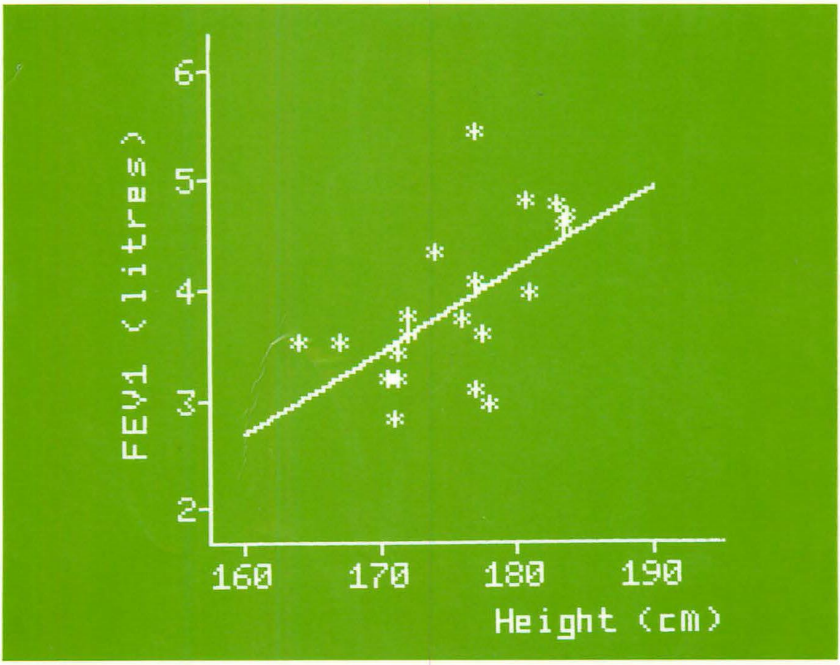
MARTIN BLAND

An Introduction to Medical Statistics



OXFORD MEDICAL PUBLICATIONS

Richard Pearson March 93

OXFORD MEDICAL PUBLICATIONS

An Introduction to Medical Statistics

An Introduction to Medical Statistics

MARTIN BLAND

Senior Lecturer in Medical Statistics
St. George's Hospital Medical School
London

Oxford New York Tokyo
OXFORD UNIVERSITY PRESS

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Petaling Jaya Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

First published 1987
Reprinted 1988 (with corrections), 1989
Reprinted 1990 (with further corrections), 1991, 1992

British Library Cataloguing in Publication Data
Bland, Martin
An introduction to medical statistics.
—(Oxford medical publications).
1. Medical statistics
I. Title
610'.212 RA407
ISBN 0–19–261502–5

Library of Congress Cataloging in Publication Data
Bland, Martin
An introduction to medical statistics.
Includes index.
1. Medical statistics. I. Title. [DNLM:
1. Biometry. 2. Statistics. WA950 B642i]
RA409.B55 1987 610'.1'5195 87–5606
ISBN 0–19–261502–5 (pbk.)

Printed and bound in Great Britain by
Biddles Ltd, Guildford and King's Lynn

To Ernest Bland

Preface

This is a textbook of statistics for medical students, doctors, medical researchers, and others concerned with medical data. I hope that it will also be of interest to students whose principal interest is statistics or mathematics, who often find that the practical application of statistical methods is the most difficult part of the subject.

The fundamental concepts of study design, data collection, and data analysis are explained by illustration and example. For those who wish to go a little further in their understanding, some of the mathematical background to the techniques described is also given, largely as appendices to the chapters rather than in the main text.

The material covered includes all the statistical work that would be required for a course in medicine and for the examinations of most of the royal colleges. It includes the design of clinical trials and epidemiological studies, data collection, summarizing and presenting data, probability, the Binomial, Normal, Poisson, t and Chi-squared Distributions, standard errors, confidence intervals, tests of significance, large sample and small sample comparisons of means, the use of transformations, regression and correlation, methods based on ranks, contingency tables, measurement error, reference ranges, mortality data, vital statistics and the choice of statistical method.

The book is firmly grounded in medical data, particularly in medical research, and the interpretation of the results of calculations in their medical context is emphasized. Except for a few obviously invented numbers used to illustrate the mechanics of calculations, all the data in the examples and exercises are real, from my own research and statistical consultation, or from the medical literature, to which reference is made where possible.

There are two kinds of exercise. Each chapter has five multiple choice questions of the true or false type. Multiple choice questions can cover a large amount of material in a short time, so are a very useful tool for revision. As MCQs are widely used in postgraduate examinations, these exercises should also be useful to those preparing for memberships. All the MCQs have solutions, with reference to the appropriate part of the text or a detailed explanation for most of the answers. Each chapter also has one long exercise. Although these usually involve calculation, I have tried to avoid merely slotting figures into formulae. These exercises include not only the applica-

tion of statistical techniques but also the interpretation of the results in the light of the source of the data.

I wish to thank many people who have contributed to the writing of this book. Firstly, there are the many medical students, doctors, research workers, and nurses whom it has been my pleasure to teach, and from whom I have learned so much. Secondly, the book contains many examples drawn from research carried out with other statisticians, epidemiologists, and social scientists, particularly Doug Altman, Ross Anderson, Mike Banks, Beulah Bewley, and Walter Holland. These studies could not have been done without the assistance of Patsy Bailey, Bob Harris, Rebecca McNair, Janet Peacock, Swatee Patel, and Virginia Pollard. Thirdly, the clinicians and scientists with whom I have collaborated or who have come to me for statistical advice have not only taught me a lot about medical data, but many of them have left me with data which are used here, including Thomas Bewley, Peter Fish, Nick Hall, Tessi Hanid, Michael Hutt, Ian Johnston, Pam Luthra, Hugh Mather, Daram Maugdal, Douglas Maxwell, Charles Mutoka, Tim Northfield, Paul Richardson, and Alberto Smith. I am particularly indebted to John Morgan, as Chapter 16 is partly based on his work. The manuscript was typed by Sue Nash, Sue Fisher, Susan Harding, and Sheila Skipp. An earlier draft of the book was read by David Jones, Doug Altman, Robin Prescott, Klim McPherson, and Stuart Pocock. Their comments have made this a better book than it would otherwise have been; the faults which remain are my own. Special thanks are due to my head of department, Ross Anderson, for all his support and to the staff of Oxford University Press. Most of all, I thank Pauline Bland for her unfailing confidence and encouragement.

London M. B.
January 1987

Contents

1. Introduction

1.1. The scope of statistics

Statistics, in the sense of numerical data, surround us in daily life. We may be told that nine out of ten cats prefer a certain catfood, or that the annual rate of inflation is 9.5 per cent, and are expected to understand what is meant. There have been acrimonious public debates about changes in the calculation of the unemployment rate, and the public are expected to decide on competing claims about the relative numbers of nuclear warheads. Statistics as an academic study is the science of assembling and interpreting numerical data, and we can see that some knowledge of this would be useful in many fields.

In clinical medicine, statistical methods are used to determine the accuracy of measurements, to compare measurement techniques, to assess diagnostic tests, to determine normal values, to estimate prognosis and to monitor patients. In the administration of medical services we are concerned with such things as bed use and perinatal mortality rates. It is in medical research, however, that statistics becomes most intimately involved, and it is with this area of application that this book is principally concerned. This is not to say that the book is addressed to the present or future researcher only. The medical profession is fond of research, but many doctors never try it. What nearly all doctors do is use the results of medical research, whether they are prescribing a new drug or advising a patient to give up smoking. In order to read the results of the enormous amount of research that pours into the medical journals, all doctors should have some understanding of the ways in which studies are designed, and data are collected, analysed and interpreted. That is what this book is about.

1.2. Statistics and medical research

In the past thirty years medical research has become deeply involved with the techniques of statistical inference. The work published in medical journals is full of statistical jargon and the results of statistical calculations. This acceptance of statistics, though gratifying to the medical statistician, may even have gone too far. More than once I have told a colleague that he did not need me to prove that his difference existed, as anyone could see it, only to be

told in turn that without the magic of the p-value he could not have his paper published.

Statistics has not always been so popular with the medical profession. Statistical methods were first used in medical research in the nineteenth century by workers such as Pierre-Charles-Alexandre Louis, William Farr, and John Snow. Snow's studies of the modes of communication of cholera, for example, made use of epidemiological techniques upon which we have still made little improvement. Despite the work of these pioneers, however, statistical methods did not become widely used in clinical medicine until the middle of the twentieth century. It was then that the methods of randomized experimentation and statistical analysis based on sampling theory which had been developed by Fisher and others were introduced into medical research, notably by Bradford Hill. It rapidly became apparent that research in medicine raised many new problems in both design and analysis, and much work has been done towards solving these by clinicians, statisticians, and epidemiologists.

Although considerable progress has been made in such fields as the design of clinical trials, there remains much to be done in developing research methodology in medicine. It seems likely that this will always be so, for every research project is something new, something which has never been done before. Under these circumstances we make mistakes. No piece of research can be perfect and there will always be something which, with hindsight, we would have changed. Furthermore, it is often from the flaws in a study that we can learn most about research methods. For this reason, the work of several researchers is described in this book to illustrate the problems into which their designs or analyses led them. I do not wish to imply that these people were any more prone to error than the rest of the human race, or that their work was not a valuable and serious undertaking. Rather, I want to learn from their experience of attempting something extremely difficult, trying to extend our knowledge, so that researchers and consumers of research may avoid these particular pitfalls in the future. We are sure to find more.

1.3. Statistics and mathematics

Many people are discouraged from the study of statistics by a fear of being overwhelmed by mathematics. It is true that many professional statisticians are also mathematicians, but not all are, and there are many very able appliers of statistics to their own fields. It is possible, though perhaps not very useful, to study some branches of statistics simply as a part of mathematics, with no concern for its application at all. Other aspects may be discussed without appearing to use any mathematics at all, as in Darrell Huff's *How to lie with statistics* (Huff 1954).

The aspects of statistics described in this book can be understood and applied with the use of simple algebra. Only the algebra which is essential for explaining the most important concepts is given in the main text. This means that several of the theoretical results used are stated without a discussion of their mathematical basis. This is done when the derivation of the result would not aid much in understanding the application. For many readers the reasoning behind these results is not of great interest. For the reader who does not wish to take these results on trust, several chapters have appendices in which simple mathematical proofs are given. These appendices are designed to help increase the understanding of the more mathematically inclined reader and to be omitted by those who find that the mathematics serves only to confuse.

1.4. Statistics and computing

Practical statistics has always involved large amounts of calculation. When the methods of statistical inference were being developed in the first half of the twentieth century, calculations were done using pencil, paper, tables, slide rules, and with luck a very expensive mechanical adding machine. Older books on statistics spend much time on the details of carrying out calculations and any reference to a 'computer' means a person who computes, not an electronic device. The development of the digital computer has brought changes to statistics as to many other fields. Calculations can be done quickly, easily and, we hope, accurately with a range of machines from pocket calculators with built-in statistical functions to powerful computers analysing data on many thousands of subjects. There is therefore no need to consider in detail the problems of manual calculation. The important thing is to know what the results of calculations actually mean. Indeed, the danger in the computer age is not so much that people may carry out complex calculations wrongly, but that they may apply very complicated statistical methods without knowing why, or without knowing what the computer output means. More than once I have been approached by a researcher bearing a computer print-out two inches thick, and asking what it all means. Sadly, too often, the answer is that another tree has died in vain.

Computers are a great benefit to statistics in that calculations which would once have taken days can now be done in minutes, and statisticians use them a lot. Most of the calculations in this book were done using a computer and all of the graphs were drawn with one. But the widespread availability of computers means that more calculations are being done, and being published, than ever before, and the chance of inappropriate statistical methods being applied may actually have increased. This arises partly because people regard their data analysis problems as computing problems, not statistical ones, and seek advice from computer experts rather than statisticians. They often get good advice on how to do it, but rather poor advice

about what to do, why to do it and how to interpret the results afterwards. It is therefore more important than ever that doctors, the consumers of research, understand something about the uses and limitations of statistical techniques.

1.5. The scope of this book

This book is intended as an introduction to some of the statistical ideas important to medicine. It does not tell you all you need to know to do medical research. Once you have understood the concepts discussed here, it is much easier to learn about the techniques of statistical analysis required to answer any particular question. There are several excellent standard works which describe the solutions to problems in the analysis of data (Armitage 1971; Colton 1974; Snedecor and Cochran 1980) and also more specialized books to which reference will be made where required.

What I hope the book will do is to give enough understanding of the statistical ideas commonly used in medicine to enable the doctor to read the medical literature competently and critically. It covers enough material for an undergraduate course in statistics for medical students and enough to answer statistical questions set in the examinations of most of the medical Colleges. At the time of writing, as far as can be established, it covers the material required for the MRCP, FRCS, FFA, MRCGP and MRCOG. It is not adequate for the MRC Psych., MFCM or FRCR, which require considerably more.

When working through a textbook, it is useful to be able to check your understanding of the material covered. Like most such books, this one has exercises at the end of each chapter, but to ease the tedium most of these are of the multiple-choice type. There is also one long exercise, usually involving calculations, for each chapter. In keeping with the computer age, intermediate results are given to avoid laborious calculation. Thus, the exercises can be completed quite quickly and the reader is advised to try them. Solutions are given at the end of the book, in full for the long exercises and as brief notes with references to the relevant sections in the text for multiple-choice questions (MCQs). Readers who would like more exercises are recommended to read Osborn (1979).

Finally, a question many students of medicine ask as they struggle with statistics: is it worth it? As Altman (1982) has argued, bad statistics leads to bad research, and bad research is unethical. Not only may it give misleading results, which can result in good therapies being abandoned and bad ones adopted, but it may also expose patients to potentially harmful new treatments for no good reason. Medicine is a rapidly changing field. In ten years' time, many of the therapies currently prescribed and many of our ideas about the causes and prevention of disease will be obsolete. They will have been

replaced by new therapies and new theories, supported by research studies and data, of the kind described in this book, and probably presenting many of the same problems in interpretation. Doctors will be expected to decide for themselves what to prescribe or advise on the basis of these studies. So a little knowledge of medical statistics would be one of the most useful things all doctors could acquire during their training.

2. The design of experiments

2.1. Comparing treatments

It is useful to distinguish between two broad types of study in medical research: observational and experimental. In observational studies, aspects of an existing situation are observed, as in a survey or a clinical case study. We then try to interpret our data to give an explanation of how the observed state of affairs has come about. In experimental studies, we do something in order to observe the result. This chapter is concerned with the way statistical thinking is involved in the design of experiments, particularly comparative experiments where we wish to study the difference between the effects of two or more treatments. These experiments may be carried out in the laboratory on animals or human volunteers, in the hospital or community on human patients, or, in the case of preventive trials, on currently healthy people. We call trials of treatments on human subjects *clinical trials*. The general principles of experimental design are the same, although there are special precautions that must be taken when experimenting with human subjects. The experiments whose results most concern clinicians are clinical trials, so the discussion will deal mainly with them.

Suppose we want to know whether a new treatment is more effective than the present standard treatment. We could approach this in a number of ways, as follows.

(a) We could compare the results of the new treatment on new patients with records of previous results using the old treatment. This is seldom convincing, because there may be many differences between the patients who received the old treatment and the patients who will receive the new. As time passes, the general population from which patients come may become healthier, standards of ancillary treatment and nursing care may improve, or the social mix in the catchment area of the hospital may change. The nature of the disease itself may change. All these factors may produce changes in the patients' apparent response to treatment. For example, Christie (1979) showed this by studying the survival of stroke patients in 1978, after the introduction of a C–T head scanner, with that of patients treated in 1974, before the introduction of the scanner. He took the records of a group of patients treated in 1978, who received a C–T scan, and matched each of them with a patient treated in 1974 of the same age, diagnosis and level of consciousness on admission. As the first column of Table 2.1 shows, patients in 1978 clearly

Table 2.1. Analysis of the difference in survival for matched pairs of stroke patients (Christie 1979)

	C-T scan in 1978	No C-T scan in 1978
Pairs with 1978 better than 1974	9 (31%)	34 (38%)
Pairs with same outcome	18 (62%)	38 (43%)
Pairs with 1978 worse than 1974	2 (7%)	17 (19%)

tended to have better survival than similar patients in 1974. The scanned 1978 patient did better than the unscanned 1974 patient in 31 per cent of pairs, whereas the unscanned 1974 patient did better that the scanned 1978 patient in only 7 per cent of pairs. It appears that the scanner was of great benefit. However, he also compared the survival of patients in 1978 who did not receive a C–T scan with matched patients in 1974. As the second column of Table 2.1 shows, these patients too showed a marked improvement in survival from 1974 to 1978. The 1978 patient did better in 38 per cent of pairs and the 1974 patients in only 19 per cent of pairs. We see that there is a general improvement in survival, over a short period of time. If we did not have the data on the unscanned patients from 1978 we might be forgiven for interpreting these data as evidence for the effectiveness of the C–T scanner. Historical controls like this are seldom very convincing. We need to compare the old and new treatments concurrently.

(b) We could ask people to volunteer for the new treatment and give the standard treatment to those who do not volunteer. The difficulty here is that people who volunteer and people who do not volunteer are likely to be different in many ways apart from the treatments we give them. We shall consider an example of the effects of volunteer bias in Section 2.4.

(c) We can allocate patients to the new treatment or the standard treatment and observe the outcome. The way in which patients are allocated to treatments can influence the results enormously. The following example (Hill 1962) illustrates this. Between 1927 and 1944 a series of trials of BCG vaccine were carried out in New York (Levine and Sackett 1944). Children from families where there was a case of tuberculosis were allocated to a vaccination group and given BCG vaccine, or to a control group who were not vaccinated. Between 1927 and 1932 a physician was told to vaccinate half the children, the choice of which children to allocate being left to him. As Table 2.2 shows, there was a clear advantage in survival for the BCG group in this part of the series. However, there was also a clear tendency for the physician to vaccinate the children of more cooperative parents, and to leave those of less cooperative parents as controls. In 1933 this was changed and allocation to treatment or control was done centrally. This was done by assigning alternate children to control and vaccine. The difference in degree of

Table 2.2. Results of studies of BCG vaccine in New York City (Hill 1962)

Period of trial	No. of children	No. of deaths from TB	Death rate (%)	Average no. of visits to clinic during 1st year of follow-up	Proportion of parents with good cooperation as judged by visiting nurses
1927–32 Selection made by physician:					
BCG group	445	3	0.67	3.6	43%
Control group	545	18	3.30	1.7	24%
1933–44 Alternate selection carried out centrally:					
BCG group	566	8	1.41	2.8	40%
Control group	528	8	1.52	2.4	34%

cooperation between the parents of the two groups of children disappeared, and so did the difference in mortality. (Note that these were a special group of children, from families where there was tuberculosis. In large trials using children drawn from the general UK population, BCG was shown to be effective in greatly reducing deaths from tuberculosis (Hart and Sutherland 1977).)

We see that different methods of allocation to treatment produce different results. This is because the method of allocation may not produce groups of subjects which are comparable, i.e. similar in every respect except the treatment. We need a method of allocation to treatments in which the characteristics of subjects will not affect their chance of being put into any particular group. The only generally satisfactory method of doing this which has been found to date is random allocation.

2.2. Random allocation

If we want to decide which of two people receive an advantage, in such a way that each has an equal chance of receiving it, we use a simple, widely accepted method. We toss a coin. This is used to decide the way football matches begin, for example, and all appear to agree that it is fair. So if we want to decide which of two subjects should receive a vaccine, we can toss a coin. Heads — and the first subject receives the vaccine; tails — and the second receives it. If we do this for each pair of subjects we build up two groups which have been assembled without any characteristics of the subjects themselves influencing the allocation in any way. The only differences between the groups will be those due to chance. As we shall see later (Chapters 8–14 inclusive), statistical methods enable us to measure the likely effects of

chance. Any difference between the groups which is larger than this is likely to be due to the treatment, since there will be no other differences between the groups. This method of dividing subjects into groups is called *random allocation* or *randomization*.

Several methods of randomizing have been in use for centuries, though not for clinical trials. Coins have been mentioned; there are also dice, cards, lots and spinning wheels. Some of the theory of probability which we shall use later to compare randomized groups was first developed as an aid to gambling. There are therefore many ways in which we can achieve random allocation. Coins, dice, and cards can all be used, but they have their disadvantages. In a clinical trial, flipping a coin in the presence of a patient may not inspire confidence. It is also too easy for the experimenter to cheat by tossing the coin again if he doesn't like the allocation. (The temptation to do this can be very strong.) It can also be very tedious in a large experiment.

A different, non-physical randomizing method uses random number tables. Table 2.3 provides an example, a table of 1000 random digits. These are more properly called pseudo-random numbers, as they are generated by a mathematical process. They are available in tables (Kendall and Babington

Table 2.3. 1000 random digits

	1–4		5–8		9–12		13–16		17–20		21–24		25–28		29–32		33–36		37–40	
1	36	45	88	31	28	73	59	43	46	32	00	32	67	15	32	49	54	55	75	17
2	90	51	40	66	18	46	95	54	65	89	16	80	95	33	15	88	18	60	56	46
3	98	41	90	22	48	37	80	31	91	39	33	80	40	82	38	26	20	39	71	82
4	55	25	71	27	14	68	64	04	99	24	82	30	73	43	92	68	18	99	47	54
5	02	99	10	75	77	21	88	55	79	97	70	32	59	87	75	35	18	34	62	53
6	79	85	55	66	63	84	08	63	04	00	18	34	53	94	58	01	55	05	90	99
7	33	53	95	28	06	81	34	95	13	93	37	16	95	06	15	91	89	99	37	16
8	74	75	13	13	22	16	37	76	15	57	42	38	96	23	90	24	58	26	71	46
9	06	66	30	43	00	66	32	60	36	60	46	05	17	31	66	80	91	01	62	35
10	92	83	31	60	87	30	76	83	17	85	31	48	13	23	17	32	68	14	84	96
11	61	21	31	49	98	29	77	70	72	11	35	23	69	47	14	27	14	74	52	35
12	27	82	01	01	74	41	38	77	53	68	53	26	55	16	35	66	31	87	82	09
13	61	05	50	10	94	85	86	32	10	72	95	67	88	21	72	09	48	73	03	97
14	11	57	85	67	94	91	49	48	35	49	39	41	80	17	54	45	23	66	82	60
15	15	16	08	90	92	86	13	32	26	01	20	02	72	45	94	74	97	19	99	46
16	22	09	29	66	15	44	76	74	94	92	48	13	75	85	81	28	95	41	36	30
17	69	13	53	55	35	87	43	23	83	32	79	40	92	20	83	76	82	61	24	20
18	08	29	79	37	00	33	35	34	86	55	10	91	18	86	43	50	67	79	33	58
19	37	29	99	85	55	63	32	66	71	98	85	20	31	93	63	91	77	21	99	62
20	65	11	14	04	88	86	28	92	04	03	42	99	87	08	20	55	30	53	82	24
21	66	22	81	58	30	80	21	10	15	53	26	90	33	77	51	19	17	49	27	14
22	37	21	77	13	69	31	20	22	67	13	46	29	75	32	69	79	37	23	32	43
23	51	43	09	72	68	38	05	77	14	62	89	07	37	89	25	30	92	09	06	92
24	31	59	37	83	92	55	15	31	21	24	03	93	35	97	84	61	96	85	45	51
25	79	05	43	69	52	93	00	77	44	82	91	65	11	71	25	37	89	13	63	87

Smith 1971) or can be produced by computer and some types of calculator. We can use tables of random numbers in many ways to achieve random allocation. Two approaches will give the general idea. First, let us randomly allocate 20 subjects to two groups, which we shall label A and B. We first choose a random starting point in the table, using one of the physical methods described above. (I used decimal dice. These are twenty-sided dice, numbered 0 to 9 twice, which fit our number system more conveniently than the traditional cube.) The random starting point was row 22, column 20, and the first twenty digits were 3, 4, 6, 2, 9, 7, 5, 3, 2, 6, 9, 7, 9, 3, 7, 2, 3, 3, 2 and 4. We now allocate subjects corresponding to odd digits to group A and those corresponding to even digits to B. The first digit, 3, is odd, so the first subject goes into group A. The second digit, 4, is even, so the second subject goes into group B, and so on. We get the following allocation:

Subject	Digit	Group	Subject	Digit	Group
1	3	A	11	9	A
2	4	B	12	7	A
3	6	B	13	9	A
4	2	B	14	3	A
5	9	A	15	7	A
6	7	A	16	2	B
7	5	A	17	3	A
8	3	A	18	3	A
9	2	B	19	2	B
10	6	B	20	4	B

The system described above gave us unequal numbers in the two groups, 12 in A and 8 in B. We sometimes want the groups to be of equal size. One way to do this would be to proceed as above until either A or B has 10 subjects in it, all the remaining subjects going into the other groups. This is satisfactory in that each subject has an equal chance of being allocated to A or B, but it has a disadvantage. There is a tendency for the last few subjects all to have the same treatment. This characteristic sometimes worries researchers, who feel that the randomization is not quite right. In statistical terms the possible allocations are not equally likely. If we use this method for the random allocation described above, the tenth subject in group A would be reached at subject 15 and the last five subjects would all be in group B. We can ensure that all randomizations are equally likely by using the table of random numbers in a different way. For example, we can use the table to draw a random sample of 10 from 20, as described in Section 3.4. These would form group A, and the remaining 10, group B.

Table 2.4. Condition of patients on admission to trial of streptomycin (MRC 1948)

General condition	S Group	Control Group	Max. evening temperature in first week (°F)	S Group	Control Group	Sedimentation Rate	S Group	Control Group
Good	8	8	98–98.9	4	4	0–10	0	0
Fair	17	20	99–99.9	13	12	11–20	3	2
Poor	30	24	100–100.9	15	17	21–50	16	20
			101+	24	19	51+	36	29
Total	55	52	Total	55	52	Total	55	51*

* Test not done in one case.

There are many ways of using random-number tables to achieve random allocation. These methods of using random numbers and the generation of the random numbers themselves are simple mathematical operations well suited to the computers which are now readily available to researchers. It is very easy to program a computer to carry out random allocation, and once a program is available it can be used over and over again for further experiments.

Having looked at the theory and techniques of random allocation, we now look at whether it works in practice. One of the first randomized experiments in medicine was the trial carried out by the Medical Research Council (MRC 1948) to test the efficacy of streptomycin for the treatment of pulmonary tuberculosis. In this study the target population was patients with acute progressive bilateral pulmonary tuberculosis, aged 15–30 years. All cases were bacteriologically proved and were considered unsuitable for other treatments then available. The trial took place in three centres and allocation was by a series of random numbers, drawn up for each sex at each centre. The streptomycin group contained 55 patients and the control group 52 cases. The condition of the patients on admission is shown in Table 2.4. The frequency distributions of temperature and sedimentation rate were similar for the two groups; if anything, the treated (S) group were slightly worse. However, this difference is no greater than could have arisen by chance, which, of course, is how it arose. The two groups are almost certain to be slightly different in some characteristics, especially with a fairly small sample, and we can take account of this in the analysis.

After six months, 93 per cent of the S group survived, compared to 73 per cent of the control group. There was a clear advantage to the streptomycin group. The relationship of survival to initial condition is shown in Table 2.5.

Table 2.5. Survival at six months in the MRC streptomycin trial, stratified by initial condition (MRC 1948)

Maximum evening temperature group during first observation week	Outcome	S group	C group
98–98.9 °F	Alive	3	4
	Dead	0	0
99–99.9 °F	Alive	13	11
	Dead	0	1
100–100.9 °F	Alive	15	12
	Dead	0	5
101 °F and above	Alive	20	11
	Dead	4	8

Survival was more likely for patients with lower temperatures, but the difference in survival between the S and C groups is clearly present within the temperature categories.

2.3. Methods of allocation without random numbers

In the second stage of the New York studies of BCG vaccine, the children were allocated to treatment or control alternately. Researchers often ask why this method cannot be used instead of randomization, arguing that the order in which patients arrive is random, so the groups thus formed will be comparable. There are two reasons for not doing this. First, although the patients may appear to be in a random order, there is no guarantee that this is the case. We could never be sure that the groups are comparable. Secondly, this method is very susceptible to mistakes, or even to cheating in the patients' perceived interest. If it is possible to cheat so that a patient who looks particularly in need will receive the allocator's preferred treatment, the temptation to do so can be very great. In the BCG studies some patients were excluded for non-cooperation, but even allowing for this there is considerable imbalance in favour of BCG. There are several examples reported in the literature of alterations to treatment allocations. Holten (1951) reported a trial of anticoagulant therapy for patients with coronary thrombosis. Patients who arrived for treatment on even dates were to be treated with anticoagulant therapy and patients arriving on odd dates were to receive no extra treatment and would form the control group. The author reports that some of the clinicians involved found it 'difficult to remember' the criterion for allocation. As a result, 50 patients admitted on even dates did not receive anticoagulant therapy and 10 who were admitted on odd dates did receive it. The results are shown in Table 2.6. Overall, the treated patients did better than the controls but, curiously, the controls on the even dates (wrongly allocated) did considerably better than control patients on the odd dates (correctly allocated) and even managed to do marginally better than those who received the treatment. In fact the best outcome, treated or not, was for those who were incorrectly allocated. All this must make us wonder

Table 2.6. Outcome of a quasi-random clinical trial with errors in allocation (Holten 1951)

	Even dates		Odd dates	
Outcome	Treated	Control	Treated	Control
Survived	125	39	10	125
Died	39 (25%)	11 (22%)	0 (0%)	81 (36%)
Total	164	50	10	206

whether any groups of patients in this trial were correctly allocated, and about the extent to which any confidence can be placed in the results.

Other methods of allocation set out to be random but can fall into this sort of difficulty. For example, we could use physical mixing to achieve randomization. This is quite difficult to do. As an experiment, take a deck of cards and order them in suits from ace of clubs to king of spades. Now shuffle them in the usual way and examine them. You will probably see many runs of several cards which remain together in order. Cards must be shuffled very thoroughly indeed before the ordering ceases to be apparent. The physical randomization method can be applied to an experiment by marking equal numbers of slips of paper with the names of the treatments, sealing them into envelopes and shuffling them. The treatment for a subject is decided by withdrawing an envelope. This method was used in another study of anticoagulant therapy by Carleton *et al.* (1960). These authors reported that in the latter stages of the trial some of the clinicians involved had attempted to read the contents of the envelopes by holding them up to the light, in order to allocate patients to their own preferred treatment.

Interfering with the randomization can actually be built into the allocation procedure, with equally disastrous results. In the Lanarkshire Milk Experiment, discussed by Student (1931), 10 000 school children received three-quarters of a pint of milk per day and 10 000 children acted as controls. The children were weighed and measured at the beginning and end of the six-month experiment. The object was to see whether the milk improved the growth of children. The allocation to the 'milk' or control group was done as follows:

> The teachers selected the two classes of pupils, those getting milk and those acting as controls, in two different ways. In certain cases they selected them by ballot and in others on an alphabetical system. In any particular school where there was any group to which these methods had given an undue proportion of well-fed or ill-nourished children, others were substituted to obtain a more level selection.

The result of this was that the control group had a markedly greater average height and weight than the milk group. Student interpreted this as follows:

> Presumably this discrimination in height and weight was not made deliberately, but it would seem probable that the teachers, swayed by the very human feeling that the poorer children needed the milk more than the comparatively well-to-do, must have unconsciously made too large a substitution for the ill-nourished among the [milk group] and too few among the controls and that this unconscious selection affected secondarily, both measurements.

Whether the bias was conscious or not, it spoiled the experiment, despite being from the best possible motives.

There is one non-random method which can be used successfully in clinical trials. It is called minimization. In this method, new subjects are allocated to

treatments so as to make the treatment groups as similar as possible in terms of the important prognostic factors. It is beyond the scope of this book, but see Pocock (1983) for a description.

2.4. Volunteer bias

One of the most interesting trials ever done was the field trial of Salk poliomyelitis vaccine carried out in 1954 in the USA (Meier 1977). This was carried out using two different designs simultaneously, due to a dispute about the correct method. In some districts, second-grade schoolchildren were invited to participate in the trial, and randomly allocated to receive vaccine or an inert saline injection. In other districts, all second-grade children were offered vaccination and the first- and third-grade left unvaccinated as controls. The argument against this 'observed-control' approach was that the groups may not be comparable, whereas the argument against the randomized-control method was that the saline injection could provoke paralysis in infected children. The results are shown in Table 2.7. In the randomized-control areas the vaccinated group clearly experienced far less polio than the control group. Since these were randomly allocated, the only difference between them should be the treatment, which is clearly preferable to saline. However, the control group also had more polio than those who had refused to participate in the trial. The difference between the control and the not-inoculated group is both in treatment (saline injection) and in selection; they are self-selected as volunteers and refusers, respectively. The observed-control areas enable us to distinguish between these two factors. The polio rates in the vaccinated children are very similar in both parts of the study, as are the rates in the not-inoculated second-grade children. It is the two control groups which differ. These were selected in different ways: in the randomized-control areas they were volunteers, whereas in the

Table 2.7. Result of the field trial of Salk poliomyelitis vaccine (Meier 1977)

		Paralytic polio	
Study group	Number in group	Number of cases	Rate per 100 000
Randomized control			
Vaccinated	200 745	33	16
Control	201 229	115	57
Not inoculated	338 778	121	36
Observed control			
Vaccinated second-grade	221 998	38	17
Control first- and third-grade	725 173	330	46
Unvaccinated second-grade	123 605	43	35

observed-control areas they are everybody eligible, both potential volunteers and potential refusers. Now suppose that the vaccine were saline instead, and that the randomized, vaccinated children had the same polio experience as those receiving saline. We would expect the following number of cases:

$$200\,745 \times \frac{57}{100\,000} = 114$$

The total number of cases in the randomized areas would be 114 + 115 + 121 = 350 and the rate per 100 000 would be 47. This compares very closely with the rate of 46 in the observed-control first- and third-grade group. Thus it seems that the principal difference between the saline-control group of volunteers and the not-inoculated group of refusers is selection, not treatment.

There is a simple explanation for this. Polio is a viral disease transmitted by the faecal – oral route. Before the development of vaccine almost everyone in the population was exposed to it at some time, usually in childhood. In the majority of cases, paralysis does not result and immunity is conferred without the child being aware of having been exposed to polio. In a small minority of cases, about one in 200, paralysis and occasionally death occurs and a diagnosis of polio is made. The older the exposed individual is, the greater the chance of paralysis developing. Hence, children who are protected from infection by high standards of hygiene are likely to be older when they are first exposed to polio than those children from homes with low standards of hygiene, and thus more likely to develop the clinical disease. There are many factors which may influence parents in their decision as to whether to volunteer or refuse their child for a vaccine trial. These may include education, personal experience, current illness, and others, but they certainly include interest in health and hygiene. Thus in this trial the high-risk children tended to be volunteered and the low-risk children tended to be refused. The high-risk, volunteer, control children had 68 per cent more cases of polio than the low-risk refusers.

In many diseases, the effect of volunteer bias is opposite to this. Poor conditions are related both to refusal to participate and to high risk, whereas volunteers tend to be low risk. The effect of volunteer bias is then to produce an apparent difference in favour of the treatment. We can see that comparisons between volunteers and other groups can never be reliable indicators of treatment effects.

In the observed control areas, quite apart from the non-random age difference, the vaccinated and control groups are not comparable. However, it is possible to make a reasonable comparison in this study by comparing all second-grade children, both vaccinated and refused, to the control group. The rate in the second-grade children is 23 per 100 000, which is less than the rate of 46 in the control group, demonstrating the effectiveness of the

vaccine. The 'treatment' which we are evaluating is not vaccination itself, but a policy of offering vaccination and treating those who accept. A similar problem can arise in a randomized trial, for example in evaluating the effectiveness of health check-ups (South-east London Screening Study Group 1977). Subjects were randomized to a screening group or to a control group. The screening group were invited to attend for an examination; some accepted and were screened and some refused. When comparing the results in terms of subsequent mortality, it was essential to compare the controls to the screening groups containing both screened and refusers. For example, the refusers may have included people who were already too ill to come for screening. The important point is this. The random allocation procedure produces comparable groups and it is these we must compare, whatever selection may be made within them.

2.5. Cross-over designs

Sometimes it is possible to use a subject as her or his own control. For example, when comparing analgesics in the treatment of arthritis, patients may receive in succession a new drug and a control treatment. The response to the two treatments can then be compared for each patient. These designs have the advantage of removing variability between subjects. We can carry out a trial with far fewer subjects than would be needed for a two-group trial.

Although all subjects receive all treatments, these trials must still be randomized. In the simplest case of treatment and control, patients may be given two different regimes: control followed by treatment and treatment followed by control. These may not give the same results, e.g. there may be a long-term carry-over effect which makes treatment followed by control show less of a difference than control followed by treatment. Subjects are, therefore, assigned to a given order at random. It is possible in the analysis of cross-over studies to estimate the size of any carry-over effects which may be present. If there are large carry-over effects the second treatment in the sequence may not give a reliable comparison. In this case the results of the first treatment can be analysed as a two-group trial, losing the advantage of the cross-over.

As an example of the advantages of a cross-over trial, consider a trial of pronethalol in the treatment of angina pectoris (Pritchard *et al.* 1963). Angina pectoris is a chronic disease characterized by attacks of acute pain. Patients in this trial received either pronethalol or an inert control treatment (or placebo) in four periods of two weeks — two periods on the drug and two on the control treatment. These periods were in random order. The outcome measure was the number of attacks of angina experienced. These were recorded by the patient in a diary. Twelve patients took part in the trial. The results are shown in Table 2.8. The advantage in favour of pronethalol is

Table 2.8. Results of a trial of pronethalol for the treatment of angina pectoris (Pritchard *et al.* 1963)

Patient number	Number of attacks while on		Difference Placebo–Pronethalol
	Placebo	Pronethalol	
1	71	29	42
2	323	348	−23
3	8	1	7
4	14	7	7
5	23	16	7
6	34	25	9
7	79	65	14
8	60	41	19
9	2	0	2
10	3	0	3
11	17	15	2
12	7	2	5

shown by 11 of the 12 patients reporting fewer attacks of pain while on pronethalol than while on the control treatment. If we had obtained the same data from two separate groups of patients instead of the same group under two conditions, it would be far from clear that pronethalol is superior because of the huge variation between subjects. Using a two-group design, we would need a much larger sample of patients to demonstrate the efficacy of the treatment.

Cross-over designs can be useful for laboratory experiments on animals or human volunteers. They can only be used in clinical trials where the treatment will not affect the course of the disease and where the patient's condition would not change appreciably over the course of the trial. A cross-over trial could be used to compare different treatments for the control of arthritis or asthma, for example, but not to compare different regimes for the management of myocardial infarction. However, a cross-over trial cannot be used to demonstrate the long-term action of a treatment, as the nature of the design means that the treatment period must be limited. As most treatments of chronic disease must be used by the patient for a long time, often several years, a two-sample trial of long duration is usually required to investigate fully the effectiveness of the treatment. Pronethalol, for example, was later found to have some unacceptable side-effects in long-term use.

2.6. Selection of experimental subjects

We have discussed the allocation of subjects to treatments at some length, but we have not considered where they come from. The way in which subjects are

selected for an experiment may have an effect on its outcome. In practice, we are usually limited to subjects which are easily available to us. For example, in an animal experiment we must take the latest batch from the animal house. In a clinical trial of the treatment of myocardial infarction, we must be content with patients who are brought into our hospital. In experiments on human volunteers we sometimes have to use the researchers themselves.

As we shall see more fully in Chapter 3, this has important consequences for the interpretation of results. In trials of myocardial infarction, for example, we would not wish to conclude that, say, the survival rate with a new treatment in a trial in London would be the same as in a trial in Edinburgh. The patients may have a different history of diet, for example, and this may have a considerable effect on the state of their arteries and hence on their prognosis. Indeed, it would be very rash to suppose that we would get the same survival rate in a hospital a mile down the road. What we rely on is the comparison between randomized groups from the same population of subjects, and hope that if a treatment reduces mortality in London it will also do so in Edinburgh. This may be a reasonable supposition, but it cannot be proved on statistical grounds alone. Sometimes in extreme cases it turns out not to be true. The BCG vaccine has been shown by large, well-conducted randomized trials to be effective in reducing the incidence of tuberculosis in children in the UK. However, in India it appears to be less effective (Lancet 1980). This may be because the amount of exposure to tuberculosis is so different in the two populations.

Given that we can only use the experimental subjects available to us, there are some principles which we use to guide our selection from them. As we shall see later, the lower the variability between the subjects in an experiment is, the better chance we have of detecting a treatment difference if it exists. This means that uniformity is desirable in our subjects. In an animal experiment this can be achieved by using animals of the same strain raised under controlled conditions. In a clinical trial we usually do this by restricting our attention to patients of a defined age group and severity of disease. The Salk vaccine trial only used children in one school year. In the streptomycin trial the subjects were restricted to patients with acute bilateral pulmonary tuberculosis, bacteriologically proved, aged between 15 and 30 years, unsuitable for other current therapy. Even with this narrow definition there was still considerable variation in the patients, as Tables 2.4 and 2.5 show.

In a clinical trial it is also important to make sure that everyone has the disease we wish to treat. Patients with a different disease are not only potentially being wrongly treated themselves, but may make the results very difficult to interpret.

Restricting attention to a particular subset of patients, useful though it may be, can lead to difficulties. For example, a treatment shown to be

effective and safe in young people may not necessarily be so in the elderly. Trials have to be carried out on the sort of patients it is proposed to treat.

2.7. Response bias and placebos

The knowledge that she or he is being treated may alter a patient's response to treatment. This is called the *placebo effect*. A *placebo* is a pharmacologically inactive substance given as if it were an active treatment. The placebo effect may take many forms, from a desire to please the doctor to measurable biochemical changes in the brain. Mind and body are intimately connected, and unless the psychological effect is actually part of the treatment we usually try to eliminate such factors from treatment comparisons. This is particularly important when we are dealing with subjective quantities such as assessment of pain or well-being.

A fascinating example of the power of the placebo effect is given by Huskisson (1974). Three active analgesics, aspirin, Codis and Distalgesic, were compared with an inert placebo. Twenty-two patients each received the four treatments in a cross-over design. The patients reported pain relief on a four-point scale, from 0 = no relief to 3 = complete relief. The changes in pain relief are shown in Fig. 2.1. All the treatments produced some pain relief, maximum relief being experienced after about two hours. The three active treatments were all superior to placebo, but not by very much. The remarkable aspect of the trial was that the four drug treatments were given in the form of tablets identical in shape and size, but each drug was given in four different colours. This was done so that patients could distinguish the drugs received to say which they preferred. Each patient received four different colours, one for each drug, and the colour combinations were allocated randomly. Thus some patients received red placebos, some blue and so on. The comparison of pain relief associated with colour of placebo are shown in Fig. 2.1. Red placebo were markedly more effective than other colours, and

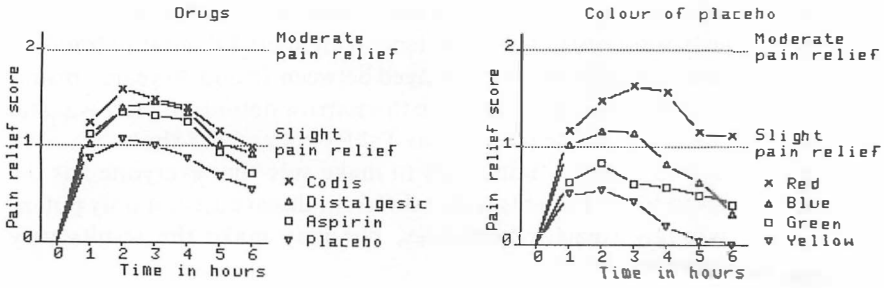


Fig. 2.1 The effect of placebo compared to three drugs and the effects of placebos in four different colours for pain relief in arthritis (Huskisson 1974).

were just as effective as the active drugs! In this study not only is the effect of a pharmacologically inert placebo in producing reported pain relief demonstrated, but so is the wide variability and unpredictability of this response. We must clearly take account of this in trial design. Incidentally, we should not conclude that red placebos always work best. There is, for example, some evidence that patients being treated for anxiety prefer tablets to be in a soothing green, and that depressive symptoms respond best to a lively yellow (Schapira *et al.* 1970).

In any study of humans it is desirable that the subjects should not be able to tell which treatment is which. In a study to compare two treatments this should be done by making the treatments as similar as possible. Where there is no treatment an inactive placebo should be used.

Placebos are not always possible or ethical. In the MRC trial of streptomycin, where the treatment involved several injections a day for several months, it was not regarded as ethical to do the same with an inert saline solution, and no placebo was given. In the Salk vaccine trial, the inert saline injections were placebos. It could be argued that paralytic polio is not likely to respond to psychological influences, but how could we be really sure of this? Further, the certain knowledge that a child had been vaccinated may have altered the risk of exposure to infection as parents allowed the child to go swimming, for example. Finally, the use of a placebo may also reduce the risk of assessment bias as we shall see in Section 2.8 below.

2.8. Assessment bias and double-blind studies

The response of subjects is not the only thing affected by knowledge of the treatment. The assessment by the researcher of the response to treatment may also be influenced by the knowledge of the treatment.

Some outcome measures do not allow for much bias on the part of the assessor. For example, if the outcome is survival or death, there is little possibility that unconscious bias may affect the observation. However, if we are interested in an overall clinical impression of the patient's progress, or in changes in an X-ray picture, the measurement may be influenced by our desire (or otherwise) that the treatment should succeed. It is not enough to be aware of this danger and allow for it. We soon have the similar problem of 'bending over backwards to be fair'. Even such an apparently objective measure as blood pressure can be influenced by the expectations of the experimenter, and special measuring equipment has been devised to avoid this (Rose *et al.* 1964).

We can avoid the possibility of such bias by using *blind assessment*, that is, the assessor does not know which treatment the subject is receiving. If a clinical trial cannot be conducted in such a way that the clinician in charge

Table 2.9. Assessment of radiological appearance at six months as compared with appearance on admission (MRC 1948)

Radiological assessment	S Group		C Group	
Considerable improvement	28	51%	4	8%
Moderate or slight improvement	10	18%	13	25%
No material change	2	4%	3	6%
Moderate or slight deterioration	5	9%	12	23%
Considerable deterioration	6	11%	6	11%
Deaths	4	7%	14	27%
Total	55	100%	52	100%

does not know the treatment, blind assessment can still be carried out by an external assessor. When the subject does not know the treatment and blind assessment is used, the trial is said to be *double blind*. Where the subject can distinguish between treatments but the assessor cannot, the trial is *single blind*.

We can see that placebos may be just as useful for avoiding assessment bias as for avoiding response bias. The subject is unable to tip the assessor off as to treatment, and there is likely to be less material evidence to indicate to an assessor what it is. In the anticoagulant study by Carleton *et al.* (1960) described above, the treatment was supplied through an intravenous drip. Control patients had a dummy drip set up, with a tube taped to the arm but no needle inserted, primarily to avoid assessment bias. In the Salk trial, the injections were coded and the code for a case was only broken after the decision had been made as to whether the child had polio and, if so, of what severity.

In the streptomycin trial, one of the outcome measures was radiological change. X-Ray plates were numbered and then assessed by two radiologists and a clinician, none of whom knew to which patient or treatment the plate belonged. They did the assessment independently, and only discussed a plate if they had not all come to the same conclusion. Only when a final decision had been arrived at was the link between plate and patient made. The results are shown in Table 2.9. The clear advantage of streptomycin is shown in the considerable improvement of over half the S group, compared to only 8 per cent of the controls.

2.9. Laboratory experiments

So far we have looked at clinical trials, but exactly the same principles apply to laboratory research on animals. It may well be that in this area the principles of randomization are not so well understood and even more critical

attention is needed from the reader of research reports. One reason for this may be that great effort has been put into producing genetically similar animals, raised in conditions as close to uniform as is practicable. The researcher using such animals as subjects may feel that the resulting animals show so little biological variability that any natural differences between them will be dwarfed by the treatment effects. This is not necessarily so, as the following examples illustrate.

A colleague was looking at the effect of tumour growth on macrophage counts in rats. The only significant difference was between the initial values in tumour-induced and non-induced rats, that is, before the tumour-inducing treatment was given. There was a simple explanation for this surprising result. The original design had been to give the tumour-inducing treatment to each of a group of rats. Some would develop tumours and others would not, and then the macrophage counts would be compared between the two groups thus defined. In the event, all the rats developed tumours. In an attempt to salvage the experiment my colleague obtained a second batch of animals, which he did not treat, to act as controls. The difference between the treated and untreated animals was thus due to differences in parentage or environment, not to treatment.

That problem arose by changing the design during the course of the experiment. Problems can arise from ignoring randomization in the design of a comparative experiment. Another colleague wanted to know whether a treatment would affect weight gain in mice. Mice were taken from a cage one by one and the treatment given, until half the animals had been treated. The treated animals were put into smaller cages, five to a cage, which were placed together in a constant-environment chamber. The control mice were in cages, also placed together in the constant-environment chamber. When the data were analysed, it was discovered that the mean initial weight was greater in the treated animals than in the control group. In a weight-gain experiment this could be quite important! It may have been that when the experimenter was picking up the animals to apply the treatment, she found the larger animals easier to pick up. What that experimenter should have done was to place the mice in the boxes, give each box a place in the constant-environment chamber, then allocate the boxes to treatment or control at random. We would then have two groups which were comparable in every respect except treatment, both in initial values and in any environmental differences which may exist in the constant-environment chamber.

These examples are given to show that even when the experimental material is as uniform as laboratory animals, biological variability is still present. Randomization was devised to cope with this and it is the most effective method we have.

2.10. Experimental units

In the weight-gain experiment described above, each box of mice contained five animals. These animals were not independent of one another, but interacted. In a box the other four animals formed part of the environment of the fifth, and might influence its growth. The box of five mice is called an *experimental unit*. An experimental unit is the smallest group of subjects in an experiment whose response cannot be affected by other subjects. This is important, as we need to know the amount of natural variation which exists between experimental units before we can decide whether the treatment effect is distinguishable from this natural variation. The accuracy with which we can estimate this depends on the number of experimental units (see Chapter 10).

The most extreme case arises when there is only one experimental unit per treatment. For example, consider a health education experiment involving two schools. In one school a special health education programme was mounted, aimed to discourage children from smoking. Both before and afterwards, the children in each school completed questionnaires about cigarette smoking. After the campaign, there were fewer smokers reported in the school where the health education had taken place than in the other, whereas before, the proportions of smokers had been similar. In this example the school is the experimental unit. The children may influence one another in their cigarette smoking habits and in their reactions to the health education programme. What happens in a school will be affected by changes in staff and pupils as well as more widespread factors such as advertising. There is no reason, therefore, to suppose that two schools should have the same proportion of smokers among their pupils, or that two schools which do have equal proportions of smokers will remain so. In this study the researchers found that smoking rates in the two schools were closer before the treatment than after. However, this may be due to the treatment or to other differences between and changes in the schools. We cannot tell from the data. This experiment would be much more convincing if we had several schools and randomly allocated them either to receive the health education programme or be control. We would then look for a consistent difference between changes in cigarette smoking in the treated and in the control schools.

2.11. Further points about trial design

There are many aspects of experimental design which we have not yet discussed. These include experiments to compare several factors at once. For example, we might wish to study the effect of a drug at different doses in the presence or absence of a second drug, with the subject standing or supine. This is usually designed as a factorial experiment, where each subject receives

every possible combination of treatments. These designs are unusual in clinical research but are sometimes used in laboratory work. They are described in more advanced texts (Armitage 1971; Snedecor and Cochran 1980).

The trials described above all had a fixed sample size, decided at the start of the experiment. Because in medicine it is desirable to expose as few patients as possible to potentially hazardous treatments, sequential designs have been developed in which the data are analysed as they are collected. As soon as the difference between treatments is large enough to be convincing, the trial is stopped (Armitage 1975).

We have yet to discuss the choice of sample size. To do this we need to see how data are analysed. We shall return to this in Chapter 9.

Finally, we have yet to mention the ethics of clinical trials. The objection to randomized experimentation may be made that we are withholding a potentially beneficial treatment from patients. However, any biologically active treatment is potentially harmful, and we are surely not justified in giving potentially harmful treatments to patients before the benefits have been demonstrated conclusively. Without properly conducted and controlled clinical trials to support it, each administration of a treatment to a patient becomes an uncontrolled experiment, whose outcome, good or bad, cannot be predicted.

For accounts of the theory and practice of clinical trials, see Pocock (1983) and Johnson and Johnson (1977).

Exercise 2M

(Each branch is either true or false.)

1. In an experiment to compare two treatments, subjects are allocated at random so that:

(a) the sample may be referred to a known population;

(b) the experimenter will not know which treatment the subjects receive;

(c) the subjects will get the treatment best suited to them;

(d) the two groups will be as similar as possible, apart from treatment;

(e) treatments may be assigned according to the characteristics of the subject.

2. In a double-blind clinical trial:

(a) the patients do not know which treatment they receive;

(b) each patient receives a placebo;

(c) the patients do not know that they are in a trial;

(d) each patient receives both treatments;

(e) the clinician making assessment does not know which treatment the patient receives.

3. In a trial of a new vaccine, children were assigned at random to a 'vaccine' and a 'control' group. The 'vaccine' group were offered vaccination, which two-thirds accepted.

(a) The group which should be compared to the controls is all children who accepted vaccination.

(b) Those refusing vaccination should be included in the control group.

(c) The trial is double blind.

(d) Those refusing vaccination should be excluded.

(e) The trial is useless because not all the treated group were vaccinated.

4. Cross-over designs for clinical trials:

(a) may be used to compare several treatments;

(b) involve no randomization;

(c) require fewer patients than do designs comparing independent groups;

(d) are useful for comparing treatments intended to alleviate chronic symptoms;

(e) use the patient as his own control.

5. Placebos are useful in clinical trials:

(a) when two apparently similar active treatments are to be compared;

(b) to guarantee comparability in non-randomized trials;

(c) because the fact of being treated may itself produce a response;

(d) because they may help to conceal the subject's treatment from assessors;

(e) when an active treatment is to be compared to no treatment.

Exercise 2E

The following is a paraphrase of a research report which appeared in a major journal. It has been extensively rewritten for the purpose of the exercise and

so does not necessarily represent the views of the original authors. Read the report and then answer the questions.

A study of infants at risk of sudden death

Introduction

Babies sometimes die unexpectedly, for no apparent reason. Many theories have been advanced to explain this, from babies sleeping so deeply that they stop breathing to deficiencies in immune response or murder by parents. None of these has won universal acceptance and indeed there may be no one reason for these deaths. These deaths are often called cot deaths or sudden infant death syndrome. This study aims to identify high-risk children with a view to preventing unexpected deaths.

In a previous paper, the authors have reported an investigation of factors related to unexpected death in babies. This was done by comparing the routine obstetric and perinatal records of babies who died unexpectedly with those of a control group who did not die. The deaths were all babies who died in a defined period of time in an English town, and for each death the control was the next birth to be registered. Using these data a scoring system was devised using a statistical method called discriminant analysis, to predict which babies would be deaths and which would be controls. It was found that a combination of eight variables best distinguished between deaths and controls. Deaths were identified by a combination of low maternal age, mother's blood group not A, urinary infection or polyhydraminos during pregnancy, long second-stage labour, high birth order (i.e. child has brothers and sisters), prematurity, and intention not to breast feed. It is not suggested that these are all directly causal factors, but they are a combination which should enable us to say whether a future child has a high risk of unexpected death. At birth, the data could be used to calculate a score which would be related to the risk of death and used to separate children into a high-risk and a low-risk group. Resources could then be concentrated on the high-risk children to try to prevent unexpected deaths. The cut-off point was chosen so that about 15 per cent of children would be in the high-risk group.

The purpose of the study reported here was two-fold:

(a) to test the ability of the scoring system to select a high-risk group;
(b) to see whether deaths can be reduced by increased surveillance of high-risk infants.

Method

Each week-day during 1973 and 1974 the score of each new baby born in the study town was calculated and infants designated as high- or low-risk. After excluding infants with gross congenital anomalies, some of whom might be

expected to die, the high-risk infants were allocated randomly to two groups: observation and control. Because of holidays, on eight days in Spring 1973 all 16 high-risk babies were allocated to control. To avoid overloading the surveillance team, between 7 July and 14 September 1974, 40 per cent of high-risk children were allocated to the observation group and 60 per cent to the control group. For the rest of the study period 50 per cent of high-risk children were allocated to observation and to control.

The observation group were invited to participate in the study. The surveillance consisted of a clinical observation within 48 hours of birth, a second at five weeks, and ten visits to the home in the first 20 weeks of life by specially appointed health visitors. Some parents refused to allow all or some of this observation.

Results

Table 2E1 shows the mortality in the different groups in the study. Unexpected death rate in the high-risk control group is 6.3 times that of the low-risk group. This difference is very unlikely to be due to chance and shows that the scoring system is highly effective. The rate in the observed group is only twice that of the low-risk group and is one-third of that in the high-risk controls.

Table 2E1. Unexpected deaths between one week and 52 weeks of life

		Unexpected deaths	
	Number in group	Number	Rate per 1000
Low risk	9630	15	1.6
High risk	1769	14	8.0
Control	922	9	9.8
Observed	627	2	3.2
Refused	210	3	14.3
Excluded	35		

If we group those who refused surveillance with the high-risk controls to give us the largest unobserved group the combined death rate is 10.6, which is 3.3 times that in the observed children. If there were really no difference in the whole population we can calculate that we would get a difference as big as this in 7.6 per cent of samples. (We usually take 5 per cent or less as an index of reasonable evidence for a difference. See Chapter 9, significance tests.)

Discussion

The possibility that surveillance reduces mortality requires comment. The primary objective was to observe the high-risk group, but serious medical

conditions could not be ignored. The health visitors were able to take infants directly to hospital if necessary and when the five-week clinical examination showed that feeds were over-concentrated, parents were taught to make up feeds correctly.

It is possible to discuss the reduction in death rate in the observed group on the grounds that it could be due to chance, but the alternative hypothesis that the death rate is reduced is more likely.

The difference between observed and control groups was also less than might be expected because all children benefited in some way from the study. For example, the campaign to explain the dangers of over-strength feeding was not restricted to the observed group. The benefit is reflected in a reduction in the death rate between ages 1 week and 52 weeks. For 1968–72 the average death rate for the study town was 8.2 per 1000, compared to 7.6 per 1000 in England and Wales. In 1973–74 it was 5.2 per 1000 compared to 7.4 for England and Wales.

Since the end of the study all children in the study town have been scored and community health authorities notified of high-risk infants. A system has been introduced whereby all children are examined in the home by a health visitor at four weeks. The mortality rate for 1975–76 was 4.7 per thousand, compared to 6.3 per 1000 for England and Wales.

It is concluded that infants with a high-risk of unexpected death can be identified at birth. The data suggest that deaths can be reduced by increased surveillance of high-risk infants.

Questions about this report:

1. Does the scoring system identify high-risk children?
2. Allocation to high- and low-risk was not random. Does this matter?
3. Would the scoring system work equally well everywhere?
4. Are the group allocated to observation and the control group comparable, apart from treatment?
5. Are the group actually observed and the control group comparable apart from treatment?
6. What was the reason for combining the refused and control groups? Was this a reasonable thing to do?
7. What comparison between groups best helps us examine the possible effects of observation?
8. What conclusion can be drawn from this study?
9. Do you think a national system of risk scoring and surveillance should be instituted?

3. Sampling and observational studies

3.1. Observational studies

In this chapter we shall be concerned with observational studies. Instead of changing something and observing the result, as in an experiment or clinical trial, we observe the existing situation and try to understand what is happening. Studying people in the wild state, as it were, can be extremely difficult and it is often impossible to draw unequivocal conclusions. We shall start by considering how to get descriptive information about populations in which we are interested. We shall go on to the problem of using such information to study disease processes and the possible causes of disease.

3.2. Censuses

One simple question we can ask about any group of interest is how many members it has. For example, for many purposes we need to know how many people live in a country and how many of them are in various age and sex categories. We need this information in order to monitor the changing pattern of disease and to plan medical services. We can obtain it by a *census*. In a census, the whole of a defined population is counted. In the United Kingdom, as in many developed countries, a population census is held every ten years. This is done by dividing the entire country into small areas called enumeration districts, usually containing between 100 and 150 households. It is the responsibility of an enumerator to identify every household in the district and ensure that a census form is completed, listing all members of the household and a few simple pieces of information. Even though completion of the census form is compelled by law, and enormous effort goes into ensuring that every household is included, there are undoubtedly some who are missed and the final data, though extremely useful, are not totally reliable.

The census is one of the oldest forms of statistical enquiry. One is mentioned in the *Old Testament* (1 Chronicles:21), where we read that Satan incited King David to count the people. This so angered the Lord and He sent a pestilence throughout Israel as a punishment and warning to presumptuous

statisticians. Censuses remain unpopular to this day, principally because of fears that census data may be used for other purposes. For this reason, great efforts are made to ensure the confidentiality of census data.

The medical profession take part in a massive, continuing census of deaths, by registering for each death which occurs not only the name of the deceased and cause of death, but also details of age, sex, place of residence and occupation. We shall have more to say about this in Chapter 16.

Census methods are not restricted to national populations. They can be used for more specific administrative purposes too. For example, we might want to know how many patients are in a particular hospital at a particular time, how many of them are in different diagnostic groups, in different age/sex groups, and so on. We can then use this information together with estimates of the death and discharge rates to estimate how many beds these patients will occupy at various times in the future (Bewley *et al.* 1975, 1980).

3.3. Sampling

A census of a single hospital can only give us reliable information about that hospital. We cannot easily generalize our results to hospitals in general. If we want to obtain information about the hospitals of the United Kingdom, two courses are open to us: we can study every single one, or we can take a representative sample of hospitals and use that to draw conclusions about hospitals as a whole.

Most statistical work is concerned with using samples to draw conclusions about some larger population. In the clinical trials described in Chapter 2, the patients act as a sample from a larger population consisting of all similar patients and we do the trial to find out what would happen to this larger group were we to give them a new treatment.

The word 'population' is used in common speech to mean 'all the people living in an area', frequently of a country. In statistics, we define the term more widely. A *population* is any collection of individuals in which we may be interested, where these individuals may be anything, and the number of individuals may be finite or infinite. Thus, if we are interested in some characteristics of the British people, the population is 'all people in Britain'. If we are interested in the treatment of diabetes the population is 'all diabetics'. If we are interested in the blood pressure of a particular patient, the population is 'all possible measurements of blood pressure in that patient'. If we are interested in the toss of two coins, the population is 'all possible tosses of two coins'. The first two examples are finite populations and could in theory if not practice be completely examined; the second two are infinite populations and could not. We could only ever look at a *sample*, which we will define as being a group of individuals taken from a larger

population and used to find out something about that population.

How should we choose a sample from a population? The problem of getting a representative sample is similar to that of getting comparable groups of patients discussed in Sections 2.1, 2.2, and 2.3. We want our sample to be representative, in some sense, of the population. We want it to have all the characteristics in terms of the proportions of individuals with particular qualities as has the whole population. In a sample from a human population, for example, we want the sample to have about the same proportion of men and women as in the population, the same proportions in different age groups, in occupational groups, with different diseases, and so on. In addition, if we use a sample to estimate the proportion of people with a disease, we want to know how reliable this estimate is, how far from the proportion in the whole population the estimate is likely to be.

It is not sufficient to choose the most convenient group. For example, if we wished to predict the results of an election, we would not take as our sample people waiting in bus queues. These may be easy to interview, at least until the bus comes, but the sample would be heavily biased towards those who cannot afford cars and thus towards lower-income groups. In the same way, if we wanted a sample of medical students we would not take the front two rows of the lecture theatre. They may be unrepresentative in having an unusually high thirst for knowledge, or poor eyesight.

How can we choose a sample which does not have a built-in bias? We might divide our population into groups, depending on how we think various characteristics will affect the result. To ask about an election, for example, we might group the population according to age, sex and social class. We then choose a number of people in each group by knocking on doors until the quota is made up, and interview them. Then, knowing the distributions of these categories in the population (from census data, etc.) we can get a far better picture of the views of the population. This is called *quota sampling*. In the same way we could try to choose a sample of rats by choosing given numbers of each weight, age, sex, etc.

There are three main difficulties in this approach:

1. It is rarely possible to think of all the relevant classifications.
2. It is still difficult to avoid bias within the classifications, by picking interviewees who look friendly, or rats which are easy to catch.
3. We can only get an idea of the reliability of findings by repeatedly doing the same type of survey, and of the representativeness of the sample by knowing the true population values (which we can actually do in the case of elections), or by comparing the results with a sample which does not have these drawbacks.

This method can be quite effective when similar surveys are made

repeatedly as in opinion polls or market research. It is less useful for medical problems, where we are continually asking new questions. We need a method where bias is avoided and where we can estimate the representativeness of the sample from the sample itself. As in Section 2.2, we use a random method: random sampling.

3.4. Random sampling

The problem of obtaining a sample which is representative of a larger population is very similar to that of allocating patients into two comparable groups. We want a way of choosing members of the sample which does not depend on their own characteristics. The only way to be sure of this is to select them at random, so that whether or not each member of the population is chosen for the sample is purely a matter of chance.

For example, to take a random sample of five students from a class of eighty, we could write all the names on pieces of paper, mix them thoroughly in a hat or other suitable container, and draw out five. All students have the same probability, 5/80, of being chosen, and so we have a random sample. All samples of five students are equally likely, too, because each is chosen quite independently of the others. This method is called *simple random sampling*.

As we have seen in Section 2.2, physical methods of randomizing are often not very suitable for statistical work. We usually use tables of random digits, such as Table 2.3, or random numbers generated by a computer program. We could use Table 2.3 to draw our sample of five from 80 students in several ways. For example, we could list the students, numbered from 1 to 80. This list from which the sample is to be drawn is called the *sampling frame*. We choose a starting point in the random-number table (Table 2.3), say row 20, column 5. This gives us the following pairs of digits:

14 04 88 86 28 92 04 03 42 99 87 08

We could use pairs of digits directly as subject numbers. We choose subjects numbered 14 and 4. There is no subject 88 or 86, so the next chosen is number 28. There is no 92, so the next is 4. We already have this subject in the sample, so we carry on to the next pair of digits, 03. The final member of the sample is number 42. Our sample of five students is thus numbers 3, 4, 14, 28 and 42. The choice of adjacent numbers 3 and 4, is something that often occurs in random number systems. They often appear to us to have pattern, perhaps because the human mind is always looking for it. On the other hand, if we try to make the sample 'more random' by replacing either 3 or 4 by a subject near the end of the list, we are imposing a pattern of uniformity on the sample and in fact destroying its randomness.

This method of using the table is fine for drawing a small sample, though it

can be tedious for drawing large samples, because of the need to check for duplicates. There are many other ways of doing it. For example, we can drop the requirement for a sample of fixed size, and only require that each member of the population will have a fixed probability of being in the sample. We could draw a 5/80 = 1/16 sample of our class by using the digits in groups to give a decimal number, say

0.1404 0.8886 0.2892 0.0403 0.4299 0.8708

We then choose the first member of the population if 0.1404 is less than 1/16. It is not, so we do not include this member, nor the second, corresponding to 0.8886, nor the third, corresponding to 0.2892. The fourth corresponds to 0.0403, which is less than 1/16 (0.0625) and so the fourth member is chosen as a member of the sample, and so on. This method is only suitable for fairly large samples, as the size of the sample obtained can be very variable in small sampling problems. In the example there is a better than 1 in 10 chance of finishing with a sample of 2 or fewer. (This is an example of the Binomial Distribution described in Chapter 6.)

Random sampling ensures that the only ways in which the sample differs from the population will be those due to chance. It has a further advantage; because the sample is random, we can apply the methods of probability theory to the data obtained. As we shall see in Chapters 8 and 9, this enables us to estimate the likely size of the errors we may get, for example, by standard errors or confidence intervals, and present them with our results.

The problem with random sampling is that we must have a list of the population from which the sample is to be drawn. Lists of populations may be hard to find, or they may be very cumbersome. For example, to sample the adult population in the UK, we could use the electoral roll. But a list of some 40 000 000 names would be difficult to handle, and in practice we would first take a random sample of electoral wards, and then a random sample of electors within these wards. This is, for obvious reasons, a *multi-stage random sample*. This approach contains the element of randomness, and so samples will be representative of the populations from which they are drawn. However, not all samples have an equal chance of being chosen, so it is not the same as simple random sampling.

We can also carry out sampling without a list of the population itself, provided we have a list of some larger units which contain all the members of the population. For example, we can obtain a random sample of school-children in an area by starting with a list of schools, which is quite easy to come by. We then draw a simple random sample of schools and all the children within our chosen schools form the sample of children. This is called a *cluster sample*, because we take a sample of clusters of individuals.

Sometimes it is desirable to divide the population into different strata, for

example into age and sex groups, and take random samples within these. This is rather like quota sampling, except that within the strata we choose at random. If the different strata have different values of the quantity we are measuring, this *stratified random sampling* can increase our precision considerably. There are many complicated sampling schemes for use in different situations.

In Section 2.3 we looked at the difficulties which can arise using methods of allocation which appear random but do not use random numbers. In sampling, two such methods are often suggested by researchers. One is to take every tenth subject from the list, or whatever fraction is required. The other is to use the last digit of some reference number, such as the hospital number, and take subjects where this is, say, 3 or 4 as the sample. These sampling methods are *systematic* or *quasi-random*. It is not usually obvious why they should not give 'random' samples, and it may be that in many cases they would be just as good as random sampling. They are certainly easier. To use them, we must be very sure that there is no pattern to the list which could produce an unrepresentative group. If it is possible, random sampling seems safer.

Volunteer bias can be as serious a problem in sampling studies as it is in trials (Section 2.4). Having drawn the sample, if we can only obtain data from a subset of them this subset will not be a random sample of the population. Its members will be self-selected. It is often very difficult to get data from every member of a sample. The proportion for whom data is obtained is called the *response rate* and in a sample survey of the general population is likely to be between 70 and 80 per cent. The possibility that those lost from the sample are different in some way must be taken into account. For example, they may tend to be ill, which can be a serious problem in disease prevalence studies. In a study of cigarette smoking and respiratory disease in Derbyshire schoolchildren, we drew a random sample of schools, and our sample of children was all children in the first secondary school year (Banks *et al.* 1978). We thus had a random cluster sample. The response rate to our survey was 80 per cent, most of those lost being absent from school on the day. Now, some of these absentees were ill and some were truants. Our sample may thus lead us to underestimate the prevalence of respiratory symptoms, by omitting sufferers with current acute disease, and the prevalence of cigarette smoking by omitting those who have gone for a quick smoke behind the bike sheds.

One of the most famous sampling disasters, the *Literary Digest* poll of 1936, illustrates these dangers (Bryson 1976). This was a poll of voting intentions in the 1936 US presidential election, fought by Roosevelt and Landon. The sample was a complex one. In some cities every registered voter was included, in others one in two, and for the whole of Chicago one in three. Ten million sample ballots were mailed to prospective voters, but only

2.3 million, less than a quarter, were returned. Still, two million is a lot of Americans, and these predicted a 60 per cent vote to Landon. In fact, Roosevelt won with 62 per cent of the vote. The response was so poor that the sample was most unlikely to be representative of the population, no matter how carefully the original sample was drawn. Two million Americans can be wrong! It is not the mere size of the sample, but its representativeness which is important. Provided the sample is truly representative, 2000 voters is all you need to estimate voting intentions to within two per cent, which is enough for election prediction if they tell the truth and don't change their minds. Chapter 8 describes the basis for this calculation.

3.5. Sampling in clinical studies

Having extolled the virtues of random sampling and cast doubt on all other sampling methods, we must admit that most medical data are not obtained in this way. This is partly because the practical difficulties are immense. To obtain a reasonable sample of the population of the UK, anyone can get a list of electoral wards, take a random sample of them, buy copies of the electoral rolls for the chosen wards and then take a random sample of names from it. You then knock on the door at your own risk. But suppose you want to obtain a sample of patients with cirrhosis of the liver, to see how many are medically qualified. You could get a list of hospitals easily enough and get a random sample of them, but then things would become difficult. The names of patients will only be released by the consultant in charge should he so wish, and you will need his permission before approaching them. Any study of human patients requires ethical approval, and you will need this from the ethical committee of each of your chosen hospitals. Getting the cooperation of so many people is a task to daunt the hardiest, and I know of no-one who has tried it on a national scale.

The result of this is that clinical studies are done on the patients to hand. We have touched on this problem in the context of clinical trials (Section 2.6) and the same applies to other types of clinical study. In a clinical trial we are concerned with the comparison of two treatments and we hope that the superior treatment in Stockport will also be the superior treatment in Southampton. If we are studying clinical measurement, we can hope that a measurement method which is repeatable in Middlesbrough will be repeatable in Maidenhead, and that two different methods giving similar results in one place will give similar results in another. Studies which are not comparative give more cause for concern. The natural history of a disease described in one place may differ in unpredictable ways from that in another, due to differences in the environment and the genetic make-up of the local population. Normal ranges for quantities of clinical interest, the limits within which values from most healthy people will lie, may well differ from place to

place, yet they are often determined on groups of subjects which are quite unrepresentative even of the local population. I know of one researcher who determined the normal range of her measurement from hospital and medical school personnel, a very common practice, and then took a second sample from the staff of a nearby office block. The second normal range was different from the first!

The same problem arises in attempts at computer diagnosis and at the identification of high-risk groups. They are usually based on local patient records, and so may omit factors which are important in other places. The prediction of high risk of unexpected death in infants described in Exercise 2E worked well in the northern town where the original data were obtained. It did not work nearly so well in an inner London district, where there is much greater variability in genetic stock and social circumstances and other factors were also important. The study did provide good evidence that the method worked. It was the data which were not applicable.

This does not mean that studies based on local groups of patients are without value. This is particularly so when we are concerned with comparisons between groups, as in a clinical trial, or relationships between different variables. However, we must always bear the limitations of the sampling method in mind when interpreting the results of such studies.

In general, most medical research has to be carried out using samples drawn from populations which are much more restricted than those about which we wish to draw conclusions. We may have to use patients in one hospital instead of all patients, or the population of a small area rather than that of the whole country or planet. We may have to rely on volunteers for studies of normal subjects, given most people's dislike of having needles pushed into them and disinclination to spend hours hooked up to batteries of instruments. Groups of 'normal' subjects contain medical students, nurses and laboratory technicians far more often than would be expected by chance. In animal research the problem is even worse, for not only does one batch of one strain of mice have to represent the whole species, it often has to represent members of a different order, namely humans.

Findings from such studies can only apply to the population from which the sample was drawn. Any conclusion which we come to about wider populations, such as all patients with the disease in question, depends on evidence which is not statistical and often unspecified, namely our general experience of natural variability and experience of similar studies. This may let us down, and results established in one population may not apply to another. We have seen this in the use of BCG vaccine in India (Section 2.6). It is very important wherever possible that studies should be repeated by other workers on other populations, so that we can sample the larger population at least to some extent.

3.6. Sampling in epidemiological studies

One of the most important and difficult tasks in medicine is to determine the causes of disease, so that we may devise methods of prevention. We are working in an area where experiments are often neither possible nor ethical. For example, to determine that cigarette smoking caused cancer, we could imagine a study in which children were randomly allocated to a 'twenty cigarettes a day for fifty years' group and a 'never smoke in your life' group. All we would have to do then would be to wait for the death certificates. Now, firstly, we could not persuade our subjects to stick to the treatment and, secondly, deliberately setting out to cause cancer is not ethical. We must therefore observe the disease process as best we can, by watching people in the wild state rather than under laboratory conditions. When we do this we must face the fact that the disease effect and putative cause do not exist in isolation but in a complex interplay of many intervening factors. We must do our best to assure ourselves that the relationship we observe is not the result of some other factor acting on both 'cause' and 'effect'. For example, it was once thought that the African fever tree, the yellow-barked acacia, caused malaria, because those unwise enough to camp under it were likely to develop the disease. Now, this tree grows by water where mosquitos breed, and provides an ideal daytime resting place for these insects, whose bite transmits the *Plasmodium* parasite which produces the disease. It was the water and the mosquitos which were the important factors, not the tree. Indeed, the name 'malaria' comes from a similar incomplete observation. It means 'bad air' and comes from the belief that the disease was caused by the air in marshy places, where the mosquitos bred. Epidemiological study designs must try to deal with the complex inter-relationships between different factors in order to deduce the true mechanism of disease causation. We also use a number of different approaches to the study of these problems, to see whether all produce the same answer.

One method is to use differences in mortality rates between countries or changes over time. Here the data are whole-population census data, so there is no sampling problem. The problem is rather to do with variations in diagnostic fashion and with the intervention of other variables. For example, it has been observed that countries with a high consumption of animal fat tend to have high mortality from coronary artery disease. However, such countries tend to have low consumption of dietary fibre also, so we must try to disentangle the effects of one from those of the other.

Another approach is the cross-sectional study. We take some sample or whole population and observe whether or not they have either disease or possible cause. For example, we wanted to know whether smoking causes respiratory symptoms in schoolchildren. We gave questionnaires to all first-year secondary schoolboys in a sample of schools in Derbyshire, as described

in Section 3.3. Among boys who had never smoked, 3 per cent reported a cough first thing in the morning, compared to 19 per cent of boys who said they smoked one or more cigarettes per week. Here we do have a sampling problem. The sample is representative of boys of this age in Derbyshire who answer questionnaires, but we want our conclusions to apply at least to the United Kingdom, if not the developed world or the whole planet. We also have the problem that smoking and respiratory symptoms may not be directly related, but may both be related to some other factor. For example, children whose parents smoke may be more likely to develop respiratory symptoms, because of passive inhalation of their parents' smoke, and also be more influenced to try smoking themselves. We can test this by looking at the relationship between the child's smoking and symptoms for those whose parents are not smokers, and for those whose parents are smokers, separately. As Fig. 3.1 shows, this relationship in fact persisted (Bland *et al.*, 1978) and there was no reason to suppose that a third causal factor was at work. The third problem is that the respondents may not be telling the truth, and we shall tackle this in Section 3.9.

Most diseases are not suited to this simple cross-sectional approach, because they are rare events. For example, lung cancer accounts for 9 per cent of male deaths in the UK (OPCS, DH2 No. 7), and so is a very important disease. However, the proportion of people who are known to have the disease at any given time, the *prevalence*, is quite low. Most deaths from lung cancer take place after the age of 45, so we shall consider a sample of men aged 45 and over. The average remaining lifespan of these men, in which they could contract lung cancer, will be about 30 years. The average time from

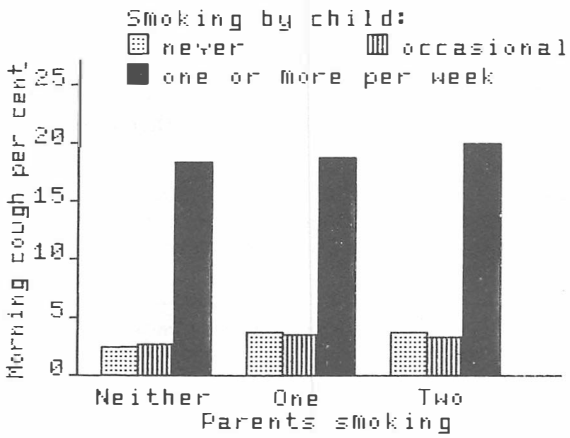


Fig. 3.1 Prevalence of self-reported morning cough in Derbyshire schoolboys, by their own and their parents' cigarette smoking (Bland *et al.* 1978).

diagnosis to death is about half a year, so of those who will contract lung cancer only 1/30 × 1/2 will have been diagnosed when the sample is drawn. Only 9 per cent of the sample will develop lung cancer anyway, so the proportion with the disease at any time is 1/30 × 1/2 × 9 per cent = 0.2 per cent or 2 per thousand. We would need a very large sample indeed to get a worthwhile number of lung cancer cases.

3.7. Case-control studies

One way of getting round the problem of the small proportion of people with the disease of interest is the *case-control study*. In this we take a group of people with the disease, the cases, and a second group without the disease, the controls. We then find the exposure of each subject to the possible causative factor and see whether this differs between the two groups. A noted case-control study was that of Doll and Hill (1950) on the aetiology of lung cancer. Twenty London hospitals notified all patients admitted with carcinoma of the lung. Notification was by various means: the admitting clerk, the house physician, the cancer registrar or the radiotherapy department, depending on the hospital. Those notified became the cases. On notification, an interviewer visited the hospital to interview the case. At the same time the interviewer selected a patient with diagnosis other than cancer, of the same sex and within the same five-year age group as the case, in the same hospital at the same time. When more than one suitable patient was available, the patient chosen was the first in the ward list considered by the ward sister to be fit for interview. Table 3.1 shows the relationship between smoking and lung cancer for these patients. A smoker was anyone who had smoked as much as one cigarette a day for as much as one year. Doll and Hill concluded that smoking is an important factor in the production of carcinoma of the lung. People have been arguing about it ever since.

The case-control study is an attractive method of investigation, because of its relative speed and cheapness compared to other approaches. However,

Table 3.1. Numbers of smokers and non-smokers among lung cancer patients and age- and sex-matched controls with diseases other than cancer (Doll and Hill 1950)

	Non-smokers	*Smokers*	*Total*
Males			
Lung cancer patients	2 (0.3%)	647	649
Controls	27 (4.2%)	622	649
Females			
Lung cancer patients	19 (31.7%)	41	60
Controls	32 (53.3%)	28	60

there are many problems which arise in the selection of cases, the selection of controls, and obtaining the data. Because of these, case-control studies sometimes produce contradictory and conflicting results.

The first problem is the selection of cases. This usually receives little consideration beyond a definition of the type of disease and a statement about the confirmation of the diagnosis. This is understandable, as there is usually little else that the investigator can do about it. He starts with the available set of patients. However, we must remember that these patients do not exist in isolation. They are the result of some process which has led to them being diagnosed as having the disease and thus being available for study.

For example, suppose we suspect that oral contraceptives might cause cancer of the breast. We have a group of patients diagnosed as having cancer of the breast. We must ask ourselves whether any of these were detected at a medical examination which took place because the woman was seeing a doctor to receive a prescription. If this were so, the pill would be associated with the *detection* of the disease rather than its cause.

Far more difficulty is caused by the selection of controls. We want a group of people who do not have the disease in question, but who are otherwise comparable to our cases. We must first decide the population from which they are to be drawn. There are two sources of controls: the general population and patients with other diseases. The latter is usually preferred because of its accessibility. Now these two populations are clearly not the same. For example, Doll and Hill gave the current smoking habits of 1014 men and women with diseases other than cancer, 14 per cent of whom were currently non-smokers. They commented that there was no difference between smoking in the disease groups — respiratory disease, cardiovascular disease, gastro-intestinal disease, and others. However, in the general population the percentage of current non-smokers was 18 per cent for men and 59 per cent for women (Todd 1972). The smoking rate in the patient group as a whole was high. Since their report, of course, smoking has been associated with diseases in each group. Smokers get more disease and are more likely to be in hospital than non-smokers.

Intuitively, the comparison we want to make is between people with the disease and healthy people, not people with a lot of other diseases. We want to find out how to prevent disease, not how to choose one disease or another! However, it is much easier to use hospital patients as controls. There may then be a bias because the factor of interest may be associated with other diseases. Suppose we want to investigate the relationship between a disease and cigarette smoking using hospital controls. Should we exclude patients with lung cancer from the control group? If we include them, our controls may have more smokers than the general population, but if we exclude them we may have less. This problem is usually resolved by choosing specific

patient groups, such as fracture cases, whose illness is thought to be unrelated to the factor being investigated.

Having defined the population we must choose the sample. There are many factors which affect disease incidence, such as age and sex, for which we wish to adjust. The most straightforward way is to take a large random sample of the control population, ascertain all the relevant characteristics, and then adjust for differences during the analysis, as described in the Derbyshire smoking study.

The alternative is to try to match a control to each case, so that for each case there is a control of the same age, sex, etc. Having done this, then we can compare our cases and controls knowing that the effects of these intervening variables are automatically adjusted for. If we wish to exclude a case we must exclude its control, too, or the groups will no longer be comparable.

Matching on some variables does not ensure comparability on all. Indeed, if it did there would be no study. Doll and Hill matched on age, sex, and hospital. They recorded area of residence and found that 25 per cent of their cases were from outside London, compared to 14 per cent of controls. If we want to see whether this influences the smoking and lung cancer relationship we must use adjustment anyway. Doll's and Hill's solution was to restrict attention to 98 pairs from district hospitals in London.

What should we match for? The more we match for, the fewer intervening variables there are to worry about. On the other hand, it becomes more and more difficult to find matches. Even matching on age and sex, Doll and Hill could not always find a control in the same hospital, and had to look elsewhere. Matching for more than age and sex can be very difficult.

Having decided on the matching variables we then find in the control population all the possible matches. If there are more matches than we need, we should choose the number required at random. Other methods, such as that used by Doll and Hill who allowed the ward sister to choose, have obvious problems of potential bias. If no suitable control can be found, we can do two things. We can widen the matching criteria, say age to within ten years rather than five, or we can exclude the case.

There is a problem of assessment bias in such studies, just as in clinical trials (Section 2.8). Interviewers will very often know whether the interviewee is a case or control and this may well affect the way questions are asked. The same problem arises in the recall of past events by the case. For example, the mother of a handicapped child may be more likely than the mother of a normal child to remember events in pregnancy which may have caused damage. These and other considerations make case-control studies extremely difficult to carry out and to interpret. The evidence from such studies can be useful, but data from other types of investigation must be considered, too, before any firm conclusions are drawn.

3.8. Cohort studies

There are many problems in interpreting the results of case-control studies. One is that the case-control design is usually *retrospective*, that is, we are starting with the present disease state, e.g. lung cancer, and relating it to the past, e.g. history of smoking. It would clearly be preferable to start with the possible cause, e.g. smoking, and see whether this leads to the disease in the future. This is a *prospective* design. We take a group of people and observe whether they have the suspected causal factor. We then follow them over time and observe whether they develop the disease. Such a sample identified at one point in time, is called a *cohort*. Compared to a case-control study, a cohort study is clearly more difficult to do. It takes longer as we must wait for the future event to occur, it involves keeping track of large numbers of people over maybe several years, and often very large numbers must be included in the sample to ensure that sufficient numbers will develop the disease between those with and without the suspected causal factor to enable comparisons to be made.

After their case-control study on smoking and cancer, Doll and Hill (1956) carried out a cohort study. They sent a questionnaire to all members of the medical profession in the UK. Respondents were asked to give their name, address, age, and details of current and past smoking habits. The deaths among this group were recorded. Only 60 per cent of doctors cooperated, so in fact, the cohort does not represent all doctors. The results for the first 53 months are shown in Table 3.2.

We have a sampling problem here, as our cohort represents doctors willing to return questionnaires, not people as a whole. We cannot use the death rates as estimates for the whole population, or even for all doctors. What we can say is that, in this group, smokers were far more likely than non-smokers to

Table 3.2. Standardized death rates per year per 1000 men aged 35 or more, in relation to most recent amount smoked, 53 months follow-up (Doll and Hill 1956)

Cause of death	No. of deaths	Non-Smokers	Smokers	Death rate among men smoking a daily average weight of tobacco of		
				1–14 g	15–24 g	25 g+
Lung cancer	84[a]	0.07	0.90	0.47	0.86	1.66
Other cancer	220	2.04	2.02	2.01	1.56	2.63
Other respiratory diseases	126	0.81	1.13	1.00	1.11	1.41
Coronary thrombosis	508	4.22	4.87	4.64	4.60	5.99
Other causes	779	6.11	6.89	6.82	6.38	7.19
All causes	1714	13.25	15.78	14.92	14.49	18.84

[a] Three deaths in which lung cancer was recorded as a contributory but not direct cause of death are recorded twice.

die from lung cancer. Now, it would be surprising if this relationship were only true for doctors, but we cannot definitely say that this would be the case for the whole population, because of the way the sample has been chosen.

We also have the problem of other intervening variables. We have not allocated doctors to be smokers or non-smokers; they have chosen themselves. The decision to begin smoking may be related to many things (social factors, personality factors, genetic factors) which may also be related to lung cancer. The great statistician, Fisher himself argued strongly against the causal interpretation. We must consider these alternative explanations very carefully before coming to any conclusion about the causes of cancer. In this study there were no data to test such hypotheses, a common problem in cohort studies. Because the sample is so large, only a little information is collected on each member of it.

There are many problems in using these observational designs, and the medical consumer of such research must be aware of them. We have no better way to tackle these questions and so we must make the best of them and look for consistent relationships which stand up to the most severe examination. We can also look for confirmation of our findings indirectly, from animal models and from dose–response relationships in the human population. However, we must accept that perfect proof is impossible in these issues and it is unreasonable to demand it. Sometimes, as with smoking and health, we must act on the balance of the evidence.

3.9. Questionnaire bias in observational studies

We have already looked at response bias in clinical trials (Section 2.7) and the same problems arise in observational studies. This is often further complicated because so many data have to be supplied by the subjects themselves.

The way in which a question is asked may influence the reply. Sometimes the bias in a question is obvious. Compare these:

(a) Do you think people should be free to provide the best medical care possible for themselves and their families, free of interference from a State bureaucracy?
(b) Should the wealthy be able to buy a place at the head of the queue for medical care, pushing aside those with greater need, or should medical care be shared solely on the basis of need for it?

Version (a) expects the answer yes, version (b) expects the answer no. We would hope not to be mislead by such blatant manipulation, but the effects of question wording can be much more subtle than this. Hedges (1978) reports several examples of the effects of varying the wording of questions. He asked two groups of about 800 subjects one of the following:

(a) Do you feel you take enough care of your health, or not?
(b) Do you feel you take enough care of your health, or do you think you could take more care of your health?

In reply to question (a), 82 per cent said that they took enough care, whereas only 68 per cent said this in reply to question (b). Even more dramatic was the difference between this pair:

(a) Do you think a person of your age can do anything to prevent ill-health in the future or not?
(b) Do you think a person of your age can do anything to prevent ill-health in the future, or is it largely a matter of chance?

Not only was there a difference in the percentage who replied that they could do something, but as Table 3.3 shows this answer was related to age for version (a) but not for version (b). Here version (b) is ambiguous, as it is quite possible to think that health is largely a matter of chance but that there is still something one can do about it. Only if it is totally a matter of chance is there nothing one can do.

Table 3.3. Replies to two similar questions about ill-health, by age (Hedges 1978)

	Age (years)			
	16–34	35–54	55 +	Total
Can do something (a)	75%	64%	56%	65%
Can do something (b)	45%	49%	50%	49%

Sometimes the respondents may interpret the question in a different way from the questioner. For example, when asked whether they usually coughed first thing in the morning, 3.7 per cent of the Derbyshire schoolchildren replied that they did. When their parents were asked about the child's symptoms 2.4 per cent replied positively, not a dramatic difference. Yet when asked about cough at other times in the day or at night 24.8 per cent of children said yes, compared to only 4.5 per cent of their parents (Bland *et al.* 1979). These symptoms all showed relationships to the child's smoking and other potentially causal variables, and also to one another. We are forced to admit that we are measuring something, but that we are not sure what!

Another possibility is that respondents may not understand the question at all, especially when it includes medical terms. In an earlier study of cigarette smoking by children, we found that 85 per cent of a sample agreed that smoking caused cancer, but that 41 per cent agreed that smoking was not harmful (Bewley *et al.* 1974). There are at least two possible explanations for

this apparent contradiction: the negative statement 'smoking is not harmful' may have confused the children; or they may not see cancer as harmful. We have evidence for both of these possibilities. In a repeat study in Kent we asked a further sample of children whether they agreed that smoking caused cancer and that 'smoking is bad for your health' (Bewley and Bland 1977). In this study 90 per cent agreed that smoking causes cancer and 91 per cent agreed that smoking is bad for your health.

In another study (Bland *et al.* 1975), we asked children what was meant by the term 'lung cancer'. Only 13 per cent seemed to us to understand and 32 per cent clearly did not, often saying 'I don't know'. They nearly all knew that lung cancer was caused by smoking, however.

Often the easiest and best method, if not the only method, of obtaining data about people is to ask them. When we do it, we must be very careful to ensure that questions are straightforward, unambiguous, and in language the respondents will understand. If this is not done then disaster is likely to follow.

Exercise 3M

(Each branch is either true or false.)

1. In statistical terms, a population:

(a) consists only of people;

(b) may be finite;

(c) may be infinite;

(d) can be any set of things in which we are interested;

(e) may consist of things which do not actually exist.

2. A one-day census of in-patients in a psychiatric hospital could:

(a) give good information about the patients in that hospital;

(b) give reliable estimates of seasonal factors in admissions;

(c) enable us to draw conclusions about the psychiatric hospitals of Britain;

(d) enable us to estimate the distribution of different diagnoses in mental illness in the local area;

(e) tell us how many patients there were in the hospital.

3. In simple random sampling:

(a) each member of the population has an equal chance of being chosen;

(b) adjacent members of the population must not be chosen;

(c) likely errors cannot be estimated;

(d) each possible sample of given size has an equal chance of being chosen;

(e) the decision to include a subject in the sample depends only on its own characteristics.

4. Advantages of random sampling include:

(a) it can be applied to any population;

(b) likely errors can be estimated;

(c) it is not biased;

(d) it is easy to do;

(e) the sample can be referred to a known population.

5. In a study of hospital patients, 20 hospitals were chosen at random from a list of all hospitals. Within each hospital, 10 per cent of patients were chosen at random:

(a) The sample of patients is a random sample.

(b) All hospitals had an equal chance of being chosen.

(c) All patients had an equal chance of being chosen.

(d) The sample could be used to make inferences about all hospital patients at that time.

(e) All possible samples of patients had an equal chance of being chosen.

6. To examine the relationship between alcohol consumption and cancer of the oesophagus, feasible studies include:

(a) questionnaire survey of a random sample from the electoral role;

(b) comparison of history of alcohol consumption between a group of oesophageal cancer patients and a group of healthy controls matched for age and sex;

(c) comparison of current oesophageal cancer rates in a group of alcoholics and a group of teetotallers;

(d) comparison by questionnaire of history of alcohol consumption between a group of oesophageal cancer patients and a random sample from the electoral role in the surrounding district;

(e) comparison of death rates due to cancer of the oesophagus in a large

sample of subjects whose alcohol consumption has been determined in the past.

Exercise 3E

Between 1977 and 1979, a series of studies were reported concerning a possible relationship between eating cornflakes and Crohn's disease. The first paper reported a strong association between the two, which subsequent authors failed to demonstrate. In this exercise we shall analyse two of these studies to see whether these contradictions can be resolved and, if we can, suggest possible explanations for the initial finding.

Crohn's disease is an inflammatory disease, usually of the last part of the small intestine. It can cause a variety of symptoms, including vague pain, diarrhoea, acute pain, and obstruction. Treatment may be by drugs or surgery, but many patients have had the disease for many years.

The suggestion that cornflakes may cause Crohn's disease arose in the study of James (1977). His initial hypothesis was that foods taken at breakfast may be associated with Crohn's disease. James studied 16 men and 18 women with Crohn's disease, aged 19–64 years, mean time since diagnosis 4.2 years. These were compared to controls, drawn from hospital patients without major gastro-intestinal symptoms. Two controls were chosen per patient, matched for age and sex. James interviewed all cases and controls himself. Cases were asked whether they ate various foods for breakfast before the onset of symptoms, and controls were asked whether they ate

Table 3E.1. Numbers of Crohn's Disease patients and controls who ate various cereals regularly and otherwise (James 1977)

		Patients	Controls	Significance test
Cornflakes	Regularly	23	17	$p < 0.0001$
	Rarely or never	11	51	
Wheat	Regularly	16	12	$p < 0.01$
	Rarely or never	18	56	
Porridge	Regularly	11	15	$0.5 > p > 0.1$
	Rarely or never	23	53	
Rice	Regularly	8	10	$0.5 > p > 0.1$
	Rarely or never	26	56	
Bran	Regularly	6	2	$p = 0.02$
	Rarely or never	28	66	
Muesli	Regularly	4	3	$p = 0.17$
	Rarely or never	30	65	

various foods before a corresponding time. Table 3E.1 shows the number of patients who reported eating various cereals regularly (i.e. at least once a week) or otherwise. The p values will be explained in detail in Chapters 9 and 13. For the moment, we only need to know that this is an index of the strength of evidence. The smaller p is, the more sure we can be that the two variables, such as eating cornflakes and having Crohn's disease, are related. We usually conclude that they are related if $p < 0.05$. In this case we say the difference or relationship is significant.

There was a significant excess of eating of cornflakes, wheat, and bran among the Crohn's patients. The consumption of different cereals was interrelated, people reporting one cereal being likely to report others. In James' opinion the principal association of Crohn's disease was with cornflakes, based on the apparent strength of the association. Only one case had never eaten cornflakes.

Several papers soon appeared in which this study was repeated, with variations. None were identical in design to James' study and none appeared to support his findings. We shall discuss that of Mayberry *et al.* (1978). They interviewed 100 patients with Crohn's disease, mean duration nine years. They obtained 100 controls, matched for age and sex, from patients and their relatives attending a fracture clinic. Cases and controls were interviewed about their current breakfast habits (Table 3E.2). The only significant difference was an excess of fruit juice drinking in controls. Cornflakes were eaten by 29 cases compared to 22 controls, which was not significant. In this study there was no particular tendency for cases to report more foods than controls. Indeed, the groups appear to be very well balanced. The paper contained some further interesting data. The authors also asked cases whether they knew of an association between food (unspecified) and Crohn's disease.

Table 3E.2. Number of patients and controls regularly consuming certain foods at least twice weekly (Mayberry *et al.* 1978)

Foods at breakfast	Crohn's patients (n = 100)	Controls (n = 100)	Significance test
Bread	91	86	
Toast	59	64	
Egg	31	37	
Fruit or fruit juice	14	30	$p < 0.02$
Porridge	20	18	
Weetabix, Shreddies, or Shredded Wheat	21	19	
Cornflakes	29	22	
Special K	4	7	
Rice Krispies	6	6	
Sugar Puffs	3	1	
Bran or All Bran	13	12	
Muesli	3	10	
Any cereal	55	55	

The association with cornflakes was reported by 29, and 12 of these had stopped eating them, having previously eaten them regularly. In their 29 matched controls, three were past cornflakes eaters. Of the 71 Crohn's patients who were unaware of the association, 21 had discontinued eating cornflakes compared to 10 of their 71 controls. The authors remarked 'seemingly patients with Crohn's disease had significantly reduced their consumption of cornflakes compared with controls, irrespective of whether they were aware of the possible association'.

1. Are the cases and controls comparable in either of these studies?
2. What other sources of bias could there be in these designs?
3. What is the main point of difference in design between the study of James and that of Mayberry *et al.*?
4. In the study of Mayberry *et al.*, how many Crohn's cases and how many controls had ever been regular eaters of cornflakes? How does this compare with James' findings?
5. Why did James think cornflakes were particularly important?
6. For the data of Table 3E.1, calculate the percentage of cases and controls who said that they ate the various cereals. Now divided the proportion of cases who said that they had eaten the cereal by the proportion of controls who reported eating it. This tells us, roughly, how much more likely cases to report the cereal than were controls. Do you think cornflakes are particularly important?
7. If we have an excess of all cereals when we ask what was ever eaten, and none when we ask what is eaten now, what possible factors could account for this?

4. Summarizing data

4.1. Types of data

In Chapters 2 and 3 we looked at ways in which data are collected. In this chapter we shall see how data can be summarized to help to reveal information they contain. We do this by calculating numbers from the data which extract the important material. These numbers are called *statistics*. A statistic is anything calculated from the data alone.

It is often useful to distinguish between three types of data: qualitative; discrete quantitative; and continuous quantitative.

Qualitative data arise when individuals may fall into separate classes. These classes may have no numerical relationship with one another at all, e.g. sex: male, female; types of dwelling: house, maisonette, flat, lodgings; eye colour: brown, grey, blue, green; etc.

Quantitative data are numerical, arising from counts or measurements. If the values of the measurements are integers (whole numbers), like the number of people in a household, or number of teeth which have been filled, these data are said to be *discrete*. If the values of the measurements can take any number in a range, such as height or weight, the data are said to be *continuous*.

In practice there is overlap between these categories. Most continuous data are limited by the accuracy with which measurements can be made. Human height, for example, is difficult to measure more accurately than to the nearest millimetre and is more usually measured to the nearest centimetre. So only a finite set of possible measurements is actually available, although the quantity 'height' can take an infinite number of possible values, and the measured height is really discrete. However, the methods described below for continuous data will be seen to be those appropriate for its analysis.

We shall refer to qualities or quantities such as sex, height, age, etc. as *variables*, because they vary from one member of a sample to another. A qualitative variable is also termed a *categorical variable* or an *attribute*. We shall use these terms interchangeably.

4.2. Frequency distributions

When data are purely qualitative, the simplest way to deal with them is to count the number of cases in each category. For example, in the analysis of

Table 4.1. Principal diagnosis of patients in Tooting Bec Hospital on 23 May 1973 (Bewley *et al.* 1974)

Diagnosis	Schizophrenia	Affective illness	Organic brain syndrome	Subnormality	Alcoholism	Other and not known	Total
Number of cases	474	277	405	58	57	196	1467

the census of a psychiatric hospital population described in Chapter 3 (Bewley *et al.* 1975) one of the variables of interest was the patient's principal diagnosis. To summarize these data, we count the number of patients having each diagnosis. The results are shown in Table 4.1.

The count of individuals having a particular quality is called the *frequency* of that quality. For example, the frequency of schizophrenia is 474. The proportion of individuals having the quality is called the *relative frequency* or *proportional frequency*. The relative frequency of schizophrenia is 474/1467 = 0.32 or 32 per cent. The set of frequencies of all the possibilities is called the *frequency distribution* of the variables.

In this census we assessed whether patients were 'likely to be discharged', 'possibly to be discharged' or 'unlikely to be discharged'. The frequencies of these categories are shown in Table 4.2. Likelihood of discharge is a qualitative variable, like diagnosis, but the categories are ordered. This enables us to use another set of summary statistics, the cumulative frequencies. The *cumulative frequency* for a value of a variable is the number of individuals with values less than or equal to that value. Thus, if we order likelihood of discharge from 'unlikely', through 'possibly' to 'likely' the cumulative frequencies are 871, 1210 (= 871 + 339) and 1467. The *relative cumulative frequency* for a value is the proportion of individuals in the sample with values less than or equal to that value. For the example they are 0.59 (= 871/1467), 0.82 and 1.00. Thus we can see that the proportion of patients for whom discharge was not thought likely was 0.82 or 82 per cent.

Table 4.2. Likelihood of discharge of patients in Tooting Bec Hospital (Bewley *et al.* 1974)

	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
Unlikely to be discharged	871	0.59	871	0.59
Possibly to be discharged	339	0.23	1210	0.82
Likely to be discharged	257	0.18	1467	1.00
Total	1467	1.00	1467	1.00

As we have noted, likelihood of discharge is a qualitative variable, with ordered categories. Sometimes this ordering is taken into account in analysis, sometimes not. Although the categories are ordered this is not quantitative data. There is no sense in which the difference between 'likely' and 'possibly' is the same as the difference between 'possibly' and 'unlikely'.

Table 4.3 shows the frequency distribution of a quantitative variable,

Table 4.3. Parity of 125 women attending antenatal clinics at St George's Hospital

Parity	Frequency	Relative frequency (per cent)	Cumulative frequency	Relative cumulative frequency (per cent)
0	59	47.2	59	47.2
1	44	35.2	103	82.4
2	14	11.2	117	93.6
3	3	2.4	120	96.0
4	4	3.2	124	99.2
5	1	0.8	125	100.0
Total	125	100.0	125	100.0

parity. This shows the number of previous pregnancies for a sample of women booking for delivery at St George's Hospital. Only certain values are possible, as the number of pregnancies must be an integer, so this variable is discrete. The frequency of each separate value is given.

Table 4.4 shows a similar distribution for a continuous variable, forced expiratory volume in one second (FEV1) in a sample of male medical students. The data themselves are shown in Table 4.5. This frequency distribution is not a very informative summary of the data, most of the values occurring only once. The cumulative frequencies are quite satisfactory, however, and we can easily pick out such things as the halfway point, 4.1 litres.

To get a useful frequency distribution we need to divide the FEV1 scale into class intervals, e.g. from 3.0 to 3.5, from 3.5 to 4.0 and so on, and count the number of individuals with FEV1's in each class interval. The class intervals should not overlap, so we must decide which interval contains the boundary point to avoid it being counted twice. It is usual to put the lower boundary of an interval into that interval and the higher boundary into the next interval.

Table 4.5. FEV1 (litres) of 57 male medical students

4.47	4.47	3.48	5.00	3.42	3.78
3.10	3.57	4.20	4.50	3.60	3.75
4.50	2.85	3.70	4.20	3.20	4.05
4.90	5.10	5.30	4.16	4.56	3.54
3.50	5.20	4.71	3.70	4.78	4.14
4.14	4.80	4.10	3.83	3.60	2.98
4.32	5.10	4.30	3.90	3.96	3.54
4.80	4.30	3.39	4.47	3.19	
3.10	4.70	3.69	3.30	2.85	
4.68	4.08	4.44	5.43	3.04	

Table 4.4. Frequency distribution of FEV1 (litres) in 57 male medical students

FEV1	Frequency	Cumulative frequency	Relative cumulative frequency (per cent)
2.85	2	2	3.5
2.98	1	3	5.3
3.09	1	4	7.0
3.10	2	6	10.5
3.19	1	7	12.3
3.20	1	8	14.0
3.30	1	9	15.8
3.39	1	10	17.5
3.42	1	11	19.3
3.48	1	12	21.1
3.50	1	13	22.8
3.54	2	15	26.3
3.57	1	16	28.1
3.60	2	18	31.6
3.69	1	19	33.3
3.70	2	21	36.8
3.75	1	22	38.6
3.78	1	23	40.4
3.83	1	24	42.1
3.90	1	25	43.9
3.96	1	26	45.6
4.05	1	27	47.4
4.08	1	28	49.1
4.10	1	29	50.4
4.14	2	31	54.9
4.16	1	32	56.1
4.20	2	34	59.6
4.30	2	36	63.2
4.32	1	37	64.9
4.44	1	38	66.7
4.47	3	41	71.9
4.50	2	43	75.4
4.56	1	44	77.2
4.68	1	45	78.9
4.70	1	46	80.7
4.71	1	47	82.5
4.78	1	48	84.2
4.80	2	50	87.7
4.90	1	51	89.5
5.00	1	52	91.2
5.10	2	54	94.7
5.20	1	55	96.5
5.30	1	56	98.2
5.43	1	57	100.0

Table 4.6. Frequency distribution of FEV1 in 57 male medical students: a more practical version

FEV1	Frequency	Relative frequency (per cent)
2.0–	0	0.0
2.5–	3	5.3
3.0–	9	15.8
3.5–	14	24.6
4.0–	15	26.3
4.5–	10	17.5
5.0–	6	10.5
5.5–	0	0.0
Total	57	100.0

Thus the interval starting at 3.0 an ending at 3.5 contains 3.0 but not 3.5. We can write this as

	3.0–
or	3.0–3.5⁻
or	3.0–3.499

Table 4.7. Frequency distribution of FEV1 in 57 male medical students: an alternative version

FEV1	Frequency	Relative frequency (per cent)
2.4–	0	0.0
2.6–	0	0.0
2.8–	3	5.3
3.0–	4	7.0
3.2–	3	3.5
3.4–	6	10.5
3.6–	7	12.3
3.8–	3	5.3
4.0–	6	10.5
4.2–	5	8.8
4.4–	7	12.3
4.6–	4	7.0
4.8–	2	3.5
5.0–	3	5.3
5.2–	2	3.5
5.4–	I	1.8
5.6–	0	0.0
5.8–	0	0.0
Total	57	100.0

If we take a starting point of 2.5 and an interval of 0.5 we get the frequency distribution shown in Table 4.6. Note that this is not unique. If we take a starting point of 2.4 and an interval of 0.2 we get a different distribution, as shown in Table 4.7.

The frequency distribution can be calculated easily and accurately using a computer. Manual calculation is not so easy but must be done carefully and systematically. One way described by many texts (e.g. Hill 1971) is to set up a tally system, as in Fig. 4.1. We go through the data and for each individual make a tally mark by the appropriate interval. We then count up the number in each interval. In practice this is very difficult to do accurately, and it needs to be checked and double-checked. Hill (1971) recommends writing each number on a card and dealing the cards into piles corresponding to the intervals. It is then easy to check that each pile contains only those cases in that interval and count them. This is undoubtedly superior to the tally system. My own preferred method is to order the observations from lowest to highest before marking the interval boundaries and counting, or to use the stem and leaf plot described below. This is rather like starting from Table 4.4.

FEV1	Tally	Frequency
2.0 ~ 2.5⁻		0
2.5 ~ 3.0⁻	|||	3
3.0 ~ 3.5⁻	𝍸 ||||	9
3.5 ~ 4.0⁻	𝍸 𝍸 ||||	14
4.0 ~ 4.5⁻	𝍸 𝍸 𝍸	15
4.5 ~ 5.0⁻	𝍸 𝍸	10
5.0 ~ 5.5⁻	𝍸 |	6
5.5 ~ 6.0⁻		0
Total		57

Fig. 4.1 Tally system for finding the frequency distribution of FEV1.

4.3. Histograms and other frequency graphs

Graphical methods are very useful for examining frequency distributions. Figure 4.2 shows a graph of the cumulative frequency distribution for the FEV1 data. This is what is called a step function, for obvious reasons. We can smooth this by joining successive points where the cumulative frequency

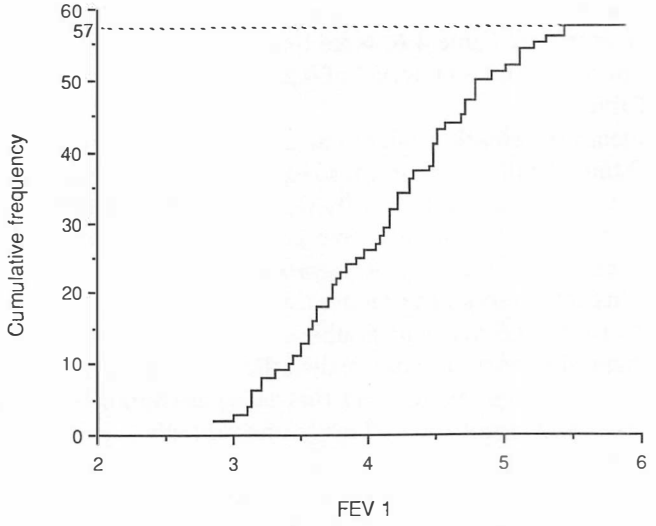


Fig. 4.2 Cumulative frequency distribution of FEV1 in a sample of male medical students.

changes by straight lines, to give a *cumulative frequency polygon*. Figure 4.3 shows this for the cumulative relative frequency distribution of FEV1. This plot is very useful for calculating some of the summary statistics referred to in Section 4.5.

The most common way of depicting a frequency distribution is by a *histogram*. This is a diagram where the class intervals are on an axis and rectangles

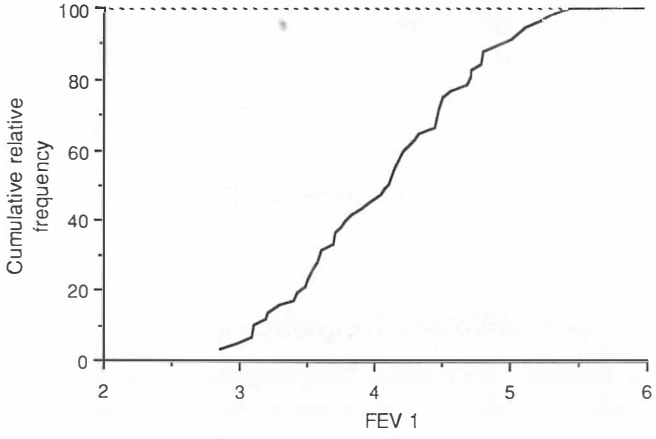


Fig. 4.3 Cumulative frequency polygon of FEV1.

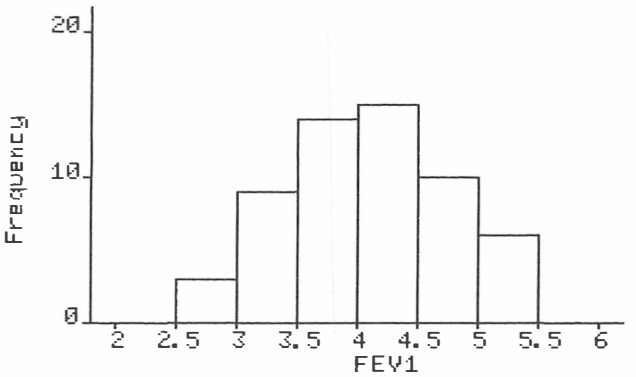


Fig. 4.4 Histogram of FEV1: frequency scale.

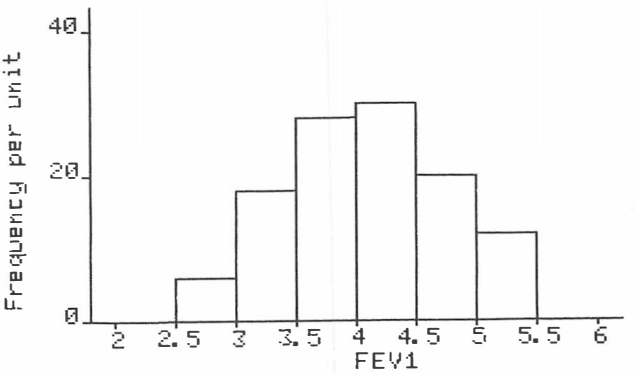


Fig. 4.5 Histogram of FEV1: frequency per unit FEV1 scale.

with heights or areas proportional to the frequencies erected on them. Figure 4.4 shows the histogram for the FEV1 distribution in Table 4.6. The vertical scale shows frequency, the number of observations in each interval. Figure 4.5 shows a histogram for the same distribution, with frequency per unit FEV1 (or frequency density) shown on the vertical axis. The distributions appear identical and we may well wonder whether it matters which method we choose. We see that it does matter when we consider a frequency distribution with unequal intervals, as in Table 4.8. If we plot the histogram using the heights of the rectangles to represent relative frequency in the interval we get Fig. 4.6, whereas if we use the relative frequency per year we get Fig. 4.7. These histograms tell different stories. Figure 4.6 suggests that the most common age for accident victims is between 15 and 44 years, whereas Fig. 4.7

Table 4.8. Distribution of age in people suffering accidents in the home (Whittington 1977)

Age group	Relative frequency (per cent)	Relative frequency per year (per cent)
0– 4	25.3	5.06
5–14	18.9	1.01
15–44	30.3	1.01
45–64	13.6	0.68
65+	11.7	0.33

suggests it is between 0 and 4. Figure 4.7 is correct, Fig. 4.6 being distorted by the unequal class intervals. It is therefore preferable in general to use the frequency per unit rather than per class interval when plotting a histogram. The frequency for a particular interval is then represented by the area of the rectangle on that interval. Only when the class intervals are all equal can the frequency for the class interval be represented by the height of the rectangle.

A different version of the histogram has been developed by Tukey (1977) in his revolutionary book *Exploratory data analysis*. This is the stem and leaf plot (Fig. 4.8). The rectangles are replaced by the numbers themselves. The 'stem' is the first digit or digits of the number and the 'leaf' the trailing digit. The first row of the figure represents the numbers 2.8, 2.8, 2.9, which in the data are 2.85, 2.85, 2.98. The plot provides a good summary of data structure, while at the same time we can see other characteristics such as a tendency to prefer some trailing digits to others, called digit preference (see Chapter 15). It is also easy to construct and much less prone to error than the tally method of finding a frequency distribution. Tukey's ideas are now becoming widely accepted and we may expect to see stem and leaf plots appearing in the medical literature.

Fig. 4.6 Age distribution of home accident victims: relative frequency scale.

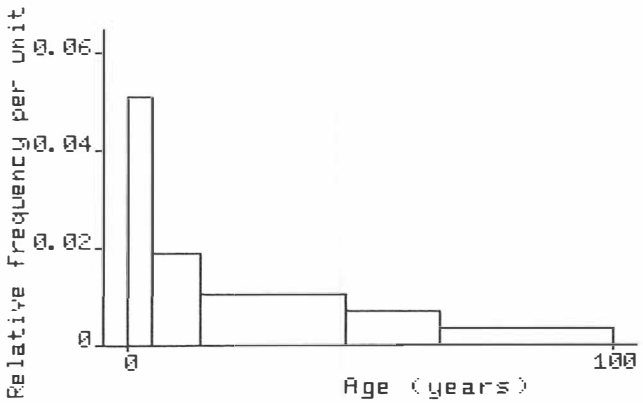


Fig. 4.7 Age distribution of home accident victims: relative frequency per year of age scale.

```
2 | 889
3 | 011123344555566677778 99
4 | 00111122333444455567778 89
5 | 011234
```

Fig. 4.8 Stem and leaf plot for the FEV1 data.

4.4. Shapes of frequency distribution

Figure 4.4 shows a frequency distribution of a shape often seen in medical data. The distribution is roughly symmetrical about its central value and has frequency concentrated about one central point. The most common value is called the *mode* of the distribution and Fig. 4.4 has one such point. It is *unimodal*. Figure 4.9 shows a very different shape. Here there are two distinct modes, one near 5 and the other near 8.5. This distribution is *bimodal*. We must be careful to distinguish between the unevenness in the histogram which results from using a small sample to represent a large population and those which result from genuine bimodality in the data. The trough between 6 and 7 in Fig. 4.9 is very marked and might represent a genuine bimodality. In this case we have children some of whom may have a condition which raises the cholesterol level and some of whom do not. We actually have two separate populations represented with some overlap between them. However, almost all distributions encountered in medical statistics are unimodal.

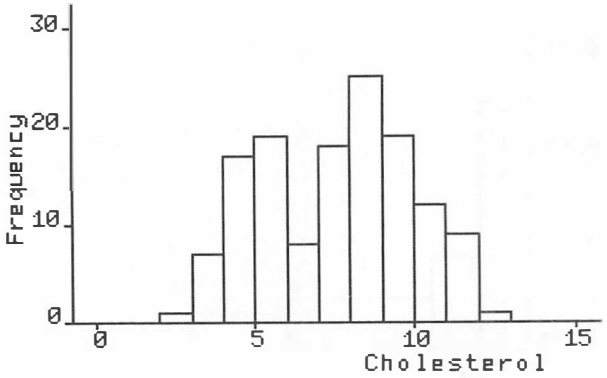


Fig. 4.9 Serum cholesterol in children from kinships with familial hypercholesterolaemia (Leonard *et al.* 1977).

Figure 4.10 differs from Figure 4.4 in a different way. We have already noted that the distribution of FEV1 is symmetrical. The distribution of serum triglyceride is *skew*, that is, the distance from the central value to the extreme is much greater on one side than it is on the other. The parts of the histogram near the extreme are called the *tails* of the distribution. If the tail on the right is longer than the tail on the left as in Fig. 4.10, the distribution is *skew to the right* or *positively skew*. If the tail on the left is longer, the distribution is *skew to the left* or *negatively skew*. If the tails are equal the distribution is *symmetrical*. Most distributions encountered in medical work are symmetrical or skew to the right, for reasons we shall discuss later (Section 7.4).

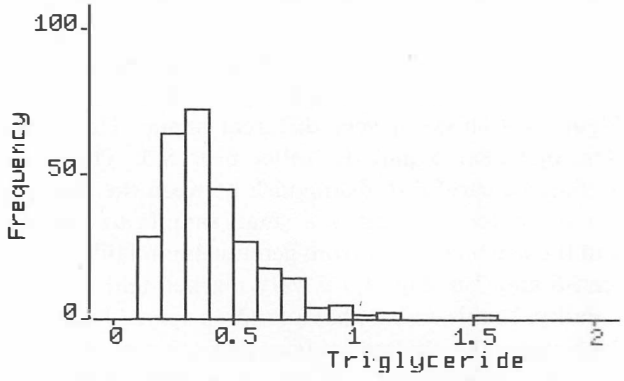


Fig. 4.10 Serum triglyceride in cord blood from 282 babies.

4.5. Medians and quantiles

We often want to summarize a frequency distribution in a few numbers, for ease of reporting or comparison. The most direct method is to use quantiles. The *quantiles* are sets of values which divide the distribution into a number of parts such that there are equal numbers of observations in each part. For example, the median is a quantile. The *median* is the central value of the distribution, such that half the points are less than or equal to it and half are greater than or equal to it. We can estimate any quantiles easily from the cumulative frequency distribution or a stem and leaf plot. For the FEV1 data the median is 4.1, the 29th value in Table 4.4. If we have an even number of points, we choose a value midway between the two central values.

In general, we estimate the q quantile, the value such that a proportion q will be below it, as follows. We have n ordered observations which divide the scale into $n + 1$ parts: below the lowest observation, above the highest and between each adjacent pair. The proportion of the distribution which lies below the ith observation is estimated by $i/(n + 1)$. We set this equal to q and get $i = q(n + 1)$. If i is an integer, the ith observation is the required quantile estimate. If not, let j be the integer part of i, the part before the decimal point. The quantile will lie between the jth and $(j + 1)$th observations. We estimate it by

$$x_j + (x_{j+1} - x_j) \times (i - j)$$

where x_j and x_{j+1} are the jth and $(j + 1)$th observations. For the median, for example, the 0.5 quantile, $i = q(n + 1) = 0.5(57 + 1) = 29$, the 29th observation as before.

Other quantiles which are particularly useful are the *quartiles* of the distribution. The quartiles divide the distribution into four equal parts. The second quartile is the median. For the FEV1 data the first and third quartiles are 3.54 and 4.53. For the first quartile, $i = 0.25 \times 58 = 14.5$. The quartile is between the 14th and 15th observations, which are both 3.54. For the third quartile, $i = 0.75 \times 58 = 43.5$, so the quartile lies between the 42nd and 43rd observations, which are 4.50 and 4.56. The quantile is given by $4.50 + (4.56 - 4.50) \times (43.5 - 43) = 4.53$. We often divide the distribution into 100 *centiles*. The median is thus the 50th centile. For the 20th centile of FEV1, $i = 0.2 \times 58 = 11.6$, so the quantile is between the 11th and 12th observation, 3.42 and 3.48, and can be estimated by $3.42 + (3.48 - 3.42) \times (11.6 - 11) = 3.46$. We can estimate these easily from Fig. 4.3 by finding the position of the quantile on the vertical axis, e.g. 0.2 for the 20th centile or 0.5 for the median, drawing a horizontal line to intersect the cumulative frequency polygon, and reading the quantile off the horizontal axis.

Tukey (1977) uses the median, quartiles, maximum and minimum as a convenient five figure summary of a distribution. He also suggested a neat graph,

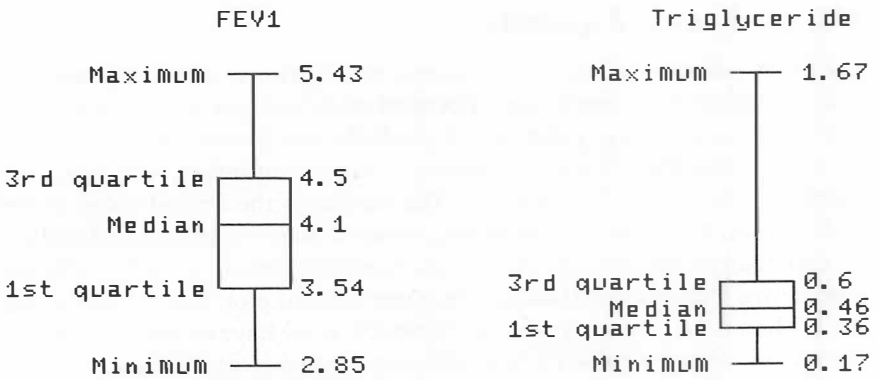


Fig. 4.11 Box and whisker plots for FEV1 and for serum triglyceride.

the *box and whisker plot* which represents this (Fig. 4.11). The box shows the distance between the quantiles, with the median marked as a line, and the 'whiskers' show the extremes. The different shapes of the FEV1 and serum triglyceride distributions is clear from the graph.

4.6. The mean

The median is not the only measure of central value for a distribution. Another is the *arithmetic mean* or average, usually referred to simply as the *mean*. This is found by taking the sum of the observations and dividing by their number. For example, consider the following artificial data:

2 3 9 5 4 0 6 3 4

The sum is 36 and there are 9 observations, so the mean is 36/9 = 4.0.

At this point we shall need to introduce some algebraic notation, widely used in statistics. We denote the observations by

$$x_1, x_2, \ldots, x_i, \ldots, x_n$$

There are n observations and the ith of these is x_i. For the example, $x_4 = 5$ and $n = 9$. The sum of all the x_i is

$$\sum_{i=1}^{n} x_i$$

The summation sign is an upper-case Greek letter, sigma, the Greek S. When it is obvious that we are adding the values of x_i for all values of i, which runs from 1 to n, we abbreviate this to

$$\sum x_i$$

or simply to Σx.

The mean of the x_i is denoted by $\bar{x}$, pronounced 'x bar', and

$$\bar{x} = \frac{1}{n} \sum x_i$$

The sum of the 57 FEV1s is 231.51 and hence the mean is 231.51/57 = 4.06. This is very close to the median, 4.1, so the median is within 1 per cent of the mean. This is not so for the triglyceride data. The median triglyceride is 0.46 but the mean is 0.51, which is higher. The median is 10 per cent away from the mean. If the distribution is symmetrical the mean and median will be about the same, but in a skew distribution they will not. If the distribution is skew to the right, as for serum triglyceride, the mean will be greater, if it is skew to the left the median will be greater. This is because the values in the tails affect the mean but not the median.

The sample mean has much nicer mathematical properties than the median and is thus more useful for the comparison methods described later. The median is a very useful descriptive statistic, but not much used for other purposes.

4.7. Variance and standard deviation

The mean and median are measures of the central tendency or position of the middle of the distribution. We shall also need a measure of the spread, dispersion or variability of the distribution.

One obvious measure is the *range*, the difference between the highest and lowest value. This is a useful descriptive measure, but it has two disadvantages. First, it depends only on the extreme values and so can vary a lot from sample to sample. Secondly, it depends on the sample size. The larger the sample is, the further apart the extremes are likely to be. We can see this if we consider a sample of size 2. If we add a third member to the sample the range will only remain the same if the new observation falls between the other two, otherwise the range will increase.

We can get round the second of these problems by using the *interquartile range*, the differences between the first and third quartiles. However, the interquartile range is quite variable from sample to sample and is also mathematically intractable. Although a useful descriptive measure, it is not the one preferred for purposes of comparison.

The most commonly used measures of dispersion are the *variance* and *standard deviation*, which we shall now describe. We start by seeing how each observation differs from its mean. Table 4.9 shows the deviations from the mean of the 9 observations of Section 4.6.

Table 4.9. Deviations from the mean of 9 observations

Observations x_i	Deviations from the mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
2	-2	4
3	-1	1
9	5	25
5	1	1
4	0	0
0	-4	16
6	2	4
3	-1	1
4	0	0
36	0	52

If the data are widely scattered, many of the observations, x_i, will be far from the mean, $\bar{x}$, and so many deviations, $x_i - \bar{x}$, will be large. If the data are narrowly scattered, very few observations will be far from the mean and so few deviations, $x_i - \bar{x}$, will be large. We need some kind of average deviation to measure the scatter. If we add all the deviations together, we get zero. This is bound to happen, because $\Sigma(x_i - \bar{x}) = \Sigma x_i - \Sigma\bar{x} = \Sigma x_i - n\bar{x}$ and $\bar{x} = \Sigma x_i/n$. Instead, we square the deviations and then add them, as shown in Table 4.9. This removes the effect of sign; we are only measuring the size of the deviation, not the direction. This gives us $\Sigma(x_i - \bar{x})^2$, in the example equal to 52, called the *sum of squares about the mean*, usually abbreviated to *sum of squares*.

Clearly, the sum of squares will depend on the number of observations as well as the scatter. We want to find some kind of average squared deviation. This leads to a difficulty. Although we want an average squared deviation, we divide the sum of squares by $(n - 1)$, not n. This is not the obvious thing to do and puzzles many students of statistical methods. The reason is that we are interested in the scatter of the population rather than that of the sample, and dividing by n would lead to small samples producing lower estimates of variability than large samples. The sum of squares is actually not proportional to n, but to $(n - 1)$. See Appendix 4A.1 for an example of this.

The estimate of variability is called the *variance*, defined as follows:

$$\text{variance} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

We have already said that $\Sigma(x_i - \bar{x})^2$ is called the sum of squares. The quantity, $n - 1$, is called the *degrees of freedom* of the variance estimate. The reason for this rather odd name is discussed in Appendix 7A.

$$\text{We have: variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

We shall usually denote the variance by s^2. In the example, the sum of squares is 52 and there are 9 observations, giving 8 degrees of freedom. Hence

$$s^2 = \frac{52}{8} = 6.5$$

The formula, $\Sigma(x_i - \bar{x})^2$, gives us a rather tedious calculation. There are two other formulae for the sum of squares, which make the calculation easier to carry out:

$$\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \text{ and } \Sigma x_i^2 - n\bar{x}^2$$

These are simply algebraic manipulations of the first form and give exactly the same answers. The algebra is quite simple and is given in Appendix 4A.2.

For example, using the second formula for the 9 observations, we have:

$$\begin{aligned}
\Sigma x_i^2 &= 2^2 + 3^2 + 9^2 + 5^2 + 4^2 + 0^2 + 6^2 + 3^2 + 4^2 \\
&= 4 + 9 + 81 + 25 + 16 + 0 + 36 + 9 + 16 \\
&= 196 \\
\Sigma x_i &= 36 \\
s^2 &= \frac{1}{n-1}\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] \\
&= \frac{1}{9-1}\left(196 - \frac{36^2}{9}\right) \\
&= \frac{1}{8} \times (196 - 144) \\
&= \frac{52}{8} \\
&= 6.5, \text{ as before}
\end{aligned}$$

On a calculator this is a much easier formula than the first, as the numbers need only be put in once. But it can be inaccurate, because we subtract one large number from another to get a small one. For this reason the first formula would be used in a computer program.

The variance is calculated from the squares of the observations. This means that it is not in the same units as the observations, which limits its use as a descriptive statistic. The obvious answer to this is to take the square root, which will then have the same units as the observations and the mean. The square root of the variance is called the *standard deviation*. Standard deviation = s, where

$$s^2 = \frac{1}{n-1}\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right]$$

Returning to the FEV1 data, we calculate the variance and standard deviation as follows:

$$\begin{aligned} n &= 57 \\ \Sigma x_i &= 231.51 \\ \Sigma x_i^2 &= 965.4499 \end{aligned}$$

$$\begin{aligned} \text{Sum of squares} &= \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \\ &= 965.4499 - \frac{231.51^2}{57} \\ &= 965.4499 - 940.296\ 142\ 1 \\ &= 25.153\ 757\ 9 \end{aligned}$$

$$\begin{aligned} s^2 &= \frac{\text{sum of squares}}{n-1} \\ &= \frac{25.1537579}{57-1} \\ &= 0.449174248 \end{aligned}$$

The standard deviation is

$$\begin{aligned} s = \sqrt{s^2} &= \sqrt{0.449174248} \\ &= 0.67024632 \\ &= 0.67 \text{ litres} \end{aligned}$$

Figure 4.12 shows the relationship for FEV1 between mean, standard deviation and frequency distribution. We see that the majority of observations are within one standard deviation of the mean, and nearly all within two

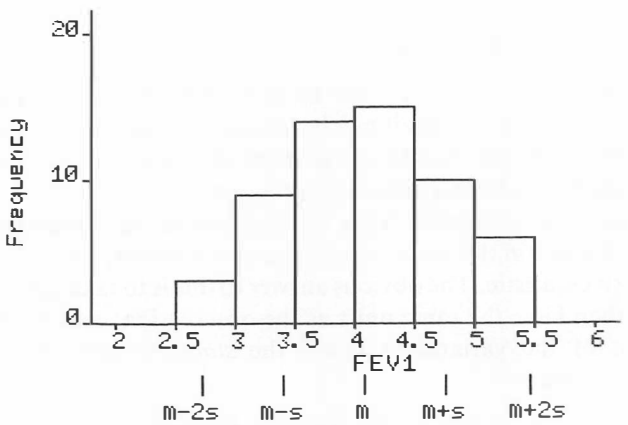


Fig. 4.12 Histogram of FEV1 with mean and standard deviation.

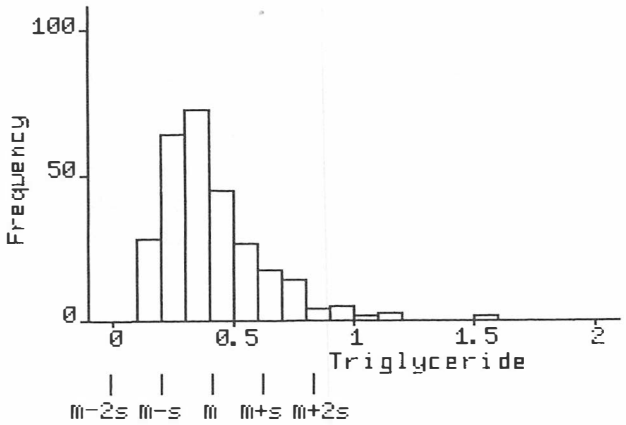


Fig. 4.13 Histogram of serum triglyceride with mean and standard deviation.

standard deviations of the mean. As Fig. 4.13 shows, this is true for the highly skew triglyceride data, too. In this case, however, the outlying observations are all in one tail of the distribution.

For large data sets calculation of mean and standard deviation may be done from a frequency distribution rather than directly from the data. The methods are described by Hill (1971), but the current ease of access to computers has made them largely obsolete.

4A. Appendix

4A.1. The divisor for the variance estimate

The variance is found by dividing the sum of squares about the sample mean by $(n - 1)$, not by n. This is because we want the scatter about the population mean, and the scatter about the sample mean is always less. The sample mean

Table 4A.1. Population of 100 random digits for a sampling experiment

9	1	0	7	5	6	9	5	8	8
1	8	8	8	5	2	4	8	3	1
2	8	1	8	5	8	4	0	1	9
1	9	7	9	7	2	7	7	0	8
7	0	2	8	8	7	2	5	4	1
1	0	5	7	6	5	0	2	2	2
6	5	5	7	4	1	7	3	3	3
2	1	6	9	4	4	7	6	1	7
1	6	3	8	0	5	7	4	8	6
8	6	8	3	5	8	2	7	2	4

is 'closer' to the data points than is the population mean. We will try a little sampling experiment to show this. Table 4A.1 shows a set of 100 random digits which we shall take as the population to be sampled. They have mean 4.74 and the sum of squares about the mean is 811.24. Hence the average squared difference from the mean is 8.1124. We can take samples of size two at random from this population using a pair of decimal dice, which will enable us to choose any digit numbered from 00 to 99. The first pair chosen was 5 and 6 which has mean 5.5. The sum of squares about the population mean 4.74 is

$$(5 - 4.74)^2 + (6 - 4.74)^2 = 1.6552$$

The sum of squares about the sample mean is

$$(5 - 5.5)^2 + (6 - 5.5)^2 = 0.5$$

The sum of squares about the population mean is greater than the sum of squares about the sample mean, and this will always be so. Table 4A.2 shows this for 20 such samples of size two. The average sum of squares about the population mean is 13.6, and about the sample mean it is 5.7. Hence, dividing by the sample size ($n = 2$) we have mean square differences of 6.8 about the population mean and 2.9 about the sample mean. Compare this to 8.1 for the population as a whole. We see that the sum of squares about the population mean is quite close to 8.1, while the sum of squares about the sample mean is

Table 4A.2. Sampling pairs from Table 4A.1

Sample	$\Sigma(x_i - \mu)^2$	$\Sigma(x_i - \bar{x})^2$
5 6	1.6552	0.5
8 8	21.2552	0.0
6 1	15.5752	12.5
9 3	21.1752	18.0
5 5	0.1352	0.0
7 7	10.2152	0.0
1 7	19.0952	18.0
9 8	28.7752	0.5
3 3	6.0552	0.0
5 1	14.0552	8.0
8 3	13.6552	12.5
5 7	5.1752	2.0
5 2	5.5752	4.5
5 7	5.1752	2.0
8 8	21.2552	0.0
3 2	10.5352	0.5
0 4	23.0152	8.0
9 3	21.1752	18.0
5 2	7.5752	4.5
6 9	19.7352	4.5
Mean	13.6432	5.7

Table 4A.3. Sums of squares about population and sample mean for sets of 100 random samples from Table 4A.2

Number in sample	Mean sums of squares				
	About population mean $\Sigma(x_i - \mu)^2$	About sample mean $\Sigma(x_i - \bar{x})^2$	$\frac{1}{n}\Sigma(x_i - \mu)^2$	$\frac{1}{n}\Sigma(x_i - \bar{x})^2$	$\frac{1}{n-1}\Sigma(x_i - \bar{x})^2$
2	16.2	9.1	8.1	4.5	9.1
3	24.5	16.2	8.2	5.4	8.1
4	31.9	23.6	8.0	5.9	7.9
5	40.2	31.0	8.0	6.2	7.7
10	79.1	71.8	7.9	7.2	8.0

much less. However, if we divide the sum of squares about the sample mean by $(n - 1)$, i.e. 1, instead of n we have 5.7, which is not much different to the 6.8 from the sum of squares about the population mean.

Table 4A.3 shows the results of a similar experiment with more samples being taken. This was done on a computer, but is otherwise identical. The Table shows the two average sums of squares for sample sizes 2, 3, 4 and 5. We see that the sum of squares about the sample mean is always too small, but if we divide it by $(n - 1)$ instead of n the estimate is good. The sum of squares about the sample mean is proportional to $n - 1$.

4A.2. Formulae for the sum of squares about the mean

The different formulae for sums of squares are derived as follows:

$$\begin{aligned}\text{sum of squares} &= \Sigma(x_i - \bar{x})^2 \\ &= \Sigma(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \Sigma x_i^2 - \Sigma 2x_i\bar{x} + \Sigma\bar{x}^2 \\ &= \Sigma x_i^2 - 2\bar{x}\Sigma x_i + n\bar{x}^2\end{aligned}$$

because $\bar{x}$ has the same value for each of the n observations.

Now, $\Sigma x_i = n\bar{x}$, so

$$\begin{aligned}\text{sum of squares} &= \Sigma x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 \\ &= \Sigma x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \Sigma x_i^2 - n\bar{x}^2\end{aligned}$$

and putting $\bar{x} = \dfrac{1}{n}\Sigma x_i$

$$\begin{aligned}\text{sum of squares} &= \Sigma x_i^2 - n\left(\frac{1}{n}\Sigma x_i\right)^2 \\ &= \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\end{aligned}$$

We thus have three formulae for variance:

$$\begin{aligned}s^2 &= \frac{1}{n-1}\Sigma(x_i - \bar{x})^2 \\ &= \frac{1}{n-1}(\Sigma x_i^2 - n\bar{x}^2) \\ &= \frac{1}{n-1}\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right]\end{aligned}$$

Exercise 4M

(Each branch is either true or false.)

1. Which of the following are qualitative variables:

(a) sex;
(b) parity;
(c) diastolic blood pressure;
(d) diagnosis;
(e) height.

2. Which of the following are continuous variables:

(a) blood glucose;
(b) peak expiratory flow rate;
(c) age last birthday;
(d) exact age;
(e) family size.

3. When a distribution is skew to the right:

(a) the median is greater than the mean;
(b) the distribution is unimodal;
(c) the tail on the left is shorter than the tail on the right;
(d) the standard deviation is less than the variance;
(e) the majority of observations are less than the mean.

4. The shape of a frequency distribution can be described using:

(a) a box and whisker plot;
(b) a histogram;
(c) a stem and leaf plot;
(d) mean and variance;
(e) a table of frequencies.

5. For the sample 3, 1, 7, 2, 2:

(a) the mean is 3;

(b) the median is 7;

(c) the mode is 2;

(d) the range is 1;

(e) the variance is 5.5.

Exercise 4E

In this exercise we shall find the frequency distribution of a set of measurements. We shall then summarize this using the mean and standard deviation and see how these relate to the frequency distribution.

The following are random blood glucose levels obtained from a group of first year medical students (mmol/l):

4.7	3.6	3.8	2.2	4.7	4.1	3.6	4.0	4.4	5.1
4.2	4.1	4.4	5.0	3.7	3.6	2.9	3.7	4.7	3.4
3.9	4.8	3.3	3.3	3.6	4.6	3.4	4.5	3.3	4.0
3.4	4.0	3.8	4.1	3.8	4.4	4.9	4.9	4.3	6.0

1. Make a stem and leaf plot for these data.
2. Find the minimum, maximum and quartiles and sketch a box and whisker plot.
3. Find the frequency distribution, using a class interval of 0.5.
4. Sketch the histogram of this frequency distribution.
5. Calculate the mean of the sample.
6. Calculate the sum of squares about the mean.
7. Calculate the degrees of freedom for this sum of squares and estimate the variance.
8. Calculate the standard deviation and find the mean $\pm$ one standard deviation and mean $\pm$ two standard deviations.
9. Indicate these points on the histogram. What do you notice about their relationship to the frequency distribution?

5. Presenting data

5.1. Rates and proportions

Having collected our data as described in Chapters 2 and 3 and extracted information from them using the methods of Chapter 4, we must find a way to convey this information to others. In this chapter we shall look at some of the methods of doing that. We begin with rates and proportions.

When we have data in the form of frequencies, we often need to compare the frequency with certain conditions in groups containing different totals. In Table 2.1, for example, two groups of patient pairs were compared, 29 where the later patient had a C–T scan and 89 where neither had a C–T scan. The later patient did better in 9 of the first group and 34 of the second group. To compare these frequencies we compare the proportions 9/29 and 34/89. These are 0.31 and 0.38, and so we can conclude that there is little difference. In Table 2.1, these were given as percentages, that is, the proportion out of 100 rather than out of 1, to avoid the decimal point. In Table 2.7, the Salk vaccine trial, the proportions contracting polio were presented as the number per 100 000 for the same reason.

A *rate* expresses the frequency of the characteristic of interest per 1000 (or per 100 000, etc.) of the population. For example, in Table 3.2, the results of the study of smoking by doctors, the data were presented as the number of deaths per 100 000 doctors per year. This is not a proportion, as a further adjustment has been made to allow for the time period observed. Furthermore, the rate has been adjusted to take account of any differences in the age distributions of smokers and non-smokers using a method described in Chapter 16. Sometimes the actual denominator for a rate may be continually changing. The number of deaths from lung cancer among men in England and Wales for 1983 was 26 502. The denominator for the death rate, the number of males in England and Wales, changed throughout 1983, as some died, some were born, some left the country and some entered it. The death rate is calculated by using a representative number, the estimated population at the end of June 1983, the middle of the year. This was 24 175 900, giving a death rate of 26 502/24 175 900, which equals 0.001096, or 109.6 deaths per 100 000 at risk per year. A number of the rates used in medical statistics are described in Section 16.5.

The use of rates and proportions enables us to compare frequencies obtained from unequal sized groups, base populations or time periods, but

we must beware of their use when their bases or denominators are not given. Victora (1982) reported a drug advertisement sent to doctors which described the antibiotic phosphomycin as being '100 per cent effective in chronic urinary infections'. This is very impressive. How could we fail to prescribe a drug which is 100 per cent effective? The study on which this was based used 8 patients, after excluding 'those whose urine contained phosphomycin-resistant bacteria'. If the advertisement has said the drug was effective on 100 per cent of 8 cases, we would have been less impressed. Had we known that it worked in 100 per cent of 8 cases selected because it might work in them, we would have been still less impressed. The same paper quotes an advertisement for a cold remedy, where 100 per cent of patients showed improvement. This was out of 5 patients! As Victora remarks, such small samples are understandable in the study of very rare diseases, but not for the common cold!

Sometimes we can fool ourselves as well as others by omitting denominators. I once carried out a study of the distribution of the soft-tissue tumour Kaposi's sarcoma in Tanzania (Bland *et al.* 1977), and while writing it up I came across a paper setting out to do the same thing (Schmid 1973). One of the factors studied was tribal group, of which there are over 100 in Tanzania. This paper reported 'the tribal incidence in the Wabende, Wambwe and Washirazi is remarkable . . . These small tribes, each with fewer than 90 000 people, constitute the group in which a tribal factor can be suspected'. This is based on the following rates of tumours per 10 000 population: national, 0.1; Wabende, 1.3; Wambwe, 0.7; Washirazi, 1.3. These are very big rates compared to the national, but the populations on which they are based are small: 8000, 14 000 and 15 000 respectively. To get a rate of 1.3/10 000 out of 8000 Wabende people we must have $8000 \times 1.3/10\,000 = 1$ case! Similarly we have one case among the 14 000 Wambwe and two among the 15 000 Washirazi. We can see that there are not enough data to draw the conclusions which the author has drawn. Rates and proportions are powerful tools and we must beware of them becoming detached from the original data.

5.2. Significant figures

When we calculated the death rate due to lung cancer among men in 1983 we quoted the answer as 0.001 096 or 109.6 per 100 000 per year. This is an approximation. The rate to the greatest number of figures my calculator will give is 0.001 096 215 653, and this number would probably go on indefinitely, turning into a recurring series of digits. The decimal system of representing numbers cannot in general represent fractions exactly. We know that $1/2 = 0.5$, but $1/3 = 0.333\ 333\ 33\ldots$ recurring infinitely. This does not usually worry us, because for most applications the difference between 0.333 and 1/3 is too small to matter. Only the first few non-zero digits of the

number are important and we call these the *significant digits* or *significant figures*. There is usually little point in quoting statistical data to more than three significant figures. After all, it hardly matters whether the lung cancer mortality rate is 0.001 096 or 0.001 097. The value 0.001 096 is given to four significant figures. The leading zeros are not significant, the first significant digit in this number being '1'. To three significant figures we get 0.001 10, because the last digit is 6 and so the 9 which precedes it is rounded up to 10. Note that significant figures are not the same as decimal places. The number 0.001 10 is given to five decimal places, the number of digits after the decimal point. When rounding to the nearest digit, we leave the last significant digit, 9 in this case, if what follows it is less than 5, and increase by one if what follows is greater than 5. When we have exactly 5, I would always round up, i.e. 1.5 goes to 2. This means that 0, 1, 2, 3, 4 go down and 5, 6, 7, 8, 9 go up, which seems unbiased. Some writers take the view that 5 should go up half the time and down half the time, since it is exactly midway between the preceeding digit and that digit plus one. Various methods are suggested for doing this but I do not recommend them myself. In any case, it is usually a mistake to round to so few significant figures that this matters.

How many significant figures we need depends on the use to which the number is to be put and on how accurate it is anyway. For example, if we have a sample of 10 sublingual temperatures measured to the nearest half degree, there is little point in quoting the mean to more than three significant figures. We shall consider this further in Chapter 8. One thing we should not do is to round numbers to a few significant figures before we have completed our calculations. In the lung cancer mortality rate example, suppose we round the numerator and denominator to two significant figures. We have 27 000/24 000 000 = 0.001 125 and the answer is only correct to two figures. This can spread through calculations causing errors to build up. We always try to retain several more significant figures than we require for the final answer.

Consider Table 5.1. This shows mortality data in terms of the exact numbers of deaths in one year. The table is taken from a much larger table (OPCS, DH2 No. 10) which shows the numbers dying from every cause of death in the *International Classification of Diseases* (ICD), which gives numerical codes to many hundreds of causes of death. The full table, which also gives deaths by age group, covers seventy A4 pages. Table 5.1 shows deaths for broad groups of diseases called ICD chapters. This table is not a good way to present these data if we want to get an understanding of the frequency distribution of cause of death, and the differences between causes in men and women. This is even more true of the seventy-page original. This is not the purpose of the table, of course. It is a source of data, a reference document from which users extract information for their own purposes. Let us see how Table 5.1 can be simplified. First, we can reduce the number of

Table 5.1. Deaths by sex and cause, England and Wales, 1983, (OPCS, DH2 No. 10)

ICD chapter and type of disease		Number of deaths	
		Males	Females
I	Infectious and parasitic	1 089	954
II	Neoplasms (cancers)	71 503	62 767
III	Endocrine, nutritional and metabolic diseases and immunity disorders	2 566	3 587
IV	Blood and blood-forming organs	592	920
V	Mental disorders	1 340	2 802
VI	Nervous system and sense organs	3 637	3 987
VII	Circulatory system	139 256	143 559
VIII	Respiratory system	42 747	43 886
IX	Digestive system	6 691	9 147
X	Genito-urinary system	3 681	4 149
XI	Complications of pregnancy, childbirth and the puerperium	0	54
XII	Skin and subcutaneous tissues	140	365
XIII	Musculo-skeletal system and connective tissues	815	2 426
XIV	Congenital anomalies	1 554	1 390
XV	Certain conditions originating in the perinatal period	1 388	1 034
XVI	Signs, symptoms, and ill-defined conditions	1 147	1 426
XVII	Injury and poisoning	11 273	7 736
Total		289 419	290 189

significant figures. Let us be extreme and reduce the data to one significant figure (Table 5.2). This makes comparisons rather easier, but it is still not obvious which are the most important causes of death. We can improve this by re-ordering the table to put the most frequent cause, diseases of the circulatory system, first (Table 5.3). We can also combine a lot of the smaller categories into an 'others' group. I did this arbitrarily, by combining all those accounting for less than 2 per cent of the total. Now it is clear at a glance that the most important causes of death in England and Wales are diseases of the circulatory system, neoplasms and diseases of the respiratory system, and that these dwarf all the others. Of course, mortality is not the only indicator of the importance of a disease. Chapter XIII of the ICD, diseases of the musculo-skeletal system and connective tissues, are easily seen from Table 5.2 to be only minor causes of death, but this group includes arthritis and rheumatism, the most important illnesses in their effects on daily activity.

5.3. Presenting tables

Tables 5.1 to 5.3 illustrate a number of useful points about the presentation of tables. Like all the tables in this book, they are designed to stand alone from the text. There is no need to refer to material buried in some paragraph

Table 5.2. Deaths by sex and cause, England and Wales, 1983, rounded to one significant figure

ICD chapter and type of disease		Number of deaths	
		Males	Females
I	Infectious and parasitic	1 000	1 000
II	Neoplasms (cancers)	70 000	60 000
III	Endocrine, nutritional and metabolic diseases and immunity disorders	3 000	4 000
IV	Blood and blood-forming organs	600	900
V	Mental disorders	1 000	3 000
VI	Nervous system and sense organs	4 000	4 000
VII	Circulatory system	100 000	100 000
VIII	Respiratory system	40 000	40 000
IX	Digestive system	7 000	9 000
X	Genito-urinary system	4 000	4 000
XI	Complications of pregnancy, childbirth and the puerperium	0	50
XII	Skin and subcutaneous tissues	100	400
XIII	Musculo-skeletal system and connective tissues	800	2 000
XIV	Congenital anomalies	2 000	1 000
XV	Certain conditions originating in the perinatal period	1 000	1 000
XVI	Signs, symptoms, and ill-defined conditions	1 000	1 000
XVII	Injury and poisoning	10 000	8 000
Total		300 000	300 000

to interpret the table. A table is intended to communicate information, so it should be easy to read and understand. A table should have a clear title, stating clearly and unambiguously what the table represents. The rows and columns must also be labelled clearly.

When proportions, rates or percentages are used in a table together with frequencies, they must be easy to distinguish from one another. This can be done, as in Table 2.9, by adding a '%' symbol, or by including a place of

Table 5.3. Deaths by sex, England and Wales, 1983, for major causes

ICD chapter and type of disease	Number of deaths	
	Males	Females
Circulatory system (VII)	100 000	100 000
Neoplasms (cancers) (II)	70 000	60 000
Respiratory system (VIII)	40 000	40 000
Injury and poisoning (XVII)	10 000	8 000
Digestive system (IX)	7 000	9 000
Others	20 000	20 000
Total	300 000	300 000

decimals. The addition in Table 2.9 of the 'total' row and the '100%' makes it clear that the percentages are calculated from the number in the treatment group, rather than the number with that particular outcome or the total number of patients.

5.4. Pie charts

It is often convenient to present data pictorially. Information can be conveyed much more quickly by a diagram than by a table of numbers. This is particularly useful when data are being presented to an audience, as here the information has to be got across in a limited time. It can also help a reader get the salient points of a table of numbers. Unfortunately, unless great care is taken, diagrams can also be very misleading and should be treated only as an addition to numbers, not a replacement.

We have already discussed methods of illustrating the frequency distribution of a qualitative variable. We will now look at the equivalent of the histogram for qualitative data, the *pie chart* or *pie diagram*. This shows the relative frequency for each category by dividing a circle into sectors, the angles of which are proportional to the relative frequency. We thus multiply each relative frequency by 360, to give the corresponding angle in degrees.

Table 5.4. Calculations for a pie chart of the distribution of cause of death

Cause of death	Frequency	Relative frequency	Angle (degrees)
Circulatory system	143 559	0.49471	178
Neoplasms (cancers)	62 767	0.21630	78
Respiratory system	43 886	0.15123	54
Injury and poisoning	7 736	0.02666	6
Digestive system	9 147	0.03152	11
Others	23 094	0.07958	29
Total	290 189	1.00000	360

Table 5.4 shows the calculation for drawing a pie chart to represent the distribution of cause of death for females, using the data of Tables 5.1 and 5.3. The resulting pie chart is shown in Fig. 5.1. This diagram is said to resemble a pie cut into pieces for serving, hence the name.

5.5. Bar charts

Histograms and pie charts depict the distribution of a single variable. A *bar chart* or *bar diagram* shows the relationship between two variables, usually

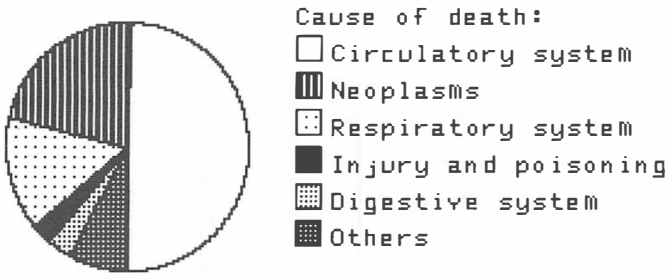


Fig. 5.1 Pie chart showing the distribution of cause of death among females, England and Wales, 1983.

one being quantitative and the other qualitative or a quantitative variable which is grouped, such as time in years. The values of the first variable are shown by the heights of bars, one bar for each category of the second variable. Table 5.5 shows the mortality due to cancer of the oesophagus in England and Wales over a ten-year period. There appears from the table to be an increase in mortality over this period. Figure 5.2 shows this relationship, the heights of the bars being proportional to the mortality.

Table 5.5. Cancer of the oesophagus: standardized mortality rate per 100 000 per year, England and Wales, 1960–69

Year	Mortality rate
60	5.1
61	5.0
62	5.2
63	5.2
64	5.2
65	5.4
66	5.4
67	5.6
68	5.8
69	6.0

Bar charts can be used to represent relationships between more than two variables. Figure 5.3 shows the relationship between children's reports of breathlessness and cigarette smoking by themselves and their parents. We can see quickly that the prevalence of the symptom increases both with the child's smoking and with that of their parents.

In the published paper reporting these respiratory-symptom data (Bland *et al.* 1978) the bar chart was not used; the data were given in the form of tables. They were thus available for other researchers to compare to their own or to carry out calculations upon. The bar chart was used to present the results

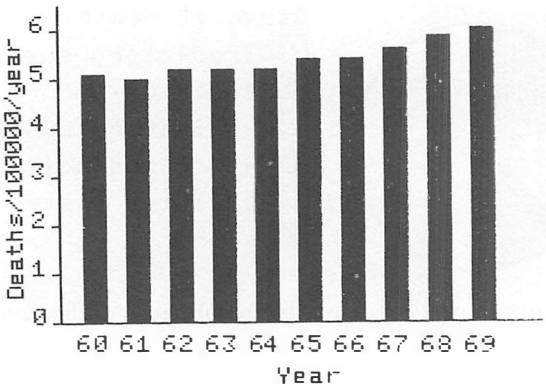


Fig. 5.2 Bar chart showing the relationship between mortality due to cancer of the oesophagus and year, England and Wales, 1960–69.

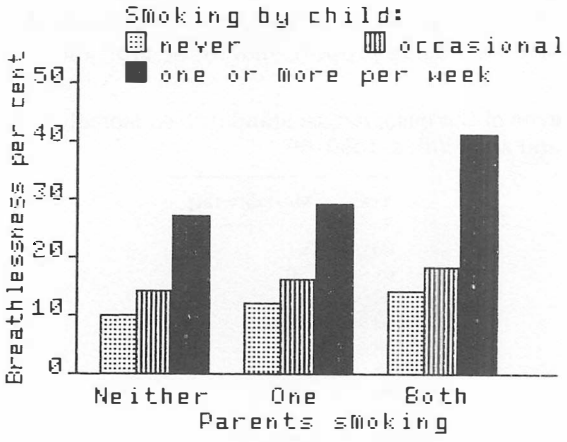


Fig. 5.3 Bar chart showing the relationship between the prevalence of self-reported breathlessness among schoolchildren and two possible causative factors.

during a conference, where the most important thing was to convey an outline of the analysis quickly.

5.6. Misleading graphs

Figures 5.1 and 5.2 are clearly titled and labelled and can be read independently of the surrounding text. The principles of clarity outlined for tables apply equally here. After all, a diagram is a method of conveying information

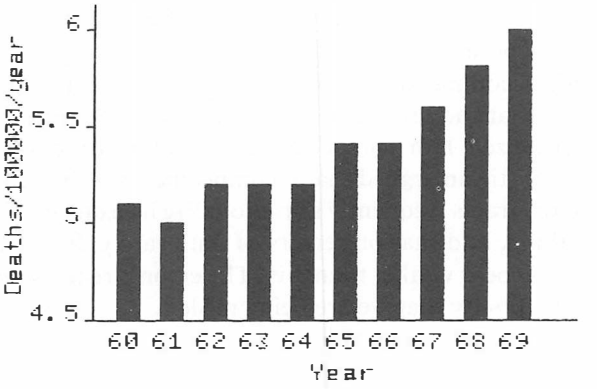


Fig. 5.4 Bar chart with zero omitted on the vertical scale.

quickly and this object is defeated if the reader or audience has to spend time trying to sort out exactly what a diagram really means. Because the visual impact of diagrams can be so great, further problems arise in their use.

The first of these is the missing zero. Figure 5.4 shows a second bar chart representing the data of Table 5.5. This chart appears to show a very rapid increase in mortality, compared to the gradual increase shown in Fig. 5.2. Yet both show the same data. Figure 5.4 omits most of the vertical scale, and instead stretches that small part of the scale where the change takes place. Even when we are aware of this, it is difficult to look at this graph and not think that it shows a massive increase in mortality. It helps if we visualize the baseline as being somewhere near the bottom of the page.

There is no zero on the horizontal axis in Figs 5.2 and 5.4, either. There are two reasons for this. There is no practical 'zero time' on the calendar; we use an arbitrary zero. Also, there is an unstated assumption that mortality rates

Table 5.6. FEV1 and height for 20 male medical students

Height (cm)	FEV1 (litres)	Height (cm)	FEV1 (litres)
174.0	4.32	167.0	3.54
180.7	4.80	171.2	3.42
183.7	4.68	177.4	3.60
177.0	5.43	171.3	3.20
177.0	3.09	183.6	4.56
172.0	3.78	183.1	4.78
176.0	3.75	172.0	3.60
177.0	4.05	181.0	3.96
164.0	3.54	170.4	3.19
178.0	2.98	171.1	2.85

vary with time and not the other way round. See Chapter 11 for further discussion of this point.

In his highly recommended book on statistical chicanery, Darrell Huff (1954) recounts that the president of a chapter of the American Statistical Association criticized him for accusing presenters of data of trying to deceive. This statistician argued that incompetence was the problem. Huff's reply was that diagrams frequently sensationalize by exaggeration and rarely minimize anything, and that presenters of data rarely distort those data to make their case appear weaker than it is. The errors are too one-sided for us to ignore the possibility that we are being misled.

When presenting data, especially graphically, be very careful that the data are shown fairly. When on the receiving end, beware!

5.7. Scatter diagrams

The bar chart would be a rather clumsy method for showing the relationship between two continuous variables, such as FEV1 and height. For this we use a *scatter diagram* or *scattergram*. This is made by marking the scales of the two variables along horizontal and vertical axes. Each pair of measurements is plotted with a cross or some other suitable symbol at the point indicated by using the measurements as coordinates. If there is more than one observation at the same coordinate we can indicate this by using the number of observation in place of the chosen symbol. Table 5.6 shows observation of two continuous variables — height and forced expiratory volume in one minute (FEV1) — for 20 male medical students. A scatter diagram for these data is shown in Fig. 5.5.

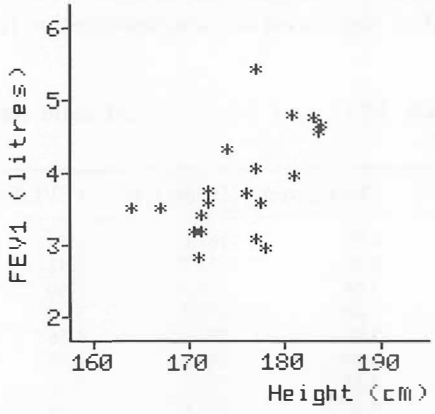


Fig. 5.5 Scatter diagram showing the relationship between FEV1 and height for a group of male medical students.

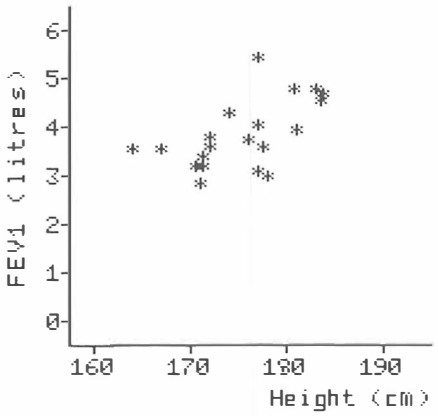


Fig. 5.6 Scatter diagram showing the relationship between FEV1 and height, with the zero included on the FEV1 scale.

Figure 5.5 commits the crime of omitting the zero. Scatter diagrams almost always do this, yet if we are to gauge the importance of the relationship between FEV1 and height by the relative change in FEV1 over the height range we need the zero on the FEV1 scale. This is shown in Fig. 5.6. We can see that height appears to be an important predictor of FEV1. Figure 5.6 does not include a zero on the height scale. As with Fig. 5.2, there is an unstated assumption that differences in height produce differences in FEV1, and although there is a true zero on the height scale, it is not of much interest.

The origin is often omitted on scatter diagrams because we are usually concerned with the existence of a relationship and the distributions followed by the observations, rather than its magnitude. We estimate the latter in a different way, described in Chapter 11.

5.8. Line graphs and time series

The data of Table 5.5 are ordered in a way that those of Table 5.6 are not, in that they are recorded at intervals in time. Such data are called a *time series*. If we plot a scatter diagram of such data, as in Fig. 5.7, it is natural to join successive points by lines to form a line graph. We do not even need to mark the points at all; all we need is the line. This would not be sensible in Fig. 5.6, as the observations are independent of one another and quite unrelated, whereas in Fig. 5.7 there is likely to be a relationship between adjacent points. Here the mortality rate recorded for cancer of the oesophagus will depend on a number of things which vary over time, including possibly causal factors, such as tobacco and alcohol consumption, and clinical factors, such as better diagnostic techniques and methods of treatment.

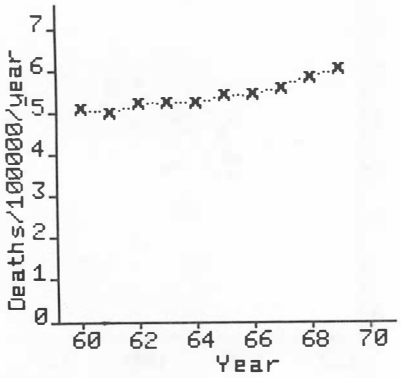


Fig. 5.7 Line graph showing changes in cancer of the oesophagus mortality over time.

Line graphs are particularly useful when we want to show the change of more than one quantity over time. Figure 5.8 shows how patients' blood pressure changed when given an active treatment and a placebo in a double-blind cross-over trial. The difference in response to the two treatments is very clear.

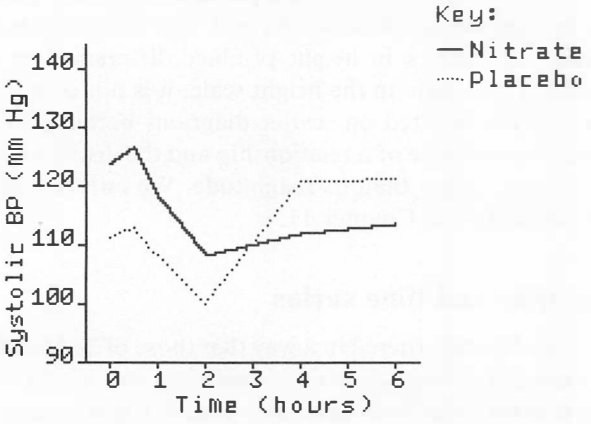


Fig. 5.8 Line graph to show the response to oral nitrate and placebo of patients who have suffered heart failure.

Unfortunately, line graphs are particularly at risk of undergoing the sort of distortion of missing zero described in Section 5.5. Figure 5.9 shows a line graph with a truncated scale, corresponding to Fig. 5.3. Just as there, the message of the graph is a dramatic increase in mortality, which the data themselves do not really support. We can make this even more dramatic by

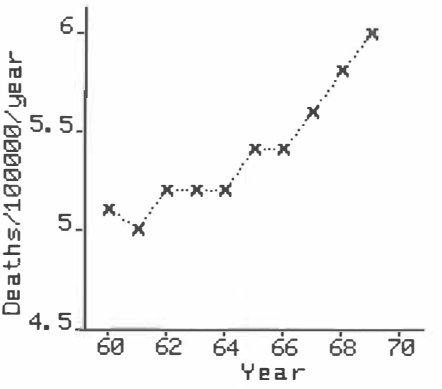


Fig. 5.9 Line graph with a missing zero.

stretching the vertical scale in Fig. 5.9, to give Fig. 5.10. The effect is now really impressive and much more likely than Fig. 5.7 to attract research funds, Nobel prizes and interviews on television. In his excellent book *How to lie with statistics*, Huff (1954) aptly names such horrors 'gee-whiz graphs'. They are even better if we omit the scales altogether and show only the soaring line, but by then we have soared out of the realms of statistics and into advertising.

This is not to say that authors who show only part of the scale are deliberately trying to mislead. There are often good arguments against graphs with vast areas of boring blank paper. In Fig. 5.8, we are not interested in blood pressures near zero and can feel quite justified in excluding them. Furthermore, it is the comparison of the two lines which is important, not their absolute magnitudes. In Fig. 5.9 we certainly are interested in zero

Fig. 5.10 Line graph with a stretched vertical scale, a 'gee-whiz' graph.

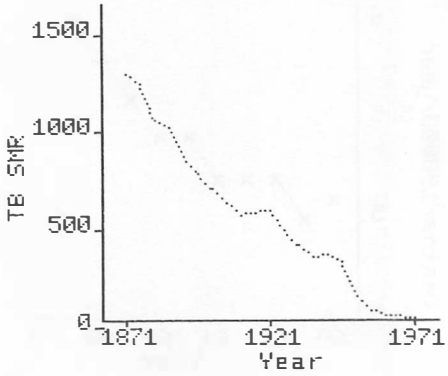


Fig. 5.11 Tuberculosis mortality in England and Wales, 1871 to 1971 (DHSS 1976).

mortality; it is surely what we are aiming for. The point is that graphs can so easily mislead the unwary reader, so let the reader beware.

5.9. Logarithmic scales

Figure 5.11 shows a line graph representing the fall in tuberculosis mortality in England and Wales over 100 years. We can see a rather unsteady curve, showing the continuing decline in the disease. Figure 5.12 shows the same data, with the mortality plotted on a logarithmic (or log) scale. A *log scale* is one where two pairs of points will be the same distance apart if their ratios are equal, rather than their differences. Thus the distance between 1 and 10 is equal to that between 10 and 100, not to that between 10 and 19. (See Appendix 5A if you do not understand this.) Figure 5.12 shows a clear kink in

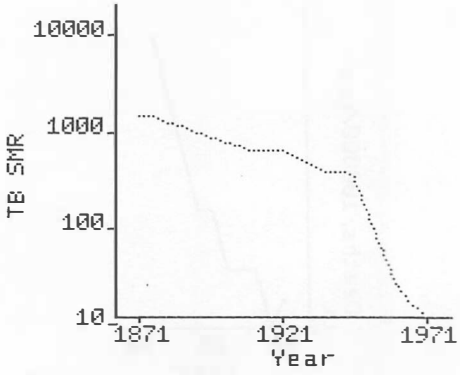


Fig. 5.12 Tuberculosis mortality shown on a log scale.

the curve about 1950, the time when a number of effective anti-TB measures, chemotherapy with streptomycin, BCG vaccine, and mass screening with X-rays were introduced. If we consider the properties of logarithms (Appendix 5A), we can see how the log scale for the tuberculosis mortality data produced such sharp changes in the curve. If the relationship is such that the mortality is falling with a constant proportion, such as 10 per cent per year, the absolute fall each year depends on the absolute level in the preceding year:

$$\text{mortality in 1960} = \text{constant} \times \text{mortality in 1959}$$

So if we plot mortality on a log scale we get:

$$\log(\text{mortality in 1960}) = \log(\text{constant}) + \log(\text{mortality in 1959})$$

For mortality in 1961, we have

$$\begin{aligned}\log(\text{mortality in 1961}) &= \log(\text{constant}) + \log(\text{mortality in 1960}) \\ &= \log(\text{constant}) + \log(\text{constant}) + \log(\text{mortality in 1959}) \\ &= 2 \times \log(\text{constant}) + \log(\text{mortality in 1959}).\end{aligned}$$

Hence we get a straight-line relationship between log mortality and time t:

$$\log(\text{mortality after } t \text{ years}) = t \times \log(\text{constant}) + \log(\text{mortality at start})$$

When the constant proportion changes, the slope of the straight line formed by plotting log (mortality) against time changes and there is a very obvious kink in the line.

Log scales are very useful analytic tools. However, a graph on a log scale can be very misleading if the reader does not allow for the nature of the scale. Figure 5.12 shows the increased rate of reduction in mortality associated with the anti-TB measures quite plainly, but it gives the impression that these measures were important in the decline of TB. This is not so. If we look at the corresponding point in Fig. 5.11, we can see that all these measures did was to accelerate a decline which had been going on for a long time (see Radical Statistics Health Group 1976).

Appendix 5A

Logarithms

Logarithms are not simply a method of calculation dating from before the computer age, but a set of fundamental mathematical functions. Because of their special properties they are much used in statistics. We shall start with logarithms (or logs for short) to base 10, the common logarithms used in calculations. The log to base 10 of a number x is y where

$$x = 10^y$$

We write $y = \log_{10}(x)$. Thus for example $\log_{10}(10) = 1$, $\log_{10}(100) = 2$, $\log_{10}(1000) = 3$, $\log_{10}(10\ 000) = 4$ and so on. If we multiply two numbers, the log of the product is the sum of their logs:

$$\log(xy) = \log(x) + \log(y)$$

For example,

$$\begin{aligned} 100 \times 1000 &= 10^2 \times 10^3 \\ &= 10^{2+3} \\ &= 10^5 \\ &= 100\ 000 \end{aligned}$$

Or in log terms:

$$\begin{aligned} \log_{10}(100 \times 1000) &= \log_{10}(100) + \log_{10}(1000) \\ &= 2 + 3 \\ &= 5 \end{aligned}$$

Hence

$$\begin{aligned} 100 \times 1000 &= 10^5 \\ &= 100\ 000 \end{aligned}$$

This means that any multiplicative relationship of the form

$$y = a \times b \times c \times d$$

can be made additive by a log transformation:

$$\log(y) = \log(a) + \log(b) + \log(c) + \log(d)$$

This is the process underlying the fit to the Lognormal Distribution described in Section 7.4.

There is no need to use 10 as the base for logarithms. We can use any number. The log of a number x to base b can be found from the log to base a by a simple calculation:

$$\log_b(x) = \frac{\log_a(x)}{\log_a(b)}$$

Ten is convenient for arithmetic using log tables, but for other purposes it is less so. For example, the gradient, slope or differential of the curve, $y = \log_{10}(x)$ is $\log_{10}(e)/x$, where e = 2.718 281 . . . is a constant which does not depend on the base of the logarithm. This leads to awkward constants spreading through formulae. To keep this to a minimum we use logs to the base e, called natural or Napierian logarithms after the mathematicians John Napier. This is the logarithm usually produced by LOG(X) functions in computer languages.

Figure 5A.1 shows the log curve for three different bases, 2, e and 10. The

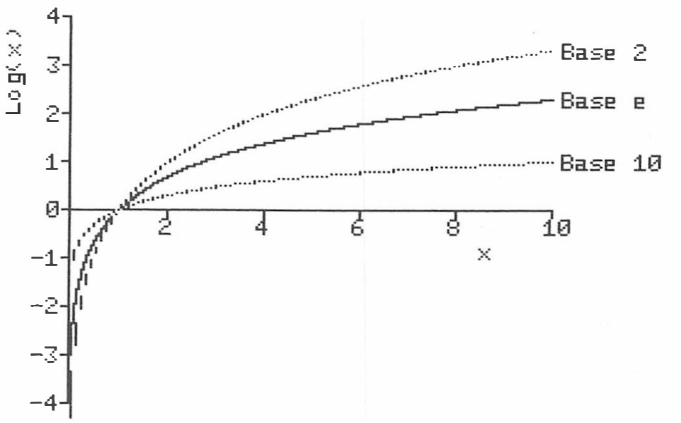


Fig. 5A.1 Logarithmic curves to three different bases.

curves all go through the point (1,0), i.e. $\log(1) = 0$. As x approaches 0, $\log(x)$ becomes a larger and larger negative number, tending towards minus infinity as x tends to zero. There are no logs of negative numbers. As x increases from 1, the curve becomes flatter and flatter. Though $\log(x)$ continues to increase, it does so more and more slowly. The curves all go through (base, 1), i.e. log(base) = 1. The curve for log to the base 2 goes through (2,1), (4,2), (8,3) because $2^1 = 2$, $2^2 = 4$, $2^3 = 8$. We can see that the effect of replacing data by their logs will be to stretch out the scale at the lower end and contract it at the upper.

We often work with logarithms of data rather than the data themselves. This may have several advantages. Multiplicative relationships may become additive, curves may become straight lines, and skew distributions may become symmetrical. See Chapters 7, 10, and 11.

Exercise 5M

(Each branch is either true or false.)

1. 'After treatment with Wondermycin, 66.67 per cent of patients made a complete recovery.'

(a) Wondermycin is wonderful.

(b) This statement may be misleading because the denominator is not given.

(c) The number of significant figures used suggest a degree of precision which may not be present.

(d) Some control information is required before we can draw any conclusions about Wondermycin.

(e) There may be only a very small number of patients.

2. The number 1729.54371:

(a) to two significant figures is 1700;

(b) to three significant figures is 1720;

(c) to six decimal places is 1729.54;

(d) to three decimal places is 1729.544;

(e) to five significant figures is 1729.5.

3. Figure 5M.1:

(a) shows a histogram;

(b) should have the vertical axis labelled;

(c) should show the zero on the vertical axis;

(d) should show the zero on the horizontal axis;

(e) should show the units for the vertical axis.

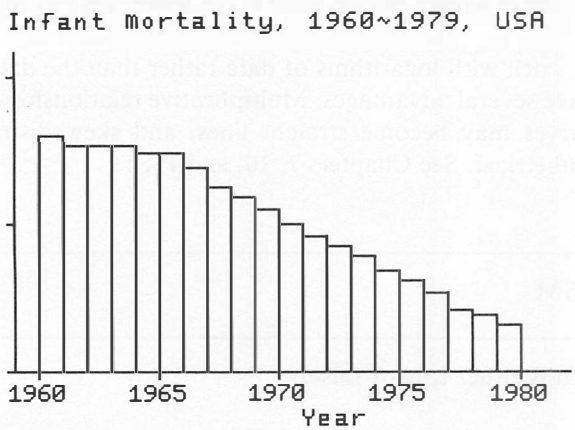


Fig. 5M.1 A dubious graph.

4. Logarithmic scales used in graphs showing time trends:

(a) show changes in the trend clearly;

(b) often produce straight lines;

(c) give a clear idea of the magnitude of changes;

(d) should show the zero point from the original scale;

(e) compress intervals between large numbers compared to those between small numbers.

5. The following methods can be used to show the relationship between two variables:

(a) histogram;

(b) pie chart;

(c) scatter diagram;

(d) bar chart;

(e) line graph.

Exercise 5E

In this exercise we shall display graphically some of the data we have studied so far.

1. Table 4.1 shows diagnoses of patients in a hospital census. Display these data as a graph.

2. Table 2.7 shows the paralytic polio rates for several groups of children. Construct a bar chart for the results from the randomized control areas.

3. Table 3.2 shows some results from the study of mortality in British doctors. Show these graphically.

4. Table 4.3 shows the parity of a group of women. Show these graphically.

5. Table 5E.1 shows the numbers of geriatric admissions in Wandsworth Health District for each week from May to September in 1982 and 1983. Show these data graphically. Why do you think the two years were different?

Table 5E.1. Weekly geriatric admissions in Wandsworth Health District from May to September, 1982 and 1983 (Fish *et al.* 1985)

Week	1982	1983	Week	1982	1983
1	24	20	12	11	25
2	22	17	13	6	22
3	21	21	14	10	26
4	22	17	15	13	12
5	24	22	16	19	33
6	15	23	17	13	19
7	23	20	18	17	21
8	21	16	19	10	28
9	18	24	20	16	19
10	21	21	21	24	13
11	17	20	22	15	29

6. Probability

6.1. Probability

We often want to use data from a sample to draw conclusions about the population from which it is drawn. For example, we might observe that a sample of patients given a new treatment respond better than patients given an old treatment on a clinical trial. We want to know whether the improvement would be seen in the whole population of patients, or if it could be due to chance. The theory of probability enables us to link samples and populations, and to draw conclusions about populations from samples. We shall start the discussion of probability with some simple randomizing devices, such as coins and dice, but the relevance to medical problems should soon became apparent.

We first ask what exactly is meant by 'probability'. There are several different approaches to this in statistics. We shall take the frequency definition. The *probability* that an event will happen under given circumstances may be defined as the proportion of trials in which the event would occur in the long run. For example, if we toss a coin it comes down either heads or tails. Before we toss it, we have no way of knowing which will happen, but we do know that it will either be heads of tails. After we have tossed it, of course, we know exactly what the outcome is. If we carry on tossing our coin, and it is a fair coin, we should get several heads and several tails. If we go on doing this for long enough, then we would expect to get as many heads as we do tails, because there is no reason to suppose the 'head' on the coin is any different to the 'tail'. So the probability of a head being thrown is half, because in the long run a head occurs on half of the throws. The number of heads which might arise in several tosses of the coin is called a *random variable*, that is, a variable which can take more than one value with given probabilities. In the same way, a die can show six faces, numbered one to six, with equal probability. We can investigate such random variables as the number of sixes in a given number of throws, the number of throws before the first six, and so on.

The same definition of probability applies to continuous measurement, such as human height. For example, suppose the median height in a population of women is 168 cm. Then half the women are above 168 cm in height. If we choose women at random (i.e. without the characteristics of the woman influencing the choice) then in the long run half the women will have heights

above 168 cm. The probability of a woman having height above 168 cm is one half. Similarly, if 1/10 of the women have height greater than 180 cm, a woman chosen at random will have height greater than 180 cm with probability 1/10. In the same way we can find the probability of height being between any given limits. When we measure a continuous quantity we are always limited by the method of measurement, and so when we say a woman's height is 170 cm we mean that it is between, say, 169.5 cm and 170.5 cm, depending on the accuracy with which we measure. So what we are interested in is the probability of the random variable taking values between certain limits rather than particular values.

6.2. Properties of probability

The following simple properties follow from the definition of probability.

1. A probability lies between 0.0 and 1.0. When the event never happens the probability is 0.0, when it always happens the probability is 1.0.

2. Suppose two events are *mutually exclusive*, i.e. when one happens the other cannot happen. Then the probability that *one or the other* happens is the *sum* of their probabilities. For example, a die may show a one or a two, but not both. The probability that it shows a one or a two = 1/6 + 1/6 = 2/6.

3. Suppose two events are *independent*, i.e. knowing one has happened tells us nothing about whether the other happens. Then the probability that *both* happen is the *product* of their probabilities. For example, suppose we toss a coin twice. Then the second toss is independent of the first toss, and the probability of two heads occurring is $1/2 \times 1/2 = 1/4$. It is not quite so easy to see why this must be so. We can list all the possibilities for the outcome of two coins:

head head
head tail
tail head
tail tail

These are equally likely events, so the probability of 'head head' must be 1/4. More generally, consider two independent events, A and B. The proportion of times A happens in the long run is the probability of A. Since A and B are independent, of those times when A happens, a proportion, equal to probability of B, will have B happen also. Hence, the proportion of times that A and B happen together is the probability of A multiplied by the probability of B.

6.3. Probability distributions and random variables

Suppose we have a set of events which are mutually exclusive and which includes all the events which can possibly happen. The sum of their probabilities is 1.0. The set of these probabilities make up a *probability distribution*. For example, if we toss a coin the two outcomes, head or tail, are mutually exclusive and these are the only events which can happen. The probability distribution is:

Event	*Probability*
Head	1/2
Tail	1/2

We can represent this with a diagram, as in Fig. 6.1. Now, let us define a variable, which we will denote by the symbol X, such that $X = 0$ if the coin shows a tail and $X = 1$ if the coin shows a head. The value X is the number of heads shown on a single toss, which must be 0 or 1. We do not know before the toss what X will be, but do know the probability of it having any possible value. We call X a *random variable*, which is defined as a quantity which may take more than one value, each with a specified probability.

What happens if we toss two coins at once? We now have four possible events: a head and a head, a head and a tail, a tail and a head, a tail and a tail. Clearly, these are equally likely and we would get each 1/4 of the trials. Now consider the number of heads, Y say. The variable Y has three possible values: 2, 1, and 0; $Y = 2$ only when we get a head and a head, so has probability 1/4. Similarly $Y = 0$ only when we get a tail and a tail and has probability 1/4. However, $Y = 1$ either when we get a head and tail, or when we have a tail and a head, and so has probability $1/4 + 1/4 = 1/2$.

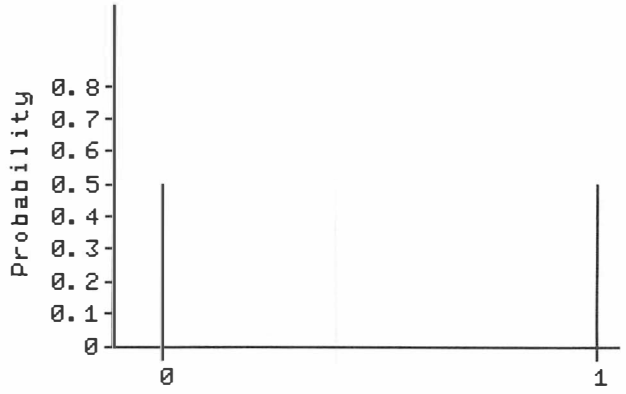


Fig. 6.1 Probability distribution for the number of heads shown in one toss of a coin.

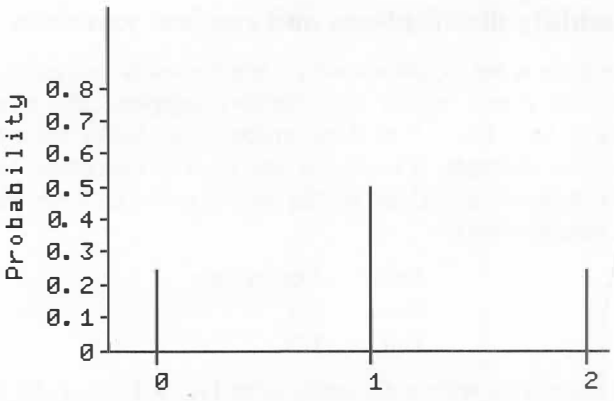


Fig. 6.2 Probability distribution for the number of heads shown in two tosses of a coin.

We can write this probability distribution as:

$$\text{Prob}(Y = 0) = 1/4$$
$$\text{Prob}(Y = 1) = 1/2$$
$$\text{Prob}(Y = 2) = 1/4$$

The probability distribution of Y is shown in Fig. 6.2.

6.4. The Binomial Distribution

We have considered the probability distributions of two random variables: X, the number of heads in one toss of a coin, taking values 0 and 1, and Y, the

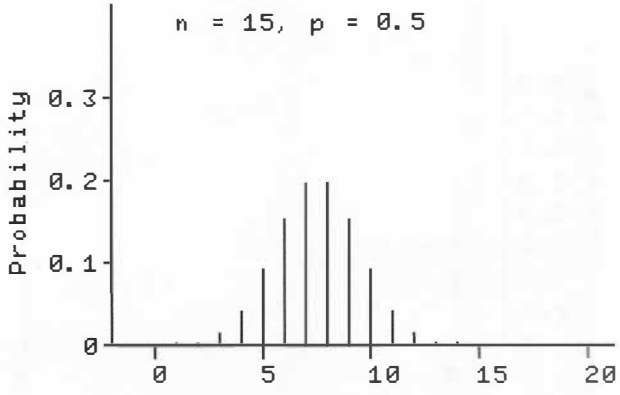


Fig. 6.3 Binomial Distribution for the number of heads shown in 15 tosses of a coin.

number of heads in two tosses of a coin, taking values 0, 1, or 2.

The distribution of X and Y are examples of the *Binomial Distribution*. It arises frequently in medical applications. There are two obvious extensions to what we have done, by taking the probability of 'head' to be other than half, and by taking more than two 'coins'. The probability of a 'head', or, more generally, of a 'success', is one parameter of the Binomial Distribution, the number of 'coins', or, more generally, the number of 'trials' is the other parameter.

For example, we could throw a die instead of a coin, and record a success if a six shows. This has probability 1/6. Figures 6.3 and 6.4 show the Binomial Distributions for 15 tosses of a coin and for 10 throws of a die. These simple randomizing devices are of interest in themselves, but not of obvious relevance to medicine. However, suppose we are carrying out a random sample prevalence survey to estimate the unknown prevalence, p, of a disease. Since members of the sample are chosen at random and independently from the population, the probability of any subject chosen having the disease is p. We thus have a series of independent trials, each with probability of success, p, and the number of successes, i.e. members of the sample with the disease, will follow a Binomial Distribution. As we shall see later, the properties of the Binomial Distribution enable us to say how accurate is the estimate of prevalence obtained.

We can calculate the probabilities for Binomial Distribution by listing all the ways in which, say, 15 coins can fall. However, there are $2^{15} = 32\ 768$ combinations of 15 coins, so this is not very practical. Instead, we find a formula for the probability in terms of the number of throws and the probability of a head.

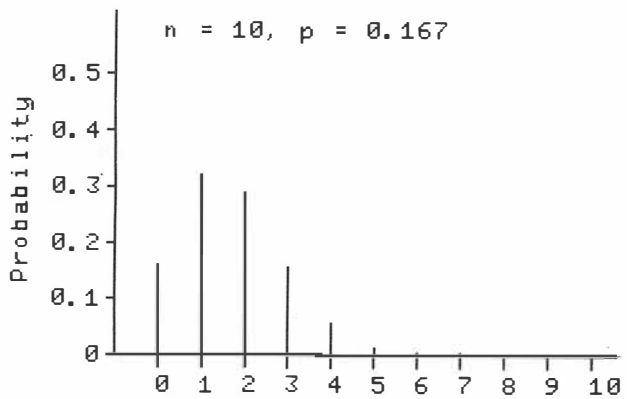


Fig. 6.4 Binomial Distribution for the number of sixes shown in 10 throws of a die.

In general, we have n independent trials with the probability that a trial is a success being p. What is the probability of r successes? For any particular series of r successes, each with probability p, and $n - r$ failures, each with probability $(1 - p)$, the probability of the series happening by chance is $p^r(1 - p)^{n-r}$, since the trials are independent and the multiplicative rule applies. The number of ways in which r things may be chosen from n things is $n!/r!(n - r)!$ (see Appendix 6A.1). Only one combination can happen at one time, so we have $n!/r!(n - r)!$ mutually exclusive ways of having r successes, each with probability $p^r(1 - p)^{n-r}$. The probability of having r successes is the sum of these:

$$\text{Prob}(r \text{ successes}) = \frac{n!}{r!(n - r)!} p^r(1 - p)^{n-r}$$

Those who remember the binomial expansion in mathematics will see that this is one term of it, hence the name Binomial Distribution. The *Binomial Distribution* is the distribution followed by the number of successes in n independent trials when the probability of a trial being a success is p.

Thus, if the probability of surviving a particular disease is 0.9 and we have a sample of 20 patients, the number who survive will be from a Binomial Distribution with $p = 0.9$ and $n = 20$.

Hence, the probability that all survive ($r = 20$) is:

$$\begin{aligned}
\text{Prob}(r = 20) &= \frac{20!}{20!(20 - 20)!} \times 0.9^{20} \times (1 - 0.9)^{20-20} \\
&= \frac{20!}{20! \times 0!} \times 0.9^{20} \times 0.1^0 \\
&= 0.9^{20} \\
&= 0.12
\end{aligned}$$

(N.B. $0! = 1$ (Appendix 6A.1) and anything to the power of zero is 1.) Similarly, the probability that 19 survive is:

$$\begin{aligned}
\text{Prob}(r = 19) &= \frac{20!}{19!(20 - 19)!} \times 0.9^{19} \times 0.1^{20-19} \\
&= \frac{20!}{19! \times 1!} \times 0.9^{19} \times 0.1^1 \\
&= 20 \times 0.9^{19} \times 0.1 \\
&= 0.27
\end{aligned}$$

Hence the probability that no more than one patient dies is the sum of these, 0.39. We can use this distribution as a model whenever we have a series of trials with two possible outcomes. If we treat a group of patients, the number who recover has a Binomial Distribution. If we measure the blood pressure of

a group of people, the number classified as hypertensive has a Binomial Distribution.

6.5. Mean and variance

The number of different probabilities in a Binomial Distribution can be very large and unwieldy. When n is large, we usually need to summarize these probabilities in some way. Just as a frequency distribution can be described by its mean and variance, so can a probability distribution and its associated random variable.

The *mean* is the average value of the random variable in the long run. It is also called the *expected value* or *expectation* and the expectation of a random variable X is usually denoted by $E(X)$.

For example, consider the number of heads in tosses of two coins. We get 0 heads in 1/4 of pairs of coins, i.e. with probability 1/4. We get 1 head in 1/2 of pairs of coins, and two heads in 1/4 of pairs. The average value we should get in the long run is found by multiplying each value by the proportion of pairs in which it occurs and adding:

$$0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1$$

If we kept on tossing pairs of coins, the average number of heads per pair would be 1. Thus for any random variable which takes discrete values the mean, expectation or expected value is found by summing each possible value multiplied by its probability.

Note that the expected value of a random variable does not have to be a value that the random variable can actually take. For example, for the mean number of heads in throws of one coin we have either no heads or one head, each with probability half, and the expected value is

$$0 \times 1/2 + 1 \times 1/2 = 1/2$$

The number of heads must be 0 or 1, but the expected value is 1/2, the average which we would get in the long run.

The *variance* of a random variable is the average squared difference from the mean. For the number of heads in two coin tosses, 0 is one unit from the mean and occurs for 1/4 of pairs of coins, 1 is zero units from the mean and occurs for 1/2 of the pairs and 2 is one unit from the mean and occurs for 1/4 of pairs, i.e. with probability 1/4. The variance is then found by squaring these differences, multiplying by the proportion of times the difference will occur (the probability) and adding:

$$\begin{aligned} & (0-1)^2 \times 1/4 + (1-1)^2 \times 1/2 + (2-1)^2 \times 1/4 \\ = \; & 1^2 \times 1/4 + 0^2 \times 1/2 + 1^2 \times 1/4 \\ = \; & \frac{1}{2} \end{aligned}$$

We denote the variance of a random variable X by $Var(X)$. In mathematical terms,

$$Var(X) = E[\{X - E(X)\}^2] = E(X^2) - \{E(X)\}^2$$

Because the variance depends on the square of the variable, it is measured in different units from the variable itself. It is often convenient to use the square root of the variance, which we call the *standard deviation*. Like the mean, this has the same units as the original variable. We often use the Greek letters μ, pronounced 'mu', and σ, pronounced 'sigma', for the mean and standard deviation of a probability distribution. The variance is then σ^2.

The mean and variance of the distribution of a continuous variable, of which more in Chapter 7, are defined in a similar way. Calculus is used to define them as integrals, but this need not concern us here. Essentially what happens is that the continuous scale is broken up into many very small intervals and the value of variable in that very small interval is multiplied by the probability of being in it, then these are added.

6.6. Properties of means and variances

When we use the mean and variance of probability distributions in statistical calculations, it is not the details of their formulae which we need to know, but some of their simple properties. The reasons for these properties are quite easy to see in a non-mathematical way.

If we add a constant to a random variable, the new variable so created has a mean equal to that of the original variable plus the constant. Suppose our random variable is human height. We can add a constant to the height by measuring the heights of people standing on a box. The mean height of people plus box will now be the mean height of the people plus the constant height of the box. However, the box will not alter the variability of the heights. The difference between the tallest and smallest, for example, will be unchanged. So the variance and standard deviation will be unchanged. We can subtract a constant by asking the people to stand in a constant hole to be measured. This reduces the mean but leaves the variance unchanged as before.

We multiply the random variable by a constant if we change our units of measurements, say from inches to centimetres. We multiply each measurement by 2.54. This has the effect of multiplying the mean by the constant, 2.54, and multiplying the standard deviation by the constant since it is in the same units as the observations. However, the variance is measured in squared units, and so is multiplied by the square of the constant. Division by a constant works in exactly the same way.

Multiplication by a negative constant is rather more difficult to model. The mean is multiplied by the constant as before, and the variance by the square

of the constant. The standard deviation, which was defined as the square root of the variance, is always positive. It is multiplied by the absolute value of the constant, i.e. the constant without the negative sign.

Another thing we often want to do is add two random variables. We can do this by measuring the height of people standing on boxes of random height. We see that the mean height is increased. In fact the mean height of people on boxes is the mean height of people + the mean height of the boxes. The variability of the heights is also increased. This is because some short people will find themselves on small boxes, and some tall people will find themselves on large boxes. To be precise, the variance of the sum of two independent random variables is the sum of their variances.

If the people do not choose their boxes at random, i.e. the two variables are not independent, something different happens. Suppose they have decided to stand on the boxes not just at a statistician's whim but for a purpose. They wish to change a light bulb, and so must reach a required height. Now the short people must pick large boxes, whereas tall people can make do with small ones. The result is a reduction in variability to almost nothing. On the other hand, if we told the tallest people to find the largest boxes and the shortest to find the smallest boxes, the variability would be increased. The mean of the sum remains the sum of the means, but the variance of the sum is not the sum of the variances. Independence is an important condition.

We can model the difference between two random variables by measuring the heights above ground level of our people standing in holes of random depth. The mean height above ground is the mean height of the people minus the mean depth of the hole. The variability is increased, because some short people stand in deep holes and some tall people stand in shallow holes. In fact, the variance of the difference between two random variables is the sum of their variances. The variables must be independent. If the variables are not independent, the additivity of the variances breaks down, as it did for the sum of two random variables. When the people try to hide in the holes, and so must find a hole deep enough to hold them, the variability is again reduced.

The effects of multiplying two random variables and of dividing one by another are much more complicated. Fortunately we very rarely need to do this.

We can now find the mean and variance of the Binomial Distribution with parameters n and p. First consider $n = 1$. Then the probability distribution is:

value	*probability*
0	$1 - p$
1	p

The mean is therefore

$$0 \times (1 - p) + 1 \times p = p$$

The variance is

$$\begin{aligned}(0-p)^2 \times (1-p) + (1-p)^2 \times p &= p^2(1-p) + p(1-p)^2 \\ &= p(1-p)(p+1-p) \\ &= p(1-p)\end{aligned}$$

Now, the Binomial variable with parameters n and p is the sum of n independent Binomial variables with parameters 1 and p. So its mean is the sum of n means all equal to p, and its variance is the sum of n variances all equal to $p(1-p)$.

Hence the Binomial Distribution has:

$$\begin{aligned}\text{mean} &= np \\ \text{variance} &= np(1-p)\end{aligned}$$

As we shall see later, these are often more useful than the Binomial probability formula.

The properties of means and variances of random variables enable us to find a formal solution to the problem of degrees of freedom for the sample variance discussed in Chapter 4. We want an estimate of variance whose expected value is the population variance. The expected value of $\Sigma(x_i - \bar{x})^2$ can be shown to be $(n-1)Var(x)$ (Appendix 6A.2) and hence we divide by $(n-1)$, not n, to get our estimate of variance.

6.7. The Poisson Distribution

The Binomial Distribution is one of many probability distributions which are used in statistics. It is a discrete distribution, that is it can take only a set of separate possible values, and is the discrete distribution most commonly encountered in medical applications. One other discrete distribution is worth discussing at this point, the Poisson Distribution. Although, like the Binomial, the Poisson Distribution arises from a simple probability model, the mathematics involved is more complicated and we shall omit it.

Suppose events happen randomly and independently in time with a constant rate, then the number of events which happen in a fixed time interval follows the Poisson Distribution. This model can be used to examine and compare annual mortality rates, for example. Deaths from many causes can be regarded as happening randomly and independently in the population and the Poisson Distribution enables us to say how far apart we would expect two mortality rates to be by chance.

The mean of the Poisson Distribution for the number of events per unit time is simply the rate, as might be expected. The variance of the Poisson Distribution also happens to be equal to the rate, so this distribution is very easy to use. The individual probability for r events happening in unit time with rate m is $e^{-m}m^r/r!$, where $e = 2.718\ldots$, the mathematical constant.

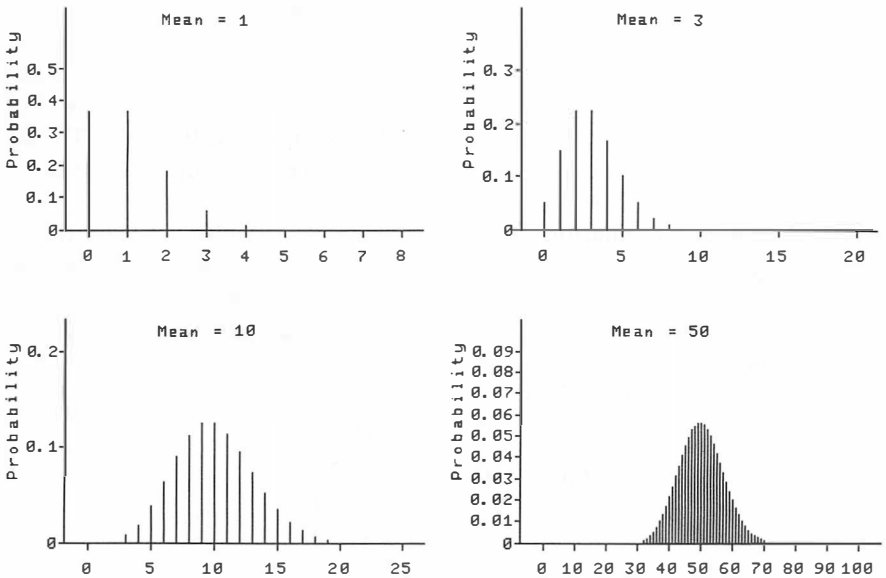


Fig. 6.5 Poisson Distributions with four different means.

However, there is seldom any need to use individual probabilities of this distribution. Its mean and variance suffice.

Figure 6.5 shows the Poisson Distribution for four different means. You will see that as the mean increases the Poisson Distribution looks rather like the Binomial Distribution in Fig. 6.3. We shall discuss this similarity further in the next chapter.

Appendix 6A

6A.1. Combinations

For those who never knew, or have forgotten, the theory of combinations, it goes like this.

First, we look at the number of permutations, i.e. ways of arranging a set of objects. Suppose we have n objects. How many ways can we order them? The first object can be chosen n ways, i.e. any object. Then only $(n - 1)$ remain, so the second object can only be chosen $(n - 1)$ ways. Hence, for each first object there are $(n - 1)$ possible second objects. Hence, there are $n \times (n - 1)$ possible first and second permutations. There are now only $(n - 2)$ choices for the third object, $(n - 3)$ choices for the fourth, and so on, until there is only one choice for the last. Hence, there are $n \times (n - 1) \times (n - 2) \times \ldots \times 2 \times 1$ permutations of n objects. We call this number the factorial of n and write it $n!$

Now we want to know how many ways there are of choosing r objects from n objects. Having made a choice of r objects, we can order those in $r!$ ways. We can also order the $(n - r)$ not chosen, in $(n - r)!$ ways. So the objects can be ordered in $r!(n - r!)$ ways without altering the objects chosen. For example, say we choose the first two from three objects, A, B, and C. Then if these are A and B, two permutations give this choice, ABC and BAC. This is, of course, $2! \times 1! = 2$ permutations.

Hence, each combination of r things accounts for $r!(n - r)!$ of the $n!$ permutations possible. Hence, there are $n!/r!(n - r)!$ possible combinations. For example, consider the number of combinations of two objects out of three, say A, B, and C. The possible choices are AB, AC, and BC. There is no other possibility.

We have $n = 3$ and $r = 2$, so

$$\begin{aligned}\frac{n!}{r!(n-r)!} &= \frac{3!}{2!(3-2)!} \\ &= \frac{3 \times 2 \times 1}{2 \times 1 \times 1} \\ &= 3\end{aligned}$$

So the formula works.

Sometimes in using this formula we come across $r = 0$ or $r = n$, giving $0!$ in the formula. This cannot be defined in the way we have chosen, but we can calculate its only possible value. Because there is only one way of choosing n objects from n, we have

$$\frac{n!}{n!(n-n)!} = 1$$

$$\frac{n!}{n! \times 0!} = 1$$

$$\frac{1}{0!} = 1$$

So $0! = 1$.

6A.2. Expected value of a sum of squares

The properties of means and variances described in Section 6.6 can be used to answer the question raised in Chapter 4 about the divisor in the variance of a sample. We ask why the variance from a sample is

$$s^2 = \frac{1}{n-1} \Sigma(x_i - \bar{x})^2$$

and not

$$\frac{1}{n} \Sigma(x_i - \bar{x})^2$$

We shall be concerned with the general properties of samples of size n, so we shall treat n as a constant and x_i and $\bar{x}$ as random variables. We shall suppose x_i has mean μ and variance σ^2.

We want the estimate of variance to be independent of the sample size, so we want the expected value of the sample variance to be σ^2.

The expected value of the sum of squares is

$$\begin{aligned} E\{\Sigma(x_i - \bar{x})^2\} &= E\{(\Sigma x_i^2) - n\bar{x}^2\} \qquad \text{(See Appendix 4A.2)} \\ &= E(\Sigma x_i^2) - E(n\bar{x}^2) \\ &= E(\Sigma x_i^2) - nE(\bar{x}^2) \end{aligned}$$

because the expected value of the difference is the difference between the expected values (Section 6.6) and n is a constant. Now, the population variance σ^2 is the average squared distance from the population mean μ, so

$$\begin{aligned} \sigma^2 &= E\{(x_i - \mu)^2\} \\ &= E(x_i^2 - 2\mu x_i + \mu^2) \\ &= E(x_i^2) + E(-2\mu x_i) + E(\mu^2) \\ &= E(x_i^2) - 2\mu E(x_i) + \mu^2 \end{aligned}$$

because μ is a constant.

Because $E(x_i) = \mu$, we have

$$\begin{aligned} \sigma^2 &= E(x_i^2) - 2\mu^2 + \mu^2 \\ &= E(x_i^2) - \mu^2 \end{aligned}$$

and so we find

$$E(x_i^2) = \sigma^2 + \mu^2$$

and so

$$\Sigma E(x_i^2) = n(\sigma^2 + \mu^2)$$

being the sum of n numbers, all of which are $\sigma^2 + \mu^2$.

We now find the value of $E(\bar{x}^2)$.

Just as

$$\begin{aligned} E(x_i^2) &= \sigma^2 + \mu^2 \\ &= Var(x_i) + \{E(x_i)\}^2 \end{aligned}$$

so

$$E(\bar{x}^2) = Var(\bar{x}) + \{E(\bar{x})\}^2$$

Now

$$\bar{x} = \frac{1}{n}\Sigma x_i$$

Hence

$$\begin{aligned} E(\bar{x}) &= E\left(\frac{1}{n}\Sigma x_i\right) \\ &= \frac{1}{n}E(\Sigma x_i) \\ &= \frac{1}{n}\Sigma E(x_i) \\ &= \frac{1}{n}\Sigma\mu \\ &= \frac{1}{n}n\mu \\ &= \mu \end{aligned}$$

$$\begin{aligned} Var(\bar{x}) &= Var\left(\frac{1}{n}\Sigma x_i\right) \\ &= \frac{1}{n^2}Var(\Sigma x_i) \\ &= \frac{1}{n^2}\Sigma Var(x_i) \end{aligned}$$

because $1/n$ is a constant, the x_i are independent and hence the variance of the sum is the sum of the variances. Hence

$$\begin{aligned} Var(\bar{x}) &= \frac{1}{n^2}\Sigma\sigma^2 \\ &= \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

We saw that

$$\begin{aligned} E(\bar{x}^2) &= Var(\bar{x}) + \{E(\bar{x})\}^2 \\ &= \frac{\sigma^2}{n} + \mu^2 \end{aligned}$$

So

$$\begin{aligned} E[\Sigma(x_i - \bar{x})^2] &= \Sigma E(x_i^2) - nE(\bar{x}^2) \\ &= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\ &= (n-1)\sigma^2 \end{aligned}$$

So the expected value of the sum of squares is $(n-1)\sigma^2$ and we must divide

the sum of squares by $(n - 1)$, not n, to obtain the best estimate of the variance, σ^2.

Exercise 6M

(Each branch is either true or false.)

1. The events A and B are mutually exclusive, so:

(a) Prob(A or B) = Prob(A) + Prob(B);
(b) Prob(A and B) = 0;
(c) Prob(A and B) = Prob(A) Prob(B);
(d) Prob(A) = Prob(B);
(e) Prob(A) + Prob(B) = 1.

2. The probability of a woman aged 50 having condition X is 0.20 and the probability of her having condition Y is 0.05. These probabilities are independent:

(a) The probability of her having both conditions is 0.01.
(b) The probability of her having both conditions is 0.25.
(c) The probability of her having either X, or Y, or both is 0.24.
(d) If she has condition X, the probability of her having Y also is 0.01.
(e) If she has condition X, the probability of her having Y also is 0.20.

3. The following variables follow a Binomial Distribution:

(a) number of sixes in 20 throws of a die;
(b) human weight;
(c) number of a random sample of patients who respond to a treatment;
(d) number of red cells in 1 ml of blood;
(e) proportion of hypertensives in a random sample of adult men.

4. If a coin is spun twice in succession:

(a) the expected number of tails is 1.5;
(b) the probability of two tails is 0.25;
(c) the number of tails follows a Binomial Distribution;
(d) the probability of at least one tail is 0.5;
(e) the distribution of the number of tails is symmetrical.

5. If X is a random variable, mean μ and variance σ^2:

(a) $E(X + 2) = \mu$;

(b) $Var(X + 2) = \sigma^2$;

(c) $E(2X) = 2\mu$;

(d) $Var(2X) = 2\sigma^2$;

(e) $Var(X/2) = \sigma^2/4$.

6. If X and Y are independent random variables:

(a) $Var(X + Y) = Var(X) + Var(Y)$;

(b) $E(X + Y) = E(X) + E(Y)$;

(c) $E(X - Y) = E(X) - E(Y)$;

(d) $Var(X - Y) = Var(X) - Var(Y)$;

(e) $Var(-X) = -Var(X)$.

Exercise 6E

In this exercise we shall apply some of the basic laws of probability to a practical exercise. The data are based on a life table. (We shall say more about these in Chapter 16.) Table 6E.1 shows the number of men, from a group numbering 1000 at birth, who we would expect to be alive at different ages. Thus, for example, after 10 years, we see that 959 survive and so 41 have died,

Table 6E.1. Number of men remaining alive at ten-year intervals (from *English Life Table* No. 11, Males)

Age in years (x)	Number surviving (l_x)
0	1000
10	959
20	952
30	938
40	920
50	876
60	758
70	524
80	211
90	22
100	0

at 20 years 952 survive and so 48 have died, 41 in age range 0–9 and 7 in age range 10–19.

1. What is the probability that an individual chosen at random will survive to age 10?

2. What is the probability that this individual will die before age 10? Which property of probability does this depend on?

3. What are the probabilities that the individual will survive to ages 10, 20, 30, 40, 50, 60, 70, 80, 90, 100? Is this set of probabilities a probability distribution?

4. What is the probability that an individual aged 60 years survives to age 70?

5. What is the probability that two men aged 60 will both survive to age 70? Which property of probability is used here?

6. If we had 100 individuals aged 60, how many would we expect to attain age 70?

7. What is the probability that a man dies in his second decade? (Use the fact that Prob(death in 2nd) + Prob(survives to 3rd) = Prob(survives to 2nd).)

8. What is the probability that a man dies in his 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th decades. This is a probability distribution — why? Sketch the distribution.

9. We can assume that the average number of years lived in the decade of death is 5. Thus, those who die in the 2nd decade will have an average lifespan of 15 years. The probability of dying in the 2nd decade is 0.007, i.e. a proportion 0.007 of men have a mean lifetime of 15 years. What is the mean lifetime of all men? This is the expectation of life at birth.

7. The Normal Distribution

7.1. Probability distributions for continuous variables

When we derived the theory of probability in the discrete case, we were able to say what the probability was of a random variable taking a particular value. As the number of possible values increases, the probability of a particular value decreases. For example, in the Binomial Distribution with $p = 0.5$ and $n = 2$, the most likely value, 1, has probability 0.5. In the Binomial Distribution with $p = 0.5$ and $n = 15$, the most likely values, 7 and 8, have probability 0.2, and when $n = 100$ the most likely value has probability 0.08. In such cases we are usually more interested in the probability of a range of values than one particular value.

When we come to continuous variables, such as height, the set of possible values is infinite. As we have already noted in Section 6.1, the probability of any particular value is zero. What we are interested in is the probability of the random variable taking values between certain limits rather than particular values. If the proportion of individuals in the population whose values are between the limits is p, and we choose an individual at random, the probability of choosing an individual who lies between the limits is equal to p. This comes from our definition of probability, the choice of each individual

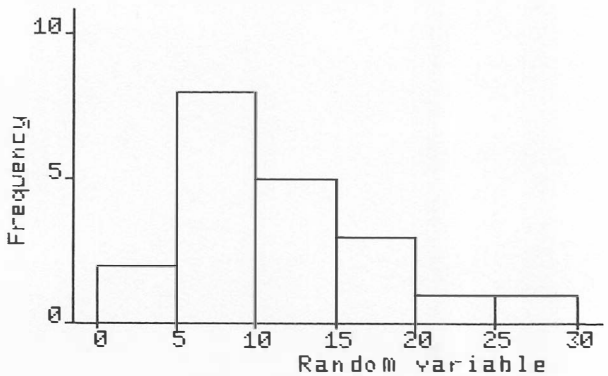


Fig. 7.1 Histogram of the frequency distribution of a random variable.

being equally likely. The problem is finding and giving a value to this proportion.

We can approach this from frequency distributions. In a frequency distribution we observe how many values in a particular sample fall within certain limits (Chapter 4). We represent this as a histogram, as in Fig. 7.1. The heights of the rectangles, representing frequency, depend on the total number in the sample and the size of the intervals as well as the shape of the distributions. To adjust for this we can develop the idea of relative or proportional frequency. Instead of considering the number of individuals falling within the interval, we find the proportion of the sample who fall within the interval. This is then the relative frequency, and is not affected by the size of the sample (Fig. 7.2). Here the heights of the rectangles represent the proportion of observations falling within certain limits. These heights now depend only on the shape of the distribution and the size of the intervals.

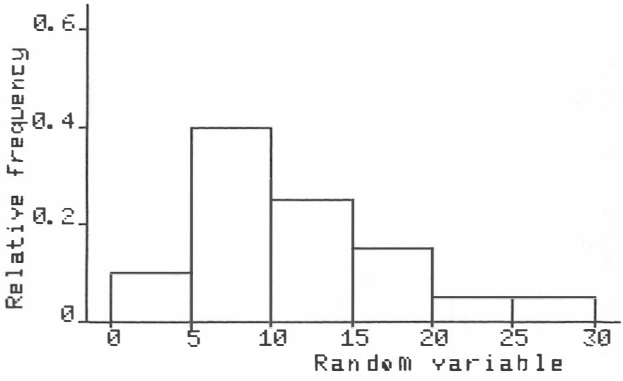


Fig. 7.2 Histogram showing relative frequency.

We can adjust for this also. To do this we move from relative frequency to *relative frequency density*. This is the proportion of observations in the interval per unit of X. Thus, when the interval size is 5, as above, the relative frequency density is the relative frequency divided by 5 (Fig. 7.3). We did this in Chapter 4, when considering histograms for distributions where the class intervals were unequal.

The relative frequency is now represented by the number of units of X multiplied by the density, which gives the area of the rectangle. Thus, the relative frequency between any two points can be found from the area under the histogram between the points. For example, to estimate the relative frequency between 10 and 20 we have the density from 10 to 15 as 0.05 and between 15 and 20 as 0.03. Hence the relative frequency is

$$0.05 \times (15 - 10) + 0.03 \times (20 - 15) = 0.25 + 0.15 = 0.40.$$

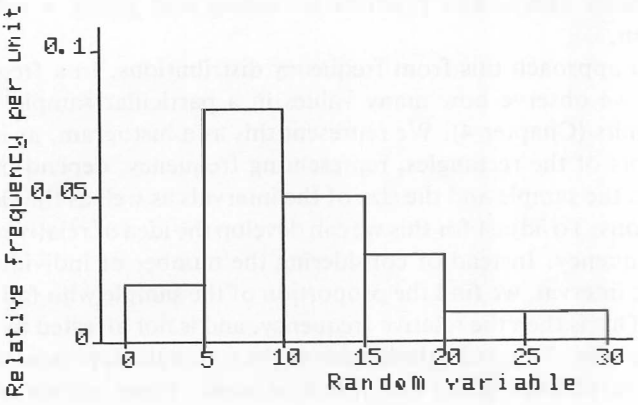


Fig. 7.3 Histogram showing relative frequency density.

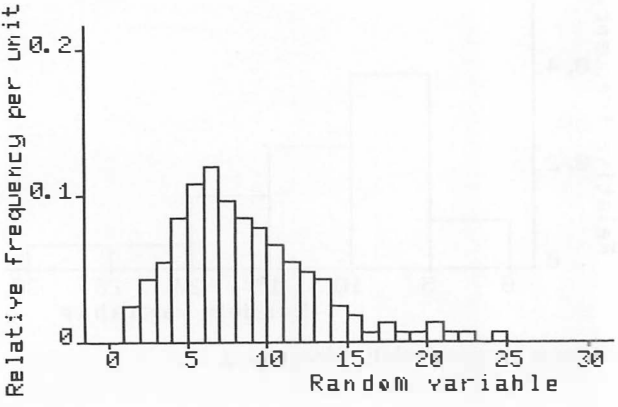


Fig. 7.4 The effect on a frequency distribution of increasing sample size.

As we take larger and larger samples, we can take smaller intervals. We get a smoother-looking histogram, as in Fig. 7.4, and as we take larger and larger samples and so smaller and smaller intervals, we get a shape very close to a smooth curve (Fig. 7.5). As the sample size approaches that of the population, which we can assume to be very large, this curve becomes the relative frequency density of the whole population. Thus, from this limiting curve we can find the proportion of observations between any two limits by finding the area under the curve, as indicated in Fig. 7.5.

If we know the equation of this curve, we can find the area under it. (Mathematically we do this by integration, but we do not need to know how to integrate to use or to understand practical statistics — all the integrals we

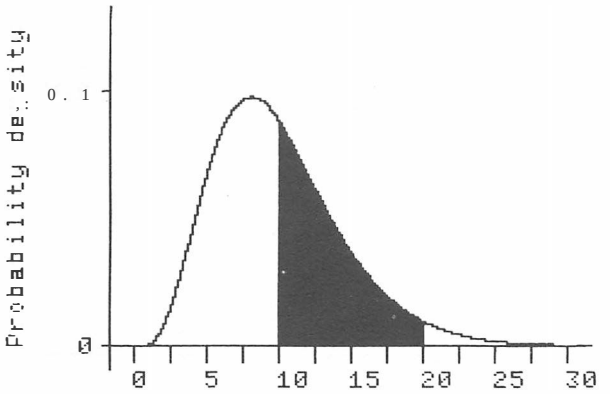


Fig. 7.5 Relative frequency density or probability density function.

need have been done and tabulated.) Now, if we choose an individual at random, the probability that X lies between given limits is equal to the proportion of individuals who fall between these limits. Hence, the relative frequency distribution for the whole population gives us the probability distribution of the variable. We call this curve the *probability density function*. To find the probability of the variable lying between any given limits, we simply find the area under the curve between the limits.

These curves have a number of general properties. For example, the total area under the curve must be one, since this is the total probability of all possible events. As was noted in Section 6.5, continuous random variables have means, variances and standard deviations defined in a similar way to those for discrete random variables and possessing the same properties. The mean will be somewhere near the middle of the curve and most of the area under the curve will be between the mean minus two standard deviations and the mean plus two standard deviations (Fig. 7.6).

The precise shape of the curve is more difficult to ascertain. There are many possible probability functions and some of these can be shown to fit simple probability models, as were the Binomial and Poisson Distributions. However, most continuous variables with which we have to deal, such as height, blood pressure, serum cholesterol, do not arise from simple, known probability situations. As a result, we do not know the probability distribution for these measurements on theoretical grounds. As we shall see, we can often find a standard distribution whose mathematical properties are known, which fits observed data well and which enables us to draw conclusions about them. Further, as sample size increases the distribution of certain statistics calculated from the data, such as the mean, become independent of the distribution of the observations themselves and follow one particular

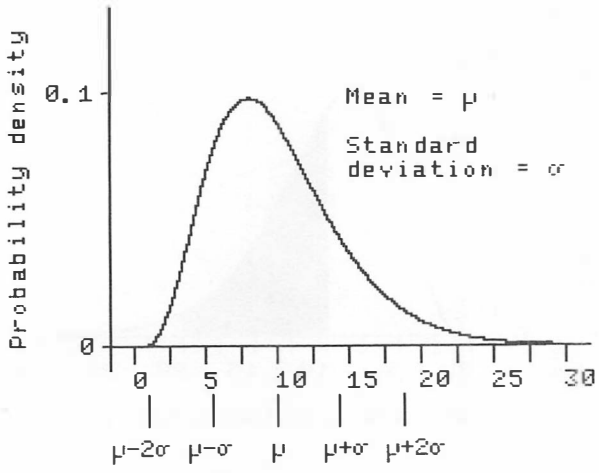


Fig. 7.6 Mean, standard deviation, and a probability density function.

distribution form, the Normal Distribution. It is on this fact that most statistical analysis depends. We shall devote the remainder of this chapter to a study of this distribution.

7.2. The Normal Distribution

The Normal Distribution, also known as the Gaussian Distribution, may be regarded as the fundamental probability distribution of statistics. The word 'normal' here is not used in its usual meaning of 'ordinary or common', or its medical meaning of 'not diseased'. The usage relates to its older meaning of 'conforming to a rule or pattern', and we shall see, the Normal Distribution is the form to which the Binomial Distribution tends as its parameter n increases. There is no implication that most variables are Normally distributed.

We shall start by considering the Binomial Distribution as n increases. Take, for example, the Binomial Distribution with $p = 0.3$. Figure 7.7 shows this distribution for six different values of n. When n is 1 or 2, the shape of the distribution is very skew and 0 is the most likely value. As n increases, the shape of the distribution changes. The most extreme possible values become less likely and the distribution becomes more symmetrical. When $n = 100$, the distribution is very close to symmetry. This happens whatever the value of p. The position of the distribution along the horizontal axis, and its spread, are still determined by p, but the shape is not. A smooth curve can be drawn which goes very close to these points. This is the Normal curve, the curve of continuous distribution which the Binomial Distribution approaches as n

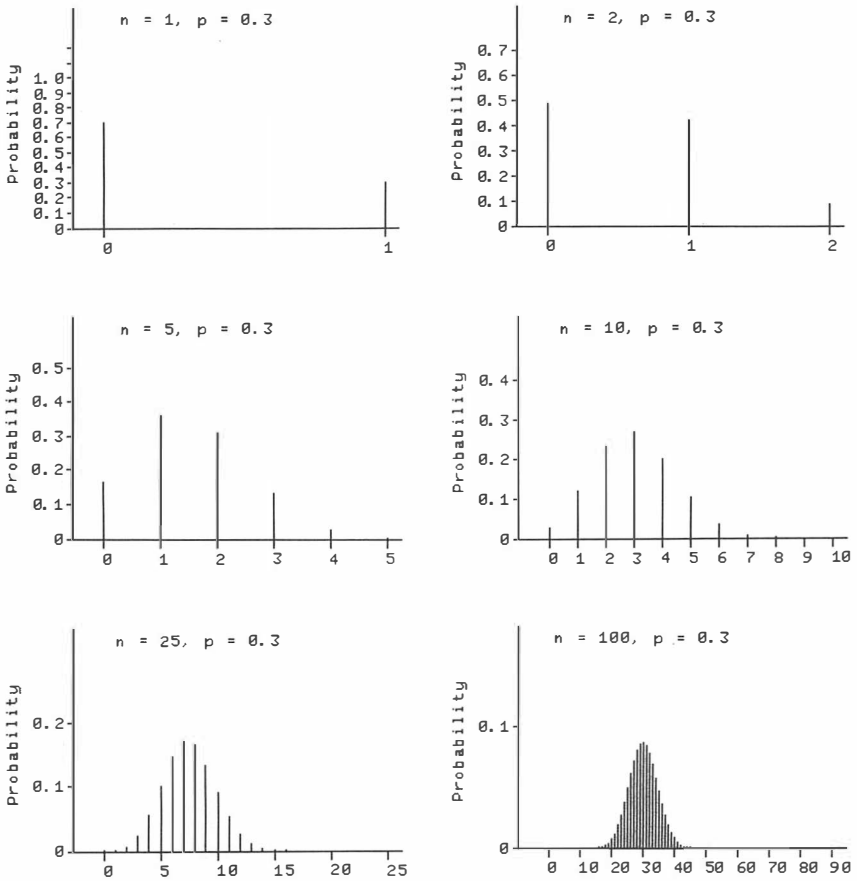


Fig. 7.7 Binomial Distribution for $p = 0.3$ and six different values of n.

increases. Any Binomial Distribution may be approximated by the Normal Distribution of the same mean and variance provided n is large enough. Figure 7.8 shows the Binomial Distributions of Fig. 7.7 with the corresponding Normal curves. From $n = 10$ onwards the two distributions are very close.

Generally, if both np and $n(1 - p)$ exceed 5 the approximation of the Binomial to the Normal Distribution is quite good enough for most practical purposes. See Section 8.4 for an application. The Poisson Distribution has the same property, as Fig. 6.4 suggests.

The Binomial variable may be regarded as the sum of n independent identically distributed random variables, each being the outcome of one trial and

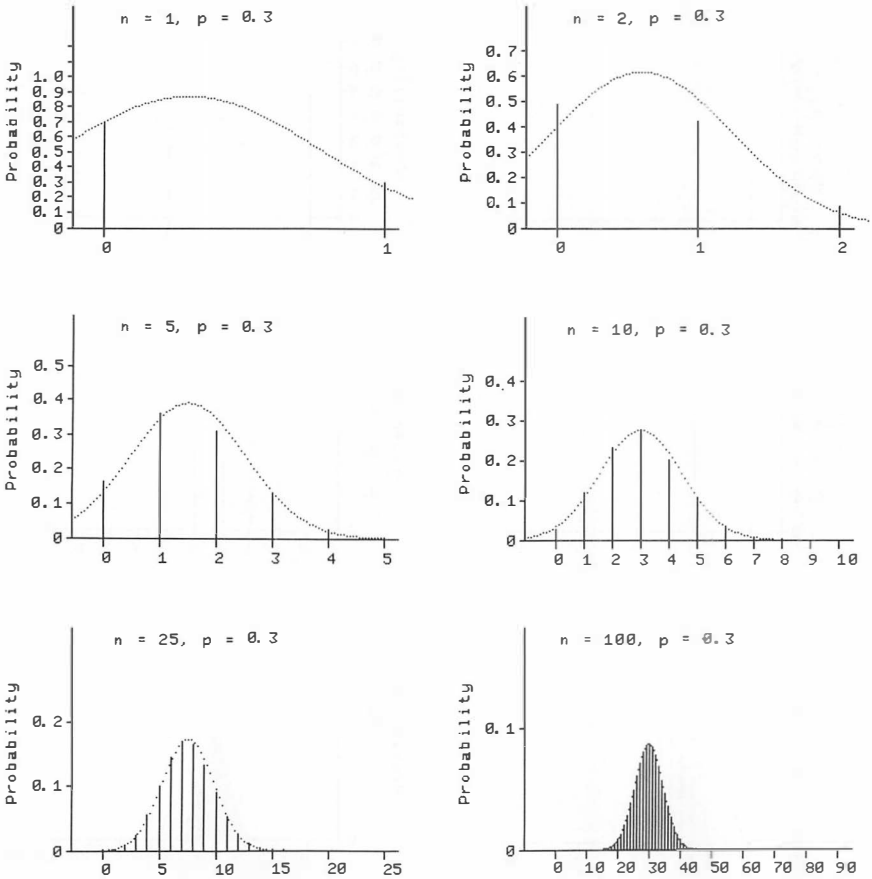


Fig. 7.8 Binomial Distributions for $p = 0.3$ and six different values of n, with corresponding normal curves.

taking value 1 with probability p. In general, if we have any series of independent, identically distributed random variables, then their sum tends to a Normal Distribution as the number of variables increases. This is known as the *central limit theorem*. As most sets of measurements are observations of such a series of random variables, this is a very important property. From it, we can deduce that the sum or mean of any large series of independent observations follows a Normal Distribution.

We can see this, for example, from the Uniform or Rectangular Distribution. This is the distribution where all values between two limits, say 0 and 1, are equally likely and no other values are possible. Observations from this arise if we take random digits from a table of random numbers such as Table

2.3, and form numbers by a decimal point followed by a string of such digits. On a microcomputer, this is usually the distribution produced by the RND(X) function in the BASIC language. Figure 7.9 shows the histogram for the frequency distribution of 500 observations from the Uniform Distribution between 0 and 1. It is quite different from the Normal Distribution. Now suppose we create a new variable by taking two Uniform variables and adding them. The histogram for 500 observations of this is also shown in Fig. 7.9. The shape of the distribution of the sum of two is quite different to the shape of the Uniform Distribution. The sum is unlikely to be close to either extreme, and observations are concentrated in the middle near the expected value. The reason for this is that to obtain a low sum, both the Uniform variables forming it must be low; to make a high sum both must be high. But we get a sum near the middle if the first is high and the second low, or the first is low and second high, or both first and second are moderate. The distribution of the sum of two is much closer to the Normal than is the Uniform Distribution itself. However, the abrupt cut off at 0 and at 2 is unlike the corresponding Normal Distribution. Figure 7.9 also shows the result of adding four Uniform variables and six Uniform variables. The similarity to the Normal Distribution increases as the number added increases and for the sum

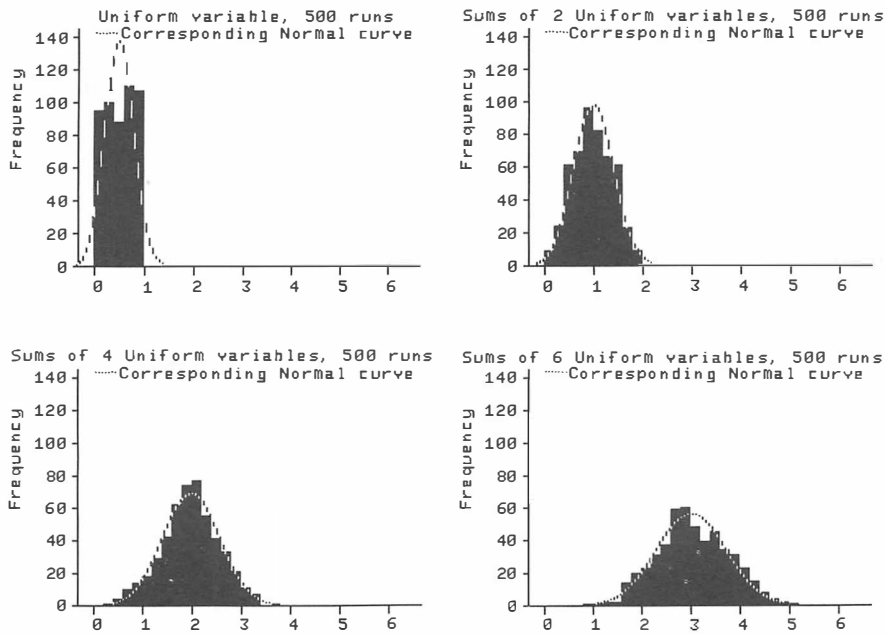


Fig. 7.9 Sums of observations from a Uniform Distribution.

of six the correspondence is so close that the distributions could not easily be told apart.

The approximation of the Binomial to the Normal Distribution is a special case of the central limit theorem. The Poisson Distribution is another. If we take a set of Poisson variables with the same rate and add them, we will get a variable which is the number of random events in a longer time interval (the sum of the intervals for the individual variables) and which is therefore a Poisson Distribution with increased mean. As it is the sum of a set of independent, identically distributed random variables it will tend towards the Normal as the mean increases. Hence as the mean increases the Poisson Distribution becomes approximately Normal. For most practical purposes this is when the mean exceeds 10. The similarity between the Poisson and the Binomial noted in Section 6.7 is a part of a more general convergence shown by many similar distributions.

7.3. Properties of the Normal Distribution

In its simplest form the equation of the Normal curve, called the Standard Normal Distribution, is

$$y = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$$

where π is the usual mathematical constant. The medical reader can be reassured that we do not need to use this forbidding formula in practice. The Standard Normal Distribution has a mean of 0, a standard deviation of 1 and a shape as shown in Fig. 7.10. The curve is symmetrical about the mean and often described as 'bell-shaped'. We can note that most of the area, i.e.

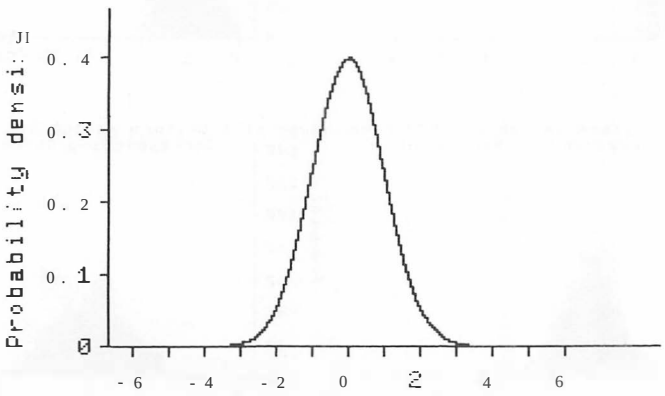


Fig. 7.10 Standard Normal Distribution.

Table 7.1. The Normal Distribution

x	$P(x)$	x	$P(x)$	x	$P(x)$
−3.0	0.001	−1.0	0.159	1.0	0.841
−2.9	0.002	−0.9	0.184	1.1	0.864
−2.8	0.003	−0.8	0.212	1.2	0.885
−2.7	0.003	−0.7	0.242	1.3	0.903
−2.6	0.005	−0.6	0.274	1.4	0.919
−2.5	0.006	−0.5	0.307	1.5	0.933
−2.4	0.008	−0.4	0.345	1.6	0.945
−2.3	0.011	−0.3	0.384	1.7	0.955
−2.2	0.014	−0.2	0.421	1.8	0.964
−2.1	0.018	−0.1	0.450	1.9	0.971
−2.0	0.023	0.0	0.500	2.0	0.977
−1.9	0.029	0.1	0.540	2.1	0.982
−1.8	0.036	0.2	0.579	2.2	0.986
−1.7	0.045	0.3	0.616	2.3	0.989
−1.6	0.055	0.4	0.655	2.4	0.992
−1.5	0.067	0.5	0.691	2.5	0.994
−1.4	0.081	0.6	0.726	2.6	0.995
−1.3	0.097	0.7	0.758	2.7	0.997
−1.2	0.155	0.8	0.788	2.8	0.997
−1.1	0.136	0.9	0.816	2.9	0.998
−1.0	0.159	1.0	0.841	3.0	0.999

The table shows the probability, P, of a Normal variable, mean 0 and variance 1 being less than x.

the probability, is between -1 and $+1$, the large majority between -2 and $+2$, and almost all between -3 and $+3$.

Although the Normal curve has many remarkable properties, it has one rather awkward one: it cannot be integrated. In other words, there is no simple formula for the probability of a Normally Distributed random variable lying between given limits. The areas under the curve can be found numerically, however, and these have been calculated and tabulated.

This mathematical intractability is shown by most of the probability distributions used in statistics, and many sets of tables are available, from a short set of those most frequently needed (e.g. Lindley and Miller 1955) to the two-volume hard-cover *Biometrika tables* (Pearson and Hartley 1969, 1972) which has 69 tables, many of which occupy over ten pages each.

Table 7.1 shows the area under the Normal curve for different values of the Normal Distribution. To be more precise, for a value, x, the table shows the area under the curve to the left of x, i.e. from minus infinity to x (Fig. 7.11). Note that half this table is not strictly necessary. We only need the half for positive x as $P(-x) + P(x) = 1$. This arises from the symmetry of the distribution. To find the probability of x lying between two values a and b, where $b > a$, we find $P(b) - P(a)$. These formulae are examples of the additive law of probability.

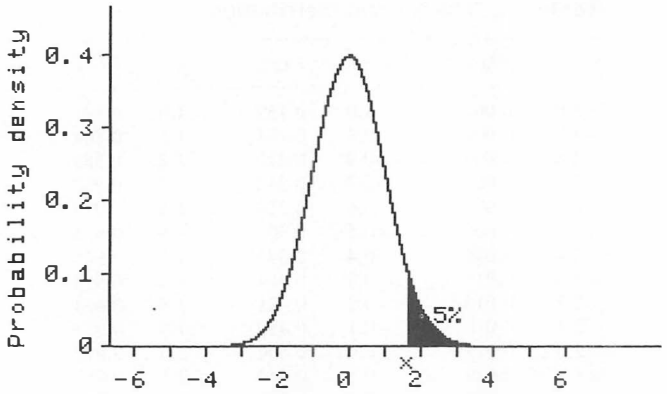


Fig. 7.11 One-sided percentage point (5 per cent) of the Standard Normal Distribution.

Table 7.1 only gives a few values of x, and much more extensive ones are available (Lindley and Miller 1955; Pearson and Hartley 1969). These values will be quite sufficient for our purposes.

There is another way of tabulating this distribution, using what is called percentage point. The *one-sided* p *percentage point* of a distribution is the value, x, such that there is a probability, p per cent, of an observation from that distribution being greater than or equal to x (Fig. 7.11). The *two-sided* p *percentage point* is the value, x, such that there is a probability, p per cent, of an observation being greater than or equal to x or less than or equal to $-x$ (Fig. 7.12). Table 7.2 shows both one sided and two sided percentage points

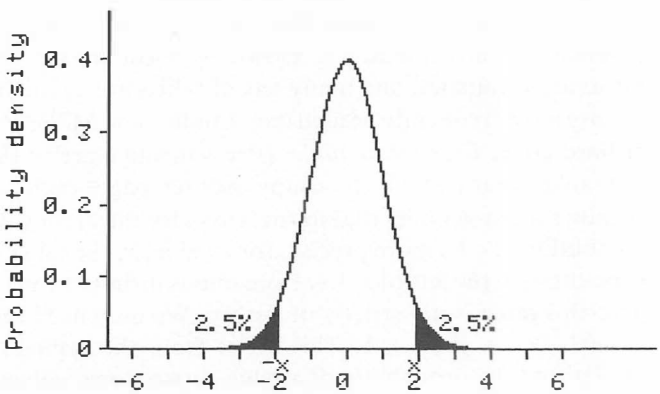


Fig. 7.12 Two-sided percentage point (5 per cent) of the Standard Normal Distribution.

Table 7.2. Percentage points of the Normal Distribution

One-sided		Two-sided	
P	*x*	*P*	*x*
50	0.00		
25	0.67	50	0.67
10	1.28		
5	1.64	10	1.64
2.5	1.96	5	1.96
1	2.33		
0.5	2.58	1	2.58
0.1	3.09		
0.05	3.29	0.1	3.29

for the Normal Distribution. The probability is quoted as a percentage because when we use percentage points we are usually concerned with rather small probabilities, such as 0.05 or 0.01, and use of the percentage form, making them 5 per cent and 1 per cent, cuts out the leading zero.

So far we have examined the Normal Distribution with mean 0 and standard deviation 1. If we add a constant μ to a Standard Normal variable, we get a new variable which has mean μ (see 6.5). Figure 7.13 shows the Normal Distribution with mean 0 and the distribution obtained by adding 1 to it together with their two-sided 5 per cent points. The curves appear identical apart from a shift along the axis. On the curve with mean 0 nearly all the probability is between -3 and $+3$. For the curve with mean 1 it is between -2 and $+4$, i.e. between the mean -3 and the mean $+3$. The probability of being a given number of units from the mean is the same for both distributions, as is also shown by the 5 per cent points.

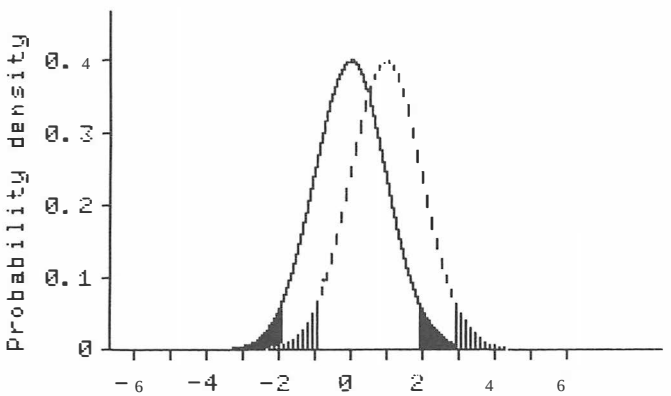


Fig. 7.13 Normal Distributions with different means, showing two-sided 5 per cent points.

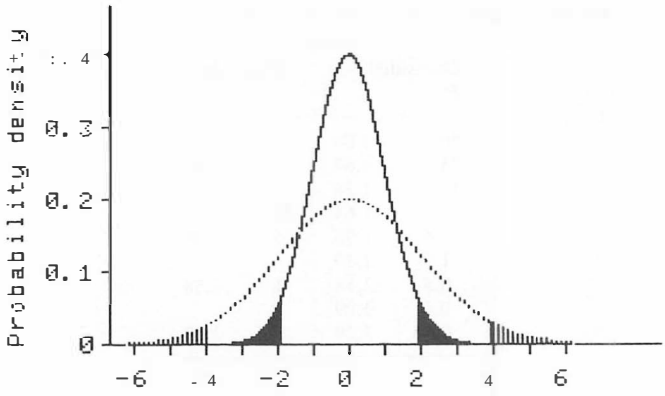


Fig. 7.14 Normal Distributions with different variances, showing two-sided 5 per cent points.

If we take a Standard Normal variable, with standard deviation 1, and multiply by a constant σ we get a new variable which has standard deviation σ. Figure 7.14 shows the Normal Distribution with mean 0 and standard deviation 1 and the distribution obtained by multiplying by 2. The curves do not appear identical. For the distribution with standard deviation 2, nearly all the probability is between -6 and $+6$, a much wider interval than the -3 and $+3$ for the standard distribution. The values -6 and $+6$ are -3 and $+3$ standard deviations. We can see that the probability of being a given number of standard deviations from the mean is the same for both distributions. This is also seen from the 5 per cent points, which represent the mean plus or minus 1.96 standard deviations in each case.

In fact, if we add μ to a Standard Normal variable and multiply by σ, we get a Normal Distribution of mean, μ, and standard deviation, σ. Tables 7.1 and 7.2 apply to it directly, if we denote by x the number of standard deviations above the mean, rather than the numerical value of the variable. Thus, for example, the upper 2.5 per cent point of a Normal Distribution with mean 10 and standard deviation 5 is found by $10 + 5 \times 1.96 = 19.8$, the value 1.96 being found from Table 7.2.

This property of the Normal Distribution, that multiplying or adding constants still gives a Normal Distribution, is not as obvious as it might seem. The Binomial does not have it, for example. Take a Binomial variable with $n = 3$, possible values 0, 1, 2, 3, and multiply by 0.5. The possible values are now 0, 0.5, 1, 1.5. Try putting these into the Binomial probability formula, and you will soon have problems with the factorial of 1.5.

We have seen that adding a constant to a Normally distributed variable gives another Normally distributed variable. If we add two Normally distributed variables together, even with different means and variances, the

sum is Normally distributed. The difference between two Normally distributed variables is also Normally distributed.

7.4. Variables which are themselves Normally distributed

So far we have discussed the Normal Distribution as it arises from sampling as the sum or limit of other distributions. However, many naturally occurring variables, such as human height, appear to follow a Normal Distribution very closely. We might expect this to happen if the variable were the result of adding variation from a number of different sources. The sort of process shown by the central limit theorem may well produce a result close to Normal. Figure 7.15 shows the distribution of height in a sample of pregnant women, and the corresponding Normal curve. The fit to the Normal curve is very good.

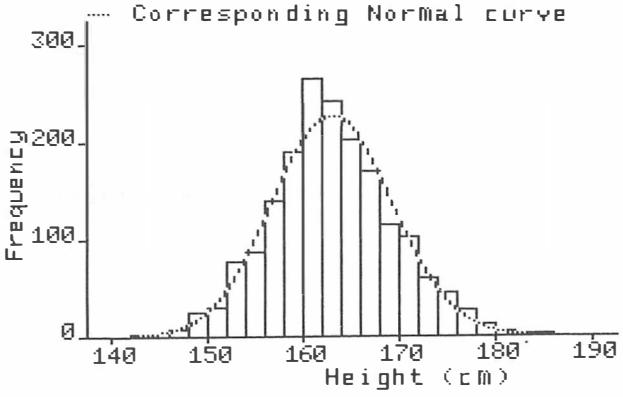


Fig. 7.15 Distribution of height in a sample of pregnant women.

If the variable we measure is the result of multiplying several different sources of variation we would not expect the result to be Normal from the properties discussed in Section 7.3, which were all based on addition of variables. However, if we take the log transformation of such a variable (see Appendix 5A) we would then get a new variable which is the sum of several different sources of variation and which may well have a Normal Distribution. This process often happens with quantities which are part of metabolic pathways, the rate at which reaction can take place depending on the concentrations of other compounds. Many measurements of blood constituents exhibit this, for example. Figure 7.16 shows the distribution of serum triglyceride measured in cord blood for 282 babies. The distribution is highly skewed and quite unlike the Normal curve. However, when we take the log

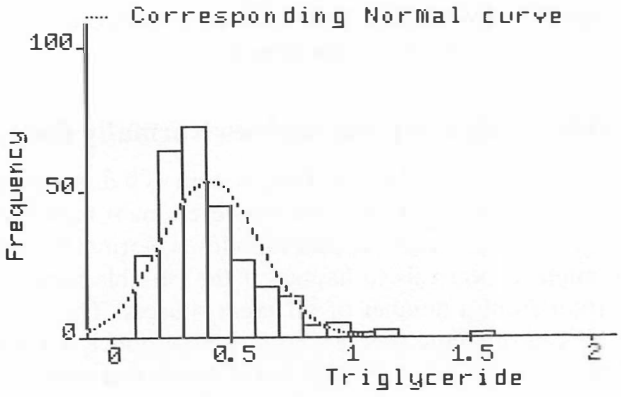


Fig. 7.16 Serum triglyceride in cord blood for 282 babies.

transformation of the triglyceride concentration, we have a remarkably good fit to the Normal Distribution (Fig. 7.17). If the logarithm of a random variable follows a Normal Distribution, the random variable itself follows a *Lognormal Distribution*.

7.5. Assessing the fit of the Normal Distribution

We often want to decide whether a sample appears Normally distributed. There are many ways of testing this and of estimating the deviation from the Normal, but most are only suited to fairly large samples. With a large sample we can inspect a histogram to see whether it looks like a Normal curve. The easiest way with a small sample is the *Normal plot*. This is a graphical

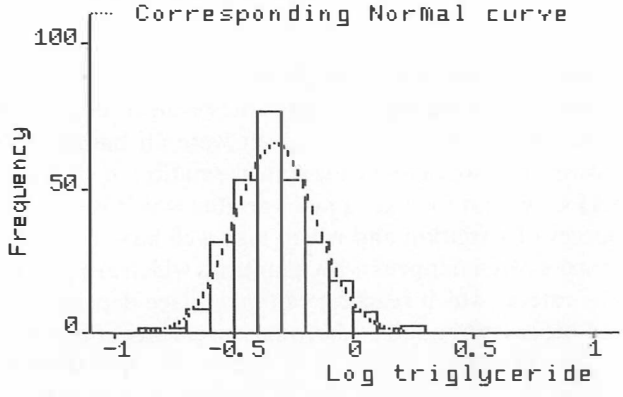


Fig. 7.17 Log transformation of serum triglyceride in cord-blood.

method, which can be done using ordinary graph paper and Table 7.1 (or a fuller version), with specially printed Normal probability paper, or using a computer. Any good general statistical package will give Normal plots; if it doesn't then it isn't a good package.

The Normal plot is a plot of the cumulative frequency distribution for the data against the cumulative frequency distribution for the Normal Distribution. To construct a Normal plot we order the data from lowest to highest. We then estimate for each observation the probability that a number from the distribution followed by the data will have a value below that observation. We can say that for n points there are $(n + 1)$ different sections of the scale, and $(n - 1)$ intervals between adjacent points and those parts of the scale below the lowest and above the highest. The probability of a number from this distribution falling between any adjacent pair of observations will be $1/n$. The probability of a number from the distribution falling below the lowest observation is $\frac{1}{2}/n$ and the probability of falling above the highest observation is also $\frac{1}{2}/n$. Note that the sum of all these probabilities is 1.0, as the events are mutually exclusive and include all possibilities. The probability of being below the lowest observation is $\frac{1}{2}/n$, of being below the second is $1\frac{1}{2}/n$, and below the ith is $(i - \frac{1}{2})/n$. There are other methods, but this is the simplest.

We then find from Table 7.1 the values of x which correspond to $P(x) = \frac{1}{2}/n$, $1\frac{1}{2}/n$, etc. For 5 points, for example, we have $P(x) = 0.1, 0.3, 0.5, 0.7, 0.9$ and $x = -1.3, -0.5, 0, 0.5, 1.3$. These are the points of the Standard Normal Distribution which correspond to the observed data. Now, if the observed data come from a Normal Distribution of mean μ and variance σ^2, the observed point should equal $\sigma x + \mu$, where x is the corresponding point of the Standard Normal Distribution. If we plot the Standard Normal points against the observed values we should get something

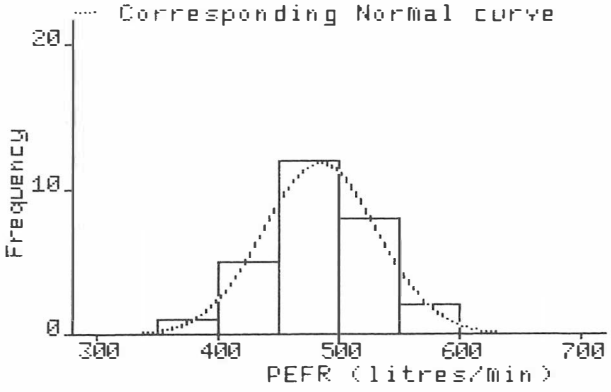


Fig. 7.18 Peak expiratory flow rate of 28 female medical students.

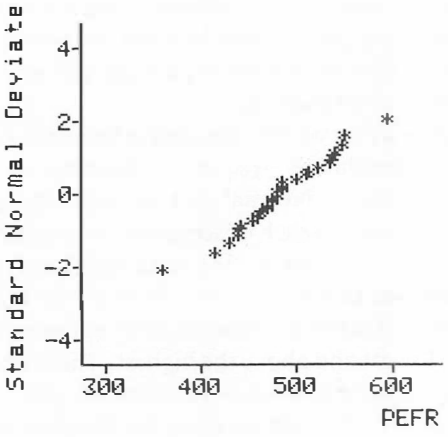


Fig. 7.19 Normal plot of the PEFR data.

close to a straight line. If the data are not from a Normal Distribution we shall get a curve of some sort. Figure 7.18 shows the distribution of peak expiratory flow rate (PEFR) in a sample of female medical students, and Fig. 7.19 shows the Normal plot. The line is quite straight and it would certainly be reasonable to use a Normal Distribution in the analysis of these data. In contrast, Fig. 7.20 shows some data obtained in a geographical study of the soft tissue tumour Kaposi's sarcoma (Bland *et al.* 1977). This distribution may not result from either a simple additive or simple multiplicative process, and so there is no reason to assume that either the Normal or Lognormal Distributions will apply. We can see that this distribution is not Normal. The Normal plot, Fig. 7.21, shows a very wavy line.

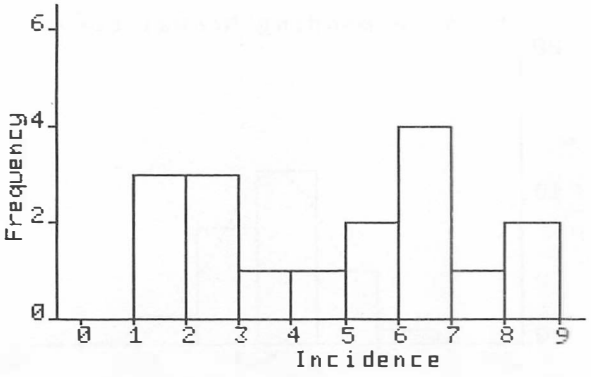


Fig. 7.20 Estimated incidence per million per year of Kaposi's sarcoma in 17 regions of mainland Tanzania (Bland et *al.* 1977).

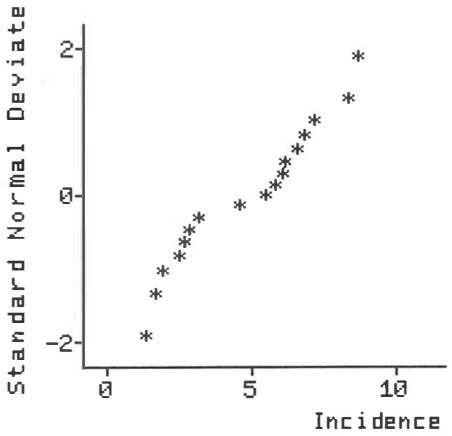


Fig. 7.21 Normal plot of Kaposi sarcoma incidence data.

The Normal plot method can be used to investigate the Normal assumption in very small samples and is a very useful check when using methods such as the *t* Distribution methods described in Chapter 10.

Appendix 7A

The Chi-squared and Student's *t* Distributions

Less mathematically inclined readers can skip this Section, but those who persevere should find that applications like chi-squared tests (Chapter 13) appear much more logical. Many probability distributions can be derived for functions of Normal variables which arise in statistical analysis. Three of these are particularly important: the Chi-squared, *t* and *F* Distributions. These have many applications, some of which we shall discuss in later chapters.

The Chi-squared Distribution is defined as follows. Suppose U is a Standard Normal variable, so having mean 0 and variance 1. Then the variable formed by U^2 follows the Chi-squared Distribution with 1 degree of freedom. If we have n such independent Standard Normal variables, U_1, U_2, . . ., U_n then the variable defined by

$$\chi^2 = \Sigma U_i^2$$

is defined to be the *Chi-squared Distribution with* n *degrees of freedom*. The letter χ is the Greek 'chi', pronounced 'ky' as in 'kite'. The distribution curves for several different numbers of degrees of freedom are shown in Fig. 7.22. The mathematical description of this curve is rather complicated,

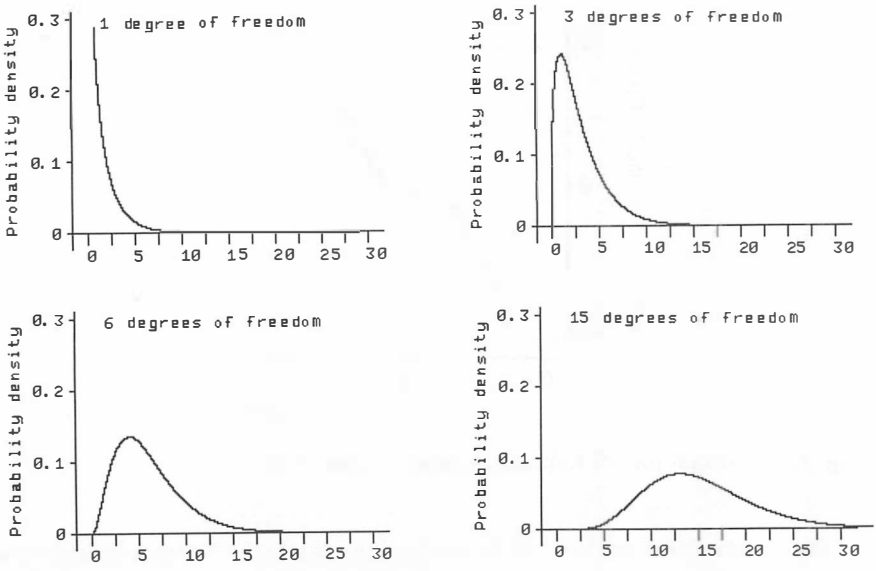


Fig. 7.22 Some Chi-squared Distributions.

but we do not need to go into this. Table 13.3 gives some percentage points of the Chi-squared Distribution.

Some properties of the Chi-squared Distribution are easy to deduce. As the distribution is the sum of n independent identically distributed random variables it tends to the Normal as n increases, following from the central limit theorem. The convergence is rather slow, however (Fig. 7.22), and the square root of chi-squared converges much more quickly.

The expected value of U^2 is of course the variance of U, the expected value of U being 0, and so $E(U^2) = 1$. The expected value of chi-squared with n degrees of freedom is thus n:

$$\begin{aligned} E(\chi^2) &= E(\Sigma U_i^2) \\ &= \Sigma E(U_i^2) \\ &= \Sigma 1 \\ &= n \end{aligned}$$

The square root of χ^2 has mean approximately equal to $\sqrt{(n - \frac{1}{2})}$ and variance approximately $\frac{1}{2}$.

The Chi-squared Distribution has a very important property. Suppose we restrict our attention to a subset of possible outcomes for the n variables U_1, . . ., U_n. The subset will be defined by those values of U_1, . . ., U_n which satisfy the equation $a_1U_1 + a_2U_2 + \ldots + a_nU_n = k$, where a_1, a_2, a_n and k are constants. (This is called a linear constraint.) Then under this restriction,

ΣU_i^2 follows a Chi-squared Distribution with $(n - 1)$ degrees of freedom. If there are m such constraints such that none of the equations can be calculated from the others, then we have a Chi-squared Distribution with $(n - m)$ degrees of freedom. This is the source of the name 'degrees of freedom'.

The proof of this is too complicated to give here, involving such mathematical abstractions as n-dimensional spheres, but its implications are very important. First, consider the sum of squares about the population mean μ of a sample of size n from a Normal Distribution, divided by σ^2. The sum of squares $\Sigma[(x_i - \mu)/\sigma]^2$ will follow a Chi-squared Distribution with n degrees of freedom, as the $(x_i - \mu)/\sigma$ have mean 0 and variance 1 and they are independent. Now suppose we replace μ by an estimate calculated from the data, $\bar{x}$. The variables are no longer independent; they must satisfy the relationship $\Sigma(x_i - \bar{x})/\sigma = 0$ and we now have $(n - 1)$ degrees of freedom. Hence

$$\frac{1}{\sigma^2}\Sigma(x_i - \bar{x})^2$$

follows a Chi-squared Distribution with $(n - 1)$ degrees of freedom. The sum of squares about the mean of any Normal sample with variance follows the distributions of a Chi-squared variable multiplied by σ^2. It therefore has expected value $(n - 1)\sigma^2$ and we divide by $(n - 1)$ to give the estimate of σ^2.

Thus, provided the data are from a Normal Distribution, not only is the sample mean Normally distributed, but the sample variance is from a Chi-squared Distribution times σ^2. Furthermore, the sample variance and sample mean are independent if, and only if, the data are from a Normal Distribution. Because the square root of the Chi-squared Distribution converges quite rapidly to the Normal, the sample standard deviation is approximately Normally distributed for $n > 20$, provided the data themselves are from a Normal Distribution.

We shall see further applications of this property when we deal with two sample t tests (Section 10.3) and chi-squared tests (Chapter 13).

Student's t Distribution with n degrees of freedom is the distribution of $U/\sqrt{\chi^2_n/n}$, where U is a Standard Normal variable and χ^2_n is independent of it and has n degrees of freedom. We shall deal with it in Chapter 10.

The F Distribution with m and n degrees of freedom is the distribution of $(\chi^2_m/m)/(\chi^2_n/n)$, the ratio of two independent χ^2 variables each divided by its degrees of freedom. This distribution is used for comparing variances. We shall not meet it in this book, but it is worth mentioning as it occurs often in slightly more advanced work.

Exercise 7M

(Each branch is either true or false.)

1. The Normal Distribution:

(a) is also called the Gaussian Distribution;

(b) is followed by many variables;

(c) is so called because it is the one which is usually followed by naturally occurring quantities;

(d) is followed by all measurements made in healthy people;

(e) is the distribution towards which the Poisson Distribution tends as its mean increases.

2. The Standard Normal Distribution:

(a) is skew to the left;

(b) has mean = 1.0;

(c) has standard deviation = 0.0;

(d) has variance = 1.0;

(e) has the median equal to the mean.

3. The PEFRs of a group of 11-year-old girls are Normally distributed with mean 300 l/min and a standard deviation 20 l/min.

(a) About 95 per cent of the girls have PEFR between 260 and 340 l/min.

(b) 50 per cent of the girls have PEFR above 300 l/min.

(c) The girls have healthy lungs.

(d) About 5 per cent of girls have PEFR below 260 l/min.

(e) All the PEFRs must be less than 340 l/min.

4. The mean of a large sample:

(a) is always greater than the median;

(b) is calculated from the formula $\Sigma x_i/n$;

(c) is from an approximately Normal Distribution;

(d) increases as the sample size increases;

(e) is always greater than the standard deviation.

5. If X and Y are independent variables which follow Standard Normal Distributions, a Normal Distribution is also followed by:

(a) $5X$;

(b) X^2;

(c) $X + 5$;

(d) $X - Y$;

(e) X/Y.

Exercise 7E

In this exercise we shall return to the blood glucose data of exercise 4E and try to decide how well they conform to a Normal Distribution.

1. From the box and whisker plot and the histogram found in exercise 4E (see 4E solution if you have not tried exercise 4E), do the blood glucose levels look like a Normal Distribution?

2. Construct a normal plot for the data. This is quite easy as they are ordered already in the stem and leaf plot of exercise 4E. Find $(i - \frac{1}{2})/n$ for $i = 1$ to 40 and obtain the corresponding cumulative normal probabilities from Table 7.1. Now plot these probabilities against the corresponding blood gucose.

3. Does the plot appear to give a straight line? Do the data follow a Normal Distribution?

8. Estimation, standard error, and confidence intervals

8.1. Sampling distributions

We have seen in Chapter 3 how samples are drawn from much larger populations. Data are collected about the sample so that we can find out something about the population. One of the things we want to do is to estimate quantities such as disease prevalence, mean blood pressure, or mean exposure to a carcinogen. We also want to know by how much these estimates might vary from sample to sample.

In Chapters 6 and 7 we saw how the theory of probability enables us to link random samples with the populations from which they are drawn. In this chapter we shall see how probability theory enables us to use samples to estimate quantities in populations, and to determine the precision of these estimates. First we shall consider what happens when we draw repeat samples from the same population. Table 8.1 shows a set of 100 random digits which we can use as the population for a sampling experiment. The distribution of the numbers in this population is shown in Fig. 8.1. The population mean is 4.7 and the standard deviation is 2.9.

The sampling experiment is done by using a suitable random sampling method to draw repeated samples from the population. In this case decimal

Table 8.1. Population of 100 random digits for a sampling experiment

9	1	0	7	5	6	9	5	8	8
1	8	8	8	5	2	4	8	3	1
2	8	1	8	5	8	4	0	1	9
1	9	7	9	7	2	7	7	0	8
7	0	2	8	8	7	2	5	4	1
1	0	5	7	6	5	0	2	1	2
6	5	5	7	4	1	7	3	3	3
2	1	6	9	4	4	7	6	1	7
1	6	3	8	0	5	7	4	8	6
8	6	8	3	5	8	2	7	2	4

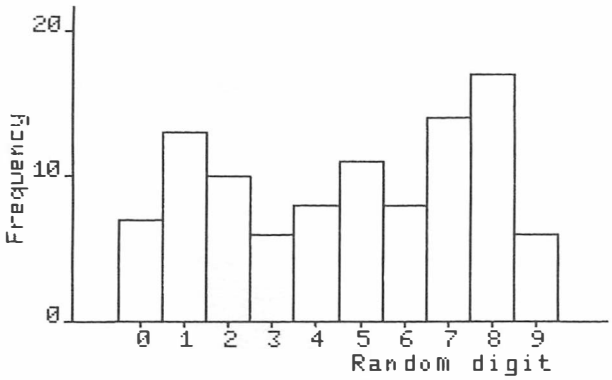


Fig. 8.1 Distribution of the population of Table 8.1.

dice were a convenient method. A sample of size four was chosen; 6, 4, 6, and 1. The mean was calculated: 17/4 = 4.25. This was repeated to draw a second sample of four numbers: 7, 8, 1, 8. Their mean is 6.00. This sampling procedure was done 20 times altogether, to give the samples and their means shown in Table 8.2.

Table 8.2. Random samples drawn in a sampling experiment

Sample	6	7	7	1	5	5	4	7	2	8
	4	8	9	8	2	5	2	4	8	1
	6	1	2	8	9	7	7	0	7	2
	1	8	7	4	5	8	6	1	7	0
Mean	4.25	6.00	6.25	5.25	5.25	6.25	4.75	3.00	6.00	2.75
Sample	7	7	2	8	3	4	5	4	4	7
	8	3	5	0	7	8	5	3	5	4
	7	8	0	7	4	7	8	1	8	6
	2	7	8	7	8	7	3	6	2	3
Mean	6.00	6.25	3.75	5.50	5.50	6.50	5.25	3.50	4.75	5.00

These samples means are not all the same. They show random variation. If we were able to draw all the 3 921 225 possible samples and calculate their means, these means themselves would form a distribution. Our 20 samples means are themselves a sample from this distribution. The distribution of all possible sample means is called the *sampling distribution* of the mean. In general, the sampling distribution of any statistic is the distribution of the values of the statistic which would arise from all possible samples.

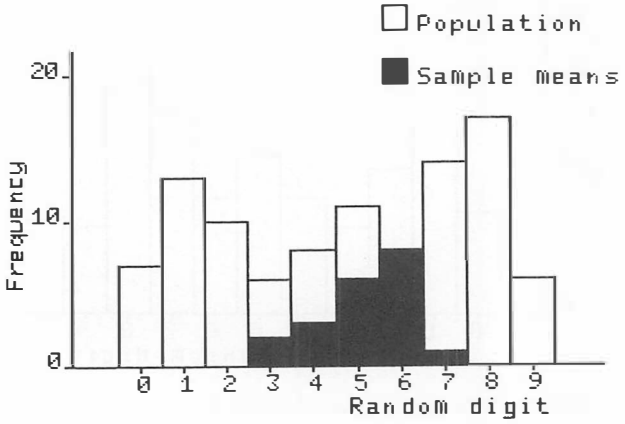


Fig. 8.2 Distribution of the population of Table 8.1 and of the sample of the means of Table 8.2.

8.2. Standard error of a sample mean

For the moment we shall consider the sampling distribution of the mean only. As our sample of twenty means is a random sample from it, we can use this to estimate some of the parameters of the distribution. The twenty means have their own mean and standard deviation. The mean is 5.1 and the standard deviation is 1.1. Now the mean of the whole population is 4.7, which is close to the mean of the samples. But the standard deviation of the population is 2.9, which is considerably greater than that of the sample means. If we plot a histogram for the sample of means (Fig. 8.2) we see that the centre of the sampling distribution and the parent population distribution are the same, but the scatter of the sampling distribution is much less.

Another sampling experiment, on a larger scale, will illustrate this further. This time our parent distribution will be the Normal Distribution with mean 0 and standard deviation 1. Figure 8.3 shows a computer simulation of 500 observations from this distribution. Figure 8.3 also shows the distribution of means from 500 random samples of size four from this population, the sample size as in Fig. 8.2. Figure 8.3 also shows the distributions of 500 means of samples of size nine and of size sixteen. In all four distributions the means are close to zero, the mean of the parent distribution. But the standard deviations are not the same. They are, in fact, approximately 1 (parent distribution); 1/2 (means of 4), 1/3 (means of 9) and 1/4 (means of 16). The relationship between the standard deviations of the parent distribution and of the sampling distribution of the mean is this: the sampling distribution of the mean has standard deviation $\sigma/\sqrt{n}$ or $\sqrt{(\sigma^2/n)}$, where σ is the standard

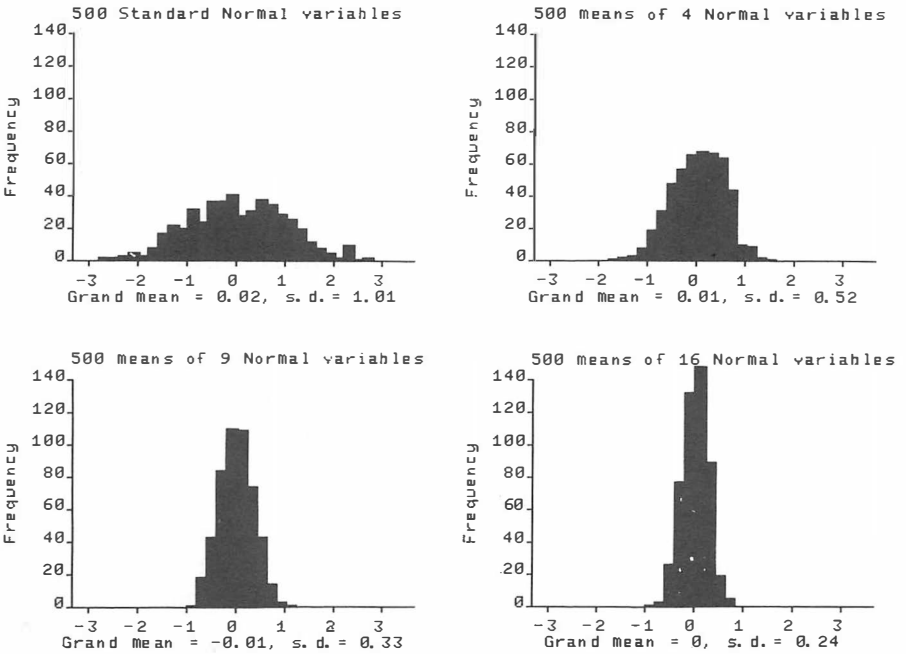


Fig. 8.3 Samples of means from a Standard Normal variable.

deviation of the parent distribution and n is the sample size. The mean of the sampling distribution is equal to the mean of the parent distribution. The actual, as opposed to simulated, distribution of the mean of four observations from a Normal Distribution is shown in Fig. 8.4. We can also see this mathematically, by a simple application of the properties of random variables as shown in Appendix 6A.2.

The sample mean is an estimate of the population mean. The standard deviation of its sampling distribution is called the *standard error* of the estimate. It provides a measure of how far from the true value the estimate is likely to be. In most kinds of estimates, the estimate is likely to be within one standard error of the true mean and unlikely to be more than two standard errors from it. We shall look at this more precisely in Section 8.3.

In most cases we do not know the true value of the population variance, σ^2, but only its estimate, s^2, which was described in Chapter 4. We can use this to estimate the standard error by $s/\sqrt{n}$. This estimate is also referred to as the standard error of the mean. It is usually clear from the context whether the standard error is the true value or estimated from the data.

When the sample size, n, is large, the sampling distribution of $\bar{x}$ tends to a Normal Distribution. Also, we can assume that s^2 is a good estimate of σ^2. So

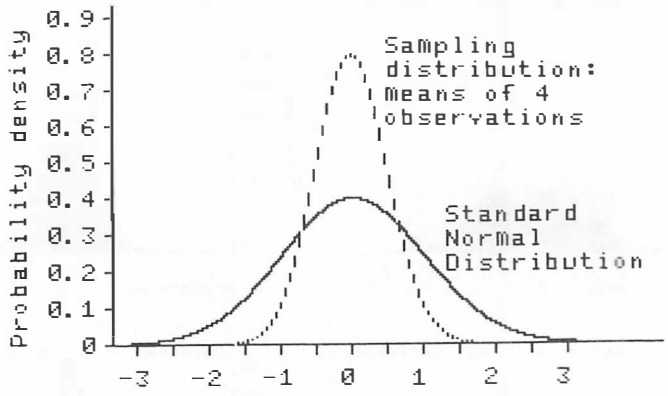


Fig. 8.4 Sampling distribution of the *mean* of four observations from a standard Normal Distribution.

for large n, $\bar{x}$ is an observation from a Normal Distribution with mean μ and standard deviation $s/\sqrt{n}$. So with probability 0.95, $\bar{x}$ is within two, or more precisely is within 1.96 standard errors of μ. With small samples we cannot assume either a Normal Distribution or, more importantly, that s^2 is a good estimate of σ^2. We shall discuss this in Chapter 10.

For an example, consider the 57 FEV1 measurements of Chapter 4. We have $\bar{x} = 4.062$ litres, $s^2 = 0.449\ 174$, $s = 0.67$ litres. Then the standard error of $\bar{x}$ is $\sqrt{(s^2/n)} = \sqrt{0.449174/57} = \sqrt{0.007\ 880} = 0.089$. The best estimate of the mean FEV1 in the population is then 4.06 litres, with standard error 0.089 litres.

The mean and standard error are often written as 4.062 ± 0.089. This is rather misleading, as the true value may be up to two standard errors from the mean with a reasonable probability. This practice is not recommended. There is often confusion between the terms 'standard error' and 'standard deviation'. This is understandable, as the standard error is a standard deviation (of the sampling distribution) and the terms are often interchanged in this context. The convention is this: we use the term 'standard error' when we measure the precision of estimates, and the term 'standard deviation' when we are concerned with the variability of populations or distributions. If we want to say how good our estimate of the mean FEV1 measurement is, we quote the standard error of the mean. If we want to say how widely scattered the FEV1 measurements are, we quote the standard deviation, s.

8.3. Confidence intervals

The estimate of mean FEV1 is a single value and so is called a *point estimate*. There is no reason to suppose that the population mean will be exactly equal

to the point estimate, the sample mean. It is likely to be close to it, however, and the amount by which it is likely to differ from the estimate can be found from the standard error. What we do is find limits between which the population mean is likely to lie, and say that we estimate the population mean to lie somewhere in the interval (the set of all possible values) between these limits. This is called an *interval estimate*.

For instance, in the above example we have a large sample, and so we can assume that the observed mean is from a Normal Distribution, and that the standard error is a good estimate of its standard deviation. We therefore expect about 95 per cent of such means to be within 1.96 standard errors of the population mean, μ. Hence, for about 95 per cent of all possible samples, the population mean must be greater than the sample mean minus 1.96 standard errors and less than the sample mean plus 1.96 standard errors. In this case we have, with probability 0.95 or 95 per cent:

$$4.062 - 1.96 \times 0.089 < \mu < 4.062 + 1.96 \times 0.089$$
$$3.89 < \mu < 4.24$$

or

$$3.9 < \mu < 4.2 \text{ litres}$$

rounding to two significant figures. The values 3.9 and 4.2 are called the 95 per cent confidence limits for the estimate, and the interval 3.9–4.2 is called the 95 per cent confidence interval. We may define a *p per cent confidence interval* as being a part of the measurement scale in which there is a probability of p per cent that the estimated quantity lies. The *confidence limits* are the ends of the confidence interval.

In this example, the sampling distribution of the mean is Normal and its standard deviation is well estimated because the sample is large. This is not always true and although it is usually possible to calculate confidence intervals for an estimate they are not all quite as simple as this. We shall look at the mean of a small sample in Chapter 10.

There is no necessity for the confidence interval to have a probability of 95 per cent. For example, we can also calculate 99 per cent confidence limits. From Table 7.2 we find that the upper 0.5 per cent point of the Standard Normal Distribution is 2.58, so the probability of a Standard Normal deviate being above 2.58 or below -2.58 is 1 per cent and the probability of being within these limits is 99 per cent. The 99 per cent confidence limits for the mean FEV1 are therefore, $4.062 - 2.58 \times 0.089$ and $4.062 + 2.58 \times 0.089$, i.e. 3.8 and 4.3. These give a wider interval than the 95 per cent limits, as we would expect since the mean is more likely to be included. The probability we choose for a confidence interval is thus a compromise between the desire to include the estimated population parameter and the desire to avoid parts of scale where there is a low probability that the mean will be

found. For most purposes, 95 per cent confidence intervals have been found to be satisfactory.

8.4. Standard error of a proportion

The standard error of a proportion estimate can be calculated in the same way. Suppose the proportion of individuals who have a particular condition in a given population is p, and we take a random sample of size n, the number observed with the condition being r. Then the estimated proportion is r/n. We have seen (Chapter 6) that r comes from a Binomial Distribution with mean np and variance $np(1-p)$. Provided n is large, this distribution is approximately Normal. So r/n, the estimated proportion, is Normally distributed with mean given by $np/n = p$, and variance given by

$$\begin{aligned}\text{Var}\left(\frac{r}{n}\right) &= \frac{1}{n^2}\text{Var}(r), \text{ since } n \text{ is constant,} \\ &= \frac{1}{n^2} np(1-p) \\ &= \frac{p(1-p)}{n}\end{aligned}$$

and the standard error is

$$\sqrt{\frac{p(1-p)}{n}}$$

We can estimate this by replacing p by r/n.

For example, in a survey of a random sample of first-year secondary schoolchildren in Derbyshire, 118 out of 2837 boys said that they usually coughed first thing in the morning. This gave a prevalence estimate of $118/2837 = 0.0416$, with standard error $\sqrt{0.0416 \times (1-0.0416)/2837} = 0.0038$. The sample is large so we can assume that the estimate is from a Normal Distribution and that the standard error is well estimated. The 95 per cent confidence interval for the prevalence is thus $0.0416 - 1.96 \times 0.0038$ to $0.0416 + 1.96 \times 0.0038 = 0.034$ to 0.049. Even with this fairly large sample the estimate is not very precise.

The standard error of the proportion is only of use if the sample is large enough for the Normal approximation to apply. A rough guide to this is that np and $n(1-p)$ should both exceed 5. This is usually the case when we are concerned with straightforward estimation.

8.5. Standard error of the difference between two means

In many studies we are much more interested in the difference between two parameters than in their absolute value. These could be means, proportions,

the slopes of lines, and many other statistics. This is usually straightforward if the parameters are estimated from two independent samples. It can be more difficult if the samples are matched or are the same.

For the difference between any two parameters estimated from independent samples the standard error can be found as follows. Suppose the estimates are x_1 and x_2, with standard errors se_1 and se_2, and hence sampling variances se_1^2 and se_2^2. The difference between the parameters is estimated by $x_1 - x_2$. Then

$$\begin{aligned}\text{Var}(x_1 - x_2) &= \text{Var}(x_1) + \text{Var}(x_2), \text{ because } x_1 \text{ and } x_2 \text{ are independent,} \\ &= se_1^2 + se_2^2\end{aligned}$$

and standard error of $x_1 - x_2$

$$= \sqrt{se_1^2 + se_2^2}$$

If we are comparing two means $\bar{x}_1$ and $\bar{x}_2$ from samples size n_1 and n_2 with variance estimates s_1^2 and s_2^2, the standard errors of the means will be $\sqrt{(s_1^2/n_1)}$ and $\sqrt{(s_2^2/n_2)}$. The standard error of the difference is then

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Provided the samples are large enough for the sampling distributions of means to be Normal and for the variances to be well estimated, say both n_1 and n_2 greater than 30, the 95 per cent confidence interval for the difference will be

$$\bar{x}_1 - \bar{x}_2 - 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ to } \bar{x}_1 - \bar{x}_2 + 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For an example, in a study of respiratory symptoms in schoolchildren we wanted to know whether children reported by their parents to have respiratory symptoms had worse lung function than children who were not reported to have symptoms. Ninety-two children were reported to have cough during the day or at night, and their mean PEFR was 294.8 litre/min with standard deviation 57.1 litre/min. The number of children not reported to have the symptom was 1643, and their mean PEFR was 313.6 litre/min with standard deviation 55.2 litre/min. We thus have two large samples, and can apply the Normal Distribution. We have

$$n_1 = 92 \qquad \bar{x}_1 = 294.8 \qquad s_1 = 57.1 \qquad se_1 = \sqrt{\frac{57.1^2}{92}}$$

$$n_2 = 1643 \qquad \bar{x}_2 = 313.6 \qquad s_2 = 55.2 \qquad se_2 = \sqrt{\frac{55.2^2}{1643}}$$

The difference between the two groups is

$$\bar{x}_1 - \bar{x}_2 = 294.3 - 313.6$$
$$= -18.8$$

The standard error of the difference is

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{57.1^2}{92} + \frac{55.2^2}{1643}}$$
$$= 6.11$$

The 95 per cent confidence limits for the difference are thus $-18.8 - 19.6 \times 6.11$ and $-18.8 + 1.96 \times 6.11$, i.e. -6.8 and -30.8 litre/min. This confidence interval does not include zero, so we have good evidence that there is a difference between mean lung function in these two groups of children. The difference itself is not very well estimated, however. It could be anything from 7 to 31 litre/min lower in children with the symptom.

8.6. Standard error of the difference between two proportions

The argument applied to the comparison of two means works equally well for the comparison of two proportions. Suppose we have two proportions estimated by p_1 and p_2 obtained from independent samples size n_1 and n_2. We want the standard error of the difference, estimated by $p_1 - p_2$. The standard errors of the two proportions are

$$se(p_1) = \sqrt{\frac{p_1(1-p_1)}{n_1}} \quad se(p_2) = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

The standard error of the difference is given by

$$se(p_1 - p_2) = \sqrt{se(p_1)^2 + se(p_2)^2}$$
$$= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Provided the conditions of Normal approximation are met (see Section 8.4) we can find a confidence interval for the difference in the usual way.

For example, in a study of respiratory disease in childhood, we wanted to know whether children with bronchitis in infancy get more respiratory symptoms in later life than others. We had 273 children with a history of bronchitis before age 5 years, 26 of whom were reported to have day or night cough at age 14. We had 1046 children with no bronchitis before age 5 years, 44 of whom were reported to have day or night cough at age 14.

Bronchitis	*No bronchitis*
$n_1 = 273$	$n_2 = 1046$
$p_1 = 26/273 = 0.095\,24$	$p_2 = 44/1046 = 0.042\,07$

The difference is

$$p_1 - p_2 = 0.095\ 24 - 0.042\ 07 = 0.053\ 17$$

The standard error of the difference is

$$\begin{aligned}
&\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\
&= \sqrt{\frac{0.095\ 24 \times (1 - 0.095\ 24)}{273} + \frac{0.042\ 07 \times (1 - 0.042\ 07)}{1046}} \\
&= \sqrt{0.000\ 315\ 638\ 6 + 0.000\ 038\ 527\ 8} \\
&= \sqrt{0.000\ 354\ 166\ 5} \\
&= 0.0188
\end{aligned}$$

The 95 per cent confidence interval for the difference is

$$\begin{aligned}
&0.053\ 17 - 1.96 \times 0.0188 \text{ to } 0.053\ 17 + 1.96 \times 0.0188 \\
&= 0.016 \text{ to } 0.090
\end{aligned}$$

Although the difference is not very well estimated, it is well away from zero and gives us clear evidence that children with bronchitis reported in infancy are more likely than others to be reported to have respiratory symptoms in later life. The data on lung function in Section 8.5 give us some reason to suppose that this is not entirely due to response bias (Chapter 3). As in Section 8.4, the standard error of the difference between two proportions is not very useful for small samples.

8.7. Standard error of a sample standard deviation

In most applications we are interested in the estimation of the standard deviation as a means to an end rather than because we want to know it for direct application. We do need to know how precisely it is estimated if we want to use it to find a clinical reference range or normal range (Section 15.5). Unlike that of the sample mean, $\bar{x}$, the standard error of the sample standard deviation, s, depends on the distribution of the observations themselves. We have seen in Appendix 7A that provided the observations come from a Normal Distribution, $(n-1)s^2/\sigma^2$ is from a Chi-squared Distribution with $(n-1)$ degrees of freedom. The square root of this Chi-squared Distribution is approximately Normal with variance 1/2 if n is large enough, so $\sqrt{(n-1)s^2/\sigma^2}$ is approximately Normally distributed with variance 1/2. Hence s is approximately Normally distributed with variance $\sigma^2/2(n-1)$. The standard error of s is thus $\sqrt{[\sigma^2/2(n-1)]}$, estimated by $\sqrt{[s^2/2(n-1)]}$. This is only true when the observations themselves are from a Normal Distribution.

8.8. Sample size for an estimate

We can use the concepts of standard error and confidence interval to help decide how many subjects should be included in a sample. If we want to estimate some parameter of a population and we know how the standard error is related to the sample size then we can calculate the sample size required to give a confidence interval with the desired size. The difficulty is that the standard error may also depend either on some other parameter of the population, such as the standard deviation, or on the parameter we wish to estimate, as for a proportion. We must estimate these either from data already available, or carry out a pilot study to obtain a rough estimate. The calculation of sample size can only be approximate anyway, so these parameter estimates used in it need not be precise.

For example, we wish to estimate the mean serum cholesterol in a population of men. We note that other workers have reported serum cholesterol to have a standard deviation, of about 40 mg/100 ml. We therefore expect the standard error of the mean to be

$$\sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{40^2}{n}} = \frac{40}{\sqrt{n}}$$

We can set the size of standard error we want and choose the sample size to achieve this. We might decide that a standard error of 5 is what we want, i.e. for the estimate to be within $2 \times 5 = 10$ mg/100 ml of the true value.

Then

$$se = \frac{40}{\sqrt{n}}$$

$$n = \frac{40^2}{(se)^2}$$

$$= \frac{40^2}{5^2}$$

$$= 64$$

We can also see what the standard error would be for different values of n:

n	10	20	50	100	200	500
standard error	13	8.9	5.7	4.0	2.8	1.8

So that if we had a sample size of 200, we would have a probability of 0.95 of being within 5.6 units of the mean (two standard errors), whereas with a sample of 50 we could only be confident of being within 11.4 units of the mean. If the maximum sample we could use is 100, and we need to be within 5 units of the true value, there is no point in proceeding.

When we wish to estimate a proportion we have a further problem. The standard error depends on the very quantity which we wish to estimate. We must guess the proportion first. For example, suppose we wish to estimate the prevalence of a disease, which we suspect to be about 2 per cent, to the nearest 1 per 1000. The value of p is about 0.02 and we want 95 per cent confidence intervals to be 0.001 on either side. So the standard error must be half this, 0.0005.

$$0.0005 = \sqrt{\frac{0.02 \times (1 - 0.02)}{n}}$$

$$n = \frac{0.02 \times (1 - 0.02)}{0.0005^2}$$

$$= 78\ 400$$

The accurate estimation of very small proportions requires very large samples! This is a rather extreme example and we do not usually need to estimate proportions with such accuracy. A wider confidence interval, obtainable with a smaller sample, is usually acceptable.

If we can only afford a sample size of 1000, what will be the standard error?

$$\sqrt{\frac{0.02 \times (1 - 0.02)}{1000}} = 0.0044$$

The 95 per cent confidence interval would thus be the observed proportion ± 0.009. We would expect the true value to be within about 50 per cent of the estimate. If this accuracy were sufficient for the purpose we could proceed.

These estimates of sample size are based on the assumption that the sample is large enough to use the Normal Distribution. If a very small sample is indicated it will be inadequate and other methods must be used which are beyond the scope of this book. Sample size for comparisons are usually estimated rather differently. The size of the confidence interval for a difference is not the important thing, but whether it contains zero. We shall discuss this in Chapter 9.

Exercise 8M

(Each branch is either true or false.)

1. The standard error of the mean of a sample:

(a) measures the variability of the observations;

(b) is the accuracy with which each observation is measured;

(c) is a measure of how far the sample mean is likely to be from the population mean;

(d) is proportional to the number of observation;

(e) is greater than the estimated standard deviation of the population.

2. 95 per cent confidence limits for the mean estimated from a set of observations:

(a) are limits between which, in the long run, 95 per cent of observations fall;

(b) are a way of measuring the precision of the estimate of the mean;

(c) are limits within which the sample mean falls with probability 0.95;

(d) are limits which exclude the population mean with probability 0.05;

(e) are a way of measuring the variability of a set of observations.

3. If the size of a random sample were increased, we would expect:

(a) the mean to decrease;

(b) the standard error of the mean to decrease;

(c) the standard deviation to decrease;

(d) the sample variance to increase;

(e) the degrees of freedom for the estimated variance to increase.

4. The prevalence of a condition in a population is 0.1. If the prevalence is estimated repeatedly from samples of size 100, these estimates will form a distribution which:

(a) is a sampling distribution;

(b) will be approximately Normal;

(c) will have mean = 0.1;

(d) will have variance = 9;

(e) will be Binomial.

5. It is necessary to estimate the mean FEV1 by drawing a sample from a large population. The accuracy of the estimate will depend on:

(a) the mean FEV1 in the population;

(b) the number in the population;

(c) the number in the sample;

(d) the way the sample is selected;

(e) the variance of FEV1 in the population.

Exercise 8E

Table 8E.1 shows data from a study of plasma magnesium in diabetics. The subjects were all patients attending one out-patient diabetes clinic over a five-month period.

Table 8E.1. Plasma magnesium levels for diabetics of different treatment regimes (Mather *et al.* 1979)

Treatment	Number of patients	Plasma magnesium (mmol/l) mean	s.d.
Insulin	227	0.719	0.068
All non-insulin therapy	352	0.748	0.070
Oral hypoglycaemic therapy	225	0.744	0.070
Dietary restriction alone	127	0.756	0.070

N.B. Fifteen patients whose blood samples could not be analysed and three receiving both insulin and oral hypoglycaemic drugs have been excluded.

1. Find the standard errors of the mean plasma magnesium for each group.

2. Find the standard error of the difference in mean plasma magnesium between patients on oral hypoglycaemic therapy and patients on dietary restriction alone. Find a 95 per cent confidence interval for the difference.

3. Find the standard error of the difference between mean plasma magnesium in insulin-treated and non-insulin-treated patients. Find a 95 per cent confidence interval for this difference.

4. What can be concluded about the relationship between therapy, which is determined by type or severity of diabetes, and plasma magnesium level?

5. How many patients would we need in a group to estimate the mean plasma magnesium to within 1 per cent?

9. Significance tests

9.1. Testing a hypothesis

In Chapter 8 we dealt with estimation and the precision of estimates. This is one form of statistical inference, the process by which we use samples to draw conclusions about the populations from which they are taken. In this chapter we will introduce a different form of inference, the significance test.

A significance test enables us to measure the strength of evidence which the data supplies for or against some proposition of interest. For example, consider the cross-over trial of pronethalol for the treatment of angina, described in Chapter 2. Table 9.1 shows the results of the trial, the number of attacks over four weeks on each treatment. These 12 patients are a sample from the population of all patients. Would the other members of this population experience fewer attacks while using pronethalol? We can see that the number of attacks is highly variable from one patient to another, and it is quite possible that this is true from one occasion to another as well. So it could be that some patients would have fewer attacks while on pronethalol than while on placebo quite by chance. In a significance test, we ask whether the difference observed was small enough to have occurred by chance. If it were so, then the evidence in favour of there being a difference between the treatment periods would be weak. On the other hand, if the difference were

Table 9.1. Results of a trial of pronethalol for the treatment of angina pectoris (Pritchard *et al.* 1963)

Number of attacks while on	
placebo	pronethalol
71	29
323	348
8	1
14	7
23	16
34	25
79	65
60	41
2	0
3	0
17	15
7	2

much larger than we would expect due to chance the evidence in favour of a real difference would be strong.

To carry out the test of significance we suppose that there is no difference between the two treatment periods. We hypothesize that there is no difference between the periods. The hypothesis of 'no difference' or 'no effect' is called the *null hypothesis*. We compare this with the alternative hypothesis of a difference between the treatments, in either direction. We do this by finding the probability of getting data as extreme as those observed if the null hypothesis were true. If this probability is large the data are consistent with the null hypothesis; if it is small the data are unlikely to have arisen if the null hypothesis were true and the evidence is in favour of the alternative hypothesis.

9.2. An example: the sign test

We now find a way of testing this null hypothesis. An obvious start is to consider the differences between the number of attacks on the two treatments for each patient, as in Table 9.2. Now, if the null hypothesis were true then differences in number of attacks would be just as likely to be positive as negative; they would be random. The probability of a change being negative would be equal to the probability of it becoming positive, 0.5. Then the number of negatives would be an observation from a Binomial Distribution with $n = 12$ and $p = 0.5$. (If there were any subjects who had the same number of attacks on both regimes we would omit them, as they provide no information about the direction of any difference between the treatments. In this test, n is the number of subjects for whom there is a difference, one way or the other.)

Table 9.2. Differences between numbers of attacks of angina while on placebo and pronethalol

Number of attacks while on placebo	pronethalol	Difference placebo–pronethalol	Sign of difference
71	29	42	+
323	348	−25	−
8	1	7	+
14	7	7	+
23	16	7	+
34	25	9	+
79	65	14	+
60	41	19	+
2	0	2	+
3	0	3	+
17	15	2	+
7	2	5	+

If the null hypothesis were true, what would be the probability of getting an observation from this distribution as extreme as the value we have actually observed? The expected number of negatives would be $np = 6$. What is the probability of getting a value as far from this as is that observed?

The number of negative differences is 1. The probability of getting 1 negative change is

$$\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r} = \frac{12!}{1!11!} \times 0.5^1 \times 0.5^{11}$$
$$= 12 \times 0.5^{12}$$
$$= 0.002\ 93$$

This is not a likely event in itself. However, we are interested in the probability of getting a value as far from the expected value, $np = 6$, as is the observed value 1. Clearly 0 is further and must be included. The probability of no negative changes is

$$\frac{12!}{0!12!} \times 0.5^0 \times 0.5^{12} = 0.000\ 24$$

So the probability of one or fewer negative changes is $0.002\ 93 + 0.000\ 24 = 0.003\ 17$. We said that the alternative hypothesis was that there was a difference in either direction. We must, therefore, consider the probability of getting a value as extreme on the other side of the mean, that is, 11 or 12 negatives. In other words we want to know the probability of the number of negatives being at least as far from its expected value as that observed. A sketch of the distribution should make this clearer (Fig. 9.1).

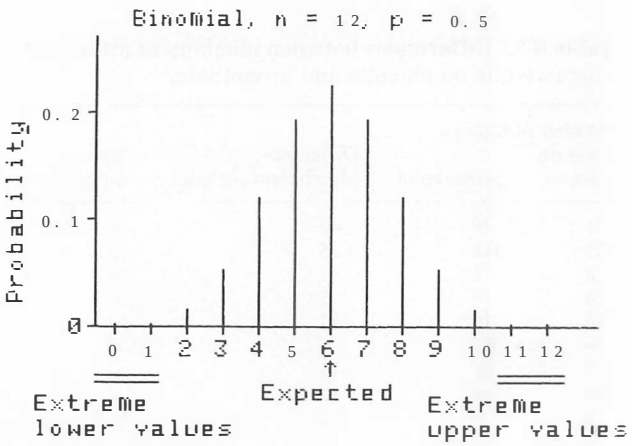


Fig. 9.1 Extreme values of the test statistic in a sign test.

The probability of 11 or 12 negatives is

$$\frac{12!}{11!1!} \times 0.5^{11} \times 0.5^{1} + \frac{12!}{12!0!} \times 0.5^{12} \times 0.5^{0}$$
$$= 0.002\,93 + 0.000\,24$$
$$= 0.003\,17$$

Hence, the probability of getting as extreme a value as that observed, in either direction, is 0.003 17 + 0.003 17 = 0.006 34. This means that if the null hypothesis were true we would have data which are so extreme that the probability of them arising by chance is 0.006, less than one in a hundred.

Thus, we would have observed a very unlikely event if the null hypothesis were true. This means that the data are not consistent with null hypothesis, so we can conclude that there is strong evidence in favour of a difference between the treatment periods. (Since this was a double-blind randomized trial, it seems reasonable to suppose that this was caused by the activity of the drug.)

This is a test of significance, the *sign test*. The number of negative differences is called the *test statistic*.

9.3. General principles of significance tests

The general procedure for a significance test is as follows:

1. Set up a null hypothesis and its alternative.
2. Find the value of the test statistic.
3. Refer the value of the test statistic to a known distribution which it would follow if the null hypothesis were true.
4. Find the probability of a value of the test statistic arising which is as or more extreme than that observed.
5. Conclude that the data are consistent or inconsistent with the null hypothesis.

We shall deal with several different significance tests in this and subsequent chapters. We shall see that they all follow this pattern.

If the data are not consistent with the null hypothesis, the difference is said to be *statistically significant*. If the data do not support the null hypothesis, it is sometimes said that we reject the null hypothesis, and if the data are consistent with the null hypothesis it is said that we accept it. Such an 'all or nothing' decision-making approach is seldom appropriate in medical research. It is preferable to think of the significance test probability as an index of the strength of evidence against the null hypothesis.

9.4. Significance levels

We must still consider the question of how small is small. A probability of 0.006, as in the example above, is clearly small and we have a quite unlikely event. But what about 0.06, or 0.1? Suppose we take a probability of 0.01 or less as constituting reasonable evidence against the null hypothesis. If the null hypothesis is true, we shall make a wrong decision one in a hundred times. Deciding against a true null hypothesis is called an *error of the first kind*. We get an *error of the second kind* if we decide in favour of a null hypothesis which is in fact false. Now the smaller we demand the probability be before we decide against the null hypothesis, the larger the observed difference must be, and so the more likely we are to miss real differences. By reducing the risk of an error of the first kind we increase the risk of an error of the second kind.

The conventional compromise is to say that differences are significant if the probability is less than 0.05. This is a reasonable guideline, but should not be taken as some kind of absolute demarkation. There is not a great difference between probabilities of 0.06 and 0.04, and they surely indicate similar strength of evidence. It is better to regard probabilities around 0.05 as providing some evidence against the null hypothesis, which increases in strength as the probability falls. If we decide that the difference is significant, the probability is sometimes referred to as the *significance level*.

9.5. One- and two-sided tests of significance

In the above example, the alternative hypothesis was that there was a difference in one or other direction. This is called a *two-sided test*, because we are interested in extreme values in both directions. It would have been possible to have the alternative hypothesis that there was a decrease in the pronethalol direction, in which case the null hypothesis would be that the number of attacks on the placebo was less than or equal to the number on pronethalol. This would give $p = 0.003\ 17$, and of course, a higher significance level than the two sided test. This would be a *one-sided test* (Fig. 9.2). The logic of this is that we should ignore any signs that the active drug is harmful to the patients. If what we were saying was 'if this trial does not give a significant reduction in angina using pronethalol we will not use it again', this might be reasonable, but the research process does not work like that. This is one of several pieces of evidence and so we should certainly use a method of inference which would enable us to detect effects in either direction. Two-sided tests are also called *two-tailed*.

The question of whether one- or two-sided tests should be the norm has been the subject of considerable debate among practitioners of statistical methods. Perhaps the position taken depends on the field in which the testing is usually done. In biological science, treatments seldom have only one effect

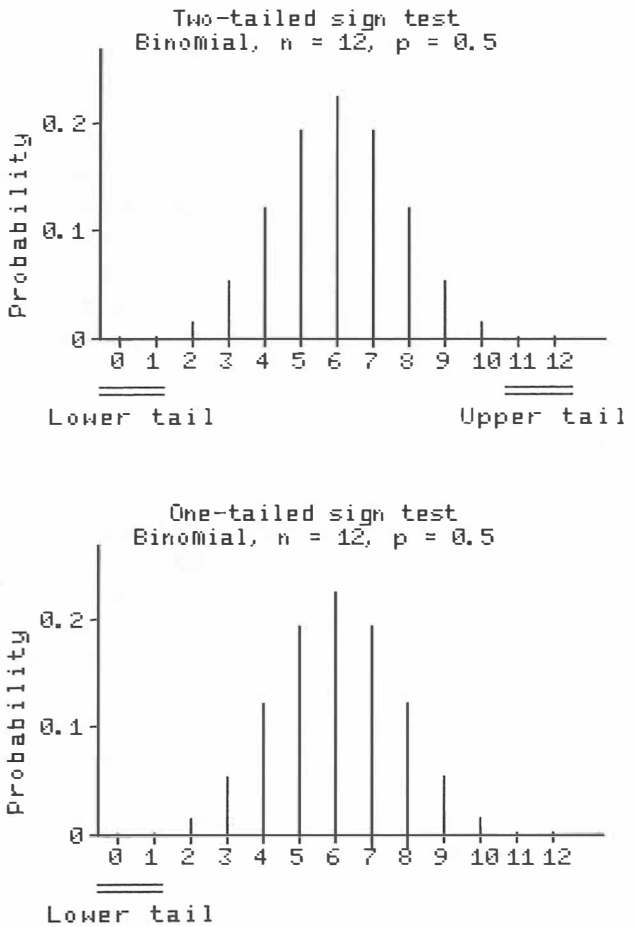


Fig. 9.2 One- and two-sided tests of significance.

and relationships between variables are usually complex. Two-sided tests would seem to be generally preferable.

There are circumstances in which a one-sided test is appropriate. In a study of the effects of an investigative procedure, laparoscopy and hydrotubation, on the fertility of sub-fertile women (Luthra *et al.* 1982), we studied women presenting at an infertility clinic. These women were observed for several months, during which some conceived, before laparoscopy was carried out on those still infertile. These were then observed for several months afterwards and some of these women also conceived. We compared the conception rate in the period before laparoscopy with that afterwards. Of course, women who conceived during the first period did not have a

laparoscopy. We argued that the less fertile a woman was, the longer it was likely to take her to conceive. Hence, the women who had the laparoscopy should have a lower conception rate (by an unknown amount) than the larger group who entered the study, because the more fertile women had conceived before their turn for laparoscopy came. To see whether laparoscopy increased fertility, we could test the null hypothesis that the conception rate afer laparoscopy was less than or equal to that before. The alternative hypothesis was that the conception rate after laparoscopy was higher than that before. A two-sided test was inappropriate because if the laparoscopy had no effect on fertility the post-laparoscopy rate was expected to be lower; chance did not come into it. In fact the post-laparoscopy conception rate was very high and the difference clearly significant.

9.6. Significant, real and important

If a difference is statistically significant, then it may well be real, but not necessarily important. For example, we may look at the effect of drug, given for some other purpose, on blood pressure. Suppose we find that the drug raises blood pressure by an average of 1 mm Hg, and that this is statistically significant. A rise in blood pressure of 1 mm Hg is not clinically significant, so, although it may be there, it does not matter. It is (statistically) significant, and real, but not important.

On the other hand, if a difference is not statistically significant, it could still be real. We may simply have too small a sample to show that a difference exists. Furthermore, the difference may still be important. The difference in mortality in the anticoagulant trial of Carleton *et al.* (1960), described in Chapter 2, was not significant, the difference in percentage survival being 5.5 in favour of the active treatment. However, the authors also quote a confidence interval for the difference in percentage survival of 24.2 in favour of heparin to 13.3 in favour of the control treatment. Thus from these data there could have been a difference in survival of as much as 24 per cent in favour of the treatment, which would certainly be important if it turned out to be the case. 'Not significant' does not imply that there is no effect. It means that we have failed to demonstrate the existence of one.

9.7. Comparing the means of large samples using the Normal Distribution

We have already seen in Section 8.5 that if we have two samples of size n_1 and n_2, with sample means $\bar{x}_1$ and $\bar{x}_2$ and sample variances s_1^2 and s_2^2, the standard error of the difference estimate $\bar{x}_1 - \bar{x}_2$ is $\sqrt{[(s_1^2/n_1 + s_2^2/n_2)]}$. Furthermore, if n_1 and n_2 are large, $\bar{x}_1 - \bar{x}_2$ will be from a Normal Distribution with mean $\mu_1 - \mu_2$, the population difference, and its standard deviation well estimated

by the standard error estimate. We used this to find a confidence interval for the difference between the means.

We can use this confidence interval to carry out a significance test of the null hypothesis that the difference between the means is zero, i.e. the alternative hypothesis is that μ_1 and μ_2 are not equal. If the confidence interval includes zero, then the probability of getting such extreme data if the null hypothesis were true is greater than 0.05 (i.e. 1 – 0.95). If the confidence interval excludes zero, then the probability of such extreme data under the null hypothesis is less than 0.05 and the difference is significant.

Another way of doing the same thing is to note that

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

is from a Standard Normal Distribution, i.e. mean 0 and variance 1. Under the null hypothesis that $\mu_1 = \mu_2$, or $\mu_1 - \mu_2 = 0$, this is

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This is the test statistic, and if it lies between -1.96 and $+1.96$ then the probability of such an extreme value is greater than 0.05 and the difference is not significant. If the test statistic is greater than 1.96 or less than -1.96, there is a less than 0.05 probability of such data arising if the null hypothesis were true, and the data not consistent with null hypothesis; the difference is significant at the 0.05 or 5 per cent level.

For an example, in the study of respiratory symptoms in schoolchildren mentioned in 8.5 above, we wanted to know whether children reported by their parents to have respiratory symptoms had worse lung function than children who were not reported to have symptoms. Ninety-two children were reported to have cough during the day or at night, and their mean PEFR was 294.8 litre/min with standard deviation 57.1 litre/min. The number of children not reported to have the symptom was 1643, and their mean PEFR was 313.6 litre/min with standard deviation 55.2 litre/min. We thus have two large samples, and can apply the Normal test. We have

$$\begin{array}{llll} n_1 = 92 & \bar{x}_1 = 294.8 & s_1 = 57.1 \\ n_2 = 1643 & \bar{x}_2 = 313.6 & s_2 = 55.2 \end{array}$$

The difference between the two groups is $\bar{x}_1 - \bar{x}_2 = 294.3 - 313.6 = -18.8$. The standard error of the difference is

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{57.1^2}{92} + \frac{55.2^2}{1643}}$$
$$= 6.11$$

The test statistic is

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{-18.8}{6.11}$$

$$= -3.1$$

Under the null hypothesis this is an observation from a Standard Normal Distribution, and the two-sided probability of this is about 0.002 from Table 7.2. If the null hypothesis were true the data which we have observed would be unlikely. We can conclude that children reported to have cough during the day or at night have lower PEFR than other children.

In this case, we have two ways of interpreting the same calculation: as a confidence interval estimate or as a significance test. The confidence interval is usually superior, because we not only demonstrate the existence of a difference but also have some idea of its size. This is of particular value when the difference is not significant. For example, in the same study only 27 children were reported to have phlegm during the day or at night. These had mean PEFR of 298.0 litre/min and standard deviation 53.9 litre/min, hence a standard error for the mean of 10.4 litre/min. This is greater than the standard error for the mean for those with cough, because the sample size is smaller. The 1708 children not reported to have this symptom had mean 312.6 litre/min and standard deviation 55.4 litre/min, giving a standard error of 1.3 litre/min. Hence the difference between the means was -14.6, with standard error given by

$$\sqrt{10.4^2 + 1.3^2} = 10.5$$

The test statistic is

$$\frac{-14.6}{10.5} = -1.4$$

This has a probability of about 0.16, and so the data are consistent with the null hypothesis. However, the 95 per cent confidence interval for the difference is

$$-14.6 - (1.96 \times 10.5) \text{ to } -14.6 + (1.96 \times 10.5)$$
$$= -35 \text{ to } 6 \text{ litre/min}$$

We see that the difference could be just as great as for cough. Because the size of the smaller sample is not so great, the test has less power for the phlegm comparison than it has for the cough comparison. We shall discuss power further below, but note for the moment that where a confidence interval can be calculated it is more informative than a test of significance.

9.8. Comparison of two proportions

For the comparison of means the test of significance and the confidence interval required the same calculation, because the standard error of the difference was the same whether the null hypothesis was true or not. When we compare two proportions this is not the case, because the standard error depends on the proportions themselves.

Suppose we wish to compare two proportions, p_1 and p_2, estimated from large independent samples of size n_1 and n_2. The null hypothesis is that the proportions in the populations from which the samples are drawn are the same. Since under the null hypothesis the proportions for the two groups are the same, we can get one common estimate of the proportion and use it to estimate the standard errors.

We estimate the common proportion from the data by

$$p = \frac{r_1 + r_2}{n_1 + n_2} \quad \text{where} \quad p_1 = \frac{r_1}{n_1},\ p_2 = \frac{r_2}{n_2}$$

We want to make inferences from the difference between samle proportions, $p_1 - p_2$, so we require the standard error of this.

$$se(p_1) = \sqrt{\frac{p(1-p)}{n_1}}, \quad se(p_2) = \sqrt{\frac{p(1-p)}{n_2}}$$

As p is based on more subjects than either p_1 or p_2, if the null hypothesis were true then standard errors would be more reliable than those estimated in Section 8.6 using p_1 and p_2 separately

$$se(p_1 - p_2) = \sqrt{se(p_1)^2 + se(p_2)^2}$$

since the samples are independent. Hence

$$\begin{aligned} se(p_1 - p_2) &= \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} \\ &= \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \end{aligned}$$

We then find the test statistic

$$\frac{p_1 - p_2}{se(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Under the null hypothesis this has expected value zero. Because the sample is large, $p(1-p)$ is a good estimate of variance, so

$$\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

is a good estimate of the standard deviation of the distribution from which $(p_1 - p_2)$ comes, i.e. the standard error, so the test statistic has variance one. Because the sample is large, $(p_1 - p_2)$ can be assumed to come from a Normal Distribution. Hence if the null hypothesis were true, the test statistic would be from a Standard Normal Distribution.

In Section 8.6, we looked at the proportions of children with bronchitis in infancy and with no such history who were reported to have respiratory symptoms in later life. We had 273 children with a history of bronchitis before age 5 years, 26 of whom were reported to have day or night cough at age 14. We had 1046 children with no bronchitis before age 5 years, 44 of whom were reported to have day or night cough at age 14. We shall test the null hypothesis that the prevalence of the symptom is the same in the populations, against the alternative that it is not.

Bronchitis | *No bronchitis*

$n_1 = 273$ | $n_2 = 1046$

$p_1 = 26/273 = 0.095\ 24$ | $p_2 = 44/1046 = 0.042\ 07$

$$p = \frac{26 + 44}{273 + 1046} = 0.053\ 07$$

$$p_1 - p_2 = 0.095\ 24 - 0.042\ 07 = 0.053\ 17$$

$$se(p_1 - p_2) = \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= \sqrt{0.053\ 07\ (1 - 0.053\ 07)\left(\frac{1}{273}+\frac{1}{1046}\right)}$$

$$= 0.015\ 24$$

$$\frac{p_1 - p_2}{se(p_1 - p_2)} = \frac{0.053\ 17}{0.015\ 24}$$

$$= 3.49$$

Referring this to the Table 7.2 of the Normal Distribution, we find the probability of such an extreme value is less than 0.01, so we conclude that the data are not consistent with the null hypothesis. We conclude that children with a history of bronchitis are more likely to be reported to have day or night cough at age 14.

Note that the standard error used here is not the same as that found in Section 8.6. It is only correct if the null hypothesis is true. If there is little difference between the proportions the two formulae will give very similar answers. If there is a difference, as here, the formula of Section 8.6 is used for finding the confidence interval.

9.9. The power of a test

The test for comparing means in Section 9.7 is more likely to detect a difference between two populations if that difference is large than it is if that difference is small. The probability that a test will produce a significant difference at a given significance level is called the *power* of the test. For a given test, this will depend on such things as the difference between the populations compared, the sample size and the significance level chosen. We have already noted in 9.4 that we are more likely to obtain a significant difference with a significance level of 0.05 than with one of 0.01. We have greater power if the critical probability is larger.

We can calculate the power of the Normal comparison of two means quite easily. The sample difference $(\bar{x}_1 - \bar{x}_2)$ is from a Normal Distribution with mean $(\mu_1 - \mu_2)$ and standard deviation $\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$, the standard error, which we shall denote by *se*. The test statistic to test the null hypothesis $\mu_1 = \mu_2$ is $(\bar{x}_1 - \bar{x}_2)/se$. The test will be significant at the 0.05 level if the test statistic is further from zero than 1.96. If $\mu_1 > \mu_2$, it is very unlikely that we will find $\bar{x}_1$ significantly less than $\bar{x}_2$, so for a significant difference we must have $(\bar{x}_1 - \bar{x}_2)/se > 1.96$.

We now find the probability that $(\bar{x}_1 - \bar{x}_2)/se$ will exceed 1.96.

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{se}$$

is an observation from a Standard Normal Distribution. We can find the probability that this exceeds any particular value x from $1 - P(x)$ in Table 7.1. If

$$\frac{\bar{x}_1 - \bar{x}_2}{se} > 1.96$$

then

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 + \mu_2)}{se} > 1.96 - \frac{\mu_1 - \mu_2}{se}$$

So the power of the test, the probability of getting a significant result, is $1 - P(x)$ where $x = 1.96 - (\mu_1 - \mu_2)/se$ and $P(x)$ is found in Table 7.1.

For the comparison of PEFR in children with and without phlegm, for example, suppose that in fact the population means were $\mu_1 = 310$ and $\mu_2 = 295$ litre/min, with standard deviation 55 in each. The sample sizes were $n_1 = 1708$ and $n_2 = 27$, so the standard error of the difference would be

$$se = \sqrt{\frac{55^2}{1708} + \frac{55^2}{27}}$$
$$= 10.67 \text{ litre/min}$$

The population difference we want to be able to detect is $\mu_1 - \mu_2 = 310 - 295 = 15$, and so

$$1.96 - \frac{\mu_1 - \mu_2}{se} = 1.96 - \frac{15}{10.67}$$
$$= 1.96 - 1.41$$
$$= 0.55$$

From Table 7.1, P for 0.55 is between 0.691 and 0.726, say 0.71. The power of the test would be $1 - 0.71 = 0.29$. If these were the population means and standard deviation, our test would have had a poor chance of detecting the difference in means, even though it existed. The test would have low power. Figure 9.3 shows the power of this test as it changes with the difference between population means. As the difference gets larger, the power increases, getting closer and closer to 1.

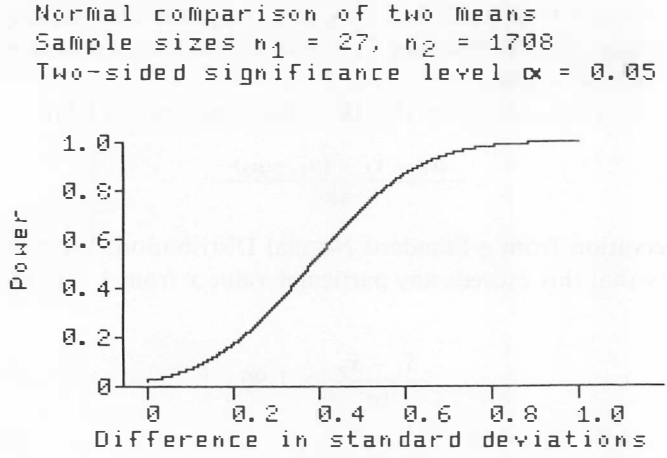


Fig. 9.3 Power curve for a comparison of two means.

9.10. Sample size for a comparison

We can use the power of a test to help choose the sample size required to detect differences if they exist. For example, suppose we want to test a treatment designed to lower serum cholesterol. We shall do this by a two sample randomized trial. From previous work we expect the standard deviation to be about 40 mg/100 ml. We must decide what sort of difference we want to detect, perhaps the sort of difference which will be clinically meaningful. Let us suppose this to be 20 mg/100 ml, half a standard deviation. Half a standard deviation is a fairly small difference: compare Fig. 7.13 which

shows the overlap between two curves one standard deviation apart. We will take the usual significance level of 0.05. We want a fairly high power, so that there is a high probability of detecting a difference of the chosen size should it exist. Usual values for the power required are 0.90 or 0.95. We shall take 0.90. Then from Table 7.1 the value of x we require corresponding to $1 - P(x) = 0.90$, $P(x) = 0.10$, is about -1.3. More precisely from Table 7.2, it is -1.28. This is the value which

$$1.96 - \frac{(\mu_1 - \mu_2)}{se}$$

must exceed to give a probability of 0.90 of obtaining a significant difference. Thus we need to find n such that

$$-1.28 = 1.96 - \frac{20}{\sqrt{\frac{40^2}{n} + \frac{40^2}{n}}}$$

If we rearrange this we get

$$\sqrt{\frac{40^2}{n} + \frac{40^2}{n}} = \frac{20}{1.96 + 1.28}$$

$$\sqrt{\frac{3200}{n}} = \frac{20}{3.24}$$

$$n = 3200 \times \left(\frac{3.24}{20}\right)^2$$

$$= 83.98$$

Since n must be an integer, $n = 84$.

Alternatively, we could start with a range of sample sizes and find out what difference we could be sure of detecting at each size. This can often tell us whether a trial is worth starting. Of course, there are a lot of assumptions made in such calculations and they can only suggest the approximate sample size required. This is discussed in more detail by Snedecor and Cochran (1980), who also give a formula for the comparison of two proportions. Altman (1982) gives a neat graphical method of calculation.

9.11. Multiple significance tests

If we test a null hypothesis which is in fact true, using 0.05 as the critical significance level, we have a probability of 0.95 of getting a 'not significant' (i.e. correct) decision. If we test two independent true null hypotheses, the probability that neither test will be significant is $0.95 \times 0.95 = 0.90$ (Section 6.2). If we test twenty such hypotheses the probability that none will be significant is $0.95^{20} = 0.36$. This gives a probability of $1 - 0.36 = 0.64$ of getting

at least one significant result; we are more likely to get one than not. The expected number of spurious significant results is $20 \times 0.05 = 1$.

Many medical research studies are published with large numbers of significance tests. These are not usually independent, being carried out on the same set of subjects, so the above calculations do not apply exactly. However, it is clear that if we go on testing long enough we shall find something which is 'significant'. We must beware of attaching too much importance to a lone significant result among a mass of non-significant ones. It may be the one in twenty which we should get by chance alone.

This is particularly important when we find that a clinical trial or epidemiological study gives no significant difference overall, but does so in a particular subset of subjects, such as women aged over 60. A remarkable paper by Lee *et al.* (1980) demonstrates this. These authors simulated a clinical trial of the treatment of coronary artery disease by allocating 1073 patient records from past cases into two 'treatment' groups at random. They then analysed the outcome as if it were a genuine trial of two treatments. The analysis was quite detailed and thorough. As we would expect, it failed to show any significant difference in survival between those patients allocated to the two 'treatments'. Patients were then subdivided by two variables which affect prognosis, the number of diseased coronary vessels and whether the left ventricular contraction pattern was normal or abnormal. A significant difference in survival between the two 'treatment' groups was found in those patients with three diseased vessels (the maximum) and abnormal ventricular contraction. As this would be the subset of patients with the worst prognosis, the finding would be easy to account for by saying that the superior 'treatment' had its greatest advantage in the most severely ill patients! As the authors show, it is in fact explained by small chance differences in other prognosis indicators between the two 'treatment' groups in this subset. The moral of this story is that if there is no difference between the treatments overall, significant differences in subsets are to be treated with the utmost suspicion.

Exercise 9M

(Each branch is either true or false.)

1. In a case-control study, patients with a given disease drank coffee more frequently than did controls, and the difference was highly significant. We can conclude that:

(a) drinking coffee causes the disease;

(b) there is evidence of a real relationship between the disease and coffee drinking in the sampled population;

(c) the disease is not related to coffee drinking;

(d) eliminating coffee would prevent the disease;

(e) coffee and the disease always go together.

2. In a comparison of two methods of measuring PEFR, 6 of 17 subjects had higher readings on the Wright peak flowmeter, 10 had higher readings on the mini-peak flowmeter and one had the same on both. If the difference between the instruments is tested using a sign test:

(a) the test statistic may be the number with the higher reading on the Wright meter;

(b) the null hypothesis is that there is no tendency for one instrument to read higher than the other;

(c) a one-tailed test of significance should be used;

(d) the test statistic should follow the Binomial Distribution ($n = 16$ and $p = \frac{1}{2}$) if the null hypothesis were true;

(e) the instruments should have been presented in random order.

3. When comparing the means of two large samples using the Normal test:

(a) the null hypothesis is that the sample means are equal;

(b) the null hypothesis is that the means are not significantly different;

(c) standard error of the difference is the sum of the standard errors of the means;

(d) the standard errors of the means must be equal;

(e) the test statistic is the ratio of the difference to its standard error.

4. In a small randomized double-blind trial of a new treatment in acute myocardial infarction, the mortality in the treated group was half that in the control group, but the difference was not significant. We can conclude that:

(a) the treatment is useless;

(b) there is no point in continuing to develop the treatment;

(c) the reduction in mortality is so great that we should introduce the treatment immediately;

(d) we should keep adding cases to the trial until the Normal test for comparison of two proportions is significant;

(e) we should carry out a new trial of much greater size.

5. In a large sample comparison between two groups, increasing the sample size will:

(a) improve the approximation of the test statistic to the Normal Distribution;

(b) decrease the chance of an error of the first kind;

(c) decrease the chance of an error of the second kind;

(d) increase the power against a given alternative;

(e) make the null hypothesis less likely to be true.

Exercise 9E

In this exercise we shall use a test of significance to compare two proportions and consider the choice of sample size for such a study.

Table 2.7 shows the results of the field trial of Salk poliomyelitis vaccine. In the randomized control areas, 200 745 children received the vaccine, of whom 33 developed paralytic polio. Placebo was given to 201 229 children, of whom 115 contracted paralytic polio.

1. Test the significance of the difference in proportion of polio cases between the groups.

2. If this difference were significant, would you conclude that the vaccine was effective in preventing polio?

3. Find a 95 per cent confidence interval for the difference.

4. What size of sample would be required to have a 90 per cent chance of detecting a reduction of 40 per cent in polio the number of polio cases, if the control rate was expected to be about 50 per 100 000? (Hint: use the method of Section 9.10 with the standard error formula of Section 9.8, putting $n_1 = n_2 = n$.)

10. Analysis of the means of small samples using the *t* Distribution

10.1. The *t* Distribution

We have seen in Chapters 8 and 9 how the Normal Distribution can be used to calculate confidence intervals for means and carry out tests of significance on means when we have large samples. In this chapter we shall see how similar methods may be used when we have small samples. We shall do this using the *t* Distribution.

So far, the probability distributions we have used have arisen because of the way data were collected either from the way samples are drawn as for the Binomial Distribution, or from the mathematical properties of large samples as for the Normal Distribution. The distribution did not depend on any property of the data themselves. To use the *t* Distribution we must make an assumption about the distribution from which the observations themselves are taken, that is, the distribution of the variable in the population. We must assume that the observations come from a Normal Distribution. As we saw in Chapter 7, many naturally occurring variables have been found to follow a Normal Distribution closely. We will discuss the effects of any deviations from the Normal later.

We have already mentioned the *t* Distribution in Chapter 7, as one of those derived from the Normal. We will now look at it in more detail. Assume we have a random sample of observations from a Normal Distribution with mean μ and variance σ^2. We denote the observations by $x_1, x_2, \ldots, x_i, \ldots, x_n$ and there are n observations in all. We estimate the variance by calculating the sum of squares about the mean, $\Sigma(x_i - \bar{x})^2$, and dividing by the degrees of freedom, $n - 1$, to give $s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$.

As we saw in Chapter 8, the distribution of all possible sample means, i.e. of all possible $\bar{x}$'s, has a standard deviation and the estimate of this is $\sqrt{(s^2/n)}$, the standard error of $\bar{x}$. If we had a large sample, we would then say that the mean $\bar{x}$ comes from a Normal Distribution and that $\sqrt{(s^2/n)}$ is a good estimate of its standard deviation (Chapter 8). Then we could say that the ratio $(\bar{x} - \mu)/\sqrt{(s^2/n)}$ was from a Normal Distribution with mean 0 and standard

deviation 1, the Standard Normal Distribution. However, this is not true for a small sample. The estimated standard deviation, s, may vary from sample to sample. Samples with small standard deviations will give very large ratios and the distribution will have much longer tails than the Normal.

The distribution of the ratio (mean over standard error) calculated from a small sample depends on the distribution from which the original observations come. As so many variables follow a Normal Distribution, it is worth looking at what happens when the observations are Normal. Provided our observations are from a Normal Distribution, $\bar{x}$ is too. But we cannot assume that $\sqrt{(s^2/n)}$ is a good estimate of its standard deviation. We must allow for the variation of s^2 from sample to sample. In fact, it can be shown that, provided the observations follow a Normal Distribution, the sampling distribution of $(\bar{x} - \mu)/\sqrt{(s^2/n)}$ is Student's t Distribution with $(n - 1)$ degrees of freedom (see Appendix 10A). We can therefore replace the Normal Distribution by the t Distribution in confidence intervals and significance tests for small samples. In fact, when we divide anything which is Normally distributed with mean zero, such as $\bar{x} - \mu$, by its standard error which is based on a single sum of squares of Normally distributed data, we get a t Distribution.

Figure 10.1 shows the t Distribution with 1, 4, and 20 degrees of freedom (d.f.). It is symmetrical, with longer tails than the Normal Distribution. For example, with 4 d.f. the probability of t being greater than 2.78 is 2.5 per cent,

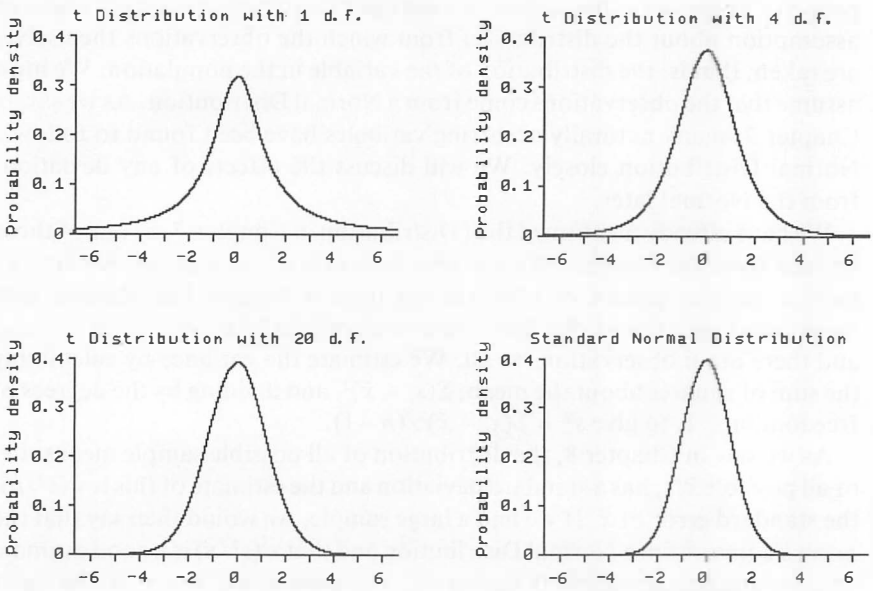


Fig. 10.1 Student's t Distribution with 1, 4, and 20 degrees of freedom.

whereas for the Standard Normal Distribution the probability of t being greater than 2.78 is only 0.5 per cent. This is what we expect, as in the expression $(\bar{x} - \mu)/\sqrt{(s^2/n)}$ the variation in s^2 from sample to sample will produce some samples with low values of s^2 and so large values of t.

As the degrees of freedom, and hence the sample size, increase, s^2 will tend to be closer to its expected value of σ^2. The variation in s^2 will be less, and hence the variation in t will be less. This means that extreme values of t will be less likely, and so the tails of the distribution, which contain the probability associated with extreme values of t, will be smaller. We have already seen that for large samples $(\bar{x} - \mu)/\sqrt{(s^2/n)}$ follows a Standard Normal Distribution. The t Distribution gets more and more like the Standard Normal Distribution as the degrees of freedom increase. This is clearly shown in Fig. 10.1, which also shows the Standard Normal Distribution.

Table 10.1. Two-tailed probability points of the t Distribution

Degrees of freedom	Probability 0.10 (10%)	0.05 (5%)	0.01 (1%)	0.001 (0.1%)
1	6.31	12.70	63.66	636.62
2	2.92	4.30	9.93	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	1.90	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	3.06	4.32
13	1.77	2.16	3.01	4.22
14	1.76	2.15	2.98	4.14
15	1.75	2.13	2.95	4.07
16	1.75	2.12	2.92	4.02
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	1.73	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
30	1.70	2.04	2.75	3.65
40	1.68	2.02	2.70	3.55
60	1.67	2.00	2.66	3.46
120	1.66	1.98	2.62	3.37
Infinite (Normal Distribution)	1.65	1.96	2.58	3.29

Like the Normal Distribution, the t Distribution function cannot be integrated algebraically and its numerical values have been tabulated. Because the t Distribution depends on the degrees of freedom, it is not usually tabulated in full like the Normal Distribution in Chapter 7. Instead, probability points are given for a range of degrees of freedom. Table 10.1 shows two sided probability points for selected degrees of freedom. Thus, with 4 degrees of freedom, we can see that, with probability 0.05, t will be 2.78 or more from its mean, zero.

Because only certain probabilities are quoted, we cannot usually find the exact probability associated with a particular value of t. For example, suppose we want to know the probability of t on 9 degrees of freedom being further from zero than 3.7. From Table 10.1 we see that the 0.01 point is 3.25 and the 0.001 point is 4.78. We therefore know that the required probability lies between 0.01 and 0.001. We could write this as $0.001 < p < 0.01$. Often the lower bound, 0.001, is omitted and we write $p < 0.01$. With a computer it is possible to calculate the exact probability every time, so this common practice is due to disappear.

The name 'Student's t Distribution' often puzzles newcomers to the subject. It is not, as is often thought, an easy method suitable for students to use. The origin of the name is part of the folklore of statistics. The distribution was discovered by W. S. Gossett, an employee of the Guinness brewery in Dublin. At that time, the company would not allow its employees to publish the results of their work, lest it should lose some commercial advantage. Gossett therefore submitted his paper under the pseudonym 'Student' (Student 1908). In this paper he not only presented the mathematical derivation of the distribution, but also gave the results of a sampling

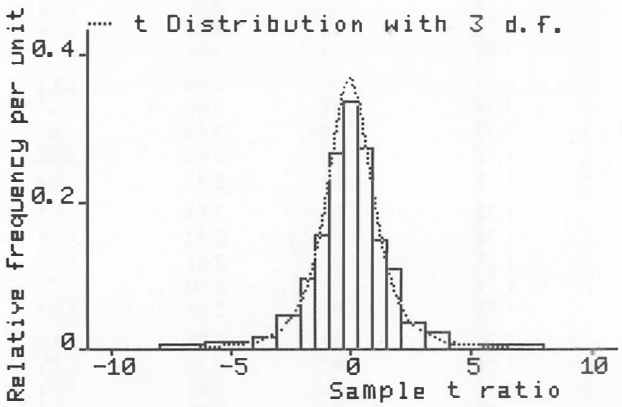


Fig. 10.2 Sample t ratios derived from 750 samples of four human heights, after Student (1908).

experiment like those described in Sections 4.7 and 8.2. He took the heights of 3000 criminals, wrote each on a piece of card, then drew 750 samples of size 4 to give 750 $(\bar{x} - \mu)/\sqrt{(s^2/n)}$ statistics. Figure 10.2 shows the very good agreement which he obtained.

10.2. The one-sample *t* method

We can use the t Distribution to find confidence intervals for parameters estimated from a small sample from a Normal Distribution. We do not usually have small samples in sample surveys, but we often do find them in clinical studies. For example, we can use the t Distribution to find confidence intervals for the size of difference between two treatment groups, or between measurements obtained from subjects under two conditions. We shall deal with the latter, single-sample problem first.

In general, for a sample from a Normal Distribution $(\bar{x} - \mu)/\sqrt{(s^2/n)}$ is from a t Distribution with $(n - 1)$ degrees of freedom. The population mean, μ, is unknown and we wish to know how far from the sample mean it is likely to be, using a 95 per cent confidence interval. We can see that, with probability 0.95, the difference between $\bar{x}$ and μ is at most t standard errors, where t is the value of the t Distribution such that 95 per cent of observations will be closer to zero than t. For a large sample this will be 1.96 as for the Normal Distribution. For small samples we must use Table 10.1. In this table, the probability that the t Distribution is further from zero than t is given, so we must first find one minus our desired probability 0.95. We have $1 - 0.95 = 0.05$, so we use the 0.05 column of the table to get the value of t. We then have the 95 per cent confidence interval for μ, which is

$$\bar{x} - t\sqrt{s^2/n} \quad \text{to} \quad \bar{x} + t\sqrt{s^2/n}$$

Consider the data of Table 10.2. These are results from a comparison of two instruments for measuring PEFR, a Wright Peak Flowmeter and a Mini Peak Flowmeter. The subjects were family and colleagues, and so not a random sample. Each gave two readings on each instrument in random order. Table 10.2 shows the second reading on each. We shall measure the amount of bias between the instruments, the amount by which one tends to read above the other. The first step is to find the differences (Wright – mini). We then find the mean difference and its standard error, as described in Chapter 8.

To find the 95 per cent confidence interval for the mean difference we must suppose that the differences follow a Normal Distribution. To calculate the interval, we first require the relevant point of the t Distribution from Table 10.1. There are 12 differences and hence $(n - 1) = 11$ degrees of freedom associated with s^2. We want a probability of 0.95 of being closer to zero than t, so we go to Table 10.2 with $p = 1 - 0.95 = 0.05$. Using the 11 d.f. row, we get $t = 2.20$. Hence the difference between the sample and population

means is at most 2.20 standard errors with probability 0.95, and the 95 per cent confidence interval is $-17.2 - (2.20 \times 11.6)$ to $-17.2 + (2.20 \times 11.6) = -42.7$ to 8.3 litre/min. (In the large sample case, we would use the Normal Distribution instead of the t Distribution, putting 1.96 instead of 2.20. We would not then need the differences to follow a Normal Distribution.)

On the basis of these data, the mini-meter could tend to over-read by as much as 43 litre/min, or to under-read by as much as 8 litre/min. An error of 43 litre/min is quite substantial, and we may have a problem. We would need a much larger sample to obtain a more precise estimate if we thought this were required.

We can also use the t Distribution to test the null hypothesis that the mean difference is zero. If the null hypothesis were true, and the differences follow a Normal Distribution, the test statistic $\bar{x}/\sqrt{(s^2/n)}$ would be from a t Distribution with $(n-1)$ degrees of freedom. This is because the null hypothesis is $\mu = 0$, hence the numerator $\bar{x} - \mu = \bar{x}$.

For the example, we have

$$\frac{\bar{x}}{\sqrt{s^2/n}} = \frac{-17.2}{11.6} = -1.48$$

If we go to the 11 d.f. row of Table 10.1, we find that the probability of such an extreme value arising is greater than 0.10, the 0.10 point of the distribution being 1.80. Using a computer we would find $p = 0.17$. The data are consistent with the null hypothesis and we have failed to demonstrate the existence of a bias. Note that the confidence interval is more informative than the significance test.

We could also use the sign test to test the null hypothesis of no bias. This gives us 3 positives out of 11 differences (one difference, being zero, gives no useful information) which gives a two-sided probability of 0.23. This is greater than the t probability, but fairly similar. The t test gives the smaller probability because, provided the assumption of Normality is true, the t test is the more powerful test.

The validity of the methods described above depends on the assumption that the differences are from a Normal Distribution. We can check the assumption of Normality by a Normal plot (Section 7.5). Figure 10.3 shows a Normal plot for the differences and also for the Wright meter reading. The Normal plot for the differences deviates only slightly from a straight line, one point in particular, subject 11, appearing rather out. The Wright meter readings show a clear kink in the line and are unlikely to be from a Normal Distribution. At first sight this is surprising, as we have remarked before that PEFR tends to be Normally distributed. However, this is not a sample from a population of similar age, or from the adult population in general. Most of these subjects were aged between 20 and 30 years, but subjects 10 and 11 (my mother and mother-in-law!) were in an older age group and so produced

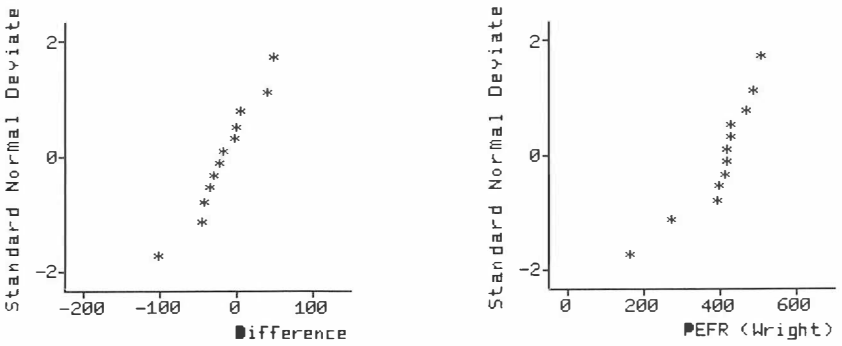


Fig. 10.3 Normal plots for the data of Table 10.2.

much lower PEFRs. Subject 11, who also produced the difference of 103 litre/min, was quite consistent in this. Her first pair of readings were 178 litre/min (Wright) and 259 (minimeter), which were very similar to those in Table 10.2.

Table 10.2. PEFR (litre/min) measured by Wright meter and mini-meter, female subjects

Subject	Wright PEFR	Mini PEFR	Difference
1	490	525	−35
2	397	415	−18
3	512	508	4
4	401	444	−43
5	470	500	−30
6	415	460	−45
7	431	390	41
8	429	432	−3
9	420	420	0
10	275	227	48
11	165	268	−103
12	421	443	−22
Sum x	4826	4996	−206
Mean $\bar{x}$	402.2	419.3	−17.2
Sum of squares about the mean $\Sigma(x_i - \bar{x})^2$	99 215.7	89 290.7	17 889.7
Variance s^2	9019.6	8117.3	1626.3
se mean $\sqrt{s^2/n}$	27.4	26.0	11.6

Another plot which is a useful check here is the difference against the sum or mean (Fig. 10.4). If the difference depends on the mean, then we should be careful of drawing any conclusion about the mean difference. We may want to investigate this further, perhaps by looking at the ratio instead of the

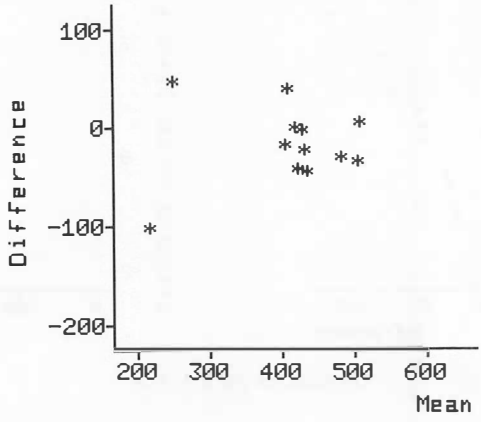


Fig. 10.4 Plot of difference against mean for the data of Table 10.2.

difference, or estimating the difference as a function of the mean of the two measurements. In this case the difference between the two readings does not appear to be related to the level of PEFR and we need not be concerned about this.

Incidentally, we should be wary of drawing any conclusions about the Mini Wright Peak Flowmeter from this sample of one meter. It had been used in a field survey of wheezy children, being kept in their homes for periods of a fortnight (Johnston *et al.* 1984). It may well have been battered about a bit.

Despite the clear non-Normality of the PEFR, the differences look like a fairly good fit to the Normal. There are two reasons for this: the subtraction removes variability due to age and height, leaving the measurement error which is more likely to be Normal, and the two measurement errors are then added, producing the tendency of sums to Normality seen in the Central Limit Theorem (Section 7.3). We can see that the assumption of Normality for the one sample case is quite likely to be met. We discuss this further in Section 10.5. When the one sample *t* test is used with differences, as in the PEFR meter example, it is also known as the *paired* t *test.*

10.3. The means of two independent samples

Suppose we have two samples from Normally distributed populations, with which we want to estimate the difference between the population means. We shall call the population means μ_1 and μ_2, the sample means $\bar{x}_1$ and $\bar{x}_2$, the variances s_1^2 and s_2^2, and the sample sizes n_1 and n_2. If the samples were large, the 95 per cent confidence interval for the difference would be

$$\bar{x}_1 - \bar{x}_2 - 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2} \text{ to } \bar{x}_1 - \bar{x}_2 + 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}$$

Unfortunately, we cannot simply replace 1.96 by a number from Table 10.1. This is because the standard error does not have the simple form described in 10.2. It is not the square root of a constant times a sum of squares, but rather is the square root of the sum of two constants multiplied by two sums of squares. Hence, it does not follow the square root of the Chi-squared Distribution as required for the denominator of a t distributed random variable (Appendix 7A).

In order to use the t Distribution we must make a further assumption about the data. Not only must the samples be from Normal Distributions, they must be from Normal Distributions with the same variance. This is not as unreasonable an assumption as it may sound. A difference in mean but not in variability is a common phenomenon. The PEFR data for children with and without symptoms analysed in Sections 8.5 and 9.7 show the characteristic very clearly.

We now estimate the common variance, s^2. First we find the sum of squares about the sample mean for each sample, which we can label SS_1 and SS_2. We form a combined sum of squares by $SS_1 + SS_2$. The sum of squares for the first group, SS_1, has $(n_1 - 1)$ degrees of freedom and, for the second, SS_2 has $(n_2 - 1)$ degrees of freedom. The total degrees of freedom is therefore $(n_1 - 1 + n_2 - 1) = (n_1 + n_2 - 2)$. We have lost 2 degrees of freedom because we have a sum of squares about two means. The combined estimate of variance is

$$s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$$

The standard error of $\bar{x}_1 - \bar{x}_2$ is

$$\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = \sqrt{s^2\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}$$

Now we have a standard error related to the square root of the Chi-squared Distribution and we can get a t distributed variable by

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s^2\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}}$$

having $(n_1 + n_2 - 2)$ degrees of freedom. We thus have the 95 per cent confidence interval for the difference between means as

$$\bar{x}_1 - \bar{x}_2 - t\sqrt{s^2(1/n_1 + 1/n_2)} \quad \text{to} \quad \bar{x}_1 - \bar{x}_2 + t\sqrt{s^2(1/n_1 + 1/n_2)}$$

where t is the 0.05 point with $(n_1 + n_2 - 2)$ degrees of freedom from Table 10.1. Alternatively, we can test the null hypothesis that the difference is zero, i.e. that $\mu_1 = \mu_2$, using the test statistic

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

which would follow the t Distribution with $(n_1 + n_2 - 2$ d.f.) if the null hypothesis were true.

For a practical example, Table 10.3 shows the data for male subjects corresponding to that in Table 10.2. We shall estimate the difference between the biases (i.e. mean difference between Wright and mini-meter) for females and males. We have already noted the approximate Normality of the differences. We must now consider the similarity of their variances. It is clear that the variance for the males is much smaller than that for females, and the assumption that in the populations the variances are the same is in question. However, the disparity is not too great to be due to chance, as can be shown using the F test or Levene test (beyond the scope of this book; see Armitage 1973; Snedecor and Cochran 1980). We shall accept it for the moment and consider its effect later.

Table 10.3. PEFR (litre/min) measured by Wright meter and mini-meter, male subjects

Subject	Wright PEFR	Mini PEFR	Difference
13	611	625	-14
14	638	642	-4
15	633	605	28
16	492	467	25
17	372	370	2
Sum Σx_i	2746	2709	37
Mean $\bar{x}$	549.2	541.8	7.4
Sum of squares about the mean $\Sigma(x_i - \bar{x})^2$	53 398.8	56 066.8	1351.2
Variance s^2	13 349.7	14 016.7	337.8
se mean $\sqrt{s^2/n}$	51.7	52.9	8.2

First we find the common variance estimate, s^2. The combined sum of squares about the sample means is

$$17889.7 + 1351.2 = 19240.9$$

The combined degrees of freedom are

$$n_1 + n_2 - 2 = 12 + 5 - 2 = 15$$

Hence

$$s^2 = 19240.9/15 = 1282.73$$

The standard error of $(\bar{x}_1 - \bar{x}_2)$ is

$$\sqrt{s^2(1/n_1 + 1/n_2)} = \sqrt{1282.73(1/12 + 1/5)}$$
$$= 19.06$$

The value of the t Distribution for the 95 per cent confidence interval is found from the 0.05 column and 15 d.f. row of Table 10.1 to be 2.13. Hence the 95 per cent confidence interval is

$$\bar{x}_1 - \bar{x}_2 - t\sqrt{s^2(1/n_1 + 1/n_2)} \text{ to } \bar{x}_1 - \bar{x}_2 + t\sqrt{s^2(1/n_1 + 1/n_2)}$$
$$= -17.2 - 7.4 - 2.13 \times 19.06 \text{ to } -17.2 - 7.4 + 2.13 \times 19.06$$
$$= -65.2 \text{ to } 16.0 \text{ litre/min}$$

Hence there could be quite a large difference between the response of males and females, from this very small sample, or there may be none at all.

To test the null hypothesis that the male–female difference is zero, the test statistic is

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}} = \frac{-17.2 - 7.4}{19.06}$$
$$= -1.29$$

If the null hypothesis were true, this would be an observation from the t Distribution with $(n_1 + n_2 - 2) = 12 + 5 - 2 = 15$ degrees of freedom. From Table 10.1, the probability of such an extreme value is greater than 0.10. The computer gives a probability of 0.22. Hence the data are consistent with the null hypothesis and we cannot conclude that the bias is different for males and females. Again, we can see the advantage which estimation by confidence interval has over significance tests.

What happens if we do not make the assumption of uniform variance? There is an approximate solution based on the t Distribution (e.g. see Davies and Goldsmith 1972; Snedecor and Cochran 1980) using the standard error formula of Section 8.5, $\sqrt{[s_1^2/n_1 + s_2^2/n_2]}$. For our data this standard error is 14.3. The difference between the variance leads to a rather complicated reduction in the degrees of freedom, in this case to 14. For this example we obtain a confidence interval of -55.1 to 6.0 litres/min, or a t test statistic of -1.7 with 14 degrees of freedom, $p = 0.11$. This is similar to what we obtained by the standard method. There are other approaches based on the t test (see Armitage 1973). Another approach is to abandon the use of variance altogether and use the Mann–Whitney U test described in Section 12.2. We shall look at these data again in Chapter 15.

10.4. The use of transformations

We have already seen (Section 7.4) that some variables which are not Normally distributed can be made so by a suitable transformation. There are

Table 10.4. Biceps skinfold thickness (mm) in two groups of patients

Crohn's disease		Coeliac disease
1.8	4.2	1.8
2.2	4.4	2.0
2.4	4.8	2.0
2.5	5.6	2.0
2.8	6.0	3.0
2.8	6.2	3.8
3.2	6.6	4.2
3.6	7.0	5.4
3.8	10.0	7.6
4.0	10.4	

several transformations that can be used for this purpose. The most commonly used is the logarithm, which is suitable for data which are quite highly skewed. This works when the standard deviations of different samples from the population are proportional to their means. A typical variable for this treatment is the serum triglyceride of Figs 7.15 and 7.16 and this is often true of such serum measurements. The square root transformation is useful when data are not so highly skewed and when the variance of a sample is proportional to its mean. Poisson variables have this property, for example. The reciprocal can be used when the standard deviation is proportional to the square of the mean, and data are very highly skewed indeed. Survival times tend to behave like this.

In large data sets, there are fairly good methods of determining the appropriate transformation (see Healy 1968, for a readable account). For small samples it is a matter of experience, trial and error. Table 10.4 shows some data from a study of anthropometry and diagnosis in patients with intestinal disease (Maugdal *et al.* 1985). We are interested in the differences in various anthropometrical measurements in patients with different diagnoses,

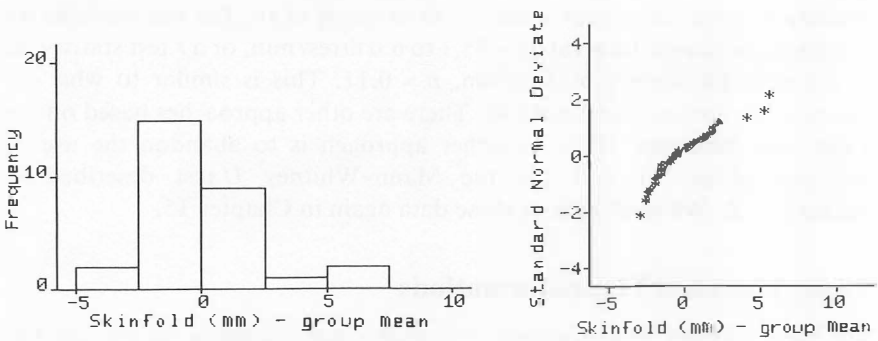


Fig. 10.5 Histogram and Normal plot for the biceps skinfold data.

and here we have the biceps skinfold measurements for 20 patients with Crohn's disease and 9 patients with coeliac disease. The data have been put into order of magnitude and it is fairly obvious that the distribution is skewed and far from Normal. Figure 10.5 shows this clearly. I have subtracted the group mean from each observation, giving what is called the within-group residual, and then found both the frequency distribution and Normal plot. The distribution is clearly skew, and this is reflected in the Normal plot, which shows a pronounced curvature.

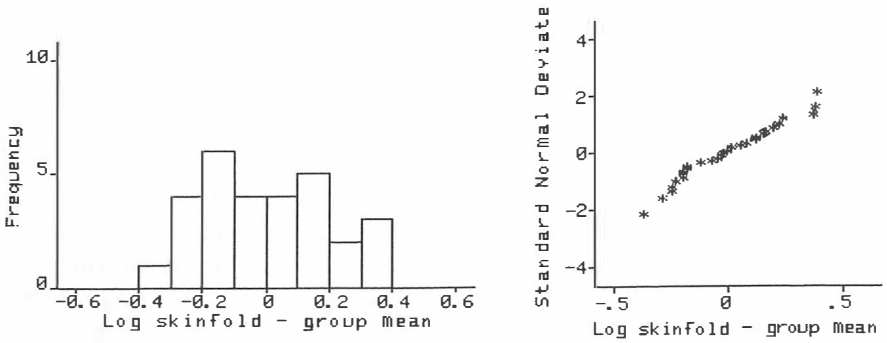


Fig. 10.6 Histogram and Normal plot for the biceps skinfold data, log transformed.

We need a Normalizing transformation, if one can be found. The first guess is the log transform, and Fig. 10.6 shows the histogram and Normal plot for the residuals after transformation. (These are logarithms to base 10.) The fit to the Normal Distribution is not perfect, but much better than in Fig. 10.5. We could use the two sample t method on these data quite happily. Compare Fig. 10.7, which shows the result of a square root transformation.

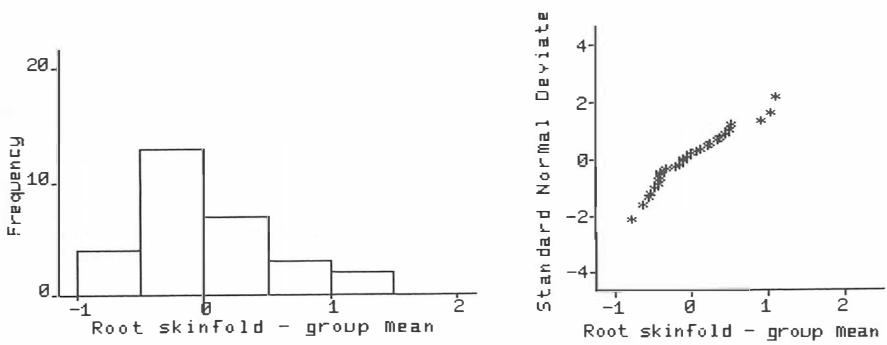


Fig. 10.7 Histogram and Normal plot for the biceps skinfold data, square root transformed.

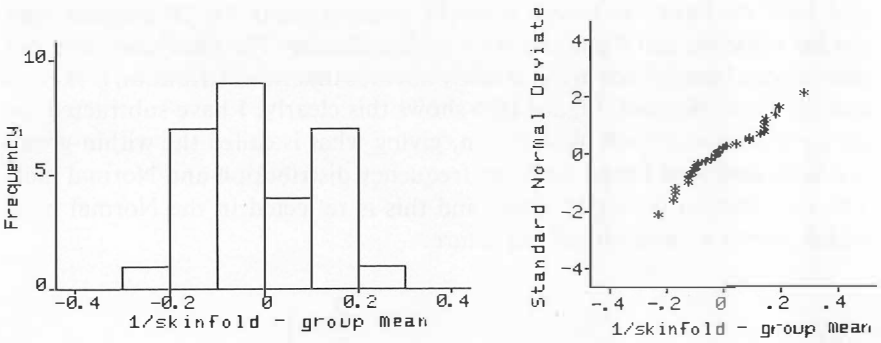


Fig. 10.8 Histogram and Normal plot for the biceps skinfold data, reciprocal transformed.

The skewness is still apparent though less than in the untransformed data. Figure 10.8 shows the results of the reciprocal transformation. The results are if anything marginally worse than those for the log transformation, though it is not easy to choose between them. In practice I would probably select the log transformation, as the resulting statistics are easier to interpret. The reciprocal transformation changes the sign of the difference, for example.

Table 10.5 shows the results of the two sample t method used with the raw, untransformed data and with each transformation. The t test statistic increases and its associated probability decreases as we move closer to a Normal Distribution, reflecting the increasing power of the t test as its assumptions are more closely met. Table 10.5 also shows the ratio of the variances in the two samples. We can see that, as the transformed data get closer to Normality, the variances tend to become more equal also.

The transformed data clearly give a better test of significance than the raw data. The confidence intervals for the transformed data are more difficult to interpret, however, so the gain here is not so apparent. The confidence limits

Table 10.5. Biceps skinfold thickness compared for two groups of patients, using different transformations

Transformation	Test of significance, t Distribution, 27 d.f.		95 per cent confidence interval for difference on transformed scale	Variance ratio, larger/smaller
	t statistic	probability		
None, raw data	1.28	0.21	−0.71 mm to 3.07 mm	1.52
Square root	1.38	0.18	−0.140 to 0.714	1.16
Logarithm	1.48	0.15	−0.050 to 0.307	1.10
Reciprocal	−1.65	0.11	−0.203 to 0.022	1.63

for the difference cannot be transformed back to the original scale. If we try it, the square root and reciprocal limits give ludicrous results. The log gives interpretable results (0.89–2.03) but these are not limits for the difference in millimetres. How could they be, for they do not contain zero yet the differences is not significant? They are in fact the 95 per cent confidence limits for the ratio of the Crohn's disease mean to the coeliac disease mean. If there were no difference, of course, the expected value of this ratio would be one, not zero, and so lies within the limits. The reason is that when we take the difference between the logarithms of two numbers, we get the logarithm of their ratio, not of their difference (Appendix 5A). However, when we take the mean of the logarithms of several numbers, we do get the logarithm of a mean of sorts, the geometric mean. The geometric mean of n numbers is the nth root of their product.

10.5. Deviations from the assumptions of Normality and uniform variance

The methods described in this chapter depend on some strong assumptions about the distributions from which the data come. This often worries users of statistical methods, who feel that these assumptions must limit greatly the use of t Distribution methods and find the attitude of many statisticians, who often use methods based on Normal assumptions almost as a matter of course, rather sanguine. We shall look at some consequences of deviations from the assumptions.

First we shall consider non-Normality. As we have seen, some variables conform very closely to the Normal Distribution; others do not. Deviations occur in two main ways: grouping and skewness.

Grouping occurs when a continuous variable, such as human height, is measured in units which are fairly large relative to the range. This happens, for example, if we measure human height to the nearest inch. The heights in Fig. 10.2 were to the nearest inch, and the fit to the t Distribution is very good. This was a very coarse grouping, as the standard deviation of heights was 2.5 inches and so 95 per cent of the 3000 observations had values over a range of 10 inches, only 10 or 11 possible values in all. We can see from this that if the underlying distribution is Normal, rounding the measurement is not going to affect the application of the t Distribution by much.

Skewness, on the other hand, can invalidate methods based on the t Distribution. For small samples of highly skew data, the t Distribution does not fit the distribution of $(\bar{x} - \mu)/\sqrt{(s^2/n)}$ at all well. Figure 10.9 shows the results of a computerized repetition of Student's experiment (Fig. 10.2), using random data from a highly skewed distribution. The fit is quite poor. The sampling distribution is skewed, the left tail being shorter and the right longer than for the t Distribution. This is because the mode, the place where

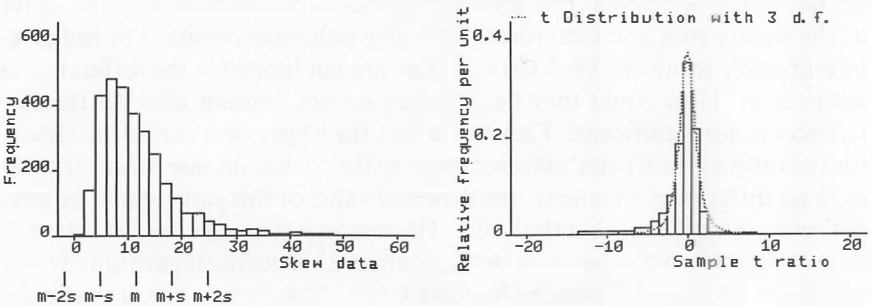


Fig. 10.9 Sampling experiment on the one-sample t statistic from a highly skewed population distribution.

observation is most frequent for the distribution of the data, is not close to the population mean. We may get a sample of observations all close to the mode, hence having a small standard deviation but a large $(\bar{x} - \mu)$, and so a large t ratio. The solution is to use a transformation to Normality, such as the logarithm. If this does not work, then we must turn to methods which do not require a Normal assumption. The sign test is one possibility. Others are described in Chapter 12.

Skewness is much more likely to arise in two-sample problems, because we do not have the Normalizing effect of differences described in Section 10.2. It also affects the two-sample t statistic of Section 10.3. Figure 10.10 shows the distribution of 750 two-sample t statistics obtained by drawing pairs of samples of three-observations from the highly skewed population of Fig. 10.9. Two samples of size three give a t statistic with 4 degrees of

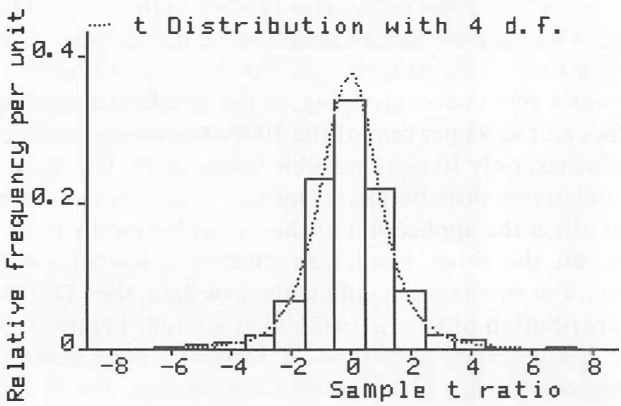


Fig. 10.10 Sampling experiment on the two-sample t statistic from a highly skewed population distribution, equal samples.

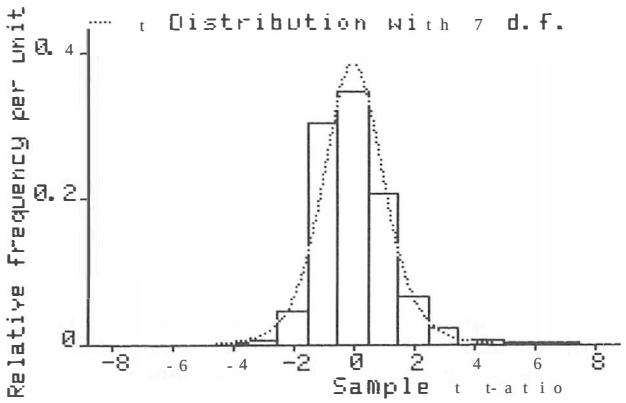


Fig. 10.11 Sampling experiment on the two-sample t statistic from a highly skewed population distribution, unequal samples.

freedom. The fit to the t Distribution is remarkably good and in general for two equal-sized samples the t method is very resistant to deviations from Normality. As the samples become less equal in size the fit becomes less good. Figure 10.11 shows what happens when we have the difference between a mean of 3 and a mean of 6. The fit is not as good as in Fig. 10.10, but still considerably better than in Fig. 10.9. We can see that even large departures from Normality are not too upsetting to the two-sample t method. This means that we need not worry about small departures from Normality. If there is an obvious departure from Normality, we should try to transform the data to Normality and then apply the t Distribution. If we can't do this we must use a different approach to the data as described in Section 12.2.

The other assumption of the two-sample t method is that the variances in the two populations are the same. If this is not correct, the t Distribution will not necessarily apply. However, the effect is usually small if the two populations are from a Normal Distribution. Figure 10.12 shows the two sample t statistic for samples size 3 and 6, and for sizes 6 and 3, where the variance of the second is four times that of the first. However, it is unusual to have unequal variances with Normal data. Unequal variance is more often associated with skewness in the data, in which case a transformation designed to correct one fault often tends to correct the other as well.

To sum up, both the one- and two-sample t methods are said to be 'robust' to most deviations from the assumptions. In other words, only large deviations are going to have much effect on the method. The main problem is with skewed data in the one sample method, but for reasons given in Section 10.2, the paired test will usually provide differences with a reasonable distribution. If the data do appear to be non-Normal, then a Normalizing transformation will improve matters.

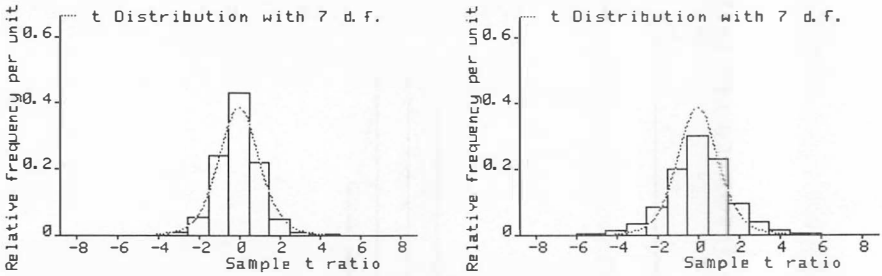


Fig. 10.12 Sampling experiments on the two-sample t statistic from Normal populations with unequal variances, samples size 3 and 6. (a) larger sample has larger variance, (b) smaller sample has larger variance.

If the assumptions are not met and the data cannot be transformed, all is not lost. In the unusual case of populations which are Normal with variances which cannot be assumed to be the same, there is an approximate t method using the formula of 8.5, as we noted in Section 10.3. We shall discuss an alternative approach which requires no assumption of Normality in Chapter 12.

10.6. What is a large sample?

In this chapter we have looked at small-sample versions of the large-sample methods of Sections 8.5 and 9.7. In Sections 8.5 and 9.7 we ignored both the distribution of the variable and the variability of s^2, on the grounds that they did not matter provided the samples were large. How small can a large sample be? This question is critical to the validity of these methods, but seldom seems to be discussed in textbooks.

Provided the assumptions of the t test apply, the question is easy enough to answer. Inspection of Table 10.1 will show that for 30 degrees of freedom the 5 per cent point is 2.04, which is so close to the Normal value of 1.96 that it makes little difference which is used. So for Normal data with uniform variance we can forget the t Distribution when we have more than 30 observations.

When the data are not in this happy state, things are not so simple. Figure 10.13 shows one sample t statistics from the highly skewed distribution of Fig. 10.9. When we have samples of size 30, the fit to the Normal Distribution is quite poor. The sampling distribution of the t statistic is noticably skew, in this case to the left. Even a sample of size 100 appears to deviate from the Normal, though in this case not by much. Although there is still some skewness present the proportion of mean/standard error ratios outside the -2 to $+2$ interval is 4.7 per cent, so here the use of the Normal approximation will not lead us far astray.

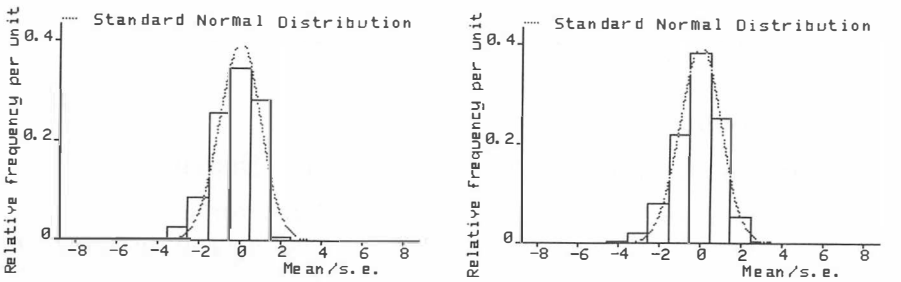


Fig. 10.13 Sample mean – population mean over standard error for 750 samples from a skewed population, sample sizes 30 and 100.

For two sample comparisons, things are better. We have already seen that equal-sized samples give a remarkably good fit to the t Distribution. Figure 10.14 shows the sampling distribution for the difference between means of sample size 10 and size 20, divided by the standard error. The standard error used is that of Section 8.5, with separate variances. As we might expect from the previous section, the fit here is better than for the single sample, but there is still room for improvement. For samples of size 33 and 67 the fit is quite good.

Of course, it would be a mistake to draw any strong conclusions from so few simulations, but they do illustrate a few principles. First, if in doubt, treat the sample as small. Secondly, transform to Normality if possible, especially in the one-sample case. In the one-sample case it is easy to transform estimates of confidence limits, etc. back to the original scale anyway. Thirdly, the more non-Normal the data, the larger the sample needs to be before we can ignore errors in the Normal approximation.

There is no simple answer to the question: 'how large is a large sample?'. If we want a rough guide, we should be reasonably safe with inferences about means if the sample is greater than 100 for a single sample, or if both samples

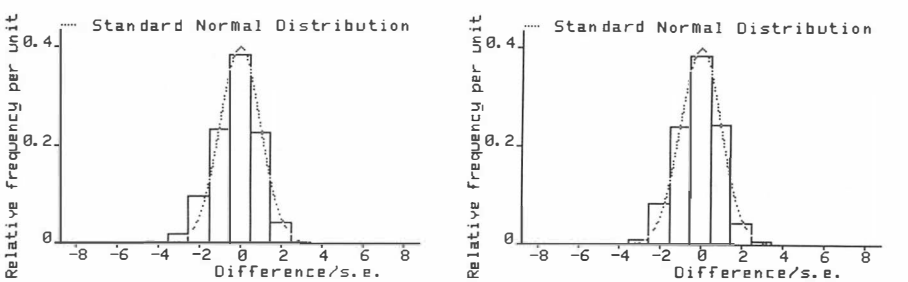


Fig. 10.14 Difference between means over standard error for 750 samples from a skewed distribution, sample sizes 10 and 20 and sample sizes 33 and 67.

are greater than 50 for two samples. We can see that the application of statistical methods is a matter of judgment as well as knowledge.

10A. Appendix

Why the mean/standard error follows the *t* Distribution

The t Distribution for the ratio 'mean over standard error' arises as follows. We know that $\bar{x}$ has a Normal Distribution with mean μ and variance σ^2/n. Hence $(\bar{x}-\mu)/\sqrt{(\sigma^2/n)}$ will be Normal with mean 0 and variance 1. The distribution of $(n-1)s^2/\sigma^2$ is Chi-squared with $(n-1)$ degrees of freedom (Appendix 7A). If we divide a Standard Normal variable by the square root of an independent Chi-squared variable over its degrees of freedom, we get the t Distribution:

$$\frac{\dfrac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\dfrac{(n-1)s^2/\sigma^2}{n-1}}} = \frac{\dfrac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\dfrac{s^2}{\sigma^2}}}$$

$$= \frac{\bar{x}-\mu}{\sqrt{\dfrac{\sigma^2}{n}} \times \sqrt{\dfrac{s^2}{\sigma^2}}}$$

$$= \frac{\bar{x}-\mu}{\sqrt{\dfrac{\sigma^2}{n} \times \dfrac{s^2}{\sigma^2}}}$$

$$= \frac{\bar{x}-\mu}{\sqrt{\dfrac{s^2}{n}}}$$

As if by magic, we have our sample mean over its standard error. We shall not bother to go into this detail for the other similar ratios which we shall encounter. Any Normally distributed quantity with mean zero (such as $\bar{x}-\mu$), divided by its standard error, will follow a t Distribution provided the standard error is based on one sum of squares and hence is related to the Chi-squared Distribution.

Exercise 10M

(Each branch is either true of false.)

1. The paired *t* test is:

(a) impractical for large samples;
(b) useful for the analysis of qualitative data;
(c) suitable for very small samples;
(d) used for independent samples;
(e) based on the Normal Distribution.

2. Which of the following conditions must be met for a valid *t* test between the means of two samples:

(a) the numbers of observations must be the same in the two groups;
(b) the standard deviations must be approximately the same in the two groups;
(c) the means must be approximately equal in the two groups;
(d) the observations must be from approximately Normal Distributions;
(e) the samples must be small.

3. In a two-sample clinical trial, one of the outcome measures was highly skewed. To test the difference between the levels of this measure in the two groups of patients, possible approaches include:

(a) a standard *t* test using the observations;
(b) a Normal approximation if the sample is large;
(c) transforming the data to Normality and using a *t* test;
(d) a sign test;
(e) the standard error of the difference between two proportions.

4. In the two-sample *t* test, deviation from the Normal Distribution by the data may seriously affect the validity of the test if:

(a) the sample sizes are equal;
(b) the distribution followed by the data is highly skewed;
(c) one sample is much larger than the other;
(d) both samples are large;
(e) the data deviate from Normality because the measurement unit is large and only a few values are possible.

5. If we take samples of size *n* from a Normal Distribution and calculate the sample mean $\bar{x}$ and variance s^2:

(a) samples with large values of $\bar{x}$ will tend to have large s^2;
(b) the sampling distribution of $\bar{x}$ will be Normal;

(c) the sampling distribution of s^2 will be related to the Chi-squared Distribution with $(n-1)$ degrees of freedom;

(d) the ratio $\bar{x}/\sqrt{(s^2/n)}$ will be from a t Distribution with $(n-1)$ degrees of freedom;

(e) the sampling distribution of s will be approximately Normal if $n > 20$.

Exercise 10E

Table 10E.1 shows the total static compliance of the respiratory system and the arterial oxygen tension ($p_a(O_2)$) in 16 patients in intensive care (Al-Saady, personal communication). The patients' breathing was assisted by a respirator and the question was whether their respiration could be improved by varying the characteristics of the air flow. Table 10E.1 compares a constant inspiratory flow waveform with a decelerating inspiratory flow waveform. In this exercise we shall examine the effect of waveform on $p_a(O_2)$ and compliance.

Table 10E.1. $p_a(O_2)$ and compliance for two inspiratory flow waveforms

	$p_a(O_2)$ (kPa)		Compliance (ml/cm H_2O)	
Patient	Constant	Decelerating	Constant	Decelerating
1	9.1	10.8	65.4	72.9
2	5.6	5.9	73.7	94.4
3	6.7	7.2	37.4	43.3
4	8.1	7.9	26.3	29.0
5	16.2	17.0	65.0	66.4
6	11.5	11.6	35.2	36.4
7	7.9	8.4	24.7	27.7
8	7.2	10.0	23.0	27.5
9	17.7	22.3	133.2	178.2
10	10.5	11.1	38.4	39.3
11	9.5	11.1	29.2	31.8
12	13.7	11.7	28.3	26.9
13	9.7	9.0	46.6	45.0
14	10.5	9.9	61.5	58.2
15	6.9	6.3	25.7	25.7
16	18.1	13.9	48.7	42.3

1. Calculate the changes in $p_a(O_2)$. Find a stem and leaf plot. (Hint: you will need both a zero and a minus zero row). Do the differences appear close

enough to the Normal Distribution to apply a t Distribution method? If not, try a transformation.

2. Calculate the mean, variance, standard deviation and standard error of the mean for the $p_a(O_2)$ differences.

3. Calculate a 95 per cent confidence interval for the mean difference. Do you think there is any effect on $p_a(O_2)$?

4. As a check on the validity of the t method, plot the difference against the mean of $p_a(O_2)$. Do they appear to be related?

5. Calculate the differences for compliance. Find a stem and leaf plot (only using the figures before the decimal point) and plot the difference against the mean.

6. Calculate mean, variance, standard deviation and standard error of the mean.

7. Even though the compliance differences are far from a Normal Distribution, calculate the 95 per cent confidence interval using the t Distribution. We shall compare this with that for transformed data.

8. Find the logarithms of the compliance and repeat step 5. Do the assumptions of the t Distribution method apply more closely?

9. Calculate the 95 per cent confidence interval for the log difference and transform back to the original scale. What does this mean and how does it compare to that based on the untransformed data?

10. What can be concluded about the effect of inspiratory waveform on $p_a(O_2)$ and compliance in intensive care patients?

11. Regression and correlation

11.1. Scatter diagrams

In this chapter we shall look at methods of analysing the relationship between two quantitative variables. Consider Table 11.1, which shows data collected by a group of medical students in a physiology class. Inspection of the data suggests that there may be some relationship between FEV1 and height. Before trying to quantify this relationship, we can plot the data and get an idea of their nature. The usual first plot is a *scatter diagram* (Section 5.7). Which variable we choose for which axis depends on our ideas as to the underlying relationship between them, as discussed below. Figure 11.1 shows the scatter diagram for FEV1 and height.

Inspection of Fig. 11.1 suggests that FEV1 increases with height. The next step is to try and draw a line which best represents the relationship. The simplest line is a straight one; we shall consider more complicated relationships later.

The equation of a straight-line relationship between variables x and y is $y = a + bx$, where a and b are constants. In coordinate geometry this is often written as $y = mx + c$, but in statistics a and b are conventional symbols for the coefficients. The first, a, is called the *intercept*. It is the value of y when x is 0. The second, b, is called the *slope* or *gradient* of the line. Their geome-

Table 11.1. FEV1 and height for 20 male medical students

Height (cm)	FEV1 (litre)	Height (cm)	FEV1 (litre)
174.0	4.32	167.0	3.54
180.7	4.80	171.2	3.42
183.7	4.68	177.4	3.60
177.0	5.43	171.3	3.20
177.0	3.09	183.6	4.56
172.0	3.78	183.1	4.78
176.0	3.75	172.0	3.60
177.0	4.05	181.0	3.96
164.0	3.54	170.4	3.19
178.0	2.98	171.2	2.85

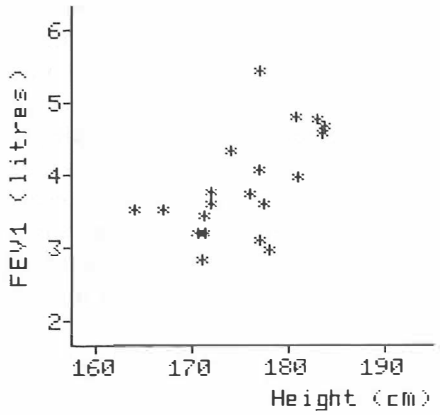


Fig. 11.1 Scatter diagram showing the relationship between FEV1 and height for a group of male medical students.

trical meaning is shown in Fig. 11.2. We can find the values of a and b which best fit the data by regression analysis.

11.2. Regression

Regression is a method of estimating the numerical relationship between variables. For example, we would like to know what is the mean or expected FEV1 for students of a given height, and what increase in FEV1 is associated with a unit increase in height.

The name 'regression' was given by Galton (1886), who developed the technique to investigate the relationship between the heights of people and the

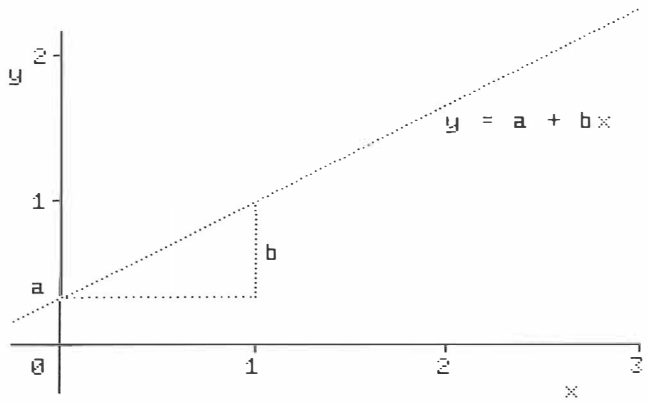


Fig. 11.2 Coefficients of a straight line.

heights of their parents. He observed that if we choose a group of parents of a given height, the mean height of their children will be closer to the mean height of the population than is the given height. In other words, tall parents tend to be taller than their children, short parents tend to be shorter. Galton termed this phenomenon 'regression', meaning 'going back'. It is now called *regression towards the mean*. The method used to investigate it was called regression analysis and the name has stuck. However, in Galton's terminology there was 'no regression' if the relationship between the variables was such that one predicted the other exactly; in modern terminology there is no regression if the variables are not related at all. We have a reversal of meaning.

In regression problems we are interested in how changes in one variable are related to changes in another. In the case of FEV1 and height, for example, we are concerned with how FEV1 changes with height rather than how height changes with lung function. We have two kinds of variables: the *predictor* variable, in this case height, and the *outcome* variable which it predicts, in this case FEV1. The predictor variable is often called the *independent* variable and the outcome variable is called the *dependent* variable. However, these terms have other meanings in probability theory, already mentioned in Chapter 6, so we shall not use them. If we denote the predictor variable by X and the outcome by Y, the relationship between them may be written as

$$Y = a + bX + E$$

where a and b are constants and E is a random variable with mean 0, called the *error*, which represents that part of the variability of Y which is not explained by the relationship with X. If the mean of E were not zero, we could make it so by changing a.

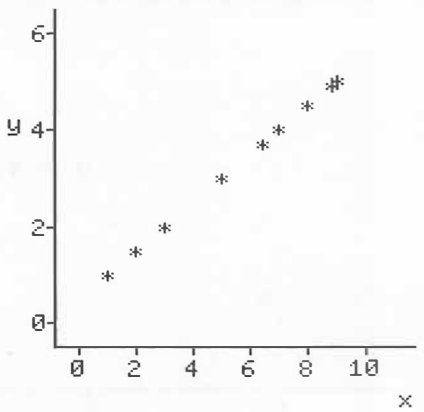


Fig. 11.3 Points lying on a line.

11.3. The method of least squares

If the points all lay along a line and there is no random variation, it would be easy to draw a line on the scatter diagram (Fig. 11.3). In Fig. 11.1 this is not the case. There are many possible values of a and b which could represent the data and we need a criterion for choosing the best line.

Figure 11.4 shows the deviation of a point from the line, the distance from the point to the line in the Y direction. The line will fit the data well if the deviations from it are small, and will fit badly if they are large. These deviations represent the error E, that part of the variable Y not explained by X. One solution to the problem of finding the best line is to choose that which leaves the minimum amount of the variability of Y unexplained, by making the variance of E a minimum. This will be achieved by making the sum of squares of the deviations about the line a minimum. This is called the *method of least squares* and the line found is the least-squares line.

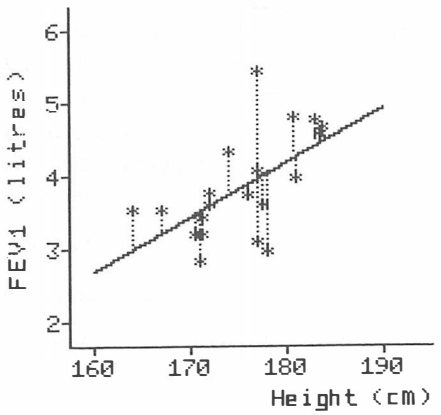


Fig. 11.4 Deviations from the line in the y direction.

The method of least squares is the best method if the deviations from the line are Normally distributed with uniform variance along the line. This is likely to be the case, as the regression tends to remove from Y the variability between subjects and leave the measurement error, which is likely to be Normal. We observed the same process in the paired t method of Section 10.2. We shall deal with deviations from this assumption later in the chapter.

Many users of statistics are puzzled by the minimization of variation in one direction only. Usually both variables are measured with some error and yet we seem to ignore that in X. Why not minimize the perpendicular distances to the line rather than the vertical, as shown in Fig. 11.5? There are two reasons

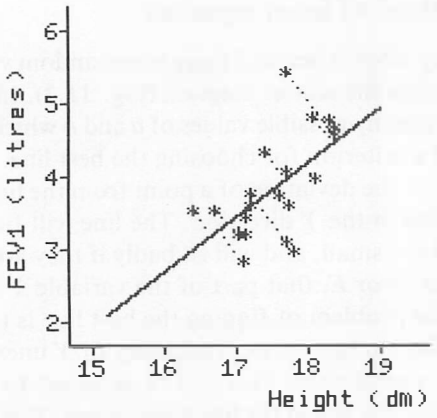


Fig. 11.5 Deviations perpendicular to the line.

for this. First, we are finding the relationship between the observed values of X and Y, not their 'true' values. The measurement error in both variables is one of the causes of deviations from the line, and is included in these deviations measured in the Y direction. Secondly the line found in this way depends on the units in which the variables are measured. We need not burden the reader with the details of unsound methods of analysis; we need only to note that for the data of Table 11.1 the line found by this method is

$$\text{FEV1} = -9.33 + 0.075 \times \text{height}$$

If we measure height in metres instead of centimetres, we get

$$\text{FEV1} = -34.70 + 22.0 \times \text{height}$$

Thus by this method the predicted FEV1 for a student of height 170 cm is 3.45 litres, but for a student of height 1.70 m it is 2.70 litres. This is clearly unsatisfactory and we shall not consider this approach further.

Returning to Fig. 11.4, the equation of the line which minimizes the sum of squared deviations from the line in the outcome variable is found quite easily. The derivation, which involves some simple calculus, is given in Appendix 11A. The solution is:

$$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{\Sigma x_i y_i - \dfrac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}}$$

$$= \frac{\text{sum of products about the mean of } X \text{ and } Y}{\text{sum of squares about the mean of } X}$$

We then find the intercept a by

$$a = \bar{y} - b\bar{x}$$

Notice that the line has to go through the mean point, $(\bar{x}, \bar{y})$. The sum of products about the mean is similar to the sum of squares about the mean derived in Section 4.6. The second form, which is easier for calculator work, is found in the same way. We shall say more about the properties of the sum of products, as it is usually termed, when we discuss correlation. Fitting a straight line by this method is called *simple linear regression*.

The equation $Y = a + bX$ is called the *regression equation of Y on X*, Y being the outcome variable and X the predictor. The gradient, b, is also called the *regression coefficient*. We shall calculate it for the data of Table 11.1. We have

$$\Sigma x_i = 3507.6 \qquad \Sigma x_i^2 = 61\ 5739.24 \qquad n = 20$$
$$\Sigma y_i = 77.12 \qquad \Sigma y_i^2 = 306.8134 \qquad \Sigma x_i y_i = 1\ 3568.18$$

$\bar{x} = 3507.6/20 = 175.38$
$\bar{y} = 77.12/20 = 3.856$

$$\text{sum of squares } X = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$
$$= 61\ 5739.24 - \frac{3507.6^2}{20}$$
$$= 576.32$$

$$\text{sum of squares } Y = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$
$$= 306.8134 - \frac{77.12^2}{20}$$
$$= 9.43868$$

$$\text{sum of products about mean} = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}$$
$$= 1\ 3568.18 - \frac{3507.6 \times 77.12}{20}$$
$$= 42.8744$$

We do not need the sum of squares for Y yet, but we shall later.

$$b = \frac{\text{sum of products about the mean of } X \text{ and } Y}{\text{sum of squares about the mean of } X}$$

$$= \frac{42.8744}{576.352}$$

$$= 0.074389 \text{ litre/cm}$$

$$a = \bar{y} - b\bar{x}$$

$$= 3.856 - 0.074389 \times 175.38$$

$$= -9.19 \text{ litre}$$

Hence the regression equation of FEV1 on height is

$$\text{FEV1} = -9.19 + 0.0744 \times \text{height}$$

Figure 11.6 shows the line drawn on the scatter diagram. The coefficients a and b have dimensions, depending on those of X and Y. If we change the units in which X and Y are measured we also change a and b, but we do not change the line. For example, if height is measured in metres we divide the x_i by 100 and we find that b is multiplied by 100 to give $b = 7.4389$ litre/m. The line is

$$\text{FEV1 (litre)} = -9.19 + 7.44 \times \text{height (m)}$$

This is exactly the same line on the scatter diagram.

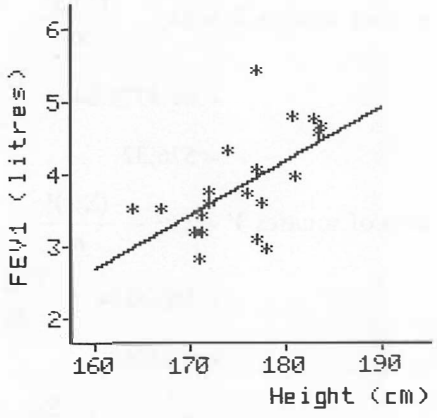


Fig. 11.6 The regression of FEV1 on height.

11.4. The regression of X on Y

What happens if we change our choice of outcome and predictor variables? The regression equation of height on FEV1 is

$$\text{height} = 158 + 4.54 \times \text{FEV1}$$

This is not the same line as the regression of FEV1 on height. For if we rearrange this equation by dividing each side by 4.54 we get

$$0.220 \times \text{height} = 34.8 + \text{FEV1}$$

or

$$\text{FEV1} = -34.8 + 0.220 \times \text{height}$$

The slope of the regression of height on FEV1 is greater than that of FEV1 on height (Fig. 11.7). In general, the slope of the regression of X on Y is greater than that of Y on X, when X is the horizontal axis. Only if all the points lie exactly on a straight line are the two equations the same. This has implications for the choice of outcome variable, which we will consider later.

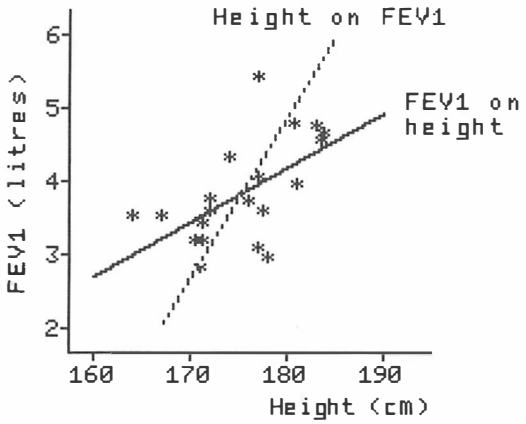


Fig. 11.7 The two regression lines.

11.5. The standard error of the regression coefficient, *b*

In any estimation procedure, we want to know how reliable our estimates are. We do this by finding their standard errors and hence confidence intervals. We can also test hypotheses about the coefficients, for example, the null hypothesis that $b = 0$ and there is no linear relationship. The details are given in Appendix 11A.2.

We first find the sum of squares of the deviations from the line, that is, the difference between the observed y_i and the values predicted by the regression line. This is

$$\Sigma(y_i - \bar{y})^2 - b^2\Sigma(x_i - \bar{x})^2$$

$\Sigma(y_i - \bar{y})^2$ is of course the total sum of squares about the mean of y_i. The term $b^2\Sigma(x_i - \bar{x})^2$ is called the *sum of squares due to the regression* on X. The difference between them is the *residual sum of squares* or *sum of squares about the regression*. The sum of squares due to the regression divided by the total sum of squares is called *the proportion of variability explained by the regression.*

In order to estimate the variance we need the degrees of freedom with which to divide the sum of squares. We have estimated not one parameter from the data, as for the sum of squares about the mean (Section 4.6), but two, a and b. We lose two degrees of freedom, leaving us with $(n - 2)$. Hence the variance of Y about the line, called the *residual variance*, is

$$s^2 = \frac{1}{n-2}\{\Sigma(y_i - \bar{y})^2 - b^2\Sigma(x_i - \bar{x})^2\}$$

For the FEV1 data we have

$$\begin{aligned} s^2 &= \frac{1}{20-2}\{9.43868 - 0.074389^2 \times 576.352\} \\ &= \frac{1}{18} \times 6.2493 \\ &= 0.34718 \end{aligned}$$

The standard error of b is given by

$$\begin{aligned} se(b) &= \sqrt{\frac{s^2}{\Sigma(x_i - \bar{x})^2}} \\ &= \sqrt{\frac{0.34718}{576.352}} \\ &= 0.02454 \text{ litre/cm} \end{aligned}$$

Now, we have already assumed that the error E is Normally distributed, so b must be, too. The standard error is based on a single sum of squares, and we can see that $b/se(b)$ is an observation from the t Distribution with $(n - 2)$ degrees of freedom. Hence we can find a 95 per cent confidence interval for b by taking t standard errors on either side of the estimate.

For the example, we have 18 degrees of freedom. From Table 10.1, the 5 per cent point of the t Distribution is 2.10, so the 95 per cent confidence interval for b is

$$0.074389 - (2.10 \times 0.02454) \text{ to } 0.074389 + (2.10 \times 0.02454)$$
$$= 0.022848 \text{ to } 0.12593$$

or 0.02 to 0.13 litre/cm, rounding off all the meaningless digits. We can see that FEV1 and height are related, though the slope is not very well estimated.

Alternatively, we can test the null hypothesis that $b = 0$ against the alternative that b is not equal to 0, a relationship in either direction. The test statistic is $b/se(b)$ and if the null hypothesis is true this will be from a t Distribution with $(n - 2)$ degrees of freedom.

For the example,

$$t = \frac{b}{se(b)}$$
$$= \frac{0.074389}{0.02454}$$
$$= 3.03$$

From Table 10.1 this has two-tailed probability of less than 0.01. The computer tells us that the probability is about 0.007. Hence the data are inconsistent with the null hypothesis and the data provide fairly good evidence that a relationship exists.

If the sample were much larger, we could dispense with the t Distribution and use the Standard Normal Distribution in its place.

11.6. Using the regression line for prediction

We can use the regression equation to predict the mean or expected Y for any given value of X. This is called the regression estimate of Y. We can use this to say whether any individual has an observed Y greater or less than would be expected given X. For example, the predicted FEV1 for students with height 177 cm is $-9.19 + 0.744 \times 177 = 3.98$ litres. Three subjects had height 177 cm. The first had observed FEV1 of 5.43 litres, 1.45 litres above that expected. The second had a rather low FEV1 of 3.09 litres, 0.89 litres below expectation, while the third with an FEV1 of 4.05 litres was very close to that predicted. We can use this clinically to adjust a measured lung function for height and thus get a better idea of the patient's status. We would, of course, use a much larger sample to establish a precise estimate of the regression equation. We can also use a variant of the method to adjust FEV1 for height in comparing different groups, where we can both remove variation in FEV1 due to variation in height and allow for differences in mean height between the groups. We may wish to do this to compare patients with respiratory disease on different therapies, or to compare subjects exposed to different environmental factors, such as air pollution, cigarette smoking, etc.

As with all sample estimates, the regression estimate is subject to sampling variation. We estimate its precision by standard error and confidence intervals in the usual way. The standard error of the expected Y for an observed value x is

$$\text{s.e.} = \sqrt{s^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma (x_i - \bar{x})^2} \right]}$$

We need not go into the algebraic details of this. It is very similar to that in the previous Section. For $x = 177$ we have

$$\text{s.e.} = \sqrt{0.34718 \left[\frac{1}{20} + \frac{(177 - 175.38)^2}{576.352} \right]}$$
$$= 0.138$$

This gives a 95 per cent confidence interval of $3.98 - (2.10 \times 0.138)$ to $3.98 + (2.10 \times 0.138)$ giving 3.69 litres to 4.27 litres. Here 3.98 is the estimate and 2.10 is the appropriate point of the t Distribution with $(n - 2) = 18$ degrees of freedom.

The standard error is a minimum at $X = \bar{x}$, and increases as we move away from $\bar{x}$ in either direction. It can be useful to plot the standard error and 95 per cent confidence interval about the line on the scatter diagram. Figure 11.8 shows this for the FEV1 data. Notice that the lines diverge considerably as we reach the extremes of the data. It is very dangerous to extrapolate beyond the data. Not only do the standard errors become very wide, but we often have no reason to suppose that the straight-line relationship would persist if we could.

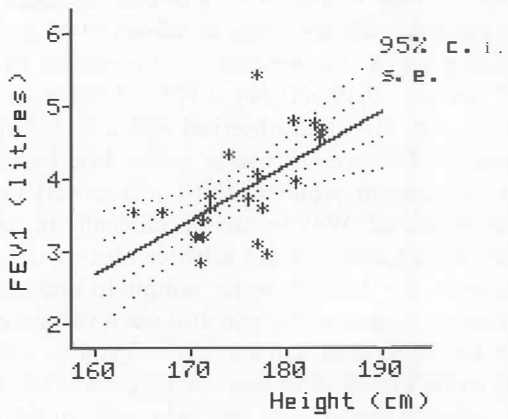


Fig. 11.8 Confidence intervals for the regression estimate.

The intercept a, the predicted value of Y when $X = 0$, is a special case of this. Clearly, we cannot actually have a medical student of height zero and with FEV1 of -9.19 litres. Figure 11.9 shows the confidence interval for the regression estimate with a much smaller scale, to show the intercept. The

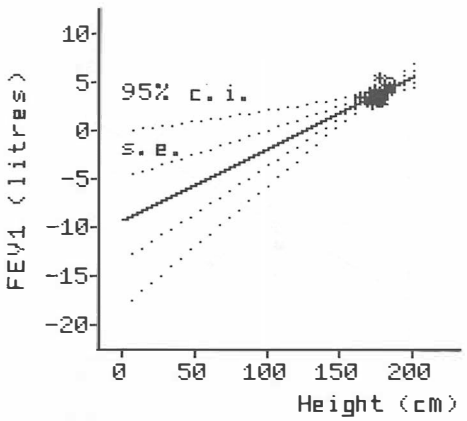


Fig. 11.9 Confidence intervals for the regression estimate, showing the intercept.

confidence interval is very wide at height = 0, and this does not take account of any breakdown in linearity.

We can also use the regression equation of Y on X to predict X from Y. However, this is much less accurate than predicting Y from X. For example, if we use the regression of height on FEV1 (Fig. 11.7) to predict the FEV1 of subjects with height 177 cm, we get a prediction of 4.21 litres, with standard error 0.255. This is almost twice the standard error obtained from the regression of FEV1 on height. Thus we can see that if in doubt about the choice of outcome and predictor variables, the outcome variable should be the one we wish to predict. Only if there is no possibility of deviations in the X direction fulfilling the assumptions of Normality should we consider predicting X from Y.

Rather than predict the expected Y for a given value of X, we may wish to predict the value of Y which we would observe for a given X. In other words, we may wish to use the value of X for a subject to estimate that subject's value of Y, a calibration problem. The estimate is the same as the regression estimate, but the standard error is much greater

$$\text{s.e.} = \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma (x_i - \bar{x})^2} \right]}$$

For a student with a height of 177 cm, the predicted FEV1 is 3.98 litres, with standard error 0.605. Figure 11.10 shows the precision of the prediction of a further observation. As we might expect, the 95 per cent confidence intervals include all but one of the 20 observations. This is only going to be a useful prediction when the residual variance s^2 is small.

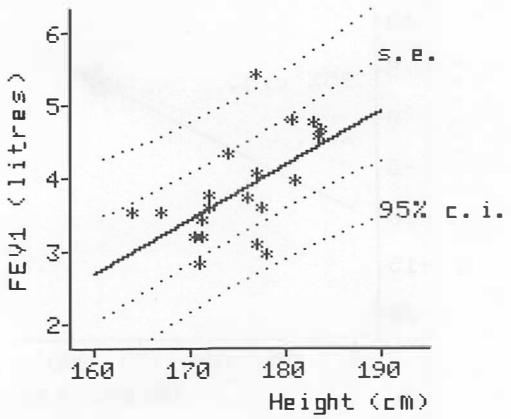


Fig. 11.10 Confidence interval for a further observation.

11.7. Analysis of residuals

It is often very useful to examine the residuals, the differences between the observed and predicted Y. This is best done graphically. We can assess the assumption of Normality by looking at the histogram and frequency distributions. Figure 11.11 shows these for the FEV1 data. The fit is not bad, though there is a suggestion that the distribution is slightly skew to the right.

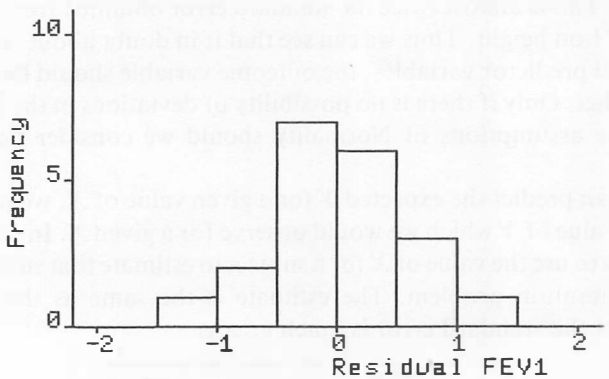


Fig. 11.11 Distribution of residuals for the FEV1 data.

Figure 11.12 shows a plot of residuals against the predictor variable. This plot enables us to examine deviations from linearity. For example, if the true relationship were quadratic, so that Y increases more and more rapidly as X increases, we should see that the residuals are related to X. Large and small X would tend to have positive residuals, whereas central values would have

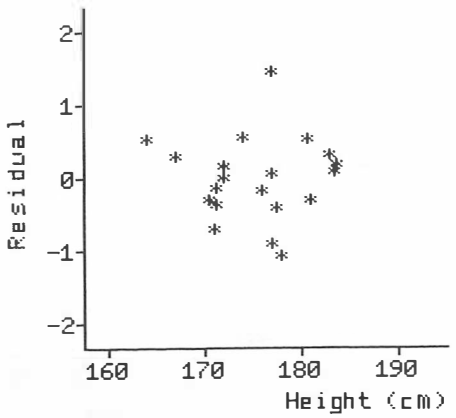


Fig. 11.12 Residuals against height for the FEV1 data.

negative residuals. Figure 11.12 shows no relationship between the residuals and height, and the linear model seems to be an adequate fit to the data.

Figure 11.12 shows something else, however. One point stands out as having a rather larger residual than the others. This may be an *outlier*, a point which may well come from a different population. It is often difficult to know what to do with such data. At least we have been warned to double-check this point for transcription errors. It is all too easy to transpose adjoining digits when transferring data from one medium to another. This may have been the case here, as an FEV1 of 4.53, rather than the 5.43 recorded, would have been more in line with the rest of the data. If this happened at the point of recording, there is not much we can do about it. We could try to measure the subject again, or exclude him and see whether this makes any difference. My own feeling is that, on the whole, we must work with all the data unless there are very good reasons for not doing so. I have accordingly retained this case here.

11.8. Deviations from assumptions in regression

Both the appropriateness of the method of least squares and the use of the t Distribution for confidence intervals and tests of significance depend on the assumption that the residuals are Normally distributed. This assumption is easily met, for the same reasons that it is in the paired t test. The removal of the variation due to X tends to remove some of the variation between individuals, leaving the measurement error. Problems can arise, however, and it is always a good idea to plot the original scatter diagram and the residuals to check that there are no gross departures from the assumptions of the method. Not only does this help preserve the validity of the statistical method used,

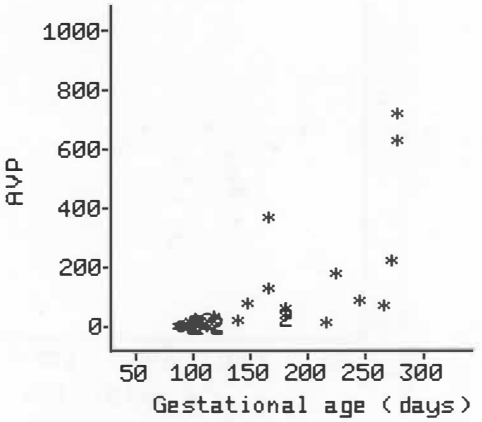


Fig. 11.13 Data which do not meet the conditions of the method of least squares.

but it may also help us learn more about the structure of the data.

Figure 11.13 shows the relationship between gestational age and cord blood levels of AVP, the antidiuretic hormone, in a sample of male fetuses. The variability of the outcome variable AVP depends on the actual value of the variable, being larger for large values of AVP. The assumptions of the method of least squares do not apply. However, we can use a transformation as we did for the comparison of means in Section 10.5. Figure 11.14 shows the data after Y has been log transformed, together with the least-squares line. The transformed data appear quite suitable for least squares linear regression.

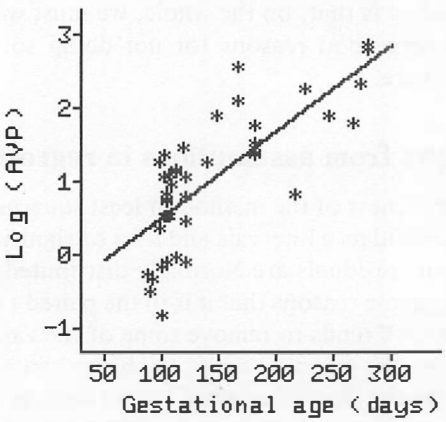


Fig. 11.14 Data of Fig. 11.13 after logarithmic transformation.

11.9. Extensions of the regression method

Often there are several predictor variables related to the outcome and we want to study the effects of all of them together. In a study of factors influencing respiratory disease, we might wish to adjust the FEV1 for other things besides height: respiratory disease in infancy, say, or cigarette smoking. We may wish to see whether differences in FEV1 between, say, registrars and consultants, can be explained by the different age distributions of the groups.

We do this by *multiple regression*, calculating the coefficients of regression equations like

$$\text{FEV1} = a + b \times \text{height} + c \times \text{age}$$

We can have as many predictors as we like, provided we have enough data to calculate the coefficients. The predictors do not have to be continuous. They can be discrete or qualitative. The formulae are complicated, but there are plenty of computer programs to do the job. We can also fit curves rather than straight lines. For example

$$\text{FEV1} = a + b \times \text{height} + \text{c} \times \text{height}^2$$

enables us to fit a quadratic curve to our data. We can then see whether this fits better than the straight line. This method is called *polynomial regression*.

We can fit similar linear expressions to many kinds of non-Normal data, too. The overall method is called the *general linear model*. The details are beyond the scope of this book, but the underlying principle is that of regression, whether the outcome variable is a binomial proportion or qualitative with several categories.

11.10. Correlation

The regression method tells us something about the nature of the relationship between two variables, how one changes with the other, but it does not tell us how close that relationship is. To do this we need a different coefficient, the correlation coefficient. The correlation coefficient is based on the sum of products about the mean of the two variables, so we shall start by considering the properties of the sum of products and why it is a good indicator of the closeness of the relationship.

Take the scatter diagram of Fig. 11.1 and draw two new axes through the mean point (Fig. 11.15). The distances of the points from these axes represent the deviations from the mean. In the top right section of Fig. 11.15, the deviations from the mean of both variables, FEV1 and height, are positive. Hence, their products will be positive. In the bottom left section, the

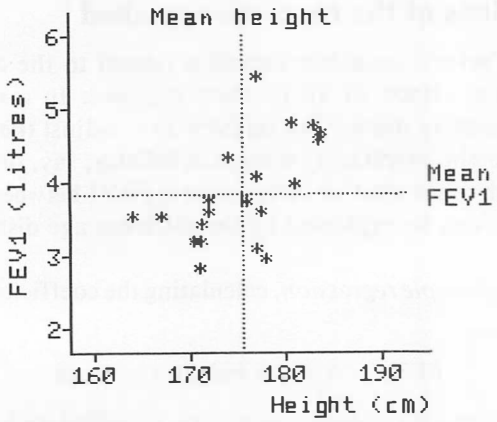


Fig. 11.15 Scatter diagram with axes through the mean point.

deviations from the mean of the two variables will both be negative. Again, their product will be positive. In the top left section of Fig. 11.15, the deviations of FEV1 will be positive, and the deviation of height from its mean will be negative. The product of these will be negative. In the bottom right section, the product will again be negative. So in Fig. 11.15 nearly all these products will be positive, and their sum will be positive. We say that there is a *positive correlation* between the two variables; as one increases so does the other. If one variable decreased as the other increased, we would have a scatter diagram where most of the points lay in the top left and bottom right sections. In this case the sum of the products is negative and there is a *negative correlation* between the variables. When the two variables are not related, we have a scatter diagram with roughly the same number of points in each of the sections. In this case, there are as many positive as negative products, and the sum is zero. There is *zero correlation* or *no correlation*. The variables are said to be *uncorrelated*.

The value of the sum of products depends on the units in which the two variables are measured. We can find a dimensionless coefficient if we divide the sum of products by the square roots of the sums of squares of X and Y. This gives us the *product moment correlation coefficient*, or the *correlation coefficient* for short, usually denoted by r.

If the pairs of observations are denoted by x_i and y_i, and there are n pairs, then r is given by

$$r = \frac{\Sigma(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\,[\Sigma(x_i - \bar{x})^2]\,[\Sigma(y_i - \bar{y})^2]}}$$

which may also be written

$$= \frac{\Sigma x_i y_i - \dfrac{\Sigma x_i \Sigma y_i}{n}}{\sqrt{\left[\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}\right]\left[\Sigma y_i^2 - \dfrac{(\Sigma y_i)^2}{n}\right]}}$$

$$= \frac{\text{sum of products about the mean of } X \text{ and } Y}{\sqrt{\text{sum of squares about the mean of } X \text{ times sum of squares about the mean of } Y}}$$

For the FEV1 and height we have

$$r = \frac{42.8744}{\sqrt{576.352 \times 9.43868}} = 0.58$$

The effect of dividing the sum of products by the root sum of squares of deviations of each variable is to make the correlation coefficient lie between -1.0 and $+1.0$. When all the points lie exactly on a straight line such that Y increases as X increases, $r = 1$. For if $Y = a + bX$, the sum of products will be

$$\begin{aligned}\Sigma(x_i - \bar{x})(y_i - \bar{y}) &= \Sigma(x_i - \bar{x})(a + bx_i - a - b\bar{x}) \\ &= \Sigma(x_i - \bar{x})(bx_i - b\bar{x}) \\ &= b\Sigma(x_i - \bar{x})^2\end{aligned}$$

The sum of squares for Y will be

$$\begin{aligned}\Sigma(y_i - \bar{y})^2 &= \Sigma(a + bx_i - a - b\bar{x})^2 \\ &= \Sigma(bx_i - b\bar{x})^2 \\ &= b^2\Sigma(x_i - \bar{x})^2\end{aligned}$$

So for the correlation coefficient we have

$$\begin{aligned}r &= \frac{b\Sigma(x_i - \bar{x})^2}{\sqrt{[\Sigma(x_i - \bar{x})^2]\,[b^2\Sigma(x_i - \bar{x})^2]}} \\ &= \frac{b\Sigma(x_i - \bar{x})^2}{b\Sigma(x_i - \bar{x})^2} \\ &= 1\end{aligned}$$

When all the points lie exactly on a straight line with negative slope, $r = -1$. When there is no relationship at all, $r = 0$, because the sum of products is zero. The correlation coefficient describes the closeness of the relationship between two variables. It does not matter which variable we take to be Y and which to be X. There is no choice of predictor and outcome variable, as there is in regression.

The correlation coefficient measures how close the points are to a straight line. Even if there is a perfect mathematical relationship between X and Y,

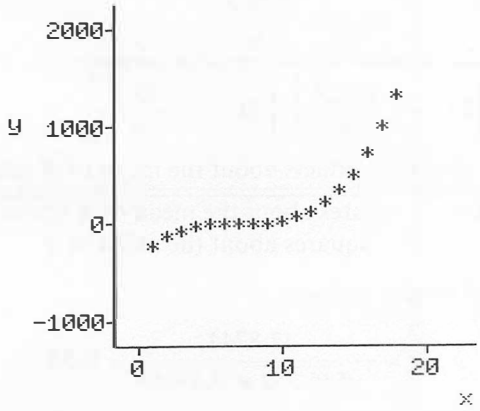


Fig. 11.16 Variables which are perfectly related but with $r < 1$.

the correlation coefficient will not be exactly 1 unless this is of the form $Y = a + bX$. For example, Fig. 11.16 shows two variables which are perfectly related but have $r = 0.86$. Figure 11.17 shows two variables which are clearly related but have zero correlation. This shows again the importance of plotting the data and not relying on summary statistics such as the correlation coefficient only. In practice, relationships like those of Fig. 11.16 and 11.17 are rare in medical statistics, although the possibility is always there. We more often have so much random variation that it is not easy to discern any relationship at all.

The correlation coefficient, r, is related to the regression coefficient, b, in a simple way. If $Y = a + bX$ is the regression of Y on X, and $X = a' + b'Y$ is

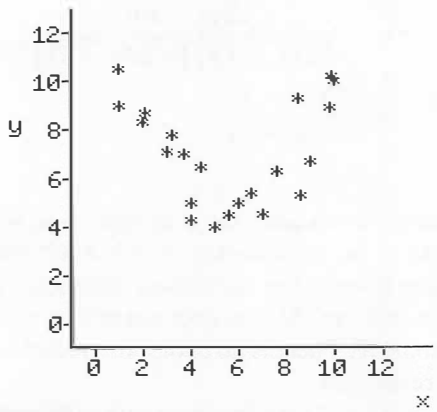


Fig. 11.17 Variables which are related but have zero correlation.

If straight line $R^2 = 1$ & $b = 1/b'$

the regression of X on Y, then $r^2 = bb'$. This arises from the formulae for r and b. For the FEV1 data, $b = 0.074\,389$ and $b' = 4.5424$, so $bb' = 0.074\,389 \times 4.5424 = 0.33790$, the square root of which is 0.58129, the correlation coefficient. We also have

$$r^2 = \frac{(\text{sum of products about mean})^2}{\text{sum of squares of } X \times \text{sum of squares of } Y}$$

$$= \frac{(\text{sum of products about mean})^2}{(\text{sum of squares of } X)^2} \times \frac{\text{sum of squares of } X}{\text{sum of squares of } Y}$$

$$= \frac{b^2 \times \text{sum of squares of } X}{\text{sum of squares of } Y}$$

This is the proportion of variability explained described in Section 11.5.

11.11. Confidence interval and significance test for the correlation coefficient

The correlation coefficient is unusual among sample statistics in having a most awkward sampling distribution. Even when X and Y are both Normally distributed, r does not itself approach a Normal Distribution until the sample size is in the thousands. Furthermore, its distribution is rather sensitive to deviations from the Normal in X and Y. However, Fisher discovered a remarkable transformation called Fisher's z-transformation, which gives a Normally distributed variable whose mean and variance are known in terms of the population correlation coefficient which we wish to estimate. From this a confidence interval can be found. We will omit the details (see Snedecor and Cochran 1980). For the FEV1 data the 95 per cent confidence interval is 0.19 to 0.81. This is very wide, reflecting the wide sampling variation which the correlation coefficient has for small samples. This means that correlation coefficients must be treated with some caution, especially when derived from small samples.

When it comes to testing the null hypothesis that $r = 0$, or that there is no linear relationship, things are much simpler. The test is numerically equivalent to testing the null hypothesis that $b = 0$, and the test is valid provided at least one of the variables is from a Normal Distribution. This condition is the same as that for testing b, where the residuals in the Y direction must be Normal. If $b = 0$, the residuals in the Y direction are simply the deviations from the mean, and these will only be Normally distributed if Y is. If the condition is not met, we can use one of the rank correlation methods described in Sections 12.4 and 12.5.

Because the correlation coefficient does not depend on the means or variances of the observations, the distribution of the sample correlation coefficient when the population coefficient is zero is easy to tabulate. Table 11.2

Table 11.2. Two-sided 5 per cent and 1 per cent points of the distribution of the correlation coefficient, r, under the null hypothesis

Degrees of freedom	5%	1%	Degrees of freedom	5%	1%
1	1.00	1.00	21	0.41	0.53
2	0.95	0.99	22	0.40	0.52
3	0.88	0.96	23	0.40	0.51
4	0.81	0.92	24	0.39	0.50
5	0.75	0.87	25	0.38	0.49
6	0.71	0.83	26	0.37	0.48
7	0.67	0.80	27	0.37	0.47
8	0.63	0.77	28	0.36	0.46
9	0.60	0.74	29	0.36	0.46
10	0.58	0.71	30	0.35	0.45
11	0.55	0.68	40	0.30	0.39
12	0.53	0.66	50	0.27	0.35
13	0.51	0.64	60	0.25	0.33
14	0.50	0.62	70	0.23	0.30
15	0.48	0.61	80	0.22	0.28
16	0.47	0.59	90	0.21	0.27
17	0.46	0.58	100	0.20	0.25
18	0.44	0.56	200	0.14	0.18
19	0.43	0.55	500	0.09	0.12
20	0.42	0.54	1000	0.06	0.08

shows the correlation coefficient at the 5 per cent and 1 per cent level of significance for various degrees of freedom. As in regression, the degrees of freedom are $(n - 2)$. For the example we have $r = 0.58$ from 20 points, so we have 18 degrees of freedom. The 1 per cent point for 18 degrees of freedom is 0.56, so we have $p < 0.01$, and the correlation is unlikely to have arisen by chance. Note that the values of r which can arise by chance with small samples are quite high. With 10 points, 8 degrees of freedom, r would have to be greater than 0.63 to be significant. On the other hand with 1000 points very small values of r, as low as 0.06, will be significant.

The ease of the significance test compared to the relative complexity of the confidence interval calculation has meant that in the past a significance test was usually given for the correlation coefficient. The increasing availability of computers with well-written statistical packages should lead to correlation coefficients appearing with confidence intervals in the future.

11.12. Uses of the correlation coefficient

The correlation coefficient has several uses. Using Table 11.2, it provides a simple test of the null hypothesis that the variables are not linearly related, with less calculation than the regression method. It is also useful as a

summary statistic for the strength of relationship between two variables. This is of great value when we are considering the inter-relationships between a large number of variables. We can set up a square array of the correlations of each pair of variables. This is called the correlation matrix. Examination of the correlation matrix can be very instructive, but we must bear in mind the possibility of non linear relationships. There is no substitute for plotting the data. The correlation matrix also provides the starting point for a number of methods for dealing with a large number of variables simultaneously. We will not discuss these further.

Of course, for the reasons discussed in Chapter 3, the fact that two variables are correlated does not mean that one causes the other.

11A. Appendix

11A.1. Derivation of the least-squares equation

This section requires knowledge of calculus. We want to find a and b so that the sum of squares about the line $y = a + bx$ is a minimum. We therefore want to minimize $\Sigma(y_i - a - bx_i)^2$. This will have a minimum when the partial differentials with respect to a and b are both zero.

$$\frac{\partial}{\partial a} \Sigma(y_i - a - bx_i)^2 = \Sigma 2(y_i - a - bx_i)(-1)$$
$$= -2\Sigma y_i + 2a\Sigma 1 + 2b\Sigma x_i$$
$$= -2\Sigma y_i + 2an + 2b\Sigma x_i$$

This must equal 0 so $\Sigma y_i = na + b\Sigma x_i$

$$\frac{\partial}{\partial b} \Sigma(y_i - a - bx_i)^2 = \Sigma 2(y_i - a - bx_i)(-x_i)$$
$$= -2\Sigma x_i y_i + 2a\Sigma x_i + 2b\Sigma x_i^2$$

This must equal 0 so $\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2$

We multiply the first equation by $\frac{1}{n}\Sigma x_i$

$$\frac{1}{n}\Sigma x_i \Sigma y_i = a\Sigma x_i + \frac{b}{n}(\Sigma x_i)^2$$

Subtracting this from the second equation we get

$$\Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i = b\Sigma x_i^2 - \frac{b}{n}(\Sigma x_i)^2$$

$$\Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i = b\left\{\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2\right\}$$

This gives us

$$b = \frac{\Sigma x_i y_i - \dfrac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}}$$

If we divide the first equation by n we get the formula for a

$$\begin{aligned} \Sigma y_i &= na + b\Sigma x_i \\ \frac{1}{n}\Sigma y_i &= a + \frac{b}{n}\Sigma x_i \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

Hence the line passes through the mean point.

11A.2. The standard error of b and the variance about the line

To find the standard error of b, we must bear in mind that in our regression model all the random variation is in Y. We first rewrite the sum of products:

$$\begin{aligned} \Sigma(x_i - \bar{x})(y_i - \bar{y}) &= \Sigma\{(x_i - \bar{x})y_i - (x_i - \bar{x})\bar{y}\} \\ &= \Sigma(x_i - \bar{x})y_i - \Sigma(x_i - \bar{x})\bar{y} \\ &= \Sigma(x_i - \bar{x})y_i - \bar{y}\Sigma(x_i - \bar{x}) \\ &= \Sigma(x_i - \bar{x})y_i \end{aligned}$$

This is because $\bar{y}$ is the same for all i and so comes out of the summation, and $\Sigma(x_i - \bar{x}) = 0$. We now find the variance of the sampling distribution of b by

$$\begin{aligned} \text{Var}\ (b) &= \text{Var}\left[\frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}\right] \\ &= \text{Var}\left[\frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}\right] \\ &= \frac{1}{[\Sigma(x_i - \bar{x})^2]^2}\Sigma\,\text{Var}\,[(x_i - \bar{x})y_i] \end{aligned}$$

The variance of a constant times a random variable is the square of the constant times the variance of the random variable (Section 6.6). The x_i are constants, not random variables. So

$$\text{Var}\ (b) = \frac{1}{[\Sigma(x_i - \bar{x})^2]^2}\Sigma(x_i - \bar{x})^2\,\text{Var}\ (y_i)$$

Var (y_i) is the same for all y_i, say Var $(y_i) = s^2$. Hence

$$\text{Var}(b) = \frac{s^2}{\Sigma(x_i - \bar{x})^2}$$

The standard error of b is the square root of this. Next we find s^2. The regression model is $Y = a + bX + E$, and a and b are constants. We are predicting Y for given X, so there is no random variation in X; all the random variation is in E. Hence $s^2 = \text{Var}(Y) = \text{Var}(E)$. We have seen in Section 11.3 that the error E is the random variable which stands for the deviations from the line in the Y direction. These deviations are $y_i - (a + bx_i)$, since $a + bx_i$ is the y value for the line at $x = x_i$. The sum of squares of these deviations is found by a mathematical trick, replacing a by $\bar{y} - b\bar{x}$.

$$\begin{aligned}
\Sigma[y_i - (a + bx_i)]^2 &= \Sigma[y_i - (\bar{y} - b\bar{x} + bx_i)]^2 \\
&= \Sigma[y_i - \bar{y} - (bx_i - b\bar{x})]^2 \\
&= \Sigma[y_i - \bar{y} - b(x_i - \bar{x})]^2 \\
&= \Sigma[(y_i - \bar{y})^2 - 2b(y_i - \bar{y})(x_i - \bar{x}) + b^2(x_i - \bar{x})^2] \\
&= \Sigma(y_i - \bar{y})^2 - 2b\Sigma(y_i - \bar{y})(x_i - \bar{x}) + b^2\Sigma(x_i - \bar{x})^2 \\
&= \Sigma(y_i - \bar{y})^2 - 2b \times b\Sigma(x_i - \bar{x})^2 + b^2\Sigma(x_i - \bar{x})^2 \\
&= \Sigma(y_i - \bar{y})^2 - b^2\Sigma(x_i - \bar{x})^2
\end{aligned}$$

This is because

$$b = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

So

$$b\Sigma(y_i - \bar{y})(x_i - \bar{x}) = b^2\Sigma(x_i - \bar{x})^2$$

Exercise 11M

(Each branch is either true or false.)

1. The product moment correlation coefficient, r:

(a) must lie between -1 and $+1$;

(b) can only have a valid significance test carried out when at least one of the variables is from a Normal Distribution;

(c) is half when there is no relationship;

(d) depends on the choice of dependent variables;

(e) measures the magnitude of the change in one variable associated with a change in the other.

2. A simple linear regression equation:

(a) describes a line which goes through the origin;

(b) describes a line with zero slope;

(c) is not affected by changes of scale;

(d) describes a line which goes through the mean point;

(e) is affected by the choice of dependent variable.

3. If a *t* test is used to test the significance of the slope of a regression line:

(a) deviations from the line in the independent variable must follow a Normal Distribution;

(b) deviations from the line in the dependent variable must follow a Normal Distribution;

(c) the variance about the line is assumed to be the same throughout the range of the predictor variable;

(d) the *y* variable must be log transformed;

(e) all the points must lie on the line.

4. In Fig. 11M.1:

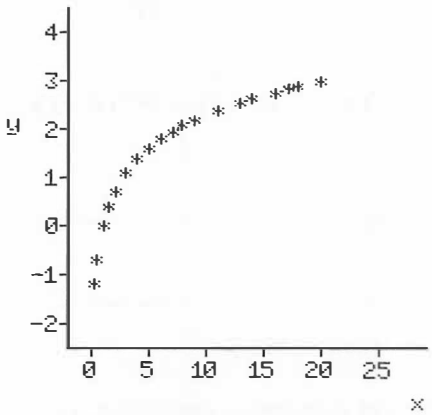


Fig. 11M.1 A scatter diagram.

(a) *x* and *y* are independent;

(b) *x* and *y* are uncorrelated;

(c) the correlation between *x* and *y* is less than 1;

(d) *x* and *y* are perfectly related;

(e) the relationship is best estimated by simple linear regression.

5. In Fig. 11M.2:

(a) x and y are independent random variables;

(b) x and y are uncorrelated;

(c) y increases as x increases;

(d) x and y are linearly related;

(e) the relationship between x and y could be studied by polynomial regression.

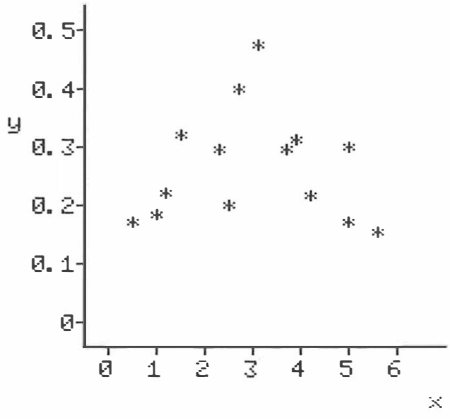


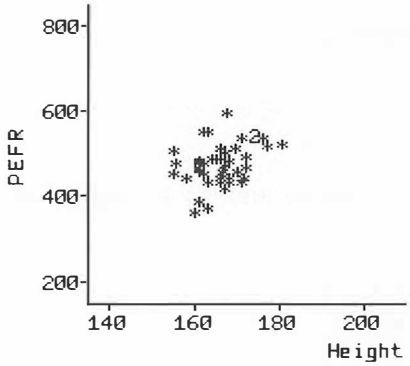
Fig. 11M.2 A scatter diagram.

Exercise 11E

Figure 11E.1 shows scatter diagrams and summary statistics for PEFR and height measured in a group of medical students. There appears to be a relationship between height and PEFR and a difference in both height and PEFR between the sexes. We shall use regression to examine whether the greater PEFR of males is explained by their greater height or whether there is a difference in PEFR quite apart from that produced by height.

We shall fit the regression line of PEFR on height for females, for males and for both together. If there is no difference in PEFR apart from that due to height then the three lines will be the same, except for random variation (Fig. 11E.2(a)). If males have a higher PEFR than females of the same height then the slope for the combined group will be greater than those for the two sexes separately (Fig. 11E.2(b)). If males have a lower PEFR than females of the same height then the slope for the combined group will be less than those for the sexes separately.

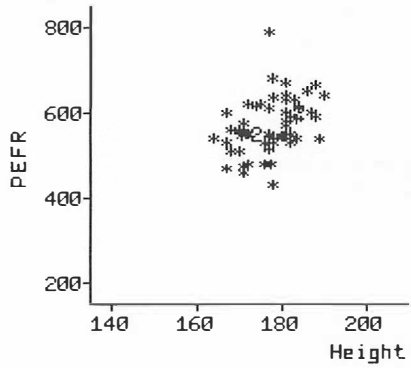
Females:



SUMMARY STATISTICS

	Height	PEFR
NUMBER	43	43
MEAN	165.937209	474.069768
MEDIAN	166	475
MINIMUM	155	360
MAXIMUM	180.6	595
VARIANCE	34.3952492	2407.32502
ST. DEV.	5.8647463	49.0644986
S.E. MEAN	.894365427	7.48226588
SUM OF SQS	1444.60047	101107.651
SUM OF PRODUCTS	4206.94837	

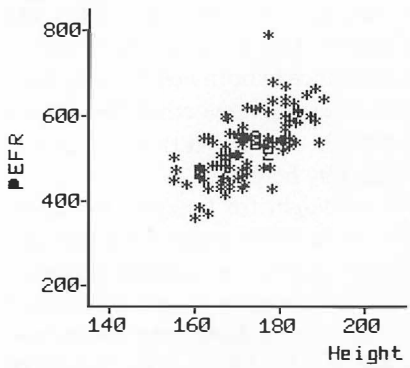
Males:



SUMMARY STATISTICS

	Height	PEFR
NUMBER	58	58
MEAN	177.303448	568.2
MEDIAN	177.2	551.5
MINIMUM	164	430
MAXIMUM	190	792
VARIANCE	39.7806897	3980.24315
ST. DEV.	6.30719349	63.0891683
S.E. MEAN	.828175079	8.28401364
SUM OF SQS	2267.49931	226873.86
SUM OF PRODUCTS	8993.36	

All:



SUMMARY STATISTICS

	Height	PEFR
NUMBER	101	101
MEAN	172.464356	528.124753
MEDIAN	172	530
MINIMUM	155	360
MAXIMUM	190	792
VARIANCE	69.0223172	5467.74469
ST. DEV.	8.30796709	73.9441999
S.E. MEAN	.826673623	7.35772289
SUM OF SQS	6902.23172	546774.469
SUM OF PRODUCTS	39619.5891	

Fig. 11E.1 PEFR (litre/min) and height (cm) for a group of medical students.

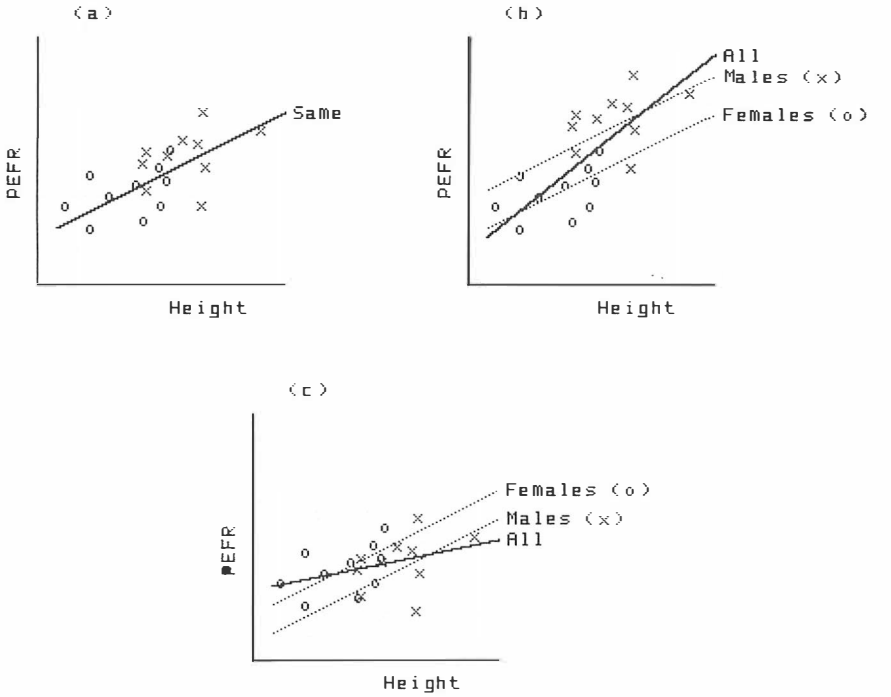


Fig. 11E.2 Hypothetical regressions of PEF on height for three possible sex effects.

1. From the means and sums of squares and products about the mean given, calculate the slope and intercept for each set of data.

2. Does the difference in PEFR between males and females appear to be explained by the difference in height?

3. For males and for females, calculate the correlation coefficient between PEFR and height.

4. If the relationship between PEFR and height is the same for males and females, why should the correlation coefficients be different?

12. Methods based on rank order

12.1. Non-parametric methods

In Chapters 10 and 11 we described a number of methods of analysis which relied on the assumption that the data came from a Normal Distribution. To be more precise, we could say the data come from one of the Normal family of distributions, the particular Normal Distribution involved being defined by its mean and standard deviation, the parameters of the distribution. In these methods we estimate the parameters of the underlying Normal Distribution.

Methods which do not assume a particular family of distributions for the data are said to be *non-parametric*. The *t* Distribution methods in Chapter 10 and regression and correlation in Chapter 11 are *parametric* methods, because the data are assumed to come from the family of Normal Distributions. The parameters are the mean and standard deviation of the particular Normal Distribution which the data follow. In this and the next chapter we shall consider some non-parametric tests of significance. There are many others, but these will illustrate the general principle.

We have already met one non-parametric test, the sign test (Section 9.2). The large-sample Normal test could also be regarded as non-parametric.

It is useful to distinguish between three types of measurement scales:

(a) Interval scale This means that the size of the difference between two values on the scale has a meaning. For example, the difference in temperature between 1 °C and 2 °C is the same as the difference between 31 °C and 32 °C.

(b) Ordinal scale Observations are ordered, but differences may not have a meaning. For example, anxiety and neuroticism are often measured using sets of questions, the number of positive answers giving the anxiety scale. A set of 36 questions would give a scale from 0 to 36. The difference in anxiety between scores of 1 and 2 is not necessarily the same as the difference between scores 31 and 32.

(c) Nominal scale We have a qualitative or categorical variable, where individuals are grouped but not necessarily ordered. Eye colour is a good example.

There are intermediate cases, such as nominal scales with ordered categories, but these will suffice for our purposes. All the methods of Chapters 10 and 11 apply to interval data, being based on differences of observations from the mean. The methods in this chapter apply to ordinal data. Any interval scale which does not meet the requirements of Chapters 10 and 11 may be treated as ordinal, since it is, of course, ordered. This is the more common application in medical work.

General texts such as Armitage (1973), Snedecor and Cochran (1980) and Colton (1974) tend not to go into a lot of detail about rank and related methods. A handbook unsurpassed for clarity and ease of use is Seigel (1956), which is highly recommended. For a more up-to-date account, try Conover (1980).

12.2. The Mann–Whitney *U* test

This is the non-parametric equivalent of the two sample *t* test (Section 10.3). It works like this. Suppose we have anxiety scores on a 37-point scale, 0 to 36, obtained from two groups of individuals (hypothetical data):

A	7	4	9	17
B	11	6	21	14

We want to know whether there is any evidence that A and B are drawn from populations with different levels of anxiety. The null hypothesis is that there is no tendency for members of one population to exceed members of the other. The alternative is that there is such a tendency, in either direction.

First we arrange the observations in ascending order, i.e. we rank them:

4	6	7	9	11	14	17	21
A	B	A	A	B	B	A	B

We now choose one group, say A. For each A, we count how many Bs precede it. For the first A, 4, no Bs precede. For the second, 7, one B precedes, for the third A, 9, one B, for the fourth, 17, three Bs. We add these numbers of preceding Bs together:

$$U = 0 + 1 + 1 + 3 = 5$$

Now, if U is very small, nearly all the As are less than nearly all the Bs. If U is large, nearly all As are greater than nearly all Bs. Moderate values of U mean that As and Bs are closely mixed. Consider two further samples:

C	2	3	3	5
D	26	30	19	25

For A and C we have:

2	3	3	4	5	7	9	17
C	C	C	A	C	A	A	A

Counting Cs for each A, we have:

$$U = 3 + 4 + 4 + 4 = 15$$

The value of U is large; A and C appear different. For A and D we have:

4	7	9	17	19	25	26	30
A	A	A	A	D	D	D	D

Here we have:

$$U = 0 + 0 + 0 + 0 = 0$$

The value of U is very small; A and D appear different.

Now, if we know the distribution of U under the null hypothesis that the samples come from the same population, we can say with what probability these data could have arisen if there were no difference. We can carry out the test of significance. The distribution of U under the null hypothesis can be found easily. The two sets of four observations can be arranged in 70 different ways, from AAAABBBB to BBBBAAAA. Under the null hypothesis these arrangements are all equally likely and, hence, have probability 1/70. Each has its value of U, from 0 to 16, and by counting the number of arrangements which give each value of U, we can find the probability of U. For example, $U = 0$ only arises from the order AAAABBBB and so has probability $1/70 = 0.014$. The result $U = 1$ only arises from AAABABBB and so has probability $1/70 = 0.014$ also. The result $U = 2$ can arise in two ways: AAABBABB and AABAABBB. It has probability $2/70 = 0.029$. The full set of probabilities is shown in Table 12.1.

Table 12.1. Distribution of the Mann–Whitney U statistic, for two samples of size 4

U	probability	U	probability
0	0.014	9	0.100
1	0.014	10	0.100
2	0.029	11	0.071
3	0.043	12	0.071
4	0.071	13	0.043
5	0.071	14	0.029
6	0.100	15	0.014
7	0.100	16	0.014
8	0.114		

We apply this to our three examples. For A and B, $U = 5$ and the probability of this is 0.071. As we did for the sign test (Section 8.2) we consider the probability of more extreme values of U, $U = 5$ or less, which is $0.071 + 0.071 + 0.043 + 0.029 + 0.014 + 0.014 = 0.242$. This gives a one-sided test. For a two-sided test, we must consider the probabilities of a difference as extreme in the opposite direction. We can see from Table 12.1 that the distribution of U is symmetrical, so the probability of an equally extreme value in the opposite direction is also 0.242, hence the two sided probability is $0.242 + 0.242 = 0.484$. This is clearly likely to have happened by chance and so the two samples could have come from the same population.

For A and C, $U = 15$, probability $U \geqslant 15$ is $0.014 + 0.014 = 0.028$, and for a two-sided test $0.028 + 0.028 = 0.056$. By the usual criterion, we would just accept the null hypothesis. For A and D, $U = 0$, probability $U \leqslant 0$ is 0.014, for a two-sided test we have $0.014 + 0.014 = 0.028$ which is significant at the 0.05 level. With such small samples, of course, only the most extreme outcomes are unlikely to happen by chance.

We shall now consider the Mann–Whitney U test in practice. There is no need to carry out the summation of probabilities described above, as these are already tabulated. Table 12.2 shows the 5 per cent points of U for each combination of sample sizes n_1 and n_2 up to 20.

For our groups A and B, $U = 5$, we turn to the $n_2 = 4$ section, and find the $n_1 = 4$ column. From this we see that the 5 per cent point for U is 0, and so $U = 5$ is not significant.

For groups A and C we have a problem. $U = 15$ is an extreme value, but only low values of U are tabulated. Note that the maximum size of U is 16:

$$\begin{array}{l} \text{C C C C A A A A} \\ U = 4 + 4 + 4 + 4 = 16 \end{array}$$

or, in general, $n_1 \times n_2$. We can see this if we consider the extreme arrangement where all n_1 Cs are less than all n_2 As:

$$\underbrace{\text{C C C} \ldots \text{C}}_{n_1} \quad \underbrace{\text{A A A} \ldots \text{A}}_{n_2}$$

Each A has n_1 Cs before it so U is given by:

$$U = \underbrace{n_1 + n_1 + n_1 + \ldots + n_1}_{n_2} = n_1 n_2$$

As the distribution is symmetrical, the probability that $U \geqslant r$ = probability that $(n_1 n_2 - U) \leqslant n_1 n_2 - r$, which is tabulated. The two possible values of U are related by $U + U' = n_1 n_2$. So we subtract U from $n_1 n_2$ to give $16 - 15 = 1$. We see that the probability is just over 5 per cent. For A and D, $U = 0$, and this is significant at the 5 per cent level.

Table 12.2. Two-sided 5 per cent points for the distribution of U, lower value, in the Mann–Whitney U test

n_1 \ n_2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	–	–	0	0	0	0	1	1	1	1	1	2	2	2	2
3	–	–	–	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	–	–	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	–	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	–	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	–	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

If U is less than or equal to the tabulated value the difference is significant.

Table 12.3. Biceps skinfold thickness (mm) in two groups of patients

Crohn's disease		Coeliac disease
1.8	4.2	1.8
2.2	4.4	2.0
2.4	4.8	2.0
2.5	5.6	2.0
2.8	6.0	3.0
2.8	6.2	3.8
3.2	6.6	4.2
3.6	7.0	5.4
3.8	10.0	7.6
4.0	10.4	

We can now turn to the practical analysis of some real data. Consider the biceps skinfold thickness data of Table 10.4, reproduced as Table 12.3. We will analyse this using the Mann–Whitney U test. Denote the Crohn's disease group by A and the coeliac group by B. The joint order is as follows:

1.8	1.8	2.0	2.0	2.0	2.2	2.4	2.5	2.8	2.8
A	B	B	B	B	A	A	A	A	A
3.0	3.2	3.6	3.8	3.8	4.0	4.2	4.2	4.4	4.8
B	A	A	A	B	A	A	B	A	A
5.4	5.6	6.0	6.2	6.6	7.0	7.6	10.0	10.4	
B	A	A	A	A	A	B	A	A	

Let us count the As before each B. Immediately we have a problem. The first A and the first B have the same value. Does the first A come before the first B or after it? We resolve this dilemma by counting $\frac{1}{2}$ for the tied A. The ties between the second, third and fourth Bs do not matter, as we can count the number of As before each without difficulty. We have for the U statistic:

$$U = \tfrac{1}{2} + 1 + 1 + 1 + 6 + 8\tfrac{1}{2} + 10\tfrac{1}{2} + 13 + 18 = 59\tfrac{1}{2}$$

This is the lower value, since $n_1 n_2 = 9 \times 20 = 180$ and so the middle value is 90. We can therefore refer it to Table 12.2. The critical value at the 5 per cent level is for groups size 9 and 20 is 48, which our value exceeds. Hence the difference is not significant at the 5 per cent level and the data are consistent with the null hypothesis that there is no tendency for members of one population to exceed members of the other. This is the same as the result of the t test of Section 10.4.

For larger values of n_1 and n_2, calculation of U can be rather tedious. A

simple formula for U can be found using the ranks. The rank of the lowest observation is 1, of the next is 2, and so on. If a number of observations are tied, each having the same value and hence the same rank, we give each the average of the ranks they would have were they ordered. For example, in the skinfold data the first two observations are each 1.8. They each receive rank $(1 + 2)/2 = 1\frac{1}{2}$. The third, fourth and fifth are tied at 2.0, giving each of them rank $(3 + 4 + 5)/3 = 4$. The sixth, 2.2, is not tied and so has rank 6.

The ranks for the skinfold data are as follows:

skinfold	1.8	1.8	2.0	2.0	2.0	2.2	2.4	2.5	2.8	2.8
group	A	B	B	B	B	A	A	A	A	A
rank	$1\frac{1}{2}$	$1\frac{1}{2}$	4	4	4	6	7	8	$9\frac{1}{2}$	$9\frac{1}{2}$
		r_1	r_2	r_3	r_4					
skinfold	3.0	3.2	3.6	3.8	3.8	4.0	4.2	4.2	4.4	4.8
group	B	A	A	A	B	A	A	B	A	A
rank	11	12	13	$14\frac{1}{2}$	$14\frac{1}{2}$	16	$17\frac{1}{2}$	$17\frac{1}{2}$	19	20
	r_5				r_6			r_7		
skinfold	5.4	5.6	6.0	6.2	6.6	7.0	7.6	10.0	10.4	
group	B	A	A	A	A	A	B	A	A	
rank	21	22	23	24	2	2	27	28	29	
	r_8						r_9			

We denote the ranks of one group by $r_1, r_2, \ldots, r_{n_1}$. The number of As preceding the first B must be $(r_1 - 1)$, since there are no Bs before it and it is the r_1th observation. The number of As preceding the second B is $(r_2 - 2)$, since it is the r_2th observation, and one preceding observation is a B. Similarly, the number preceding the third B is $(r_3 - 3)$, and the number preceding the ith B is $(r_i - i)$. Hence we have:

$$\begin{aligned} U &= \sum_{i=1}^{n_1} (r_i - i) \\ &= \sum_{i=1}^{n_1} r_i - \sum_{i=1}^{n_1} i \\ &= \sum_{i=1}^{n_1} r_i - \frac{n_1(n_1 + 1)}{2} \end{aligned}$$

That is, we add together the ranks of all the n_1 observations, subtract $n_1(n_1 + 1)/2$ and we have U. For the example, we have

$$\begin{aligned} U &= 1\tfrac{1}{2} + 4 + 4 + 4 + 11 + 14\tfrac{1}{2} + 17\tfrac{1}{2} + 21 + 27 - \frac{9 \times (9 + 1)}{2} \\ &= 104\tfrac{1}{2} - 45 = 59\tfrac{1}{2} \end{aligned}$$

as before.

This formula is sometimes written

$$U' = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum_{i=1}^{n_1} r_i$$

But this is simply based on the other group, since $U + U' = n_1 n_2$. For testing we use the smaller value, as before.

As n_1 and n_2 increase, the calculation of the exact probability distribution becomes more difficult. When we cannot use Table 12.2, we use a large-sample approximation instead. Because U is found by adding together a number of independent, identically distributed random variables, the central limit theorem (Section 7.2) applies. The distribution of U approximates to a Normal Distribution. The mean is $\frac{1}{2}n_1 n_2$ and the standard deviation is

$$\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

Hence

$$\frac{U - \dfrac{n_1 n_2}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$$

is an observation from a Standard Normal Distribution. For the example, $n_1 = 9$ and $n_2 = 20$, we have

$$\frac{n_1 n_2}{2} = \frac{9 \times 20}{2} = 90$$

and

$$\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{9 \times 20 \times 30}{12}}$$
$$= \sqrt{450}$$
$$= 21.21$$

$$\frac{U - \dfrac{n_1 n_2}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}} = \frac{59.5 - 90}{21.21}$$
$$= -1.44$$

This gives a two-sided probability from Table 7.1 of 0.15, which we can compare with 0.21 for the two sample t test on the untransformed data and 0.15 for the log-transformed data.

Neither Table 12.2 nor the above formula for standard deviation take ties

into account; both assume the data can be fully ranked. Their use for data with ties is an approximation. For small samples we must accept this. For the Normal approximation there is a truly daunting formula which takes this into account (see Seigel 1956). Thus the Mann–Whitney U test is not free of assumptions which may be violated. We assume that the data can be fully ordered, which in the case of ties is not so.

The Mann–Whitney U test is a non-parametric equivalent of the two-sample t test. The advantage over the t test is that the only assumption about the distribution of the data is that the observations can be ranked, whereas for the t test we must assume the data are from Normal Distributions with uniform variance. There are three disadvantages. For data which are Normally distributed, the U test is less powerful than the t test, i.e. the t test, when valid, can detect smaller differences for given sample size. However, the U test is almost as powerful for moderate and large sample sizes, and usually this difference is not important. The U test gives no idea of the size of the difference. It is purely a test of significance. The t test also enables us to estimate the size of the difference and give confidence intervals.

There are other non-parametric tests which test the same or similar null hypotheses. Two of these, the Wilcoxon two-sample test and the Kendall tau test, are different versions of the Mann–Whitney U test which were developed around the same time and later shown to be identical.

12.3. The Wilcoxon matched-pairs test

This test is an equivalent of the paired t test. We have a sample measured under two conditions and the null hypothesis is that there is no tendency for

Table 12.4. Results of a trial of pronethalol for the prevention of angina pectoris, (Pritchard *et al.* 1963)

Number of attacks while on		Difference
Placebo	Pronethalol	Placebo – Pronethalol
71	29	42
323	348	−25
8	1	7
14	7	7
23	16	7
34	25	9
79	65	14
60	41	19
2	0	2
3	0	3
17	15	2
7	2	5

the outcome on one condition to be higher or lower than the other. The alternative hypothesis is that the outcome on one condition tends to be higher or lower than the other.

Consider the data of Table 12.4, previously discussed in Sections 2.4 and 8.2, where we used the sign test for the analysis. In the sign test, we have ignored the magnitude of differences, and only considered their signs. If we can use information about the magnitude, we would hope to have a more powerful test. To avoid making assumptions about the distribution of the differences, we use their rank order in a similar manner to the Mann–Whitney U test.

First, we rank the differences by their absolute values, i.e. ignoring the sign. From the data of Table 12.4 we have:

difference	2	2	3	5	7	7	7	9	14	19	−25	42
rank	$1\frac{1}{2}$	$1\frac{1}{2}$	3	4	6	6	6	8	9	10	11	12
ranks of +ve differences	$1\frac{1}{2}$	$1\frac{1}{2}$	3	4	6	6	6	8	9	10		12
ranks of −ve differences											11	

We now sum the ranks of the positive differences, $1\frac{1}{2} + 1\frac{1}{2} + 3 + 4 + 6 + 6 + 6 + 8 + 9 + 10 + 12 = 67$, and the ranks of the negative differences, 11. If the null hypothesis were true and there was no difference, we would expect the rank sums for positive and negative differences to be about the same, equal to 39 (their average). The test statistic is the lesser of these sums, T. The smaller T is, the lower the probability of the data arising by chance.

The distribution of T when the null hypothesis is true can be found by enumerating all the possibilities, as described for the Mann–Whitney U statistic. Table 12.5 gives the 5 per cent and 1 per cent points for this distribution, for sample size N up to 25. For the example, $N = 12$ and so the difference would be significant at the 5 per cent level if T were less than or equal to 14. We have $T = 11$, so the data are not consistent with the null hypothesis. The data support the view that there is a real tendency for patients to have fewer attacks while on the active treatment.

From Table 12.5, we can see that the probability ($T < 11$) lies between 0.05 and 0.01. This is greater than the probability given by the sign test, which was 0.006 (Section 9.2). This is surprising, as we would expect greater power, and hence lower probabilities when the null hypothesis is false, when we use more of the information. This greater probability reflects the fact that the one negative difference, −25, is large. Examination of the original data shows that this individual had very large numbers of attacks on both treatments, and it seems at least possible he may belong to a different population from the other eleven.

Table 12.5. Two-sided 5 per cent and 1 per cent points of the distribution of T (lower value) in the Wilcoxon one-sample test

Sample size n	Probability that T is as far or further from the expected than the tabulated value 5%	1%
6	1	none
7	2	none
8	4	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	19
17	35	23
18	40	28
19	46	32
20	52	37
21	59	43
22	66	49
23	73	55
24	81	61
25	90	68

Like Table 12.2, Table 12.5 is based on the assumption that the differences can be fully ranked and there are no ties. Ties may occur in two ways in this test. First, ties may occur in the ranking sense. In the example we had two differences of +2 and three of +7. These were ranked equally: $1\frac{1}{2}$ and $1\frac{1}{2}$, and 6, 6 and 6. When ties are present between negative and positive differences, Table 12.5 only approximates to the distribution of T.

Ties may also occur between the observations under the two conditions, where the observed difference is zero. In the same way as for the sign test, we omit zero differences (Section 8.2). The test is done using the number of non-zero differences only to enter Table 12.5. This seems odd, in that a lot of zero differences would appear to support the null hypothesis. For example, if in Table 12.4 we had another dozen patients with zero differences, the calculation and conclusion would be the same. However, the observed difference would be smaller and the Wilcoxon test tells us nothing about the size of the difference, only about its existence. This illustrates the danger of allowing significance tests to outweigh all other ways of looking at the data.

As N increases, the distribution of T under the null hypothesis tends

towards a Normal Distribution, as does that of Mann–Whitney U statistic. The sum of all the ranks, irrespective of sign is $\frac{1}{2}N(N+1)$, so the expected value of T under the null hypothesis is $\frac{1}{4}N(N+1)$, since the two sums should be equal. The standard deviation of T is

$$\sqrt{\frac{N(N+1)\,(2N+1)}{24}}$$

Hence

$$\frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{N(N+1)\,(2N+1)}{24}}}$$

is from a Standard Normal Distribution if the null hypothesis is true. For the example of Table 12.4, we have:

$$\frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{N(N+1)\,(2N+1)}{24}}} = \frac{11 - \frac{1}{4} \times 12 \times 13}{\sqrt{\frac{12 \times 13 \times 25}{24}}}$$
$$= -2.197$$

From Table 7.1 this gives a two-tailed probability of 0.028, similar to that obtained from Table 12.5.

If the differences are Normally distributed, the t test is the most powerful test. The Wilcoxon test is almost as powerful, however, and in practice the difference is not great. The sign test is similar in power to the Wilcoxon for very small samples, but as the sample size increases, the Wilcoxon test becomes much more powerful. This might be expected since the Wilcoxon test uses more of the information.

12.4. Spearman's rank correlation coefficient, ρ

We noted in Chapter 11 the sensitivity to assumptions of Normality of the product moment correlation coefficient, r. This led to the development of a non-parametric alternative based on ranks. Spearman's approach was direct. First we rank the observations, then we calculate the product moment correlation of the ranks, rather than the observations themselves. The resulting statistic has a distribution which does not depend on the distribution of the original variables. It is usually denote by the Greek letter ρ, pronounced 'rho'.

Table 12.6 shows data from a study of the geographical distribution of a tumour, Kaposi's sarcoma, in mainland Tanzania. The incidence rates were calculated from cancer registry data and there was considerable doubt that all

Table 12.6. Incidence of Kaposi's sarcoma and access of population to health centres for each region of mainland Tanzania (Bland *et al.* 1977)

Region	Incidence cases/million/year	Percent population with 10 km of health centre	Rank order	
			Incidence	Pop %
Coast	1.28	4.0	1	3
Shinyanga	1.66	9.0	2	7
Mbeya	2.06	6.7	3	6
Tabora	2.37	1.8	4	1
Arusha	2.46	13.7	5	13
Dodoma	2.60	11.1	6	10
Kigoma	4.22	9.2	7	8
Mara	4.29	4.4	8	4
Tanga	4.54	23.0	9	16
Singida	6.17	10.8	10	9
Morogoro	6.33	11.7	11	11
Mtwara	6.40	14.8	12	14
Westlake	6.60	12.5	13	12
Kilimanjaro	6.65	57.3	14	17
Ruvuma	7.21	6.6	15	5
Iringa	8.46	2.6	16	2
Mwanza	8.54	20.7	17	15

cases had been notified. The degree of reporting of cases may have been related to population density or availability of health services. In addition, incidence was closely related to age and sex (where recorded) and so could be related to the age and sex distribution in the region. To check that none of these were producing artefacts in the geographical distribution, I calculated the rank correlation of disease incidence with each of the possible explanatory variables. Table 12.6 shows the relationship of incidence to the percentage of the population living within 10 km of a health centre. Figure 12.1 shows the scatter diagram of these data, suggesting that there may be a slight relationship. The percentage within 10 km of a health centre is very highly skewed, whereas the disease incidence appears somewhat bimodal (Fig. 7.20). The assumptions of the product moment correlation do not appear to be met, so rank correlation was preferred.

The calculation of Spearman's ρ proceeds as follows. The ranks for the two variables are found (Table 12.6). Now, we can easily apply the formula for the product moment correlation to these ranks. We define:

$$\rho = \frac{\text{sum of products about mean of ranks}}{\sqrt{\text{sum of squares of ranks for first variable} \times \text{sum of squares of ranks for second variable}}}$$

The calculation proceeds as follows (denoting ranks in one variable by x and the other by y)

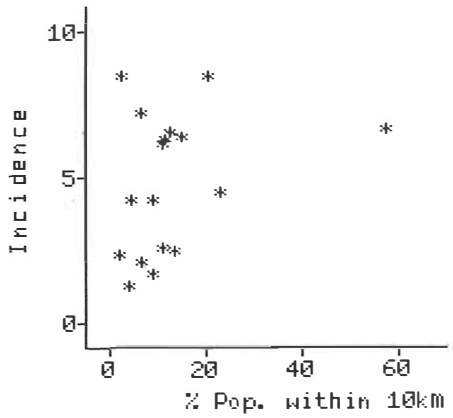


Fig. 12.1 Incidence of Kaposi's sarcoma per million per year and percentage of population within 10 km of a health centre, for 17 regions of mainland Tanzania.

$$\begin{aligned}
\Sigma x_i y_i &= 1 \times 3 + 2 \times 7 + 3 \times 6 + 4 \times 1 + 5 \times 13 + 6 \times 10 + 7 \times 8 + 8 \times 4 \\
&\quad + 9 \times 16 + 10 \times 9 + 11 \times 11 + 12 \times 14 + 13 \times 12 + 14 \times 17 \\
&\quad + 15 \times 5 + 16 \times 2 + 17 \times 15 \\
&= 1531 \\
\Sigma x_i &= 1 + 2 + 3 + \ldots + 15 + 16 + 17 = 153 \\
\Sigma y_i &= 153, \text{ similarly} \\
\Sigma x_i^2 &= 1^2 + 2^2 + 3^2 + \ldots + 15^2 + 16^2 + 17^2 = 1785 \\
\Sigma y_i^2 &= 1785, \text{ similarly}
\end{aligned}$$

Sum of products about mean

$$= \Sigma x_i y_i - \frac{(\Sigma x_i)\,(\Sigma y_i)}{n} = 1531 - \frac{153 \times 153}{17}$$
$$= 154$$

Sum of squares for x

$$= \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = 1785 - \frac{153^2}{17}$$
$$= 408$$

Sum of squares $y = 408$, similarly
Hence

$$\rho = \frac{154}{\sqrt{408 \times 408}} = 0.38$$

We can now use ρ to test the null hypothesis that the variables are independent, the alternative being that either one variable increases as the other

Table 12.7. Two-sided 5 per cent and 1 per cent points of the distribution of Spearman's ρ

Sample size n	Probability that ρ is as far or further from the expected than the tabulated value 5%	1%
4	none	none
5	1.00	none
6	0.89	1.00
7	0.82	0.96
8	0.79	0.93
9	0.70	0.83
10	0.68	0.81

increases, or that one decreases as the other increases. As usual with ranking statistics, the distribution of ρ for small samples can be found by listing all the possible permutations and their values of ρ. For a sample size of n there are, of course, $n!$ possibilities. Table 12.7 shows the critical value of ρ for sample sizes up to 10. As n increases, so ρ tends to a Normal Distribution when the null hypothesis is true, with expected value 0 and variance $1/(n-1)$. Thus $\rho/\sqrt{[1/(n-1)]} = \rho\sqrt{(n-1)}$ is from a Standard Normal Distribution.

For our data we have $0.38\sqrt{(17-1)} = 1.52$, which from Table 7.1 has two-sided probability of 0.13. Hence we have not found any evidence of a relationship between the observed incidence of Kaposi's sarcoma and access to health centres. In this study there was no significant relationship with any of the possible explanatory variables and we concluded that the observed geographical distribution did not appear to be an artefact of population distribution or diagnostic provision.

We have ignored the problem of ties in the above. We treat observations with the same value as described in Section 12.2. We give them the average of the ranks they would have if they were separable and apply the rank correlation formula as described above. In this case the distribution of Table 12.7 is only approximate.

There are several ways of calculating this coefficient, resulting in formulae which appear quite different, though they give the same result (see Siegel 1956).

12.5. Kendall's rank correlation coefficient, τ

Spearman's rank correlation is quite satisfactory for testing the null hypothesis of no relationship, but is difficult to interpret when the null hypothesis is not true. Kendall developed a different rank correlation coefficient, Kendall's τ, which has some advantages over Spearman's. (The Greek

letter τ is pronounced 'tau'.) It is rather more tedious to calculate than Spearman's, but in the computer age this hardly matters. To see how it works we shall consider some artificial data for two variables, X and Y, measured on five subjects, A to E:

Subject	X	Y	*rank* X	*rank* Y
A	3	12	2	3
B	7	17	4	5
C	6	9	3	2
D	2	7	1	1
E	9	16	5	4

We find the rank order of the subjects on the two scales. We wish to examine the strength of the relationship between X and Y. Consider any pair of individuals, say A and B. We ask: are they in the same order in the two rankings? For X, A has rank 2 and B rank 4, so A precedes B. For Y, A has rank 3 and B rank 5, so A precedes B. Hence, A and B are in the same order. Now consider A and C. For X, A has rank 2 and C has rank 3, so A precedes C. For Y, A has rank 3 and C has rank 2, so C precedes A. Thus, A and C are in a different order. We can do this for every possible pair of subjects:

AB, same	AC, diff.	AD, same	AE, same	BC, same
BD, same	BE, diff.	CD, same	CE, same	DE, same

Now, if the ranks are identical, each pair of individuals will be in the same order on both rankings. If the ranks are exactly opposite, each pair will be in a different order on the two rankings. If there is no relationship at all between the rankings, half the pairs will be in the same order and half will be different. We count the number of pairs in the same order, which we shall call P, and the number in different order, which we shall call Q. The total number of pairs is $(P + Q)$. For the example, $P = 8$ and $Q = 2$, there are 10 pairs. Now define S, the difference between the number of pairs in the same and the number of pairs in different order, $S = P - Q$. For the example, $S = 8 - 2 = 6$. Clearly the maximum possible value of S is $(P + Q)$, and this occurs when the two rankings are the same. The minimum possible value is $-(P + Q)$, when the rankings are exactly opposite, and S is zero when there is no relationship at all. Now define

$$\tau = \frac{S}{P + Q}$$

This is Kendall's rank correlation coefficient. For the example

$$\tau = \frac{6}{8 + 2} = \frac{6}{10} = 0.6$$

As the number of subjects increases, the number of pairs increases rapidly

and this method of calculation becomes unwieldy. First we note that the total number of pairs, $P + Q$, is the number of ways of choosing 2 things from n things, i.e.

$$\frac{n!}{2!\,(n-2)!} = \frac{n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1}{2 \times 1 \times (n-2) \times (n-3) \times \ldots \times 3 \times 2 \times 1} = \tfrac{1}{2}n(n-1)$$

Hence

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}$$

Also, $Q = \frac{1}{2}n(n-1) - P$ and so $S = 2P - \frac{1}{2}n(n-1)$.

The simplest way to calculate P is to write down the ranks with one ranking in order:

	D	A	C	B	E
X	1	2	3	4	5
Y	1	3	2	5	4

Now consider the second ranking. The first item, D, has rank 1. All the 4 individuals to the right of this have greater rank, so they will all be in the correct order. Hence 4 pairs containing D are in the same order on both rankings. The second item, A, has 2 individuals to the right of it with greater rank (all but C) but so contributes 2 further pairs in the correct order. Note that the pair AD has already been counted. The third item has 2 ranks greater than it on the right, the fourth has no ranks greater than it, nor of course, has the fifth and last. Hence $P = 4 + 2 + 2 + 0 + 0 = 8$ as before. Since $n = 5$, $P + Q = \frac{1}{2}n(n-1) = \frac{1}{2} \times 5 \times 4 = 10$, $Q = 10 - 8 = 2$ and $S = P - Q = 8 - 2 = 6$. Hence $\tau = 6/10 = 0.6$ as before.

Kendall's τ is intuitively very simple when there are no ties. When there are tied ranks things become more complicated. First we shall calculate the numerator.

Consider the following artificial data:

Subject	X	Y	*X rank*	*Y rank*
A	0	0	$1\frac{1}{2}$	2
B	5	4	$5\frac{1}{2}$	$4\frac{1}{2}$
C	5	6	$5\frac{1}{2}$	6
D	2	4	3	$4\frac{1}{2}$
E	0	0	$1\frac{1}{2}$	2
F	4	0	4	2

Before discussing the practical calculation of τ, we shall look at all the possible pairs:

AD, same	AC, same	AD, same	AE, none	AF, none
BC, same	BD, none	BE, same	BF, same	CD, same
CE, same	CF, same	DE, same	DF, diff.	EF, none

There are 15 pairs. A and B have ranks 1 and 5 on the first ranking and 2 and 4 on the second ranking, and so are in the same order. So are pairs AC and AD. Consider A and E. A and E have ranks 1 and 1 on the first ranking, and ranks 2 and 2 on the second. They are not in any order. We do not know whether A or E really ranks higher. Hence there is no known order and we cannot say whether they are ordered in the same way or differently. We record 'none'. Pair A and F have ranks 1 and 4 on the first ranking and 2 and 2 on the second. We cannot distinguish between them on the second ranking, so we cannot say whether the two orders are the same or different, despite the fact that they are ordered on the first ranking. Hence we now have three kinds of pairs: those where the orders are the same, those where the orders are different, and those where the order is undetermined. For the example, the number the same is $P = 10$, the number different is $Q = 1$ and the number with no order is $N = 4$. Here $P + Q$ does not add up to $\frac{1}{2}n(n-1)$. We have $P + Q + N = \frac{1}{2}n(n-1)$. We define $S = P - Q$ as before, in this case $S = 10 - 1 = 9$.

For practical calculation of S, we first order the individuals by one of the rankings, say the X rank:

	A	E	D	F	B	C
X	$1\frac{1}{2}$	$1\frac{1}{2}$	3	4	$5\frac{1}{2}$	$5\frac{1}{2}$
Y	2	2	$4\frac{1}{2}$	2	$4\frac{1}{2}$	6

It is no longer sufficient to find P, as we cannot calculate S from it. We must calculate Q as well. The first case, A, has the same rank as the second, E, on the first rank, so we do not count the pair AE. We have 3 greater ranks to the right (D, B, C), and 0 smaller. For E, we also have 3 greater and 0 smaller. For D, we have 2 greater, 1 smaller, and 1, B, with no order. For F we have 2 greater and 0 smaller. The final two, B and C have the same rank and so contribute nothing to P or Q. We now have

$$P = 3 + 3 + 2 + 2 + 0 + 0 = 10$$
$$Q = 0 + 0 + 1 + 0 + 0 + 0 = 1$$
Hence $S = P - Q = 10 - 1 = 9$, as before.

Now consider the denominator. There are $\frac{1}{2}n(n-1)$ possible pairs. If there are t individuals tied at a particular tank for variable X, no pairs from these t individuals contribute to S. There are $\frac{1}{2}t(t-1)$ such pairs. If we consider all the groups of tied individuals we have $\Sigma\frac{1}{2}t(t-1)$ pairs which do not contribute to S, summing over all groups of tied ranks. In the example we have two tied at $1\frac{1}{2}$, and two tied at $5\frac{1}{2}$, so

$$\Sigma\tfrac{1}{2}t(t-1) = \tfrac{1}{2} \times 2 \times 1 + \tfrac{1}{2} \times 2 \times 1 = 2$$

Hence the total number of pairs which can contribute to S is $\frac{1}{2}n(n-1) - \Sigma\frac{1}{2}t(t-1)$. For the example this is $(15-2) = 13$. Hence S cannot be greater than $\frac{1}{2}n(n-1) - \Sigma\frac{1}{2}t(t-1)$. The size of S is also limited by ties in the second ranking. If we denote the number of tied individuals in a group by u, then the number of pairs which can contribute to S is $\frac{1}{2}n(n-1) - \Sigma\frac{1}{2}u(u-1)$. For the example, there are three at 2 and two at $4\frac{1}{2}$. Hence

$$\begin{aligned}\tfrac{1}{2}n(n-1) - \Sigma\tfrac{1}{2}u(u-1) &= (\tfrac{1}{2} \times 9 \times 8 - \tfrac{1}{2} \times 3 \times 2 - \tfrac{1}{2} \times 2 \times 1) \\ &= 15 - 4 \\ &= 11\end{aligned}$$

So S must be less than 11.

There are three different ways of handling ties when using τ, denoted by τ_a, τ_b, and τ_c. By far the most useful formula is that for τ_b, which is the proportion of untied pairs having the same ordering. We define τ_b by

$$\tau_b = \frac{S}{\sqrt{[\tfrac{1}{2}n(n-1) - \Sigma\tfrac{1}{2}t(t-1)]\,[\tfrac{1}{2}n(n-1) - \Sigma\tfrac{1}{2}u(u-1)]}}$$

We note that if there are no ties, $\Sigma\frac{1}{2}t(t-1) = 0$ and $\Sigma\frac{1}{2}u(u-1) = 0$, so

$$\tau_b = \frac{S}{\sqrt{[\tfrac{1}{2}n(n-1) \times \tfrac{1}{2}n(n-1)]}} = \tau$$

We can also see that if the numbers of ties are the same, it is possible for τ_b to equal 1 or -1, since the two limits on the size of S are the same. If the numbers of ties are not the same, τ_b must be less than 1. In our example S must be less than or equal to 11, but the denominator is $\sqrt{(13 \times 11)}$, greater than 11. This is reasonable, since if the numbers of ties are different, the rankings cannot be identical. When the rankings are identical $\tau_b = 1$, no matter how many ties there are. For the example:

$$\tau_b = \frac{9}{\sqrt{(13 \times 11)}} = 0.75$$

There are good reasons for this choice of denominator, based on the general theory of correlations (Kendall 1970). Kendall also discusses two other ways of dealing with ties, obtaining coefficients τ_a and τ_c, but their use is restricted.

The value of τ_b for the Kaposi's sarcoma data can be calculated directly from Table 12.6, since the incidence is presented in rank order. There are no ties, so we can use the simple method:

$$\begin{aligned}P &= 14 + 10 + 10 + 13 + 4 + 6 + 7 + 8 + 1 + 5 + 4 + 2 + 2 + 0 + 1 + 1 + 0 \\ &= 88\end{aligned}$$

The number of pairs is $\frac{1}{2}n(n-1) = \frac{1}{2} \times 17 \times 16 = 136$, so $Q = 136 - 88 = 48$ and $S = P - Q = 88 - 48 = 40$. Hence $\tau = S/\frac{1}{2}n(n-1) = 40/136 = 0.29$.

We often want to test the null hypothesis that there is no relationship between the two variables in the population from which our sample was drawn. As usual, we are concerned with the probability of S being the same as or more extreme (i.e. far from zero) than the observed value. Table 12.8 was calculated in the same way as Table 12.1. It shows the probability of being as extreme as the observed value of S for n up to 10. For convenience, S is tabulated rather than τ. When ties are present this is only an approximation.

Table 12.8. Two-sided 5 per cent and 1 per cent points of the distribution of S for Kendall's τ

Sample size n	Probability that S is as far or further from the expected than the tabulated value 5%	1%
4	none	none
5	10	none
6	13	15
7	15	19
8	18	22
9	20	26
10	23	29

When the sample size is greater than 10, S is approximately Normally distributed. The mean of the distribution is zero. If there are no ties, the variance is

$$\text{Var}(S) = \frac{1}{18} n(n-1)(2n+5)$$

When there are ties, the variance formula is horrifically complicated, though if there are not many ties it will not make much difference if the simple form above is used. Seigel (1956) gives it. See Kendall (1970) for a full discussion.

For the example, $S = 40$, $n = 17$ and so the Standard Normal variate is

$$\frac{S}{\sqrt{\text{Var}(S)}} = \frac{S}{\sqrt{\frac{1}{18} n(n-1)(2n+5)}}$$

$$= \frac{40}{\sqrt{\frac{1}{18} \times 17 \times 16 \times 39}}$$

$$= \frac{40}{\sqrt{589.33}}$$

$$= \frac{40}{24.276}$$

$$= 1.65$$

From Table 7.1 of the Normal Distribution we find that the two-sided probability of a value as extreme as this is $0.05 \times 2 = 0.1$, which is very similar to that found using Spearman's ρ. The product moment correlation, r, gives $r = 0.30$, $p = 0.24$. This illustrates the considerable reduction in power for r when its assumptions are not met.

Why have two different rank correlation coefficients? Spearman's ρ is older than Kendall's τ, and can be thought of as a simple analogue of the product moment correlation coefficient, Pearson's r. The coefficient, τ is a part of a more general and consistent system of ranking methods.

In general, the numerical value of ρ is greater than that of τ. It is not possible to calculate ρ from τ or τ from ρ, they measure different sorts of correlation. The coefficient ρ gives more weight to reversals of order when data are far apart in rank than when there is a reversal close together in rank, τ does not. However, in terms of tests of significance both have the same power to reject a false null hypothesis, so for this purpose it does not matter which is used.

12.6. Continuity corrections

In most of the methods in this chapter we have used a continuous distribution, the Normal, to approximate to a discrete distribution, U, T, or S. For example, Fig. 12.2 shows the distribution of the Mann–Whitney U statistic for $n_1 = 4$, $n_2 = 4$ (Table 12.1) with the corresponding Normal curve. From the exact distribution, the probability that $U < 2$ is $0.014 + 0.014 + 0.029 = 0.057$. The corresponding Standard Normal deviate is

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{2 - \frac{4 \times 4}{2}}{\sqrt{\frac{4 \times 4 \times 9}{12}}}$$

$$= -1.732$$

This has a probability of 0.048, interpolating in Table 7.1. This is smaller than the exact probability. This disparity arises because the continuous distribution gives probability to values other than the integers 0, 1, 2, etc. The estimated probability for $U = 2$ can be found by the area under the curve between $U = 1.5$ and $U = 2.5$. The corresponding Normal deviates are

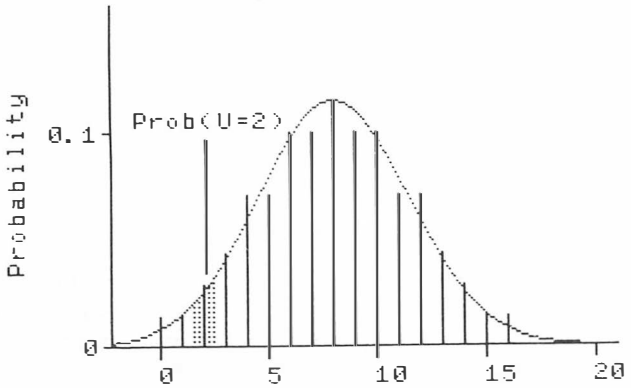


Fig. 12.2 Distribution of the Mann-Witney U statistic, $n_1 = 4$, $n_2 = 4$, when the null hypothesis is true, with the corresponding Normal curve, and area estimating Prob($U = 2$).

-1.876 and -1.588, which have probabilities from Table 7.1 of 0.030 and 0.056. This gives an estimated probability for $U = 2$ of $0.056 - 0.030 = 0.026$, which compares quite well with the exact figure of 0.029. Thus, to estimate the probability that $U < 2$, we estimate the area below $U = 1.5$, not below $U = 2$. This gives us a Standard Normal deviate of -1.588, as already noted, and hence a probability of 0.056. This corresponds remarkably well with the exact probability of 0.057, especially when we consider how small n_1 and n_2 are.

We shall get a better approximation from our Standard Normal deviate if we make U closer to its expected value by $\frac{1}{2}$. In general, we get a better fit if we make the observed value of the statistic closer to its expected value by half of the interval between adjacent discrete values. This is a *continuity correction*.

For S, this interval is 2, not 1, For $S = 2P - \frac{1}{2}n(n + 1)$, and P is an integer. A change of one unit in P produces a change of two units in S. The continuity correction is therefore half of 2, which is 1. We make S closer to the expected value of 0 by 1 before applying the Normal approximation. For the Kaposi's sarcoma data, we had $S = 40$, with $n = 17$. Using the continuity correction gives

$$\frac{|S| - 1}{\sqrt{\text{Var}(S)}} = \frac{40 - 1}{\sqrt{\frac{1}{18} \times 17 \times 16 \times 39}}$$

$$= \frac{39}{24.276}$$

$$= 1.607$$

where $|S|$ means the absolute value or modulus, without sign. This gives a two-sided probability of $0.054 \times 2 = 0.11$, slightly greater than the uncorrected value of 0.10.

Continuity corrections are important for small samples; for large they are negligible. We shall meet another in Chapter 13.

12.7. Parametric or non-parametric methods?

For many statistical problems there are several possible solutions, just as for many diseases there are several treatments, similar perhaps in their overall efficacy but displaying variation in their side-effects, in their interactions with other diseases or treatments, and in their suitability for different types of patients. There is often no one right treatment, but rather treatment is decided by the prescriber's judgement of these effects, past experience, and plain prejudice. Many problems in statistical analysis are like this. In comparing the means of two small groups, for instance, we could use a t test, a t test with a transformation, a Mann–Whitney U test, or one of several others. Our choice of method depends on the plausibility of Normal assumptions, the importance of obtaining a confidence interval, the ease of calculation, and so on. It depends on plain prejudice, too. some users of statistical methods are very concerned about the implications of Normal assumptions and will advocate non-parametric methods wherever possible; others are too careless of the errors that may be introduced when assumptions are not met.

I sometimes meet people who tell me that they have used non-parametric methods throughout their analysis as if this is some kind of badge of statistical purity. It is nothing of the kind. It may mean both that their significance tests have less power than they might have, and that results are left as 'not significant' when, for example, a confidence interval for a difference might be more informative.

On the other hand, such methods are very useful when the necessary assumptions of the t Distribution method cannot be made, and it would be equally wrong to eschew their use. Rather, we should choose the method most suited to the problem, bearing in mind both the assumptions we are making and what we really want to know. We shall say more about what method to use when, in Chapter 14.

Exercise 12M

(Each branch is either true or false.)

1. For comparing the responses to a new treatment of a group of patients with the responses of a control group to a standard treatment, possible approaches include:

(a) the two-sample t method;

(b) the sign test;

(c) the Mann–Whitney U test;

(d) the Wilcoxon matched-pairs test;

(e) rank correlation between responses to the treatments.

2. Kendall's rank correlation coefficient:

(a) depends on the choice of dependent variable;

(b) is zero when there is no relationship;

(c) cannot have a valid significance test when there are tied observations;

(d) must lie between -1 and $+1$;

(e) is not affected by a log transformation of the variables.

3. Tests of significance based on ranks:

(a) are always to be preferred to methods which assume the data to be Normally distributed;

(b) are less powerful than methods based on the Normal Distribution when data are Normally distributed;

(c) enable confidence intervals to be estimated easily;

(d) require no assumptions about the data;

(e) are often to be preferred when data cannot be assumed to follow any particular distribution.

4. Ten men with angina were given an active drug and a placebo on alternate days in random order. Patients were tested using the time in minutes for which they could exercise until angina or fatigue stopped them. The existence of an active drug effect could be examined by:

(a) paired t test;

(b) Mann–Whitney U test;

(c) sign test;

(d) Wilcoxon matched-pairs test;

(e) Spearman's ρ

5. An observed value of zero for the following test statistics would lead to a conclusion of 'not significant':

(a) S for Kendall's τ;

(b) Mann–Whitney U;

(c) T in the Wilcoxon matched-pairs test;

(d) Spearman's ρ;

(e) the number of negatives in the sign test.

Exercise 12E

In this exercise we shall analyse the respiratory compliance and arterial oxygen pressure data of Table 10E.1 using non-parametric methods.

1. For the data of Table 10E.1, use the sign test to test the null hypothesis that changing the waveform has no effect on $p_a(0_2)$.

2. Test the same null hypothesis using a test based on ranks.

3. How do these compare with one another and with the results of 10E part 3?

4. Use the sign test to test the null hypothesis that changing the waveform has no effect on compliance.

5. Test the same null hypothesis using a test based on ranks.

6. Repeat step 4 using log-transformed compliance. Does the transformation make any difference?

7. Repeat step 5 using log compliance. Why do you get a different answer?

8. What do you conclude about the effect of waveform from the non-parametric tests?

9. How do the conclusions of the parametric and non-parametric approaches differ?

13. The analysis of cross-tabulations using the Chi-squared Distribution

13.1. The chi-squared test for association

Table 13.1 shows the results of the clinical trial of streptomycin for treatment of pulmonary tuberculosis (MRC 1948), described in Chapter 2. We have the assessment of the patients' condition tabulated by the treatment. This kind of cross-tabulation of frequencies is also called a *contingency table* or *cross-classification*. This chapter concerns the analysis of such tables, principally using the Chi-squared Distribution.

This is an area where non-parametric methods are mainly used. It can be quite difficult to measure the strength of the association between two qualitative variables, but it is easy to test the null hypothesis that there is no relationship or association between the two variables. If the sample is large, we do this by a chi-squared test.

The chi-squared test for association in a contingency table works like this. The null hypothesis is that there is no association between the two variables, the alternative being that there is an association of some type. We find for each cell of the table the frequency which we would expect if the null hypothesis were true. To do this we use the row and column totals, so we are

Table 13.1. Contingency table showing radiological appearance at six months as compared with appearance on admission in the MRC streptomycin trial

Radiological assessment	Streptomycin	Control	Total
Considerable improvement	28	4	32
Moderate or slight improvement	10	13	23
No material change	2	3	5
Moderate or slight deterioration	5	12	17
Considerable deterioration	6	6	12
Deaths	4	14	18
Total	55	52	107

finding the expected frequencies for tables with these totals, called the *marginal* totals.

There are 107 patients, of whom 32 showed considerable improvement, a proportion 32/107. If there were no relationship between the variables treatment and outcome, we would expect each column of the table to have the same proportion, 32/107, of its members in the first row. Thus the 55 patients in the first column would be expected to have $55 \times 32/107 = 16.4$ in the first row. (By expected we mean the average frequency we would get in the long run. We could not actually observe 16.4 subjects.) The 52 patients in the second column would be expected to have $52 \times 32/107 = 15.6$ in the first row. The sum of these two expected frequencies is, of course, 32, the row total. Similarly, there are 23 patients in the second row and so we would expect $55 \times 23/107 = 11.8$ in the second row, first column and $52 \times 23/107 = 11.2$ in the second row, second column. We calculate the expected frequency for each row and column combination, or *cell*. The 12 cells of Table 13.1 give us the expected frequencies shown in Table 13.2. Notice that the row and column totals are the same as in Table 13.1.

Table 13.2. Expected frequencies under the null hypothesis for Table 13.1

Radiological assessment	Streptomycin	Control	Total
Considerable improvement	16.4	15.6	32
Moderate or slight improvement	11.8	11.2	23
No material change	2.6	2.4	5
Moderate or slight deterioration	8.7	8.3	17
Considerable deterioration	6.2	5.8	12
Deaths	9.3	8.7	18
Total	55.0	52.0	107

In general, the expected frequency for a cell of the contingency table is found by

$$\frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

It does not matter which variable is the row and which the column.

We now compare the observed and expected frequencies. If the two variables are not associated, the observed and expected frequencies should be close together, any discrepancy being due to random variation. We need a test statistic which measures this. The differences between observed and expected frequencies are a good place to start. We cannot simply sum them as the sum would be zero, both observed and expected frequencies having the same grand total, 107. We can resolve this as we resolved a similar problem

with differences from the mean (Section 4.7), by squaring them. The size of the difference will also depend in some way on the number of patients. When the row and column totals are small, the difference between observed and expected is forced to be small. It turns out, for reasons discussed in Appendix 13A.1, that the best statistic is

$$\sum_{\text{all cells}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

This is often written as

$$\sum \frac{(O-E)^2}{E}$$

For Table 13.1 this is

$$\begin{aligned}\sum \frac{(O-E)^2}{E} &= \frac{(28-16.4)^2}{16.4} + \frac{(4-15.6)^2}{15.6} \\ &+ \frac{(10-11.8)^2}{11.8} + \frac{(13-11.2)^2}{11.2} \\ &+ \frac{(2-2.6)^2}{2.6} + \frac{(3-2.4)^2}{2.4} \\ &+ \frac{(5-8.7)^2}{8.7} + \frac{(12-8.3)^2}{8.3} \\ &+ \frac{(6-6.2)^2}{6.2} + \frac{(6-5.8)^2}{5.8} \\ &+ \frac{(4-9.3)^2}{9.3} + \frac{(14-8.7)^2}{8.7} \\ &= 27.2\end{aligned}$$

As is explained in 13A.1, the distribution of this test statistic when the null hypothesis is true and the sample is large enough is the Chi-squared Distribution, with degrees of freedom given by

$$(\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

We shall discuss what is meant by 'large enough' in Section 13.2.

For Table 13.1 we have $(6-1)\times(2-1)=5$ degrees of freedom. Table 13.3 shows some percentage points of the Chi-squared Distribution for selected degrees of freedom. We see that for 5 degrees of freedom the 1 per cent point is 15.1, which our observed value of 27.2 exceeds. The data are not consistent with the null hypothesis and we can conclude that there is good evidence of a relationship between treatment and condition.

It is worth pointing out that the chi-squared statistic is not an index of the

Table 13.3. Percentage points of the Chi-squared Distribution

	Probability that the tabulated value is exceeded (Fig. 13.1)			
Degrees of freedom	10%	5%	1%	0.1%
1	2.71	3.84	6.63	10.83
2	4.61	5.99	9.21	13.82
3	6.25	7.81	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.09	20.52
6	10.64	12.59	16.81	22.46
7	12.02	14.07	18.48	24.32
8	13.36	15.51	20.09	26.13
9	14.68	16.92	21.67	27.88
10	15.99	18.31	23.21	29.59
11	17.28	19.68	24.73	31.26
12	18.55	21.03	26.22	32.91
13	19.81	22.36	27.69	34.53
14	21.06	23.68	29.14	36.12
15	22.31	25.00	30.58	37.70
16	23.54	26.30	32.00	39.25
17	24.77	27.59	33.41	40.79
18	25.99	28.87	34.81	42.31
19	27.20	30.14	36.19	43.82
20	28.41	31.41	37.57	45.32

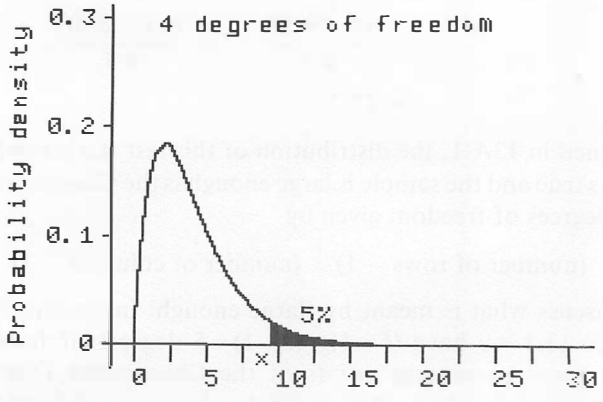


Fig. 13.1 Percentage point of Chi-squared Distribution.

strength of the association. If we double the frequencies in Table 13.1, this will double the chi-squared statistic but the strength of the association is unchanged.

13.2. Validity of the chi-squared test for small samples

We have seen that the test statistic $\Sigma(O - E)^2/E$, which we can call the chi-squared statistic, follows the Chi-squared Distribution provided the expected values are large enough. This is a large sample test, like those of Sections 9.7 and 9.8. The smaller the expected values become, the more dubious will be the test.

The conventional criterion for the test to be valid is usually attributed to the statistician W. G. Cochran. The rule is this: the chi-squared test is valid if at least 80 per cent of the expected frequencies exceed 5 and all the expected frequencies exceed 1. We can see that Table 13.2 satisfies this requirement, since only 2 out of 12 expected frequencies, 17 per cent, are less than 5 and none are less than 1. Note that this condition applies to the expected frequencies, not the observed frequencies. It is quite acceptable for an observed frequency to be 0, provided the expected frequencies meet the criterion.

This criterion is open to question. Simulation studies appear to suggest that the condition may be too conservative and that the chi-squared approximation works for smaller expected values, especially for larger numbers of rows and columns. At the time of writing the analysis of tables based on small sample sizes, particularly 2 by 2 tables, is the subject of hot dispute among statisticians. As yet, no-one has succeeded in devising a better rule than Cochran's, so I would recommend keeping to it until the theoretical questions are resolved. Any chi-squared test which does not satisfy the criterion is always open to the charge that its validity is in doubt.

If the criterion is not satisfied we can usually combine or delete rows and columns to give bigger expected values. Of course, this cannot be done for 2 by 2 tables, which we consider in more detail below. This editing must be done with regard to the meaning of the various categories. In Table 13.1, there would be no point in combining rows 1 and 6 to give a new category of 'considerable improvement or death' to be compared to the remainder, as the comparison would be absurd.

For an example, consider Table 13.4. These data, from the MRC streptomycin trial, show the results of radiological assessment for a subgroup of patients, defined by a prognostic variable. We want to know whether there is evidence of a streptomycin effect within this subgroup, so we want to test the null hypothesis of no effect using a chi-squared test. There are 6 out of 8 expected values less than 5, so the test on this table would not be valid. We must combine the rows so as to raise the expected values. Since there are no observations in the 'no change' row, it does not matter what we do with it, so

Table 13.4. Observed and expected frequencies of categories of radiological appearance at six months as compared with appearance on admission in the MRC streptomycin trial, patients with an initial temperature of 100–100.9 °F

	Streptomycin		Control		
Radiological assessment	Observed	Expected	Observed	Expected	Total
Improvement	13	8.4	5	9.6	18
No change	0	0.0	0	0.0	0
Deterioration	2	4.2	7	4.8	9
Death	0	2.3	5	2.7	5
Total	15	15	17	17	32

we could combine it with the 'improvement' row to give 'no change or improvement'. We can also combine the 'deterioration' and 'death' rows to give a 'deterioration or death' row. The expected values are then all greater than 5 and we can do the chi-squared test with 1 degree of freedom. The new table is shown in Table 13.5. We have

Table 13.5. Reduction of Table 13.4 to a 2 by 2 table

	Streptomycin		Control		
Radiological assessment	Observed	Expected	Observed	Expected	Total
Improvement or no change	13	8.4	5	9.6	18
Deterioration or death	2	6.6	12	7.4	14
Total	15	15	17	17	32

$$\sum \frac{(O-E)^2}{E} = \frac{(13-8.4)^2}{8.4} + \frac{(5-9.6)^2}{9.6} + \frac{(2-6.6)^2}{6.6} + \frac{(12-7.4)^2}{7.4}$$

$$= 10.8$$

Under the null hypothesis this is from a Chi-squared Distribution with one degree of freedom, and from Table 13.3 we can see that the probability of getting a value as extreme as 10.8 is less than 1 per cent. We have data inconsistent with the null hypothesis and we can conclude that the evidence suggests a treatment effect in this subgroup.

If the table does not meet the criterion even after reduction to a 2 by 2 table, we can apply either a continuity correction to improve the approximation to the Chi-squared Distribution, or an exact test based on a discrete distribution

like those of Sections 12.2 to 12.5. These methods are described in Sections 13.5 and 13.6.

13.3. Tests for 2 by 2 tables

Consider the data on cough symptom and history of bronchitis discussed in Section 9.8. We had 273 children with a history of bronchitis of whom 26 were reported to have day or night cough, and 1046 children without history

Table 13.6. Cough during the day or at night at age 14 for children with and without a history of bronchitis before age 5 (Holland *et al.* 1978)

	Bronchitis	No bronchitis	Total
Cough	26	44	70
No cough	247	1002	1249
Total	273	1046	1319

of bronchitis, of whom 44 were reported to have day or night cough. We can set these data out as a cross-classification table, as in Table 13.6.

Table 13.7. Expected frequencies for Table 13.6

	Bronchitis	No bronchitis	Total
Cough	14.49	55.51	70
No cough	258.51	990.49	1249
Total	273	1046	1319

Let us use the chi-squared test to test the null hypothesis of no association between cough and history. The expected values are shown in Table 13.7. The test statistic is

$$\sum \frac{(O-E)^2}{E} = \frac{(26-14.49)^2}{14.49} + \frac{(44-55.51)^2}{55.51} + \frac{(247-258.51)^2}{258.51} + \frac{(1002-990.49)^2}{990.49} = 12.2$$

We have $r = 2$ rows and $c = 2$ columns, so there are $(r-1)(c-1) = (2-1) \times (2-1) = 1$ degree of freedom. We see from Table 13.3 that the 5 per cent point is 3.84, and the 1 per cent point is 6.63, so we have observed something very unlikely if the null hypothesis were true. Hence we reject the

null hypothesis of no association and conclude that there is evidence for a relationship between present cough and history of bronchitis.

Now the null hypothesis 'no association between cough and bronchitis' is the same as the null hypothesis 'no difference between the proportions with cough in the bronchitis and no-bronchitis groups'. If there were a difference, the variables would be associated. Thus we have tested the same null hypothesis in two different ways. In fact these tests are exactly equivalent. If we take the Normal deviate from Section 9.8, which was 3.49, and square it, we get 12.2, the chi-squared value. The method of Sections 9.8 and 8.6 has the advantage that it can also give us a confidence interval for the size of the difference, which the chi-squared method does not.

13.4. The chi-squared test for trend

Consider the data of Table 13.8. Using the chi-squared test described in Section 13.1, we can test the null hypothesis that there is no relationship between reported cough and smoking against the alternative that there is a relationship of some sort. The chi-squared statistic is 64.1, with 2 degrees of freedom. This is a very unlikely value and the data are not consistent with the null hypothesis.

Now, we would have got the same value of chi-squared whatever the order of the columns. We might expect, however, that if there were a relationship between reported cough and smoking, the prevalence of cough would be greater for greater amounts of smoking. In other words, we look for a trend in cough prevalence from one end of the table to the other. We can test for this using the *chi-squared test for trend*.

First, we define two variables, X and Y, whose values depend on the categories of the row and column variables. For example, we could put $X = 1$ for non-smokers, $X = 2$ for occasional smokers and $X = 3$ for regular smokers, and put $Y = 1$ for yes and $Y = 2$ for no. Then for a non-smoker who coughs, the value of X is 1 and the value of Y is 1. If there are N individuals, we have N pairs of observations (x_i, y_i). If there is a linear trend across the table, there will be linear regression of Y on X which has non-zero slope.

Table 13.8. Cough during the day or at night and cigarette smoking by 12-year-old boys (Bland *et al.* 1978)

	Boy's smoking			
	Non-smoker	Smokes occasionally	Smokes regularly	Total
Cough	266	395	80	741
No cough	1037	977	92	2106
Total	1303	1372	172	2847

We fit the usual least-squares regression line, $Y = a + bX$, where:

$$b = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

$$se(b) = \sqrt{\frac{s^2}{\Sigma(x_i - \bar{x})^2}}$$

where s^2 is the estimated variance of Y. In simple linear regression, as described in Chapter 11, we are usually concerned with estimating b and making statements about its precision. Here we are only interested in testing the null hypothesis that $b = 0$. Under the null hypothesis, the variance about the line is equal to the total variance of Y, since the line has zero slope. We use the estimate

$$s^2 = \frac{1}{N}\Sigma(y_i - \bar{y})^2$$

(We use N as the denominator here, not $(N - 1)$, because the test is conditional on the row and column totals as described in Appendix 13A. There is a good reason for it, but it is not worth going into here.) Hence the standard error of b is

$$se(b) = \sqrt{\frac{s^2}{\Sigma(x_i - \bar{x})^2}} = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{N\Sigma(x_i - \bar{x})^2}}$$

Now, b is the sum of many independent, identically distributed random variables $(y_i - \bar{y})(x_i - \bar{x})$, and so follows a Normal Distribution by the central limit theorem. As N is large, $se(b)$ should be a good estimate of the standard deviation of this distribution. Hence, if the null hypothesis is true and $E(b) = 0$, $b/se(b)$ is an observation from a Standard Normal Distribution. Hence the square of this, $b^2/se(b)^2$, is from a Chi-squared Distribution with one degree of freedom.

$$\frac{b^2}{se(b)^2} = \frac{\left[\dfrac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}\right]^2}{\dfrac{\Sigma(y_i - \bar{y})^2}{N\Sigma(x_i - \bar{x})^2}}$$

$$= \frac{[\Sigma(y_i - \bar{y})(x_i - \bar{x})]^2}{\Sigma(x_i - \bar{x})^2\,\Sigma(y_i - \bar{y})^2/N}$$

$$= \frac{N[\Sigma(y_i - \bar{y})(x_i - \bar{x})]^2}{\Sigma(x_i - \bar{x})^2\,\Sigma(y_i - \bar{y})^2}$$

For practical calculations we use the alternative forms of the sums of squares and products:

$$\frac{b^2}{se(b)^2} = \frac{N\left[\Sigma y_i x_i - \dfrac{(\Sigma y_i)(\Sigma x_i)}{N}\right]^2}{\left[\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{N}\right]\left[\Sigma y_i^2 - \dfrac{(\Sigma y_i)^2}{N}\right]}$$

Note that it does not matter which variable is X and which is Y. The sums of squares and products are easy to work out. For example, for the column variable, X, we have 1303 individuals with $X = 1$, 1372 with $X = 2$ and 172 with $X = 3$. For our data we have

$$\begin{aligned}
\Sigma x_i^2 &= 1^2 \times 1303 + 2^2 \times 1372 + 3^2 \times 172 = 8339 \\
\Sigma x_i &= 1 \times 1303 + 2 \times 1372 + 3 \times 172 = 4563 \\
\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N} &= 8339 - \frac{4563^2}{2847} = 1025.7 \\
\Sigma y_i^2 &= 1^2 \times 741 + 2^2 \times 2106 = 9165 \\
\Sigma y_i &= 1 \times 741 + 2 \times 2106 = 4953 \\
\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{N} &= 9165 - \frac{4953^2}{2847} = 548.1 \\
\Sigma x_i y_i &= 1 \times 1 \times 266 + 2 \times 1 \times 395 + 3 \times 1 \times 80 \\
&\quad + 1 \times 2 \times 1037 + 2 \times 2 \times 977 + 3 \times 2 \times 92 \\
&= 7830 \\
\Sigma y_i x_i - \frac{(\Sigma y_i)(\Sigma x_i)^2}{N} &= 7830 - \frac{4563 \times 4953}{2847} \\
&= -108.4
\end{aligned}$$

Hence

$$\chi^2 = \frac{2847 \times (-108.4)^2}{1025.7 \times 548.1} = 59.5$$

If the null hypothesis is true, χ^2 is an observation from the Chi-squared Distribution with 1 degree of freedom. The value 59.5 is highly unlikely from this distribution and the trend is significant.

We can also assess the departure from a linear regression. We subtract the chi-squared for trend statistic from the usual contingency table chi-squared. This can be used to test the null hypothesis that there is no deviation from the linear trend. It follows a Chi-squared Distribution with $(r-1)(c-1) - 1$ degrees of freedom. For the example, the chi-squared for deviation from the trend is $64.1 - 59.5 = 4.6$, with 1 degree of freedom. This indicates that a linear trend does not completely explain the relationship.

There are several points to note about this method. The choice of values for X and Y is arbitrary. By putting $X = 1, 2$, or 3 we assumed that the difference between non-smokers and occasional smokers is the same as that between occasional smokers and smokers. This need not be so and a different choice

Table 13.9. Smoking by 12-year-old boys and smoking by their parents (Bland *et al.* 1978)

	Parents' smoking			
Boy's smoking	Neither smoke	One smokes	Both smokes	Total
Never smoked	480	432	391	1303
Smoked only once	256	393	327	976
Smoked occasionally	90	147	159	396
Smokes > 1/week	13	34	45	92
Smokes > 1/day	9	25	46	80
Total	843	1031	968	2847

of X would give a different chi-squared for trend statistic. The choice is not critical, however.

The trend may be significant even if the overall contingency table chi-squared is not. This is because the test for trend has greater power for detecting trends than has the ordinary chi-squared test. On the other hand, if we had an association where those who were occasional smokers had far more symtoms than either non-smokers or regular smokers, the trend test would not detect it.

There can be more than two categories for X and for Y. For example, consider Table 13.9. The overall chi-squared is 59.8 with 8 degrees of freedom and the trend chi-squared is 50.6 with 1 degree of freedom, based on dummy variables with equal intervals. We put $X = 1, 2,$ or 3 for the parents' smoking and $Y = 1, 2, 3, 4,$ or 5 for the child's smoking. The chi-squared about trend is $59.8 - 50.6 = 9.2$ with $8 - 1 = 7$ degrees of freedom. This is not significant, showing that the trend for the child to report smoking more as the number of parents smoking increases can explain all the association in the table.

We can also apply the trend test to Table 13.1. The chi-squared for trend is 17.9 with 1 d.f., $p < 0.001$, and the chi-squared about trend is 9.0 with 4 d.f., $p > 0.05$. The main association is an increasing tendency to be in the control group as we go from considerable improvement to death.

13.5. Fisher's exact test

The chi-squared test described in Section 13.1 is a large-sample test. When the sample is not large and expected values are less than 5, we can turn to an exact distribution like those for the Mann–Whitney U statistic, rank correlation coefficient, etc. This method is called Fisher's exact test.

The exact probability distribution for the table can only be found when the

row and column totals are given. Just as with the large-sample chi-squared test, we restrict our attention to tables with these totals. This difficulty has led to much controversy about the use of this test. We shall see how the test works first, then discuss its applicability.

Fisher's exact test works like this. Suppose we carry out a clinical trial and randomly allocate 4 patients to treatment A and 4 to treatment B. The outcome is as follows:

	Survived	*Died*	*Total*
Treatment A	3	1	4
Treatment B	2	2	4
Total	5	3	8

We want to know whether the difference in mortality between the two treatments is due to chance or is evidence of a difference between the treatments. To do this, we first ask how many randomizations would give this outcome? We can see that if we keep the row and column totals constant, there are only 4 possible tables:

(i)

	S	D	T
A	4	0	4
B	1	3	4
T	5	3	8

(ii)

	S	D	T
A	3	1	4
B	2	2	4
T	5	3	8

(iii)

	S	D	T
A	2	2	4
B	3	1	4
T	5	3	8

(iv)

	S	D	T
A	1	3	4
B	4	0	4
T	5	3	8

These tables are found by putting the values 0, 1, 2, 3 in the 'A and D' cell. Any other values would make the D total greater than 3.

Now, let us label our subjects a to h. The survivors we shall call a to e, and the deaths f, g, h. How many ways can these patients be arranged in two groups of 4 to give tables i, ii, iii and iv?

Table i can arise in 5 ways. Patients f, g, and h would have to be in group B, to give 3 deaths, and the remaining member of B could be a, b, c, d or e.

Table ii can arise in 30 ways. The 3 survivors in group A can be abc, abd, abe, acd, ace, ade, bcd, bce, bde, cde, 10 ways. The death in A can be f, g or h, 3 ways. Hence the group can be made up in $10 \times 3 = 30$ ways. Table iii is the same as table ii, with A and B reversed, so arises in 30 ways. Table iv is the same as table i with A and B reversed, so arises in 5 ways.

Hence we can arrange the 8 patients into 2 groups of 4 in $5 + 30 + 30 + 5 = 70$ ways. Now, the probability of any one arrangement arising by chance is

1/70, since they are all equally likely. If there are 3 deaths, table i arises from 5 of the 70 arrangements, so had probability 5/70 = 0.071. Table ii arises from 30 out of 70 arrangements, so has probability 30/70 = 0.429. Similarly, table iii has probability 30/70 = 0.429, and table iv has probability 5/70 = 0.071.

Hence, under the null hypothesis that there is no association between treatment and survival, table ii, which we observed, has a probability of 0.429. It could easily have arisen by chance and is consistent with the null hypothesis. We should also consider tables more extreme than the observed. In this case, the only more extreme table is table i, where all the deaths occur in one group, so the probability of the observed table or a more extreme one is 0.071 + 0.429 = 0.5.

This is the rationale for Fisher's exact test. We calculate the probability of each possible table arising, under the null hypothesis. We then find the probability of the observed or more extreme tables arising by chance and if this total probability is small (say less than 0.05) the data are inconsistent with the null hypothesis and we can conclude that there is evidence that an association exists.

There is no need to enumerate all the possible tables, as above. The probability can be found from a simple formula. If the row and column totals are R_1, R_2, C_1, and C_2, the probability of observing frequencies O_{11}, O_{12}, O_{22}, O_{21} is

$$\frac{R_1! \times R_2! \times C_1! \times C_2!}{N! \times O_{11}! \times O_{12}! \times O_{22}! \times O_{21}!}$$

(For the derivation of this see Appendix 13A.2.) We can calculate this easily for each possible table and so find the probability for the observed table and each more extreme one. For the example, we have:

$$\text{table i: } \frac{5! \times 3! \times 4! \times 4!}{8! \times 4! \times 0! \times 1! \times 3!} = \frac{5! \times 4!}{8!} = \frac{4 \times 3 \times 2 \times 1}{8 \times 7 \times 6} = 0.071$$

$$\text{table ii: } \frac{5! \times 3! \times 4! \times 4!}{8! \times 3! \times 1! \times 2! \times 2!} = 0.429$$

giving a total of 0.50 as before.

Unlike the exact distributions for the rank statistics, this is fairly easy to calculate but difficult to tabulate. A good table of this distribution required a whole book (Finney *et al.* 1963).

Fisher's exact test is essentially one sided. We have only considered more extreme values in one direction. It is not clear what the corresponding deviations in the other direction would be, especially when all the marginal totals

are different. This is because the distribution is asymmetrical, unlike those of Sections 12.2 to 12.5. One solution is to double the one-sided probability to get a two-sided test when this is required, but this is more to have uniformity with other tests, such as chi-squared, than because two-sided probabilities are more meaningful than one-sided ones.

We can apply this test to Table 13.5. The 2 by 2 tables to be tested and their probabilities are:

Table	*Probability*
13 5 2 12	0.001 378 2
14 4 1 13	0.000 075 7
15 3 0 14	0.000 001 4

The total one-sided probability is 0.001 455 3, which for an approximate two-sided test gives 0.0029. This is considerably bigger than the probability for the χ^2 value of 10.6, which is 0.0011.

13.6. Yates' continuity correction for the 2 by 2 table

The discrepancy in probabilities between the chi-squared test and Fisher's exact test arises because we are estimating the discrete distribution of the test statistic by the continuous Chi-squared Distribution. A continuity correction like those of Section 12.6, called *Yates' correction*, can be used to improve the fit. The observed frequencies change in units of one, so we make them closer to their expected values by one half. Hence the formula for the corrected chi-squared statistic for a 2 by 2 table is

$$\sum \frac{(|O - E| - \frac{1}{2})^2}{E}$$

where $|O - E|$ means the absolute value or modulus, without sign. For Table 13.5 we have:

$$\begin{aligned}\sum \frac{(|O - E| - \tfrac{1}{2})^2}{E} &= \frac{(|13 - 8.4| - \tfrac{1}{2})^2}{8.4} + \frac{(|5 - 9.6| - \tfrac{1}{2})^2}{9.6} \\ &\quad + \frac{(|2 - 6.6| - \tfrac{1}{2})^2}{6.6} + \frac{(|12 - 7.4| - \tfrac{1}{2})^2}{7.4} \\ &= \frac{(4.6 - \tfrac{1}{2})^2}{8.4} + \frac{(4.6 - \tfrac{1}{2})^2}{9.6} \\ &\quad + \frac{(4.6 - \tfrac{1}{2})^2}{6.6} + \frac{(4.6 - \tfrac{1}{2})^2}{7.4} \\ &= 8.6\end{aligned}$$

This has probability 0.0037, which is closer to the exact probability, though there is still a considerable discrepancy. Of course, at such extremely low probabilities any approximation is liable to break down. In the critical area between 0.10 and 0.01, the continuity correction usually gives a very good fit to the exact probability.

13.7. The validity of Fisher's exact test and Yates' correction

There has been much dispute among statisticians about the validity of the exact test and the continuity correction which approximates to it. Among the more argumentative of the founding fathers of statistical inference, such as Fisher and Neyman, this was quite acrimonious. Unfortunately, the problem is still unresolved and still generating almost as much heat as light.

Note that Tables 13.5 and 13.6 arose in different ways. In Table 13.5, the column totals were fixed by the design of the experiment and only the row totals are from a random variable. In Table 13.6 neither row nor column totals were set in advance. Both are from the Binomial Distribution, depending on the incidence of bronchitis and prevalence of chronic cough in the population. There is a third possibility, that both the row and column totals are fixed. This is rare in practice, but it can be achieved by the following experimental design. We want to know whether a subject can distinguish an active treatment from a placebo. We present him with 10 tablets, 5 of each, and ask him to sort the tablets into the 5 active and 5 placebo. This would give a 2 by 2 table, subject's choice versus truth, in which all row and column totals are preset to 5. There are several variations on these types of table, too.

It can be shown that the same chi-squared test applies to all these cases when samples are large. When samples are small, this is not necessarily so. A discussion of the problem is well beyond the scope of this book, but it suffices to say that this is at root of all the conflicting statements which you may come across about the validity of various tests of significance in the 2 by 2 table.

When the row and column totals are fixed, Fisher's exact test and Yates' correction are undoubtedly correct. For other cases they may be conservative, that is, give rather larger probabilities than they should, or they may not. My own opinion is that Yates' correction and Fisher's exact test should be used. If we must err, it seems better to err on the side of caution.

13.8. McNemar's test for matched samples

The chi-squared test described above enables us, among other things, to test the null hypothesis that binomial proportions estimated from two independent samples are the same. We shall close this chapter with the one sample or matched sample problem.

For example, Holland *et al.* (1978) obtained respiratory symptom questionnaires for 1319 Kent schoolchildren at ages 12 and 14. One question we asked was whether the prevalence of reported symptoms was different at the two ages. At age 12, 356 (27 per cent) children were reported to have had severe colds in the past 12 months compared to 468 (35 per cent) at age 14. Was there evidence of a real increase?

Just as in the one-sample or paired t test (Section 10.2) we would hope to improve our analysis by taking into account the fact that this is the same sample. We might expect, for instance, that symptoms on the two occasions will be related. The method which enables us to do this is McNemar's test, another version of the sign test. We need to know that 212 children were reported to have colds on both occasions, 144 to have colds at 12 but not at 14, 256 to have colds at 14 but not at 12 and 707 to have colds at neither age. Table 13.8 shows the data in tabular form.

Table 13.10. Severe colds reported at two ages for Kent schoolchildren (Holland *et al.* 1978)

		Severe colds at age 14		
		Yes	No	Total
Severe colds at age 12	Yes	212	144	356
	No	256	707	963
Total		468	851	1319

The null hypothesis is that the proportions saying yes on the first and second occasions are the same, the alternative being that one exceeds the other. This is a hypothesis about the row and column totals, quite different from that in Section 13.1. If the null hypothesis were true we would expect the frequencies for 'yes, no' and 'no, yes' to be equal. In other words as many should go up as down. (Compare this with the sign test, Section 9.2.) If we denote these frequencies by O_{yn} and O_{ny}, then the expected frequencies will be $(O_{yn} + O_{ny})/2$. In the same way as Section 13.2 we get the test statistic:

$$\sum \frac{(O-E)^2}{E} = \frac{\left(O_{yn} - \dfrac{O_{yn}+O_{ny}}{2}\right)^2}{\dfrac{O_{yn}+O_{ny}}{2}} + \frac{\left(O_{yn} - \dfrac{O_{yn}+O_{ny}}{2}\right)^2}{\dfrac{O_{yn}+O_{ny}}{2}}$$

which follows a Chi-squared Distribution provided the expected values are large enough. There are two observed frequencies and one constraint

(Appendix 13A.1), that the sum of the observed frequencies = the sum of the expected frequencies. Hence there is one degree of freedom.

The test statistic can be simplified considerably. Each term in brackets simplifies like this:

$$O_{yn} - \frac{O_{yn} + O_{ny}}{2} = \frac{2O_{yn} - O_{yn} - O_{ny}}{2}$$

$$= \frac{O_{yn} - O_{ny}}{2}$$

Hence the test statistic is

$$\sum \frac{(O - E)^2}{E} = \frac{\left(\frac{O_{yn} - O_{ny}}{2}\right)^2}{\frac{O_{yn} + O_{ny}}{2}} + \frac{\left(\frac{O_{yn} - O_{ny}}{2}\right)^2}{\frac{O_{yn} + O_{ny}}{2}}$$

$$= \frac{2 \times \frac{1}{4} \times (O_{yn} - O_{ny})^2}{\frac{1}{2}(O_{yn} + O_{ny})}$$

$$= \frac{(O_{yn} - O_{ny})^2}{O_{yn} + O_{ny}}$$

For Table 13.8, we have

$$\frac{(O_{yn} - O_{ny})^2}{O_{yn} + O_{ny}} = \frac{(144 - 256)^2}{144 + 256}$$

$$= \frac{112^2}{400}$$

$$= 31.4$$

This can be referred to Table 13.3 with one degree of freedom and is clearly highly significant. There was a difference between the two ages. As there was no change in any of the other symptoms studied, we thought that this was possibly due to an epidemic of upper respiratory tract infection just before the second questionnaire.

There is a continuity correction, again due to Yates. If the observed frequency, O_{yn}, increases by 1, O_{ny} decreases by 1 and $(O_{yn} - O_{ny})$ increases by 2. Thus, half the difference between adjacent possible values is 1 and we make the observed difference nearer to the expected difference (zero) by 1. Thus the continuity corrected test statistic is

$$\frac{(|O_{yn} - O_{ny}| - 1)^2}{O_{yn} + O_{ny}}$$

For Table 13.8 this is

$$\frac{(|O_{yn} - O_{ny}| - 1)^2}{O_{yn} + O_{ny}} = \frac{(|144 - 256| - 1)^2}{144 + 256}$$

$$= \frac{(112 - 1)^2}{400} = 30.8$$

There is very little difference because the expected values are so large, but if the expected values are small, say less than 20, the correction is advisable. For very small samples, we can taken O_{ny} as an observation from the binomial distribution with $p = 1/2$ and $n = O_{yn} + O_{ny}$ and proceed as for the sign test in Section 9.2.

Appendix 13A

A13.1. Why the chi-squared test works

We noted some of the properties of the Chi-squared Distribution in Appendix 7A. In particular, it is the sum of the squares of a set of independent Standard Normal variables, and if we look at a subset of values defined by independent linear relationships between these variables we lose one degree of freedom for each constraint. It is on these two properties that the chi-squared test depends.

Suppose we did not have a fixed size to our streptomycin experiment, but allocated and observed patients as they arrived randomly. Then, in any given time interval the number in a given cell of the table would be from a Poisson Distribution and the set of Poisson variables corresponding to the cell frequency would be independent of one another. Our table is one set of samples from these Poisson Distributions. However, we do not know the expected values of these distributions under the null hypothesis; we only know their expected values if the table has the row and column totals we observed. We can only consider the subset of outcomes of these variables which has the observed row and column totals. The test is said to be *conditional* on these row and column totals.

The mean and variance of a Poisson variable are equal (Section 6.7). If the null hypothesis is true, the means of these variables will be equal to the expected frequency calculated in Section 13.1. Thus O, the observed cell frequency, is from a Poisson Distribution with mean E, the expected cell frequency, and standard deviation $\sqrt{E}$. Provided E is large enough, this Poisson Distribution will be approximately Normal. Hence $(O - E)/\sqrt{E}$ is from a Normal Distribution mean 0 and variance 1. Hence if we find

$$\sum \left(\frac{O - E}{\sqrt{E}}\right)^2 = \sum \frac{(O - E)^2}{E}$$

this is the sum of a set of Normally distributed random variables, mean 0 and variance 1, and so is from a Chi-squared Distribution.

We shall now find the degrees of freedom. Although the underlying variables are independent, we are only considering a subset defined by the row and column totals. Consider the following table:

			Total
	O_{11} O_{21}	O_{12} O_{22}	R_1 R_2
Total	C_1	C_2	N

The values O_{11} to O_{22} are the observed frequencies, R_1, R_2 the row totals, etc. Denote the corresponding expected values by E_{11} to E_{22}. There are three linear constraints on the frequencies:

$$O_{11} + O_{12} + O_{21} + O_{22} = N$$
$$O_{11} + O_{12} = R_1$$
$$O_{11} + O_{21} = C_1$$

Any other constraint can be made up of these. For example, we must have

$$O_{21} + O_{22} = R_2$$

This can be found by subtracting the second equation from the first. On the left side

$$O_{11} + O_{12} + O_{21} + O_{22} - (O_{11} + O_{12}) = O_{21} + O_{22}$$

On the right side

$$N - R_1 = R_2$$

Each of these linear constraints on O_{11} to O_{22} is also a linear constraint on $(O_{11} - E_{11})/\sqrt{E_{11}}$ to $(O_{22} - E_{22})/\sqrt{E_{22}}$. We can see this by replacing O_{11} by

$$O_{11} = \sqrt{E_{11}} \times \frac{(O_{11} - E_{11})}{\sqrt{E_{11}}} + E_{11}$$

etc. in the equations. This gives the required linear constraints.

There are four observed frequencies and so four $(O - E)/\sqrt{E}$ variables, with three constraints. We lose one degree of freedom for each constraint and so have $4 - 3 = 1$ degree of freedom.

If we have r rows and c columns, then we have one constraint that the sum of the frequencies is N. Each row must add up, but when we reach the last row the constraint can be obtained by subtracting the first $(r - 1)$ rows from the

grand total. The rows contribute only $(r-1)$ further constraints. Similarly the columns contribute $(c-1)$ constraints. Hence, there being rc frequencies, the degrees of freedom are

$$\begin{aligned} rc - 1 - (r-1) - (c-1) &= rc - 1 - r + 1 - c + 1 \\ &= rc - r - c + 1 \\ &= (r-1)(c-1) \end{aligned}$$

So we have degrees of freedom given by the number of rows minus one times the number of columns minus one.

13A.2. Derivation of the formula for Fisher's exact test

The derivation of Fisher's formula is not difficult for the algebraically minded. Remember that the number of ways of choosing r things out of n things (Appendix 6A) is $n!/r!(n-r)!$ Now, suppose we have a 2 by 2 table made up of N individuals:

			Total
	O_{11} O_{21}	O_{12} O_{22}	R_1 R_2
Total	C_1	C_2	N

First, we ask how many ways N individuals can be arranged to give marginal totals, R_1, R_2, C_1 and C_2. They can be arranged in columns in $N!/C_1!C_2!$, since we are choosing C_1 objects out of N, and in rows $N!/R_1!R_2!$ ways. (Remember $N - C_1 = C_2$ and $N - R_1 = R_2$.) Hence they can be arranged in

$$\frac{N!}{C_1!C_2!} \times \frac{N!}{R_1!R_2!} = \frac{N!N!}{C_1!C_2!R_1!R_2!}$$

ways. For example, the table with totals

		4
		4
5	3	8

can happen in

$$\frac{8!}{5! \times 3!} \times \frac{8!}{4! \times 4!} = 56 \times 70 = 3620 \text{ ways}$$

As we saw in Section 13.5, the columns can be arranged in 70 ways. Now we ask, of these ways how many make up the particular table? We are now dividing the N into four groups of sizes O_{11}, O_{12}, O_{21} and O_{12}. We can

choose the first group in $N!/O_{11}!\ (N - O_{11})!$ ways, as before. We are now left with $N - O_{11}$ individuals, so we can choose O_{12} in $(N - O_{11})!/O_{12}!\ (N - O_{11} - O_{12})!$ ways. We are left with $N - O_{11} - O_{12}$, and so we choose O_{21} in $(N - O_{11} - O_{12})!/O_{21}!\ (N - O_{11} - O_{12} - O_{21})!$ ways. This leaves $N - O_{11} - O_{12} - O_{21}$, which is, of course, equal to O_{22} and so O_{22} can only be chosen in one way. Hence we have altogether:

$$\frac{N!}{O_{11}! \times (N - O_{11})!} \times \frac{(N - O_{11})!}{O_{12}! \times (N - O_{11} - O_{12})!} \times \frac{(N - O_{11} - O_{12})!}{O_{21}! \times (N - O_{11} - O_{12} - O_{12})!}$$

$$= \frac{N!}{O_{11}! \times O_{12}! \times O_{21}! \times (N - O_{11} - O_{12} - O_{12})!}$$

$$= \frac{N!}{O_{11}! \times O_{12}! \times O_{21}! \times O_{22}!}$$

because $N - O_{11} - O_{12} - O_{12} = O_{22}$. So out of the

$$\frac{N! \times N!}{C_1! \times C_2! \times R_1! \times R_2!}$$

possible tables, the given table arises in

$$\frac{N!}{O_{11}! \times O_{12}! \times O_{21}! \times O_{22}!}$$

ways. The probability of this table arising by chance is

$$\frac{\dfrac{N!}{O_{11}! \times O_{12}! \times O_{21}! \times O_{22}!}}{\dfrac{N! \times N!}{C_1! \times C_2! \times R_1! \times R_2!}} = \frac{C_1! \times C_2! \times R_1! \times R_2!}{N! \times O_{11}! \times O_{12}! \times O_{21}! \times O_{22}!}$$

Exercise 13M

(Each branch is either true or false.)

1. In a chi-squared test for a 5 by 3 contingency table:

(a) variables must be quantitative;

(b) observed frequencies are compared to expected frequencies;

(c) there are 15 degrees of freedom;

(d) at least 12 cells must have expected values greater than 5;

(e) all the observed values must be greater than 1.

2. The standard chi-squared test for a 2 by 2 contingency table is not valid unless:

(a) all the expected frequencies are greater than five;

(b) both variables are continuous;

(c) at least one variable is from a Normal Distribution;

(d) all the observed frequencies are greater than five;

(e) the sample is very large.

3. In Table 13M.1:

Table 13M.1. Cough first thing in the morning in a group of schoolchildren, as reported by the child and by the child's parents (Bland *et al.* 1979)

	Child's report		
Parents' report	Yes	No	Total
Yes	29	104	133
No	172	5097	5269
Total	201	5201	5402

(a) the association between reports by parents and children can be tested by a chi-squared test;

(b) the difference between symptom prevalence as reported by children and parents can be tested by McNemar's test;

(c) if McNemar's test is significant, the contingency chi-squared test is not valid;

(d) the contingency chi-squared test has one degree of freedom;

(e) it would be important to use the continuity correction in the contingency chi-squared test.

4. McNemar's test could be used:

(a) to compare the numbers of cigarette smokers among cancer cases and age- and sex-matched healthy controls;

(b) to examine the change in respiratory symptom prevalence in a group of asthmatics from winter to summer;

(c) to look at the relationship between cigarette smoking and respiratory symptoms in a group of asthmatics;

(d) to examine the change in PEFR in a group of asthmatics from winter to summer;

(e) to compare the number of cigarette smokers among a group of cancer cases and a random sample of the general population.

5. Fisher's exact test for a contingency table:

(a) applies to 2 by 2 tables;
(b) gives a larger probability than the ordinary chi-squared test;
(c) gives about the same probability as the chi-squared test with Yates' continuity correction;
(d) is suitable when expected frequencies are small;
(e) is difficult to calculate when the expected frequencies are large.

Exercise 13E

In this exercise we shall look at some data assembled to test the hypothesis that there is a considerable increase in the number of admissions to geriatric

Table 13E.1. Mean peak daily temperatures for each week from May to September of 1982 and 1983, with geriatric admissions in Wandsworth (Fish *et al.* 1985)

	1982		1983	
Week	Mean peak temperature (°C)	Admissions	Mean peak temperature (°C)	Admissions
1	12.4	24	15.3	20
2	18.2	22	14.4	17
3	20.4	21	15.5	21
4	18.8	22	15.6	17
5	25.3	24	19.6	22
6	23.2	15	21.6	23
7	18.6	23	18.9	20
8	19.4	21	22.0	16
9	20.6	18	21.0	24
10	23.4	21	26.5	21
11	22.8	17	30.4	20
12	21.7	11	25.0	25
13	22.5	6	27.3	22
14	25.7	10	22.9	26
15	23.6	13	24.3	12
16	20.4	19	26.5	33
17	19.6	13	25.0	19
18	20.2	17	21.2	21
19	22.2	10	19.7	28
20	23.3	16	16.6	19
21	18.1	24	18.4	13
22	17.3	15	20.7	29

wards during heatwaves. Table 13E.1 shows the number of admissions to geriatric wards in a health district for each week during the summers of 1982, which was cold, and 1983, which was hot. Also shown are the average of the daily peak temperatures for each week.

1. When do you think the heatwave began and ended?

2. How many admissions were there during the heatwave and in the corresponding period of 1982? Would this be sufficient evidence to conclude that heatwaves produce an increase in admissions?

3. We can use the periods before and after the heatwaves weeks as controls for changes in other factors between the years. Divide the years into three periods — before, during, and after the heatwave — and set up a two-way table of numbers of admissions, period by year.

4. We can use this table to test for a heatwave effect. State the null hypothesis and calculate the frequencies expected if the null hypothesis were true.

5. Test the null hypothesis. What conclusions can you draw?

6. What other information could be used to test the relationship between heatwaves and geriatric admissions?

14. Choosing the statistical method

14.1. Method-orientated and problem-orientated teaching

Most statistical textbooks are method orientated. They present related statistical methods together, rather than related statistical problems. This book is no exception. Thus the comparison of two groups is dealt with in Sections 8.5, 8.6, 9.7, 9.8, 10.3, 12.2, 13.1, 13.5, and 13.6, depending on whether the sample is large or small, the data Normally distributed, ordinal, nominal or dichotomous. On the other hand, all the methods involving rank statistics are together in Chapter 12. This structure is almost dictated by the subject matter, as it is easier to introduce some methods in the context of one problem, others in another. The use of the t Distribution, for example, is easier to introduce with the one-sample problem, but rank methods seem to me more obvious in the two-sample comparison.

This leads to difficulties for two groups of readers: the applier of statistics searching for the right method of analysis for the data and the student trying to answer a question in an exam. This and the next two chapters use a problem-orientated approach instead. We start with the problem and develop the statistical method required for its solution. Chapter 15 deals with some problems in clinical medicine and Chapter 16 mainly with problems in population study. In this chapter we deal with the three most common problems in statistical inference:

(a) comparison of two independent groups of subjects, for example, two groups of patients given different treatments;

(b) comparison of the response of one group under different conditions, as in a cross-over trial, or of matched pairs of subjects, as in some case-control studies;

(c) investigation of the relationship between two variables measured on the same sample of subjects.

This chapter acts as a sort of map of the methods described in Chapters 8–13. As we discussed in Section 12.7, there are often several different approaches to even a simple statistical problem. The methods described here and recommended for particular types of question may not be the only methods, and

may not always be universally agreed as the best method. Statisticians are at least as prone to disagree as clinicians. However, these would usually be considered as valid and satisfactory methods for the purposes for which they are suggested here.

14.2. Types of data

The study design is one factor which determines the method of analysis, the variable being analysed is another. We shall therefore classify variables into five types as follows.

(a) *Interval scales* The interval or distance between points on the scale has precise meaning, and a change in one unit at one scale point is the same as a change in one unit at another. For example, temperature and time are interval scales, whereas anxiety score calculated from a questionnaire is not. We can add and subtract on an interval scale.

(b) *Ordinal scale* The scale enables us to order the subjects, from that with the lowest value to that with the highest. Any ties which cannot be ordered are assumed to be because the measurement is not sufficiently precise.

(c) *Ordered nominal scale* We can group subjects into several categories, which have an order. For example, we can ask patients if their condition is much improved, improved a little, no change, a little worse, much worse.

(d) *Nominal scale* We can group subjects into categories which need not be ordered in any way. Eye colour is measured on a nominal scale.

(e) *Dichotomous scales* Subjects are grouped into only two categories, for example: survived or died. This is a special case of the nominal scale.

Clearly these classes are not mutually exclusive, and an interval scale is also ordinal. Sometimes it is useful to apply tests for a lower level of measurement, ignoring some of the information.

Interval scales allow us to calculate means and variances, and to find standard errors and confidence intervals for these. For example, in comparing two groups we can find the difference in mean between them, and estimate limits within which this should lie in the population from which the sample was drawn. This is clearly a great advantage over simply saying a difference is likely to exist, or that it may not. In particular, if a difference is 'not significant' we want to know what the maximum size of the difference could reasonably be expected to be.

For large samples, the estimation of confidence intervals presents no problem, as the means will be Normally distributed and the variances reasonably good estimates of their population values. For small samples, say less than 100 in a sample, we must assume that the observations themselves are from a Normal Distribution. Many interval scales do follow a Normal Distribution, and if not they can often be made to do so by a suitable transforma-

tion. Provided the assumption of a Normal Distribution is valid, methods based on this are the most powerful available. If Normal assumptions do not apply, methods based on ranks can be used.

For the ordinal and lower levels of measurement, most simple analyses produce tests of significance only which, as we have indicated, are less satisfactory. The only exception so far discussed is the confidence interval for the difference between two proportions.

14.3. Comparing two groups

The methods used for comparing two groups are summarized in Table 14.1.

(a) *Interval data* For large samples, say more than 50 in each group, confidence intervals for the mean can be found by the Normal approximation (Section 8.5). For smaller samples, confidence intervals for the mean can be found using the t Distribution provided the data follow or can be transformed to a Normal Distribution (Section 10.3, 10.4). If not, a significance test of the null hypothesis that the means are equal can be carried out using the Mann–Whitney U test (Section 13.1). This can be useful when the data are

Table 14.1. Methods for comparing two samples

Type of data	Size of sample	Method	Section
Interval	large, > 50 each sample	Normal Distribution for means	(8.5, 9.7)
	small, < 50 each sample, with Normal Distribution	t Distribution for means	(10.3)
	small, < 50 each sample, non-Normal	Mann–Whitney U test	(12.1)
Ordinal	any	Mann–Whitney U test	(12.1)
Nominal, ordered	large, most expected frequencies > 5	chi-squared for trend	(13.4)
Nominal, not ordered	large, most expected frequencies > 5	chi-squared test	(13.1)
	small, more than 20% expected frequencies < 5	reduce number of categories by combining or excluding as appropriate	(13.2)
Dichotomous	large, all expected frequencies > 5	Confidence interval for proportions, chi-squared test	(8.6, 9.8) (13.1)
	small, at least one expected frequency < 5	chi-squared test with Yates' correction, Fisher's exact test	(13.6) (13.5)

censored, that is, there are values too small or too large to measure. This happens, for example, when concentrations are too small to measure and are labelled 'not detectable'.

Provided that data are Normally distributed, it is possible to compare the variances of the groups. This is done by the F test, not included in this book (see Armitage 1971; Snedecor and Cochran 1980).

(b) *Ordinal data* The tendency for one group to exceed members of the other is tested by the Mann–Whitney U test (Section 12.1).

(c) *Ordered nominal data* First the data are set out as a two-way table, one variable being group and the other the ordered nominal data. A chi-squared test (Section 13.1) will test the null hypothesis that there is no relationship between group and variable, but takes no account of the ordering. This is done by using the chi-squared test for trend, which takes the ordering into account and provides a much more powerful test (Section 13.4).

(d) *Nominal data* Set the data out as a two-way table as described in (c) above. The chi-squared test for a two-way table is the appropriate test (Section 13.1). The condition for validity of the test, that at least 80 per cent of the expected frequencies should be greater than 5, must be met by combining or deleting categories as appropriate (Section 13.2). If the table reduces to a 2 by 2 table without the condition being met, use Fisher's exact test as described in (e) below.

(e) *Dichotomous data* For large samples, either present the data as two proportions and use the Normal approximation to find the confidence interval for the difference (Section 8.6), or set the data up as a 2 by 2 table and do a chi-squared test (Section 13.1). These are equivalent methods. If the sample is small, the fit to the Chi-squared Distribution can be improved by using Yates' correction (Section 13.6). Alternatively, use Fisher's exact test (Section 13.5).

14.4. One sample and paired samples

Methods of analysis for paired samples are summarized in Table 14.2.

(a) *Interval data* Inferences are on differences between the variable as observed on the two conditions. For large samples, say > 100, the confidence interval for the mean difference is found using the Normal approximation (Section 8.3). For small samples, provided the differences are from a Normal Distribution, use the paired t test (Section 10.2). This assumption is often very reasonable, as most of the variation between individuals is removed and random error is largely made up of measurement error. Furthermore, the error is the result of two added measurement errors and so tends to Normality anyway. If not, transformation of the original data will often Normalize differences (Section 10.4). If no assumption of Normality can be made, use the Wilcoxon signed-rank matched-pairs test (Section 12.2). Here the

Table 14.2. Methods for differences in one or paired sample

Type of data	Size of sample	Method	Section
Interval	large, > 100	Normal Distribution	(8.3)
	small, < 100, Normal differences	Paired *t* method	(10.2)
	small, < 100, non-Normal differences	Wilcoxon matched-pairs test	(12.2)
Ordinal	any	sign test	(9.2)
Nominal, ordered	any	sign test	(9.2)
Nominal	any	see Maxwell (1970)	(not in book)
Dichotomous	any	McNemar's test	(13.8)

assumption is that the differences are ordinal.

It is rarely asked whether there is a difference in variability in paired data. This can be tested by finding the differences between the two conditions and their sum. Then if there is no change in variance the correlation between difference and sum has expected value zero. This is by no means obvious, but it is true. Think about it.

(b) *Ordinal data* If the data do not form an interval scale, as noted in Section 14.2 the difference between conditions is not meaningful. However, we can say what direction the difference is in, and this can be examined by the sign test (Section 9.2).

(c) *Ordered nominal data* Use the sign test, with changes in one direction being positive, in the other negative, no change as zero (Section 9.2).

(d) *Nominal data* With more than two categories, this is difficult. There is a test (see Maxwell 1970). The calculation is difficult, as it involves inverting a matrix, so I have not included it. It is a generalization to more than two categories of McNemar's test (Section 13.8).

(e) *Dichotomous data* Here we are comparing the proportions of individuals in a given state under the two conditions. The appropriate test is McNemar's test (Section 13.8).

14.5. Relationship between two variables

The methods for studying relationships between variables are summarized in Table 14.3. Relationships with dichotomous variables are studied as the difference between two groups in Section 14.3, the groups being defined by the two states of the dichotomous variable. Dichotomous data have been excluded from this section.

(a) *Interval and interval data* Two methods are used: regression and

Table 14.3. Methods for relationships between variables

	Interval, normal	Interval, non-Normal	Ordinal	Nominal, ordered	Nominal	Dichotomous
Interval Normal	regression (11.3) correlation (11.10)	regression (11.3) rank correction (12.3, 12.4)	rank correlation (12.3, 12.4)	rank correlation (12.4)	analysis of variance (not in book)	*t* test (10.3) Normal test (8.5)
Interval, non-Normal		rank correlation (12.3, 12.4)	rank correlation (12.3, 12.4)	rank correlation (12.4)	analysis of variance by ranks (not in book)	large sample Normal test (8.5) Mann–Whitney *U* test (12.1)
Ordinal			rank correlation (12.3, 12.4)	rank correlation (12.4)	analysis of variance by ranks (not in book)	Mann–Whitney *U* test (12.1)
Nominal, ordered				chi-squared test for trend (13.4)	chi-squared test (13.1)	Chi-squared test for trend (13.4)
Nominal					chi-squared test (13.1)	chi-squared test (13.1)
Dichotomous						chi-squared (13.1, 13.6) Fisher's exact test (13.5)

correlation. Regression (Sections 11.3, 11.5) is usually preferred, as it gives information about the nature of the relationship as well as about its existence. Correlation (Section 11.10) measures the strength of the relationship. For regression, residuals about the line must be Normally distributed with uniform variance. The correlation coefficient requires an assumption that both variables follow a Normal Distribution, but to test the null hypothesis only one variable needs to be Normally distributed.

If neither variable can be assumed to be Normally distributed nor transformed to it (Section 11.8), use rank correlation (Sections 12.3, 12.4).

(b) *Interval and ordinal data* Rank correlation coefficient (Sections 12.3, 12.4).

(c) *Interval and ordered nominal data* This can be approached by rank correlation, using Kendall's τ (Section 12.4) because it copes with the large number of ties better than does Spearman's ρ, or by analysis of variance as described in (d) below. The latter requires an assumption of Normal Distribution and uniform variance for the interval variable. These two approaches are not equivalent.

(d) *Interval and nominal data* If the interval scale is Normally distributed, use one-way analysis of variance. This is not included in this book (see Armitage 1971; Snedecor and Cochran 1980). The assumption is that within categories the interval variable is Normally distributed with uniform variance. If this assumption is not reasonable, use analysis of variance by ranks, which has also been omitted (see Seigel 1956; Conover 1980).

(e) *Ordinal and ordinal data* Use a rank correlation coefficient, Spearman's ρ (Section 12.3) or Kendall's τ (Section 12.4). Both will give very similar answers for testing the null hypothesis of no relationship in the absence of ties. For data with many ties and for comparing the strengths of different relationships, Kendall's τ is preferable.

(f) *Ordinal and ordered nominal data* Use Kendall's rank correlation coefficient, τ (Section 12.4).

(g) *Ordinal and nominal data* Use one-way analysis of variance by ranks, not included in this book (see Seigel 1956; Conover 1980).

(h) *Ordered nominal and ordered nominal data* Use chi-squared for trend (Section 13.4).

(i) *Ordered nominal and nominal data* Use the chi-squared test for a two-way table (Section 13.1).

(j) *Nominal and nominal data* Use the chi-squared test for a two-way table (Section 13.1), provided the expected values are large enough. Otherwise use Yates' correction (Section 13.6) or Fisher's exact test (13.5).

Exercise 14M

(Each branch is either true or false.)

1. The following variables have interval scales of measurement:

(a) height;
(b) presence or absence of asthma;
(c) Apgar score;
(d) age;
(e) forced expiratory volume.

2. The following methods may be used to investigate a relationship between two continuous variables:

(a) paired t test;
(b) the correlation coefficient, r;
(c) simple linear regression;
(d) Kendall's τ;
(e) Spearman's ρ.

3. Ten men were given a drug and a placebo on alternate days in random order. The exact time for which the patients could exercise until angina or fatigue stopped them was measured. Methods which could be used to investigate the existence of a treatment effect include:

(a) Mann–Whitney U test;
(b) paired t method;
(c) sign test;
(d) Normal confidence intervals for the mean difference;
(e) Wilcoxon matched-pairs test.

4. When analysing categorical variables the following statistical methods may be used:

(a) simple linear regression;
(b) correlation coefficient, r;
(c) paired t test;

(d) Kendall's τ;

(e) chi-squared test.

5. To compare levels of a continuous variable in two groups, possible methods include:

(a) the Mann–Whitney U test;

(b) Fisher's exact test;

(c) a t test;

(d) Wilcoxon matched-pairs test;

(e) the sign test.

Exercise 14E

In this exercise we shall look at a number of statistical problems. The object is not to carry out calculations, but to decide which statistical method is appropriate. If you wish, you can carry out the calculations for practice, but only brief solutions are given. Sometimes more than one of the methods discussed in this book are possible, as well as others which we have not discussed.

1. In a cross-over trial to compare two appliances for ileostomy patients, of 14 patients who received system A first, 5 expressed a preference for A, 9 for system B and none had no preference. Of the patients who received system B first, 7 preferred A, 5 preferred B and 4 had no preference. How

Table 14E.1. Gastric pH and urinary nitrite concentrations in 26 subjects (Hall and Northfield 1985)

pH	Nitrite	pH	Nitrite
5.71	21.9	5.55	83.8
5.18	0.0	1.93	7.13
2.94	6.53	2.17	1.48
2.11	0.19	4.94	55.6
6.03	19.5	2.03	15.7
2.64	2.33	2.73	52
4.07	22.7	1.94	12.1
5.86	3.26	1.72	1.64
4.91	17.8	5.31	43.9
2.17	9.36	5.29	50.6
5.5	35.2	5.90	63.4
5.91	81.2	5.77	48.9
5.59	81.8	5.59	52.5

would you decide whether one treatment was preferable? How would you decide whether the order of treatment influenced the choice?

2. Table 14E.1 shows the pH and nitrite concentrations in samples of gastric fluid from 26 patients. A scatter diagram is shown in Fig. 14E.1. How would you assess the evidence of a relationship between pH and nitrite concentration?

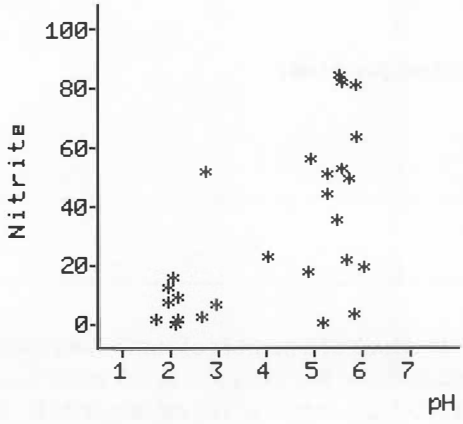


Fig. 14E.1 Gastric pH and urinary nitrite.

3. There is some concern about the number of instrumental deliveries experienced by women receiving epidural anaesthesia during labour. It is hoped that an improved epidural regime will reduce this. How do you decide how many women will be needed for a trial?

4. In a trial of screening and treatment for mild hypersion (Reader *et al.* 1980), 1138 patients completed the trial on active treatment, with 9 deaths, and 1080 completed on placebo, with 19 deaths. A further 583 patients allocated to active treatment withdrew, of whom 6 died, and 626 allocated to placebo withdrew, of whom 16 died during the trial period. How would you decide whether screening and treatment for mild hypertension reduces the risk of dying?

5. Burr *et al.* (1976) tested a procedure to remove house-dust mites from the bedding of adult asthmatics in attempt to improve subjects' lung function, which they measured by PEFR. The trial was a two-period cross-over design, the control or placebo treatment being thorough dust removal from the living room. The mean PEFRs in the 32 subjects were:

active treatment:	335 litre/min, s.e. = 19.6 litre/min
placebo treatment:	329 litre/min, s.e. = 20.8 litre/min

differences within subjects (treatment – placebo): 6.45 litre/min, s.e. = 5.05 litre/min

How would you decide whether the treatment improves PEFR?

6. Table 14E.2 shows the relationship between age of onset of asthma in children and maternal age at the child's birth. How would you test whether these were related? The children were all born in one week in March, 1958. Apart from the possibility that young mothers in general tend to have children prone to asthma, what other possible explanations are there for this finding?

Table 14E.2. Asthma or wheeze by maternal age

	Mother's age at child's bith		
AW reported	15–19	20–29	30+
Never	261	4017	2146
Onset by age 7	103	984	487
Onset from 8 to 11	27	189	95
Onset from 12 to 16	20	157	67

7. The lung function of 79 children with a history of hospitalization for whooping cough and 178 children without a history of whooping cough, taken from the same school classes, was measured. The mean transit time for the whooping cough cases was 0.49 s (s.d. = 0.14 s) and for the controls 0.47 s (s.d. = 0.11 s), (Johnston *et al.* 1983). Was there a difference in lung function between children who had had whooping cough and those who had not? Each case had two matched controls. If you had all the data, how could you use this information?

15. Clinical measurement

15.1. Repeatability and precision in measurement

In this chapter we shall look at a number of problems associated with clinical measurement. These include how precisely we can measure, how different methods of measurement can be compared, how measurements can be used in diagnosis, and how to deal with incomplete measurements of survival.

We have already discussed some factors which may produce bias in measurements (Sections 2.7, 2.8, 3.9). We have not yet considered the natural biological variability, in subject and in measurement method, which may lead to measurement error. For example, in the measurement of blood pressure we are dealing with a quantity that varies continuously, not only from heart beat to heart beat but from day to day, season to season, and even with the sex of the measurer. The measurer, too, will show variation in the perception of the sound and reading of the manometer. Because of this, most clinical measurements cannot be taken at face value without some consideration being given to their error.

The measurement of measurement error is not difficult in principle. To do it we need a set of duplicate readings, obtained, say, by measuring each member of a sample of subjects twice. We can then estimate the standard deviation of repeated measurements on the same subject, which is one possible method of representing measurement error. If the pairs of measurements are x_i and y_i for $i = 1$ to n, the best estimate of the error standard deviation (see Appendix 15A) is given by

$$s = \sqrt{\frac{1}{2n} \Sigma (x_i - y_i)^2}$$

Table 15.1 shows some replicated measurements of peak expiratory flow rate, made with a Wright Peak Flow Meter (see Section 10.2). It also shows the differences between the first and second measurements and their sum of squares. We get a very small mean difference, of 4.9 litre/min, suggesting that there is little tendency for the second or first reading to be larger. The standard deviation of the measurement error is 15 litre/min, to two significant figures. There are a number of ways in which the measurement error may be presented to the user of the measurement. It may be as the standard deviation calculated above, or it may be, as recommended by the British Standards Institution (1979), the value below which the difference between

Table 15.1. Pairs of readings made with a Wright Peak Flow Meter on 17 healthy volunteers, with the calculation of repeatability

Subject number	First PEFR (litre/min)	Second PEFR (litre/min)	First – Second	Difference squared
1	494	490	4	16
2	395	397	−2	4
3	516	512	4	16
4	434	401	33	1089
5	476	470	6	36
6	557	611	−54	2916
7	413	415	−2	4
8	442	431	11	121
9	650	638	12	144
10	433	429	4	16
11	417	420	−3	69
12	656	633	23	529
13	267	275	−8	64
14	478	492	−14	196
15	178	165	13	169
16	423	372	51	2601
17	427	421	6	36
Total	15 228		84	7966
Mean	447.9		4.9	
Total/2n = s^2				234.3
s				15.3

two measurements will lie with probability 0.95. Provided the measurement errors are from a Normal Distribution, this is estimated by $1.96 \times \sqrt{(2s^2)}$, or $2.8s$.

It may also be reported as the *coefficient of variation*, which is the standard deviation divided by the mean, often multiplied by 100 to give a percentage. I do not like this statistic much, as its value depends on both the standard deviation and the mean. Small means produce large coefficients of variation. For our data the coefficient of variation is 15.3/447.9 = 0.034 or 3.4 per cent. The difference between the observed value, with measurement error, and the subject's true value will be at most two standard deviations with probability 0.95, provided the sample is large enough, and so we may have the accuracy quoted as to within $2s$, $2 \times 15.3 \simeq 30$ litre/min, or $2 \times 15.3/447.9 = 0.068$, or 7 per cent. The trouble with quoting the error as a percentage is that 7 per cent of the smallest observation, 165 litres, is only 12 litre/min, compared to 7 per cent of the largest, 656, which is 46 litre/min. This is not a good method if the range is great compared to the size of the smallest observations and the error does not depend on the value of the measurement. It is a good method if the standard deviation is proportional to the

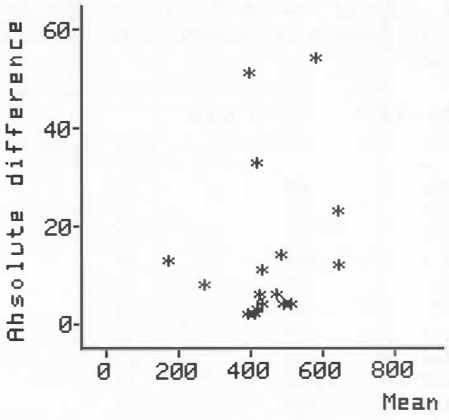


Fig. 15.1 Absolute difference versus mean for 17 pairs of Wright peak flow meter measurements.

mean. In that case a logarithmic transformation can be used, and the antilog of two standard deviations gives the number corresponding to 0.068 in the example. If we multiply this by 100, it will give the measurement error as a percentage of the measurement wherever we are on the scale. The use of logarithmic transformations is discussed in Section 10.4.

We should check to see whether the error does depend on the value of the measurement, usually being larger for larger values. We can do this by plotting a scatter diagram of the absolute value of the difference and the sum or mean of the two observations (Fig. 15.1). For the PEFR data, there is no obvious relationship. We can check this by calculating a correlation (Section 11.10) or rank correlation coefficient (Sections 12.4, 12.5). For Fig. 15.1 we have $r = 0.20, p = 0.4$, so there is little to suggest that the measurement error is related to the size of the PEFR.

15.2. Digit preference

One of the constraints on the accuracy of measurement is the measurement instrument itself. A clinical thermometer might be graduated in fifths of a degree, and attempts to read temperature to a tenth of a degree will be at best approximations. Even reading to the nearest fifth of a degree may be suspect, as it will often be difficult to decide whether a point between two divisions is nearer one line or another. This is complicated by the fact that most of us under these circumstances tend to prefer some terminal digits over others. In Table 15.1 there is little evidence of this and only 9 seems at all under-represented. Table 15.2 shows corresponding data, obtained for the same subjects at the same time, using a different instrument, the Mini Wright Peak

Table 15.2. Pairs of readings made with a mini-Wright Peak Flow Meter on 17 healthy volunteers

Subject number	First PEFR (litre/min)	Second PEFR (litre/min)
1	512	525
2	430	415
3	520	508
4	428	444
5	500	500
6	600	625
7	364	460
8	380	390
9	658	642
10	445	432
11	432	420
12	626	605
13	260	227
14	477	467
15	259	268
16	350	370
17	451	443

Flow Meter, an instrument with a cruder scale. Table 15.3 shows the terminal digits of the readings in Tables 15.1 and 15.2. Clearly 0, 2, 5, 7 and 8 are preferred in Table 15.2, a trend just discernable in Table 15.1 also, where these make up 20 of the 34 rather than the 17 expected. This is despite the measurer, myself, being fully aware of the possibility of digit preference. Table 15.3 also shows the last digits of the 57 FEV1 measurements in Table 4.5. These were each measured by a different student, and the vast excess of zeros is due in part to some recording the answer to only one decimal place. Even so, the lack of 'ones' is clear.

Table 15.3. Terminal digits of three sets of observations

Table 15.1	Table 15.2	Table 4.5
0000	000000000000	0000000000000 0000000000000
1111	1	1
2222	2222	22
3333	3	33
44	44	444444
5555	55555	5555
666	6	666
77777	777	7777
888	8888	888888
9	9	999

Does digit preference matter? It does if differences in the last digit are of importance to the outcome, as it might be in Table 15.1, where we are dealing with the difference between two similar numbers. Because of this it is a mistake to have one measurer take readings under one set of conditions and a second under another, as their degree of digit preference may differ. It is also important to agree the number of figures to be recorded and to ensure that instruments have sufficiently fine scales for the job in hand. The last digit in Table 15.2 is almost meaningless, for example. Of course, for clinical purposes and in view of the measurement error, this does not matter.

15.3. Comparing two methods of measurement

In clinical measurement, most of the things we want to measure — hearts, lungs, livers and so on — are deep within living bodies and out of reach. This means that many of the methods we use to measure them are indirect and we cannot be sure how closely they are related to what we really want to know. When we develop a new method of measurement, rather than compare its outcome to a set of known values we must often compare it to another method just as indirect. This is a common type of study, and one which is often badly done (Bland and Altman 1986).

Tables 15.1 and 15.2 show measurements of PEFR by two different

Table 15.4. Comparison of two methods of measuring PEFR

Subject number	PEFR (litre/min) Wright meter	Mini-meter	Difference
1	494	512	−18
2	395	430	−35
3	516	520	−4
4	434	428	6
5	476	500	−24
6	557	600	−43
7	413	364	49
8	442	380	62
9	650	658	−8
10	433	445	−12
11	417	432	−15
12	656	626	30
13	267	260	7
14	478	477	1
15	178	259	−81
16	423	350	73
17	427	451	−24
Total			−36
Mean			2.1
Standard deviation			38.8

methods. For simplicity, we shall use only one measurement by each method in the following (Table 15.4). We could make use of all the data by using the average for each method, but this introduces an extra stage in the calculation. Bland and Altman (1986) give details.

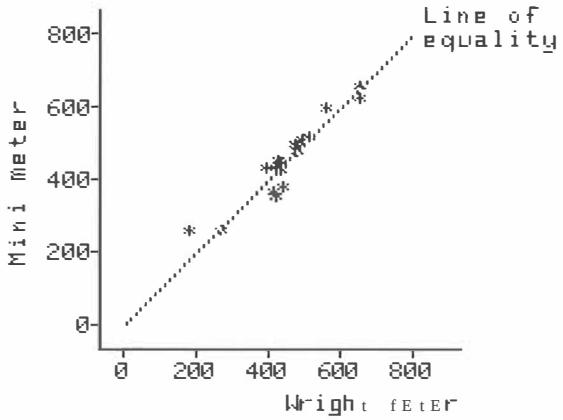


Fig. 15.2 PEFR measured by two different instruments.

The first step in the analysis is to plot the data on a scatter diagram (Fig. 15.2). If we draw the line of equality, along which the two measurements would be exactly equal, this gives us an idea of the extent to which the two methods compared. This is not the best way of looking at data of this type, because much of the graph is empty space and the interesting information is clustered along the line. A better approach is to plot the difference between the methods against the sum or average. This time the sign of the difference is important, as there is a possibility that one method may give higher values than the other and this may be related to the true value we are trying to measure. This plot is shown in Fig. 15.3, together with a histogram of the differences. There is no clear evidence of a relationship between difference and mean, and we can check this by a test of significance using the correlation coefficient. We get $r = 0.19$, $p = 0.5$. There is little evidence of overall bias, the mean difference being close to zero. We can find a confidence interval for the mean difference as described in Section 10.2. The differences have a mean of -2.1 litre/min, and a standard deviation of 38.87. The standard error of the mean is thus $s/\sqrt{n} = 38.8/\sqrt{17} = 9.41$ litre/min and the corresponding value of t with 16 degrees of freedom is 2.12. The 95 per cent confidence interval for the bias is thus $-2.1 \pm 2.12 \times 9.41 = -22$ to $+18$ litre/min. Thus on the basis of these data we could have a bias of as much as 22 litre/min, which could be clinically important. The original comparison of these instruments used a much larger sample and found that any bias was very small (Oldham *et al.* 1979).

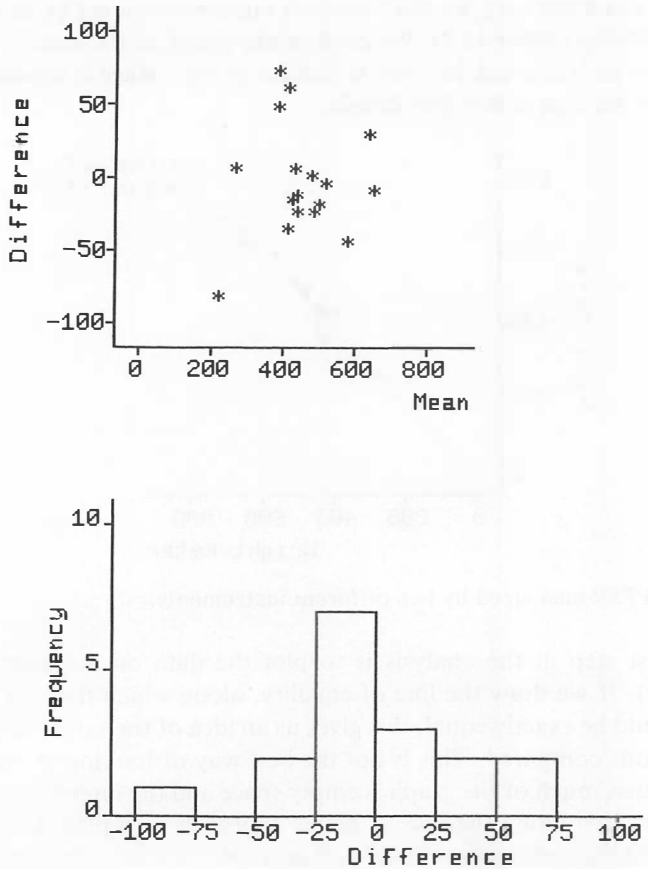


Fig. 15.3 Difference versus mean for PEFR measured by two different instruments.

The standard deviation of the differences between measurements made by the two methods provides a good index of the comparability of the methods. If we can estimate the mean and standard deviation reliably, with small standard errors, we can then say that the difference between methods will be at most two standard deviations on either side of the mean except with a small probability. We can check how closely the differences follow a Normal Distribution from their histogram.

The standard deviation of the differences is estimated to be 38.8 litre/min and the mean is −2 litre/min. Two standard deviations is therefore 78 litre/min. The reading with the mini-meter is expected to be 80 litres below to 76 litres above for most subjects. Certainly on the basis of these data we would not conclude that the two methods are comparable or that the mini-

meter could reliably replace the Wright peak flow meter. As remarked in Section 10.2, this meter had received considerable wear.

15.4. Sensitivity and specificity

One of the main purposes of making clinical measurements is to aid in diagnosis. This may be to identify one of several possible diagnoses in a patient, or to find people with a particular disease in an apparently healthy population. The latter is known as *screening*. In either case the measurement provides us with a test which we may be able to compare later with a true diagnosis. The test may be based on a continuous variable and the disease indicated if it is above or below a given level, or it may be a qualitative observation such as carcinoma in-situ cells on a cervical smear. In either case we shall call the test positive if it indicates the disease and negative if not, and the diagnosis positive if the disease is later confirmed, negative if not.

How do we measure the effectiveness of the test? Table 15.5 shows three artificial sets of test and diagnosis data. We could take as an index of test

Table 15.5. Some artificial test and diagnosis data

	True diagnosis		
Test 1	Positive	Negative	Total
positive	4	5	9
negative	1	90	91
Total	5	95	100

	True diagnosis		
Test 2	Positive	Negative	Total
positive	0	0	0
negative	5	95	100
Total	5	95	100

	True diagnosis		
Test 3	Positive	Negative	Total
positive	2	0	2
negative	3	95	98
Total	5	95	100

effectiveness the proportion giving the true diagnosis from the test. For Test 1 in the example it is 94 per cent. Now consider Test 2, which always gives a negative result. Test 2 will never detect any cases of the disease. We are now right for 95 per cent of the subjects! However, the first test is useful, in that it detects some cases of the disease, and the second is not, so this is clearly a poor index. We could use a coefficient of agreement, for example the number positive on both tests over the number positive on at least one test. For Test 1 this is $4/(4 + 5 + 1) = 0.4$; for Test 2 it is $0/(0 + 0 + 5) = 0$. This is better, but still not good enough. Compare Test 3, which has the same coefficient of agreement as Test 1, $2/(2 + 0 + 3) = 0.4$. However, Test 3 is not as good as Test 1 in one respect: it only detects 2 of the 5 true positives, compared to 4. On the other hand, it is a better test in another way: it does not diagnose as positive any true negatives.

There is no one simple index which enables us to compare different tests in all the ways we would like. This is because there are two things we need to measure. One is how good the test is at finding true positives i.e. those with the condition. The other is how good the test is at excluding true negatives, i.e. those who do not have the condition. The indices conventionally employed to do this are:

$$\text{sensitivity} = \frac{\text{true + ve who are also test + ve}}{\text{all true + ve}}$$

$$\text{specificity} = \frac{\text{true − ve who are also test − ve}}{\text{all true − ve}}$$

In other words, the sensitivity is a proportion of true positives who are test positive, and the specificity is the proportion of true negatives who are test negatives. For our three tests these are:

	Sensitivity	*Specificity*
Test 1	0.80	0.95
Test 2	0.00	1.00
Test 3	0.40	1.00

Test 2, of course, misses all the true positives and finds all the true negatives, by saying all are negative. The difference between Tests 1 and 3 is brought out by the greater sensitivity of 1 and the greater specificity of 3. We are comparing tests in two dimensions. We can see that Test 3 is better than Test 2, because its sensitivity is higher and its specificity is the same. However, it is more difficult to see whether Test 3 is better than Test 1. We must come to a judgement based on the relative importance of sensitivity and specificity in the particular case.

When the test is based on a continuous variable, we can alter the sensitivity and specificity by changing the cut-off point. If high values indicate the disease, raising the cut-off point will mean fewer cases will be detected and so the sensitivity will be decreased. However, there will be fewer false positives, positives on test but in fact not diseased, and the specificity will be increased. On the other hand, if we lower the cut-off point we shall detect more cases and the sensitivity will be increased, but we shall have more false positives and the specificity will be decreased.

For a practical example, Maxwell *et al.* (1983) observed that a remarkable number of alcoholics had evidence at X-ray of past rib fractures. We asked whether this would be of any value in the detection of alcoholism in patients. Among 74 patients with alcoholic liver disease, 20 had evidence of at least one past fracture on chest X-ray and 11 had evidence of bilateral or multiple fractures. In a control group of 181 patients with non-alcoholic liver disease or gastrointestinal disorders, 6 had evidence of at least one fracture and 2 of bilateral or multiple fractures.

For any fractures as a test for alcoholism, the sensitivity was $20/74 = 0.27$, and the specificity $(181 - 6)/181 = 0.97$. For bilateral or multiple fractures the sensitivity was $11/74 = 0.15$ and the specificity was $(181 - 2)/181 = 0.99$. Hence both tests were very specific; very few non-alcoholics would be indicated as alcoholics by them. On the other hand, neither was very sensitive; many alcoholics would be missed. As might be expected, the more stringent test of bilateral or multiple fractures was more specific and less sensitive than the test of any fracture.

Sensitivity and specificity are often multiplied by 100 to give percentages. They are both binomial proportions, so their standard errors and confidence intervals are found as described in Section 8.4 and the sample size required for their reliable estimation can be calculated as described in Section 8.8.

15.5. Normal or reference ranges

In Section 15.4 we were concerned with the diagnosis of particular diseases. In this section we look at it the other way round and ask within what range of values measurements on normal, healthy people will lie. We should then be able to say that measurements outside this range are indicative of disease.

There are great difficulties in doing this. Who is 'normal' anyway? In the UK population almost everyone has hard fatty deposits in their coronary arteries, which result in death for about half of them. Very few Africans have this; they die from other causes. So it is normal in the UK to have an abnormality. We can set this problem aside and say that normal people are the apparently healthy members of the local population. We can draw a

sample of these as described in Chapter 3 and make the measurement on them.

The next problem is to calculate the range. If we use the range as defined in Section 4.7, the difference between the two most extreme values, we can be fairly confident that if we carry on sampling we will eventually find observations outside it, and the range will get bigger and bigger. To avoid this we use a range between two quantiles (Section 4.7), usually the 2.5 percentile and the 97.5 percentile. This leaves 5 per cent of normals outside the 'normal range'. Thus the normal range or reference range is the set of values within which 95 per cent of measurements from apparently healthy individuals will lie.

A third difficulty comes from confusion between 'normal' as used in Medicine and 'Normal Distribution' as used in Statistics. This had led some people to develop approaches which say that all data which do not fit under a Normal curve are abnormal! Such conclusions are simply absurd; there is no reason to suppose that all variables are Normally distributed (Sections 7.2, 7.4, 7.5). The term 'reference range', which is becoming widely used, has the advantage of avoiding this confusion. However, the most commonly used method of calculation rests on the assumption that the variable is, in fact, Normally distributed.

We have already seen that in general most observations fall within two standard deviations of the mean, and that for a Normal Distribution 95 per cent are within these limits with 2.5 per cent below and 2.5 per cent above. If we estimate the mean and standard deviation of data from a Normal population we can estimate the reference range as $(\bar{x} - 2s)$ to $(\bar{x} + 2s)$.

Consider the FEV1 data of Table 4.5. We shall estimate the reference range for FEV1 in male medical students. We have 57 observations, mean 4.06 and standard deviation 0.67 litres. The reference range is thus 2.7 to 5.4 litres. From Table 4.4 we see that in fact only one student, 2 per cent, is outside these limits although the sample is rather small.

Standard errors and confidence intervals for these limits are easy to find, provided the observations are from a Normal Distribution. The estimates $\bar{x}$ and s are independent with standard errors $\sqrt{(s^2/n)}$ and $\sqrt{[s^2/2(n-1)]}$ (Sections 8.2, 8.7). The value of $\bar{x}$ follows a Normal Distribution and s a distribution which is approximately Normal. Hence $(\bar{x} - 2s)$ is from a Normal Distribution with variance:

$$Var(\bar{x} - 2s) = Var(\bar{x}) + Var(2s) = Var(\bar{x}) + 4Var(s)$$
$$= \frac{s^2}{n} + 4 \times \frac{s^2}{2(n-1)} = s^2\left(\frac{1}{n} + \frac{2}{n-1}\right)$$

Hence, provided Normal assumptions hold, the standard error of the limit of the reference range is

$$\sqrt{s^2\left(\frac{1}{n} + \frac{2}{n-1}\right)}$$

If n is large, this is approximately

$$\sqrt{\frac{3s^2}{n}}$$

For the FEV1 data, this is $\sqrt{(3 \times 0.67^2/57)} = 0.15$. Hence the 95 per cent confidence intervals for these limits are $2.7 \pm 1.96 \times 0.15$ and $5.4 \pm 1.96 \times 0.15$ i.e. from 2.4 to 3.0 and 5.1 to 5.7 litres. These confidence intervals seem quite wide to me, yet reference ranges based on much smaller samples are often reported in the literature without any estimates of their precision.

Compare the triglyceride levels of Fig. 4.13. Here the data are highly skewed, and the Normal model does not fit. The lower limit is 0.07, well below any of the observations, and the upper limit is 0.94, greater than which are 5 per cent of the observations. It is possible for such data to give a negative lower limit!

The triglyceride values are highly skewed to the right, which suggests that a logarithmic transformation would help Normalize the data (Fig. 7.16). Figure 7.17 shows the $\log_{10}$-transformed data, which give a breathtakingly symmetrical distribution ($\bar{x} = -0.331$, s = 0.171). The lower limit in the transformed data is -0.67, corresponding to a triglyceride level of 0.21, below which are 2.1 per cent of observations. The upper limit is 0.01, corresponding to 1.02, above which are 2.5 per cent of observations. The fit to the log-transformed data is excellent. For the standard error of the reference limit we have $\sqrt{(3 \times 0.171^2/282)} = 0.0176$. The 95 per cent confidence intervals are thus $-0.673 \pm 2 \times 0.0176$ and $0.011 \pm 2 \times 0.0176$, i.e. -0.707 to -0.637 and -0.025 to 0.046. In the untransformed data this gives 0.196 to 0.231 and 0.945 to 1.112, found by taking the antilogs. These confidence limits can be transformed back to the original scale, unlike those in Section 10.4, because no subtraction of means has taken place.

Because of the obviously unsatisfactory nature of the Normal model for some data, some authors have advocated the estimation of the percentiles directly as in Section 4.5, without any distributional assumptions. This is an attractive idea. We want to know the point below which 2.5 per cent of values will fall. Let us simply rank the observations and find the point below which 2.5 per cent of the observations fall. For the 282 triglicerides, the 2.5 and 97.5 percentiles are found as follows. For the 2.5 percentile, we find $i = q(n + 1) = 0.025 \times (282 + 1) = 7.08$. The required quantile will be between the 7th and 8th observation. The 7th is 0.21, the 8th is 0.22 so the 2.5 percentile would be estimated by $0.21 + (0.22 - 0.21) \times (7.08 - 7) = 0.211$. Similarly the 97.5 percentile is 1.049.

This approach gives an unbiased estimate whatever the distribution. The log-transformed triglyceride would give exactly the same results. Note that the Normal theory limits from the log-transformed data are very similar. We

now look at the confidence interval. The 95 per cent confidence interval for the q quantile, here q being 0.025 or 0.975, estimated directly from the data, is found by an application of the Binomial Distribution (Sections 6.4, 6.6) (see Conover 1980). (The number of observations less than the q quantile will be an observation from a Binomial Distribution with parameters n and q.)

$$j = nq - 1.96\sqrt{nq(1-q)}$$
$$k = nq + 1.96\sqrt{nq(1-q)}$$

We round j and k up to the next integer. Then the 95 per cent confidence interval is between the jth and the kth observations in the ordered data. For the triglyceride, $n = 282$ and so for the lower limit, $q = 0.025$, we have

$$j = 282 \times 0.025 - 1.96\sqrt{282 \times 0.025 \times 0.975}$$
$$k = 282 \times 0.025 + 1.96\sqrt{282 \times 0.025 \times 0.975}$$

This gives $j = 1.9$ and $k = 12.2$, which we round up to $j = 2$ and $k = 13$. In the triglyceride data the second observation, corresponding to $j = 2$, is 0.19 and the 13th is 0.26. Thus the 95 per cent confidence interval for the lower reference limit is 0.19 to 0.26. The corresponding calculation for $q = 0.975$ gives $j = 270$ and $k = 281$. The 270th observation is 0.98 and the 281st is 1.62, giving a 95 per cent confidence interval for the upper reference limit of 0.98 to 1.62. These are wider confidence intervals than those found by the Normal method, those for the long tail particularly so. This suggests that this method of estimating percentiles in long tails is imprecise.

15.6. Survival data

Survival data arise in many ways in medical research. The most obvious is in studying the length of time patients live after a treatment or after the onset of a disease. We are concerned with the length of time elapsed between entry (start of disease, start of treatment, randomization in a trial) and exit from the population (death). There are other processes which have the same characteristic. For example, in the chemotherapy of gall stones, we can observe via ultra-sound the length of time between the start of treatment and the disappearance of the stone. In the study of infertility we can observe the length of time between treatment and conception. In this section we shall talk about times from entry to death, but other applications should be obvious.

Problems arise in the measurement of survival because often we do not know the exact survival times of all cases. This is because some will still be surviving when we want to analyse the data. When cases have entered the study at different times, some of the recent entrants may be surviving, but only have been observed for a short time. Their survival time may be less than those cases admitted early in the study and who have since died. The method of calculating survival curves described below takes this into account. When

Table 15.6. Survival time in years of patients after diagnosis of parathyroid cancer

Alive	Deaths
less than 1	less than 1
less than 1	2
1	6
1	6
4	7
5	9
6	9
8	11
10	14
10	
17	

we know some of the observations exactly, and only that others are greater than some value, we say that the data are *censored*. We overcome this difficulty by the construction of a life table.

Table 15.6 shows some typical survival data, for patients with parathyroid cancer. The survival times are recorded in completed years. A patient who survived for 6 years and then died can be taken as having lived for 6 years and then died in the seventh. In the first year from diagnosis, one patient died. Three patients were observed for only part of this year and 17 survived into the next year. The 2 who have only been observed for part of the year are said to be *withdrawn from follow-up*. There is no information about their survival after the first year, because it has not happened yet. These patients are only at risk of dying for part of the year and we cannot say that 1 out of 20 died as they may yet contribute another death in the first year. We can say that such patients will contribute half a year of risk, on average, so the number of patient years at risk in the first year is 18 (17 who survived and 1 who died) plus 2 halves for those withdrawn from follow-up, giving 19 altogether. We get an estimate of the probability of dying in the first year of 1/19, and an estimated probability of surviving of 1 – 1/19. We can do this for each year until the limits of the data are reached. We thus trace the survival of these patients estimating the probability of death or survival at each year and the cumulative probability of survival to each year. This set of probabilities is called a *life table*.

To carry out the calculation, we first set out for each year the number alive at the start, the number withdrawn during the year and the number at risk and the number dying (Table 11.7). Thus in year 1 the number at the start is 20, the number withdrawn is 2, the number at risk 19 and the number of deaths is 1. As there were 2 withdrawals and 1 death the number at the start of year 2 is 17. For each year we calculate the probability of dying in that year for patients who have reached the beginning of it, and hence the probability of

Table 15.7. Life table calculation for parathyroid cancer survival

Year x	Number at start n_x	Withdrawn during year w_x	At risk $r_x = n_x - \frac{1}{2}w_x$	Deaths d_x	Prob. of death $q_x = d_x/r_x$	Prob. of surviving year x $p_x = 1 - q_x$	Cumulative Prob. of surviving x years $P_x = p_x P_{x-1}$
1	20	2	19	1	0.0526	0.9474	0.9474
2	17	2	16	0	0	1	0.9474
3	15	0	15	1	0.0667	0.9333	0.8842
4	14	0	14	0	0	1	0.8842
5	14	1	$13\frac{1}{2}$	0	0	1	0.8842
6	13	1	$12\frac{1}{2}$	0	0	1	0.8842
7	12	1	$11\frac{1}{2}$	2	0.1739	0.8261	0.7304
8	9	0	9	1	0.1111	0.8889	0.6493
9	8	1	$7\frac{1}{2}$	0	0	1	0.6493
10	7	0	7	2	0.2857	0.7143	0.4638
11	5	2	4	0	0	1	0.4638
12	3	0	3	1	0.3333	0.6667	0.3092
13	2	0	2	0	0	1	0.3092
14	2	0	2	0	0	1	0.3092
15	2	0	2	1	0.5	0.5	0.1546
16	1	0	1	0	0	1	0.1546
17	1	0	1	0	0	1	0.1546
18	1	1	$\frac{1}{2}$	0	0	1	0.1546

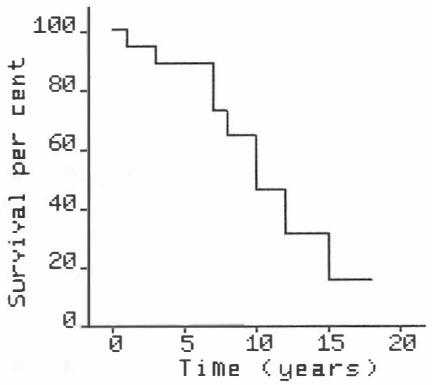


Fig. 15.4 Survival curve for parathyroid cancer patients.

surviving to the next year. Finally we calculate the cumulative survival probability. For the first year, this is the probability of surviving that year, $P_1 = p_1$. For the second year, it is the probability of surviving up to the start of the second year, P_1, times the probability of surviving that year, p_2, to give $P_2 = p_2 P_1$. The probability of surviving for 3 years is similarly $P_3 = p_3 P_2$, and so on. From this life table we can estimate the *five-year survival rate*, a useful measure of prognosis in cancer. For the parathyroid cancer, the five-year survival rate 0.8842, or 88 per cent. We can see that the prognosis for this cancer is quite good.

We can draw a graph of the cumulative survival probability, the *survival curve*. This is usually drawn in steps, with abrupt changes in probability (Fig. 5.4). This convention emphasizes the relatively poor estimation at the long survival end of the curve, where the small numbers at risk produced large steps.

The standard error for the survival probabilities can be found (see Armitage 1973) and two survival curves can be compared by several significance tests, of which the best known is the logrank test (Peto *et al.* 1977).

15.7. Computer-aided diagnosis

Reference ranges (Section 15.5) are one area where statistical methods are involved directly in diagnosis, computer-aided diagnosis is another. Computers are machines which can store large amounts of information and process them very quickly. In statistics they are very widely used to store data, calculate, draw graphs, etc. They are used in many other areas of medicine, from medical records to patient interviews, but none cause so much argument as those used in computer-aided diagnosis. The 'aided' is put in to persuade

clinicians that the main purpose is not to do them out of a job, but, naturally, they have their doubts.

This is not a book about medical computing (see Kember 1982; Norris *et al.* 1985), but computer-aided diagnosis is partly a statistical exercise. In fact, there are two types of computer-aided diagnosis: statistical methods, where diagnosis is based on a set of data obtained from past cases; and decision-tree methods, which try to imitate the thought processes of an expert in the field. We shall look briefly at each approach.

There are several methods of statistical computer aided diagnosis. One uses *discriminant analysis*. In this we start with a set of data on subjects whose diagnosis was subsequently confirmed, and calculate one or more discriminant functions. A discriminant function has the form:

$$\text{constant}_1 \times \text{variable}_1 + \text{constant}_2 \times \text{variable}_2 + \ldots + \text{constant}_n \times \text{variable}_n$$

The constants are calculated so that the values of the functions are as similar as possible for members of the same group and as different as possible for members of different groups. In the case of only two groups, we have one discriminant function and all the subjects in one group will have high values of the function and all subjects in the other will have low values. For each new subject we evaluate the discriminant function and use it to allocate the subject to a group or diagnosis. We can say what the probability is of the subject falling in that group, and in any other. We have already come across an example of this technique in Exercise 2E. Many forms of discriminant analysis have been developed to try and improve this form of computer diagnosis, but it does not seem to make much difference which is used.

A different approach is *Bayesian analysis*. This is based on *Bayes' Theorem*, a result about probability which may be stated in terms of the probability of diagnosis A having observed data B, as:

$$\text{Prob(diagnosis A if data B)} = \frac{\text{Prob(data B if diag A)} \times \text{Prob(diag A)}}{\text{Prob(data B)}}$$

If we have a large data set of known diagnoses and their associated symptoms and signs, we can determine the Prob(diagnosis A) easily. It is simply the proportion of times A has been diagnosed. The problem of finding the probability of a particular combination of symptoms and signs is more difficult. If they are all independent, we can say that the probability of a given symptom is the proportion of times it occurs, and the probability of the symptom for each diagnosis is found in the same way. The probability of any combination of symptoms can be found by multiplying their individual probabilities together, as described in Section 6.2. In practice the assumption that signs and symptoms are independent is most unlikely to be met and a more complicated analysis would be required to deal with this. However, some systems

of computer-aided diagnosis have been found to work quite well with the simple approach.

Expert or knowledge-based systems work in a different way. Here the knowledge of a human expert or group of experts in the field is converted into a series of decision rules, e.g. 'if the patient has post-bilateral rib fractures then the patient is an alcoholic, if not then on to the next decision'. These systems can be modified by asking further experts to test the system with cases from their own experience and to suggest further decision rules if the program fails. They also have the advantage that the program can 'explain' the reason for its 'decision' by listing the series of steps which led to it.

Most of Chapter 14 consists of rules of just this type and could be turned into an expert system for statistical analysis. Indeed some of my colleagues are already discussing the possibility of an expert system for medical statistics.

Although there have been some impressive achievements in the field of computer-aided diagnosis, it has to date made little progress towards acceptance in routine medical practice. As computers become more familiar to clinicians, more common in their surgeries and more powerful in terms of data storage and processing speed, we may expect computer-aided diagnosis to become as well established as computer-aided statistical analysis is today.

Appendix 15A

Standard deviation for measurement error

The standard deviation for the error in repeated measurements is found as follows. We want to find the variance within subjects, as we found the variance within groups for the two-sample t test (Section 10.3). The sum of squares for one subject is

$$\left(x_i - \frac{x_i + y_i}{2}\right)^2 + \left(y_i - \frac{x_i + y_i}{2}\right)^2 = \left[\frac{2x_i - (x_i + y_i)}{2}\right]^2 + \left[\frac{2y_i - (x_i + y_i)}{2}\right]^2$$

$$= \left(\frac{x_i - y_i}{2}\right)^2 + \left(\frac{y_i - x_i}{2}\right)^2$$

$$= \tfrac{1}{4}(x_i - y_i)^2 + \tfrac{1}{4}(y_i - x_i)^2$$

$$= \tfrac{1}{2}(x_i - y_i)^2$$

since $(x_i - y_i)^2 = (y_i - x_i)^2$

This sum of squares has $2 - 1 = 1$ degree of freedom. We add n of these together to get

$$\Sigma \tfrac{1}{2}(x_i - y_i)^2$$

and this has n degrees of freedom so we divide by n to get the estimate of variance. The square root gives us the standard deviation.

Exercise 15M

(Each branch is either true or false.)

1. The specificity of a test for a disease:

(a) has a standard error derived from the Binomial Distribution;

(b) measures how well the test detects cases of the disease;

(c) measures how well the test excludes subjects without the disease;

(d) measures how often a correct diagnosis is obtained from the test;

(e) is all we need to tell us how good the test is.

2. The repeatability or precision of measurements may itself be measured by:

(a) the coefficient of variation of repeated measurements;

(b) the correlation coefficient between pairs of measurements;

(c) the standard deviation of the difference between pairs of measurements;

(d) the standard deviation of repeated measurements;

(e) the difference between the means of two sets of measurements on the same set of subjects.

3. If the normal or reference range for haematocrit in men is 43.2–49.2:

(a) any man with haematocrit outside these limits is abnormal;

(b) haematocrits outside these limits are proof of disease;

(c) a man with a haematocrit of 46 must be very healthy;

(d) a woman with a haematocrit of 48 has a haematocrit within normal limits;

(e) a man with a haematocrit of 42 may be ill.

4. When a survival curve is calculated from censored survival times:

(a) the estimated proportion surviving becomes less reliable as survival time increases;

(b) individuals withdrawn during the first time interval are excluded from the analysis;

(c) survival estimates depend on the assumption that survival rates remain constant over the study period;

(d) it may be that the survival curve will not reach zero survival;

(e) the five-year survival rate can be calculated even if many of the subjects were identified less than five years ago.

5. Terminal digits in measurements which are likely to occur more often than expected include:

(a) 0;

(b) 1;

(c) 2;

(d) 5;

(e) 9.

Exercise 15E

In this exercise we shall estimate a reference range. Mather *et al.* (1979) measured plasma magnesium in 140 apparently healthy people, to compare with a sample of diabetics. The normal sample was chosen from blood donors and people attending day centres for the elderly in the area of St George's Hospital, to give 10 male and 10 female subjects in each age decade from 15–24 to 75 years and over. Questionnaires were used to exclude any subject with persistent diarrhoea, excessive alcohol intake or who were on regular

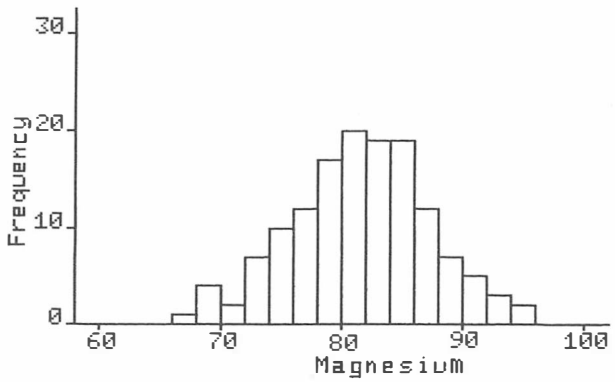


Fig. 15.E.1 Distribution of plasma magnesium in 140 apparently healthy people.

drug therapy other than hypnotics and mild analgesics in the elderly. The distribution of plasma magnesium is shown in Fig. 15E.1. The mean was 0.810 mmol/litre and the standard deviation 0.057 mmol/litre.

1. What do you think of the sampling method? Why use blood donors and elderly people attending day centres?

2. Why were some potential subjects excluded? Was this a good idea? Why were certain drugs allowed for the elderly?

3. Does plasma magnesium appear to follow a Normal Distribution?

4. What is the reference range for plasma magnesium, using the Normal Distribution method?

5. Find confidence intervals for the reference limits.

6. Would it matter if mean plasma magnesium in normal people increased with age? What method might be used to improve the estimate of the reference range in this case?

16. Mortality statistics and the structure of human populations

16.1. Mortality rates

One of our principal sources of information about the changing pattern of disease within a country and the differences in disease between countries is mortality statistics. In most developed countries, any death must be certified by a doctor, who records the cause, date and place of death and some data about the deceased. In Britain, these include the date of birth, area of residence and last known occupation. These death certificates form the raw material from which mortality statistics are compiled by a national bureau of censuses, in Britain the Office of Population Censuses and Surveys. The numbers of deaths can be tabulated by cause, by sex, by age, by types of occupation, by area of residence and by marital status. Table 5.1 shows one such tabulation, of deaths by cause and sex.

For purposes of comparison we must relate the number of deaths to the number in the population in which they occur. We have this information fairly reliably at ten-year intervals, from the decennial census of the country. We can estimate the size and age and sex structure of the population between censuses from registration of births and deaths. Each birth or death which takes place is notified to an official registrar, and so we can keep some track of changes in the population. There are other, less well-documented changes taking place, such as immigration and emigration, which mean that population size estimates between the census years are only approximations. Some estimates, such as the numbers in different occupations, are so unreliable that mortality data are only tabulated by them for census years.

If we take the number of deaths over a given period of time and divide it by the number in the population and the time period, we get a mortality rate, the number of deaths per unit time per person. We usually take the number of deaths over one calendar year, although when the number of deaths is small we may take deaths over several years, to increase the precision of the numerator. The number in the population is changing continually, and we take as the denominator the estimated population at the mid-point of the

time period. Mortality rates are usually very small numbers, so we usually multiply them by a constant, such as 1000 or 100 000, to avoid strings of zeros after the decimal point.

When we are dealing with deaths in the whole population, irrespective of age, the rate we obtain is called the *crude mortality rate* or *crude death rate.* The terms 'death rate' and 'mortality rate' are used interchangeably. We calculate the crude mortality rate for a population as:

$$\frac{\text{deaths occurring over given period}}{\text{number in population at mid-point of period} \times \text{length of period}} \times 1000$$

If the period is in years, this gives the crude mortality rate as deaths per 1000 population per year.

The crude mortality rate is so called because no allowance is made for the age distribution of the population, or for comparisons between populations with different age structures. For example, in 1901 the crude mortality rate among adult males (aged over 15 years) in England and Wales was 15.7 per 1000 per year, and in 1971 it was 15.5 per 1000 per year. It seems strange that with all the improvements in medicine, housing and nutrition between these times there has been so little improvement in the crude mortality rate. To see why we must look at the *age-specific mortality rates*, the mortality rates within narrow age groups. Age-specific mortality rates are usually calculated for one-, five-, or ten-year age groups. In 1901 the age-specific mortality rate for men aged 15–19 was 3.5 deaths per 1000 per year, whereas in 1971 it was only 0.9. As Table 16.1 shows, the age-specific mortality rate in 1901 was greater than that in 1971 for every age group. However, in 1901 there was a much greater proportion of the population in the younger age groups, where

Table 16.1. Age-specific mortality rates and age distribution in adult males, England and Wales, 1901 and 1971

Age group (years)	Age-specific death rate 1000 per year		% adult population in age group	
	1901	1971	1901	1971
15–19	3.5	0.90	15.36	9.61
20–24	4.7	0.95	14.07	10.62
25–34	6.2	0.99	23.76	17.45
35–44	10.6	2.32	18.46	16.16
45–54	18.0	7.09	13.34	16.63
55–64	33.5	20.20	8.68	15.48
65–74	67.8	50.80	4.57	9.90
75–84	139.8	114.20	1.58	3.52
85–	276.5	234.60	0.17	0.62

mortality was low, than there was in 1971. Correspondingly, there was a smaller proportion of the 1901 population than the 1971 population in the higher mortality older age groups. Although mortality was lower at any given age in 1971, the greater proportion of older people meant that there were as many deaths as in 1901.

16.2. Age-standardized mortality rates using the direct method

To eliminate the effects of different age structures in the populations which we want to compare, we can look at the age-specific death rates. But if we are comparing several populations, this is rather a cumbersome procedure, and it is often more convenient to calculate a single summary figure from the age-specific rates. There are many ways of doing this, of which three are frequently used: the direct and indirect methods of age standardization, and the life table.

We shall describe the direct method first. We use a standard population structure, i.e. a standard set of proportions of people in each age group. We then calculate the overall mortality rate which a population with the standard age structure would have if it experienced the age specific mortality rates of the observed population, the population whose mortality rate is to be adjusted. We shall take the 1971 population as the standard and calculate the mortality rate the 1901 population would have experienced if it had the 1971 age distribution. We do this by multiplying each 1901 age specific mortality rate by the proportion in that age group in the standard population, and adding. This then gives us an average mortality rate for the whole population, the *age-standardized mortality rate.* For example, the 1901 mortality rate in age

Table 16.2. Calculation of the age-standardized mortality rate by the direct method

Age group	Observed mortality rate per 1000 (a)	Standard proportion in group (b)	(a) × (b)
15–19	3.5	0.0961	0.336
20–24	4.7	0.1062	0.499
25–34	6.2	0.1745	1.082
35–44	10.6	0.1616	1.713
45–54	18.0	0.1663	2.993
55–64	33.5	0.1548	5.186
65–74	67.8	0.0990	6.712
75–84	139.8	0.0352	4.921
85–	276.5	0.0062	1.714
Sum			25.157

group 15–19 was 3.5 per 1000 per year and the proportion in the standard population in this age group is 9.61 per cent or 0.0961. The contribution of this age group is $3.5 \times 0.0961 = 0.336$. The calculation is set out in Table 16.2.

If we used the population's own proportions in each age group in this calculation we would get the crude mortality rate. Since 1971 has been chosen as the standard population, its crude mortality rate of 15.5 is also the age-standardized mortality rate. The age-standardized mortality rate for 1901 was 25.2 per 1000 men per year. We can see that there was a much higher age-standardized mortality in 1901 than 1971, reflecting the difference in age specific mortality rates.

16.3. Standardized mortality ratios by the indirect method

The direct method relies upon age-specific mortality rates for the observed population. If we have very few deaths, these age-specific rates will be very poorly estimated. This will be particularly so in the younger age groups, where we may even have no deaths at all. Such situations arise when considering mortality due to particular conditions or in relatively small groups, such as those defined by occupation. The indirect method of standardization is used for such data. We calculate the number of deaths we would expect in the observed population if it experienced the age-specific mortality rates of a standard population. We then compare the expected number of deaths with that actually observed.

We shall take as an example the deaths due to cirrhosis of the liver among male qualified medical practitioners recorded around the 1971 census. There were 14 deaths among 43 570 doctors aged below 65, a crude mortality rate of $14/43\,570 = 321$ per million, compared to 1423 out of 15 247 980 adult males (aged 15–64), or 93 per million. The mortality among doctors appears high, but the medical population may be older than the population of men as a whole, as it will contain relatively few below the age of 25. Also the actual number of deaths among doctors is small and any difference not explained by the age effect may be due to chance. The indirect method enables us to test this.

Table 16.3 shows the age-specific mortality rates for cirrhosis of the liver among all men aged 15–65, and the number of men estimated in each ten-year age group, for all men and for doctors. We can see that the two age distributions do appear to be different.

The calculation of the expected number of deaths is similar to the direct method, but different populations and rates are used. For each age group, we take the number in the observed population, and multiply it by the standard age-specific mortality rate, which would be the probability of dying if the mortality in the observed population were the same as that in the standard

Table 16.3. Age-specific mortality rates due to cirrhosis of the liver and age distributions of all men and medical practitioners, England and Wales, 1971

Age group (years)	Mortality per million men per year	Number of men	Number of doctors
15–24	5.859	3 584 320	1080
25–34	13.050	3 065 100	12 860
35–44	46.937	2 876 170	11 510
45–54	161.503	2 965 880	10 330
55–64	271.358	2 756 510	7790

population. This gives us the number we would expect to die in this age group in the observed population. We add these over the age groups and obtain the expected number of deaths. The calculation is set out in Table 16.4.

The expected number of deaths is 4.4965, which is considerably less than the 14 observed. We usually express the result of the calculation as the ratio of observed to expected deaths, called the *standardized mortality ratio* or SMR. Thus the SMR for cirrhosis among doctors is

$$\text{SMR} = \frac{14}{4.4965} = 3.11$$

We usually multiply by 100 to get rid of the decimal point. We say the SMR with all men = 100 is 311.

If we do not adjust for age at all, the ratio of the crude death rates is 3.44, compared to the age-adjusted figure of 3.11, so the adjustment has made some, but not much, difference.

We can calculate a confidence interval for the SMR quite easily. Denote the observed deaths by O and expected by E. It is reasonable to suppose that the deaths are independent of one another and happening randomly in time, so

Table 16.4. Calculation of the expected number of deaths due to cirrhosis of the liver among practitioners, using the indirect method

Age group	Standard mortality rate, all men (a)	Observed population: Number of doctors (b)	(a) × (b)
15–24	0.000005859	1080	0.0063
25–34	0.000013050	12 860	0.1678
35–44	0.000046937	11 510	0.5402
45–54	0.000161503	10 330	1.6683
55–64	0.000271358	7790	2.1139
Total			4.4965

the observed number of deaths is from a Poisson Distribution (Section 6.7). The standard deviation of this Poisson Distribution is the square root of its mean and so can be estimated by the square root of the observed deaths, $\sqrt{O}$. The expected number is calculated from a very much larger sample and is so well estimated it can be treated as a constant, so the standard deviation of $100 \times O/E$, which is the standard error of the SMR, is estimated by $100 \times \sqrt{O}/E$.

Provided the number of deaths is large enough, say more than 10, an approximate 95 per cent confidence interval is given by

$$100 \times \frac{O}{E} \pm 1.96 \times 100 \times \frac{\sqrt{O}}{E}$$

For small observed frequencies, tables based on the exact probabilities of the Poisson Distribution are available (Pearson and Hartley 1970). For the cirrhosis data the formula gives

$$311 - 1.96 \times 100 \times \frac{\sqrt{14}}{4.4965} \text{ to } 311 + 1.96 \times 100 \times \frac{\sqrt{14}}{4.4965}$$
$$= 311 - 163 \text{ to } 311 + 163$$
$$= 148 \text{ to } 474$$

The confidence interval clearly excludes 100 and the high mortality cannot be ascribed to chance.

The news is not all bad for medical practitioners, however. Their SMR for cancer of the trachea, bronchus and lung is only 32. Doctors may drink, but they don't smoke!

16.4. Demographic life tables

We have already discussed a use of the life table technique for the analysis of clinical survival data (Section 15.6). The life table was found by following the survival of a group of subjects from some starting point to death. In *demography*, which means the study of human populations, life tables are generated in a different way. Rather than charting the progress of a group from birth to death, we start with the present age-specific mortality rates. We then calculate what would happen to a cohort of people from birth if these age-specific mortality rates applied unchanged throughout their lives. We denote the probability of dying between ages x and $(x + 1)$ years (which is the age-specific mortality rate) by q_x. As in Table 15.6, the probability of surviving from age x to $(x + 1)$ is $p_x = 1 - q_x$. We now suppose that we have a cohort of size l_0 at age 0, i.e. at birth. The size of l_0 is usually 100 000 or 10 000. The number who would still be alive after x years is l_x. We can see that the number alive after $(x + 1)$ years is $l_{x+1} = p_x \times l_x$, so given all the p_x from $(x = 0)$ onwards we can calculate the l_x. The cumulative survival probability

Table 16.5. Extract from English Life Table Number 11, 1950–52, males

Age (in years) x	Expected number alive at age x l_x	Probability an individual dies between ages x and $x + 1$ q_x	Expected life at age x (years)' e_x
0	100 000	0.03266	66.42
1	96 734	0.00241	67.66
2	96 501	0.00141	66.82
3	96 395	0.00102	65.91
4	96 267	0.00084	64.98
.	.	.	.
.	.	.	.
.	.	.	.
100	23	0.44045	1.67
101	13	0.45072	1.62
102	7	0.46011	1.58
103	4	0.46864	1.53
104	2	0.47636	1.50

to age x is then $P_x = l_x/l_0$. We have already used this in Exercise 6E.

Table 16.5 shows an extract from Life Table Number 11, 1950–52, for England and Wales. With the exception of 1941, a life table like this has been produced every ten years since 1871, based on the decennial census year. The life table is based on the census year because only then do we have a good

Table 16.6. Abridged Life Table 1969–71, England and Wales

Age	Males		Females	
x	l_x	e_x	l_x	e_x
0	10 000	68.8	10 000	75.1
5	9766	65.4	9819	71.4
10	9746	60.5	9806	66.5
15	9728	55.6	9795	61.6
20	9683	50.9	9776	56.7
25	9638	46.1	9755	51.8
30	9595	41.3	9731	47.0
35	9542	36.5	9696	42.1
40	9467	31.8	9639	37.4
45	9327	27.2	9538	32.7
50	9079	22.9	9372	28.3
55	8673	18.9	9127	24.0
60	8016	15.2	8768	19.8
65	7012	12.0	8227	16.0
70	5625	9.4	7403	12.5
75	3982	7.2	6191	9.4
80	2355	5.5	4544	6.9
85	1072	4.0	2696	5.0

measure of the number of people at each age, the denominator in the calculation of q_x. A three-year period is used to increase the number of deaths for a year of age and so improve the estimation of q_x. Separate tables are produced for males and females because the mortality of the two sexes is very different. Age-specific deaths rates are higher in males than females at every age. Between census years life tables are still produced but are only published in an abridged form, giving l_x at five-year intervals (Table 16.6).

The final column in Tables 16.5 and 16.6 is the *expected life*, *expection of life* or *life expectancy*, e_x. This is the average length of life still to be lived by those reaching age x. We have already calculated this as the expected value of the probability distribution of year of death (Exercise 6E). We can do the calculation in a number of other ways. For example, if we add l_{x+1}, l_{x+2}, l_{x+3}, etc. we shall get the total number of years to be lived, because the l_{x+1} who survive to $(x + 1)$ will have added l_{x+1} years to the total, the l_{x+2} of these who survive from $(x + 1)$ to $(x + 2)$ will add a further l_{x+2} years, and so on. If we divide this sum by l_x we get the average number of whole years to be lived. If we then remember that people do not die only on their birthdays, but scattered throughout the year, we can add half to allow for the average of half a year lived in the year of death. We thus get

$$e_x = \frac{1}{l_x} \sum_{i=x+1} l_i + \frac{1}{2}$$

If many people die in early life, with high age-specific death rates for children, this has a great effect on expectation of life at birth. In 1971, for example, expectation of life at birth for males was 69 years, compared to only 41 years in 1871, an improvement of 28 years. However, expectation of life at age 45 in 1971 was 27 years compared to 22 years in 1871, an improvement of only 5 years. At age 65, male expectation of life was 12 years in 1971 and 11 years in 1871, an even smaller change. Hence the change in life expectancy at birth was due to changes in mortality in early life, not late life.

Life tables have a number of uses, both medical and non-medical. Expectation of life provides a useful summary of mortality without the need for a standard population. The table enables us to predict the future size of and age structure of a population given its present state; this prediction is called a *population projection*. This can be very useful in predicting such things as the future requirement for geriatric beds in a health district. Life tables are also invaluable in non-medical applications, such as the calculation of insurance premiums, pensions and annuities.

The main difficulty with prediction from a life table is finding a table which applies to the populations under consideration. For the general population of, say, a health district, the national life table will usually be adequate, but for special populations this may not be the case. If we want to predict the future need for care of an institutionalized population, such as in a long-stay

psychiatric hospital or old peoples' home, the mortality may be considerably greater than that in the general population. Predictions based on the national life table can only be taken as a very rough guide. If possible, life tables calculated on that type of population should be used.

16.5. Vital statistics

We have seen a number of occasions where ordinary words have been given quite different meanings in statistics from those they have in common speech; 'normal' and 'significant' are good examples. 'Vital statistics' is the opposite, a technical term which has acquired a completely unrelated popular meaning. As far as the medical statistician is concerned, vital statistics have nothing to do with the dimensions of female bodies. They are the statistics relating to life and death: birth rates, fertility rates, marriage rates and death rates. We have already dealt with the crude mortality rate, age-specific mortality rates, age-standardized mortality rates, standardized mortality ratio, and expectation of life. In this section we shall define a number of other statistics which are often quoted in the medical literature.

The *infant mortality rate* is the number of deaths under one year of age divided by the number of live births, usually expressed as deaths per 1000 live births. The *neonatal mortality rate* is the same thing for deaths in the first 4 weeks of life.

The *stillbirth rate* is the number of stillbirths divided by the total number of births, live and still. A stillbirth is a child born dead after 28 weeks' gestation. The *perinatal mortality rate* is the number of stillbirths and deaths in the first week of life divided by the total births, again usually presented per 1000 births. Infant and perinatal mortality rates are regarded as particularly sensitive indicators of the health status of the population. The *maternal mortality rate* is the number of deaths of mothers ascribed to problems of pregnancy and birth, divided by the total number of births.

The *attack rate* for a disease is the proportion of people exposed to infection who develop the disease. The *case fatality rate* is the proportion of cases who die. The *prevalence* of a disease is the proportion of people who have it at one point in time. The *incidence* is the number of new cases in one year divided by the number at risk.

The *birth rate* is the number of live births per year divided by the total population. The *fertility rate* is the number of live births per year divided by the number of women of childbearing age, taken as 15–44 years.

16.6. The population pyramid

The age distribution of a population can be presented as histogram, using the methods of Section 4.3. However, because the mortality of males and

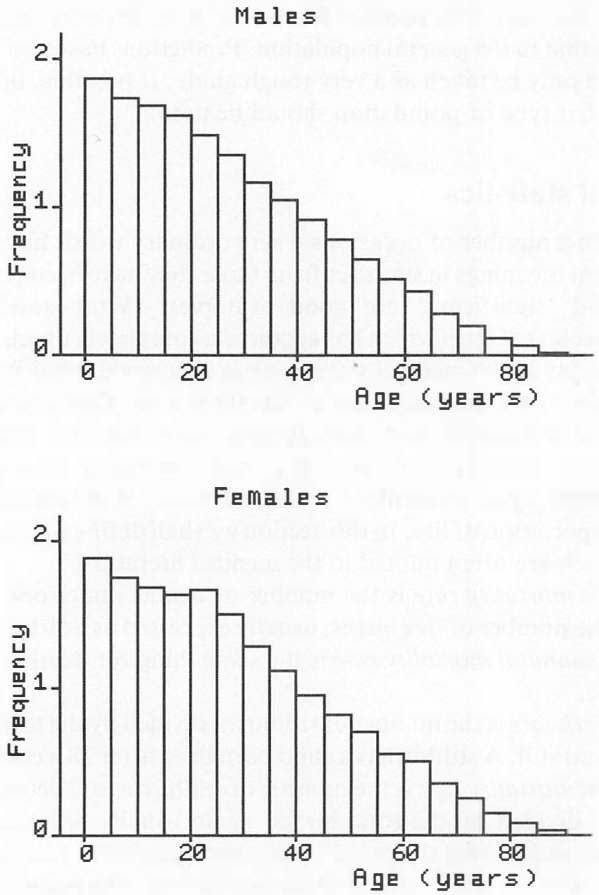


Fig. 16.1 Age distributions for the population of England and Wales, by sex, 1901.

females is so different the age distributions for males and females are also different. It is usual to present the age distributions for the two sexes separately. Figure 16.1 shows the age distributions for the male and female populations of England and Wales in 1901. Now, these histograms have the same horizontal scale. The conventional way to display them is with the age scale vertically and the frequency scale horizontally as in Fig. 16.2. The frequency scale has zero in the middle and increases to the right for females and to the left for males. This is called a *population pyramid*, from the shape.

Figure 16.3 shows the population pyramid for England and Wales in 1971. The shape is quite different. Instead of a triangle we have an irregular figure with almost vertical sides which begin to bend very sharply inwards at about

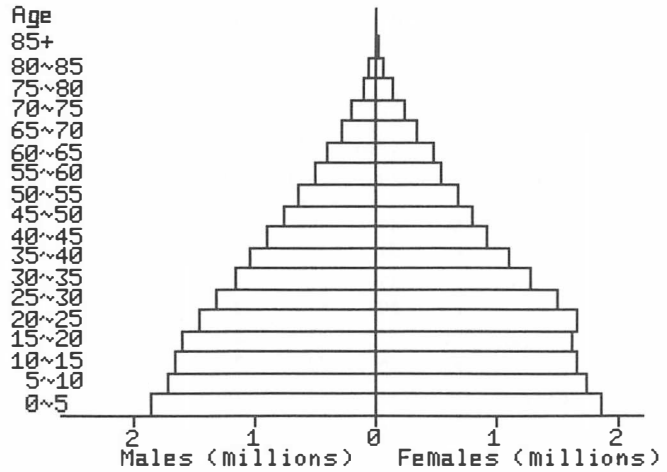


Fig. 16.2 Population pyramid for England and Wales, 1901.

age 65. A major change in population structure has taken place, with a vast increase in the proportion of elderly. This has major implications for medicine, as the care of the elderly has become a large proportion of the work of doctors, nurses and their colleagues. It is interesting to see how this has come about.

It is popularly supposed that people are now living much longer as a result of modern medicine, which prevents deaths in middle life. This is only partly

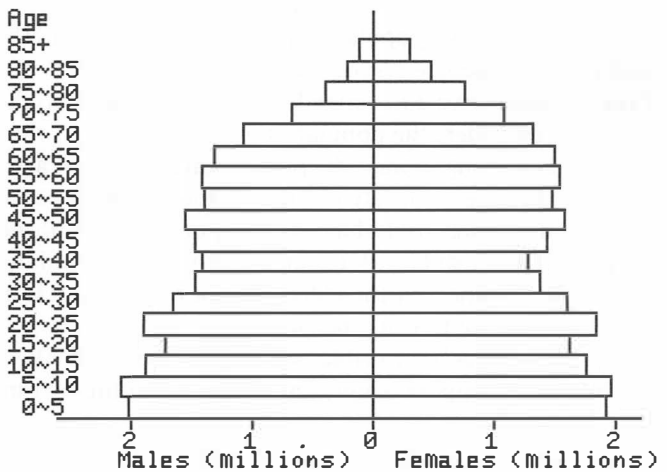


Fig. 16.3 Population pyramid for England and Wales, 1971.

Table 16.7. Life expectancy in 1901 and 1971, England and Wales

Age	Sex	Expectation of life in years		Increase
		1901	1971	1901 to 1971
Birth	M	49	69	20
	F	52	75	23
15 years	M	47	56	9
	F	50	62	12
45 years	M	23	27	4
	F	26	33	7
65 years	M	11	12	1
	F	12	16	5

true. Table 16.7 shows the life expectancy at different ages in 1901 and 1971. Life expectancy at birth has increased dramatically, but the increase in later life is much less. Thus we can see that the change is not an extension of every life by 20 years, which would be seen at every age, but mainly a reduction in mortality in childhood and early adulthood. Mortality in later life has changed by relatively little.

Now, a big reduction in mortality in childhood would result in an increase in the base part of the pyramid, as more children survived, unless there was a corresponding fall in the number of babies being born. In the nineteenth century, women were having many children and despite the high mortality in childhood the number who survived into adulthood to have children of their own exceeded that of their own parents. The population expanded and this history is embodied in the 1901 population pyramid. In the twentieth century, infant mortality fell and people responded to this by having fewer children. The base of the pyramid ceased to expand. As those who were in the base of the 1901 pyramid grew older, the population in the top half of the pyramid increased. The 0–4 age group in the 1901 pyramid are the 70–74 age group in the 1971 pyramid. Had the birth rate not fallen, the population would have continued to expand and we would have as great or greater a proportion of young people in 1971 as we did in 1901, and a vastly larger population. Thus the increase in the proportion of the elderly is not because adult lives have been extended, but because fertility has declined.

Most developed countries have stable populatiion pyramids like Fig. 16.3 and those of most developing countries have expanding pyramids like Fig. 16.2.

Exercise 16M

(Each branch is either true or false.)

1. Age-specific mortality rate:

(a) is a ratio of observed to expected deaths;
(b) can be used to compare mortality between different age groups;
(c) is an age-adjusted mortality rate;
(d) measures the number of deaths in a year;
(e) measures the age structure of the population.

2. Expectation of life:

(a) is the number of years most people live;
(b) is a way of summarizing age-specific death rates;
(c) is the expected value of a particular probability distribution;
(d) varies with age;
(e) is derived from life tables.

3. In 1971, the SMR for cirrhosis of the liver for men was 773 for publicans and inn-keepers and 25 for window cleaners, both being significantly different from 100 (Donnan and Haskey 1978). We can conclude that:

(a) publicans are more than seven times as likely as the average person to die from cirrhosis of the liver.
(b) the high SMR for publicans may be because they tend to be found in the older age groups.
(c) being a publican causes cirrhosis of the liver.
(d) window cleaning protects men from cirrhosis of the liver.
(e) window cleaners are at high risk of cirrhosis of the liver.

4. The age and sex structure of a population may be described by:

(a) a life table;
(b) a correlation coefficient;
(c) a standardized mortality ratio;
(d) a population pyramid;
(e) a bar chart.

5. The following statistics are adjusted to allow for the age distribution of the population:

(a) age-standardized mortality rate;

(b) fertility rate;

(c) perinatal mortality rate;

(d) crude mortality rate;

(e) expectation of life at birth.

Exercise 16E

Anderson *et al.* (1985) studied mortality associated with volatile substance abuse (VSA), often called glue sniffing. In this study all known deaths associated with VSA from 1971 to 1983 inclusive were collected, using sources including three press-cuttings agencies and a six-monthly systematic survey of all coroners. Cases were also notifed by the Office of Population Censuses and Surveys for England and Wales and by the Crown Office and procurators fiscal in Scotland.

Table 16E.1 shows the age distribution of these deaths for Great Britain and for Scotland alone, with the corresponding age distributions at the 1981 decennial census.

Table 16E.1. Volatile substance abuse mortality and population size, Great Britain and Scotland, 1971–83 (Anderson *et al.* 1985)

	Great Britain		Scotland	
Age group (years)	VSA deaths	Population (thousands)	VSA deaths	Population (thousands)
0–9	0	6770	0	653
10–14	44	4271	13	425
15–19	150	4467	29	447
20–24	45	3959	9	394
25–29	15	3616	0	342
30–39	8	7408	0	659
40–49	2	6055	0	574
50–59	7	6242	0	579
60 +	4	10 769	0	962

1. Calculate age-specific mortality rates for VSA per year and for the whole period. What is unusual about these age-specific mortality rates?

2. Calculate the SMR for VSA deaths for Scotland.

3. Calculate the 95 per cent confidence interval for this SMR.

4. Does the number of deaths in Scotland appear particularly high? Apart from a lot of glue sniffing, are there any other factors which should be considered as possible explanations for this finding?

17. Solutions to exercises

Exercise 2M

T means 'true', and F means 'false', throughout the answers to the multiple-choice questions.

1. (a) F, it is done for the comparability of the groups, (2.2). (b) F. (c) F. (d) T. (e) F, (2.2).

2. (a) T, (2.8). (b) F, refers to cross-over trial, (2.5) (c) F, patients do not know their treatment. They usually do know that they are in a trial. (d) F, cross-over trial, (2.5). (e) T, (2.8).

3. (a) F, must be true to randomization, vaccinated and refusing children are self-selected, (2.4). (b) F. (c) F, control group not offered vaccination. (d) F. (e) F, we can compare effect of a vaccination programme by comparing whole vaccination group, vaccinated and refusers to the controls.

4. (a) T, (2.5). (b) F, order is randomized, (2.5). (c) T, (2.5). (d) T, (2.5). (e) T, (2.5).

5. (a) F, the purpose of placebos is make dissimilar treatments appear similar (2.7). (b) F, only in randomized trials can we rely on comparability, and then only within the limits of random variation, (2.2). (c) T, (2.7). (d) T, (2.8). (e) T, (2.7).

Exercise 2E

1. Yes, it appears to. The death rate in the high-risk control group was 6.3 times greater than in low-risk children.

2. No. It is the allocation procedure itself that we are testing.

3. We cannot be sure. There may be factors which are more important in other areas than they are in the study town. For example, if the study town was overwhelmingly populated by one ethnic group, factors relevant to other

ethnic groups would be missed. In fact, when the scoring system was tried in an inner city area of London, it was not so effective as in this study.

4. No. The month of birth distribution is different. This may be important if the month of birth is related to mortality in the first year of life. As it happens, it is (Weatherall 1976) and August is the birth month with second highest mortality. This biases the comparison. There are two ways round this. At the allocation stage, children who cannot be equally allocated could be excluded. This would be the 16 allocated to control on holidays and 20 per cent of those born between 7 July and 14 September 1974.

Alternatively at the analysis stage comparison should be made within allocation periods. This is complicated, but we would find the differences in rates for the 50–50 and 60–40 periods separately and then combine them. The 16 'controls' born on holidays have no observation children to compare to them, and so contribute no useful information. They should still be excluded.

5. No. Apart from the effect of the deviations from randomization, which is probably small and which we shall ignore from now on, the observed group consists of volunteers. They have all agreed to the surveillance, but the control group have not had the opportunity to agree or refuse. They will therefore not be comparable (see Section 2.3).

6. The authors say that they did this to show that even the most extreme comparison did not achieve statistical significance. In other words, to show that the difference is not large enough to provide convincing evidence that surveillance lowers mortality. This is not a reasonable thing to do because we expect these groups to be different simply by the selection procedure, irrespective of whether the treatment has any effect.

7. The only comparison which is true to the allocation procedure is to compare the full observation group, both observed and refused, with the controls. The mortality rate in the observation group is

$$\frac{2 + 3}{627 + 210} \times 1000 = 6.0 \text{ per } 1000$$

compared to 9.8 for the controls.

8. The study shows that this method of identifying high-risk children works quite well. The study is based on too few deaths for any firm conclusions to be drawn about the effectiveness of surveillance, although the data suggest that it may reduce mortality. The fall in mortality in the study town is impressive, but as the national data show there is a downward trend in mortality rates anyway.

9. A definitive answer as to whether surveillance is effective could only be shown by a much larger study, perhaps covering several towns. This would be

expensive, but a national programme would be far more so. Health visitors represent resources which would have to be transferred from some other health service function.

Exercise 3M

1. (a) F, can be anything (3.3). (b) T, e.g. people of Britain (3.3). (c) T, e.g. all possible tosses of coins (3.3). (d) T, (3.3). (e) T, e.g. all possible tosses of a coin, or all patients as they would be if given a new treatment. The latter does not physically exist but it is a population in which we may be very interested.

2. (a) T, (3.3). (b) F, only tells us who is there on that day. (c) F, the hospital could be quite unusual. (d) F, only applies to current inpatients. Some diagnoses are less likely than others to lead to admission or to long stay. (e) T, (3.2).

3. (a) T, (3.4). (b) F, we must stick to the sample the random process produces (3.4). (c) F, they can be (3.4), using standard errors, confidence intervals and significance tests (Chapters 8 and 9). (d) T, (3.4). (e) F, it does not depend on the subject's characteristics at all, except for its being in the population.

4. (a) F, some populations are unidentifiable. (b) T, (3.4). (c) T, (3.4). (d) F, it can be very difficult. (e) T, (3.4).

5. (a) T, it is a random cluster sample (3.4). (b) T, (3.4). (c) T, each patient had the same chance of their hospital being chosen and then the same chance of being chosen within the hospital. This would not be so if we chose a fixed number from each hospital rather than a fixed proportion, as those in small hospitals would be more likely to be chosen than those in large hospitals. (d) T, it is a random sample. (e) F, what about a sample with patients in every hospital?

6. (a) F, we would not get enough cases of cancer of the oesophagus (3.7). (b) T, a case control study (3.7). (c) F, we would not get enough cases (3.7). (d) T, another form of case control study (3.7). (e) T, a cohort study (3.7).

Exercise 3E

1. Both control groups are drawn from populations which were easy to get to, one being hospital patients without gastro-intestinal symptoms, the other

being fracture patients and their relatives. Both are matched for age and sex. Mayberry *et al.* also matched for social class and marital status. Apart from the matching factors, we have no way of knowing whether cases and controls are comparable, or any way of knowing whether controls are representative of the general population. This is usual in case control studies and is a major problem with this design.

2. There are two obvious sources of bias: interviews were not blind and information is being recalled by the subject. The latter is particularly a problem for data about the past. In James' study subjects were asked what they used to eat several years in the past. For the cases this was before a definite event, onset of Crohn's disease, but for the controls it was not, the time being time of onset of the disease in the matched case.

3. The question in James' study was 'What did you use to eat in the past?' That in Mayberry *et al.* was 'what do you eat now?'

4. Of the 100 patients with Crohn's disease, 29 were current eaters of cornflakes. Of 29 cases who knew of the cornflakes association, 12 were ex-eaters of cornflakes, and among the other 71 cases 21 were ex-eaters of cornflakes, giving a total of 33 past but not present eaters of cornflakes. Combining these with the 29 current consumers, we get 62 cases who had at some time been regular eaters of cornflakes. If we carry out the same calculation for the controls, we obtain (3 + 10) = 13 past eaters and with 22 current eaters this gives 35 sometime regular cornflakes eaters. Cases were more likely than controls to have eaten cornflakes regularly at some time, the proportion of cases reporting having eaten cornflakes being almost twice as great as for controls. Compare this to James' data, where 17/68 = 25 per cent of controls and 23/34 = 68 per cent of cases, 2.7 times as many, had eaten cornflakes regularly. The results are similar.

5. The relationship between Crohn's disease and reported consumption of cornflakes had a much smaller probability for the significance test and hence stronger evidence that a relationship existed (see Chapter 9). Also, only one case had never eaten cornflakes (it was also the most popular cereal among controls).

6. Of the Crohn's cases, 67.6 per cent (i.e. 23/34) reported having eaten cornflakes regularly compared to 25.0 per cent of controls. Thus cases were 67.6/25.0 = 2.7 times as likely as controls to report having eaten cornflakes. The corresponding ratios for the other cereals are: wheat, 2.7; porridge, 1.5; rice, 1.6; bran, 6.1; muesli, 2.7. Cornflakes does not stand out when we look at the data in this way. The small probability simply arises because it is the most popular cereal.

7. We can conclude that there is no evidence that eating cornflakes is more closely related to Crohn's disease than is consumption of other cereals. The tendency for Crohn's case to report excessive eating of breakfast foods

before onset of the disease may be the result of greater variation in diet than in controls, as they try different foods in response to their symptoms. They may also be more likely to recall what they used to eat, being more aware of the effects of diet because of their disease.

Exercise 4M

1. (a) T, (4.1). (b) F, parity is quantitative and discrete (4.1). (c) F, continuous (4.1). (d) T, (4.1). (e) F, continuous (4.1).

2. (a) T, (4.1). (b) T, (4.1). (c) F, this is discrete (4.1). (d) T, this includes years and fraction of a year (4.1). (e) F, this is discrete (4.1).

3. (a) F, the mean is greater (4.6). (b) F, it could have more than one mode, we cannot say. (c) T, (4.4). (d) F, this depends only on whether the variance is greater than 1 (4.7). (e) T, because the median is less than the mean (4.5, 4.6),

4. (a) T, (4.5), (b) T, (4.3). (c) T, (4.3). (d) F, these only tell us the location and spread of the distribution (4.6, 4.7). (e) T, (4.2).

5. (a) T, (4.6). (b) F, it is 2. The observations must be ordered before the central one is found (4.5). (c) T, the most common observation is 2 which appears twice. (d) F, it is $7 - 1 = 6$ (4.7). (e) T, the deviations from the mean are $0, -2, 4, -1, -1$ so the sum of squares is $0 + 4 + 16 + 1 + 1 = 22$. These are $(n - 1) = 5 - 1 = 4$ degrees of freedom so variance $= 22/4 = 5.5$ (4.7).

Exercise 4E

1. Stem and leaf plot:

2	2 9
3	3 3 3 4 4 4 6 6 6 6 7 7 8 8 8 9
4	0 0 0 1 1 1 2 3 4 4 4 5 6 7 7 7 8 9 9
5	0 1
6	0

2. Minimum = 2.2, maximum = 6.0. The median is the average of the 20th and 21st ordered observations, since the number of observations is even. These are both 4.0, so the median is 4.0. The first quartile is between the 10th and 11th, which are both 3.6. The third quartile has $i = 0.75 \times 41$

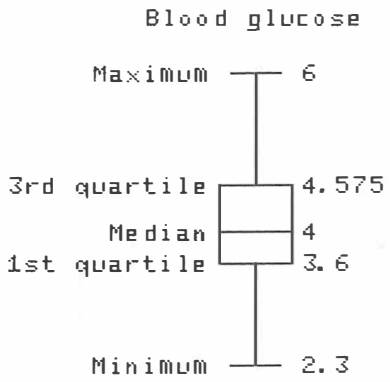


Fig. 17.1 Box and whisker plot of blood glucose.

= 30.75 and lies between the 30th and 31th observations, which are 4.5 and 4.6, giving $4.5 + 0.75 \times 0.1 = 4.575$. The box and whisker plot is shown in Fig. 17.1.

3. The frequency distribution is derived easily from the stem and leaf plot:

Interval	*Frequency*
2.0–2.4	1
2.5–2.9	1
3.0–3.4	6
3.5–3.9	10
4.0–4.4	11
4.5–4.9	8
5.0–5.4	2
5.5–5.9	0
6.0–6.4	1
Total	40

4. The histogram is shown in Fig. 17.2.
5. $\Sigma x = 162.2$; $\bar{x} = 162.2/40 = 4.055$
6. $\Sigma x^2 = 676.74$

$$\Sigma x^2 - \frac{(\Sigma x)^2}{n} = 676.74 - \frac{162.2^2}{40} = 19.019$$

7. There are $n - 1 = 40 - 1 = 39$ degrees of freedom.

$$\text{Variance, } s^2 = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$
$$= \frac{19.109}{39}$$
$$= 0.4876\ 67$$

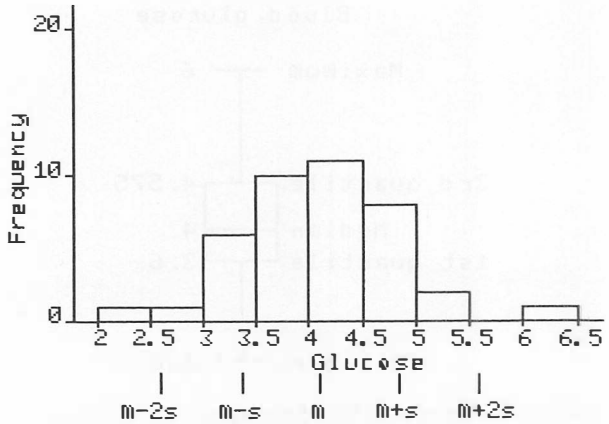


Fig. 17.2 Histogram of blood glucose.

8. $s = \sqrt{s^2} = \sqrt{0.487667} = 0.698$

$$\begin{aligned}
\bar{x} - 2s &= 4.055 - 2 \times 0.698 = 2.654 \\
\bar{x} - s &= 4.055 - 0.698 = 3.357 \\
\bar{x} &= 4.055 \\
\bar{x} + s &= 4.055 + 0.698 = 4.753 \\
\bar{x} + 2s &= 4.055 + 2 \times 0.698 = 5.451
\end{aligned}$$

9. Fig. 17.2 also shows the mean and standard deviation marked on the glucose scale. The majority of points fall within one standard deviation of the mean and nearly all within two standard deviations of the mean.

Exercise 5M

1. (a) F, we have no idea how many would get better anyway. (b) T, (5.1). (c) T, (5.2). (d) T, (2.1). (e) T, 66.67 per cent is 2/3. We may only have 3 patients.

2. (a) T, (5.2). (b) F, it should be 1730. We round up because of the 9. (c) F, this is given to six significant figures. To six decimal places it is 1729.543710. (d) T, we round up because of the 7. (e) T, (5.2).

3. (a) F, it is a bar chart (5.5). A histogram shows a frequency for a single variable, this shows the relationship between two variables. (b) T, (5.5), see Fig. 17.3. (c) T, (5.5), see Fig. 17.3. (d) F, the time has no true zero to show. (e) T, (5.5), see Fig. 17.3.

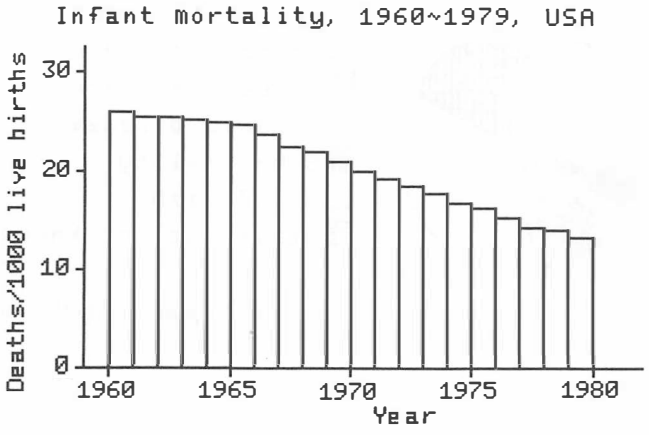


Fig. 17.3 A dubious graph revised.

4. (a) T, (5.8). (b) T, (5.8). (c) F, (5.8). (d) F, there is no logarithm of zero and a log scale cannot do this. (e) T, (5.8, 5A).

5. (a) F, shows the distribution of a single variable. (b) F, shows the distribution of a single categorical variable. (c) T, (5.6). (d) T, (5.5). (e) T, (5.7).

Exercise 5E

1. This is the frequency distribution of a qualitative variable, so a pie chart can be used to display it. The calculations are as follows:

Category	*Frequency*	*Relative frequency*	*Angle*
schizophrenia	474	0.323 11	116
affective illness	277	0.188 82	68
organic brain syndrome	405	0.276 07	99
subnormality	58	0.039 54	14
alcoholism	57	0.038 85	14
other	196	0.133 61	48
Total	1467	1.000 00	359

Notice that we have lost one degree through rounding errors. We could work to fractions of a degree, but the eye is unlikely to spot the difference. The pie chart is shown in Fig. 17.4.

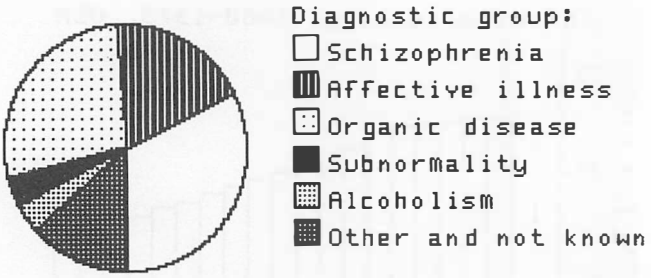


Fig. 17.4 Pie chart showing the distribution of patients in Tooting Bec Hospital by diagnostic group.

2. See Fig. 17.5.

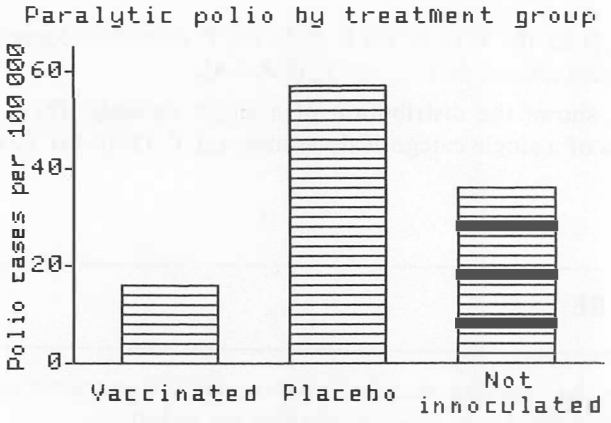


Fig. 17.5 Bar chart showing the results of the Salk vaccine trial.

3. There are several possibilities. In the original paper, Doll and Hill used a separate bar chart for each disease, as shown in Fig. 17.6.

4. This is a frequency distribution of a quantitative variable, so a histogram is appropriate. See Fig. 17.7.

5. Line graphs can be used here, as we have simple time series (Fig. 17.8). For an explanation of the difference between years, see Exercise 13E.

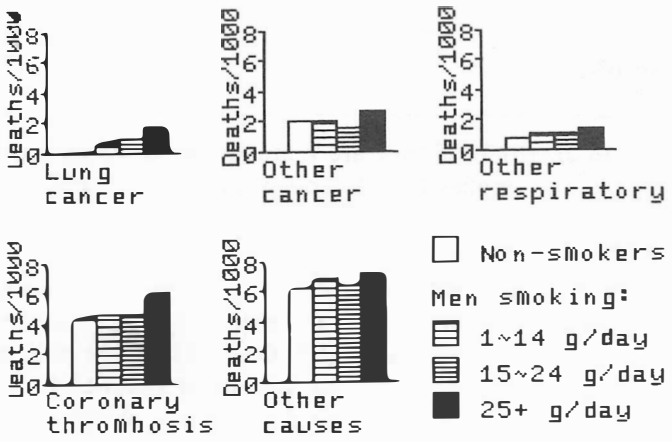


Fig. 17.6 Mortality in British doctors by smoking habits after Doll and Hill (1956).

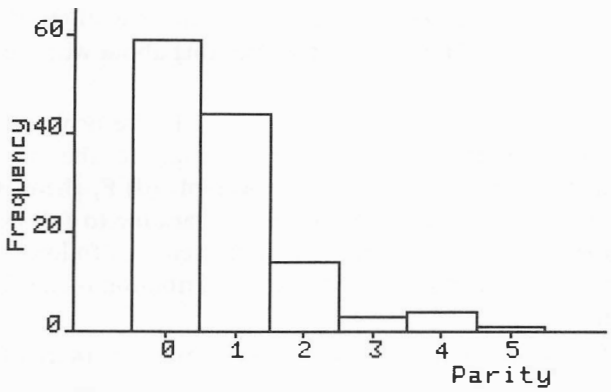


Fig. 17.7 Histogram showing parity of women attending antenatal clinics at St George's Hospital.

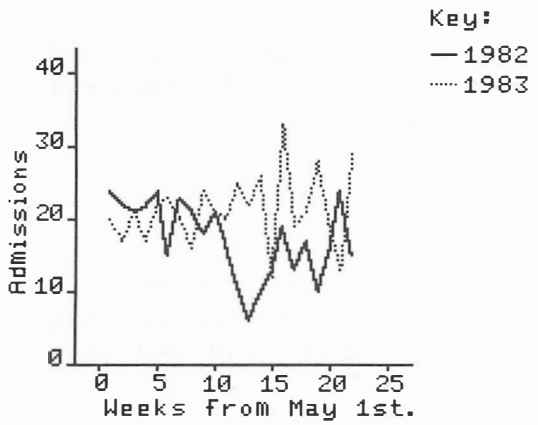


Fig. 17.8 Line graphs for geriatric admissions in Wandsworth in the summers of 1982 and 1983.

Exercise 6M

1. (a) T, (6.2). (b) T, if they are mutually exclusive they cannot both happen. (c) F, this applies to independent events (6.2). (d) F, there is no reason why this should be so. (e) F, only true if these are exhaustive, the only events which can happen (6.3).

2. (a) T, $0.2 \times 0.05 = 0.01$ (6.2). (b) F, the probabilities are multiplied. Clearly the probability of both must be less than that for each one. (c) T, a difficult question. The probability of both is 0.01, so the probability of X alone is $0.20 - 0.01 = 0.19$ and the probability of Y alone is $0.05 - 0.01 = 0.04$. The probability of having X or Y is the probability of X alone + probability of Y alone + probability of X and Y together, because these are three mutually exclusive events. Having X and having Y are not mutually exclusive as she can have both. There are several ways of arriving at this result. (d) F, if she has X the probability of having Y is still 0.05, because X and Y are independent. Having X tells us nothing about whether she has Y. (e) F, see (d).

3. (a) T, (6.4). (b) F, this is continuous. (c) T, the probability of each random choice producing a responder is equal to the proportion of responders in the population, which is constant. (d) F, there is no set of independent trials here. We might expect the variable to follow a Poisson Distribution (6.7). (e) F, the number of hypertensives follows a Binomial Distribution, not the proportion, though its distribution is closely related to the Binomial.

4. (a) F, it is one (6.6). (b) T, independent (6.2). (c) T, (6.4). (d) F, at least one tail means one tail (0.5) or two tails (0.25). These are mutually exclusive, so the probability of at least one tail is $0.5 + 0.25 = 0.75$ (e) T, (6.3).

5. (a) F, should be $\mu + 2$ (6.6). (b) T, (6.6). (c) T, (6.6). (d) F, should be $4\sigma^2$, (6.6). (e) T, (6.6).

6. (a) T, (6.6). (b) T, (6.6). (c) T, (6.6). (d) F, the variance of a difference is the sum of the variances (6.6). (e) F, variances cannot be negative. $Var(-X) = (-1)^2 \times Var(X) = Var(X)$.

Exercise 6E

1. Probability of survival to age 10. This illustrates the frequency definition of probability. The number out of 1000 surviving is 959, so the probability is $959/1000 = 0.959$.

2. Survival and death are mutually exclusive, exhaustive events. So

$$\text{Prob(survives)} + \text{Prob(dies)} = 1$$
$$\text{Prob(dies)} = 1 - 0.959 = 0.041$$

3. As in 1, these are just number surviving/1000.

Survive to age	*Probability*
10	0.959
20	0.952
30	0.938
40	0.920
50	0.876
60	0.758
70	0.524
80	0.211
90	0.022
100	0.000

The events are not mutually exclusive, e.g. a man cannot survive to age 20 if he does not survive to age 10. This does not form a probability distribution.

4. Prob(aged 60 survives to 70) = number alive at 70 divided by number alive at 60

$$= \frac{524}{758} = 0.691$$

5. Independent events. Prob(survival 60 to 70) = 0.691, probability both survive = $0.691 \times 0.691 = 0.478$.

6. The proportion surviving on average is the probability of survival = 0.691. So a proportion of 0.691 of the 100 survive. We expect $0.691 \times 100 = 69.1$ to survive.

7. Prob(death in 2nd)

$$= \text{Prob(survives to 2nd)} - \text{Prob(survives to 3rd)}$$
$$= 0.959 - 0.952$$
$$= 0.007$$

8. As in 7, we find

Decade	*Probability of dying*
1st	0.041
2nd	0.007
3rd	0.014
4th	0.018
5th	0.044
6th	0.118

Decade	*Probability of dying*
7th	0.234
8th	0.313
9th	0.189
10th	0.022

This is a set of mutually exclusive events and they are exhaustive — there is no other decade in which death can take place. The sum of the probabilities is therefore 1.0. The distribution is shown in Fig. 17.9.

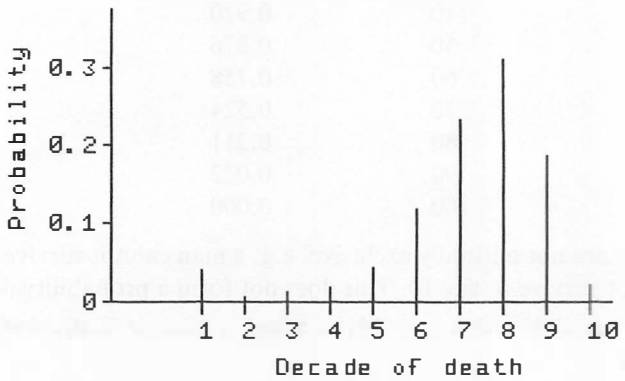


Fig. 17.9 Probability distribution of decade of death.

9. We find the expected values or mean of a probability distribution by summing each value times its probability (Section 6.4):

$$
\begin{aligned}
5 \times 0.041 &= 0.205 \\
15 \times 0.007 &= 0.105 \\
25 \times 0.014 &= 0.350 \\
35 \times 0.018 &= 0.630 \\
45 \times 0.044 &= 1.980 \\
55 \times 0.118 &= 6.490 \\
65 \times 0.234 &= 15.210 \\
75 \times 0.313 &= 23.475 \\
85 \times 0.189 &= 16.065 \\
95 \times 0.022 &= 2.090 \\
&\underline{\quad\quad} \\
&66.600
\end{aligned}
$$

Life expectancy at birth 66.6 years

Exercise 7M

1. (a) T, (7.2). (b) T, (7.2, 7.4, 7.5). (c) F, (7.2). (d) F, (7.2), see also (15.5). (e) T, (7.2).

2. (a) F, it is symmetrical (7.3). (b) F, it is 0 (7.3). (c) F, it is 1 (7.3). (d) T, (7.3). (e) T, because it is symmetrical (7.3, 4.6).

3. (a) T, (7.2). (b) T, as this is the median (7.2). (c) F, we do not know. The Normal Distribution has nothing to do with normal physiology (7.2). (d) F, 2.5 per cent do (7.2). (e) F, 2.5 per cent will be greater (7.2).

4. (a) F, this depends on the skewness (4.6), not the sample size. (b) T, (4.6). (c) T, because of the central limit theorem (7.4). (d) F, the sample size should not affect the mean. (e) F, depends on the shape of the frequency distribution and the nature of the variable.

5. (a) T, multiply by a constant (7.3). (b) F, follows a very skew Chi-squared Distribution with one degree of freedom (7A). (c) T, add a constant (7.3). (d) T, difference between two independent Normal variables (7.3). (e) F, the Normal Distribution is only preserved by adding or subtracting variables or constants and multiplying or dividing by constants. In fact this follows the t Distribution with one degree of freedom (7A).

Exercise 7E

1. The box and whisker plot shows a very slight degree of skewness, the lower whisker being shorter than the upper and the lower half of the box smaller than the upper. From the histogram it appears that the tails are a little longer than the Normal curve of Fig. 7.10 would suggest. Figure 17.10 shows the Normal Distribution with the same mean and variance superimposed on the histogram, which also indicates this.

2. We have $n = 40$. For $i = 1$ to 40 we want to calculate $(i - \frac{1}{2})/n = (2i - 1)/2n$. This gives us a probability. We use Table 7.1 to find the value of the Normal Distribution corresponding to this probability. For example, for $i = 1$ we have

$$\frac{(2i - 1)}{2n} = \frac{2 - 1}{2 \times 40} = \frac{1}{80} = 0.0125$$

From Table 7.1 we cannot find the value of x corresponding to $P = 0.0125$ directly, but we see that $x = -2.3$ corresponds to $P = 0.011$ and $x = -2.2$ to

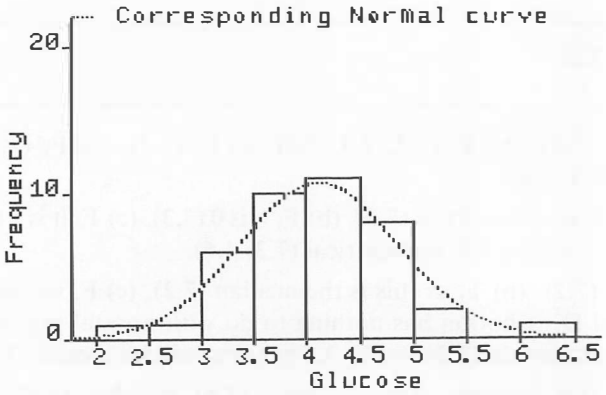


Fig. 17.10 Histogram of the blood glucose data with the corresponding Normal curve.

$P = 0.014$. $P = 0.0125$ is midway between these probabilities so we can estimate the value of x as midway between -2.3 and -2.2, giving -2.25. This corresponds to the lowest blood glucose, 2.2. For $i = 2$ we have $P = 0.0375$. Referring to the table we have $x = -1.8$, $P = 0.036$ and $x = -1.7$, $P = 0.045$. For $P = 0.0375$ we must have x just above -1.8, about -1.78. The corresponding blood glucose is 2.9. We do not have to be very accurate because we are only using this plot for a rough guide. We get a set of probabilities as follows:

i	$(2i-1)/2n$	$= P$	x	blood glucose
1	1/80	= 0.0125	−2.25	2.2
2	3/80	= 0.0375	−1.78	2.9
3	5/80	= 0.0500	−1.65	3.3
4	7/80	= 0.0875	−1.36	3.3

and so on. Because of the symmetry of the Normal Distribution, from $i = 21$ onwards the values of x are those corresponding to $40 - i + 1$, but with a positive sign. The Normal plot is shown in Fig. 17.11.

3. The points do not lie on a straight line. The central part of the line is not very far from it, but there are pronounced bends near each end. These bends reflect rather long tails of the distribution of blood glucose. If the line showed a steady curve, rising less steeply as the blood glucose value increased, this would show simple skewness which can often be corrected by a log transformation. This would not work here; the bend at the lower end would be made worse.

The deviation from a straight line is not very great, compared, say, to Fig. 7.21. As we shall see in Chapter 10, such small deviations from Normality do not usually matter.

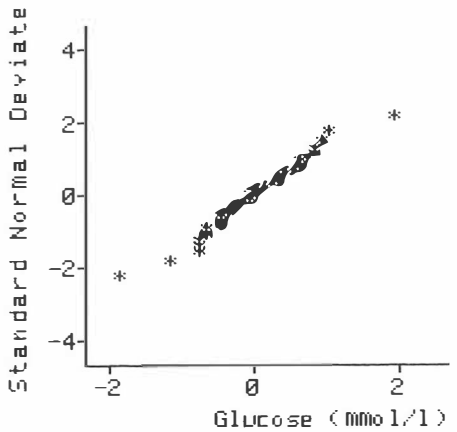


Fig. 17.11 Normal plot of the blood glucose data.

Exercise 8M

1. (a) F, standard deviation measures this. (b) F, see Chapter 15. (c) T (8.2). (d) F, it is proportional to the square root of the number of observations. (e) F, it must be less, or equal when the sample size is 1.

2. (a) F (8.3). (b) T (8.3). (c) F, the sample mean is always in the middle of limits. (d) T (8.3). (e) F.

3. (a) F. (b) T, $s/\sqrt{n}$. (c) F. (d) F. (e) T, $n - 1$.

4. (a) T, (8.1). (b) T, Chapter 7. (c) T (8.2). (d) F, this is $100 \times 0.1 \times 0.9 = np(1-p)$, it should be $p(1-p)/n = 0.1 \times 0.9/100 = 0.0009$. (e) F, the number in the sample with the condition follows a Binomial Distribution, not the proportion.

5. (a) F, it depends on the variability of FEV1, not the mean, (8.2). (b) F, it depends on the number in the sample only. (c) T, (8.2). (d) T, the sample should be random, (3.2). (e) T, 8.2.

Exercise 8E

1. The standard error of the mean is $s/\sqrt{n}$.

Insulin: $s/\sqrt{n} = 0.068/\sqrt{227} = 0.0045$

Oral: $s/\sqrt{n} = 0.070/\sqrt{225} = 0.0047$
Diet: $s/\sqrt{n} = 0.070/\sqrt{127} = 0.0062$
All non-insulin: $s/\sqrt{n} = 0.070/\sqrt{352} = 0.0037$

2. Difference = 0.744 − 0.756 = −0.012
standard error is

$$\sqrt{se_1^2 + se_2^2} = \sqrt{0.0047^2 + 0.0062^2} = 0.0078$$

The samples are large, so the 95 per cent confidence interval for the difference is

$$-0.012 - 1.96 \times 0.0078 \text{ to } -0.012 + 1.96 \times 0.0078 = -0.027 \text{ to } 0.003$$

3. Difference = 0.719 − 0.748 = −0.029
Standard error is

$$\sqrt{se_1^2 + se_2^2} = \sqrt{0.0045^2 + 0.0037^2} = 0.0059$$

The 95 per cent confidence interval for the difference is

$$-0.029 - 1.96 \times 0.0059 \text{ to } -0.029 + 1.96 \times 0.0059 = -0.040 \text{ to } -0.018$$

4. Magnesium levels are related to treatment. Among non-insulin treated patients there may be no difference, though the data suggest that the patients given oral hypoglycaemics may have lower plasma magnesium levels than those treated by diet alone. The difference could be as great as 0.03 mmol/l. Patients receiving insulin have clearly lower plasma magnesium levels than non-insulin treated patients, the difference being between 0.02 and 0.04 mmol/l.

5. To estimate mean plasma magnesium to within 1 per cent, we require the 95 per cent confidence interval to be mean ± 0.01 × mean. So for a mean of 0.72, say, we require $1.96 \times s/\sqrt{n} = 0.01 \times 0.72$ and from the table we expect s to be about 0.07 so

$$n = \frac{(1.96 \times 0.070)^2}{0.01 \times 0.72} = 363$$

We should add a few to allow for lost blood samples etc., so 400 would be a good number to aim for.

Exercise 9M

1. (a) F, there may be other differences related to coffee drinking, such as smoking (3.5). (b) T, (9.6). The relationship may not be causal, however. (c) F, there is a relationship (9.6). (d) F, not necessarily causal. (e) F, we only know that they are related.

2. (a) T, the number with lower readings could also be used (9.2). (b) T. (c) F, it is quite possible for either to be higher and deviations in either direction are important (9.5). (d) T, (9.2). $n = 16$ because the subject giving the same reading on both gives no information about the difference and is excluded from the test. (e) T, (2.4). In fact, they were.

3. (a) F, it is that the population means are equal (9.7). (b) F, that is what we are trying to find out. (c) F, (8.5). (d) F, there is no need for this. (e) T, (9.7).

4. (a) F, it may be very effective (9.6). (b) F, (9.6). (c) F, the trial is small and it may be due to chance. We must do a bigger trial. (d) F, this would completely invalidate the test. If the null hypothesis is true, the test will give a 'significant' result one in 20 times. If we keep adding cases and doing many tests we have a very high chance of getting a 'significant' result on one of them, even though there is no treatment effect. (e) T, we need to increase the power (9.9).

5. (a) T, the large sample methods depend on estimates of variance obtained from the data. This estimate gets closer to the population value as the sample size increases (Sections 9.7, 9.8). (b) F, the chance of an error of the first kind is the significance level set in advance, say 5 per cent. (c) T, the larger the sample the more likely we are to detect a difference should one exist (9.9). (d) T, (9.9). (e) F, the null hypothesis depends on the phenomena we are investigating, not on the sample size.

Exercise 9E

1. The null hypothesis is that the proportion of vaccinated and non-vaccinated children who develop polio are the same. We use the test for two proportions, Section 9.8. First we calculate the two proportions, p_1 and p_2, and the combined proportion, p.

Vaccinated group: $n_1 = 200\,745 \quad r_1 = 33$

$$p_1 = \frac{33}{200\,745} = 0.000\,164\,39$$

Control group: $n_2 = 201\ 229 \quad r_2 = 115$

$$p_2 = \frac{115}{201\ 229} = 0.000\ 571\ 49$$

Combined:

$$p = \frac{33 + 115}{200\ 745 + 201\ 229} = \frac{148}{401\ 974} = 0.000\ 368\ 18$$

Test statistic:

$$\frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.000\ 164\ 39 - 0.000\ 571\ 44}{\sqrt{0.000\ 368\ 18 \times (1 - 0.000\ 368\ 18) \times \left(\frac{1}{200\ 745} + \frac{1}{201\ 229}\right)}}$$

$$= \frac{-0.000\ 407\ 10}{0.000\ 060\ 52}$$

$$= -6.72$$

If the null hypothesis were true, this would be an observation from the Standard Normal Distribution. From Table 7.2 we see that the probability of such an extreme value is much less than 0.1 per cent or 0.001. Hence the difference is highly significant.

2. This is a randomized double-blind trial and it is reasonable to suppose that any difference which occurred must be due to the treatment.

3. To find the 95 per cent confidence interval we must see the standard error formula of Section 8.6, which does not assume that there is no difference. The standard error of the difference is

$$se(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$= \sqrt{\frac{0.000\ 164\ 39 \times (1 - 0.000\ 164\ 39)}{200\ 745} + \frac{0.000\ 571\ 49 \times (1 - 0.000\ 571\ 49)}{201\ 229}}$$

$$= 0.000\ 060\ 47$$

This is very similar to the standard error assuming equality of p_1 and p_2, because the sample sizes are so similar and the proportions so small that the $1 - p_1$ and $1 - p_2$ terms could be omitted, being almost one.

The 95 per cent confidence interval is found by

$$p_1 - p_2 \pm 1.96 se(p_1 - p_2)$$
$$= 0.000\ 164\ 39 - 0.000\ 571\ 49 \pm 1.96 \times 0.000\ 060\ 47$$

$$= -0.000\ 407\ 10 \pm 0.000\ 118\ 53$$
$$= -0.000\ 525\ 63 \text{ to } -0.000\ 288\ 57$$

Rounding to 2 significant figures gives us a reduction in proportion contracting polio of between 0.000 29 and 0.000 53, or between 29 and 53 cases per 100 000. With a control polio rate of 57 per 100 000, the vaccine was clearly effective.

4. The determination of the sample size is a challenge, as we have not considered this for the comparison of two proportions, but only for two means. We follow the method of Section 9.10. As we saw in Section 9.9, the relationship between the power of a test for a given difference and a significance level of 0.05, and the expected value of the test statistic is that the power is $1 - P(x)$ where

$$x = 1.96 - \text{expected test statistic}$$

since the absolute value of the test statistic must exceed the critical value $P(x)$ is the cumulative Normal Distribution function of Table 7.1. For a power of 90 per cent or 0.90, $x = -1.28$ (Section 9.10). The test statistic is

$$\frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

For this trial the sample size are equal, $n_1 = n_2 = n$, so $p = (p_1 + p_2)/2$. We expect the proportion in the control group, p_1, to be about 50 per 100 000, or 0.0005. We assume p_2 to be 60 per cent of this, 0.0003, to give the reduction in the number of cases of 40 per cent. So we want to have a highly probability of detecting a difference when $p = (0.0005 + 0.0003)/2 = 0.0004$ and $(1 - p) = 1 - 0.0004 = 0.9996$.

$$x = 1.96 - \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$-1.28 = 1.96 - \frac{0.0005 - 0.0003}{\sqrt{0.0004 \times 0.9996 \times \frac{2}{n}}}$$

$$-3.24 = \frac{-0.0002}{\sqrt{\frac{0.0008}{n}}}$$

$$\sqrt{\frac{0.0008}{n}} = \frac{-0.0002}{-3.24} = 0.000\ 061\ 729$$

$$n = \frac{0.0008}{0.000\ 061\ 729^2} = 209\ 948$$

So 200 000 children in each group is a very reasonable sample size for this study. After all, a vaccine which reduced the disease by less than 40 per cent would not really be worth using.

Exercise 10M

1. (a) F, it is equivalent to the Normal Distribution method of 8.3. (b) F, it is for quantitative data. (c) T. (d) F, it is for a single or two matched samples. (e) T, (10.2).

2. (a) F, (10.3). (b) T, (10.3). (c) F, this is what we are trying to find out. (d) T, (10.3). (e) F, the large sample case is like the Normal test of 9.7, except for the common variance estimate. It is valid for any sample size.

3. (a) F, the assumption of a Normal Distribution would not be met. (b) T, the distribution followed by the data would not matter (9.7). (c) T, (10.4). (d) F, the sign test is for paired data. (e) F, we have measurements, not qualitative data.

4. (a) F, (10.5). (b) T, (10.5). (c) T, the more different the sample sizes are, the worse is the approximation to the t Distribution (10.5). (d) F, this becomes a large-sample Normal Distribution test (9.7). (e) F, grouping of data is not a serious problem (10.5).

5. (a) F, for a Normal Distribution $\bar{x}$ and s^2 are independent (7A). (b) T, (7.4). (c) T, it will follow this distribution multiplied by $\sigma^2/(n-1)$, where σ^2 is the population variance. (d) F, this is only true if the mean of the population distribution is zero (10.1). (e) T, (7A).

Exercise 10E

1. The differences for $p_a(O_2)$ and compliance are shown in Table 17.1. The stem and leaf plot is:

4	2
3	
2	0
1	
0	2 6 6 7

−0	1 3 5 5 6 8
−1	6 7
−2	8
−3	
−4	6

The distribution is fairly symmetrical, though the tails are rather long for a Normal Distribution. The t Distribution should be a reasonable approximation here, as there is little skewness.

2.

$$\Sigma x = -5.2 \quad \Sigma x^2 = 58.94 \quad \bar{x} = -0.325$$

$$\Sigma x^2 - \frac{(\Sigma x)^2}{n} = 58.94 - \frac{(-5.2)^2}{16} = 57.25$$

$$s^2 = \frac{1}{15} \times 57.25 = 3.8167 \qquad s = 1.9536$$

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{3.8167}{16}} = 0.48841$$

3. We have 15 degrees of freedom, so, from Table 10.1, $t = 2.31$. The 95 per cent confidence interval is

$$\bar{x} - t\sqrt{\frac{s^2}{n}} \text{ to } \bar{x} + t\sqrt{\frac{s^2}{n}}$$

$$-0.325 - 2.13 \times 0.48841 \text{ to } -0.325 + 2.13 \times 0.48841$$
$$= -10.36531 \text{ to } 0.71531$$
$$= -1.4 \text{ to } 0.7.$$

There is little evidence of an effect of waveform on $p_a(O_2)$. Any effect which exists is quite small.

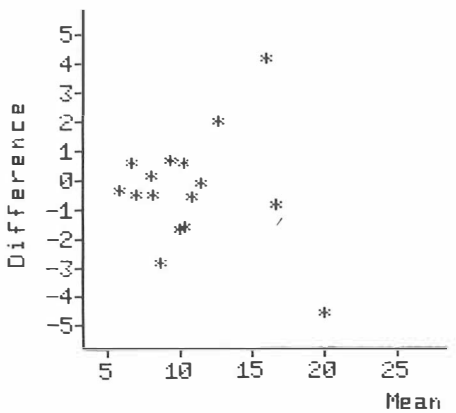


Fig. 17.12 Difference versus mean for p_a (O_2).

Table 17.1. Differences and means for $p_a(O_2)$ and compliance

	$p_a(O_2)$				Compliance			
Patient	Constant	Decelerating	Difference	Mean	Constant	Decelerating	Difference	Mean
1	9.1	10.8	−1.7	9.95	65.4	72.9	−7.5	69.15
2	5.6	5.9	−0.3	5.75	73.7	94.4	−20.7	84.05
3	6.7	7.2	−0.5	6.95	37.4	43.3	−5.9	40.35
4	8.1	7.9	0.2	8.0	26.3	29.0	−2.7	27.65
5	16.2	17.0	−0.8	16.6	65.0	66.4	−1.4	65.7
6	11.5	11.6	−0.1	11.55	35.2	36.4	−1.2	35.8
7	7.9	8.4	−0.5	8.15	24.7	27.7	−3.0	26.2
8	7.2	10.0	−2.8	8.6	23.0	27.5	−4.5	25.25
9	17.7	22.3	−4.6	20.0	133.2	178.2	−45.0	155.7
10	10.5	11.1	−0.6	10.8	38.4	39.3	−0.9	38.85
11	9.5	11.1	−1.6	10.3	29.2	31.8	−2.6	30.5
12	13.7	11.7	2.0	12.7	28.3	26.9	1.4	27.6
13	9.7	9.0	0.7	9.35	46.6	45.0	1.6	45.8
14	10.5	9.9	0.6	10.2	61.5	58.2	3.3	59.85
15	6.9	6.3	0.6	6.6	25.7	25.7	0.0	25.7
16	18.1	13.9	4.2	16.0	48.7	42.3	6.4	45.5

4. Fig. 17.12 shows the difference against the mean. There is some indication that the difference increases with the mean, but that the relationship is not strong. A logarithmic transformation of $p_a(O_2)$ may improve matters, but given the robustness of the t test this does not seem necessary.

5. Stem and leaf plot for compliance:

0	0 1 1 3 6
−0	0 1 1 2 2 3 4 5 7
−1	
−2	0
−3	
−4	5

The plot of difference against mean is Fig. 17.13. The distribution is highly skewed and the difference closely related to the mean.

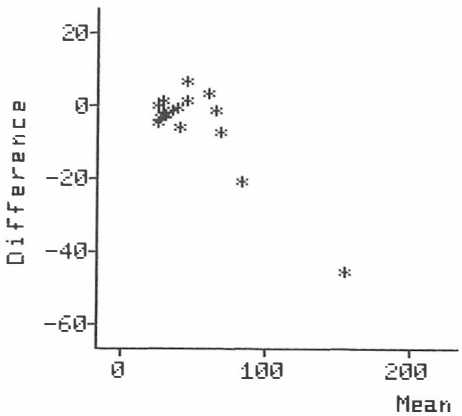


Fig. 17.13 Difference versus mean for compliance.

6. $\Sigma x = -82.7 \quad \Sigma x^2 = 2648.43 \quad \bar{x} = -5.168\ 75$

$$\Sigma x^2 - \frac{(\Sigma x)^2}{29} = 2648.43 - \frac{(-82.7)^2}{16} = 2220.974\ 38$$

$$s^2 = \frac{1}{15} \times 2220.974\ 38 = 148.064\ 96 \qquad s = 12.168$$

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{148.064\ 96}{16}} = 3.0420$$

7. As in 3 above, $t = 2.13$. The 95 per cent confidence interval is

$$-5.168\ 75 - 2.13 \times 3.0420 \text{ to } -5.168\ 75 + 2.13 \times 3.0420$$
$$= -11.6482 \text{ to } 1.3107$$
$$= -12 \text{ to } +1$$

Table 17.2. Difference and mean for log-transformed compliance (to base 10)

	Log compliance			
Patient	Constant	Decelerating	Difference	Mean
1	1.816	1.863	−0.047	1.8395
2	1.867	1.975	−0.108	1.921
3	1.573	1.636	−0.063	1.6045
4	1.420	1.462	−0.042	1.441
5	1.813	1.822	−0.009	1.8175
6	1.547	1.561	−0.014	1.554
7	1.393	1.442	−0.049	1.4175
8	1.362	1.439	−0.077	1.4005
9	2.125	2.251	−0.126	2.188
10	1.584	1.594	−0.010	1.589
11	1.465	1.502	−0.037	1.4835
12	1.452	1.430	0.022	1.441
13	1.668	1.653	0.015	1.6605
14	1.789	1.765	0.024	1.777
15	1.410	1.410	0.000	1.410
16	1.688	1.626	0.062	1.657

8. Table 17.2 shows the log-transformed data, using logs to base 10, with their differences and sums. The stem and leaf plot is

0.06	2
0.05	
0.04	
0.03	
0.02	2 4
0.01	5
0.00	0
−0.00	9
−0.01	0 4
−0.02	
−0.03	7
−0.04	2 7 9
−0.05	
−0.06	3
−0.07	7
−0.08	
−0.09	
−0.10	8
−0.11	
−0.12	6

This is a little unwieldy and we can condense it by pairing the first significant digits:

0.06	2
0.04	
0.02	2 4
0.00	0 5
−0.00	9 0 4
−0.02	7
−0.04	2 7 9
−0.06	3 7
−0.08	
−0.10	8
−0.12	6

The difference against the mean is shown in Fig. 17.14. The differences are still related to the mean but not nearly so strongly as in Fig. 17.13. The distribution is more symmetrical and the use of the *t* Distribution seems much more reasonable than for the untransformed data.

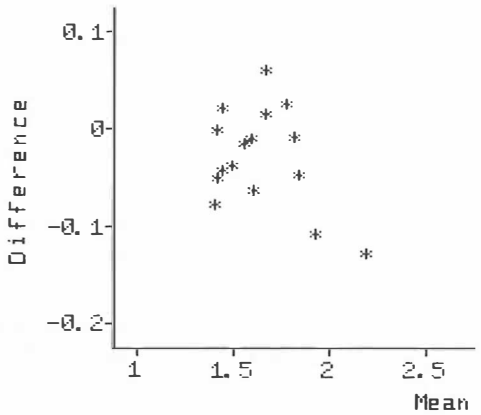


Fig. 17.14 Difference versus mean for log compliance.

9. $\Sigma x = -0.45 \quad \Sigma x^2 = 0.049\ 886 \quad \bar{x} = -0.028\ 125$

$$\Sigma x^2 - \frac{(\Sigma x)^2}{n} = 0.049\ 886 - \frac{(-0.45)^2}{16} = 0.037\ 229\ 75$$

$$s^2 = \frac{1}{15} \times 0.037\ 229\ 75 = 0.002\ 481\ 98 \qquad s = 0.049\ 820$$

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{0.002\ 481\ 98}{16}} = 0.012\ 453$$

The 95 per cent confidence interval is

$$-0.028\ 125 - 2.13 \times 0.012\ 455 \text{ to } -0.028\ 125 + 2.13 \times 0.012\ 455$$
$$= -0.054\ 654 \text{ to } -0.0015\ 959$$

If we transform these limits back by taking the antilogs we get 0.882 to 0.996. This means that the compliance with a decelerating waveform is between 0.882 and 0.996 times that with a constant waveform. There is some evidence that waveform has an effect, whereas with the untransformed data the confidence interval for the difference included zero. Because of the skewness of the raw data the confidence interval was too wide.

10. We can conclude that there is little evidence of any effect on $p_a(O_2)$, although we cannot exclude the possibility that the decelerating waveform produces a mean reduction of up to 1.4 kPa, or an increase of up to 0.7 kPa. There is some evidence of a reduction in mean compliance, which could be up to 12 per cent (from $(1 - 0.882) \times 100$), but could also be negligibly small.

Exercise 11M

1. (a) T, (11.10). (b) T, (11.11). (c) F, should be 0, (11.10). (d) F, (11.10). (e) F, this is the regression coefficient (11.3).

2. (a) F, usually has non-zero intercept, (11.3). (b) F, (11.3). (c) F, the slope and the intercept have dimensions, (11.3). (d) T, we calculate a by $\bar{y} = a + b\bar{x}$, (11.3). (e) T, (11.2, 11.3, 11.4).

3. (a) F, the independent variable has no error in the regression model (11.3). (b) T, (11.5). (c) T, (11.6). (d) F, only if necessary to achieve (b) and (c). (e) F, there is a scatter about the line, (11.3).

4. (a) F, they are closely related. In fact $y = \log(x)$ exactly. (b) F, the correlation coefficient is 0.89, (11.10). (c) T, see (b). (d) T, see (a). (e) F, this is not a straight line. Polynomial regression would be better (11.9).

5. (a) F, knowledge of x tells us something about y (6.1). (b) T, the correlation coefficient is zero (11.10). (c) F, for part of the scale y decreases as x increases. (d) F, this is not a straight line. (e) T, (11.9).

Exercise 11E

1. females:

$$b = \frac{\text{sum of products about mean}}{\text{sum of squares height}} = \frac{4206.948\ 37}{1444.600\ 47}$$
$$= 2.912\ 19$$
$$a = \text{mean PEFR} - b \times \text{mean height}$$
$$= 474.069\ 768 - 2.912\ 19 \times 165.937\ 209$$
$$= -9.170\ 91$$
$$\text{PEFR} = -9.17 + 2.91 \times \text{height}$$

males:

$$b = \frac{8993.36}{2267.499\ 31} = 3.966\ 20$$
$$a = 568.2 - 3.96620 \times 177.303\ 448$$
$$= -135.021\ 266$$
$$\text{PEFR} = -135.02 + 3.97 \times \text{height}$$

total:

$$b = \frac{39\ 619.5891}{6902.231\ 72} = 5.740\ 11$$
$$a = 528.124\ 753 - 5.740\ 11 \times 172.464\ 356$$
$$= -461.839\ 62$$
$$\text{PEFR} = -461.84 + 5.74 \times \text{height}$$

2. The slope for the combined group is considerably greater than for either sex separately, which suggests an increase in PEFR for males apart from that due to height. Note that the intercepts, which are well outside the range of the data, are meaningless on their own.

3. females:

$$r = \frac{\text{sum of products about mean}}{\sqrt{(\text{sum of squares height} \times \text{sum of squares PEFR})}}$$
$$= \frac{4206.948\ 37}{\sqrt{(1444.600\ 47 \times 101\ 107.651)}}$$
$$= 0.35$$

males:

$$r = \frac{8993.36}{\sqrt{(2267.499\ 31 \times 226\ 873.86)}}$$
$$= 0.40$$

4. We would expect the correlation for females to be less because they have a smaller range of height than do males. The smaller the range of height, the smaller will be the proportion of variability in PEFR which can be explained by variations in height. This proportion is r^2 (Section 11.10).

Exercise 12M

1. (a) T, (10.3). (b) F, this is for paired data (9.2). (c) T, (12.2). (d) F, for paired data, (12.3). (e) F, this looks for the existence of relationships between two ordinal variables, not a comparison between two groups (12.4, 12.5).

2. (a) F, there is no dependent variable in correlation (12.5, 11.12). (b) T, (12.5). (c) F, this copes well with ties (12.5). (d) T, (12.5). (e) T, this would not affect the rank order of the observation.

3. (a) F, if Normal assumptions are met the methods using them are better (12.7). (b) T, (12.7). (c) F, estimation of confidence intervals using these methods is difficult. (d) F, they required the assumption that the scale is ordinal, i.e. that the data can be ranked. (e) T, this is what they are for.

4. (a) T, (10.2). (b) F, for two samples (12.2). (c) T, (9.2). (d) T, (12.3). (e) F, this would look for a relationship between responses on the two treatments.

5. (a) T, (12.5). (b) F, U does not have expected value zero if the null hypothesis is true (12.2). (c) F, (12.3). (d) T, (12.4). (e) F, this would be an extreme value (8.2).

Exercise 12E

1. Sign test for $p_a(O_2)$. The differences are shown in Table 17.1. We have 6 positive and 10 negative differences, with no zeros. The null hypothesis is that there is no tendency for the constant waveform to produce higher or lower $p_a(O_2)$ than the decelerating waveform. Under this null hypothesis the number of positives is from a Binomial Distribution with $n = 16$ and $p = \frac{1}{2}$. We have

$$\text{Prob}(r = 6) = \frac{16!}{6! \times 10!} \times (\tfrac{1}{2})^{16} = 0.122\ 19$$

$$\text{Prob}(r = 5) = \frac{16!}{5! \times 11!} \times (\tfrac{1}{2})^{16} = 0.066\ 65$$

$$\text{Prob}(r = 4) = \frac{16!}{4! \times 12!} \times (\tfrac{1}{2})^{16} = 0.027\ 77$$

$$\text{Prob}(r = 3) = \frac{16!}{3! \times 13!} \times (\tfrac{1}{2})^{16} = 0.008\ 54$$

$$\text{Prob}(r = 2) = \frac{16!}{2! \times 14!} \times (\tfrac{1}{2})^{16} = 0.001\ 83$$

$$\text{Prob}(r = 1) = \frac{16!}{1! \times 15!} \times (\tfrac{1}{2})^{16} = 0.000\ 24$$

$$\text{Prob}(r = 0) = \frac{16!}{0! \times 16!} \times (\tfrac{1}{2})^{16} = 0.000\ 02$$

$$\text{Prob}(r \leqslant 6) = 0.227\ 24$$

Since this Binomial Distribution is symmetrical, the probability of an equally extreme result in the opposite direction is $\text{Prob}(r > 10) = 0.227\ 24$ and the two-sided probability is the sum of these, $0.227\ 24 + 0.227\ 24 = 0.454\ 48$. The difference is not significant and we have no evidence of an effect.

2. We use the Wilcoxon test for matched data. We first rank the differences irrespective of sign, and then sum the ranks of the negative differences.

Difference	−0.1	0.2	−0.3	−0.5	−0.5	−0.6	0.6	0.6
Rank	1	2	3	$4\frac{1}{2}$	$4\frac{1}{2}$	7	7	7
Difference	0.7	−0.8	−1.6	−1.7	2.0	−2.8	4.2	−4.6
Rank	9	10	11	12	13	14	15	16

Sum of ranks for positive differences, $T = 2 + 7 + 7 + 9 + 13 + 15 = 53$. From Table 13.5 the 5 per cent point is 30 and T exceeds this. The difference is not significant.

3. All three methods tell us that we have no evidence of a difference, but the t Distribution method gives us upper and lower estimates for any difference which may exist.

4. The differences are shown in Table 17.1. We have 4 positive, 11 negative and 1 zero. Under the null hypothesis of no difference, the number of positives is from the Binomial Distribution with $p = \frac{1}{2}$, $n = 15$. We have $n = 15$ because the single zero contributes no information about the direction of the difference. For $\text{Prob}(r < 4)$ we have

$$\text{Prob}(r = 4) = \frac{15!}{4! \times 11!} \times (\tfrac{1}{2})^{15} = 0.041\ 66$$

$$\text{Prob}(r = 3) = \frac{15!}{3! \times 12!} \times (\tfrac{1}{2})^{15} = 0.013\ 89$$

$$\text{Prob}(r = 2) = \frac{15!}{2! \times 13!} \times (\tfrac{1}{2})^{15} = 0.003\ 20$$

$$\text{Prob}(r = 1) = \frac{15!}{1! \times 14!} \times (\tfrac{1}{2})^{15} = 0.000\ 46$$

$$\text{Prob}(r = 0) = \frac{15!}{0! \times 15!} \times (\tfrac{1}{2})^{15} = 0.000\ 03$$

$$\text{Prob}(r \leqslant 4) = 0.059\ 24$$

If we double this for a two-sided test we get 0.118 48, again not significant.

5. Using the Wilcoxon matched-pairs test we get

Difference	−0.9	−1.2	−1.4	1.4	1.6	−2.6	−2.7	−3.0
Rank	1	2	$3\frac{1}{2}$	$3\frac{1}{2}$	5	6	7	8
Difference	3.3	−4.5	−5.9	6.4	−7.5	−20.7	−45.0	
Rank	9	10	11	12	13	14	15	

As for the sign test, the zero is omitted. Sum of ranks for positive differences is

$$T = 3\tfrac{1}{2} + 5 + 9 + 12 = 29.5$$

From Table 13.5 the 5 per cent point is 25, which T exceeds, so the difference is not significant at the 5 per cent level. The three tests give similar answers.

6. Using the log-transformed differences in Table 17.2, we still have 4 positives, 11 negatives and 1 zero, with a sign test probability of 0.118 48. The transformation does not alter the direction of the changes and so does not affect the sign test.

7. For the Wilcoxon matched-pairs test on the log compliance:

Difference	−0.009	−0.010	−0.014	0.015	0.022	0.024
Rank	1	2	3	4	5	6
Difference	−0.037	−0.042	−0.047	−0.049	0.062	−0.063
Rank	7	8	9	10	11	12
Difference	−0.077	−0.108	−0.126			
Rank	13	14	15			

$$T = 4 + 5 + 6 + 11 = 26$$

This is just above the 5 per cent point of 25 and is different from that in the untransformed data. This is because the transformation has altered the relative size of the differences. This test assumes interval data. By changing to a log scale we have moved to a scale where the differences are more comparable, because the change does depend on the magnitude of the original value. This does not happen with the other rank tests, the Mann–Whitney U test and rank correlation coefficients, which involve no differencing.

8. We have found no evidence of an effect on $p_a(O_2)$ and although there is a possibility of a reduction in compliance it does not reach the conventional level of significance.

9. The conclusions are broadly similar, but the effect on compliance is more strongly suggested by the t method. Provided the data can be transformed to approximate Normality the t Distribution analysis is more

powerful, and as it also gives confidence intervals more easily, I would prefer it.

Exercise 13M

1. (a) F, (13.1). (b) T, (13.1). (c) F, $(5 - 1) \times (3 - 1) = 8$, (13.1). (d) T, 80 per cent $\times$ 15 = 12 cells must be >5, (13.2). (e) F, (13.2).

2. (a) T, 80 per cent of 4 is greater than 3 so all must be >5, (13.2). (b) F, for categorical data, (13.1). (c) F, as b. (d) F, (13.2). (e) F, can be as small as 20, if all row and column totals are 10.

3. (a) T, (13.1). (b) T, (13.8). (c) F, the tests are independent. (d) T, $(2 - 1) \times (2 - 1) = 1$, (13.1). (e) F, with such large numbers it does not make much difference. Without the continuity correction we get chi-squared = 124.5, with it we get chi-squared = 119.4, (13.6).

4. (a) T, we look at the smoking of each matched pair. (b) T, (13.8). (c) F, we use the chi-squared test (13.1). (d) F, this is continuous variable, we use the paired *t* method (10.2). (e) F, there are two independent samples, we use the chi-squared test (13.1).

5. (a) T, (13.5). (b) T, (13.5). (c) T, (13.6). (d) T, this is its usual application, (13.5). (e) T, the factorials of large numbers can be difficult to calculate.

Exercise 13E

1. The heatwave appears to begin in week 10 and continue to include week 17. This period was much hotter than the corresponding period of 1982.

2. There were 178 admissions during the heatwave in 1983 and 110 in the corresponding weeks of 1982. We could test the null hypothesis that these came from distributions with the same admission rate and we would get a significant difference. This would not be convincing, however. It could be due to other factors, such as the closure of another hospital with resulting changes in catchment area.

3. The cross-tabulation is shown in Table 17.3.

4. The null hypothesis is that there is no association between year and period, in other words that the distribution of admissions between the

Table 17.3. Cross-tabulation of time period by year for geriatric admissions

	Period			
Year	before heatwave	during heatwave	after heatwave	Total
1982	190	110	82	382
1983	180	178	110	468
Total	370	288	192	850

periods will be the same for each year. The expected values are shown in Table 17.4.

Table 17.4. Expected frequencies for Table 17.3

	Period			
Year	before heatwave	during heatwave	after heatwave	Total
1982	166.3	129.4	86.3	382
1983	203.7	158.6	105.7	468
Total	370	288	192	850

5. The chi-squared statistic is given by:

$$\sum \frac{(O-E)^2}{E} = \frac{(190-166.3)^2}{166.3} + \frac{(110-129.4)^2}{129.4} + \frac{(82-86.3)^2}{86.3} + \frac{(180-203.7)^2}{203.7} + \frac{(178-158.6)^2}{158.6} + \frac{(110-105.7)^2}{105.7} = 11.806$$

There are 2 rows and 3 columns, giving us $(2-1) \times (3-1) = 2$ degrees of freedom. Thus we have chi-squared = 11.8 with 2 degrees of freedom. From Table 14.3 we see that this has probability of less than 0.01. The data are not consistent with the null hypothesis. The evidence supports the view that admissions rose by more than could be ascribed to chance during the 1983 heatwave. We cannot be certain that this was due to the heatwave and not some other factor which happened to operate at the same time.

6. We could see whether the same effect occurred in other districts between 1982 and 1983. We could also look at older records to see whether there was a similar increase in admissions, say for the heatwaves of 1975 and 1976.

Exercise 14M

1. (a) T. (b) F, this is dichotomous, (14.2). (c) F, this is only ordinal, (14.2). (d) T. (e) T.

2. (a) F, (14.5). (b) T, (14.5). (c) T, (14.5). (d) T, though only the rank order is used, (14.5). (e) T, though only the rank order is used, (14.5).

3. (a) F, for independent samples, (14.3). (b) T, (14.4). (c) T, (14.4). (d) F, the sample is too small. (e) T, (14.4).

4. (a) F, for continuous data, (12.3). (b) F, for continuous data, (12.10). (c) F, for continuous data, (10.2). (d) F, for ordinal data, (12.5). (e) T, (13.1).

5. (a) T, (14.2). (b) F, for dichotomous data, (14.2). (c) T, the two-sample t test, (14.2). (d) F, for a single or matched sample, (14.3). (e) F, for a single or matched sample, (14.3).

Exercise 14E

1. Overall preference: we have one sample of patients, of whom 12 preferred A, 14 preferred B and 4 did not express a preference. We can use a Binomial or sign test (Section 8.1), only considering those who expressed a preference. Those for A are positives, those for B are negatives. We get two-sided $p = 0.85$, not significant.

Preference and order: we have the relationship between two variables, preference and order, both nominal. We set up a two-way table and do a chi-squared test. For the 3 by 2 table we have two expected frequencies less than five, so we must edit the table. There are no obvious combinations, so we delete those who expressed no preference, leaving a 2 by 2 table, $\chi^2 = 1.3$, 1 d.f., $p > 0.05$.

2. Both variables have very non-Normal Distributions. The pH is bimodal and nitrite is highly skew. It might be possible to transform the nitrites to a Normal Distribution but the transformation would not be a simple one. The zero prevents a simple logarithmic transformation, for example. Because of this, regression and correlation are not appropriate and rank correlation can be used. Spearman's $\rho = 0.58$ and Kendall's $\tau = 0.40$ both give a probability of 0.004.

3. The trial will have to be a two-group comparison, as we cannot wait for our subjects to have two labours. The outcome, mode of delivery, is

categorical and we are particularly interested in the proportion of instrumental deliveries. We therefore plan our trial as a comparison of two proportions. We need to know what proportion of epidurals have instrumental deliveries and what sort of reduction in the rate we are looking for. The first is fairly easy to get, the second more difficult. One approach is to use the methods of 9.10 to show the size of reduction we can detect with reasonable power for a range of sample sizes. Then we can use the proportion of women given epidurals and the number of deliveries per year to estimate how long a trial these different sample sizes would require and decide whether it is feasible to detect the sort of difference we may expect or hope for.

4. We must use the total number of patients we randomized to treatments, because these are the comparable groups. Thus we have 1711 active treatment patients including 15 deaths, and 1706 placebo patients with 35 deaths. A chi-squared test gives us $\chi^2 = 8.2$, d.f. = 1, $p < 0.01$. A comparison of two proportions gives a difference of -0.0117 with 95 per cent confidence interval -0.0198 to -0.0037 (8.6) and test of significance using the Standard Normal Distribution gives a value of 2.86, $p < 0.01$, (9.8).

5. The data are paired so we use a paired t test (10.2). The assumption of a Normal Distribution for the differences should be met as PEFR itself follows a Normal Distribution fairly well. We get $t = 6.45/5.05 = 1.3$, d.f. = 31, which is not significant, and a 95 per cent confidence interval of -3.85 to 16.75 litre/min.

6. We want to test for the relationship between two variables, which are both presented as categorical. We use a chi-squared test for a contingency table, $\chi^2 = 38.1$, d.f. = 6, $p < 0.001$. One possibility is that some other variable, such as the mother's smoking, or poverty, is related to both maternal age and asthma. Another is that there is a cohort effect. All the age 14–19 mothers were born during the Second World War, and some common historical experience may have produced the asthma in their children.

7. We have two large samples and can do the Normal comparison of two means (8.5). The standard error of the difference is 0.0178 s and the observed difference is 0.02 s, giving a 95 per cent confidence interval of -0.015 to 0.055 for the excess mean transit time (MTT) in the controls. For matched cases only, for each case we could calculate the mean MTT for the two controls matched to each case, find the difference between case MTT and control mean MTT, and use the one sample method of Section 8.3.

Exercise 15M

1. (a) T, (15.4). (b) F, this is measured by sensitivity (15.4). (c) T, (15.4). (d) F, this is the proportion agreeing (15.4). (e) F, we need the sensitivity as well (15.4). There are other things, dependent on the population studied, which may be important too, like the false positive rate.

2. (a) T, (15.1). (b) F, this would depend only on how variable were the true values which we were trying to measure. (c) T, (15.1). (d) T, (15.1). (e) F, unless the measurement process changes the subject, we would expect this to be zero.

3. (a) F, we expect 5 per cent of 'normal' men to be outside these limits (15.5). (b) F, see (a). (c) F, he may have a disease which does not produce an abnormal haematocrit. (d) F, this reference range is for men. Women may have a different distribution of haematocrit and it is dangerous to extrapolate the reference range to a different population. In fact, for women the reference range quoted was 35.8 to 45.4, putting a woman with a haematocrit of 48 outside the reference range. (e) T, the haematocrit outside the range suggests it, although it does not prove it.

4. (a) T, it is based on fewer potential survivors (15.6). (b) F, they contribute half an interval at risk (15.6). (c) T, if survival rates change those subjects starting later, and so more likely to be withdrawn, will have a different survival to those starting earlier. The first part of the curve will represent a different population to the second. (d) T, the longest survivor may still be alive and so become a withdrawal. (e) T, (15.6).

5. (a) T, (15.2). (b) F, (15.2). (c) T, (15.2). (d) T, (15.2). (e) F, (15.2).

Exercise 15E

1. The blood donors were used because it was easy to get the blood. This would produce a sample deficient in older people, so it was supplemented by people attending day centres. This would ensure that these were reasonably active, healthy people for their age. Given the problem of getting blood and the limited resources available, this seems a fairly satisfactory sample for the purpose. The alternative would be to take a random sample from the local population and try to persuade them to give the blood. There might have been so many refusals that volunteer bias would make the sample unrepresentative anyway.

The sample is also biased geographically, being drawn from one part of London. In the context of the study, where we wanted to compare diabetics with normals this did not matter so much, as both groups came from the same place. For a reference range, if there were a geographical factor, the range would be biased in other places. To look at this we would have to repeat the study in several places, compare the resulting ranges and pool as appropriate.

2. We want normal, healthy people for the sample, so we want to exclude people with obvious pathology and especially those with disease known to affect the quantity being measured. However, if we excluded all elderly people currently receiving drug therapy we would find it very difficult to obtain a sufficiently large sample. It is indeed 'normal' for the elderly to be taking analgesics and hypnotics, so these were permitted.

3. From the shape of the histogram, the distribution of plasma magnesium does indeed appear Normal. Figure 17.15 shows the superimposed Normal curve.

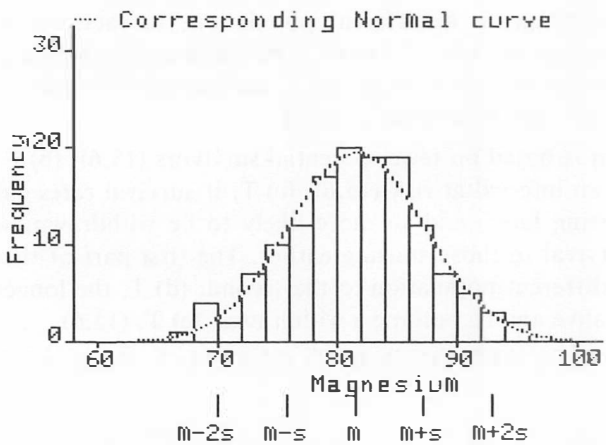


Fig. 17.15 Distribution of plasma magnesium in 140 apparently healthy people, with superimposed normal curve, mean, and standard deviation.

4. The reference range, outside which about 5 per cent of normal values are expected to lie, is $(\bar{x} - 2s)$ to $(\bar{x} + 2s)$, or $(0.810 - 2 \times 0.057)$ to $(0.810 + 2 \times 0.057)$, which is 0.696 to 0.924, or 0.70–0.92 mmol/litre.

5. As the sample is large and the data Normally distributed the standard error of the limits is approximately

$$\sqrt{\frac{3s^2}{n}} = \sqrt{\frac{3 \times 0.057^2}{140}}$$
$$= 0.008\ 343\ 9$$

For the 95 per cent confidence interval we take 1.96 standard errors on either side of the limit, $1.96 \times 0.008\,343\,9 = 0.016$. The 95 per cent confidence interval for the lower reference limit is $0.696 - 0.016$ to $0.696 + 0.016 =$ 0.680 to 0.712 or 0.68 to 0.71 mmol/litre. The confidence interval for the upper limit is $0.924 - 0.016$ to $0.696 + 0.016 = 0.908$ to 0.940 or 0.91 to 0.94 mmol/litre. We can see that the reference range is well estimated as far as sampling errors are concerned.

6. Plasma magnesium did indeed increase with age. The variability did not. This would mean that for older people the lower limit would be too low and the upper limit too high, as the few above this would all be elderly. We could simply estimate the reference range separately at different ages. We could do this using separate means but a common estimate of variance, obtained like that for the two sample t test in Section 10.3. Or we could use the regression of magnesium on age to get a formula which would predict the reference range for any age. The method chosen would depend on the nature of the relationship.

Exercise 16M

1. (a) F, this is the SMR. (b) T, (16.1). (c) F, it is for a specific age group, not age adjusted. (d) F, it measures the number of deaths per person at risk, not the total number. (e) F, it tells us nothing about age structure.

2. (a) F, (16.4). (b) T, this is how the life table is calculated. (c) T, the distribution of age at death if these mortality rates apply (Exercise 6E). (d) T, (16.4). (e) T, (16.4).

3. (a) T, in fact 7.7 times as likely. (b) F, age effects have been adjusted for. (c) F, it may be true, but it may also be that heavy drinkers become publicans. It is difficult to infer causation from observational data. (d) F, men at high risk of cirrhosis of the liver, i.e. heavy drinkers, may not become window cleaners, or window cleaners who drink may change their occupation, which requires good balance. (e) F, they have a low risk. Here the 'average' ratio is 100, not 1.0.

4. (a) F, this tells us about mortality, not population structure. (b) F. (c) F, see a. (d) T, (16.6). (e) F, a bar chart shows the relationship between two variables, not their frequency distribution (5.5).

5. (a) T, (16.2). (b) F, (16.5). (c) F, this is a rate per 1000 live births (16.5). (d) F, (16.1). (e) T, does not depend on age distribution (16.4).

Exercise 16E

1. We obtain the rates for the whole period by dividing the number of deaths in an age group by the population size. Thus for ages 10–14 we have 44/4271 = 0.010 30 cases per thousand population. This is for a thirteen-year period so the rate per year is 0.010 30/13 = 0.000 79 per 1000 per year, or 0.79 per million per year. Table 17.5 shows the rates for each age group.

Table 17.5. Age-specific mortality rates for volatile substance abuse, Great Britain, and calculation of SMR for Scotland

	Great Britain a.s.m.r.s			
Age group	per million per year	per thousand per 13 years	Scotland population (thousands)	Scotland expected deaths
0–9	0.00	0.000 00	653	0.000 00
10–14	0.79	0.010 30	425	4.377 50
15–19	2.58	0.033 58	447	15.010 26
20–24	0.87	0.011 37	394	4.479 78
25–29	0.32	0.004 15	342	1.419 30
30–39	0.08	0.001 08	659	0.711 72
40–49	0.03	0.000 33	574	0.189 42
50–59	0.09	0.001 12	579	0.648 48
60+	0.03	0.000 37	962	0.355 94
TOTAL				27.192 490

The rates are unusual because they are highest among the adolescent group, where mortality rates for most causes are low. Anderson *et al.* (1985) note that '. . . our results suggest that among adolescent males abuse of volatile substances currently account for 2 per cent of deaths from all causes . . .' The rates are also unusual because we have not calculated them separately for each sex. This is partly for simplicity and partly because the number of cases in most age groups is small as it is.

2. The expected number of deaths by multiplying the number in the age group in Scotland by the death rate for the period, i.e. per thirteen-years, for Great Britain. We then add these to get 27.19 deaths expected altogether. We observed 48, so the SMR is 48/27.19 = 1.76, or 176 with Great Britain as 100.

3. We find the standard error of the SMR by $\sqrt{O/E} = \sqrt{48/27.19} = 0.2548$. The 95 per cent confidence interval is then $1.76 - 1.96 \times 0.2548$ to $1.76 + 1.96 \times 0.2548$, or 1.26 to 2.25. With Great Britain as 100 we get 126 to

225. The observed number is quite large enough for the Normal approximation to the Poisson Distribution to be used.

4. Yes, the confidence interval is well away from zero. Other factors relate to the data collection, which was from newspapers, coroners, death registrations etc. Scotland has different newspapers and other news media and a different legal system to the rest of Great Britain. It may be that the association of deaths with VSA is more likely to be reported there than in England and Wales.

References

Altman, D.G. (1982). Statistics and ethics in medical research. In *Statistics in practice*, (ed. S.M. Gore and D.G. Altman). British Medical Association, London.

Anderson, H.R., MacNair, R.S. and Ramsay, J.D. (1985). Deaths from abuse of volatile substances: a national epidemiological study. *British Medical Journal*, **290**, 304–7.

Armitage, P. (1971). *Statistical methods in medical research*. Blackwell, Oxford.

Armitage, P. (1975). *Sequential medical trials*, 2nd edn, Blackwell, Oxford.

Banks, M.H., Bewley, B.R., Bland J.M., Dean, J.R. and Pollard, V.M. (1978). A long-term study of smoking by secondary schoolchildren. *Archives of Diseases in Childhood*, **53**, 12–9.

Bewley, B.R. and Bland, J.M. (1976). Academic performance and social factors related to cigarette smoking by schoolchildren. *British Journal of Preventive Social Medicine*, **31**, 18–24.

Bewley, B.R., Bland, J.M. and Harris, R. (1974). Factors associated with the starting of cigarette smoking by primary schoolchildren. *British Journal of Preventive Social Medicine*, **28**, 37–44.

Bewley, T., Bland, J.M., Ilo, M., Walch, E. and Willington, G. (1975). Census of mental hospital patients and life expectancy of those unlikely to be discharged. *British Medical Journal*, **4**, 671–5.

Bewley, T., Bland, J.M., Mechen, D. and Walch, E. (1981). 'New chronic' patients. *British Medical Journal*, **283**, 1161–4.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307–10.

Bland, J.M., Bewley, B.R. and Banks, M.H. (1979). Cigarette smoking and children's respiratory symptoms: validity of questionnaire method. *Revue d'Epidémiolgie et de Santé Publique*, **27**, 69–76.

Bland, J.M., Bewley, B.R., Banks, M.H. and Pollard, V.M. (1975). Schoolchildren's beliefs about smoking and disease. *Health Education Journal*, **34**, 71–8.

Bland, J.M., Bewley, B.R., Pollard, V. and Banks, M.H. (1978). Effect of children's and parents' smoking on respiratory symptoms. *Archives of Diseases in Childhood*, **53**, 100–5.

Bland, J.M., Mutoka, C. and Hutt, M.S.R. (1977). Kaposi's sarcoma in Tanzania. *East African Journal of Medical Research*, **4**, 47–53.

British Standards Institution. (1979). Precision of test methods 1: Guide for the

determination and reproducibility for a standard test method (BS5497, part 1). BSI, London.

Bryson, M.C. (1976). The *Literary Digest* poll: making of a statistical myth. *The American Statistician*, **30**, 184–5.

Burr, M.L., St Leger, A.S. and Neale, E. (1976). Anti-mite measures in mite-sensitive adult asthma: a controlled trial. *Lancet*, **i**, 333–5.

Carleton, R.A., Sanders, C.A. and Burack, W.R. (1960). Heparin administration after acute myocardial infarction. *New England Journal of Medicine*, **263**, 1002–4.

Christie, D. (1979). Before-and-after comparisons: a cautionary tale. *British Medical Journal*, **2**, 1629–30.

Colton, T. (1974). *Statistics in medicine*. Little Brown, Boston.

Conover, W.J. (1980). *Practical Nonparametric statistics*. 2nd edn, John Wiley and Sons, New York.

Davies, O.L. and Goldsmith, P.L. (1972). *Statistical methods in research and production*. Oliver and Boyd, Edinburgh.

DHSS (1978). *Prevention and health: everybody's business*, HMSO, London.

Doll, R. and Hill, A.B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, **ii**, 739–48.

Doll, R. and Hill, A.B. (1956). Lung cancer and other causes of death in relation to smoking: a second report on the mortality of British doctors. *British Medical Journal*, **2**, 1071–81.

Donnan, S.P.B. and Haskey, J. (1977). Alcoholism and cirrhosis of the liver. *Population Trends*, **7**, 18–24.

Finney, D.J., Latscha, R., Bennett, B.M. and Hsa, P. (1963). *Tables for testing significance in a 2 × 2 contingency table*. Cambridge University Press, London.

Fish, P.D., Bennett, G.C.J., Millard, P.H. (1985). Heatwave morbidity and mortality in old age. *Age and Aging*, **14**, 243–45.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, **15**, 246–63.

Hart, P.D. and Sutherland, I. (1977). BCG and vole bacillus vaccine in the prevention of tuberculosis in adolescent and early adult life. *British Medical Journal*, **2**, 293–295.

Healy, M.J.R. (1968). Disciplining medical data. *British Medical Bulletin*, **24**, 210–14.

Hedges, B.M. (1978). Question wording effects: presenting one or both sides of a case. *The Statistician*, **28**, 83–99.

Hill, A.B. (1962). *Statistical methods in clinical and preventive medicine*. Churchill Livingstone, Edinburgh.

Hill, A.B. (1977). *A short textbook of medical statistics*. Hodder and Stoughton, London.

Holland, W.W., Bailey, P. and Bland, J.M. (1978). Long-term consequences of respiratory disease in infancy. *Journal of Epidemiology and Community Health*, **32**, 256–9.

Holten, C. (1951). Anticoagulants in the treatment of coronary thrombosis. *Acta Medica Scandinavica*, **140**, 340–8.

Huff, D. (1954). *How to lie with statistics*. Gollancz, London.

Huskisson, E.C. (1974). Simple analgesics for arthritis. *British Medical Journal*, **4**, 196–200.

James, A.H. (1977). Breakfast and Crohn's disease. *British Medical Journal*, **1**, 943–7.

Johnson, F.N. and Johnson, S. (eds) (1977). *Clinical trials*. Blackwell, Oxford.

Johnston, I.D.A., Anderson, H.R., Lambert, H.P. and Patel, S. (1983). Respiratory morbidity and lung function after whooping cough. *Lancet*, **ii**, 1104–8.

Johnston, I.D.A., Anderson, H.R. and Patel, S. (1984). Variability of peak flow in wheezy children. *Thorax*, **39**, 583–7.

Kendall, M.G. and Babington Smith, B. (1971). *Tables of random sampling numbers*, Cambridge University Press, Cambridge.

Lancet (1980). BCG: bad news from India. *Lancet*, **1**, 73–4.

Lee, K.L., McNeer, J.F., Starmer, F.C., Harris, P.J. and Rosati, R.A. (1980). Clinical judgement and statistics: lessons from a simulated randomised trial in coronary artery disease. *Circulation*, **61**, 508–15.

Leonard, J.V., Whitelaw, A.G.L., Wolff, O.H., Lloyd, J.K. and Slack, S. (1977). Diagnosing familial hypercholesterolaemia in childhood by measuring serum cholesterol. *British Medical Journal*, **1**, 1566–8.

Levine, M.I. and Sackett, M.F. (1946). Results of BCG immunization in New York City. *American Review of Tuberculosis*, **53**, 517–32.

Lindley, D.L. and Miller, J.C.P. (1955). *Cambridge Elementary Statistical Tables*, Cambridge University Press.

Luthra, P., Bland, J.M. and Stanton, S.L. (1982). Incidence of pregnancy after laparoscopy and hydrotubation. *British Medical Journal*, **284**, 1013.

Mather, H.M., Nisbet, J.A., Burton, G.H., Poston, G.H., Bland, J.M., Bailey, P.A. and Pilkington, T.R.E. (1979). Hypomagnesaemia in diabetes. *Clinica Chemica Acta*, **95**, 235–42.

Maugdal, D.P., Ang, L., Patel, S., Bland, J.M. and Maxwell, J.D. (1985). Nutritional assessment in patients with chronic gastro-intestinal symptoms: comparison of functional and organic disorders. *Human Nutrition: Clinical Nutrition*, **39(C)**, 203–12.

Maxwell, A.E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, **116**, 651–5.

Maxwell, J.D., Patel, S.P., Bland, J.M., Lindsell, D.R.M. and Wilson, A.G. (1983). Chest radiography compared to laboratory markers in the detection of alcoholic liver disease. *Journal of the Royal College of Physicians, London*, **17**, 220–3.

Mayberry, J.F., Rhodes, J. and Newcombe, R.G. (1978). Breakfast and dietary aspects of Crohn's disease. *British Medical Journal*, **2**, 1401.

Meier, P. (1977). The biggest health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine. In *Statistics: a guide to the biological and health sciences* (ed. J.M. Tanur *et al.*). Holden-Day, San Francisco.

MRC (1948). Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal*, **2**, 769–82.

Norris, D.E., Skilbeck, C.E., Hayward, A.E. and Torpy, D.M. (1985). *Microcomputers in clinical practice*, John Wiley and Sons, Chichester.

Osborn, J.F. (1979). *Statistical exercises in medical research*, Blackwell, Oxford.

Pearson, E.S. and Hartley, H.O. (1970). *Biometrika tables for statisticians, volume 1*, Cambridge University Press, Cambridge.

Pearson, E.S. and Hartley, H.O. (1972). *Biometrika tables for statisticians, volume 2*, Cambridge University Press, Cambridge.

Pocock, S.J. (1983). *Clinical trials: a practical approach*, John Wiley and Sons, Chichester.

Prichard, B.N.C., Dickinson, C.J., Alleyne, G.A.O., Hurst, P., Hill, I.D., Rosenheim, M.L. and Laurence, D.R. (1963). Report of a clinical trial from Medical Unit and MRC Statistical Unit, University College Hospital Medical School, London. *British Medical Journal*, **2**, 1226–7.

Radical Statistics Health Group. (1976). *Whose priorities?*, Radical Statistics, London.

Reader, R. *et al.* (1980). The Australian therapeutic trial in mild hypertension: report by the management committee. *Lancet*, **1**, 1261–7.

Rose, G.A., Holland, W.W. and Crowley, E.A. (1964). A sphygmomanometer for epidemiologists. *Lancet*, **1**, 296–300.

Schapira, K., McClelland, H.A., Griffiths, N.R. and Newell, D.J. (1970). Study on the effects of tablet colour in the treatment of anxiety states. *British Medical Journal*, **2**, 446–9.

Schmid, H. (1973). Kaposi's sarcoma in Tanzania: a statistical study of 220 cases. *Tropical Geographical Medicine*, **25**, 266–76.

Seigel, S. (1956). *Non-parametric statistics for the behavioural sciences*. McGraw-Hill, Kagakusha, Tokyo.

Snedecor, G.W., Cochran, W.G. (1980). *Statistical methods*, 7th edn. Iowa State University Press, Ames, Iowa.

South-east London Screening Study Group (1977). A controlled trial of multiphasic screening in middle-age: results of the south-east London screening study. *International Journal of Epidemiology*, **6**, 357–63.

'Student'. (1908). The probable error of a mean. *Biometrika*, **6**, 1–24.

'Student'. (1931). The Lanarkshire milk experiment. *Biometrika*, **23**, 398–406.

Todd, G.F. (1972). *Statistics of smoking in the United Kingdom*, 6th ed., Tobacco Research Council, London.

Tukey, J.W. (1977). *Exploratory data analysis*. Addison-Wesley, Massachusetts.

Victora, C.G. (1982). Statistical malpractice in drug promotion: a case-study from Brazil. *Social Science and Medicine*, **16**, 707–9.

Weatherall, R. (1976). Recent seasonal patterns of infant mortality in England and Wales. In *Child Health: a collection of studies*, OPCS, HMSO, London.

Whittington, C. (1977). Safety begins at home. *New Scientist*, **76**, 340–2.

Index

This is a textbook in medical statistics for medical students, doctors, medical researchers, and others concerned with medical data. It should also be of interest to students of statistics who wish to learn about the practical application of statistical methods. It contains all the material required for a medical degree and for most post-graduate qualifications.

The fundamental concepts of study design and statistical inference are explained by illustration and example, and, for those who wish to go further, the mathematical background is also described. The material covered includes the design of clinical trials and epidemiological studies, summarizing and presenting data, probability, standard errors, regression and correlation, rank methods, measurement error, reference ranges, mortality data, vital statistics, and the choice of statistical method.

The book is firmly grounded in medical research and the interpretation of the results of statistical calculations is emphasized. All the data in the many examples are real, from the author's own research and statistical consultation or from the medical literature, to which reference is made where possible. There are 75 multiple-choice questions, with annotated solutions, and 15 exercises in study design and data analysis, with fully explained solutions.

Reviews

. . . a book which I think anyone teaching an introductory course in medical statistics should seriously consider as the main text to accompany their course. . . . It covers all the material which is likely to be needed at medical undergraduate level and for the various professional exams.

Statistics in Medicine

At last I have a book on medical statistics that I can safely recommend to my students!

Journal of the Royal Statistics Society

If you want to understand some of the statistical ideas important to medicine but fear being overwhelmed by mathematics you will welcome *An Introduction to Medical Statistics* by M Bland. . . . Altogether a useful introduction to medical statistics.

British Medical Journal

ISBN 0-19-261502-5

OXFORD UNIVERSITY PRESS