# Statistics and Epidemiology 



- the core questions and answers you need

O the key concepts you should understand
O the key facts you should know
in one concise integrated revision guide

$$
\begin{array}{r}
\text { Emily Ferenczi \& Nina Muirhead } \\
\text { Editorial Advisor: Lucy Carpenter }
\end{array}
$$

## ONE STOP DOC

 Statistics and Epidemiology
## One Stop Doc

Titles in the series include:
Cardiovascular System - Jonathan Aron
Editorial Advisor - Jeremy Ward
Cell and Molecular Biology - Desikan Rangarajan and David Shaw
Editorial Advisor - Barbara Moreland
Endocrine and Reproductive Systems - Caroline Jewels and Alexandra Tillett Editorial Advisor - Stuart Milligan

Gastrointestinal System - Miruna Canagaratnam
Editorial Advisor - Richard Naftalin
Musculoskeletal System - Wayne Lam, Bassel Zebian and Rishi Aggarwal
Editorial Advisor - Alistair Hunter
Nervous System - Elliott Smock
Editorial Advisor - Clive Coen
Metabolism and Nutrition - Miruna Canagaratnam and David Shaw Editorial Advisors - Barbara Moreland and Richard Naftalin

Respiratory System - Jo Dartnell and Michelle Ramsay
Editorial Advisor - John Rees
Renal and Urinary System and Electrolyte Balance - Panos Stamoulos and Spyridon Bakalis Editorial Advisors - Alistair Hunter and Richard Naftalin

Gastroenterology and Renal Medicine - Reena Popat and Danielle Adebayo Editorial Advisor - Steve Pereira

## Coming soon:

Cardiology - Rishi Aggarwal, Nina Muirhead and Emily Ferenczi Editorial Advisor - Darrell Francis

Respiratory Medicine - Rameen Shakur and Ashraf Khan Editorial Advisors - Nikhil Hirani and John Simpson

Immunology - Stephen Boag and Amy Sadler Editorial Advisor - John Stewart

## ONE STOP DOC

## Statistics and Epidemiology

## Emily Ferenczi $\mathrm{BA}($ (Cantab)

Sixth Year Medical Student, Oxford University Clinical School, Oxford, UK
Nina Muirhead ba(0xon)
Sixth Year Medical Student, Oxford University Clinical School, Oxford, UK
Editorial Advisor: Lucy Carpenter bA msc PhD
Reader in Statistical Epidemiology, Department of Public Health, Oxford University, Oxford, UK
Series Editor: Elliott Smock MBBS BSc(Hons)
House Officer (FY1), Eastbourne District General Hospital, Eastbourne, UK

## Hodder Arnold

First published in Great Britain in 2006 by
Hodder Arnold, an imprint of Hodder Education and a member of the Hodder Headline Group, 338 Euston Road, London NW1 3BH

## http://www.hoddereducation.com

Distributed in the United States of America by
Oxford University Press Inc.,
198 Madison Avenue, New York, NY10016
Oxford is a registered trademark of Oxford University Press
(c) 2006 Edward Arnold (Publishers) Ltd

All rights reserved. Apart from any use permitted under UK copyright law, this publication may only be reproduced, stored or transmitted, in any form, or by any means with prior permission in writing of the publishers or in the case of reprographic production in accordance with the terms of licences issued by the Copyright Licensing Agency. In the United Kingdom such licences are issued by the Copyright Licensing Agency: 90 Tottenham Court Road, London W1T 4LP.

Whilst the advice and information in this book are believed to be true and accurate at the date of going to press, neither the author[s] nor the publisher can accept any legal responsibility or liability for any errors or omissions that may be made. In particular, (but without limiting the generality of the preceding disclaimer) every effort has been made to check drug dosages; however it is still possible that errors have been missed. Furthermore, dosage schedules are constantly being revised and new side-effects recognized. For these reasons the reader is strongly urged to consult the drug companies' printed instructions before administering any of the drugs recommended in this book.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library
Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress
ISBN-10 034092554 X
ISBN-13 9780340925546

## 12345678910

Commissioning Editor: Christina De Bono
Project Editor: Clare Weber, Jane Tod
Production Controller: Lindsay Smith
Cover Design: Amina Dudhia
Indexer: Jane Gilbert, Indexing Specialists (UK) Ltd
Typeset in 10/12pt Adobe Garamond/Akzidenz GroteskBE by Servis Filmsetting Ltd, Manchester Printed and bound in Spain

Hodder Headline's policy is to use papers that are natural, renewable and recyclable products and made from wood grown in sustainable forests. The logging and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

## CONTENTS

PREFACE ..... vi
ABBREVIATIONS ..... vii
PART 1 EPIDEMIOLOGY
SECTION 1 STUDYING HEALTH AND DISEASE IN POPULATIONS ..... 3
SECTION 2 OBSERVATIONAL STUDIES: ECOLOGICAL STUDIES ..... 11
SECTION 3 OBSERVATIONAL STUDIES: CROSS-SECTIONAL STUDIES ..... 17
SECTION 4 OBSERVATIONAL STUDIES: CASE-CONTROL STUDIES ..... 25
SECTION 5 OBSERVATIONAL STUDIES: COHORT STUDIES ..... 33
SECTION 6 INTERVENTION STUDIES: RANDOMIZED CONTROLLED TRIALS ..... 43
SECTION 7 META-ANALYSIS ..... 51
SECTION 8 CLINICAL EPIDEMIOLOGY ..... 59
PART 2 STATISTICAL TOOLKIT
SECTION 9 DESCRIBING DATA ..... 73
SECTION 10 ESTIMATION ..... 83
SECTION 11 HYPOTHESIS TESTING ..... 91
SECTION 12 INTERPRETATION OF DATA ..... 99
SECTION 13 SOURCES OF ERROR ..... 111
APPENDIX ..... 119
INDEX ..... 127

## From the Series Editor, Elliott Smock

Are you ready to face your looming exams? If you have done loads of work, then congratulations; we hope this opportunity to practise SAQs, EMQs, MCQs and Problem-based Questions on every part of the core curriculum will help you consolidate what you've learnt and improve your exam technique. If you don't feel ready, don't panic - the One Stop Doc series has all the answers you need to catch up and pass.

There are only a limited number of questions an examiner can throw at a beleaguered student and this text can turn that to your advantage. By getting straight into the heart of the core questions that come up year after year and by giving you the model answers you need, this book will arm you with the knowledge to succeed in your exams. Broken down into logical sections, you can learn all the important facts you need to pass without having to wade through tons of different textbooks when you simply don't have the time. All questions presented here are 'core'; those of the highest importance have been highlighted to allow even sharper focus if time for revision is running out. In addition, to allow you to organize your revision efficiently, questions have been grouped by topic, with answers supported by detailed integrated explanations.

On behalf of all the One Stop Doc authors I wish you the very best of luck in your exams and hope these books serve you well!

From the Authors, Emily Ferenczi and Nina Muirhead

In our first year of medical school, we remember groaning at the thought of having a statistics lecture. It all seemed so irrelevant and abstract at the time. However, after several years of essays, critical reviews and projects, we have come to appreciate the value of statistics. So much so in fact, that we were inspired to write a book about it! In the hospital, hearing doctors talk to patients about the evidence they have for offering one particular treatment over another, we realised that 'evidence-based medicine' is not just a fantasy, but a real and important aspect of the way we should approach medical practice throughout our careers.

In this book, we have used examples from recent medical literature to provide both inspiration and practical examples of the way statistics and epidemiological methods are used in clinical studies to guide clinical practice. The aim of this book is to equip medical students with an understanding and a tool guide for reading and reviewing clinical studies so that, as practising doctors, they can arrive at valid conclusions and make justifiable clinical decisions based upon the available evidence. It also aims to provide a basis by which a medical student or junior doctor can learn about starting a clinical study and how to access the information and resources that they need.

We have chosen published studies to illustrate important epidemiological and statistical concepts. Please bear in mind that the studies are chosen on the basis of their ability to demonstrate key issues that arise when analysing different study designs, not necessarily on the basis of their quality.

We would like to thank Adrian Smith for his very helpful comments on the draft document.

## ABBREVIATIONS

| ANOVA | analysis of variance | MI |
| :--- | :--- | :--- |
| BMI | body mass index | MMR |
| CFTR | cystic fibrosis transmembrane <br>  <br> conductance regulator | NHS |
| CI | confidence interval | NNT |
| df | degrees of freedom | NPV |
| FEV | forced expiratory volume | OR |
| FN | false negative | PPV |
| FP | false positive | PSA |
| F/T PSA | free-to-total prostate-specific antigen | RSI |
| GP | SD |  |
| H $_{0}$ | null hyporal practitioner | SE |
| $H_{1}$ | alternative hypothesis | SEM |
| HbA $_{1 c}$ | haemoglobin A | SE $(p)$ |
| HIV | human immunodeficiency virus | SSRI |
| MHRA | Medicines and Healthcare products | TN |
|  | Regulatory Agency |  |

myocardial infarction
measles, mumps and rubella
National Health Service
numbers needed to treat
negative predictive value odds ratio
positive predictive value prostate-specific antigen repetitive strain injury standard deviation standard error standard error of the mean standard error of the proportion selective serotonin reuptake inhibitor true negative
true positive

This page intentionally left blank


This page intentionally left blank

- PRINCIPLES OF EPIDEMIOLOGY
- MEASURING DISEASE6
- MEASURING ASSOCIATIONS 8


1. What is the definition of epidemiology and what are its uses?
2. What is meant by the following terms and how do they differ from each other?
a. The distribution of disease
b. The determinants of disease
3. Which type of information would provide you with an idea of the distribution of the disease in developing versus developed countries?
a. A case-control study
b. A randomized controlled trial
c. The National Infant Mortality Register
d. The National Cancer Register
e. An ecological study into the correlation between infectious disease rate and population density
4. Which type of information would help you to understand the determinants of breast cancer?
a. A case-control study investigating the correlation between use of hormonal contraception and the risk of breast cancer
b. The National Cancer Register
c. A cohort study into the incidence of breast cancer in two groups of women: in one group, all the women have a family history of breast cancer, in the second group, there is no family history
d. An ecological study comparing the rates of breast cancer in the UK and in France
e. A randomized controlled trial investigating the efficacy of a new drug treatment for breast cancer
5. Identify the numerators and denominators in the following scenarios
a. In a school of 670 children, 380 eat lunch in the canteen, 8 children have been identified as having gastroenteritis as a result of one of the canteen's chicken dishes
b. A country has a population of 20 million people. Of these, 10 million live in highly polluted cities. 450000 have been diagnosed with pollution-induced asthma
c. A study wants to investigate the association between smoking and infertility using data on couples. There are 340 couples enrolled in a fertility clinic; 120 couples are defined as smokers (one or both partners smoke)

## EXPLANATION: PRINCIPLES OF EPIDEMIOLOGY

Epidemiology is the quantitative study of the distribution and determinants of health and disease in a population (1).

Analytic epidemiological studies typically involve four components: the definition of disease and identification of the 'at risk' population; the measurement of disease; the measurement of exposure and the examination of the association between disease and exposure.

Understanding of the distribution and determinants of health problems in populations can help direct public health strategies, for the prevention and treatment of disease, to improve the health of a population. It can ensure that money is spent in the right way on the people who are at risk (1).

Any epidemiological parameter requires two numbers: a numerator, such as the number of individuals who have been defined as having a disease and a denominator, the defined population from which these individuals have been taken. Information on both the numerator and the denominator is crucial in epidemiology.

To illustrate: ' 10 people have been diagnosed with skin cancer in one month' - this figure is meaningless if one does not know from what size population these 10 people have been identified. A population of only 20 individuals may raise more concern than a population of 20 million individuals.

Different populations and subgroups of populations are affected by different health problems to different extents. Information about the distribution of health problems can be obtained through routinely collected data such as censuses and registers for death and disease, and through cross-sectional prevalence surveys (2a).

Establishing the determinants of health and disease is based upon identifying the association between an individual's exposure to specific risk or protective factors and the subsequent health outcome for that individual (2b). Ecological studies investigate exposure and disease at the level of population groups, rather than at the level of the individual. Studies that record exposure and disease status of individuals within a population include:

- Cross-sectional studies, which measure disease exposure
- Case-control studies
- Cohort studies.

Epidemiological evidence provides an idea of the extent and burden of health problems in a population, and thus can be used to direct public health strategies and treatment programmes aimed at improving health and reducing disease. Studies that investigate the effects of an intervention on health status include randomized controlled clinical trials of individual communities.

## Answers

1. See explanation
2. See explanation
3. F F T T T
4. T F T T F
5. a - Numerator: 8 cases, denominator: 380 children at risk from canteen food, b - Numerator: 450000 cases, denominator: 10 million in 'at risk' population, c - Numerator: 120 couples who are smokers (cases), denominator: 340 infertile couples (population of interest)
6. Define prevalence and incidence

## 7. Calculate the prevalence of

a. Smoking in medical students: sample of 170 medical students, 38 smokers, 132 nonsmokers
b. Repetitive strain injury in secretaries: 340 secretaries, 65 with repetitive strain injury
8. Calculate the annual incidence of
a. Work-related injuries in a car factory: 680 workers, four injuries per month
b. Leukaemias in primary school children in a town near a nuclear energy plant, town population: 32000 ; number of primary school children: 5800; number of leukaemia cases reported per year: 46
9. Calculate the age-specific mortality rate for the over-65-year age group in England: midyear population for over 65 years is 9.2 million of which 30914 died in one year
10. The prevalence of a disease
a. Can only be calculated by a cohort study
b. Is the number of new cases per unit time in a defined population
c. Describes the balance between incidence, mortality and recovery
d. Can be standardized for age and sex
e. Can be used to compare the burden of a disease across different geographical areas

## EXPLANATION: MEASURING DISEASE

Disease occurrence can be measured in different ways and using different sources. One example is that of routinely collected data, i.e. data collected not for the specific purpose of conducting an epidemiological study, which can give estimates of the prevalence and incidence of a disease in a population.

## 1. Denominator data (defining 'at risk' populations): census

2. Numerator data (defining cases) - falls into several categories:

- Mortality: e.g. death registers and certificates
- Morbidity: e.g. NHS contact or disease registers
- Wider impact: e.g. cost to the NHS or days missed from work for health reasons.

3. Prevalence is the number of cases in a population at a single point in time divided by the total number of individuals in that population at the same point in time (6). Prevalence is often expressed as a percentage (\%) but for rarer diseases it can be expressed in larger population units such as per 1000 population or per 10000 population.

The prevalence of disease at any time is determined by the incidence of new disease, the duration of the disease and changes in the population at risk, e.g. births and deaths. Prevalence measures the overall disease burden in a population at a particular point in time.
4. Incidence measures the number of new cases occurring in a defined time period divided by the number in the population at risk of becoming a case (6). To estimate incidence, one needs:

- A defined population at risk of an event
- A defined time period
- The number of events occurring in that period.

Incidence is often considered by epidemiologists to be the most informative measure of disease occurence. It is expressed as the number of events per 1000 or per 100000 population. For example, it is easier to think in terms of 12 deaths per 1000 than 0.012 deaths per person. The denominator for incidence can be refined and measured using 'person-time', e.g. person-years at risk, and this measure is often called the 'incidence rate'.

[^0]11. Which of the study designs are being described in the examples?

## Options

A. Cohort study
C. Case-control study
E. Randomized controlled trial
B. Meta-analysis
D. Ecological study
F. Cross-sectional study

1. A group of Gulf War veterans is followed over the course of 10 years to determine the association between the exposure to life-threatening experiences and the risk of psychiatric disturbance
2. The prevalence of HIV is compared in two African countries, one with a national 'safesex' education programme in place and the other with no such programme
3. A group of patients with liver disease is questioned on its daily consumption of alcohol over the past year. Consumption rates are compared to those of a group of patients in the same hospital but without liver disease
4. Thirty women with breast cancer are given a new drug treatment and 30 similar women are given an existing treatment. Neither the doctors involved in the care of the women nor the women themselves are aware of which treatment they are taking. The women are followed over a period of five years and at the end the five-year survival rates for the two groups of women are calculated and compared
5. The prevalence of leukaemia in children living near power lines is compared with the prevalence in children living far away from power lines
6. All the existing evidence for the effectiveness of a new laparoscopic technique for resection of large bowel tumours from multiple different studies is collected together
7. Match the study designs below to the following scenarios (each option can be used once, more than once or not at all)

## Options

A. Cohort study
C. Case-control study
E. Randomized controlled trial
B. Meta-analysis
D. Ecological study
F. Cross-sectional study

1. A rare disease
2. A rare risk factor
3. More than one outcome
4. Multiple risk factors
5. The temporal relationship between a risk factor and a disease
6. To prove the effect of a new drug for asthma
7. To test the hypothesis that hypertension is a risk factor for cardiovascular disease
8. When time and money are limited

## EXPLANATION: MEASURING ASSOCIATIONS

Special studies are used to assess the effects of exposure to particular risk or protective factors on a particular health outcome, such as a disease of interest.

The choice of a particular study design may depend upon a number of factors such as the prevalence of the condition of interest, the frequency of the exposure of interest or the amount of time and money available.

The table below summarizes the key characteristics, uses and disadvantages of the main types of epidemiological study.

| Approach | Category | Type | Timing | Uses | Problems |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Observational | Ecological study | Study of groups <br> or populations <br> using routinely <br> collected data | Usually <br> retrospective | Data on <br> distribution of <br> disease across <br> population groups | No data about <br> individuals |
| Observational | Cross-sectional <br> study | Special health <br> survey of <br> individuals | One point in time | To measure <br> prevalence | No incidence |

## Answers

11. $1-\mathrm{A}, 2-\mathrm{D}, 3-\mathrm{C}, 4-\mathrm{E}, 5-\mathrm{F}, 6-\mathrm{B}$
12. $1-\mathrm{C}, 2-\mathrm{A}, 3-\mathrm{A}, 4-\mathrm{C}, 5-\mathrm{AE}, 6-\mathrm{BE}, 7-\mathrm{A}, 8$ - BDF

This page intentionally left blank

# SEcTION OBSERVATIONAL STUDIES: ECOLOGICAL STUDIES 

- ECOLOGICAL STUDIES

12, 16

- ADVANTAGES AND DISADVANTAGES OF ECOLOGICAL STUDIES


## OBSERVATIONAL STUDIES: ECOLOGICAL STUDIES

Seagroatt V. MMR vaccine and Crohn's disease: ecological study of hospital admissions in England, 1991 to 2002. BMJ 2005;330:1120-1121 (extracts and figures reproduced with permission from BMJ Publishing Group).

## INTRODUCTION

It has been hypothesised that the measles, mumps, and rubella vaccine (MMR vaccine) increases the risk of autism and Crohn's disease. Although a possible link with autism has been extensively studied and refuted, a link with Crohn's disease has not. I tested this hypothesis by analysing trends in age specific admission rates for Crohn's disease in children and adolescents to determine if the introduction of MMR vaccine in 1988 increased rates in those populations that were offered the vaccine as infants.'

1. What is the question being investigated by this study?
2. What type of comparison is being performed?
3. What types of data are being used to answer the question?

## RESULTS

'Age specific rates per 10000 population per year for emergency hospital admissions for Crohn's disease in England, 1991 to 2002. Rates in children aged < 4 years were relatively low and so were excluded from the figure. Three-year

 groups, rather than the more conventional five-year groups, were used in order to discriminate between rates in children born before and after the introduction of MMR.'
'There were 4463 admissions for Crohn's disease, 923 of which occurred in populations with a vaccination rate of $\geq 84$ per cent (those born in 1988-89 or later). Although the age specific rates increased over the study period, no obvious changes occurred that coincided with the introduction of MMR vaccine. The estimated rate ratio for the MMR vaccination programme (rates in populations with a vaccination rate of $\geq 84$ per cent compared with those with a rate of $\leq 7$ per cent) was 0.95 ( 95 per cent confidence interval 0.84 to 1.08 ).'
4. What can we infer from the rate ratio and the confidence interval?

## 5. What potential confounding factors may influence the results?

## EXPLANATION: ECOLOGICAL STUDIES

In an ecological study the units of observation are populations or groups of people, rather than individuals. For example, an individual does not have a life expectancy or an income distribution, but a population of a city, state or country does. Ecological studies allow statements to be made about the populations being compared. While they may suggest associations between a disease and exposure, these usually require confirmation from studies involving individuals.

For data collection, disease rates and exposure are measured in each of a series of populations and their relationship is analysed. Ecological studies often use routinely collected aggregate statistics, usually published for other purposes, such as mortality rates or hospital admissions rates.

Populations can be compared in a variety of different ways:

- Geographical comparison: comparison of disease rates and prevalence of risk factors in different geographical areas
- Temporal comparison: ecological studies can be used to analyse trends in disease patterns over time by taking routinely collected statistics from the same population group over successive time intervals
- Migrant studies: the study of migrant populations helps to disentangle the influence of genetics and environmental exposures in determining disease processes. It can also help to establish the age at which environmental influences exert their effect. For example, studies looking at multiple sclerosis prevalence in migrant populations have shown that populations from places close to the equator maintain their low prevalence rates when they migrate to higher latitudes. However the offspring of the migrants adopt the high prevalence rates associated with the higher latitude location
- Occupation and socio-economic group: statistics on exposure and disease are widely available for specific groups in society. For example, occupational risk factors, such as the stress associated with working in the medical profession can be correlated with morbidity statistics such as rates of alcoholism in doctors.

The question being asked in the example study (page 12) is: 'Is there an association between the rate of Crohn's disease in children and the introduction of the MMR vaccine in 1988?' The study is a temporal comparison $(1,2)$. The data used in the study come from two sources of population-level data (3):

1. Routinely collected statistics for the age-specific rates of emergency hospital admissions for Crohn's disease for children under the age of 18 years from April 1991 to March 2003
2. Percentages of children completing a primary course of MMR vaccine in their second year of life (in the first two years of the MMR vaccination programme, these were 7 per cent and 68 per cent; thereafter they were at least 84 per cent).

[^1]6. Which one of the following definitions best describes the 'ecological fallacy'?
a. The weakness of ecological studies compared to case-control or cohort studies
b. The mistaken interpretation of a study as ecological when really it is a cross-sectional study
c. An association found in an ecological study does not exist at the individual level
7. What are the advantages of an ecological study?
a. It can be used to study associations at the individual level
b. It can study large and very different population groups
c. It does not rely on existing published statistics which may contain errors
d. It uses aggregate data on exposure and disease in population groups, increasing the power of the study
e. It helps to formulate hypotheses on aetiological factors in disease
f. It is easy to minimize confounding factors
8. What are the disadvantages of an ecological study?
a. It cannot make inferences about individuals
b. There is a risk of the ecological fallacy
c. It is costly and time consuming to conduct
d. It is less reliable than a case-control study that lacks within-population exposure variability
e. It cannot compare populations that have very different characteristics

## EXPLANATION: ADVANTAGES AND DISADVANTAGES OF ECOLOGICAL STUDIES

Advantages of an ecological study are:

- It is quick, simple and cheap to conduct due to the availability of routinely collected data that have already been published
- It has more power than individual-level studies, such as case-control and cohort studies where there is less exposure variability
- Data can be used to compare populations with widely differing characteristics, for example the Chinese and the Americans
- It provides a useful starting point for more detailed epidemiological work by helping to formulate hypotheses about the aetiology of disease.


## Disadvantages are:

- The risk of the 'ecological fallacy'. This is when inappropriate conclusions are drawn on the basis of ecological data. An association seen at the group level does not necessarily represent an association at the individual level, therefore an ecological study cannot make inferences about individual level associations
- Inability to control potential confounding factors other than age and sex. For example, in geographical comparisons, although it may be possible to adjust for age and sex, data for other potential confounders, such as dietary or cultural habits, may not be available. In temporal studies, changes in diagnostic or treatment techniques may influence disease statistics over time. In migration studies, factors associated with the act of migration itself, such as psychological stress, may influence disease processes, confounding the influence of new environmental risk factors. In occupational studies, socio-economic factors may confound the results and vice versa
- Reliance upon existing published statistics may limit the breadth and type of studies conducted.

[^2]
## EXPLANATION: ECOLOGICAL STUDIES Cont'd from page 13

The analysis of data from ecological studies depends upon the mode of comparison being used, for example in geographical studies, associations between disease occurrence and exposure are often presented graphically in the form of scatter plots (see page 75). For temporal comparisons, trends may also be displayed graphically, and correlation coefficients (see page 107) or rate ratios with confidence intervals (as in the example study on page 12 and also see page 89) may be calculated.

A rate ratio of 0.95 suggests that the rate of emergency Crohn's admissions in children born after the introduction of the MMR vaccine (population group with $\geq 84$ per cent vaccinated) is almost the same as the rate in the group born before the vaccine was introduced ( $\leq 7$ per cent vaccinated). The narrow confidence inter$\operatorname{val}$ (which includes the value of 1 ) indicates that the rate ratio estimate is precise: we can be 95 per cent certain that the true rate ratio being estimated lies in the range 0.84 to 1.08 (4).

Potential confounders (see page 117) could include changes in dietary habit, new medical treatments or another immunological-type factor with a protective effect against Crohn's disease in order to counteract an added risk from the MMR (5). However, quoting from the example study: ‘. . . some factor(s) would have to be negatively associated with Crohn's disease, be introduced over the same three-year period, and be targeted at the same population of infants as MMR vaccine to mask a true association. This seems highly unlikely.'

# $-3$ <br> <br> OBSERVATIONAL STUDIES: <br> <br> OBSERVATIONAL STUDIES: CROSS-SECTIONAL STUDIES 

 CROSS-SECTIONAL STUDIES}

- CROSS-SECTIONAL STUDIES (i) 18
- CROSS-SECTIONAL STUDIES (ii) 20
- ADVANTAGES AND DISADVANTAGES OF CROSS-SECTIONAL STUDIES22


Chinn S and Rona RJ. Prevalence and trends in overweight and obesity in three cross sectional studies of British children, 1974-94. BMJ 2001;322:24-26 (summary of study reproduced with permission from the BMJ Publishing Group).

Study participants were primary school children, 10414 boys and 9737 girls in England and 5385 boys and 5219 girls in Scotland aged 4 to 11 years. The height and body mass of all the children were measured and body mass index (BMI) calculated as weight/height ${ }^{2}$, where weight was measured in kg and height in m .

Using linear interpolation between cut-off points for each six months of age, the percentages of children who were overweight or obese, as defined by the International Obesity Task Force (internationally agreed cut-off points equivalent to BMIs of 25 and 30 respectively at age 18 years) were calculated.

1. Answer the following questions with reference to the above study:
a. What are the research questions being asked in this cross-sectional study?
b. What features of the study design characterize it as cross-sectional?
c. What type of data does it generate?
d. What are the advantages and disadvantages of conducting the study in England and Scotland?
2. From data in the table below, which shows a summary of results for 1974, what statements are correct regarding overweight and obesity in the UK?

|  | Percentage prevalence of <br> overweight children in 1974 | Percentage prevalence of <br> obese children in 1974 |
| :--- | :--- | :--- |
| English boys | 6.4 | 1.4 |
| English girls | 9.1 | 1.5 |
| Scottish boys | 5.4 | 1.7 |
| Scottish girls | 8.8 | 1.9 |

a. The incidence of obesity is higher than the incidence of being overweight
b. Girls and boys have a similar risk of being overweight
c. The dietary habits of the Scots explain the higher rates of obesity in Scotland
d. Nearly one in ten primary school-aged girls in England are considered overweight as defined by the International Obesity Task Force

## EXPLANATION: CROSS-SECTIONAL STUDIES (i)

A cross-sectional study usually involves a health survey of a group of individuals in a specified population (1b). This is in contrast to an ecological study which uses routinely collected data about groups. Crosssectional studies are particularly useful for estimating the prevalence of a condition or characteristic in a population, where the condition is a disease status defined in a standard way. These studies are conducted at a single point in time or over a very short time period (1b), and are therefore unable to measure disease incidence. Prevalence data rely upon strict definitions. The definition of a disease state will influence prevalence data collected in a cross-sectional study. Taking the example opposite, different cut-off values of BMI would produce different prevalence estimates for obesity.

The validity of the prevalence data obtained from cross-sectional studies depends on study participants providing a representative sample of the relevant population. When the aim is to measure disease prevalence according to presence or absence of an exposure, they are not picked according to whether they have a disease or condition, or according to their exposure to any postulated factor (1c). The number of 'cases' studied is therefore not specified in advance but is simply the number present in the sample. The best studies are those in which the participants are selected by a random method, for example from the electoral roll, school registers or general practitioners' lists, rather than by using volunteers (1c). Non-random methods of recruiting participants, such as requesting volunteers or reliance on responses to postal questionnaires where the response rate is poor, may introduce bias into the sample population.

Cross-sectional studies can also be conducted simultaneously in different geographical places, so that prevalence can be compared in different areas. They may also compare different population groups, such as males versus females, different age groups, different socio-economic or ethnic groups. The questions asked in the example study are: 'What are the rates of obesity and overweight in the UK?' and 'Are there differences in prevalence between girls and boys/between England and Scotland/between age subgroups? (1a)'.

Advantages of conducting such a study in England and Scotland are that it is possible to obtain representative data for the whole of the UK and one can compare rates in different geographical areas to formulate hypotheses about causative environmental factors (1d).

Disadvantages are that extra expense and co-ordination are needed to conduct more than one study simultaneously and there may be differences in methods of conducting the study in different geographical areas, leading to confounding factors (1d).

This study is an example of a descriptive cross-sectional study. An analytical cross-sectional study collects data on both disease and exposure. Examples of analytical studies are:

- BRCA1 gene and diagnosis of breast cancer in women attending breast cancer screening clinics
- Prevalence of cancer in a representative sample of individuals in Chernobyl before and after the Chernobyl nuclear reactor disaster.

[^3]Chinn S and Rona RJ. Prevalence and trends in overweight and obesity in three cross sectional studies of British children, 1974-94. BMJ 2001;322:24-26 (tables adapted with permission from the BMJ Publishing Group).

The example study described on page 18 was actually repeated at three sequential points in time, 1974, 1984 and 1994. Results are shown in the table below.

|  | Change in prevalence of overweight (95 per cent CI$)$ |  |
| :--- | :--- | :--- |
| $1974-84$ |  |  |$)$

3. What do the results tell us about changes in prevalence of overweight children between 1974 and 1984, and 1984 and 1994?
4. How confident can we be in these results?
5. What conclusions can be drawn from the obesity data?
6. How could one improve the study, particularly regarding the obesity data?
7. What, if any, action would these results prompt you to take?

## EXPLANATION: CROSS-SECTIONAL STUDIES (ii)

Cross-sectional studies are used to calculate the prevalence of a disease or condition at one particular point in time.

$$
\text { Prevalence }=\frac{\text { Number of people with disease at a single point in time }}{\text { Total number studied at the same time point }}
$$

However, cross-sectional studies can be repeated at different points in time to obtain sequential prevalence estimates and thus comment on trends in the population.

The example study tells us:

- Overweight 1974 to 1984: prevalence decreased in one group (English boys) but increased in others. In all groups the confidence interval includes 0 , which means that one cannot exclude the possibility that there is no change in prevalence of overweight
- Overweight 1984 to 1994: prevalence in all groups has increased. This change is greatest in Scottish girls and the 95 per cent confidence intervals do not include 0 suggesting real increases
- Obesity 1974 to 1984: prevalence decreased in all groups. For English boys the estimated change in prevalence is -0.8 and the confidence interval does not include zero and suggests this is a real reduction
- Obesity 1984 to 1994: for all groups, the difference in prevalence shows an increase and the confidence intervals all exclude 0 suggesting this is a real increase (3).

Confidence intervals (e.g. 95 per cent) can be calculated for changes in prevalence over time, providing a value for the statistical uncertainty associated with the estimated change in prevalence (see page 89). From the example given, the prevalence of overweight in English boys increased by 3.6 between 1984 and 1994 with a 95 per cent confidence interval of 2.3 to 5.0 . This means that the true value of the increase in prevalence has a 95 per cent probability of being between 2.3 and 5.0 (4). The prevalence changes for obesity are smaller. The estimated increase in prevalence in English boys between 1984 and 1994 was 1.2 with a 95 per cent confidence interval of 0.6 to 1.7 . As this interval excludes zero we can conclude from these findings that there has been a real increase in obesity over the time frame investigated (5).

The prevalence of obesity is much lower than that of overweight. The smaller number of cases reduces the statistical precision of the study for obesity. To increase the precision of obesity results, the size of the study would need to be increased (see pages 85 and 87) (6).

The study has highlighted that the prevalence of overweight in children is increasing and suggests that there is a similar rising trend in the prevalence of obesity. Public health initiatives could be implemented for the primary prevention of overweight and for helping overweight children to maintain their weight, in order to prevent any further increases in the prevalence of obesity (7).

[^4]8. What are the advantages of a cross-sectional study?
a. It is quick and inexpensive
b. It can be used to calculate incidence
c. It is good for studying common conditions
d. It is good for looking at rare diseases
e. It can be used to test a hypothesis
f. It can be used to compare the prevalence data in different subsets of a population
g. It can compare prevalence in populations in different geographical regions
9. Which of the following questions would be best investigated by using a crosssectional study?
a. Whether a new drug for hyperlipidaemia is effective at reducing serum triglyceride level
b. Whether there is an association between obesity and childhood asthma
c. Whether there has been a change in the incidence of heart disease in the under-40-year age group over the last 20 years
d. Whether hay fever is more common in cities than rural areas
e. Whether alcohol-related accidents are more common in France than in Britain
10. What are the disadvantages of a cross-sectional study?
a. It cannot be used to estimate incidence
b. There is a potential for recall bias
c. Definitions of disease may influence prevalence estimates
d. It can always demonstrate true trends in disease
11. In which of the following situations would a cross-sectional study be inappropriate?
a. To test the hypothesis that obesity leads to an increased risk of asthma
b. To calculate the number of new cases of obesity per year in the UK
c. To study the prevalence of obesity in different countries of Europe
d. To study whether the incidence of allergies changes with season
e. To study whether exposure to mobile phone signals precedes development of brain tumours

## EXPLANATION: ADVANTAGES AND DISADVANTAGES OF CROSS-SECTIONAL STUDIES

Advantages of cross-sectional studies are:

- They are relatively quick: they are only conducted at a single point in time (unless a series of repeated studies is done) and therefore can be relatively inexpensive
- They can give estimates of prevalence
- They are flexible: they can be used for studying both rare and common conditions or diseases, depending on the sample size recruited
- They can be used to compare prevalence in different populations and thus used to formulate hypotheses about disease, based on characteristics such as sex, age or geographical location of a population
- Sequentially repeated cross-sectional studies can be used to estimate changes in prevalence over time and thus are used to demonstrate trends in health and disease.

Disadvantages of cross-sectional studies are:

- They cannot measure incidence: cross-sectional studies usually collect data only from a single point in time; including both new cases and those diagnosed in the past
- They can be used to formulate but not formally test hypotheses regarding associations between subsequent disease following previous exposure: the subjects are not chosen according to their exposure to a factor of interest or their disease status. It is only possible to use cross-sectional studies to study whether exposure and outcome may be associated. The results obtained may be interpreted in a number of different ways, giving rise to alternative explanations for the associations observed
- Long-term outcomes: cross-sectional studies do not provide information about the long-term outcomes of particular trends in health and disease status. Study participants are not followed up to a defined end-point and the consequences are not recorded, only the characteristic(s) of interest at the time of the survey. For example, the overweight and obesity study can only speculate on the future consequences of a trend of increasing BMI in children. It cannot provide hard proof as to the long-term medical complications of such a trend, as demonstrated by the following quote from the paper detailed on pages 18 and 20: 'Rising trends are likely to be reflected in increases in adult obesity and associated mortality'.

[^5]This page intentionally left blank

# -4 <br> <br> OBSERVATIONAL STUDIES: <br> <br> OBSERVATIONAL STUDIES: CASE-CONTROL STUDIES 

 CASE-CONTROL STUDIES}

- CASE-CONTROL STUDIES (i)
- CASE-CONTROL STUDIES (ii) 28, 32
- ADVANTAGES AND DISADVANTAGES OF CASE-CONTROL STUDIES


## SECTION OBSERVATIONAL STUDIES: CASE-CONTROL STUDIES

Doll R and Hill AB. Smoking and carcinoma of the lung. Preliminary report. BMJ 1950;2:739-748 (extracts reproduced with permission from the BMJ Publishing Group).
'The great increase in the number of deaths attributed to cancer of the lung in the last 25 years justifies the search for a cause in the environment. An investigation was therefore carried out into the possible association of carcinoma of the lung with smoking, exposure to car and fuel fumes, occupation, etc. The preliminary findings with regard to smoking are reported.
'The material for the investigation was obtained from twenty hospitals in the London region which notified patients with cancer of the lung, stomach and large bowel. Almoners then visited and interviewed each patient. The patients with carcinoma of the stomach and large bowel served for comparison and, in addition, the almoners interviewed a non-cancer control group of general hospital patients, chosen so as to be of the same sex and age as the lung-carcinoma patients.'

## 1. Answer the following with reference to the above study:

a. What question was this study designed to answer?
b. What features of the study characterize it as a case-control study?
c. What is the definition of a case in this study?
d. Who are the controls?
e. On what basis are cases and controls compared?
f. What is the source population for the cases and controls?
g. Can you think of any disadvantages of using this population as a source?
h. How was information collected in this study?
i. What are the disadvantages of using this method of data collection?

## EXPLANATION: CASE-CONTROL STUDIES (i)

A case-control study is an observational epidemiological study where subjects are selected on the basis of whether they have (cases) or do not have (controls) a particular disease. A comparison is made between the proportion of subjects in the case and control groups who have been exposed to the aetiological factor under study. Case-control studies are always retrospective, i.e. they look back in time to determine past exposure in cases and controls (1b).

The example study set out to answer the following question:
'Is there an association between smoking and lung cancer?'; more precisely: 'Is it more likely that you were a smoker in the past if you have lung cancer now than if you don't have lung cancer now?'. In this study a 'case' is a person diagnosed with the disease of interest, i.e. a patient with lung cancer (1a,1c).
'Controls' do not have lung cancer. They are not necessarily healthy - in this study, the controls had diseases such as cancer of the oesophagus or stomach, or non-cancer diseases not then thought to be associated with smoking (1d). Cases and controls were compared with regard to their past smoking habits (1e). The data were collected from patients from 20 London hospitals by personal interviews ( $1 \mathbf{f}, 1 \mathrm{~h}$ ). There are disadvantages of using patients only from London hospitals, as this may overlook other possible aetiological factors for lung cancer, such as pollution or overcrowding (1g). Interviewers may consciously or subconsciously encourage particular answers or responders may not give accurate responses when asked questions face to face, introducing observer or responder bias, respectively (1i).


In a matched study, each case has one or more controls to which it is matched on an individual basis. Typical matching factors may include gender, age, occupation, region of residence, etc. The object of matching is to obtain a more accurate estimate of differences by 'removing' the possible influences of variables other than the exposure under investigation. The matching is done on confounding factors that are already established risk or protective factors, but are not under investigation themselves. There is a danger of over-matching, where cases and controls are matched on so many different confounding variables that it makes them over-similar with respect to potential aetiological factors of interest.

## Answers

1. See explanation

Doll R and Hill AB. Smoking and carcinoma of the lung. Preliminary report. BMJ 1950;2:739-748 (table adapted with permission from the BMJ Publishing Group).

The table below shows results obtained from this study.

| Patients | Total number of patients | Number who are smokers |
| :--- | :---: | :---: |
| Males |  |  |
| Lung cancer | 649 | 647 |
| No lung cancer | 649 | 622 |
| Females |  |  |
| Lung cancer | 60 | 41 |
| No lung cancer | 60 | 28 |

2. Which method is commonly used to present the results of a case-control study?
3. Use this method for the above data on males and females
4. What measure of association is used to summarize the results of a case-control study?
5. Calculate the value of this measure of association for the above data
6. Comment on the meaning of these calculated values
7. State the null hypothesis for a case-control study
8. What simple hypothesis test can be applied to the results of a case-control study?
9. Apply this test to the above data
10. What value is calculated to determine the statistical significance of the test statistic?
11. Why does this value differ for men and women?
[^6]
## EXPLANATION: CASE-CONTROL STUDIES (ii)

Results of a case-control study can be presented in a $2 \times 2$ table (2):

|  | Cases | Controls | Total |
| :--- | :--- | :--- | :--- |
| Exposed | $a$ | $b$ | $a+b$ |
| Unexposed | $c$ | $d$ | $c+d$ |
| Total | $a+c$ | $b+d$ | $a+b+c+d$ |

The data from this study can be tabulated as follows (3):

|  | Lung cancer | No lung cancer | Total |
| :--- | :---: | :---: | ---: |
| Males |  |  |  |
| Smokers | 647 | 622 | 1269 |
| Non-smokers | 2 | 27 | 29 |
| Total | 649 | 649 | 1298 |
| Females |  |  |  |
| Smokers | 41 | 28 | 69 |
| Non-smokers | 19 | 32 | 51 |
| Total | 60 | 60 | 120 |

The measure of effect that can be used to measure the association between disease and exposure in a case-control study is the odds ratio (see page 103) (4). Odds ratio $(\mathrm{OR})=(a / c) /(b / d)=a d l b c$

For this study this gives an OR (5) of: males: $(647 \times 27) /(622 \times 2)=14.0$; females: $(41 \times 32) /(28 \times 19)=$ 2.5

For rare events, the value of the odds ratio is close to that of the relative risk (see pages 103 and 105), and is thus sometimes considered as an estimate of relative risk.

- Odds ratio = 1: suggests the odds (or risk) is the same in exposed and unexposed groups
- Odds ratio > 1: suggests an increased risk of disease associated with exposure
- Odds ratio $<\mathbf{1}$ : suggests a reduced risk of disease associated with exposure, i.e. exposure is protective.

In this study an odds ratio of 14.0 is highly suggestive of an increased risk of lung cancer in males associated with smoking. For females, an odds ratio of 2.5 also suggests an increased risk but this is less striking than for males. This might reflect a true difference in the risk of smoking for developing lung cancer in men and women, or that women have not been smoking for as long. The odds ratio estimate for women will be less precise due to the lower number of cases (6).

## Answers

2. A $2 \times 2$ table
3. Odds ratio $=1$
4. See explanation
5. The chi-squared test
6. The odds ratio
7. See explanation (page 32)
8. See explanation
9. The $P$-value
10. See explanation
11. See explanation (page 32)
12. What are the advantages of a case-control study?
a. It is quick and relatively inexpensive
b. It can be used to determine the temporal relationship between risk factor and disease
c. It may reveal new associations not previously considered
d. It is good for studying rare diseases
e. It can be used to test a hypothesis
13. What are the disadvantages of a case-control study?
a. Are particularly subject to recall bias
b. They cannot directly measure disease risk
c. There is a risk of loss of cases and controls to follow up
d. A larger sample size is needed than for a cohort study
14. A case-control study of the suspected association between breast cancer and oestrogen therapy
a. Can directly measure the attributable risk to an individual of developing the disease as a result of therapy
b. Will require controls selected randomly from the general population
c. Will require follow-up of a group of women on oestrogen therapy and a control group not on therapy
d. Will investigate the temporal relationship between oestrogen therapy and development of breast cancer
e. Will prove the existence of a causal relationship

## EXPLANATION: ADVANTAGES AND DISADVANTAGES OF CASE-CONTROL STUDIES

Advantages of case-control studies are:

- They are relatively quick
- They are relatively inexpensive
- They can test a hypothesis. For example, is there an association between smoking and lung cancer?
- They are good for studying rare diseases as only a relatively small sample size is needed
- They can be used for simultaneously investigating several potential aetiological factors
- Dose dependence of the aetiological factor can be analysed. For example, from Doll and Hill's paper: 26 per cent of men with lung cancer smoked more than 25 cigarettes a day whereas only 13.5 per cent of men without lung cancer smoked the same amount. For women the pattern was the same, 14.6 per cent versus 0 per cent, respectively. This is illustrated by these quotes from the example study: '. . . the risk of developing the disease . . . may be approximately 50 times as great among those who smoke 25 or more cigarettes a day as among non-smokers.' and 'The greater prevalence of carcinoma of the lung in men compared with women leads naturally to the suggestion that smoking may be a cause since smoking is predominantly a male habit.'

Disadvantages of case-control studies are:

- Cases may be likely to remember past exposures better than controls and are therefore subject to recall bias
- They cannot determine the temporal relationship between exposure and disease
- It is difficult to be sure of cause and effect: they cannot directly calculate risk of a disease only estimate odds ratios associated with an exposure. For example, in Doll and Hill's study, their conclusion relied upon the fact that smoking appeared to precede the development of lung cancer and that there were no other confounders that could be thought of by the study designers
- They cannot reveal associations that are not specifically looked for, i.e. they do not ask open-ended questions
- They are not good for rare exposures. For example, in Doll and Hill's study the odds ratio for smoking in men was much greater than in women. This may be partly explained by the lower prevalence of smoking in women
- There is a risk of selection bias through selection of unsuitable cases or selection of controls that are inadequately or overly matched to controls. For example, in the example study by Doll and Hill, differences with regard to social class, place of residence and the part of England from which subjects were drawn may have influenced the results
- Confounding factors: in the example study, potential confounding factors of London life considered at the time included atmospheric pollution, type of housing and occupation. Matching cases and controls for such factors or measuring potential risk factors and adjusting for them in the analysis may help to reduce this source of error.

[^7]
## EXPLANATION: CASE-CONTROL STUDIES (ii) cont'd from page 29

The hypothesis that there is an association between exposure to the aetiological factor of interest and subsequently developing the disease of interest can be tested using the data generated by a case-control study.

The null hypothesis is that the true odds ratio is 1 (7). The chi-squared $\left(\chi^{2}\right)$ test statistic provides a $P$-value to measure the strength of evidence against the null hypothesis $(8,10)$. Results of applying this test are shown in the table below (9).

|  | Lung cancer <br> Observed |  | Expected | No lung cancer <br> Observed |  |
| :--- | :---: | :--- | :--- | :--- | ---: |
|  | Expected |  | Total |  |  |
| Males |  |  |  |  |  |
| Smokers | 647 | $=1269 \times(649 / 1298)=634.5$ | 622 | $=1269 \times(649 / 1298)=634.5$ | 1269 |
| Non-smokers | 2 | $=29 \times(649 / 1298)=14.5$ | 27 | $=29 \times(649 / 1298)=14.5$ | 29 |
| Total | 649 |  | 649 |  | 1298 |
| Females |  |  |  |  |  |
| Smokers | 41 | $=69 \times(60 / 120)=34.5$ | 28 | $=69 \times(60 / 120)=34.5$ | 69 |
| Non-smokers | 19 | $=51 \times(60 / 120)=25.5$ | 32 | $=51 \times(60 / 120)=25.5$ | 51 |
| Total | 60 |  | 60 |  | 120 |

Doll R and Hill AB. Smoking and carcinoma of the lung. Preliminary report. BMJ 1950;2:739-748 (table adapted with permission from the BMJ Publishing Group).
$\chi^{2}$ for males is 20.3 and $\chi^{2}$ for females is 4.90 .
The formula for obtaining $\chi^{2}$ is provided in the appendix (page 122) and below is a worked example:
For males:

$$
\begin{aligned}
\chi^{2}= & \Sigma\left[(|O-E|-0.5)^{2} / E\right] \\
= & {\left.\left[(|647-634.5|-0.5)^{2} / 634.5\right]+[(|622-634.5|)-0.5)^{2} / 634.5\right]+\left[(|2-14.5|-0.5)^{2} / 14.5\right] } \\
& +\left[(|27-14.5|-0.5)^{2} / 14.5\right] \\
= & 0.227+0.227+9.93+9.93 \\
= & 20.3
\end{aligned}
$$

( $O$ is observed, $E$ is expected and modulus $(I \mathrm{O}-\mathrm{El})$ is the absolute magnitude of the difference between $O$ and $E$ regardless of the sign.)

Referring to the statistical table of the chi-squared distribution in the Appendix (page 126), the test statistic follows the chi-squared distribution with 1 degree of freedom ( 2 proportions, $\mathrm{df}=2-1=1$ ): for males, $P<0.001$; for females, $0.01<P<0.05$. The $P$-value is smaller for males than for females. This means that the statistical significance of the results is less strong for females. This is partly because there are fewer cases of females with lung cancer, reducing the power of the study for females (11).

## sEction (5 OBSERVATIONAL STUDIES: COHORT STUDIES

- COHORT STUDIES (i) 34
- COHORT STUDIES (ii) 36
- COHORT STUDIES (iii) 38
- ADVANTAGES AND DISADVANTAGES OF COHORT STUDIES



## OBSERVATIONAL STUDIES: COHORT STUDIES

Song Y-M, Sung J, Lawlor DA et al. Blood pressure, haemorrhagic stroke, and ischaemic stroke: the Korean national prospective occupational cohort study. BMJ 2004;328:324-325 (extracts reproduced with permission from BMJ Publishing Group).
'We examined the association of blood pressure with subtype of stroke in a large cohort of Korean civil servants. Blood pressure was measured for individuals between 1986 and 1996 during their biennial health examinations. We included deaths attributed to all strokes, haemorrhagic and ischaemic between 1991 and 2000 in these analyses. We categorized nonfatal strokes using data on the use of medical care, and found an accuracy of 83.4 per cent and 85.7 per cent for ischaemic stroke and haemorrhagic stroke.
'In 9.5 million person years of observation of 955271 people; they had 14057 strokes, giving crude and age standardized incidences of 1.48 and 2.24 for every 1000 person years. Of these, 10716 ( 76 per cent) strokes had complete information on major exposure variables and we included these in our analyses; we classified 2695 strokes as haemorrhagic, 5326 as ischaemic, 1731 as undetermined, and 964 as subarachnoid haemorrhage.
'Multiple characteristics were measured as variables that are known or potential risk factors for ischaemic or haemorrhagic strokes. These included age, sex, body mass index (BMI), height, blood glucose, blood cholesterol, haemoglobin concentration, ethanol consumption, smoking, monthly pay level, and area of residency.
'We calculated fully adjusted relative risks and 95 per cent confidence intervals using logistic regression. . . . Both ischaemic stroke and haemorrhagic stroke had strong gradients with blood pressure, but these were much steeper for haemorrhagic stroke. . . . For each higher 20 mmHg of systolic blood pressure, the relative risk of ischaemic and haemorrhagic stroke increased by 2.23 (2.17 to 2.30) and 3.18 ( 3.06 to 3.30), z test for difference between odds ratios 11.40, $P<0.00001$. . . Our findings emphasize the importance of controlling blood pressure, particularly in countries with a high risk of haemorrhagic stroke.'

## 1. Answer the following with reference to the study above

a. What question has this study been designed to answer?
b. What features of this study suggest that it is a cohort study?
c. Is this a prospective or retrospective cohort study?
d. What are the likely sources of error in this study?

## EXPLANATION: COHORT STUDIES (i)

A cohort study is a type of observational epidemiological study.
A cohort study can be either prospective (current) or retrospective (historical). In a prospective study, the data on exposure are collected prior to the occurrence of disease and subjects are followed up over time to observe occurrence of the disease. In a retrospective study, past exposure data are assembled for a defined cohort using historical records. The disease outcome may or may not be known at this time; ideally it should be obtained independent from or subsequent to assembly of exposure data.

The figure below gives a diagrammatic representation of different cohort study designs.


The example given is a huge prospective cohort study (1c) designed to investigate if there is any association between blood pressure and subtype of stroke (haemorrhagic or ischaemic) (1a). The factors that made this typical of a cohort study were that a large 'cohort' of study participants was defined according to participants' exposure (high blood pressure) rather than according to their disease status, and that participants were followed over time to determine subsequent stroke risk (1b).

A major limitation of cohort studies, particularly those done over a long time period, is loss of subjects due to factors such as death or migration. In this example, complete information was only collected on 76 per cent of recorded strokes, therefore almost a quarter of outcomes have not been related to exposure; this could be a potentially huge source of error, especially if relative risks were smaller. Some of the tools available to reduce error due to loss of subjects when analysing the results of a cohort study are discussed on page 39 .

Likely sources of error in the study in question include bias, confounding, human error in measuring the initial blood pressures and error due to incomplete information (1d).

## Answers

1. See explanation

## 2. Using the results table opposite, answer the following true or false

a. You are less likely to have a stroke if you have normal blood pressure
b. If you have very high blood pressure you are almost 20 times as likely to have a fatal stroke than if you have normal blood pressure
c. If you have a very high blood pressure you are almost 20 times as likely to have an ischaemic stroke than if you have normal blood pressure
d. The larger the number (of people having strokes) the smaller the range of the confidence intervals

## 3. Using the results table on page 37 answer the following as percentages

a. What is the probability that less than 16.41 times as many people with very high blood pressure as normal blood pressure have a fatal stroke?
b. What is the probability that between 24.89 and 33.40 times as many people with very high blood pressure as normal blood pressure have a haemorrhagic stroke?
4. Choose from the following options the most appropriate test for the situations given

## Options

A. Chi-squared tests (see page 122)
B. $z$-test for numeric data (see page 95)
C. Cox proportional hazards (see page 39)
D. Hypothesis testing (see page 93)
E. Calculating confidence intervals (see page 89)

1. Determining if results are statistically significant
2. Estimating the size of the effect on the outcome of the exposure
3. Comparing the means of continuous variables (height, BMI, age, etc.) in two groups
4. Comparing the dichotomous variables (male/female, etc.) in the groups
5. To account for differences other than exposure (blood pressure) or risk of outcome (stroke)

## EXPLANATION: COHORT STUDIES (ii)

The risk estimates obtained from this cohort study have been presented as relative risk rather than absolute values. See page 105 on relative risk and absolute risk. This is sometimes a useful way of representing the information as it can be seen at a glance that, whatever subcategory of stroke is considered, a given individual is between 8 and 34 times as likely to have a stroke if their blood pressure is very high.

The disadvantage of representing the data as relative risk values is that the absolute risk of ischaemic and haemorrhagic stroke in each of the blood pressure groups cannot be compared directly. The increase in relative risk of haemorrhagic stroke is much greater than the increase in relative risk of ischaemic stroke with increased blood pressure; 28.83 times compared to 9.56 times normal values, respectively. The absolute risk of an ischaemic stroke if one has very high blood pressure might still be higher than haemorrhagic stroke. This depends on the magnitude of the risk of each condition for those with normal blood pressure.

The best cohort studies will identify potential confounding variables and consider whether they differ between the two groups. If there are statistically significant differences, e.g. higher cholesterol levels in the very high blood pressure group, these can be adjusted for in the data analyses. In the example study, the confounding variables adjusted for included age, gender, BMI, height, blood glucose, blood cholesterol, haemoglobin concentration, ethanol consumption, smoking, monthly pay level, and area of residency.

The table below shows relative risks (95 per cent confidence intervals) for mean blood pressure from the Korean National Health System Study, 1986-2000.

|  | All strokes <br> fatal | Non-fatal | Stroke subtypes <br> Ischaemic | Haemorrhagic |
| :--- | :--- | :--- | :--- | :--- |
| Number of strokes <br> Blood pressure $(\mathrm{mmHg})$ | 2073 | 8643 | 5326 | 2695 |
| $<140 /<90$ Normal | 1 | 1 | 1 | 1 |
| $\geq 180 / \geq 110$ Very high | $19.39(16.41$ to 22.90$)$ | $11.21(10.17$ to 12.36$)$ | $9.56(8.46$ to 10.80$)$ | 28.83 (24.89 to 33.40$)$ |

Song Y-M, Sung J, Lawlor DA et al. Blood pressure, haemorrhagic stroke, and ischaemic stroke: the Korean national prospective occupational cohort study. BMJ 2004;328:324-325 (table adapted with permission from the BMJ Publishing Group).

[^8]A group of 20 young adults went on a 2 -week-long skiing trip. On the first day one person broke a finger and another broke an arm. On the third day someone tore a ligament in their knee; two further injuries occurred in the first week, one on the fifth day and one on the sixth day. On the penultimate day of the holiday one of the members of the group broke their leg. Each of the members of the group who was injured stopped skiing from the time they were injured onwards.
5. What is the risk, per day, of a skier in the group of 20 having an injury?
a. On the first day
b. If the holiday were only 1 week long
c. Over the 2 weeks
6. What was the average number of days that each member of the group spent skiing during the 2 -week holiday?
7. Which of the following compares 'time to event' in two groups?
a. Log rank
c. Odds ratio
b. Cox proportional hazards
d. Kaplan-Meier survival curves
8. The curve below showing the probability of survival tells us:
a. Less than 25 per cent survive 5 years
b. 75 per cent survive 10 years
c. 90 per cent survive 5 years
d. More than 50 per cent survive 10 years


## EXPLANATION: COHORT STUDIES (iii)

A Kaplan-Meier survival curve compares the prognosis of different conditions and different treatments used at different times, with different periods of follow-up. A survival curve allows us to estimate the cumulative probability of an event occurring, using data from the subjects who are still alive and available at the last follow-up. This allows us to calculate the expected time period for a particular event to occur. Although survival curves were originally used to assess mortality, they can be used to assess any outcome, for example, time for a fracture to heal, time to discontinuation of a psychiatric drug and many others.

An example survival curve for two childhood cancers is shown below.


Each step on the curve represents the change in probability as individual events occur. For example, if two children out of 20 die between 2 and 6 weeks, the probability of survival will fall by 10 per cent over that 4 -week interval. As the sample size increases, the steps will become smaller and closer together, smoothing out the curve.

The Cox proportional hazard model is a useful analytical tool in cohort studies as it helps reduce error due to loss to follow-up from deaths and dropout of subjects. Supposing that in the example study (page 34) all the people with very high blood pressure had their strokes in the first 2 years, while those with normal blood pressure had their strokes 7 or 8 years afterwards. The method used there (logistic regression) would not take this into account. It would then be more useful to compare the time between exposure (having high blood pressure) and the outcome (stroke).

The Cox proportional hazards model is the most widely used method of comparing 'time to event' in medical research. Other appropriate methods include log rank and the Wilcoxon two-sample tests (7).

[^9]9. What are the advantages of a prospective cohort study?
a. It is quick and inexpensive
b. It can be used to determine the temporal relationship between risk factor and disease
c. It is good for looking at rare diseases
d. The outcome following rare exposures can be measured
e. Information can be collected during the study and will not be biased by the outcome
f. It can be used to calculate incidence
10. What are the disadvantages of a cohort study?
a. Only one outcome can be measured
b. The cohort may be a biased sample due to the health-related selection
c. It is expensive and time consuming
d. Only values for relative risk can be calculated

## EXPLANATION: ADVANTAGES AND DISADVANTAGES OF COHORT STUDIES

Advantages of cohort studies are:

- They are longitudinal and, unless a retrospective study, exposure data can be collected from the start and throughout the course of the study. This permits the collection of complete, unbiased information about subjects' exposure before the outcome is known
- A more accurate conclusion of the temporal relationship between the exposure and the following incidence of disease can be drawn
- Multiple outcomes related to a specific or single exposure can be measured
- Incidence rates (absolute risk) can be calculated as well as relative risk
- They can be used to study the outcome of relatively rare exposures.

Disadvantages of cohort studies are:

- Difficulty with the method of selection. Often the way that people are enrolled in such studies is via advertising or choosing groups of people who present to healthcare services frequently. Some famous cohort studies have avoided this problem by recruiting all members of well-defined groups, such as physicians, civil servants or even medical students. There may, however, be problems in generalising the results obtained to the wider population. Initially participants are usually healthy, but over the course of a long study their health may deteriorate as part of a natural progression. This is sometimes referred to as health-related selection (see page 117 on bias)
- Cohort studies are not suitable for rare disease as a huge initial cohort would be needed. Large studies are expensive to carry out, follow-up appointments are time consuming, and a large team is usually involved in data collection
- Loss to follow-up such as migration or death is a huge disadvantage, and the longer the follow-up the greater the loss tends to be. If there is a high dropout rate (> 20 per cent), it becomes difficult to draw accurate conclusions
- Changes over time, such as advances in treatment, may vary exposure over the course of the study and make the results irrelevant
- Exposure may be linked to a hidden confounder that is not measured
- Blinding is difficult in cohort studies; if people know they have very high blood pressure this may have either a positive or negative effect on outcome. They may start to exercise, stop smoking or take medication, thus decreasing their risks, or conversely the added stress of knowing their risk is higher may become a confounding factor that increases their overall risk.

Specific advantages and disadvantages of retrospective cohort studies are:

- If the time taken from exposure to disease outcome is very long, a prospective cohort study may have a high dropout rate; an advantage of a retrospective cohort study is that it only includes subjects for whom complete data from exposure to outcome is available
- A disadvantage is there may be bias in measuring exposure when disease outcome is known.

[^10]This page intentionally left blank


- RANDOMIZED CONTROLLED TRIALS (i) 44
- RANDOMIZED CONTROLLED TRIALS (ii) 46
- ADVANTAGES AND DISADVANTAGES OF RANDOMIZED CONTROLLED TRIALS 48


# INTERVENTION STUDIES: RANDOMIZED CONTROLLED TRIALS 

Ballard C, Margallo-Lana M, Juszczak E, et al. Quetiapine and rivastigmine and cognitive decline in Alzheimer's disease: randomised double blind placebo controlled trial. BMJ 2005; 330:874 (extracts reproduced with permission from the BMJ Publishing Group).
'Ninety-three patients with Alzheimer's disease, dementia, and clinically significant agitation were chosen from care facilities in the north east of England to determine the respective efficacy of quetiapine and rivastigmine for agitation in people with dementia in institutional care. The trial was also designed to evaluate these treatments with respect to change in cognitive performance.
'Patients were randomly sorted into three groups of equal numbers. Each group was assigned the atypical antipsychotic quetiapine plus a placebo, the cholinesterase inhibitor rivastigmine plus placebo, or placebo (double dummy) for 26 weeks. Agitation and cognition were measured at baseline and at six weeks and 26 weeks.
'Seventy-one ( 89 per cent) of the patients tolerated the maximum protocol dose (22 rivastigmine, 23 quetiapine, 26 placebo). Compared with placebo, neither group showed significant differences in improvement on the agitation inventory either at six weeks or 26 weeks. For quetiapine the change in cognitive score from baseline was estimated as an average of -14.6 points ( 95 per cent $\mathrm{Cl}-25.3$ to -4.0 ) lower than in the placebo group at six weeks ( $P=0.009$ ), and -15.4 points ( -27.0 to -3.8 ) lower at 26 weeks ( $P=0.01$ ). The corresponding changes with rivastigmine were -3.5 points ( -13.1 to 6.2 ) lower at six weeks ( $P=0.5$ ), and -7.5 points ( -21.0 to 6.0 ) lower at 26 weeks ( $P=0.3$ ).
'Conclusion: Neither quetiapine nor rivastigmine are effective in the treatment of agitation in people with dementia in institutional care. Compared with placebo, quetiapine is associated with significantly greater cognitive decline.'

1. Answer the following with reference to the study above:
a. What questions was this study designed to answer?
b. What features of the design show it is a randomized controlled trial?
c. Is it an observational or intervention study?
d. What are the likely sources of error in this study?

## 2. The design features listed below reduce error in the study. True or False?

a. Randomizing patients into treatment groups
c. Using placebo treatments
b. The use of blinding
d. The use of double blinding

## EXPLANATION: RANDOMIZED CONTROLLED TRIALS (i)

Randomized controlled trials are considered the 'gold standard' of clinical and epidemiological studies. They are intervention studies that choose a group of patients who are suitable for one or more types of drug or intervention (e.g. surgery). If one treatment is being tested then it can be compared with either placebo or the existing gold standard. Further comparisons may be made between treatments or population subtypes by addition of 'arms' to the trial. Additionally, more than one treatment or combinations of treatments could be compared directly to placebo. A conclusion drawn from a carefully conducted randomized controlled trial that is of adequate size and has used techniques such as double blinding or placebo treatments is considered the most reliable, sometimes referred to as 'top level' or 'level 1', evidence as to whether a treatment is effective or not.

The example on page 44 is an intervention study (1c) to compare quetiapine, rivastigmine and a placebo with respect to (a) levels of agitation, and (b) decline in cognition in patients with Alzheimer's disease, dementia and agitation (1a). Patients were assigned to one treatment group at random: quetiapine plus placebo, rivastigmine plus placebo or placebo only (1b). The likely sources of error in the study arise mainly from small treatment groups and that anxiety and cognition are inherently difficult to assess (1d).

In the best studies all patients are suitable for each of the options, and randomization is done using 'names out of a hat' or random number generation. Generally the aim is for there to be equal numbers in each group. With a large enough trial, random allocation should lead to a very similar distribution of characteristics in each group (mean age, male to female ratio, etc).

Blinding can be single, where either the patient or the doctor is unaware of which treatment has been given; or it can be double in which case neither the patient nor the doctor knows which treatment is being given. In the example given, double blinding was carried out: 'The randomizing clinician faxed a form to the statistician, who communicated allocation to the pharmacy, ensuring concealment.' The term 'double blinding' may be misleading, as often more than two people in the study are blinded, for example the statisticion performing the analysis may be blinded as well as the doctor and patient.

In crossover trials comparisons are not made between patients but within patients when, over a period of time, each patient receives more than one treatment. Each patient is randomly assigned to one treatment 'arm' involving a sequence of two or more interventions given consecutively. One example is that of an $A B / B A$ study, where A and B may be types of drug or treatment, or one a treatment and the other a placebo. In the first round, half of the group is randomly assigned to A, while the other half is randomly assigned to B. After a 'dry out' period the two groups swap. This has the advantage of comparing each subject's response to both treatments and also reveals if the order of treatment has an effect.

Sometimes the effect measured when a patient thinks they are on treatment is significant, this is known as the placebo effect. If part of the group is randomized to 'no treatment', they will be given an identical looking placebo tablet and contact time with health professionals will be the same.

[^11]
## 3. The following are advantages of having a large sample

a. More data for statistical comparison
b. It is cheap and easy to perform a trial on large numbers of people
c. A large initial sample allows for there to be enough data once non-compliance has been taken into account

## 4. What are the disadvantages of having a large sample?

a. It is easier to make type 1 and type 2 errors (see page 113)
b. It is ethically wrong to treat a lot of people with a substandard treatment that can be proved so by exposing far fewer people
c. Even though the conclusions are statistically significant they are weak
5. Using the compliance adjustment formula, adjust the calculated sample size per arm $(N)=100$ for a trial that expects 80 per cent compliance $(c=0.8)$ in each group

Ballard C, Margallo-Lana M, Juszczak E, et al. Quetiapine and rivastigmine and cognitive decline in Alzheimer's disease: randomised double blind placebo controlled trial. BMJ 2005; 330:874 (extract reproduced with permission from the BMJ Publishing Group).

The following explains how the sample size was chosen:
'To detect an average difference of a 6-point (SD 6) change in agitation inventory score from baseline to six weeks between active treatment and placebo with a power of 90 per cent at the 5 per cent (two sided) level of significance, we needed a sample size of 23 in each group, assuming similar efficacy of active treatments. These parameters are based on the effect reported for carbamazepine in a similar study. With allowance for a drop out rate of 25 per cent, we therefore needed 31 patients per treatment group ( $n=93$ ).'
6. In the example above how does the non-compliance affect the sample size?
a. If the non-compliance was 20 per cent how many patients per group would have been needed?
b. If the non-compliance was 30 per cent how many patients per group would have been needed?
c. If the power was 95 per cent at the 2.5 per cent level of significance would this require a larger or smaller sample?

## EXPLANATION: RANDOMIZED CONTROLLED TRIALS (ii)

There are several important reasons to get the sample size right. Firstly the numbers need to be big enough so that if the treatment really does affect clinical outcome, the trial will be able to detect a statistically significant difference. However, the bigger the sample size, the more time and money the trial will cost. Randomized controlled trials may be designed as large-scale, multicentre trials if two very similar treatments are being compared and the difference between them cannot be determined by smaller studies. From an ethical perspective, the closer the two treatments are to each other, or indeed to the 'gold standard', the more reasonable it is to involve hundreds or thousands of people in the trial.

This is a checklist devised for working out how big the sample size should be:

- Estimate the event rate in the control group by extrapolating from a population similar to the population enrolled in the trial
- Determine the smallest difference that will be of clinical importance
- Determine the power (see page 115) for the particular trial
- Determine the significance level or probability of a type 1 error (see page 115) that is acceptable or reasonable (see page 97 on significance levels and $P$-values)
- Adjust the calculated sample size ( $N$ ) for the expected dropout rate or level of non-compliance with treatment.

The level of non-compliance takes into account the proportion of patients allocated to treatment who fail to take it as well as those on placebo who do end up on treatment. A placebo-controlled study needing 100 patients per treatment arm, with 100 per cent compliance, would require over 278 patients per arm if compliance were only 80 per cent in each group. If compliance fell to 70 per cent in each group, the trial would require 625 patients.
The compliance adjustment formula is as follows: Adjusted $n$ per arm $=\frac{N}{\left(c_{1}+c_{2}-1\right)^{2}}$ where $c_{1}$ and $c_{2}$ are
the average compliance or inverse dropout rates per arm.
The figure opposite depicts the exponential increase in numbers needed in the initial sample as the dropout rate or non-compliance increases. The three different lines represent compliance of 70 per cent, 80 per cent and 90 per cent respectively in the control group.


[^12]7. What are the advantages of randomized controlled trials?
a. They can study an intervention for which there is no observational information
b. Large trials are cheap and easy to organize
c. They can be designed with features to eliminate bias
8. What are the disadvantages of randomized controlled trials?
a. They are not considered very highly as a type of epidemiological study
b. By randomizing the subjects the study is open to bias
c. Blinding can cause ethical problems
9. How might study design take advantages and disadvantages into account?
10. Put the options in the order that would suggest 'goodness' of study design from the worst to the best, assuming each feature is both possible and appropriate

## Options

A. Non-randomized
B. Randomized, placebo, crossover and double blinding
C. Randomized, placebo, crossover and single blinding
D. Randomized
E. Randomized, placebo and crossover
F. Randomized with placebo

## EXPLANATION: ADVANTAGES AND DISADVANTAGES OF RANDOMIZED CONTROLLED TRIALS

Advantages of randomized controlled trials are:

- The distribution of confounders is more likely to be even between groups
- Blinding can be incorporated into the study design
- Randomizing subjects to trial groups reduces bias and facilitates statistical analysis.

If a randomized controlled trial is conducted well, it is often possible to be fairly sure that the results are correct, at least for the type of patients who took part in the study. Good clinical trials are almost always randomized and controlled. Randomized controlled trials are therefore usually quoted most frequently when building strong arguments for or against a type of treatment. A carefully conducted, large clinical trial has the power to overturn currently accepted theories. The most limited studies are non-randomized, while the best are usually considered to be the ones with all of the features intended to reduce bias incorporated into the study design; such a study would be randomized, placebo, crossover and double blinded.

Disadvantages of randomized controlled trials are:

- They are generally expensive; a lot of time and money are required
- Volunteer bias still exists in some situations
- There are several issues that lead randomized controlled trials to be ethically problematic at times, such as blinding.

Blinding is often difficult to achieve as it has several ethical and practical problems. The ethical issue is that, for any treatment, a patient should be aware of all the options, understand their implications and give consent for each of them, including a placebo treatment. In order to give consent the patient would need to know any side-effects, thus may be able to guess which treatment they have been given. The practical problems with blinding occur when the treatment options are vastly different, for example a surgeon must know which operation he or she is performing. Patients may also need to change their behaviour according to which group they have been randomized to. An example of this is the randomized controlled trial to decide if a proper warm-up reduces sports injuries; subjects were randomized to either perform a specially devised warm-up routine or not.

A major limitation of many randomized controlled trials is getting a large enough initial sample size. Many fail to account for patients' predictable lack of compliance with their allocated treatments, and also for dropout rates due to mortality, co-morbidity and thus conflicting treatments, or even simply being lost to the trial due to relocation. As losses directly affect the size of the achievable treatment difference, they also affect the required sample size in a non-linear fashion.

[^13]This page intentionally left blank

## SEcTion 7 META-ANALYSIS

- META-ANALYSIS (i)
- META-ANALYSIS (ii)

54, 58

- ADVANTAGES AND DISADVANTAGES OF META-ANALYSIS56


## META-ANALYSIS

Gunnell D, Saperia J and Ashby D. Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review. BMJ 2005;330:385 (extracts reproduced with permission from the BMJ Publishing Group).
'Objective: To investigate whether selective serotonin reuptake inhibitor (SSRI) antidepressants are associated with an increased risk of suicide-related outcomes in adults.
'Design: Meta-analysis of randomized controlled trials of SSRIs compared with placebo in adults submitted by pharmaceutical companies to the safety review of the Medicines and Healthcare products Regulatory Agency (MHRA).
'Participants: Over 40000 individuals participating in 477 randomised controlled trials.
'Main outcome measures: Suicide, non-fatal self-harm, and suicidal thoughts.'

1. What research question is being asked by this study?
2. What is involved in doing a meta-analysis?
3. True of false? The main features that distinguish a good meta-analysis from a poor one are:
a. Only the highest-quality studies are included
b. All studies carry equal weight in the analysis
c. Studies are widely diverse in their design and patient population in order to embrace a variety of different research approaches
d. Unpublished data are identified and included in the analysis
4. Which of the following would you consider strengths or limitations of the study above?

## Options

A. Strengths
B. Limitations

1. Published and unpublished trials were included
2. It makes an important clinical distinction between the three outcomes of suicide, nonfatal self-harm and suicidal thoughts
3. Data were pooled across different SSRIs, assuming that they have similar effects

## EXPLANATION: META-ANALYSIS

Meta-analysis provides a systematic overview by drawing together quantitative primary research that has examined a particular question. Ideally using data on individuals included in each study, meta-analysis statistically combines numerical results of comparable studies or trials. By increasing the numbers of observations and the statistical power, meta-analysis aims to improve the estimates of the size of an effect or intervention or the strength of an association. The formal process of combining results from all relevant studies looking at the same health question to arrive at a final conclusion is essential for the practice of evidence-based medicine: 'the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients ${ }^{1}$.

The study on page 52 is asking if treatment with SSRIs increases an individual's risk of suicide, self-harm or suicidal thoughts (1). Meta-analysis requires $(2,3)$ :

- Similar patient populations used in every study
- The use of comparable studies with quantitative results obtained in similar ways
- All relevant studies, published and unpublished, must be included and not just well-known ones. The problem is that work may have been done in an area, but if no significance was found it might not have been published. In theory this work should be included
- Inclusion and exclusion criteria must be defined before starting the study. The correct balance should be found, as being too restrictive limits the ability to generalize when drawing conclusions, however being too inclusive weakens the confidence that can be placed in the findings
- Larger studies should carry more 'weight' than smaller studies when combining results. A simple approach to weighting is by sample size, as a study based on 100 subjects is assumed to produce a more 'precise' estimate of the population parameter than a study based on 10 subjects. However, a more accurate approach is to weight each study by its 'inverse variance.' A study with a small variance is more precise. Therefore, the inverse variance ( $1 /$ variance) would provide a greater weighting for that study.

It is rare for all these conditions to be met, therefore meta-analysis remains a controversial statistical method for reviewing the literature. For example, in the study on page 52:

- A systematic review of the literature was not carried out, as only the MHRA report was used, therefore some relevant trials may have been excluded
- Publication bias: the trials included were done by drug companies which may have had a vested interest in publishing a particular result. Independent research may show the drugs in a less favourable light
- Most of the component studies were short-term randomized controlled trials, therefore the meta-analysis cannot comment on the longer-term effects of SSRIs.

[^14][^15]Gunnell D, Saperia J and Ashby D. Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review. BMJ 2005;330:385 (summary of results reproduced with permission from the BMJ Publishing Group).

Results of example study: Forest plots of suicide, non-fatal self-harm and suicidal thoughts in placebo controlled trials of SSRIs can be found on page 58.

The results were:

- An estimated 16 suicides, 172 episodes of non-fatal self-harm, and 177 episodes of suicidal thoughts were reported
- There was no evidence that SSRIs increased the risk of suicide, but important protective or hazardous effects cannot be excluded (odds ratio $=0.85$, 95 per cent credible interval $=$ 0.20 to 3.40 )
- There was weak evidence of an increased risk of self-harm (odds ratio $=1.57$, 95 per cent credible interval $=0.99$ to 2.55 )
- Risk estimates for suicidal thoughts were compatible with a modest protective or adverse effect (odds ratio $=0.77$, 95 per cent credible interval $=0.37$ to 1.55 )
- The relative frequency of reported self-harm and suicidal thoughts in the trials compared with suicide indicates non-fatal end-points were under-recorded.

5. What do the forest plots indicate about the estimated risk with different SSRIs of
a. Suicide?
b. Self-harm?
c. Suicidal thoughts?
6. Would these results lead you to change your clinical practice?
7. What, if anything, would you tell patients who are starting SSRIs, having seen this evidence?
[^16]
## EXPLANATION: META-ANALYSIS

To perform an effective meta-analysis:

- Evaluate the results from individual studies. Ideally look at the raw data but if this is not available for each study, results can be assessed using published results
- Estimate the degree of variability in the results between studies. This variability is known as the 'statistical heterogeneity'
- Obtain a numerical estimate for the average effect by combining results from all the studies. When comparing studies the 'size of the effect' should be comparable, and will often be expressed as an odds ratio, a relative risk or a standardized mean difference
- Determine the strength and significance of the overall effect by obtaining a 95 per cent or 99 per cent confidence interval and performing a hypothesis test
- Present the findings. This is commonly done using a graphical display known as a forest plot. This shows the estimated effect size, such as relative risk or risk ratio, obtained from each contributing study. Each relative risk estimate is plotted using a square shape, where the size of the square is proportional to the size of the study. In general, the larger the study, the larger the square used to represent the estimate from that study. Confidence intervals for each estimate are also plotted, as shown in the figure below. Forest plots also show the overall estimate obtained from analyses of results from all studies combined. This overall estimate is indicated by a diamond shape near the bottom of the graph. Here the length of the diamond indicates the width of the confidence interval for the overall estimate.

The forest plot allows readers to find quickly the answer to the following questions:

- Is the effect from each study on the same side of the vertical line?
- How precise are the results from each study, i.e. the width of each confidence interval?
- Are the results from the different studies compatible, i.e. do the confidence intervals overlap?
- Is the effect significant? Do confidence intervals cross the vertical line?

Whether or not these results would lead to a change in a doctor's clinical practice is a subjective matter (6). Quoting from the conclusion of the study (page 54):


Graph reproduced wth kind permission from Everitt BS and Palmer CR (eds), Encydopaedic Companion to Medical Statistics, Great Britain: Hodder Education, 2005
'Increased risks of suicide and self harm caused by SSRIs cannot be ruled out, but larger trials with longer follow up are required to assess the balance of risks and benefits fully. Any such risks should be balanced against the effectiveness of SSRIs in treating depression.'
Since there is no definite evidence either way to say that SSRIs increase the risks of suicide, the study recommends: 'When prescribing SSRIs, clinicians should warn patients of the possible risk of suicidal behaviour and monitor patients closely in the early stages of treatment.' (7).

[^17]8. Which of the following are advantages of meta-analysis?
a. Less power
c. Generation of new data
b. Smaller sample size needed
d. More precise
9. Which of the following are possible disadvantages of meta-analysis?
a. Dependence on publication
d. Risk of excluding unpublished data
b. Costly
e. Incompatibility between component study
c. Loss to follow-up designs
10. Match up the advantages to the study designs from the options given

## Options

A. Ecological study
C. Case-control study
E. Randomized controlled trial
B. Cross-sectional study
D. Cohort study
F. Meta-analysis

1. Can compare widely differing populations
2. Can provide an estimate of relative risk
3. Can summarize results of all the studies asking a similar question
4. Can demonstrate a temporal relationship between exposure to a risk factor and development of disease
5. Can directly measure the effects of an intervention
6. Can estimate prevalence
7. Match up the following diasdvantages to the study designs from the options given

## Options

A. Ecological study
C. Case-control study
E. Randomized controlled trial
B. Cross-sectional study
D. Cohort study
F. Meta-analysis

1. Risk of loss to follow-up
2. Cannot make inferences about individual-level risks
3. Cannot estimate incidence
4. Cannot prove the existence of a temporal relationship between risk factor and effect
5. May mask heterogeneity between individual studies and thus produce an invalid conclusion
6. Expensive and time consuming

## EXPLANATION: ADVANTAGES AND DISADVANTAGES OF META-ANALYSIS

Advantages of meta-analysis are:

- Condenses large amounts of information into managable portions
- Greater power and precision to detect an effect than the individual component studies, due to its larger sample size
- Faster and cheaper than performing a whole new study
- Avoids unnecessary repetition of previous studies
- Can be used to generalize results to a wider population
- Can be used to assess consistency between studies
- Permits assessment of the quality of existing evidence and identifies outstanding research questions to be addressed
- May help to resolve conflicts or uncertainties in original trials
- May be the study of choice if it is not practical or ethical to conduct a definitive trial.

Disadvantages of meta-analysis are:

- Publication bias: some data may not have been published if the results were insignificant. A full metaanalysis should identify unpublished as well as published studies to provide a non-biased review
- Inconsistency between studies in characteristics such as patient population, methodology, outcome measures and follow-up procedure may mean that the results of the different studies cannot be compared directly
- Studies may differ in the quality of their design. Although this should be accounted for by weighting, the weighting value may be subjective and arbitrary
- Dependence: the results of one study may be published more than once, and therefore the data from different studies in the meta-analysis may not be entirely independent
- Important qualitative information may be obscured by 'averaging' simple numerical representations across studies. This is especially the case if there are two outcomes or a bimodal population.


## Answers

8. F F F T
9. T F F T T
10. 1 - A, $2-\mathrm{C}, 3-\mathrm{F}, 4-\mathrm{D}, 5-\mathrm{E}, 6-\mathrm{B}$
11. $1-\mathrm{D}, 2-\mathrm{A}, 3-\mathrm{B}, 4-\mathrm{C}, 5-\mathrm{F}, 6-\mathrm{E}$

Gunnell D, Saperia J and Ashby D. Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review. BMJ 2005;330:385 (graphs reproduced with permission from the BMJ Publishing Group).

Forest plots of suicide, non-fatal self-harm and suicidal thoughts in placebo controlled trials of SSRIs.
(a) Meta-analysis of suicide data ${ }^{1}$

| Study | Odds ratio (95\% Cl) | Odds ratio (95\% Cl) |
| :---: | :---: | :---: |
| Citalopram |  | 0.47 (0.03 to 7.54) |
| Escitalopram |  | 0.79 (0.05 to 12.61) |
| Fluvoxamine |  | 0.81 (0.11 to 5.76) |
| Paroxetine | - | 0.23 (0.02 to 2.19) |
| Sertraline |  | 6.42 (0.35 to 119.20) |
| Overall |  | 0.74 (0.25 to 2.21) |
| 0.01 | 0.10 .51251020 | 080 |
| Favours | SSRI Favours | acebo |

(b) Meta-analysis of non-fatal self-harm data (excluding and including paroxetine)

| Study | Odds ratio (95\% Cl) | Odds ratio (95\% Cl) | Study | Odds ratio (95\% Cl) | Odds ratio (95\% Cl) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Citalopram | $\xrightarrow[+1]{\square}$ | 1.04 (0.36 to 3.00) | Citalopram | $\xrightarrow[+]{1}$ | 1.04 (0.36 to 3.00) |
| Escitalopram |  | 4.74 (0.57 to 39.40) | Escitalopram | $1 \square$ | 4.74 (0.57 to 39.40) |
| Fluoxetine |  | 1.03 (0.48 to 2.20) | Fluoxetine | - | 1.03 (0.48 to 2.20) |
| Fluvoxamine |  | 1.95 (0.93 to 4.09) | Fluvoxamine |  | 1.95 (0.93 to 4.09) |
| Sertraline |  | 1.78 (0.78 to 4.05) | Paroxetine |  | 0.87 (0.52 to 1.45) |
| Overall | T | 1.50 (1.00 to 2.25) | Sertraline |  | 1.78 (0.78 to 4.05) |
| 0.01 | $0.1 \quad 0.51251020$ | 4080 | Overall |  | 1.26 (0.88 to 1.80 ) |
| Favours SSRI Favours placebo |  |  |  |  |  |
|  |  |  | Favours SSRI Favours placebo |  |  |

(c) Meta-analysis of suicidal thoughts data (excluding and including paroxetine)

| Study | Odds ratio (95\% Cl) | Odds ratio (95\% Cl) | Study | Odds ratio (95\% Cl) | Odds ratio (95\% Cl) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Citalopram | + | 1.18 (0.37 to 3.78) | Citalopram | $1$ | 1.18 (0.37 to 3.78) |
| Escitalopram | - | 0.39 (0.04 to 4.35) | Escitalopram |  | 0.39 (0.04 to 4.35) |
| Fluoxetine |  | 0.45 (0.26 to 0.77) | Fluoxetine | - | 0.45 (0.26 to 0.77) |
| Fluvoxamine |  | 1.56 (0.77 to 3.14) | Fluvoxamine |  | 1.56 (0.77 to 3.14) |
| Sertraline |  | 0.71 (0.22 to 2.21) | Paroxetine |  | 0.84 (0.50 to 1.41$)$ |
| Overall |  | 0.80 (0.43 to 1.51) | Sertraline |  | 0.71 (0.23 to 2.21) |
| 0.01 | 0.10 .51251020 |  | Overall | ! | 0.81 (0.53 to 1.26) |
| Favours SSRI Favours |  | ebo | 0.01 | $0.1 \quad 0.5125102040$ |  |
|  |  |  | Favour | SSRI Favours | acebo |



- PRINCIPLES OF SCREENING

60

- IMPLEMENTATION OF SCREENING PROGRAMMES62
- SENSITIVITY AND SPECIFICITY (i) ..... 64
- SENSITIVITY AND SPECIFICITY (ii) ..... 66
- LIKELIHOOD RATIOS AND BAYES' THEOREM ..... 68

1. What is the purpose of a screening programme?
2. Match the target population to the appropriate screening test from the options given

## Options

A. Genetic test for CFTR mutation
B. Mammography every 3 years
C. Guthrie test
D. Genetic test for sickle mutation
E. Smear test every 3 years

1. All women in UK aged $25-60$ years
2. Southern Mediterranean/African-Caribbean descent
3. Family history of cystic fibrosis
4. All women in UK aged $50-64$ years
5. Neonates
6. Regarding the Guthrie test
a. What does it involve?
b. When is it carried out?
c. What does it test for?
7. Which potential new screening programmes are currently under evaluation and which tests could be used for each of them?
8. What epidemiological study designs can be used to evaluate a screening programme?

## EXPLANATION: PRINCIPLES OF SCREENING

A definition of screening from Wald and Cuckle, $1989^{1}$ is:
'The identification, among apparently healthy individuals, of those who are sufficiently at risk of a specific disorder to justify a subsequent diagnostic test or procedure, or in certain circumstances, direct preventative action.'

Screening is not a diagnosis in itself. Its aim is to identify individuals who are at risk of disease at a pre-symptomatic stage (1). These individuals should then should undergo further investigation so that if early-stage disease is present, it can be diagnosed and treated early to improve prognosis. Screening is an example of a primary prevention strategy.

Examples of UK screening programmes are given in the following table.

| Level | Whom for | Condition detected | Test |
| :---: | :---: | :---: | :---: |
| National | Newborn | Phenylketonuria Congenital hypothyroidism | Guthrie test: heel prick at $5-10$ days (3) |
|  | Women aged 50-64 years | Early breast cancer | Mammography every 3 years |
|  | Women aged 25-60 years | Cervical intraepithelial neoplasia | Cervical smear test every 3 years |
|  | Women aged 61-65 years | Cervical intraepithelial neoplasia | Cervical smear test every 5 years |
| Targeted 'at risk' groups | Antenatal (ethnic origin/ positive family history) | $\left.\begin{array}{l} \text { Sickle cell disease } \\ \text { Cystic fibrosis } \\ \text { Duchenne muscular } \\ \quad \text { dystrophy } \end{array}\right\}$ | Genetic tests |
| Local | Rubber/dye industry workers | Bladder cancer | Cystoscopy |
| Coming soon | Newborn <br> Antenatal | Congenital deafness Congenital hepatitis B | Otoacoustic emissions Fetal blood sampling |
| Under evaluation (4) |  | Colorectal cancer Ovarian cancer Down's syndrome Prostate cancer Abdominal aortic aneurysm | Colonoscopy <br> CA125 <br> Triple test <br> Prostate-specific antigen <br> Ultrasound scan |

${ }^{1}$ Wald N and Cuckle H. Reporting the assessment of screening and diagnostic tests. British Journal of Obstetrics and Gynaecology 1989;96(4):389-396.

## Answers

1. See explanation
2. $1-E, 2-D, 3-A, 4-B, 5-C$
3. See explanation
4. See explanation
5. All study designs: randomized controlled trials are gold standard but cross-sectional, case-control and cohort studies are also used
6. These criteria that must be fulfilled for a screening programme to be introduced. True or false?
a. The screening test must be a diagnostic investigation
b. The test must have a high specificity, even at the expense of sensitivity
c. It must be the first primary prevention strategy to be implemented for that condition
d. There must be a known disease marker or early symptomatic stage
e. The programme should fulfill a demand for more evidence on the efficacy of existing treatments
7. Make a list of the pros and cons of screening programmes
8. What ethical issues surround the implementation of a new screening programme?
9. What would be the pitfalls of a screening programme for prostate cancer?
10. What would be the issues surrounding a screening programme for Down's syndrome?

## EXPLANATION: IMPLEMENTATION OF SCREENING PROGRAMMES

## National Screening Committee criteria are as follows:

## The condition:

- Should be an important health problem
- Should have well-understood epidemiology and natural history
- Should have a detectable risk factor, disease marker or early symptomatic period
- Other alternative primary prevention methods should have already been implemented.


## The test:

- Should be safe, simple, precise and validated
- Should have high sensitivity and specificity, i.e. low false-positive and false-negative rates
- The distribution of test values in the population should be known and suitable cut-off values defined
- It should be acceptable to the population
- There should be an agreed policy on further diagnostic investigation of individuals with a positive result.


## The treatment:

- An effective treatment should be available (with outcomes better than if not treated).


## The programme:

- There should be evidence that the screening programme reduces morbidity and mortality
- It should be acceptable to health professionals and the general population
- The benefit should outweigh physical and psychological harm caused by the screening programme
- It should be cost-effective with adequate staffing and facilities to carry out the programme.

The pros and cons of screening programmes are shown in the table below (7).

| Pros | Cons |
| :--- | :--- |
| Early detection of disease: better prognosis | Longer morbidity time: earlier diagnosis, unaltered <br> prognosis |
| Cost-effective: primary prevention cheaper than <br> expensive treatment | Expensive and time consuming to run |
| Less radical treatment may be required | Risks associated with screening test itself |
| Method of ensuring contact with health services | Anxiety and distress caused by false positives |
| Reassurance of normality | Some cases may be missed: false negatives, not all <br> the 'at-risk' population included in programme <br> Incidental findings that might otherwise cause no harm <br> need investigating and cause distress |

## Answers

6. F F F T F
7. See explanation
8. Ownership of information, antenatal screening, consent, medicalization of healthindividuals
9. Prostate-specific antigen (PSA) is not specific, treatment may not alter prognosis, there are insufficient resources
10. Maternal age to start of screening, risk of invasive diagnostic tests for a positive screening result, selective termination, parental anxiety and distress, test not 100 per cent specific or sensitive

Use the following example to answer the questions on sensitivity and specificity.
Im Y-J, Kim J-W, Hong S-J and Chung B-H. Proper cut-off value of free to total PSA ratio for detection of prostate cancer in Korean men. Yonsei Medical Journal 2004;45(5):873-878 (extracts reproduced with permission from the Yonsei Medical Journal).
'To assess whether the free-to-total prostate specific antigen (PSA) ratio (F/T PSA ratio) would enhance prostate cancer detection, 240 Korean men, whose serum PSA levels were between 4 and $20 \mathrm{ng} / \mathrm{ml}$, were enrolled in a two-year study. All patients underwent ultrasound-guided transrectal biopsies of the prostate gland. The F/T PSA ratio was measured using the Roche immunoassay. Overall, when the cut-off value of the F/T PSA ratio was 0.10 , the sensitivity and specificity were 75.0 per cent and 76.5 per cent, while for the cut-off value of 0.15 they were 83.3 per cent and 39.7 per cent respectively.
'Results: Of the 240 patients, 202 ( 84 per cent) had benign histology, while 38 ( 16 per cent) had prostate cancer. The two patient groups were well matched for age. The mean F/T PSA ratio showed a statistically significant difference between the two groups.
'Conclusion: Our data demonstrated the usefulness of the free to total PSA ratio in distinguishing benign prostate disease and cancer disease, hence eliminating unnecessary biopsies.'
11. What is sensitivity?
12. What is specificity?

FN, false negative; FP, false positive; PSA, prostate-specific antigen; F/T PSA, free-to-total prostate-specific antigen; TN, true negative; TP, true positive

## EXPLANATION: SENSITIVITY AND SPECIFICITY (i)

Sensitivity and specificity are measures of the reliability of clinical tests to detect correctly disease status. The closer the sensitivity and specificity are to 100 per cent, the more reliable the test. Sensitivity and specificity are important as they give both doctors and patients an idea of how much faith to put in test results and how much they should use them to guide their clinical decisions.

In an ideal world, we would like our tests for disease to be 100 per cent sensitive and 100 per cent specific but in reality there is usually a trade-off whereby sensitivity may be increased at the expense of specificity or vice versa. This is often achieved through altering the cut-off value for a test result at which we believe that an individual has a high chance of having the disease. This is illustrated in the figure below.


Sensitivity, specificity and predictive values can be calculated most easily by representing the different types of test results (true positive, false positive, true negative, false negative) in a table, as shown below.

| Test result | Disease present | Disease not present | Total |
| :--- | :--- | :--- | :--- |
| Positive | True positive (TP) | False positive (FP) | TP + FP |
| Negative | False negative (FN) | True negative (TN) | FN + TN |
| Total | TP + FN | FP + TN | TP + FP + TN + FN |

Sensitivity $=\mathbf{T P} /(\mathbf{T P}+\mathbf{F N})(11)$. This is the probability of correctly detecting the presence of disease or the proportion of true positives in the diseased population. A very sensitive test has few false negatives. High sensitivity is desirable for serious, treatable conditions where false negatives could lead to preventable deaths or serious morbidity. It might help to remember this by using the word 'SNOUT' (high SeNsitivity rules OUT).

Specificity $=\mathbf{T N} /(\mathbf{T N}+\mathbf{F P})$ (12). This is the probability of correctly concluding that an individual is not affected or the proportion of true negatives in the healthy population. A very specific test has very few false positives. This is desirable for serious but non-treatable conditions, or for when the diagnostic test is risky or expensive and it would be harmful psychologically or physically for the patient to receive a false-positive result. This can be remembered using the word 'SPIN' (high SPecificity rules IN)

[^18]13. From the example on page 64 why did sensitivity increase and specificity and decrease when the cut-off value was raised to 0.15 ?
14. In detecting prostate cancer is it more important to have a highly sensitive test or a highly specific test?
15. What are the ethical issues surrounding having a high false-negative rate?
16. What are the problems with having a high-false positive rate?
17. Are the following true or false?
a. A specific test brings about few false positives
b. High specificity is important when it would be dangerous to have a false-negative result
c. Specificity is not usually affected by the prevalence of the disease but may affect the detection rate
d. Specificity is the proportion of those who are truly non-diseased to those who are identified as so by the screening test
e. A sensitive test is a measure of the probability of incorrectly diagnosing a case
f. Sensitivity is the true positive rate
g. A highly sensitive test is best used when false-positive errors are undesirable due to expensive or invasive further investigations
h. A study that uses a sensitive test is less likely to make a type 1 error
18. What are the problems associated with comparing the sensitivity and specificity of a new test relative to the existing 'gold standard'?
19. Regarding the predictive value
a. What is the positive predictive value?
b. What is the negative predictive value?
c. What information is used to calculate these?

## EXPLANATION: SENSITIVITY AND SPECIFICITY

'Overall, when the cut-off value of the F/T PSA ratio was 0.10 , the sensitivity and specificity were 75.0 per cent and 76.5 per cent, while for the cut-off value of 0.15 they were 83.3 per cent and 39.7 per cent respectively.'

When the cut-off value was raised in the study example, sensitivity was increased as more true positives were picked up. Specificity decreased as there were more false positives (13). False positives can occur if free PSA increases transiently in proportion to total PSA, for example during a urinary tract infection.

In the care of prostate cancer, a highly sensitive test would be better than a highly specific test as it is important to pick up disease that might be life threatening (14). A high false-negative rate would lead to underdiagnosis and a delay in treatment. In these people a screening test with a false-negative result may be more dangerous than not being screened at all (15). The problem with a high false-positive rate is that people without the disease will undergo unnecessary anxiety, tests, treatments or even procedures (16).

Sensitivity and specificity of a test can only be calculated if the true diagnosis can also be known. Often the closest we can get to a true diagnosis is the best available diagnostic test or current 'gold standard' used for detecting the disease. Care must be taken when calculating sensitivity and specificity of new tests by comparing them to 'gold standards' which in themselves are not 100 per cent accurate (18). In the example given, the 'gold standard' used was a transrectal ultrasound biopsy of the prostate. This test is highly specific but not quite as sensitive, as cancer cells can be missed when samples are taken, leading to false-negative results.

Positive predictive value $(\mathbf{P P V})=\mathbf{T P} /(\mathbf{T P}+\mathbf{F P})(19 a, \mathbf{c})$. This is the probability of a person with a positive result actually having the disease. PPV is related to sensitivity in that the more sensitive the test, the greater the predictive value of a positive test result. For a very sensitive test, a negative result effectively rules out the presence of a disease. (SNOUT, see page 65)

Negative predictive value $(\mathbf{N P V})=\mathbf{T N} /(\mathbf{T N}+\mathbf{F N})(19 b, \mathbf{c})$. This is the probability of a person with a negative result being disease-free. NPV is related to specificity in that the more specific the test, the greater the predictive value of a negative test result. For a very specific test, a positive result effectively confirms the presence of disease. (SPIN, see page 65)

Predictive values are affected by the prevalence of disease. High prevalence results in high PPV or low NPV. This is because if a disease has high prevalence the number of true positives will be much greater than the number of false positives, creating a larger numerator when calculating the PPV. The converse is true for low prevalence where the number of false positives is greater relative to the number of true positives.

[^19]Im Y-J, Kim J-W, Hong S-J and Chung B-H. Proper cut-off value of free to total PSA ratio for detection of prostate cancer in Korean men. Yonsei Medical Journal 2004;45(5):873-878 (extract reproduced with permission from Yonsei Medical Journal).
'Korean men whose serum total PSA levels were between 4 and $20 \mathrm{ng} / \mathrm{ml}$ were enrolled in a two-year study, [the known background prevalence of prostate cancer in men with PSA levels between 4 and $20 \mathrm{ng} / \mathrm{ml}$ is 15 per cent]. Of the 240 patients, 202 ( 84 per cent) had benign histologies, while 38 ( 16 per cent) had prostate cancer. The two patient groups were well matched for age. The mean F/T PSA ratio showed a statistically significant difference between the two groups. Overall, when the cut-off value of the F/T PSA ratio was 0.10 , the sensitivity and specificity were 75.0 per cent and 76.5 per cent, while for the cut-off value of 0.15 they were 83.3 per cent and 39.7 per cent respectively.'

## 20. Regarding prostate cancer in this group of men what is

a. the pre-test or prior probability?
b. the post-test or posterior probability?

Look at the following Bayesian graphs of post-test probability as a function of test result and pre-test probability.
The middle (solid) line shows that if there are no test results or if the test is completely irrelevant (e.g. the results in the diseased population are equal to the background population) there is no difference in pre- and

 post-test probability. The top (dashed) line shows how the post-test probability of disease is increased given a positive test result. The bottom (dotted) line shows how the post-test probability of disease is decreased given a negative test result.
21. With regard to the graphs above one shows how the results would alter pre- and posttest probabilities when the cut-off value of the free-to-total prostate-specific antigen ( $\mathrm{F} / \mathrm{T}$ PSA) ratio was 0.10 and the other when the cut-off value was 0.15 .
a. Use the sensitivity and specificity values obtained in the study to work out which graph is which
b. Which cut-off ratio increases the positive post-test probability the most?
c. Which cut-off ratio of the two would be best if the F/T PSA ratio is used as a new screening test?

## EXPLANATION: LIKELIHOOD RATIOS AND BAYES' THEOREM

## Likelihood ratio (LR) for a positive result = sensitivity/(1 - specificity)

This is the ratio of the probability of a positive result if a patient has the disease to the probability of a positive result if the patient does not have the disease. It is helpful in making clinical decisions and is applied in Bayes' theorem as it tells us how much more likely a person is to have a disease if they have a positive test.

Bayes' theorem can be described as follows:
When a patient first walks through the door of a doctor's clinic and the symptoms are unknown, the risk of them having a certain disease is referred to as the 'prior probability'. If nothing else is known about the patient, the prior probability is equal to the prevalence of that disease in the general population. Once the patient's symptoms, signs or test results are known, their risk of having the disease is referred to as the 'posterior probability'. This represents the probability that the patient has the disease, given a positive clinical finding or test result. In this sense it is similar to the positive predictive value, but it has the added advantage of taking into account the prior probability that an individual has the disease. The posterior probability is affected by the sensitivity and the specificity of the symptom, sign or test result for that disease.

One example might be the occurrence of anaemia in the population of patients presenting to their doctor. Using the hypothetical population prevalence of 5 per cent, anyone walking through the door may have a 5 per cent chance of a haemoglobin level below the normal range. However when the patient starts giving a history of symptoms such as tiredness or breathlessness, or on examination the patient is found to have pale conjunctiva, their likelihood of having anaemia increases. Then, if the doctor does a blood test and the full blood count shows that the patient's haemoglobin level is below the normal range, the posterior probability that the patient is anaemic becomes even greater.

Bayes' theorem can be expressed using the likelihood ratio to give an idea of the 'odds' of disease. The odds of a disease after learning about a symptom or test result, are equal to the odds of the disease before the test, multiplied by the likelihood ratio.

> Posterior odds of disease $=$ prior odds $\times$ likelihood ratio of a positive test result
> Prior odds $=$ prior probability $/(1-$ prior probability $)$

Posterior probability $=$ posterior odds/(1 + posterior odds)

[^20]This page intentionally left blank


This page intentionally left blank

## sECTION 9 DESCRIBING DATA

- DISPLAYING DATA 74
- CENTRAL LOCATION 76
- SPREAD 78
- DISTRIBUTION 80


## 0 DESCRIBING DATA

1. How would you display the following sets of data, choose from the options given

## Options

A. Scatter plot
B. Pie chart
C. Histogram

1. The relationship between proximity to site of explosion and levels of contaminant in urine
2. The distribution of mean red blood cell volume
3. Causes of the wrong blood being transfused in patients in a UK hospital
4. The relationship between blood creatinine and urea levels in a series of patients
5. Causes of stroke by subtype
6. The distribution of weights of all the staff in a GP practice
7. Match the following distributions of data to how they should be represented graphically

Options
A. J-shaped
B. Skewed to
C. Skewed to
D. Bimodal
E. Normal right (positive) left (negative)






1. The heights of men and women in a small village
2. The rate of growth of a child
3. The incidence of heart disease to units of red wine consumption
4. Period of gestation
5. The blood pressures recorded over a year by a GP

## 3. Consider skewed distributions

a. When the mean is greater than the median the skew is positive
b. When the skew is positive the mean is less than the mode
c. If the skew is negative the mean is greater than the median
d. If the skew is negative the mode is greater than the median

## EXPLANATION: DISPLAYING DATA

Histograms, scatter plots and pie charts: usually one or more of these methods would be appropriate for displaying information. The one that is chosen should produce the clearest overview, so that the data can be communicated 'at a glance', however it is often just a matter of the preference of the author. As a general rule:

- Pie charts are useful for emphasizing how each category compares with the sample as a whole, and can be used for a
 single series of numbers shown as percentages of the total, the figure opposite shows causes of stroke by subtype
- Histograms are good for showing the shape and distribution, as the area under each histogram block is proportional to the number of subjects in that particular group, the figure opposite shows the weights of all staff in a GP practice
- Scatter plots can show the association between quantitative
 variables. If one variable, $a$, is dependent on the other, $b$, then it is usual to plot $a$ on the $y$ axis and $b$ on the $x$ axis. The figure opposite of blood urea and creatinine levels in a series of patients suggests that there is a positive association between these two measurements.
If the majority of results are closer to the top of the range but there are a few spread out at the bottom of the range, it results in a 'negative skew'. If most of the results are near the bottom
 of the range the skew is 'positive'.


Graphs reproduced with permission from Puri BK, Statistics for the Health Sciences, London: WB Saunders, 1996
When data are skewed positively there are a few large numbers or 'outliers' that will affect the calculated average (mean). The effect on the mean of a positive skew is to increase it, because the very large numbers are included in the calculation. When ordered according to size, the middle value of the results (median) is less affected by a skewed distribution. The number that occurs most (mode) is least affected by a skewed distribution (3).

## Answers

1. $1-\mathrm{A}, 2-\mathrm{C}, 3-\mathrm{B}, 4-\mathrm{A}, 5-\mathrm{B}, 6-\mathrm{C}$
2. $1-\mathrm{D}, 2-\mathrm{B}, 3-\mathrm{A}, 4-\mathrm{C}, 5-\mathrm{E}$
3. T F F T
4. This is a set of haemoglobin $\mathrm{A}_{1 \mathrm{c}}\left(\mathrm{HbA}_{1 c}\right)$ measurements (in mmol/L) taken at a children's diabetic clinic, used to gauge their blood sugar control over the preceding months. The target $\mathrm{HbA}_{1 \mathrm{c}}$ for this group of children is $<8 \mathrm{mmol} / \mathrm{L}$. The consultant wants to know the extent to which the levels are reaching their target
5.9
15.1
5.9
7.3
6.5
6.2
6.3
6.7
6.9
10.0
5.9
7.1
8.2
5.5
6.6
6.7
a. What is the mean?
b. What is the median?
c. What is the mode?
d. Which of these results gives the best picture of how blood glucose is being controlled in this population of children?
5. The consultant now wants to know what insulin injection devices he should order for his next clinic. He can get individual ones but it would be cheaper to bulk buy the ones he needs the most so he notes down the doses (units) the children are on

| 16 | 22 | 30 | 30 | 34 | 15 | 30 | 56 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 28 | 30 | 26 | 14 | 18 | 20 | 30 | 30 |

a. What is the mean?
b. What is the median?
c. What is the mode?
d. Which would be the most useful value for him to know if the pre-filled disposable injections are sold at volumes of 30 units or 78 units?
6. Match the following statements to the most appropriate measure of location

## Options

A. Mean
B. Median
C. Mode

1. It takes into account all the data values
2. It is the value that occurs most frequently in a set of data
3. It is used in the calculation of standard deviation
4. There are equal numbers of observations above and below this value
5. It is vulnerable to outliers (values much higher or lower than the average)
6. It is obtained by putting the data in the order from smallest to largest
7. Outliers least affect it
8. It is useful in categorical data to describe the group with the highest frequency

## EXPLANATION: CENTRAL LOCATION

In most situations simply listing a series of results does not provide a good quantitative analysis. The 'central location' of the results can be described better by using the mean, median or mode, depending on which is most suitable.

The mean is the calculated average. In a set of data all the values are used to calculate the mean by adding them together and then dividing by the number of data values, sometimes referred to as $n$ or the sample size. The mean is represented as $\bar{x}$; and can be calculated using the formula below.

$$
\bar{x}=\frac{\text { Sum of individual observations }}{\text { Total number of observations }}=\frac{\Sigma x}{n}
$$

In the example given, $n$ represents the 16 children who had a blood test at the clinic that morning and the sum of the blood result values is 116.8 . The mean is therefore 116.8 divided by 16 which is equal to $7.3 \mathrm{mmol} / \mathrm{L}(4 \mathrm{a})$.

Advantages of the mean include that it takes into account all of the data and is used as the central point around which the standard deviation is calculated. The main disadvantage is that the mean can be greatly affected by outliers, especially in a small sample. In the example most children's $\mathrm{HbA}_{1 \mathrm{c}}$ levels were below the mean but because one child presented that morning with a figure that was high ( $15.1 \mathrm{mmol} / \mathrm{L}$ ) the glucose control of the rest of the patients would not be well reflected if the mean value was the only one reported.

The median is the measure of the centre of the distribution. To calculate the median the numbers simply need to be put in order and the middle one is chosen. If the number of data values is even the mean of the two middle ones is taken. In the example the numbers 6.6 and 6.7 lie either side of the middle of the data sample so the median is $6.65 \mathrm{mmol} / \mathrm{L}$ (4b).

The main advantage of the median is that it is more robust to outliers and in small samples like the $\mathrm{HbA}_{1 \mathrm{c}}$ measurements it is much more representative of the glucose control in the group. It is also useful in skewed distributions. However, when the data follow a normal distribution (see page 81) the mean is preferable to the median as a measure of central location.

The mode is the value that occurs most frequently. It is very rarely used in statistical analysis of epidemiological data, however the example given shows a useful practical application. The mode is more useful in grouped or categorical data such as age ranges, glove sizes or catheter lengths. It is least affected by outliers, but may not represent values close to the mean or median either.

[^21]7. Choose the correct words ( $\mathrm{i}-\mathrm{v}$ ) and diagrams ( $\mathrm{A}-\mathrm{E}$ ) to match the description of the range of data given (1-5), assuming normal distribution

## Options

A.

B.

C.

D.

E.

i. Range
ii. Interquartile range
iii. One standard deviation
iv. Two standard deviations
v. Standard error

1. The values that are 34 per cent either side of the mean, representing 68 per cent of the sample population
2. The smallest value to the largest value
3. The value at the top of the first quarter to the bottom of the third quarter when the numbers are put in order from smallest to largest
4. Measures the degree of sampling variability when calculating an estimate of the population mean
5. The values that are 47.5 per cent either side of the mean, representing 95 per cent of the sample population
6. What is the difference between standard deviation and standard error?

## EXPLANATION: SPREAD

The range is the whole dataset from the minimum to the maximum values. It takes into account the outliers and is usually used when representing quantitative data. Interquartile range is calculated by putting the data in numerical order and then discarding the lower and upper quarters of the data. The advantage of the interquartile range is that it shows the range of the middle half of the set of measurements, and thus removes the risk of misrepresenting the data distribution due to outliers.

For a continuous measure which follows a normal distribution, the spread is best measured using the standard deviation (SD). This measures the average distance that observations are located from the mean and is derived from the variance (8). The variance is defined as the sum of the squared differences between the individual measurements and the true mean, divided by the total population size. By subtracting each data point from the mean and then squaring it, the distance from the mean is amplified, while the direction - whether it is smaller or larger than the mean - is removed (squared real numbers are always non-negative). The sample variance is represented by $s^{2}$ and is calculated from the data using a similar method as that described above. Instead of dividing by the total population size, $s^{2}$ is dividing by the degrees of freedom (df). Because a sample mean has to be calculated in order to estimate a variance or standard deviation, the degrees of freedom are obtained by subtracting 1 from the sample size. The square root of sample variance gives the standard deviation. The data points are expressed as $x$ and $\Sigma$ is 'sum of'. For further information on notation see the Appendix (page 120).

$$
\begin{gathered}
s^{2}=\frac{\text { Sum of all squared differences from the mean }}{\text { Sample size }-1} \\
s^{2}=\frac{\Sigma(x-\bar{x})^{2}}{n-1}
\end{gathered}
$$

The standard deviation is measured in the same units as the mean (e.g. mmHg for blood pressure). When data follow a normal distribution, we would expect that approximately 68 per cent of values will lie within one standard deviation of the mean. Similarly, we would also expect 95 per cent of values will lie within two standard deviations of the mean. To be more exact, 95 per cent of data from a normal distribution lie within plus or minus 1.96 standard deviations of the mean, rather than 2. This is often called the 'normal range'. Standard error (SE) does not measure the degree of spread of a distribution but rather the range of error associated with the estimate that is used to describe the data (e.g. standard error of the sample mean) (8). See page 87 for an explanation of standard error. It is included here as it is often mixed up with standard deviation, especially when given in relation to the mean when reporting data in a study.

[^22]9. For each data group, assign the distribution that best describes the population/ measurements

## Options

A. Normal
B. Binomial
C. Poisson

1. The weights of all the women in a village
2. The probability of a child in a class having asthma
3. The probability of a child in the school dying from asthma next Tuesday
4. The number of babies born on a particular day who are girls
5. The number of babies conceived this week by all men who had a vasectomy last year
6. Haemoglobin levels in full blood counts of blood donors
7. Using the binomial distribution answer the following questions

In a theoretical situation there are $n$ patients where $n$ is the number of people presenting to the emergency department with right iliac fossa pain (who haven't had a previous appendicectomy) in one week. Each patient has two possible outcomes, appendicectomy or no appendicectomy in the 24 hours after admission. Patients are independent, so the outcome for one patient has no effect on the outcome for another. The probability of having an appendicectomy when presenting with right iliac fossa pain is constant from one patient to another and is 20 per cent or 0.2
a. In the first week $n=10$, what is the most likely number that have an appendicectomy?
b. What is the probability that five people had an appendicectomy in that week?
c. What is the probability that eight people had an appendicectomy that week?
d. The following week $n=20$, what is the probability that 10 had an appendicectomy?
e. What is the probability that eight of those had an appendicectomy?

The hospital changed its policy and the decision was to give antibiotics and delay appendicectomy in non-urgent cases. The probability of having an appendicectomy was then 10 per cent or 0.1 . What is the probability that half of the patients had an appendicectomy
f. If $n=10$ ?
g. If $n=20$ ?

## 11. Regarding the Poisson distribution

a. It tends towards the binomial if the probability of an outcome is very large
b. It is most useful for diseases or outcomes that occur commonly
c. It is often used in the study of rare events
d. It is used to describe discrete quantitative data.

## EXPLANATION: DISTRIBUTION

Most data on human health and disease are distributed in what is known as the normal (or Gaussian) distribution. In a large population the data will appear as the classic 'bell-shaped' curve with a symmetrical distribution, with the majority of values clustered around the mean. Most statistical descriptions of data rely on the assumption that the measurements in the sample conform to the normal distribution. When the numbers are large binomial and Poisson distributions can tend towards the normal distribution.

The binomial distribution is a discrete probability distribution and is useful in describing distribution of data where there are two outcomes, such as disease or no disease, death or survival.

The binomial probability for obtaining $r$ successes in $N$ trials is shown below, where $P(r)$ is the probability of exactly $r$ successes, $N$ is the number of events, $\pi$ is the probability of success on any one trial and ! is factorial ${ }^{1}$.

$$
P(r)=\frac{N!}{r!(N-r)!} \pi^{r}(1-\pi)^{n-r}
$$

This formula assumes that the events are dichotomous (fall into only two categories), randomly selected, independent of one another and mutually exclusive (you cannot die and survive!). These outcomes are coded as 1 and 0 , respectively. In the following graph the example from question 10 has been used where the probability of appendicectomy was 0.2 and the number of patients 10 . For example, the probability of only one of them having an appendectomy equals

$$
\frac{10!}{1!(10-1)!} 0.2^{1}(1-0.2)^{10-1}=10 \times 0.2 \times(0.8)^{9}=2 \times(0.134)=0.268
$$



The more patients that present to hospital, the more this distribution tends towards the normal and the less likely it becomes that half of the patients need an appendicectomy.
${ }^{1}$ Factorial is the product of all the positive integers less than or equal to $n$. For example if $n=4,4!=4 \times 3 \times 2 \times 1$, or 24 . When $N=0,0!=1$.

## Answers

9. $1-\mathrm{A}, 2-\mathrm{B}, 3-\mathrm{C}, 4-\mathrm{B}, 5-\mathrm{C}, 6-\mathrm{A}$
10. $a-2, b-0.026, c-0.0001, d-0.002, e-0.022, f-0.0015, g-<0.0001$
11. F F T T

## EXPLANATION: DISTRIBUTION Cont'd from page 81

The Poisson distribution is useful when the probability of the outcome is very low and/or the population is relatively small. Take the example of men who have had a vasectomy. If 20000 men have had a vasectomy in the last year, which has a 1 in 2000 failure rate per year, then it is expected that from this population roughly ten babies will be concieved this year. The number of conceptions in any one week is an integer and is therefore usually 0 , but in practice may be as many as 3 . The mean number of conceptions in a 52 -week year is calculated as the sum of the number of weeks multiplied by the number of babies conceived that week, divided by the total number of weeks. For example:

$$
=\frac{(44 \times 0)+(5 \times 1)+(2 \times 2)+(1 \times 3)}{52}=\frac{5+4+3}{52}=0.23 \text { conceptions per week }
$$

Suppose that before the numbers were counted for the year it was estimated that there was a conception by a man who had had a vasectomy in the previous year roughly once every five weeks. This would estimate the probability of one in a week as 0.2 .

Using the formula for the Poisson distribution where $\mu$ is the population rate, $x$ is the number of conceptions per week and ! is factorial, a comparison can be made between what was expected and what was observed:

$$
\begin{aligned}
& P(x)=\frac{\mu^{x} \times e^{-\mu}}{x!} \\
& P(0)=0.2^{0} \times e^{-0.2} \div 0!=0.8187 \\
& P(1)=0.2^{1} \times e^{-0.2} \div 1!=0.1638 \\
& P(2)=0.2^{2} \times e^{-0.2} \div 2!=0.0164 \\
& P(3)=0.2^{3} \times e^{-0.2} \div 3!=0.0010 \\
& P(4)=0.2^{4} \times e^{-0.2} \div 4!=0.0001
\end{aligned}
$$



The figure top right shows the Poisson distribution of numbers of conceptions per week, by all men who have had a vasectomy in the last year.

Suppose we increase the population to include all women and all men who have been sterilized, by the method of tubal occlusion and vasectomy respectively, in the last year. Tubal occlusion has a failure rate of 1 in 200.


Suppose the estimated risk of conception in any week is now 5 per week. As can be seen by the graph immediately above of the Poisson distribution for $\mu=5$, the greater the value of $\mu$, the more the distribution tends to the normal or Gaussian distribution.
$e$ is a mathematical constant and is the base of natural logarithms. The value of $e$ is approximately 2.718 and most calculators have a function for calculating powers of $e$. For values of $x$ that are much smaller than 1.0 but negative, $e^{-x}$ is approximately equal to $1-x$. If, as in the example above, $x=-0.2$ this approximation is $e^{-0.2}=1-0.2=0.8$. Note that this is close to the actual value of 0.8187 .

## SECTION 10 ESTIMATION

- ACCURACY AND PRECISION AND STATISTICAL SAMPLING 84
- STANDARD ERROR 86
- CONFIDENCE INTERVALS88

2. Define precision
3. A blood sample was taken from a patient and four different assays were used to measure blood glucose. The true value of the blood glucose was known to be $4.5 \mathrm{mmol} / \mathrm{L}$

State whether the accuracy and the precision of each assay is high or low
a. Assay A: 4.6, 4.6, 4.8, 4.5, 4.5, 4.4
b. Assay B: 4.3, 3.5, 5.3, 4.6, 5.5, 3.7
c. Assay C: 3.5, 3.6, 3.3, 3.5, 3.4, 3.5
d. Assay D: 8.5, 6.4, 5.3, 7.6, 4.8, 9.3
4. For the following situations, would it be better to have high accuracy or high precision?
a. A scientific study investigating the effect of electrolyte concentrations on nerve cell function
b. A blood pH -measuring assay
c. An epidemiological study into the relationship between passive smoking and lung cancer
5. Of the following examples, choose the one that gives the most precise information and the one that gives the most accurate information
a. The weight of newborn babies measured on a recently calibrated weighing machine and recorded to the nearest 0.1 kg
b. The weight of adults in outpatients weighed using a machine calibrated several years ago but recorded to the nearest gram
c. Temperatures recorded by a thermometer that always measures $0.8^{\circ} \mathrm{C}$ above the patient's actual temperature

## EXPLANATION: ACCURACY AND PRECISION AND STATISTICAL SAMPLING

Data can be described in terms of:

- Accuracy: how close the measured estimates are to the exact or true value (1). For example, when measuring the prevalence of cancer, different studies may produce different values for prevalence. The most accurate studies are those in which the measured value is as close as possible to the true prevalence of cancer
- Precision: the reproducibility of a set of measurements (2). A precise sample estimate will have a very small random error of estimation. For example, a machine that measures blood glucose levels in blood samples is precise if it will consistently give the same measurement of blood glucose from the same individual's blood sample. However, this machine may not be accurate if it is calibrated such that it always underestimates the value of the blood glucose. Thus a precise but inaccurate method of gathering data will produce a systematic bias in the results.

A dartboard analogy of accuracy and precision is illustrated below.


Scientific experiments should be highly accurate, in order to obtain true values. Clinical studies often need to be more precise than accurate, as trends and associations are of interest rather than absolute values. Assays and analysers should be highly precise but can be calibrated in order to give accurate readings.

In the case of the blood glucose assay given in question 3 (page 84), the answers to the question are shown in the table opposite (3).

|  | Accuracy | Precision |
| :--- | :--- | :--- |
| Assay A | High | High |
| Assay B | High | Low |
| Assay C | Low | High |
| Assay D | Low | Low |

Statistical data usually represent a sample of a population of interest as it is often impractical or impossible to study the entire population. Through only studying a part of the population, we introduce a sampling error. In order to draw valid conclusions it is important to quantify this error. The precision of the sample estimate can be represented by the standard error or confidence intervals.

[^23]
## 6. What is the standard error of the following measures?

a. Mean cholesterol levels in 40-50 year olds in a GP practice in northern England where:

- Mean = $5.1 \mathrm{mmol} / \mathrm{L}$
- Sample size $=1600$
- Standard deviation $=2.3 \mathrm{mmol} / \mathrm{L}$
b. The proportion of women with skin cancer in women in a small town in Australia where:
- Number of cases of women with skin cancer $=30$
- Number of women in study sample $=5700$
c. Mean blood glucose levels on a ward where:
- Mean = $5.1 \mathrm{mmol} / \mathrm{L} \quad$ - Sample size $=16$
- Standard deviation $=2.3 \mathrm{mmol} / \mathrm{L}$
d. Mean peak expiratory flow rates on a respiratory ward where:
- Mean $=400 \mathrm{~mL} / \mathrm{min}$ - Sample size $=16$
- Standard deviation $=120 \mathrm{~mL} / \mathrm{min}$

7. What does a small standard error tell us about the sample estimate of the mean?
a. That it is highly variable
b. That it is accurate
c. That the population standard deviation may be small
d. That the sample size is probably small
e. That it is imprecise
8. What will tend to make the standard error larger?
a. A small variance
b. A large standard deviation
c. Imprecise data
d. Inaccurate data

## 9. As the size of a random sample increases

a. The standard deviation decreases
b. The standard error of the mean decreases
c. The mean increases
d. The range may increase
e. The precision of the parameter estimate increases

## EXPLANATION: STANDARD ERROR

Example: a study was conducted to investigate the mean systolic blood pressure of a group of hypertensive patients who were treated with a new antihypertensive drug. The study was repeated several times, taking successive samples from the same population treated with the same new drug.

Each sample returned a slightly different value for the mean systolic blood pressure: these values are estimates of the true mean of the blood pressure in the whole population and can be displayed as a histogram, as shown opposite.

If the distribution of blood pres-
 sure in the population follows a normal distribution, the estimates of the means from each sample should also follow a normal distributions known as the sampling distribution of the mean. If the overall sample size is very large, it will follow a normal distribution, regardless of the distribution of the original data in the population.

The variability of this distribution is known as the standard error of the mean (SEM). As it is not usually practical to repeat studies many times to obtain estimates of the mean from a number of different samples, the SEM can be estimated using the standard deviation (SD).

$$
\mathrm{SEM}=\mathrm{SD} / \mathcal{V}_{n}
$$

where $n$ is the sample size.
From this calculation it is evident that the standard error becomes smaller when the variability of the data (SD) remains constant and the sample size increases.

The smaller the standard error, the more precise is the sample estimate. This is in contrast to the SD which represents the variability of the primary observations. The standard error does not reflect the accuracy of the data.

If a study investigating the proportion of individuals with a given characteristic is repeated several times using different samples from the same population, the estimates of the true proportion will approximate a normal distribution, known as the sampling distribution of a proportion. The estimates have a variability that is represented by the standard error of the proportion $(\operatorname{SE}(p))$. This value can be approximated from a single sample using the following equation:

$$
\operatorname{SE}(p)=\sqrt{ }(p(1-p) / n)
$$

[^24]10. Using the normal distribution provided in the Appendix (page 124), calculate the confidence intervals for the following datasets
a. The 95 per cent confidence interval (CI) for the forced expiratory volume $\left(\mathrm{FEV}_{1}\right)$ of children with asthma, where the mean is 1.47 L , the standard deviation is 0.34 L and the sample size is 1089
b. The 99 per cent Cl for the proportion of babies born with birth defects to mothers who had an infection during pregnancy, where the number of babies with birth defects is 7 and the total number of babies in the study is 43 . (NB: In the normal distribution, 99 per cent of the data lie within 2.58 standard deviations of the mean.)
11. Using the $t$-distribution provided in the Appendix (page 125), calculate the CI for the length of time to recovery to full fitness: a study of 10 athletes with sports injuries, mean length of time of recovery to full fitness of 36 days, standard deviation of 21.4 days

## 12. Interpret the following results

A study was conducted to investigate whether there was a difference in mean systolic blood pressure between a group of army recruits who smoke and a group who do not smoke. The means in two groups were adjusted for

| Smoking <br> status | Mean systolic blood <br> pressure $(\mathbf{m m H g})$ | Standard <br> deviation | Sample <br> size |
| :--- | :--- | :--- | :--- |
| Yes | 149 | 17 | 70 |
| No | 131 | 10 | 70 | age and other risk factors for hypertension.

a. What is the standard error of each of the two means?
b. What are the 95 per cent Cls for the two means?
c. What can you conclude from these Cls ?
13. A 95 per cent CI for a mean
A. Is narrower than a 99 per cent Cl
B. Is a useful way of describing the precision of a study
C. Includes 95 per cent of the observations in a study
D. Will include the population mean 95 per cent of the time in repeated samples
14. Several different studies investigated the effect of a new statin drug on cholesterol levels in samples of patients with similar characteristics. Choose the set of results that shows the most convincing reduction in cholesterol level
A. $2 \mathrm{mmol} / \mathrm{L} 95$ per cent
Cl 1.1 to 2.2
D. $5 \mathrm{mmol} / \mathrm{L} 99$ per cent Cl of 0.5 to 3.1
B. $2 \mathrm{mmol} / \mathrm{L} 99$ per cent Cl 1.1 to 2.2
E. $5 \mathrm{mmol} / \mathrm{L} 95$ per cent Cl of 4.6 to 5.3
C. $5 \mathrm{mmol} / \mathrm{L} 95$ per cent $\mathrm{Cl}-0.1$ to 9.5

## EXPLANATION: CONFIDENCE INTERVALS

A confidence interval (CI) represents both the precision and the accuracy of an estimate. It is derived from the standard error but is expressed in such a way to help us to picture the sampling variation in the estimate more easily. The interval represents the range of values within which we can be confident that the true value of our population parameter lies. The confidence with which, to a certain degree, we can believe in this interval is specified by a percentage representing the probability that the CI contains the true population parameter.

The degree of confidence often used is 95 per cent. Here there is a 95 per cent chance that the theoretical interval contains the true population parameter. It is often claimed that one can be 95 per cent certain that the true population parameter lies within the 95 per cent CI obtained from a single sample. This is not, however, strictly correct. As the population parameter is a fixed unknown quantity, it either lies inside or outside the 95 per cent CI. If 100 independent random samples were drawn from the same population, we would expect 95 of these to include the true population parameter. It is also possible to obtain intervals using other degrees of confidence. For example, 99 per cent CIs would make a more certain statement about the likely magnitude of the true population parameter.

$$
95 \text { per cent CI for the mean }=\bar{x}-(1.96 \times \text { SEM }), \bar{x}+(1.96 \times \text { SEM })
$$

The CI for the example given in question 10a (page 88) is calculated as follows:

$$
1.47 \pm(1.96 \times 0.34 / \sqrt{ } 1089)=1.47 \pm 0.020
$$

where mean $=1.47 \mathrm{~L}, \mathrm{SD}=0.34 \mathrm{~L}$, sample size is 1089 , and so $\mathrm{SEM}=0.34 / \sqrt{ } 1089$.
The 95 per cent CI is 1.45 to 1.49 L .
The CI for the example given in question 10b (page 88) is calculated as follows:
Proportion $p=7 / 43=0.163$, SE of $p=\sqrt{ }[0.163 \times(1-0.163) / 43]=0.056$
Lower limit of $99 \% \mathrm{CI}$ is $0.163-(2.58 \times 0.056)=0.019$, upper limit of $99 \% \mathrm{CI}$ is $0.163+(2.58 \times 0.056)$ $=0.307$

The 99 per cent CI for the proprotion of babies born with birth defects is 0.019 to 0.307 or $1.9 \%$ to $30.7 \%$
When the sample size is large, the sampling distribution of the mean approximates the normal distribution. In the normal distribution, 95 per cent of values lie within plus or minus 1.96 standard deviations of the mean. The CI is calculated from the equation below, where SEM is the standard error of the mean and the multiplier, 1.96, is the percentile of the normal distribution that gives a 95 per cent CI.

Continued on page 90

[^25]
## EXPLANATION: CONFIDENCE INTERVALS Cont'd from page 90

When the sample size is small (e.g. less than 20), the data may not be so normally distributed, and the variance of the population may not be known. Therefore a different probability curve, the Student's $t$-distribution is used. The $t$-distribution is similar to the normal distribution, but is more spread out with longer tails. The multiplier, $t_{0.05}$, is the percentile of the $t$-distribution that gives a two-tailed probability of 0.05 .

$$
95 \text { per cent CI for the mean }=\bar{x}-\left(t_{0.05} \times \text { SEM }\right), \bar{x}+\left(t_{0.05} \times \mathrm{SEM}\right)
$$

To calculate 99 per cent CIs, one would multiply the SEM by a $t$-value corresponding to a two-tailed probability of 0.01 .

The $t$-distribution can be used to calculate CIs (11) for mean length of time to recovery to full fitness. For a $t$-distribution with $n-1$ degrees of freedom $(10-1=9), t_{0.05}=2.26$ :

$$
95 \text { per cent } \mathrm{CI} \text { for the mean }=36 \pm 2.26 \times(21.4 / \sqrt{ } 10)=36 \pm 15.3 \text { days }
$$

95 per cent CI is 20.7 to 51.3 days.
As for the mean, the sampling distribution of the proportion approximates the normal distribution when the sample size is large. Therefore the $\mathbf{9 5}$ per cent CI for the proportion is given by:

$$
p-[1.96 \times \sqrt{ }(p(1-p) / n)], p+[1.96 \times \sqrt{ }(p(1-p) / n)])
$$

where $n$ is the sample size.
If the sample size is small, the binomial distribution is used to calculate exact CIs for proportions.
The width of the CI represents the precision of the sample estimate:

- Wide CI: imprecise estimate
- Narrow CI: precise estimate.

As the width of the CI is derived from the standard error, increasing the sample size or decreasing the variability of the original data will decrease the width of the CI, reflecting the increased precision of the estimate. The higher the degree of confidence, such as a 99 per cent CI, the stronger our belief that the true parameter lies between the interval values given.

## section 11 HYPOTHESIS TESTING

- PRINCIPLES OF HYPOTHESIS TESTING 92
- CALCULATING THE TEST STATISTIC: DIFFERENT TYPES OF TEST 94
- $P$ VALUES 96

1. Define null hypothesis and alternative hypothesis
2. Put the stages of doing a hypothesis test into the correct order
A. Interpret the $P$-value
B. Calculate the test statistic
C. Define the null and alternative hypotheses
D. Collect the relevant data
E. Obtain the $P$-value from the appropriate probability distribution curve

## 3. Comparing hypothesis tests and confidence intervals

A. Hypothesis tests are easier to calculate
B. A hypothesis test can give an exact value for the $P$-value
C. A confidence interval does not tell you about the significance of the results
D. A hypothesis test gives an idea of the precision of the results whereas the confidence interval cannot
E. A hypothesis test gives you an idea of the likely magnitude of a population parameter whereas a confidence interval cannot
4. What is the difference between a one-tailed test and a two-tailed test?
5. What are the advantages of a two-tailed test?
6. Why would one use a one-tailed test?

## EXPLANATION: PRINCIPLES OF HYPOTHESIS TESTING

The purpose of an epidemiological study is often to investigate the relationship between a variable of interest and an outcome of interest. The investigator usually begins the study with an idea or a theory as to the possible nature of the relationship. This theory can be approached by hypothesis testing.

A statistical method, known as a hypothesis test, is used to quantify the strength of evidence against the null hypothesis, based on the data that have been collected in the study. The hypothesis test involves comparing the mean sample estimate, $\bar{x}$, with a hypothetical value of a true population parameter.

Stages of hypothesis testing are as follows (2):

1. Define the null and alternative hypotheses ( $H_{0}$ and $H_{1}$ respectively)
2. Collect relevant data
3. Calculate the test statistic specific for the null hypothesis
4. Compare the value of the test statistic to values from a known probability distribution to obtain a $P$-value
5. Interpret the $P$-value and the results.

Definitions are as follows:

- Null hypothesis, $\boldsymbol{H}_{0}$ : a statement about the true population parameter of interest, for example the true mean, $\mu(1)$. The null hypothesis might state that there is no difference in the true means of two populations: $\mu_{1}=\mu_{0}$
- Alternative hypothesis, $\boldsymbol{H}_{1}$ : any hypothesis that differs from $H_{0}$, for example the true population means are different: $\mu_{1} \neq \mu_{0}$ (1)
- Two-tailed test: the direction of the relationship stated by the alternative hypothesis is not specified (4). This is normally preferred as it is often difficult to predict the direction of a true association in advance of the study, as by definition statistical studies involve a degree of uncertainty. A two-tailed test therefore has stronger statistical credentials (5)
- One-tailed test: the direction of the relationship is specified by the altenative hypothesis (4). A one-tailed test is used occasionally when the results can only occur in one direction. An example might be when testing a new drug treatment for a condition in which the only outcome without treatment is death, i.e. the drug can have either no effect or can increase survival (6).

A hypothesis test provides an exact value of the $P$-value and therefore the exact degree of certainty we can hold in our results, i.e. the accuracy of the results. The $\mathbf{9 5}$ per cent CI specifies the range over which we can be 95 per cent certain that the true population parameter lies relative to our estimate of the population parameter. This is equivalent to the certainty provided by a $P$-value $<0.05$, but also indicates the degree of precision of the estimate. CIs tell us about the likely magnitude of the estimated parameter, whereas hypothesis tests do not.

[^26]
## 7. Match the study questions (1-5) to the most appropriate statistical test

## Options

A. ANOVA
B. Chi-squared test for $3 \times 2$ table
C. 2 -sample unpaired $z$-test
D. 2-sample paired $z$-test
E. Chi-squared test for $2 \times 2$ table

1. A new drug for lowering mean blood pressure is being compared with a placebo in a crossover trial
2. Is there a higher risk of breast cancer in women who take the pill versus women who do not?
3. Is the proportion of smokers different in France, Britain and Germany?
4. Is the mean BMI of 15 -year-olds in Scotland greater than the mean BMI of 15 -year-olds in England?
5. Three different drugs for the treatment of hypercholesteraemia are being compared using mean cholesterol levels across the groups

## 8. Calculate the test statistic for the following data

An unmatched study into a new lipid-lowering drug (use the data in the table below)

- $H_{0}$ : triglyceride levels are the same in the population of individuals receiving lipidlowering treatment as in the population who are not receiving treatment
- $H_{1}$ : triglyceride levels are different in the treated population versus the untreated population

|  | Mean serum triglyceride $(\mathrm{mmol} / \mathrm{L})$ | Standard deviation $(\mathrm{mmol} / \mathrm{L})$ | Sample size |
| :--- | :--- | :--- | :--- |
| Treated | 4.6 | 1.5 | 54 |
| Untreated | 5.4 | 1.5 | 67 |

## EXPLANATION: CALCULATING THE TEST STATISTIC AND DIFFERENT TYPES OF TEST

The test statistic is calculated by entering values obtained from the study data into a formula specific to the test that has been chosen. The choice of test depends upon study design, type of variable and distribution of the data.

Parametric data are data in which the population distribution is known or the sample size is large enough to assume that the data approximate a normal distribution. Hypothesis testing proceeds by calculating a $z$-statistic which, if the null hypothesis is true, follows a normal distribution with zero mean and $\mathrm{SD}=1$. If the sample size is small, Student's $t$-distribution can be used. The tests summarized in the diagrams below are all for parametric data. See summary tables in the Appendix (page 121) for more details.


For non-parametric data the population distribution is not known, and the sample size is too small to approximate the normal distribution. In such cases, other tests can be used such as the:

- Wilcoxon signed rank test (paired numerical data)
- Wilcoxon rank sum test or the Mann Whitney $U$-test (unpaired numerical data)
- Kruskal-Wallis test (non-parametric equivalent of analysis of variance (ANOVA)).

[^27]9. What does a $P$-value of $>0.05$ mean?
10. The $P$-value
a. Is the probability that the null hypothesis is false
b. Is small for small studies
c. Is the probability of the observed result or a more extreme result if the null hypothesis were true
d. Can only be one of two values: 0.05 or 0.01
e. Is equal to ( 1 - type 2 error)
11. A new treatment regime for breast cancer was introduced, producing improved cure rates. The proportion of survivors in the 2 years following introduction of the new treatment was significantly higher than in the preceding 2 years: $\chi^{2}=4.2$, degrees of freedom (df) $=1, P<0.05$
a. The probability of getting this difference or one more extreme, if there had been no real change in cure rates, is less than 1 in 20
b. The improvement in treatment outcome is statistically significant
c. The change in outcome may be due to a confounding variable
12. Using the statistical distribution tables in the Appendix, what is the $P$-value for the following data?
a. $z=2.4$, two-tailed test
b. $\chi^{2}=10.5$, two-tailed test, $\mathrm{df}=1$
13. What is the $P$-value of the test statistic calculated in question 8 (page 94) and what conclusions can we draw?

## EXPLANATION: P-VALUES

Test statistics follow known theoretical probability distributions. This means that any given value of a test statistic can be allocated a probability, i.e. the likelihood that that value would have occurred by chance alone.

A theoretical probability distribution curve for a standard normal distribution is shown below.


The area under the curve represents the probability of obtaining the test statistic calculated from the data if the null hypothesis were true. In a two-tailed test, the total probability is the sum of the probabilities under both tails of the curve.

It is convention to quote a significance level $\alpha$ of 5 per cent. This means that a $P$-value of less than 0.05 is considered sufficiently rare to disprove the null hypothesis, i.e. the same value for the test statistic would be obtained by chance alone in fewer than 5 times out of 100 .

If the $P$-value is greater than 0.05 , we can say that we do not reject the null hypothesis as there is a statistical likelihood that the same test statistic would be obtained by chance alone in more than 5 out of 100 occasions. Thus the results are not significant at the 5 per cent level (9).

The $P$-value can be obtained from the test statistic distribution table (see Appendix, page 124). Most computer packages will provide the $P$-value automatically from the test statistic.

When reporting the results of a hypothesis test it is better to quote the exact $P$-values if they are known. This is important for distinguishing highly significant results, such as $P$-values of $<0.01$, from less significant results, such as $P$-values of between 0.04 and 0.05 .

## Answers

9. See explanation
10. F F T F F
11. T T T
12. $\mathrm{a}-P=0.016, \mathrm{~b}-0.001<P<0.01$
13. Two-tailed $P$-value $=0.9203$ : no significant difference in triglyceride level between the two groups

This page intentionally left blank

## SECTION 12 INTERPRETATION OF DATA

- ASSOCIATION OR CAUSATION
- ODDS RATIO 102,110
- RISK 104
- CORRELATION AND LINEAR REGRESSION 106
- STATISTICAL INFERENCE IN CORRELATION AND REGRESSION 108

1. The following nine terms ( $\mathrm{A}-\mathrm{I}$ ) refer to criteria to consider when deciding whether findings from observational studies (i.e. those which do not involve randomization) suggest a causal relationship. Match each term to one of the nine definitions (1-9) listed below

## Options

A. Temporality
B. Strength of association
C. Dose-response
D. Plausibility
E. Consistency
F. Specificity
G. Coherence
H. Experiment
I. Analogy

1. The relationship can be modified by experimental interference
2. Agrees with prior strong evidence of a similar association
3. The risk factor must precede disease
4. A single exposure or risk factor has a specific effect
5. Consistent with current understanding of biological or pathological processes
6. Increased exposure to risk factor increases the risk of disease
7. The results are replicable or in accordance with those of other, similar studies
8. Large relative risk or correlation coefficient, with small standard error
9. Agrees with existing knowledge and widely accepted theories
10. Of the following, choose which values give a measure of the strength of an association
a. Odds ratio
b. Relative risk
c. Standard deviation
d. Confidence interval
e. Correlation coefficient

## EXPLANATION: ASSOCIATION OR CAUSATION

When considering a clearly statistically significant association found between a disease and an exposure, interest will often focus on whether the relationship is causal. While this may be a relatively straightforward issue for a well conducted randomized controlled trial, it is much more challenging when considering results from an observational study. Even after adjusting for all potential confounding factors, several important issues need to be considered. In a now classic address to the Royal Society of Medicine in 1965, Sir Austin Bradford Hill proposed that several aspects of an association should be given special consideration. These are summarized below: ${ }^{1}$

1. Strength of association: large relative risk or correlation coefficient, with small standard error
2. Consistency: the results are replicable or in accordance with those of other, similar studies
3. Specificity: a single exposure or risk factor has a specific effect
4. Temporality: the risk factor must precede disease
5. Dose-response: increased exposure to the risk factor increases the risk of disease
6. Plausibility: consistent with current understanding of biological or pathological processes
7. Coherence: agrees with existing knowledge and widely accepted theories
8. Experiment: the relationship can be modified by experimental interference
9. Analogy: confounding factors have been considered and excluded.

The following are measures used by scientists and epidemiologists to investigate the relationship between two or more sets of data:

- Odds ratio (see page 103): this is the ratio between the odds of being exposed if you have disease versus the odds of not being exposed if you do not have the disease. The odds ratio can give an estimate of relative risk
- Relative risk (see page 105): the increased risk of disease associated with exposure to factor of interest.
Relative risk = risk of disease in exposed group/risk of disease in unexposed group
- Correlation (see page 107): correlation measures the strength of association between two quantitative variables.
${ }^{1}$ Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine 1965;58:295-300.

[^28]3. What are odds? What type of epidemiological study often uses an odds ratio as an estimate of relative risk?
4. Define the term 'odds ratio'
5. What assumptions are required for an odds ratio to be a reliable estimate of relative risk?
6. How would you calculate an odds ratio and what do the following odds ratios imply: 4.6, 0.3 and 1.0?
7. What is the difference between the odds ratio and relative risk?
8. A retrospective case-control study investigated the association between bladder cancer and smoking:

- 145 people with bladder cancer: 87 were smokers, 58 were non-smokers
- 178 controls without bladder cancer: 62 were smokers, 116 were non-smokers
a. Calculate the odds ratio for the association between smoking and bladder cancer
b. Calculate the test statistic for the odds ratio
c. Using the statistical tables in the Appendix (page 126), interpret the $P$-value for the test statistic


## EXPLANATION: ODDS RATIO

Odds are the ratio of the probability of occurrence of an event to the probability of non-occurrence (3):

$$
\text { Odds }=P /(1-P)
$$

Odds ratio (OR) is a measure of the effect of a risk factor. For rare events its value approximates that of the relative risk (4). For an odds ratio to be a reliable estimate of relative risk the following assumptions are required (5):

- The proportion of individuals with the disease is small, i.e. the disease is infrequent in both exposed and unexposed individuals
- The cases and controls are not selected in favour of either exposed or non-exposed individuals, i.e. they are representative samples from each of the relevant population groups (those with the disease (cases) and those without (controls)).

To calculate the odds ratio (6): construct a $2 \times 2$ table as shown below.

|  | Cases | Controls | Total |
| :--- | :--- | :--- | :--- |
| Exposed | $a$ | $b$ | $a+b$ |
| Unexposed | $c$ | $d$ | $c+d$ |
| Total | $a+c$ | $b+d$ | $a+b+c+d$ |

OR = odds of being exposed if you are a case/odds of being exposed if you are a control

$$
\begin{gathered}
=[a /(a+c)] /[1-(a /(a+c))] /[b /(b+d)] /[1-(b /(b+d))] \\
=[a /(a+c)] /[c /(a+c)] /[b /(b+d)] /[d /(b+d)] \\
=(a / c) /(b / d) \\
=a d l b c
\end{gathered}
$$

Odds ratios are interpreted as follows (6):

- Odds ratio $=1$ : no association between exposure and disease
- Odds ratio < 1: negative association between exposure and disease; exposure reduces the odds of developing the disease - exposure
- Odds ratio > 1: positive association between exposure and disease; exposure increases the odds of developing the disease.

The odds ratio has often been shown to approximate well the relative risk calculated from a cohort study if the above assumptions are maintained (3). It is therefore considered a valuable measure of association. Unlike relative risk, which takes into account the number of individuals in the study population, the odds ratio remains constant if the number of controls were to change (7).

Continued on page 110

## Answers

3. See explanation
4. See explanation
5. See explanation
6. See explanation
7. See explanation
8. See $2 \times 2$ table page 110 . a $-\mathrm{OR}=(87 \times 116) /(62 \times 58)=2.8, \mathrm{~b}-\chi^{2}=\Sigma(5.88+4.60+5.01+3.92)=$
19.41, c - For $\mathrm{df}=1$, the $P$-value for $\chi^{2}=19.41$ is $<0.001$, therefore is highly significant
9. In a town with a population of 100000 the risk of dying from disease $X$ is two per 1000, or 0.2 per cent. From the following list of outcomes choose the best option

## Options

A. A treatment that will result in a relative reduction in mortality of 30 per cent
B. A treatment that will mean that 60 lives are saved
C. A treatment that offers an absolute risk reduction of 0.06 per cent
D. A treatment that requires 17 people to be treated to save one life
E. A treatment that reduces the risk of dying from the disease to 0.14 per cent
10. In another town there are 20000 people. Half of them drink the well water and half of them drink reservoir water with fluoride in it. Government officials pointed out that the rest of the town should receive the reservoir water that is fluoridated as it leads to less tooth decay. Suppose 60 people were found to have tooth decay who drank the well water and 54 people had tooth decay who drank the fluoridated water.
a. What is the absolute risk (well water)?
b. What is the relative risk reduction?
c. What is the absolute risk reduction?
d. How many people would have to switch to fluoridated reservoir water to prevent one person from getting tooth decay?
e. Which figures do you think the government used to persuade the people to switch over to reservoir water?
11. One hundred and thirty male body builders in a local gym are offered a try out of a new back support designed to reduce injury. Fifty men (group A) use the new support for a year and the remaining 80 (group B) decide not to. If there are five injuries in group A and ten in group B over the year:
a. What is the absolute risk reduction?
b. What is the relative risk reduction?
c. How many men would have to wear the support for a year to prevent injury in one man?

## EXPLANATION: RISK

Risk is the probability of an outcome, usually a harmful one, over a fixed period of time. It is calculated from the results of a cohort study as the number of events during the defined study period divided by the number of individuals at risk of that event.

Absolute risk = number developing disease in study period/total number in cohort. For example, the absolute risk of a 50 -year-old man having a myocardial infarction (MI) in 1 year is the number of men aged 50 years having an MI in 1 year divided by the number of men in the whole population who are 50 years old.

Relative risk = absolute risk in group 1/absolute risk in group 2. The relative risk compares the risk in two groups. One is referred to as the 'exposed' group and the other, usually the initial or control group, is the 'unexposed' group. For example, if the absolute risk of an MI within 1 year for men aged 50 is 0.01 per cent and the absolute risk for men aged 60 is 0.02 per cent, the relative risk for men aged 60 compared to men aged 50 would be $0.02 / 0.01=2.00$.

Relative risk reduction = absolute risk reduction/absolute risk in group 1. When a 'new' treatment is introduced, a trial is often carried out to compare it to the 'gold standard', namely the 'old' treatment that is currently most widely accepted as being the best. The objective is to calculate the relative risk of disease in those taking the 'old' treatment compared to that in the 'new' treatment group. The relative risk reduction is the proportion of the risk with the 'old' treatment that has been eliminated by using the 'new' treatment. For example, if the risk of an MI in a population taking 'old' treatment for hypertension is 0.05 per cent and the 'new' treatment brings the risk down to 0.03 per cent, the relative risk reduction is $(0.05-0.03) / 0.05 \times 100$ or 40 per cent.

Absolute risk reduction $=$ absolute risk in group 1 - absolute risk in group 2 . The absolute risk reduction can sometimes give a more realistic indication of the effect of a new treatment than the relative risk reduction. Using the same example above, any person who is told they could reduce their risk by 40 per cent would probably take the 'new' treatment. However, if the reduction in risk is converted to absolute risk reduction this equals 0.02 per cent $(0.05-0.03)$. Such a small reduction in absolute risk may appear less favourable to an individual in the light of the cost or any additional side-effects of the 'new' drug.

Numbers needed to treat $(\mathrm{NNT})=\mathbf{1}$ /absolute risk reduction. When presenting a choice of treatment options to a patient there may be cases when neither the relative risk reduction nor the absolute risk reduction gives them a good idea as to what it means to them as an individual. Sometimes by putting it into the user friendly' context of numbers needed to treat, the difference between old and new treatments can be understood more readily. Using the example above, $\mathrm{NNT}=1 / 0.02=50$, therefore the doctor can tell the patient that if he or she gives 50 people the new medication for 1 year he or she will prevent 1 myocardial infarction (MI).

[^29]12. For what type of data is correlation a useful measure of association?
13. What would the correlation coefficient, $r$, be for the following theoretical linear relationships?
(a)

(b)

(c)

14. What do these values of $r$ imply about the relationships displayed in graphs (a), (b) and (c)?
15. What is the equation for a linear relationship between two variables, $x$ and $y$ ?
16. A study into the effects of a new drug on serum cholesterol was conducted on 20 subjects and the following results obtained:

Regression line: $y=0.7-0.6 x$; standard error of the slope $=0.28$
Answer the following questions:


Drug dose (mg)
a. Define the null and alternative hypotheses for this study
b. Does this drug have a significant effect on serum cholesterol level at the 5 per cent level? Would you accept or reject the null hypothesis?
c. What would be the expected change in serum cholesterol if the following doses of the new drug were given to the patient: $1 \mathrm{mg}, 2 \mathrm{mg}, 5 \mathrm{mg}, 10 \mathrm{mg}$ ?
d. How could you explain the increase in serum cholesterol at low drug doses?

## EXPLANATION: CORRELATION AND LINEAR REGRESSION

Correlation measures strength of linear association between two numerical variables. If two variables have a linear relationship the data points will form a straight line when plotted on Cartesian axes (12). Degree of correlation can be calculated using the Pearson correlation coefficient, $r$, which can have a value between +1 and -1 , where $+1=$ positive correlation (risk increases linearly with increasing exposure), or $-1=$ negative correlation (risk decreases linearly with increasing exposure); $r=0$ represents absence of any linear correlation. This is illustrated below.


Graphs reproduced with kind permission from Everitt BS and Palmer CR (eds), Encyclopaedic Companion to Medical Statistics, Great Britain: Hodder Education, 2005

The square of the correlation coefficient, $r^{2}$, represents how much of the variability in $y$ can be accounted for by variability in $x$, in other words, how well the data fit the line. For example, a study shows that there is a positive association between the number of cigarettes smoked and systolic blood pressure with correlation coefficient $r=+0.5$. Therefore $r^{2}=0.25$, suggesting that 25 per cent of the increase in blood pressure can be accounted for by the increased number of cigarettes smoked. It is very important to remember that two variables having a high correlation coefficient does not mean that one causes the other.

When one of the variables is considered to be dependent on the other, the relationship can be analysed using linear regression. This is the process of attempting to fit a straight line to the observed data points, such that the data can be represented by the equation

$$
\begin{equation*}
y=b x+a \tag{15}
\end{equation*}
$$

where $x$ is the independent variable whose value can be measured accurately, $y$ is the dependent variable, $b$ is the slope of the line and $a$ is the intercept on the y axis when
 $x=0$.

Linear regression analysis is often achieved by finding the best fit: the straight line that minimizes the sum of the squares of the vertical deviations of each data point from the line. These deviations are known as the residuals and they are squared so as to avoid any cancellations between data points that are above and below the line. The resulting regression equation is valid for values of $y$ and $x$ that are within the observed range of data. Extrapolating to predict values of $y$ from values of $x$ beyond the data range can lead to inappropriate and incorrect answers.

[^30]17. Assuming a correlation exists in the following examples, which examples are appropriate for testing correlation by linear regression (true or false)
a. The correlation between resting heart rate and square root of the number of miles cycled by different cyclists on a weekly basis, assuming the relationship is linear and there are no confounding variables
b. The relationship between blood pressure and blood cholesterol levels of the people who went to their GP last week; some people presented as many as three times
c. The relationship between height, and a rare disease, $y$, that does affect stature but is not found to be normally distributed in the population
d. The relationship between blood levels of 'liver markers' and liver damage; where liver markers rise exponentially with increasing liver damage
18. For the examples a-d in question 17 give the main reason, if there is one, using the list of assumptions below (1-4), why it is not appropriate to use linear regression analysis

1. The relationship must be linear
2. Observations must be independent, there cannot be different numbers of observations made on the same individuals
3. Other, confounding variables, must be excluded
4. The $y$ values are normally distributed in the population
5. Are the following true or false?
a. Predictions can be derived from information that is correlated linearly
b. If you know $x$ and the equation for the relationship you can calculate $y$
c. If you know $y$ and the equation you cannot calculate $x$
d. The results can be 'extrapolated' or extended to predict values for $x$ and $y$ outside the original range

## EXPLANATION: STATISTICAL INFERENCE IN CORRELATION AND REGRESSION

A number of assumptions are made when doing a regression analysis:

- The relationship between $x$ and $y$ is approximately linear
- The observations are independent of each other, i.e. there is only one value of $y$ for every value of $x$
- The residuals about the fitted line are normally distributed with a mean of zero
- The residuals are independent of each other, i.e. they have a constant variability for all the fitted values of $y$
- The $x$ variable can be measured without error.

There may be a number of circumstances in which the assumptions above are not met, for example: when the relationship is non-linear, when different numbers of observations are made on the same individuals (more than one value for $y$ for each value of $x$ ), or when there is more than one independent variable.

The linearity and independence assumptions are the most important. If $x$ and $y$ do not have a linear relationship, it is often possible to convert it into a linear relationship by transforming one of the variables, for example through squaring or performing a logarithmic function on all the values of one variable. If this is not possible, a non-linear regression analysis can be performed. If one is interested in the relationship between one dependent variable, $y$, and several different independent variables, $x_{1}, x_{2}, x_{3}$, a multiple linear regression analysis can be performed.

To test whether there is a statistically significant linear relationship between $x$ and $y$, a hypothesis test can be conducted to disprove the null hypothesis that the slope of the line $=0$.

$$
\text { Test statistic }=b(\text { slope }) / \text { SE of } b
$$

The $P$-value is derived from the z statistic or, for small samples, from the $t$-distribution with $n-2 \mathrm{df}$.
The 95 per cent confidence interval for the slope is given by:
$b \pm 1.96 \times \mathrm{SE}(b)$ (or by replacing 1.96 with the equivalent value from the $t$-distribution with $n-2 \mathrm{df}$ )
The regression line can be used to predict values of $y$ for values of $x$ that are within the range covered by the observed data points, by substituting values of $x$ into the equation for the line. Predicting values of $y$ outside the tested range of $x$ will produce increasingly unreliable values of $y$, and should not be done, as it will produce invalid results. An example is arm span correlated to height in adults; by measuring arm span this can give a relatively good prediction of height.

[^31]
## EXPLANATION: ODDS RATIO Cont'd from page 103

Question 8 (page 102) $2 \times 2$ table (the observed frequencies $(O)$ are printed in bold).

|  | Bladder cancer | No bladder cancer | Total |
| :--- | :--- | :--- | :--- |
| Smokers | 87 | 62 | 149 |
| Proportions | $87 / 149=0.58$ | $62 / 149=0.42$ | $149 / 323=0.46$ |
| Expected frequency | $0.46 \times 145=66.7$ | $0.46 \times 178=81.9$ |  |
| $(\|O-E\|-1 / 2)^{2} / E$ | $(\|87-66.7\|-1 / 2)^{2} / 66.7=5.88$ | $(\|62-81.9\|-1 / 2)^{2} / 81.9=4.60$ |  |
| Non-smokers | 58 | 116 | 174 |
| Proportions | $58 / 174=0.33$ | $116 / 174=0.67$ | $174 / 323=0.54$ |
| Expected frequency | $0.54 \times 145=78.3$ | $0.54 \times 178=96.1$ |  |
| $(\|O-E\|-1 / 2)^{2} / E$ | $(\|58-78.3\|-1 / 2)^{2} / 78.3=5.01$ | $(\|116-96.1\|-1 / 2)^{2} / 96.1=3.92$ | $\mathbf{3 2 3}$ |
| Total | 145 | 178 |  |

Modulus $(|\mathrm{O}-\mathrm{E}|)=$ the absolute magnitude of the difference betwen $O$ and $E$, regardless of its sign

## SECTION 13 SOURCES OF ERROR

- TYPE 1 AND TYPE 2 ERRORS
- POWER 114
- BIAS AND CONFOUNDING116

1. In a hypothetical study the null hypothesis is that there is no association between rate of lung cancer and shoe size. The calculations lead to the conclusion that there is a statistically significant association
a. If there truly is no significant association between rate of lung cancer and shoe size, to reject the null hypothesis would be a type 1 error
b. The probability of making a type 1 error is represented as $\alpha$
c. Type 1 errors are not related to the $P$-value
d. The more stringent the $P$-value the more likely it is that a type 1 error will be made
e. Type 1 errors are directly related to type 2 errors
2. In another study the null hypothesis is that there is no association between lung cancer and smoking. The calculations lead to the conclusion that there is no statistically significant association and the null hypothesis is not rejected
a. This is also an example of a type 1 error
b. The likelihood of this error happening would be reduced if there were a larger sample size
c. The probability of making this error depends upon the power of the test
d. A type 2 error is equivalent to the $P$-value
3. In the following examples identify the type of error

## Options

A. Type 1 error
B. Type 2 error
C. No error

1. There is no correlation between number of carrots consumed per week and visual acuity. A small study of local villagers shows that there is a significant difference and you reject the null hypothesis
2. There is a correlation between mental health disorders and levels of alcohol consumed. A large population study reveals that this is significant at the 5 per cent level and you reject the null hypothesis, which is defined as 'no correlation between the two'
3. There is a relationship between obesity and diabetes. In the study population you find no significant difference and cannot reject the null hypothesis at the 5 per cent level
4. There is no correlation between gender of the baby and mode of delivery. A midwife makes note of all the babies born over a few weeks on the labour ward. A significantly larger number of female babies were born by Caesarian section and you reject the null hypothesis

## EXPLANATION: TYPE 1 AND TYPE 2 ERRORS

If a null hypothesis is rejected when it is in fact true, then this is a type 1 error. In the example given (1) there is no relationship whatsoever between lung cancer and shoe size, however the results of the study meant that the null hypothesis was wrongly rejected.

The probability of making a type 1 error is often referred to as $\alpha$, which is equivalent to the $P$-value (see page 97). If the $P$-value $<0.001$, the chance of a type 1 error will be lower than if the $P$-value were $<0.5$.

If a null hypothesis is not rejected when it is in fact false this would be an example of a type 2 error. In the example (2) the data suggest there is no significant difference between rates of lung cancer when compared to smoking, whereas we all know that there is. The probability of making a type 2 error decreases as the sample size, and therefore the power of the study, increases (see page 115 for Power).

## When designing a study to avoid type 1 and type 2 errors:

- Choose a larger sample size. Both type 1 and type 2 errors decrease with increasing sample size
- Consider the type of variable that you are trying to measure (e.g. mean, proportion) and its level of statistical variation (e.g. standard error)
- Consider how strong the association (or difference) is that you are trying to detect
- Choose a sensible $P$-value; the stricter the requirements for rejecting the null hypothesis the lower the probability of a type 1 error.

[^32]
## 4. Power is

a. The probability of making an incorrect decision to reject the null hypothesis
b. 1 minus the probability of making a type 2 error
c. Increased with increasing sample size
d. Influenced by the study design
e. Represented as $1-\alpha$
5. In a hypothetical study the null hypothesis is that there is no difference in the reduction in blood pressure between patients receiving antihypertensive drug A and those receiving placebo drug B. Which of the following will increase power?
a. Choosing a blood pressure test/machine/operator that is highly sensitive and gives repeatable results
b. The fact that drug A causes a huge reduction in true blood pressure compared to drug B
c. A high prevalence of hypertension
d. Choosing the larger sample size of 1256 people who agreed to enter the trial when enrolled by the GP rather than the 16 who were in the waiting room
e. Deciding to use $P<0.05$ rather than $P<0.001$ as the threshold to reject the null hypothesis

## EXPLANATION: POWER

Power is the ability to demonstrate an association if one exists. A study with more power is less likely to have type 2 error. Statistically, power is defined as the correct rejection of the null hypothesis.

Taking the example of smoking and lung cancer where there is truly an association between rates of lung cancer and smoking. The probability that a study rejects the null hypothesis, to correctly conclude that there is an association between smoking and lung cancer, is known as the power of a study.

The relationship between the conclusion of a study and the true reality can be represented by type 1 and type 2 errors and power as follows:

| Conclusion of significance test | Reality with respect to null hypothesis |  |
| :--- | :--- | :--- |
|  | True | False |
| Reject $H_{0}$ | Type 1 error $(\alpha)$ | Power $(1-\beta)$ |
| Do not reject $H_{0}$ | $(1-\alpha)$ | Type 2 error $(\beta)$ |

When designing a study it is wise to focus on factors that will increase the power, one of the benefits of which is to reduce the probability of making a type 2 error. One factor that minimizes this is having a larger sample size.

The example used in question 5 demonstrates that many factors relating to (a) study design and (b) the true difference between drugs $A$ and $B$ can affect the power of a study.

Larger sample sizes will have more power but they will also be more costly, time consuming and can be difficult to obtain for practical reasons, for example, in the study of rare diseases. If the other factors in the study design prove to increase power, such as drug A being significantly better than placebo B, the sample size needed to demonstrate a statistically significant difference may not need to be as large. Needlessly large samples are considered a waste of resources. There is also an ethical argument against treating too many controls with placebo, when the real drug is substantially better. Some trials are even brought to an early end once an obvious conclusion is reached.

## Answers

4. F T T T F
5. T T T T T

## 6. Match up the type of bias to the study descriptions

## Options

A. Publication bias
C. Recall bias
B. Selection bias
D. Observer bias

1. Lung cancer patients remember their smoking habits better than patients with other cancers
2. A surgeon, with a large private practice, has decided to report the success rate of his operations
3. A study investigating the amount of alcohol consumption in the general population was conducted on a group of first year undergraduates
4. Doctors in a non-blinded placebo trial for a new treatment for back pain
5. How might each of the types of bias be reduced? Choose the best method(s) from the following list to tackle each of the main causes of bias

## Options

A. Publication bias
C. Recall bias
B. Selection bias
D. Observer bias

1. Using an 'outside', unbiased, assessor or auditor to regulate or check outcomes
2. Recording information on exposure before the outcome is known
3. Ensure that the person measuring or recording the outcome does not know about the exposure
4. Using methods such as random sampling to choose a representative cross-section of the population of interest
5. Can you think of possible confounding factors for the following associations?
a. Colorectal cancer and diet
b. Type 2 diabetes mellitus and heart disease
c. Alcohol consumption in pregnancy and fetal abnormalities
d. Smoking in the household and childhood asthma
e. Fractured hips in the elderly
f. Anxiety and depression

## EXPLANATION: BIAS AND CONFOUNDING

Bias is a consistent inaccuracy in the results of a study as a consequence of the method used to collect or interpret the data. It may be due to:

- Non-random selection of subjects, for example, a study using medical students as subjects may be biased as the study population is healthier than the general population
- Health-related selection is a specific type of selection bias: it describes the phenomenon of apparent worsening of health during the course of a prospective study. When most subjects enrol in a cohort study, they are in good health. Over several years, the apparent deterioration in health may be simply a reflection of the wearing off of this selection bias
- Inappropriate study design, such that results are forced in a particular direction. For example, if doctors are not blinded as to whether patients are receiving a treatment or placebo, the expectation of a treatment effect can result in observer bias
- Recall bias is a particular problem in retrospective studies. With the hindsight of a particular health outcome, subjects may remember or exaggerate their own past exposure to a particular risk factor, leading to a bias in the reporting of exposure
- Biased interpretation of data, for example, lack of blinding of the investigators and a vested interest in producing certain results, known as publication bias, may lead to results that do not reflect a genuine effect or association.

Confounding occurs when factors, other than the risk factor under identification, influence the outcome. The confounding factor:

- Influences the outcome independently from the exposure
- Is associated with the exposure independently from the outcome
- Differs between the cases and the controls.


Confounding in observational studies can be minimized by the use of effective controls or matching. Randomizaton is the best way to minimize confounding. If a trial is large enough, randomization should lead to equal distribution of confounding factors between groups being compared. This is particularly useful when confounders exist that have not yet been identified.

[^33]This page intentionally left blank


## APPENDIX

## NOTATION AND FORMULAE

The following is a summary of some of the more useful, or frequently required formulae for statistical interpretation of epidemiological results. Some of the simpler notations such as $n$ for number of observations and $\bar{x}$ for mean are used ubiquitously in maths, statistics and epidemiology. In general, the aim of this book has been to introduce ideas and principles, therefore equations, and 'nitty-gritty' details of their calculation, have been avoided. For more detailed explanations see Kirkwood, BR and Sterne, JAC. Essential Medical Statistics, 2nd edn. Blackwell Scientific Publications, 2003.

| Sample statistic | Symbol | Formulae |
| :--- | :--- | :--- |
| Number of observations | $n$ |  |
| Sum of values | $\Sigma x$ | $x_{1}+x_{2}+x_{3}+\ldots+x_{n}$ |
| Mean | $\bar{x}$ | $(\Sigma x) / n$ |
| Standard deviation | SD | $\sqrt{ }\left\{\Sigma(x-\bar{x})^{2}\right\} /(n-1)$ |
| Coefficient of variation | CV | $(\mathrm{SD} / \bar{x}) \times 100$ |
| Normal range | $\bar{x}-1.96 \times \mathrm{SD}$ to $\bar{x}+1.96 \times \mathrm{SD}$ |  |
| Standard error | SE | $\mathrm{SD} / \sqrt{n}$ |
| 95 per cent confidence interval <br> for single mean | 95 per cent CI | $\bar{x}-1.96 \times \mathrm{SE}$ to $\bar{x}+1.96 \times \mathrm{SE}$ |
| Test statistic for single mean (or <br> comparing two means in paired <br> data) | $z$ | $z=(\bar{x}-\mu) / \mathrm{SE}$ |
| Odds ratio | OR | $(a / c) /(b / d)=\mathrm{ad} / \mathrm{bc}$ |

## SUMMARY OF STATISTICAL TESTS

## QUANTITATIVE DATA

| Test | Circumstances | Assumptions | Null hypothesis | Test statistic |
| :---: | :---: | :---: | :---: | :---: |
| One-sample $z$-test (or $t$-test if small sample) | One group of numerical data. 'Is the mean different from expected?' | Normal distribution; reasonable sample size | Sample mean ( $\bar{x}$ ) <br> = hypothesized <br> value for mean ( $\mu$ ) | $\begin{aligned} & z=\frac{(\bar{x}-\mu)}{s / \sqrt{n}} \\ & n=\text { sample size, } \\ & s=\text { estimated } \\ & \text { standard deviation } \end{aligned}$ |
| Two-sample paired $z$-test (or $t$-test if small sample) | Two groups of paired numerical data. 'Is there a difference in means between the pairs?' | Normal distribution; reasonable sample size; the two groups are the same size | Mean of differences between sample pairs $=$ zero | $\begin{aligned} & z=\frac{\bar{d}}{s / \sqrt{n}} \\ & n=\text { number of } \\ & \text { paired differences, } \\ & \bar{d}=\text { mean difference } \\ & \text { in pairs, } \\ & s=\text { estimated standard } \\ & \text { deviation of paired } \\ & \text { differences } \end{aligned}$ |


| Two-sample unpaired $z$-test (or $t$-test if small sample) | Two groups of unpaired numerical data. 'Is there a difference between the means of the two groups?' | Normal distribution of data in both groups; groups have same variance; reasonable sample sizes | Difference between sample means $=$ zero | $z=\frac{\bar{x}_{1}-\bar{x}_{2}}{\sqrt{\left(\mathrm{SE}_{1}{ }^{2}+\mathrm{SE}_{2}{ }^{2}\right)}}$ |
| :---: | :---: | :---: | :---: | :---: |
| ANOVA (analysis of variance) | Multiple groups of numerical data: single test. 'Is there a difference between the means in each group?' | Mean is same in each group |  | $F$ distribution: a complex calculation which can be carried out using computer packages. |

CATEGORICAL DATA

| Test | Circumstances | Assumptions | Null hypothesis | Notation |
| :---: | :---: | :---: | :---: | :---: |
| One-sample z-test | Two categories, one group of data 'Is there a difference in the proportion of the sample with the characteristic of interest compared to a hypothesized value for the proportion?' | Data follow the binomial distribution which approximates the normal distribution if the sample size is large enough | Population proportion, $\pi=$ hypothesized proportion, $\pi_{1}$ | $\begin{aligned} & z=\frac{\|p-\pi\|-(1 / 2 n)}{\sqrt{(p(1-p) / n)}} \\ & p=\text { population with } \\ & \text { characteristic of } \\ & \text { interest }=r / n ; \\ & r=\text { number with } \\ & \text { characteristic, } \\ & n=\text { sample size } \end{aligned}$ |
| Two-sample chi-squared test $\chi^{2}$ | Two categories with $>2$ groups of data. (See table on page 123) | Groups are independent and mutually exclusive; the expected frequency in each category > 5 | There is no difference between the observed frequency and the expected frequency (frequency expected if there was no difference between the two groups) | $\chi^{2}=\frac{\Sigma(\|O-E\|-1 / 2)^{2}}{E}$ $\begin{aligned} \mathrm{df} & =n-1 \\ \mathrm{O} & =\text { observed } \end{aligned}$ <br> frequency, <br> $E=$ expected frequency |
|  | > 2 categories; data expressed in an $r \times c$ table | Mutually exclusive categories; expected frequency > 5 in 80 per cent of the categories |  | $\chi^{2}=\frac{\Sigma(\|O-E\|)^{2}}{E}$ $\mathrm{df}=(r-1) \times(c-1)$ <br> when using chi-squared test df should be given |
| McNemar's test | Two categories; two groups of data; categories not mutually exclusive (paired) |  |  |  |
| Fisher's exact test | Two categories; 2+ groups of | Mutually exclusive categories |  |  |

data; expected frequency $<5$

There is no difference between the observed fequency and the frequency expected cted difference between the two groups)
$p=$ population with haracteristic of $r=$ number with characteristic, $n=$ sample size
$\chi^{2}=\frac{\Sigma(|O-E|-1 / 2)^{2}}{E}$
$\mathrm{df}=n-1$
O observed
frequency, $E=$ expected frequency
$\chi^{2}=\frac{\Sigma(|O-E|)^{2}}{E}$
$\mathrm{df}=(r-1) \times(c-1)$
when using chi-squared test df should be given

CHI-SQUARED CONTINGENCY TABLE

|  | Group 1 | Group 2 | Total |
| :--- | :--- | :--- | :--- |
| Characteristic present <br> (observed frequency, O) | $a$ | $b$ | $a+b$ |
| Characteristic absent | $c$ | $d$ | $c+d$ |
| Total | $a+c$ | $b+d$ | $a+b+c+d$ |
| Proportion with characteristic | $p_{1}=a /(a+c)$ | $p_{2}=b /(b+d)$ | $p=a+b /(a+b+c+d)$ |
| Expected frequency $(E)$ | $E_{1}=(a+c) \times p$ | $E_{2}=(b+d) \times p$ |  |

## STATISTICAL TABLES

THE STANDARD NORMAL DISTRIBUTION


[^34]STUDENT'S $t$-DISTRIBUTION

| df | Two-tailed $P$-value |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | 0.1 | 0.05 | 0.01 | 0.001 |
| 1 | 6.31 | 12.71 | 63.66 | 636.58 |
| 2 | 2.92 | 4.30 | 9.92 | 31.60 |
| 3 | 2.35 | 3.18 | 5.84 | 12.92 |
| 4 | 2.13 | 2.78 | 4.60 | 8.61 |
| 5 | 2.02 | 2.57 | 4.03 | 6.87 |
| 6 | 1.94 | 2.45 | 3.71 | 5.96 |
| 7 | 1.89 | 2.36 | 3.50 | 5.41 |
| 8 | 1.86 | 2.31 | 3.36 | 5.04 |
| 9 | 1.83 | 2.26 | 3.25 | 4.78 |
| 10 | 1.81 | 2.23 | 3.17 | 4.59 |
| 11 | 1.80 | 2.20 | 3.11 | 4.44 |
| 12 | 1.78 | 2.18 | 3.05 | 4.32 |
| 13 | 1.77 | 2.16 | 3.01 | 4.22 |
| 14 | 1.76 | 2.14 | 2.98 | 4.14 |
| 15 | 1.75 | 2.13 | 2.95 | 4.07 |
| 16 | 1.75 | 2.12 | 2.92 | 4.01 |
| 17 | 1.74 | 2.11 | 2.90 | 3.97 |
| 18 | 1.73 | 2.10 | 2.88 | 3.92 |
| 19 | 1.73 | 2.09 | 2.86 | 3.88 |
| 20 | 1.72 | 2.09 | 2.85 | 3.85 |
| 21 | 1.72 | 2.08 | 2.83 | 3.82 |
| 22 | 1.72 | 2.07 | 2.82 | 3.79 |
| 23 | 1.71 | 2.07 | 2.81 | 3.77 |
| 24 | 1.71 | 2.06 | 2.80 | 3.75 |
| 25 | 1.71 | 2.06 | 2.79 | 3.73 |
| 26 | 1.71 | 2.06 | 2.78 | 3.71 |
| 27 | 1.70 | 2.05 | 2.77 | 3.69 |
| 28 | 1.70 | 2.05 | 2.76 | 3.67 |
| 29 | 1.70 | 2.05 | 2.76 | 3.66 |
| 30 | 1.70 | 2.04 | 2.75 | 3.65 |
| 40 | 1.68 | 2.02 | 2.70 | 3.55 |
| 50 | 1.68 | 2.01 | 2.68 | 3.50 |
| 100 | 1.66 | 1.98 | 2.63 | 3.39 |
| 1000 | 1.65 | 1.96 | 2.58 | 3.30 |

THE CHI-SQUARED DISTRIBUTION


1-tailed tests 93
$2 \times 2$ tables 29, 103, 110
2-tailed tests 93
absolute risk 37, 104-5
absolute risk reduction 104-5
accuracy 84-5
alternative hypothesis $\left(H_{1}\right) 93$
analysis of variance (ANOVA) 95, 121
analytical studies
case-control 5, 8-9, 25-32
cohort 5, 8-9, 33-41
cross-sectional 5, 8-9, 17-23
meta-analysis $8-9,51-8$
randomized controlled trials $8-9$, 43-9
ANOVA see analysis of variance associations 8-9, 100-1, 103, 106-7
assumptions $109,121,122$
Bayes' theorem 68-9
best fit 107
bias 19, 41, 49, 57, 116-17
publication 53, 57, 116-17
responder 27
selection 31, 116-17
volunteer 49
bimodal distribution 74
binomial distribution 80-1, 90
blinding 41, 45, 49
Cartesian axes 107
case-control studies 5, 8-9, 25-32
categorical data 95
causation 100-1
central location 76-7
chi-squared contingency table 123
chi-squared distribution tables 126
chi-squared test $32,95,122$
CI see confidence intervals
clinical epidemiology 59-69
coefficient of variation 120
cohort studies 5, 8-9, 33-41
compliance 46-7, 49
compliance adjustment formula 46-7 confidence intervals (CI) 88-90, 93, 120
confounding factors 116-17
case-control studies 27, 31
cohort studies 37, 41
cross-sectional studies 19
ecological studies 15,16
contingency table 123
correlation 16, 101, 106-9
costs of studies
case-control 31
cohort 41
cross-sectional 23
ecological 15
meta-analysis 57
randomized controlled trials 49
screening programmes 63
Cox proportional hazard model 39
credible interval 54, 58
crossover designs 45
cross-sectional studies 5, 8-9, 17-23
cut-off values 65-9
data description 73-82
denominators 4-5, 7
dependence 57
descriptive cross-sectional studies 19
determinants of disease 4-5
distribution 4-5, 80-2
bimodal 74
binomial 80-1, 90
Gaussian 81
J-shaped 74
normal 74, 80-1, 90, 95, 97, 124
Poisson 80-2
probability distribution curves 97
dose dependence 31
double blinding 45
dropout rates 41, 47, 49
'ecological fallacy' 14-15
ecological studies 5, 8-9, 11-16
enrolment 41
error
sampling 85
sources 111-17
standard error 78-9, 86-7, 89, 109, 120
standard error of the mean 87,89 , 90
standard error of the proportion 87
types 1 and $2112-13,115$
estimation 83-90
ethical issues 47, 49
exclusion criteria 53
factorial 81
false negatives 65-7
false positives 65-7
Fisher's exact test 122
follow-up loss 35, 39, 41
forest plots 54-5, 58
formulae 120
Gaussian distribution 81
geographical comparisons 13, 16, 19
gold standards 9, 45, 67
$H_{0}$ see null hypothesis
$H_{1}$ see alternative hypothesis
health-related selection 41, 117
hidden confounders 41
histograms 74-5
hypothesis testing 31, 32, 91-7
incidence rates 6-7, 41
inclusion criteria 53
inconsistency between meta-analysis studies 57
independence assumptions 109
interpretation bias 117
interpretation of data 99-109
interquartile range 78-9
intervention studies 9, 43-9
inverse variance 53
J-shaped distribution 74
Kaplan-Meier survival curves 39
Kruskal-Wallis test 95
likelihood ratios 68-9
linear regression 106-9
line, slope of 107,109
location 76-7
log rank 39
long-term outcomes 23
loss of subjects $35,39,41,49$
McNemar's test 95, 122
Mann Whitney $U$-test 95
matched studies 27, 117
mean 74-7, 120
median 74-7
meta-analysis 8-9, 51-8
migrant studies 13
mode 75, 76-7
modulus 32, 110
multiple outcomes 41
National Screening Committee criteria 63
negative predictive values 67
negative skew 74-5
NNT see numbers needed to treat
non-compliance 46-7
non-parametric data 95
non-random selection 117
normal distribution 74, 80-1, 90, 95, 97, 124
normal range 79, 120
notation 120, 122
null hypothesis $\left(H_{0}\right) 92-3,97,112-15$, 121, 122
numbers needed to treat (NNT) 105
numerators 4-5, 7
numerical data 95
observational studies
case-control 5, 8-9, 25-32
cohort 5, 8-9, 33-41
cross-sectional 5, 8-9, 17-23
ecological 5, 8-9, 11-16
observer bias 116-17
occupational groups 13
odds ratio (OR) 101, 102-3, 120
case-control studies 29
meta-analysis 55, 58
one-tailed tests 93
OR see odds ratio
outliers 75
over-matching 27
overview approach 9, 51-8
parametric data 95
Pearson correlation coefficient 107
pie charts 74-5
placebos 45, 115
Poisson distribution 80-2
population-level data 13
positive predictive values 67
positive skew 74-5
posterior odds of disease 69
posterior probability 68-9
power 15, 57, 114-15
precision 57, 84-5
predictive values 66-7
pre-symptomatic stage 61
prevalence 6-7, 18-21, 23
primary prevention 61
principles of epidemiology 4-5
prior odds 69
prior probability 68-9
probability 68-9, 80, 81, 113
probability distribution curves 97
proportional hazard model 39
prospective studies 9,35
publication bias 53, 57, 116-17
published data/statistics $13,15,53,55$
$P$-values 32, 93, 96-7, 113, 124-6
randomized controlled trials 8-9, 43-9
range 78-9
rare diseases and exposures 31, 41
rate ratio 12,16
recall bias 31, 116-17
registers 7
regression 106-9
relative risk 101, 103, 104-5
case-control studies 29
cohort studies 37
meta-analysis 55
relative risk reduction 104-5
repetition avoidance 57
residuals 107
responder bias 27
retrospective studies 9, 27, 35, 41
risk $29,37,55,101,103,104-5$
routinely collected data 7,13
sample size 21, 46-7, 49, 87, 115
sampling distribution of the mean 87
sampling distribution of the proportion 87
sampling error 85
scatter plots 16, 74-5
screening programmes 60-3
SD see standard deviation
SE see standard error
selection bias 31, 116-17
SEM see standard error of the mean
sensitivity 64-9
SE $(p)$ see standard error of the proportion
significance level 97
skewed data 74-5
slope of the line 107, 109
socio-economic groups 13,19
specificity 64-9
spread of data 78-9
standard deviation (SD) 78-9, 87, 120
standard error (SE) 78-9, 86-7, 89, 109, 120
standard error of the mean (SEM) 87, 89, 90
standard error of the proportion $(\operatorname{SE}(p))$ 87
standard normal distribution tables 124
statistical heterogeneity 55
Student's $t$-distribution/test 90, 95, 121, 125
study designs 8-9, 117
sum of the squares 107
sum of values 120
survival curves 38,39
tables $(2 \times 2) 29,103,110$
$t$-distribution/test 90, 95, 121, 125
temporal comparisons/relationships 13, 16, 41
test statistic calculation 94-5
'time to event' comparison 38, 39
trends 16, 21
true negatives 65-7
true positives 65-7
$2 \times 2$ tables 29, 103, 110
two-tailed tests 93
type 1 errors 112-13, 115
type 2 errors 112-13, 115
United Kingdom screening programmes 61
unpublished data 53
variability 55,107
variance 53,79
volunteer bias 49
weighting 53, 57
Wilcoxon rank sum test 95
Wilcoxon signed rank test 95
Wilcoxon two-sample tests 39
$z$-tests $95,121,122$


[^0]:    Answers
    6. See explanation
    7. a - 38/170; 22 per cent; 22 smokers per 100 students, $b-65 / 340 ; 19$ per cent; 19 per 100 secretaries get RSI
    8. a - Injuries per year $=4 \times 12=48 ; 48 / 680=0.07 ; 7$ injuries per 100 workers per year, $b-46$ cases per year/5800 primary school children $=0.0079 ; 79$ leukaemia cases per 10000 primary school children per year
    9. $30914 / 9200000=0.0034 ; 34$ deaths per 10000 population per year in the over-65-year age group
    10. F F T T T

[^1]:    Answers

    1. See explanation
    2. See explanation
    3. See explanation
    4. See explanation (page 16)
    5. See explanation (page 16)
[^2]:    Answers
    6. F F T
    7. FTFTTF
    8. T T F F F

[^3]:    Answers

    1. a - See explanation, b-See explanation, c - Prevalence data, d - See explanation
    2. F F F T
[^4]:    Answers
    3. See explanation
    4. See explanation
    5. See explanation
    6. Increase sample size
    7. See explanation

[^5]:    Answers
    8. T F T F T T T
    9. FT T T T
    10. T F T F
    11. T T F T T

[^6]:    Cl , confidence interval; OR , odds ratio

[^7]:    Answers
    12. T F F T T
    13. T T F F
    14. F F F F F

[^8]:    Answers
    2. T T F T
    3. $\mathrm{a}-2.5$ per cent, $\mathrm{b}-95$ per cent (look at the confidence intervals)
    4. $1-\mathrm{D}, 2-\mathrm{E}, 3-\mathrm{B}, 4-\mathrm{A}, 5-\mathrm{C}$

[^9]:    Answers
    5. $a-(2 / 20) / 1=10$ per cent, $b-(5 / 20) / 7=3.6$ per cent, $c-(6 / 20) / 14=2.1$ per cent
    6. Over the possible 280 days of skiing $(20 \times 14)$ only 225 were spent skiing. This is 11.25 days each
    7. T T F T
    8. F F F T

[^10]:    Answers
    9. F T F T T T
    10. F T T F

[^11]:    Answers

    1. See explanation
    2. T T T T
[^12]:    Answers
    3. T F T
    4. F T F
    5. Adjusted $n=100 /(0.8+0.8-1)^{2}=278$
    6. $a-64, b-144, c-$ Larger

[^13]:    Answers
    7. T F T
    8. F F T
    9. By incorporating features to reduce bias and confounding as well as choosing a sample size sufficient to detect clinically significant differences
    10. A D F E C B

[^14]:    ${ }^{1}$ Sackett DL, Rosenberg WMC, Gray JAM, Haynes, RB and Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71-72.

[^15]:    Answers

    1. See explanation
    2. See explanation
    3. F F F T
    4. $1-\mathrm{A}, 2-\mathrm{A}, 3-\mathrm{B}$
[^16]:    OR, odds ratio; SSRI, selective serotonin reuptake inhibitor; CI, confidence interval
    NB: credible interval is the Bayesian equivalent of confidence intervals, as the Bayesian random effects model was used to synthesize data across all the different SSRIs

[^17]:    Answers
    5. $a-N o$ increased risk of suicide ( $O R=0.85, C l i n c l u d e s ~ t h e ~ v a l u e ~ o f ~ 1), ~ b-A ~ s l i g h t ~ i n c r e a s e ~ i n ~ t h e ~ r i s k ~ o f ~ s e l f-h a r m ~(~ O R ~=~ 1.57, ~ C l ~>~ 1), ~$ $c-$ No increased risk of suicidal thoughts $(O R=0.77, \mathrm{Cl}$ includes the value of 1)
    6. See explanation
    7. See explanation

[^18]:    Answers
    11. See explanation
    12. See explanation

[^19]:    Answers
    13. See explanation
    17. T F T T F T F F
    14. See explanation
    18. See explanation
    15. See explanation
    19. See explanation
    16. See explanation

[^20]:    Answers
    20. a - 15 per cent, $\mathrm{b}-16$ per cent
    21. a - Graph $A$ has a cut-off value of 0.1 and graph $B$ has a cut-off value of 0.15 . The cut-off of 0.15 has low specificity and graph $B$ reflects this as the post-test probability for a positive result (dashed) is low (i.e. close to the solid line), b-0.1, c-0.1

[^21]:    Answers
    4. $a-7.3 \mathrm{mmol} / \mathrm{L}, \mathrm{b}-6.65 \mathrm{mmol} / \mathrm{L}, \mathrm{c}-5.9 \mathrm{mmol} / \mathrm{L}, \mathrm{d}$ - The median
    5. $a-27, b-30, c-30, d-$ The mode
    6. $1-\mathrm{A}, 2-\mathrm{C}, 3-\mathrm{A}, 4-\mathrm{B}, 5-\mathrm{A}, 6-\mathrm{B}, 7-\mathrm{C}, 8-\mathrm{C}$

[^22]:    Answers
    7. 1 - Diii, 2 - Ei, 3 - Aii, 4 - Cv, 5 - Biv
    8. See explanation

[^23]:    Answers

    1. See explanation
    2. See explanation
    3. See explanation
    4. a - Accuracy, b - Precision, c - Precision
    5. a - Most accurate, b-Most precise
[^24]:    Answers
    6. $\mathrm{a}-\mathrm{SEM}=2.3 / \sqrt{ } 1600=0.0575 \mathrm{mmol} / \mathrm{L} ; \mathrm{b}-$ Proportion, $p=30 / 5700=0.0053 ; \mathrm{SE}(p)=\sqrt{ }(0.0053 \times(1-0.0053) / 5700)=0.00096$ women; c - SEM $=2.3 / \sqrt{ } 16=0.58 \mathrm{mmol} / \mathrm{L} ; \mathrm{d}-$ SEM $=120 / \sqrt{ } 16=30 \mathrm{~mL} / \mathrm{min}$
    7. F F T F F
    8. F T T F
    9. F T F T T

[^25]:    Answers
    10. See explanation 11. See explanation (page 90)
    12. $a-$ Smokers: $S E=17 / \sqrt{ } 70=2.0$; non-smokers: $S E=10 / \sqrt{ } 70=1.2, b-$ Smokers: 95 per cent $C I=149 \pm(1.96 \times 2.0)=149 \pm 3.92$. 95 per cent Cl is 145 to 153 ; non-smokers: 95 per cent $\mathrm{Cl}=131 \pm(1.96 \times 1.2)=131 \pm 2.35 .95$ per cent Cl is 129 to 133 , c - The SEM is smaller for non-smokers than for smokers, suggesting that the estimate of the mean blood pressure is more precise for non-smokers than smokers. The narrower Cl for non-smokers suggests that the estimate is also more accurate than for smokers
    13. F T F T 14. E

[^26]:    Answers

    1. See explanation
    2. See explanation
    3. $C D B E A$
    4. See explanation
    5. $F$ T F F F
    6. See explanation
[^27]:    Answers
    7. 1 - D, 2 - E, $3-\mathrm{B}, 4-\mathrm{C}, 5-\mathrm{A}$
    8. Two-sample unpaired comparison of means, $z=(5.4-4.6) /[1.5 / \sqrt{ }(1 / 67+1 / 54)]=0.098$

[^28]:    Answers

    1. $1-\mathrm{H}, 2-\mathrm{I}, 3-\mathrm{A}, 4-\mathrm{F}, 5-\mathrm{D}, 6-\mathrm{C}, 7-\mathrm{E}, 8-\mathrm{B}, 9-\mathrm{G}$
    2. T T F F T
[^29]:    Answers
    9. $D$ is ten times as effective as the other options as they all convert to a number needed to treat of 170
    10. $a-0.6$ per cent, $b-((60-54) / 60) \times 100)=10$ per cent, $c-0.06$ per cent, $d-1 / 0.0006=1667$ people, $e-$ Probably the relative risk reduction (10 per cent)
    11. $a$ - Absolute risk group $B=10 / 80=12.5$ per cent; absolute risk group $A=5 / 50=10$ per cent; absolute risk reduction $=12.5-10$
    $=2.5$ per cent, $b-$ Relative risk reduction $=$ absolute risk reduction/absolute risk group $B=2.5 / 12.5=0.2$ or 20 per cent, $c-N N T=40$

[^30]:    Answers
    12. See explanation 13. a - Positive, $r=+1, \mathrm{~b}$ - zero $r=0, \mathrm{c}-$ Negative $r=-1$ 14. a - Positively correlated, b - No correlation, c Negatively correlated 15. See explanation
    16. a - Null hypothesis: drug has no effect on serum cholesterol; alternative hypothesis: drug dose influences serum cholesterol, b-Test statistic $=0.6 / 0.28=2.14, \mathrm{df}=20-2=18, P<0.05$, i.e. results are significant at the 5 per cent level: reject the null hypothesis, $c-$ An increase of $0.1 \mathrm{mmol} / \mathrm{L}$, a decrease of $0.5 \mathrm{mmol} / \mathrm{L}$, a decrease of $2.3 \mathrm{mmol} / \mathrm{L}$, a decrease of $5.3 \mathrm{mmol} / \mathrm{L}$, d - With no drug treatment, $(x=$ 0 ), cholesterol levels can be expected to rise by $0.7 \mathrm{mmol} / \mathrm{L}$. Very low drug doses are insufficient to coonteract this rise.

[^31]:    Answers
    17. T F F F
    18. $a$ - No reason; it is appropriate to use linear regression, $b-2, c-4, d-1$
    19. T T F F

[^32]:    Answers

    1. T T F F (more likely a type 2 error) F (inversely related rather than directly)
    2. $F T T F$
    3. $1-\mathrm{A}, 2-\mathrm{C}, 3-\mathrm{B}, 4-\mathrm{A}$
[^33]:    Answers
    6. $1-C, 2-A, 3-B, 4-D$
    7. $1-\mathrm{A}, \mathrm{D}, 2-\mathrm{C}, 3-\mathrm{A}, \mathrm{D}, 4-\mathrm{B}$
    8. a - Age, gender, family history, smoking, b - Age, gender, obesity, vascular disease, reduced exercise, c - Smoking, drug use, d Reduced income, overcrowding, more dusty environment, no heating, e - Increased osteoporosis, increased confusion or dementia, reduced balance, more falls, f - Increased life stressors, increased alcohol or drug abuse, unemployment, poverty, other forms of mental illness

[^34]:    df, degrees of freedom

