 30, *111*
True value 20
Two-tailed test 69, *111*
Two-way table 35, 63, 92
Type I and II errors *111*

U value 33

Variable(s) *111*
Variance *111*

Wilcoxon signed rank test 33, *112*

χ^2 34, 76, *97, 112*

Yates' continuity correction 36, *112*