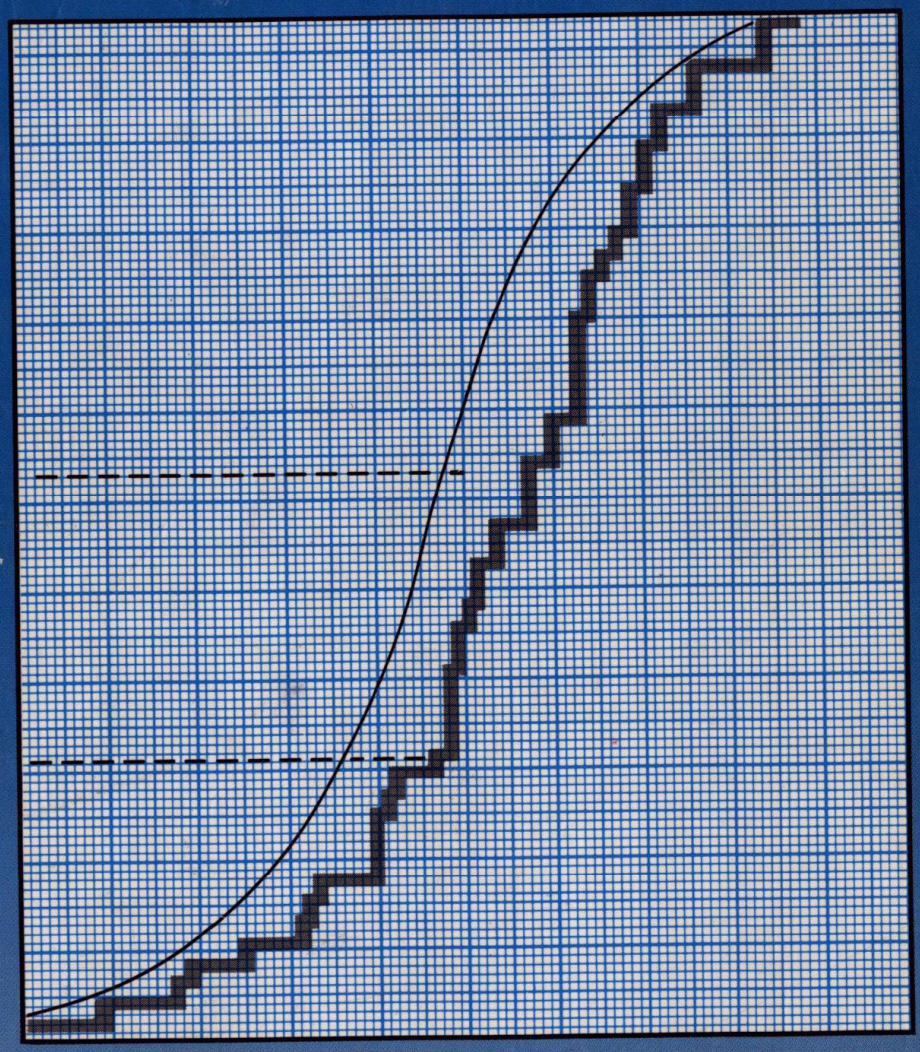
DISTRIBUTION-FREE TESTS



H.R. NEAVE & P.L. WORTHINGTON

Samsung v. Four Batons
IPR2025-00496
Exhibit 1006

Neave & Worthington

DISTRIBUTION-FREE TESTS

Math
QA278
.8
.N43
1988

HYMAN

This comprehensive treatment of distribution-free (sometimes called 'nonparametric') tests provides students in a wide range of subjects with an introductory text and reference book. These easy-to-use and widely applicable methods have made substantial inroads into popular statistical practice in recent years. Although several books are already available, none gives the inexperienced or nonspecialist user as thorough a coverage as does this book: it is more complete and detailed than the one or two brief volumes pitched at the same introductory level while being much more user-orientated and no less authoritative than the more plentiful books written primarily for the specialist statistician.

The book concentrates on tests and refrains from an extensive discussion of such nonparametric work as estimation and the formation of confidence levels and density estimation. Clarity of presentation is a key feature throughout, with the text divided into self-contained introductions to the techniques, including the full treatment of a numerical example, and, set in smaller type, 'optional extra' sections elaborating on particular problems or issues. Part I introduces the most essential ideas of probability & statistics needed for a proper understanding of the techniques introduced in the book. Chapter 2, in particular, deals with the crucial issue of hypothesis testing and provides the central foundation for the rest of the book. Parts II, III, IV and V deal respectively with tests suitable for when one sample, two samples, and more than two samples, are available for analysis. Most chapters present two tests (or procedures) for the same or similar situation; often one test is of a 'quick-and-simple' nature and the other is rather more sophisticated. Finally, Part VI contains a brief summary of additional tests and additional applications of tests introduced earlier, and concludes with some distribution-free methods of forming confidence intervals.

Not since Siegel's 1956 book has there been a user-orientated comprehensive treatment of distribution-free tests for the many students and researchers not specializing in statistics. Neave & Worthington should become widely used as both text and reference throughout the science community, especially among behavioural, social and management scientists.

Henry R. Neave is a Lecturer in the Department of Mathematics at the University of Nottingham; he is also Director of Research of the British Deming Association and he works as a statistical quality control consultant. He is well known as the author of *Statistics Tables* and *Elementary Statistics Tables* (both published by Unwin Hyman). Peter L. Worthington is Senior Lecturer in Statistics in the Department of Mathematics and Head of the Centre for Quality Improvement at North Staffordshire Polytechnic. His consultancy work involves statistical quality improvement programmes and Statistical Process Control training courses. He has been appointed the UK statistical consultant to an international company, and is a director of Quality Improvement Consultants Ltd.



ISBN 0-04-519019-4

9 780045 190195

ISBN 0 04 519019 4

THE PENNSYLVANIA STATE
UNIVERSITY LIBRARIES

MATHEMATICS LIBRARY

DISTRIBUTION-FREE TESTS

TITLES OF RELATED INTEREST

Advanced statistics
M. Hugill
Data construction in social surveys
N. Bateson
Documents of life
K. Plummer
Doing secondary analysis
A. Dale *et al.*
Elementary statistics tables
H. R. Neave
Field research
R. G. Burgess (ed.)
In the field
R. G. Burgess
Introduction to causal analysis
O. Hellevik
Modelling society
G. N. Gilbert
Nonsexist research methods
M. Eichler
Quantity and quality in social research
A. Bryman
Research design
C. Hakim
Research methods for elite studies
G. Moyser & M. Wagstaffe (eds)
The research relationship
G. C. Wenger
Social causality
J. Hage et al.
Social measurement and social indicators
M. Carley
Social science and social policy
M. Bulmer
Statistical concepts in geography
J. Silk
Statistics tables
H. R. Neave
The survey method
C. Marsh
Surveys in social research
D. A. de Vaus
Understanding maths and statistics in business
B. Edwards
The uses of social research
M. Bulmer

DISTRIBUTION-FREE TESTS

H. R. NEAVE
and
P. L. WORTHINGTON

London
UNWIN HYMAN
Boston Sydney Wellington

Exhibit 1006
Page 008 of 453

© H.R. Neave & P.L.B. Worthington, 1988
This book is copyright under the Berne Convention. No reproduction without permission. All rights reserved.

Published by the Academic Division of
Unwin Hyman Ltd
15/17 Broadwick Street, London W1V 3FP, UK

Unwin Hyman Inc.,
8 Winchester Place, Winchester, Mass. 01890, USA

Allen & Unwin (Australia) Ltd,
8 Napier Street, North Sydney, NSW 2060, Australia

Allen & Unwin (New Zealand) Ltd in association with the
Port Nicholson Press Ltd
60 Cambridge Terrace, Wellington, New Zealand

First published in 1988

British Library Cataloguing in Publication Data

Neave, Henry R. (Henry Robert) *1942–*
Distribution-free tests.
1. Distribution free statistical methods
I. Title. II. Worthington, P.L.
519.5
ISBN 0-04-519019-4

Library of Congress Cataloging in Publication Data

Neave, Henry R.
Distribution-free tests.

Bibliography: p.
Includes index.
1. Nonparametric statistics. 2. Statistical hypothesis testing. I. Worthington, P.L. II. Title.
QA278.8.N43 1988 519.5 88-17204
ISBN 0-04-519019-4 (alk. paper)
ISBN 0-04-519020-8 (pbk.). alk. paper)

Typeset in 10 on 12 point Times by Columns of Reading
and printed in Great Britain by Biddles of Guildford

It's Distribution Free

by G. E. P. Box

To be sung to the music of Gilbert and Sullivan's 'When The Foeman Bares His Steel'
(TR = Ta-ran-ta-ra)

It's distribution free! TR TR
Due to Mann and Whit-en-nee TR
And to them we give our thanks TR TR
As we fiddle with the Ranks TR
In complex calculations you need never be involved
A child of three could do it easily
But the thing we like the best
About this ad-mir-a-ble test
Is it's distribution free!

When your data's rather skew TR TR
And you don't know what to do TR
When decision is a must TR TR
And the calculator's bust TR
And although you're pretty sure they dropped the beaker on the floor
And the data isn't all it ought to be
Your alternative hypothesis will feel a rosy glow
Chosen by a test that's distribution free!

Oh it's distribution free! TR TR
(Well . . . I think it *ought* to be) TR
No problems need delay TR TR
We've assumed them all away TR
We declared our independence in this country long ago . . .
And isn't independence here the key?
So as they sound the bugle at the ending of the day
Be proud your test was distribution free!

When your alpha must be right TR TR
And your limits must be tight TR
If, with care, you rank the ties TR TR
So that none can criticize TR
Then whatever kind of errors you have made along the way
The first kind we will gladly guaranty
It is clearly evident
It's exactly five percent
If it's distribution free!

When you want a Ph.D. TR TR
Full of Mathematistry TR
In this you should invest TR TR
A NON-PARAMETRIC TEST TR
A thesis full of lemmas and theorems by the score
Your committee all will welcome that with glee
For you, their heads they'll bare
And create a special chair
If it's DISTRIBUTION FREE!

It's distribution free TR TR
It's not like Students' *t* TR
No Normality implied TR TR
Takes the Cauchy in its stride TR
You could have a distribution like a stalagmitic mass
(Inverted stalactitic, don't you see?)
And I very greatly doubt
That you'd be a fraction out
For it's Distribution Free!

When outliers abound TR TR
And John Tukey can't be found TR
Missing data everywhere TR TR
Here's a test beyond compare TR
This procedure's not most powerful, but we know how power corrupts
We therefore, rightly, shun this property
But we're going to mark the spot
Where this concept was begot
To be distribution free!

(With thanks to Professor G.E.P. Box)

Preface

We intend this book to be read (and used!) by people from a wide range of disciplines: the social sciences, business studies, sport and recreational studies, medicine, management science, engineering, science and mathematics. The examples and exercises have been carefully selected to show the versatility of distribution-free statistics in these various disciplines.

All the necessary statistical background material (including a very important chapter on the philosophy of statistical tests) is given, so that no previous knowledge is required. Furthermore, the presentation of the material in the chapters has been designed so that those readers who have studied mathematics to an elementary level can derive the same benefit as those having a deeper mathematical background.

We have taken care to develop the various ideas by intuitive reasoning so that the reader readily grasps 'what is going on'. In fact, the very nature of distribution-free statistics encourages such a common-sense approach to analysing statistical problems.

For the main procedures and tests in the book, we have provided BASIC computer programs. These will be of particular use to owners of 'home computers' and others who have access to a computer. However, many distribution-free tests are so easy to apply that usually all one needs is a pencil and paper!

We are grateful to Kathryn, Fiona and Richard Worthington for their assistance in the preparation of this book. Our thanks go to Jack Gillett for his helpful suggestions and to the staff of Unwin Hyman for their patience and forebearance during the development of this book.

Peter Worthington & Henry Neave

Contents

Preface *page* vii

How to use this book xv

Part I Essential background 1

1 Some Statistical Ideas 3
- Introduction 3
- Statistical populations 5
- 'Real-life' populations 6
- Measures of location 10
- Measures of dispersion 12
- Samples 15
- Theoretical distributions 18
- Random variables 23
- Some standard theoretical distributions 28
- The binomial distribution 29
- The normal distribution 34
- The normal approximation to the binomial distribution 41
- The chi-squared distribution 43
- A note on the use of the median instead of the mean 44

2 The Underlying Ideas of Statistical Tests 46
- Introduction 46
- Concepts of statistical testing 48
- One-sided and two-sided tests 54
- Best conservative critical values 58
- Power of tests 58
- Conclusion 61

Part II Tests involving one sample 63

3 One-sample location tests 65
- The sign test 65
- The Wilcoxon signed-rank test 69
- A note on the computer programs 75
- Assignments 76
- Program 3.1 The Wilcoxon signed-rank test 77

4 Goodness-of-fit tests 79
- Introduction 79

The chi-squared goodness-of-fit test 80
A note on small expected frequencies 84
Example on the normal distribution with mean and standard deviation specified 85
Example involving the normal distribution – mean and standard deviation not specified 88
Kolmogorov–Smirnov test 89
Discrete distributions 98
Lilliefors' test for normality 100
Assignments 103
Program 4.1 Chi-squared goodness-of-fit test 104
Program 4.2 Lilliefors' test for normality 105

Part III Tests involving two samples 107

5 Two independent samples – tests for differences in location 109
The Mann–Whitney test 109
Wilcoxon's version of the test 113
Tukey's quick test 119
Assignments 126
Program 5.1 Mann–Whitney two-sample location test 127

6 Two independent samples – tests for differences in spread 129
Introduction 129
The Siegel–Tukey test 130
A Tukey-type test for difference in spread 136
Assignments 139
Program 6.1 Siegel–Tukey two-sample test for spread 141

7 Two independent samples – tests for general differences 143
Introduction 143
Rosenbaum's test 144
The Kolmogorov–Smirnov test 149
Assignments 157
Program 7.1 The Kolmogorov–Smirnov two-sample test for general differences 158

8 Two related samples 160
Introduction 160
Wilcoxon's signed-rank test for matched pairs 161
The sign test for matched pairs 164
Assignments 165
Program 8.1 Wilcoxon's matched-pairs test 166

Exhibit 1006
Page 015 of 453

Part IV Correlation and regression 167

9 Correlation 169
Introduction 169
Spearman's rank correlation coefficient 169
Kendall's rank correlation coefficient 177
Further discussion on correlation 182
Rank correlation with non-numeric data 184
Assignments 186
Program 9.1 Spearman's rank correlation coefficient 187
Program 9.2 Kendall's rank correlation coefficient 188

10 Regression 189
Introduction 189
Theil's 'incomplete' method 192
Hypothesis test for the slope, *b* 198
The Iman–Conover method for monotonic regression 201
Assignments 206
Program 10.1 Theil's incomplete regression method 208
Program 10.2 The Iman–Conover monotonic regression method 210

11 Quick methods for correlation and regression 212
Introduction 212
The Olmstead–Tukey 'corner' test 212
Quick calculation of medians 219
Brown and Mood's quick regression method 220
A quick hypothesis test for the slope 224
A quick hyothesis test for the intercept 228
Assignments 230

12 Association between two variables 231
Introduction 231
The chi-squared test 232
A note on small expected frequencies 237
Yates' correction 238
Assignments 239
Program 12.1 Chi-squared test for association 240

Part V Tests involving more than two samples 241

13 More than two (*k*) independent samples 243
Introduction 243
The Kruskal–Wallis test 244
The median test 253

Multiple comparisons 256
Assignments 260
Program 13.1 The Kruskal–Wallis test 262

14 More than two (k) related samples 263
Introduction 263
Friedman's test 265
The match tests 273
Multiple comparisons 278
Assignments 281
Program 14.1 Friedman's test 282
Program 14.2 The match tests 284

15 More than two (k) samples – ordered alternative hypothesis 285
Introduction 285
The Terpstra–Jonckheere test for k (>2) independent samples 286
k (>2) related samples 291
Page's test 291
The match tests for ordered alternatives 295
Assignments 299
Program 15.1 Terpstra–Jonckheere test for ordered alternatives 300
Program 15.2 Page's test for ordered alternatives 301
Program 15.3 Match tests for ordered alternatives 302

16 Tests for interaction and Latin squares designs 304
Introduction 304
Interaction 304
Wilcoxon's interaction test 307
Match tests for interaction 310
Latin squares designs 314
Program 16.1 Match tests for interaction 319

Part VI Miscellaneous procedures 321

17 Tests based on runs 323
Introduction 323
Runs test for randomness 324
Runs up and down 328
The Wald–Wolfowitz two-sample runs test 332
Assignments 336

18 More tests using frequency data 337
Introduction 337
Fisher's exact test 338
Mood's two-sample median test 341
The Mood–Westenberg two-sample dispersion test 344

19 Confidence intervals 348
Introduction 348
Confidence intervals for the median 351
Confidence intervals for location differences 354
Confidence intervals for a matched-pairs difference 359
Assignments 363
Program 19.1 Confidence interval for difference between two samples 364
Program 19.2 Confidence interval for matched-pairs differences 365

Part VII Tables 367

Table A Probabilities and percentage points of the standard normal distribution 369
Table B Upper-tail probabilities for the chi-squared (χ^2) distribution 371
Table C Critical values for the sign test 372
Table D Critical values for the Wilcoxon signed-rank test 373
Table E Critical values for the Kolmogorov–Smirnov goodness-of-fit test 374
Table F Critical values for Lilliefors' normality test 374
Table G Critical values for the Mann–Whitney test 375
Table H Critical values for the Tukey test for location differences 377
Table I Critical values for the Tukey-type test for dispersion differences 379
Table J Critical values for the Rosenbaum two-sample test 383
Table K Critical values of the Kolmogorov–Smirnov two-sample test 387
Table L Critical values for Spearman's rank correlation coefficient 390
Table M Critical values for Kendall's rank correlation coefficient 391
Table N Critical values for the Kruskal–Wallis test 392
Table O Critical values for Friedman's test 395
Table P Critical values for the M_1 match test 396
Table Q Critical values for the M_2 match test 397
Table R Critical values for the Terpstra–Jonckheere test 398
Table S Critical values for Page's test 400
Table T Critical values for the L_1 match test for ordered alternatives 401
Table U Critical values for the L_2 match test for ordered alternatives 402
Table V Critical values for the S_1 match test for interaction 403
Table W Critical values for the S_2 match test for interaction 404
Table X Critical values for the runs test 405
Table Y Critical values for the runs up and down test 409
Table Z Critical values for Fisher's exact test 410

Outline solutions to assignments 413

References 425

Index 427

How to use this book

Naturally the authors would like to think that you will avidly read this book with all-consuming interest from cover to cover. However, we are realistic enough to appreciate that some readers will have neither the time nor the inclination to manage that. So here are a few comments about the structure of the book which may help you to gain the most benefit if you only have time for an abbreviated reading.

If you are essentially a beginner at statistics, having had either no formal training or just a brief introductory course in the subject, you should take the time to read carefully through Part I. This is an introduction to the most essential ideas of probability and statistics which will be needed for a proper understanding of the techniques described in the rest of the book. On the other hand, if you already have a sound basic knowledge of the subject, you could just skim through the pages of Chapter 1, making sure that everything covered there looks reasonably familiar. However, we do strongly recommend that you carefully read Chapter 2 on statistical testing, even if you already know something about this important area of statistics. The reason is that the main ideas of hypotheses testing are often badly taught and, as a result, misunderstood – sometimes by people who have been using them for years. To ensure you do not fall into that trap, please read Chapter 2, no matter what is the level of your previous experience, as it is the foundation on which virtually all the rest of the book is built.

Parts II, III, IV and V deal, respectively, with tests suitable for when one sample, two samples, association between two samples and more than two samples (or data of a more complex form) are available for analysis. Most chapters present two tests (or procedures) for the same or similar situations; often one test is of a 'quick and simple' nature and the other is rather more sophisticated. You should not try to read about the second test in a chapter before reading about the first, as the practical situation and illustrative numerical example are usually the same for both tests and so are only described once. On the other hand, you do not have to read chapters in their entirety. The coverage of most of the tests and other techniques is divided into normal-type and small-type sections. The normal-type part provides a fairly self-contained introduction to the technique, including the full treatment of a numerical example, and may well be quite sufficient for a first reading. The small-type sections contain 'optional extras', which you may dip into as and when you wish. These include such topics as the treatment of tied observations, adaptation of the tests for one-sided or two-sided alternatives, some basic theory, and what to do if the samples are so large that the tables in the back of the

book or elsewhere are insufficient. All of these aspects are important for full understanding of a particular technique; however, it will probably be sensible for you to skip, at least initially, most or all of the small type, thus giving you more time to read about a wider variety of the methods contained in the book and so getting a better 'feel' for the ideas of distribution-free statistics.

Finally, Part VI contains three chapters. The first two contain a brief summary of some other tests for situations met in earlier chapters, and also some applications to new situations of tests covered in the main part of the book. The other chapter describes some distribution-free methods of forming confidence intervals; these methods are quite interesting but often rather cumbersome in practice, and so have not been included earlier in the book.

One problem about a book of this kind, which covers a wide variety of statistical methods, is that a very substantial part of it can be taken up with tables. However, one of the authors has separately produced a set of tables (H. R. Neave's *Elementary statistics tables*, Allen & Unwin, 1981), which gives very comprehensive coverage of many of the topics contained in this book. As these tables are now in wide circulation, particularly as the standard book of tables for new statistics courses in the Open University, it seemed unnecessarily extravagant to reproduce them here. Accordingly, any required table that is included in *Elementary statistics tables* (indicated henceforth by *EST*) appears only in an abbreviated form in this book. Page references to *EST* are given throughout this book wherever appropriate.

Exhibit 1006
Page 021 of 453

PART I

Essential background

1
Some statistical ideas

Introduction

In this chapter we provide the background for the main business of the book. We do not intend to give a comprehensive treatment of all aspects of elementary statistics and probability; we shall just concentrate on what is required for understanding the procedures in the rest of the book. No doubt some readers will be quite familiar with many of the ideas presented here – nonetheless we do recommend such readers to glance through this chapter to see which main concepts of basic statistics are most relevant to distribution-free methods.

To set the scene, let us first look at some situations that are typical of those often encountered by statisticians.

SITUATION 1

A dietician ponders about the effectiveness of a new weight-reducing diet. The weights (kg) of 10 patients were recorded before and after use of the diet:

Patient	A	B	C	D	E	F	G	H	I	J
before	98.0	105.0	107.3	99.0	84.0	102.2	115.1	103.4	119.3	116.7
after	96.0	101.5	106.9	97.1	84.5	98.0	113.7	103.4	112.0	117.0

We ask: is there sufficient evidence here to support the claim that the new diet is effective in reducing weight?

SITUATION 2

The contents of a family-size packet of Corn Flakes are quoted as weighing 750 g. However, because of a suspected malfunction in the machinery, the management are concerned that the average weight might be less than 750 g. Sixteen packets are weighed giving the following results:

749.3	754.4	748.3	756.4	746.4	752.7	752.2	749.1
757.4	758.8	747.1	748.6	753.8	749.1	750.5	754.3

We ask: is there evidence to support the management's concern?

SITUATION 3

As part of a survey on shopping habits, information is sought on the number of items purchased by customers in supermarkets and local shops. On a particular day, the following data were collected:

supermarket	10	23	13	27	36	5	20	15	42	21	39
local shop	11	2	9	15	4	12	13				

We ask: what can we conclude about customers' purchasing habits?

SITUATION 4

A market-research organization seeks information on consumer preferences regarding the colour of bread wrappers. Six people are asked for their opinions on four different coloured wrappers. (In practice, of course, the number of people questioned would be far greater.) The information gathered is shown below, where 1 corresponds to the preferred choice, 2 to the second choice, and so on.

Consumer	Red	Blue	Yellow	Green
A	3	1	2	4
B	1	2	3	4
C	3	1	4	2
D	1	2	3	4
E	2	1	3	4
F	2	3	1	4

We ask: is there evidence of any colours being preferred or disliked compared with others?

Let us now spend a few moments discussing the different types of data featured in the above situations.

In situations 1 and 2, the data were *measurements* of something (weights in both of these cases). In other situations, measurement data might be, for instance, volumes, or lengths, or heights, or densities, or times, or tempertatures. Such data can (if measured precisely) generally take any values within certain limits; thus, for example, the weights of the dieters could be anywhere between, say, 28 kg and 140 kg. Such measurement quantities are said to be *continuous* variables because their values can be anywhere within a *continuous* range.

In contrast to this, consider situation 3, where the variable being recorded was the number of purchases. Here the values were obtained by

counting and so they could only be *whole numbers*, (i.e. *integers*). Other such integer variables are (a) the number of teeth in a cat, (b) a team's baseball scores, (c) the number of cups of coffee sold at Heathrow Airport on a particular day, (d) the number of times the word 'Bond' occurs in the James Bond novels and (e) the number of fleas on your pet dog. These are typical of what are known as *discrete* variables, where the value of the variable is usually obtained by counting something*.

In situation 4, the data just consisted of preferences, and so were not obtained by either accurate measuring or counting. Such data, although possessing much important information about what has been sampled, are obviously less detailed and less precise than the previous types of data. A big advantage of distribution-free techniques is that they can easily cope with such less-detailed data, unlike more traditional statistical methods. Other instances where preference data can arise are in (a) the grading of wine from different regions, (b) the weekly top-ten pop records and (c) judging the Miss World contest. The essential feature is that we are able to list the data in some kind of order – a process commonly known as *ranking*. Thus, for example, a judge testing four wines (A, B, C and D) might grade them as: 'A is the best, C is next best, D is third best, and B is worst'. To these grades we could assign the numerical ranks 1, 2, 3 and 4; so we would say that A has rank 1, B has rank 4, C has rank 2 and D has rank 3. Note that we do not attempt to attribute any actual 'marks' or 'scores' to the items being compared. As long as we can rank them, we have enough information to use distribution-free techniques. (Again, this is something of a simplification; a few distribution-free techniques do need numerical information. However, the large majority do not and, in these cases, numerical information has to be reduced to ranked data in order to use the techniques.)

Before beginning to analyse problems such as these, we will first discuss some basic statistical ideas which will help your understanding of what follows later.

STATISTICAL POPULATIONS

The 'population' is the collection of things about which our investigation and analysis are designed to provide some insight and information. More specifically, we may be concerned with the way in which a particular

* This is only a very restricted idea of 'discreteness', but it covers the majority of applications. The general meaning of discreteness is that there are definite gaps between the possible values, unlike *continuous* variables whose values may be arbitrarily close together (apart from any limitations imposed by the imprecision of the measuring device). So, for example, the variable that represents the proportion of heads when 10 coins are tossed (the possible values being 0, 0.1, 0.2, . . . , 0.9 and 1) is *discrete*, although it takes some non-integer values.

quantity or characteristic *varies*, i.e. is *distributed*, in the things comprising the population (for example, in the above situations, weights or numbers of items bought or opinions concerning colours of wrappers). We attempt to obtain information about such *distributions* by analysing *samples* drawn from these populations.

For our purposes, there are two meanings to the term 'population'. First, we have populations that arise directly from practical situations, such as the one dissussed below. Such populations may be called 'real-life' populations. We also need 'ideal' or 'model' populations, whose distributions may not really exist in real life but which nonetheless are used by statisticians to model real-life population distributions. Later in this chapter we shall examine the most important features of some of these ideal population distributions.

'REAL-LIFE' POPULATIONS

To a lay person, the word 'population' usually means the inhabitants of a region or country. However, statisticians impart a much broader meaning to the word, so that it does not have to relate just to groups of humans (or animals). In its most general form, a population is the collection of *all* items under investigation. More specifically, it is the collection of the values of the characteristic of those items that we wish to study.

Examples of such populations could be:

(a) The average petrol consumptions of Ford automobiles manufactured in Great Britain last year.
(b) The yield (kg) of tomatoes from each plant of a certain variety.
(c) The current ages of everybody living in Great Britain.
(d) The weight of contents of nominally 750 g packets of Corn Flakes filled by a particular machine.

At this point, several words of caution are appropriate. As we indicated above, a population is ideally a specific collection of 'things' on which unambiguous and relevant measurements or counts can be made on the characteristics of interest. It is also explicitly or implicitly assumed that so-called 'random samples' (see later in this chapter) can be taken of these things or their values. Although many teachers of statistics and books on the subject seem to assume (apparently, with scarcely a second thought) that the world is like that, a few moments of mature consideration will reveal that this is often not the case. For instance, consider the above four examples.

(a) *The average petrol consumptions of Ford automobiles manufactured in Great Britain last year*

How is 'average petrol consumption' measured? What is meant by that expression? Perhaps we are referring to one of the standard Government or

consumer-organization tests; but which one? Is the consumption at a steady 56 mph or at a steady 70 mph or under 'simulated city driving' (whatever that is)? How about some real consumption figures from actual driving? But that depends very much on who is driving, how they are driving and where they are driving. Maybe we *want* to take such features into account and maybe we do not. Maybe Ford are a particularly popular choice as company cars, which might imply a rather large proportion of high-speed driving. How is that to be taken into account in comparing consumption figures for other makes of car (assuming we want it to)?

Let us turn to the other part of the definition of this population: 'manufactured in Great Britain last year'. This sounds simple enough – but wait! These days, different parts of cars are made in different companies and countries. Also, assembly may take place in still further locations. Thus it is difficult to define exactly what is meant by 'Made in Great Britain'. Perhaps it does not really matter, in which case why include it in the description of the population?

Even 'manufactured . . . last year' is hard to define. Different components are made at different times and different sections are assembled at different times. Perhaps we mean 'completed last year'. But when is a car 'complete'? Is it when it is painted or when its accessories are incorporated or when it is delivered to a distributor or when it is sold or when it is registered.

It may even be difficult to define what is meant by 'Ford'. There are various companies (e.g. Iveco-Ford and Ford New Holland) which are only partly owned by Ford, and that status may change during the year. Do we involve them in our study?

All these may be irritating questions, but until they are answered, how can we know which vehicles we should regard as candidates for our sample?

(b) *The yield (kg) of tomatoes from each plant of a certain variety*

In some respects, this is even more difficult to pin down than the Ford cars in (a). At least in that case there was an attempt (albeit rather unsuccessful) to define the population in terms of time and location. Here there is no such attempt. For one thing, this 'population' is forever increasing, at least until this variety of tomato plant becomes extinct. Also the yield may change over time because of changes in growing techniques. Even at any fixed time, yield (however one attempts to define it at a 'fixed time') will depend on a number of factors, such as location, climate, soil preparation, frequency and amount of watering, frequency of picking, etc. To put it mildly, it will be rather difficult to define what we mean by the selection of a random sample from this 'population', let alone actually to take it!

(c) *The current ages of everybody living in Great Britain*

You should by now be getting the drift of our argument, so we will make our discussions on both (c) and (d) quite brief. In (c) the difficulty is that the population of individuals is perpetually changing since rather a lot of babies are born and lots of people die each day. In any case, can we define exactly who lives in Britain and who does not? What about dual-nationals, or those who are in the process of emigrating, or those 'of no fixed abode', or those whose life is mostly spent at sea? As regards the actual ages to be recorded, there is also the little problem that everybody is continually growing older!

(d) *The weights of contents of nominally 750 g packets of Corn Flakes filled by a particular machine*

The particular problems here are that (as in (b)) the number of things (packets) to be included in the population keeps on growing, and that the contents may depend very much on factors such as who is operating the machine, the machine settings used, and so on. Even for a given collection of flakes, the total weight will vary according to, say, humidity of the environment, which affects moisture content.

So, in summary, there can be (and usually are) many problems associated with the basic conditions assumed for statistical sampling and inference procedures. The population of 'things' may be continually changing, with extra items being created and, possibly, others dropping out; consequently, in practice, a 'random sample' is difficult to define, let alone take. Further, the characteristic(s) being measured may be continually changing for any given member of the population. Also, they may be very dependent on extraneous factors, as opposed to being a property of the population member itself.

These problems are usually impossible to solve completely. Therefore, the best that the able statistician can do, if he is going to use standard statistical inference techniques, is to be continually aware of such problems, to evaluate their seriousness, and to alleviate their effects as far as possible. He takes particular care to reduce looseness in the definitions as far as possible, and to define measurements on population items which are meaningful and informative. Also, if populations are to be compared with each other, he tries to ensure that these comparisons are as 'fair' as is feasible. He will also 'grasp the nettle' of saying that standard statistical techniques are inappropriate and should not be carried out when conditions are too far from what the theory assumes. In particular when either the population, or the measurements of the characteristics of interest, are varying in an unstable and unpredictable way, a sampling experiment carried out at the present time may have little bearing on the future, and may mislead. Indeed, the concept of stability is paramount for the use of the techniques covered in this book, and for most other statistical procedures as well; after all, if the subjects of our study are unstable over time, how can the 'now' tell us anything about what is to come?

All the examples and illustrations in this book are to be read with the understanding that these matters have been fully considered, and the researchers are satisfied that the situations are sufficiently stable for inference techniques to be meaningful.

One type of situation that is relatively 'safe' as regards the problems we have been discussing is anything based on census data. The population is usually fairly clear-cut and stability problems are not normally relevant since census studies, by definition, relate to a fixed time or time period, and do not pretend to forecast the future.

Accordingly, in order to illustrate a few ideas about populations, we have chosen data collected by the Office of Population Censuses and Surveys. The data relate to numbers of live births in England and Wales

during 1980. The table below gives information on the distribution of births according to the age of the mother:

Age of mother (years)	Number of births
15 to 19	60 754
20 to 24	201 541
25 to 29	223 438
30 to 34	129 908
35 to 39	33 893
40 to 44	6 075
45 to 49	625
Total	656 234

Such a table is often called a *frequency distribution* table since, as in this case, it shows how the frequencies of births are *distributed* according to the age of the mother. From it, we can glean a lot of information about the population. For instance, only about 1% (i.e. 6075/656 234 + 625/656 234 = 0.0102) of all births were to mothers aged 40 or above. About 9% (60 754/656 234 = 0.0926) of the mothers were under 20 years old. Also, not surprisingly, the majority of the births were to women aged between 20 and 35. This type of information can be nicely illustrated in a picture called a histogram, in which the frequencies are represented diagrammatically by rectangular blocks. A histogram for the births data is shown in Figure 1.1.

Our statement about the majority of births being to women between 20 and 35 years can now be seen 'at a glance'. A particularly important feature of the distribution is also immediately revealed by the histogram, namely its lack of symmetry. Indeed, the actual shape of the distribution agrees with common sense – most births are to mothers in their 20s, and the numbers decline steadily but not abruptly as age increases. The shape



Figure 1.1

of the distribution would only be symmetric if birth rates rose in the first half of the maternal age range in a similar pattern to which they decline in the second half. There is really no reason why this should be so, and yet many traditional statistical techniques depend on distributions being symmetric. Of course, in other countries and cultures, the shape of the distribution of ages might be quite different, though it is unlikely to be symmetric!

In addition to extracting information regarding proportions and shape of the distribution, it is also useful to obtain some *numerical measures* that summarize the main features of the population. The two typical numerical measures of greatest importance are: (a) *averages*, which, in some sense, locate the 'centre' of the distribution – such averages are therefore called *measures of location*; and (b) measures of *dispersion*, which indicate, for example, whether the majority of the population are closely clustered around the average or are more widely dispersed.

Measures of location

There are many ways of measuring the 'centre' or 'average' of a population. The two most popular measures are the *mean* and the *median*.

The *mean* of a population (usually denoted by the Greek letter (mu) μ, pronounced 'myew' as in 'music') is equal *to the sum of all the values divided by the size of the population*. For the births data we obtain

$$\mu = \frac{17\ 521\ 448}{656\ 234} = 26.7 \text{ years}$$

Although the mean is a much-used measure of location, it does suffer some disadvantages. We will go into details later; suffice to say here that some of these disadvantages make it quite unsuitable for most distribution-free work, and so therefore we will generally use the median as our location measure in this book.

The *median* (which we will denote by the Greek letter (phi) ϕ)* is very simple in concept: it is just the middle value of the population in the sense that half the population lies below it and half above it. This is, of course, an eminently reasonable measure of location; in fact, many would say that it is a more natural measure than the mean – it certainly satisfies the criterion of being 'central' more obviously than does the mean. If the population is not too large, then the median is easily obtained after simply arranging the numbers into ascending order; for example, the numbers 4, 8, 7, 5, 8, 5 and 6 are arranged as

* We have chosen to represent the median of a population by ϕ as this represents a population split into two equal parts – which is just what the median does!

4 5 5 6 7 8 8

from which we clearly see that the median ϕ is equal to 6.

Should there be an even number of observations, for example, 5, 5, 7, 7, 8, 8, 8, and 11, then the median is taken to be halfway between the middle pair, i.e. $\phi=\frac{1}{2}(7+8)=7.5$. However, most populations are too large for this simple technique to be suitable and a more appropriate method of obtaining the median is from the so-called *cumulative distribution function* (cdf). First, the *cumulative* frequencies are calculated by simply summing successive frequencies in the frequency table. A table of cumulative frequencies for the above data on maternal ages is given below:

Age of mother (years)	Frequency	Cumulative frequency	Percentage cumulative frequency
15 to 19	60 754	60 754	9.26
20 to 24	201 541	60 754+201 541 = 262 295	39.97
25 to 29	223 438	262 295+223 438 = 485 733	74.02
30 to 34	129 908	485 733+129 908 = 615 641	93.81
35 to 39	33 893	615 641+33 893 = 649 534	98.98
40 to 44	6 075	649 534+6 075 = 655 609	99.90
45 to 49	625	655 609+625 = 656 234	100
Total	656 234		

You will see that we have added a final column of percentage frequencies; these will make the identification of the middle value (the 50% value) easier. Before we calculate the median using this table, it is worth observing that facts concerning proportions of births in the different age groups, such as those illustrated above, are very readily computable from this cumulative distribution function. Thus, we immediately see that 9.26% of the mothers were in their teens, that (100−98.98)%, or about 1%, were in their 40s and that (93.81−9.26)% = 84.55% (i.e. the large majority) were between 20 and 34 years old inclusive. We mention this importance of the cdf in readiness for its use in later chapters. Once the cumulative distribution table has been obtained, a close approximation to the median can be easily obtained by drawing a graph of the cdf. Figure 1.2a shows the correct representation of the cdf of this population while Figure 1.2b shows the sort of graph that is usually drawn when one wants to obtain an estimate of ϕ by locating the 50% point.

In Figure 1.2b, the broken line 'shows the way' and gives a median value for the population of about 26.5 years. Of course, this estimate of the median is not absolutely accurate, but will generally be good enough

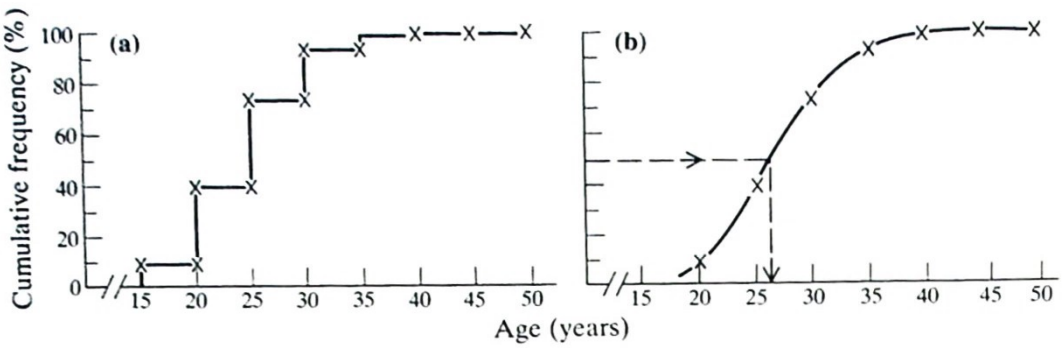


Figure 1.2

'for most practical purposes' – and is certainly to be preferred to putting 656 234 items into ascending order!

Measures of dispersion

The second vitally important characteristic of a population is the *dispersion* (or *spread*, or *variability*) of its members. It might be that the members of a population are tightly bunched, in which case there is little dispersion in the population (e.g. see Fig. 1.3a). Low dispersion is most desirable in many populations; for instance, those relating to the manufacture of scientific instruments, the weights of foodstuffs, filling operations – indeed, virtually every kind of measurement on virtually every production process. On the other hand, large variability, or dispersion, between items in any such situations (see Fig. 1.3b) may indicate faulty machinery, slap-dash methods, erratic quality of raw material, etc. – which certainly would cause displeasure to the customer. In other contexts, high dispersion of a population might be an entirely expected and natural phenomenon, such as, for example, with the weights or heights of a human population.

For the purposes of this book, it will be sufficient for the reader to understand the general ideas of dispersion in populations, rather than becoming deeply involved in calculating values of particular measures of dispersion. However, we should briefly say something about the *standard deviation* of a population, which is the quantity most commonly used in classical statistics to measure dispersion.

As we shall see shortly, the standard deviation does not arise in a particularly obvious manner. One 'natural' measure of dispersion is the average of the differences between the values of the population and the 'centre' of the population. Now, in classical statistics, 'averages' and 'centres' are usually described in terms of means and so, if a population has N members (denoted by $X_1, X_2, X_3, \ldots, X_N$), a 'natural' measure of dispersion is

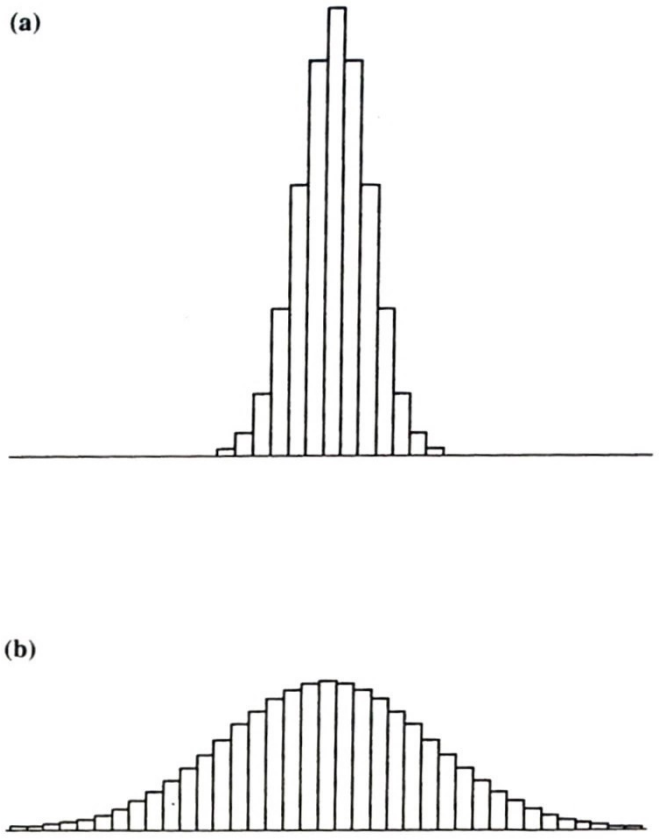


Figure 1.3

$$\frac{|X_1 - \mu| + |X_2 - \mu| + |X_3 - \mu| + \ldots + |X_N - \mu|}{N}$$

where the quantities $|X_1 - \mu|$, $|X_2 - \mu|$, etc. are the distances of each member from the mean μ of the population. (The vertical bars denote what are called 'absolute values' or 'sizes'. So, for example, if $\mu=10$ and $X_1=13$ then $|X_1-\mu| = |13-10| = 3$; and if $X_2 = 7$ then $|X_2-\mu| = |7-10| = |-3| = 3$. Both of these sizes are 3, since 13 and 7 are each 3 units from $\mu=10$.) The above measure of dispersion is sometimes referred to as the 'mean absolute deviation' and is often abbreviated by MAD. By using some standard mathematical notation, the above expression for the MAD can be abbreviated to

$$\text{MAD} = \frac{\sum_{i=1}^{N} |X_i - \mu|}{N}$$

where $\sum_{i=1}^{N}$ is a summation notation which stands for 'in the expression that follows, set i equal to 1, then 2, then 3, and so on up to N, and then add up (sum) all those expressions obtained'. You may like to check that

this 'short-hand' version of the MAD expression is indeed equivalent to the previous 'long-hand' version.

Unfortunately, mathematical statisticians have never been too fond of the MAD; the absolute values $|X_i-\mu|$ are a bit awkward to handle algebraically. But the use of absolute values in the MAD expression is necessary in order to prevent positive and negative differences (deviations) cancelling each other out. (In fact, it is quite easy to show that $\sum(X_i-\mu)$ is always zero, and so it could never form the basis for a measure of anything!)

However, there is an alternative way of preventing this – and that is to square the deviations. The resulting measure is

$$\frac{(X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_N-\mu)^2}{N}=\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$$

This measure is known as the *variance* of the population. However, this is really not at all a proper measure of dispersion. For example, if you are comparing two populations where the deviations in one are three times as big as in the other, then you will find that the variance of that population is nine times as big as the other. This we can easily see from the following two populations:

Population 1	X_i	7	8	10	15	(μ=10)
	$X_i-\mu$	−3	−2	0	5	

variance = (9 + 4 + 0 + 25)/4 = 38/4

Population 2	X_i	1	4	10	25	(μ=10)
	$X_i-\mu$	−9	−6	0	15	

variance = (81 + 36 + 0 + 225)/4 = 342/4 = 9 × 38/4

Similarly, if the deviations were four times as big, the variance would be 16 times that of the other. The remedy is simple – we take the *square root* of the variance to give a valid measure of dispersion; this overcomes the problem just mentioned (check this out with the above example). It is this measure that is known as the *standard deviation* of the population. The standard deviation is usually denoted by the Greek letter (sigma) σ. Thus we have

$$\sigma=\sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$$

The standard deviation of the ages of mothers is equal to 5.4 years. A much smaller value would have indicated that most of the ages were

closer to the mean value (so there would not be much dispersion); while a much greater value (for example, $\sigma=10$) would have indicated that the ages were spread more evenly and broadly over the possible range of 15 to 50 years, for example.

SAMPLES

More often than not, it is impracticable for us to obtain the various population features we have mentioned: the histogram, the median, the mean and the standard deviation. This is likely to be because of the high cost or sheer impossibility of measuring or observing every member of a large population. So we are obliged to take a *sample* from the population in the hope that the sample characteristics (sample median, sample mean, etc.) will be close to the corresponding population characteristics. Indeed, this is really what much of the subject of statistics is all about: we are finding out things about populations by investigating samples taken from them.

To illustrate a few of these sample characteristics, we have taken a random sample of 40 from the population of mothers. (The word 'random' implies that all possible selections from the population are equally likely.) The ages of the 40 mothers (measured to the nearest tenth of a year) were:

32.1	29.8	29.3	24.9	29.5	30.8	23.5	26.0
23.1	22.3	27.0	25.8	28.6	17.6	15.0	26.8
28.9	34.9	31.2	25.8	30.5	34.2	27.2	27.1
34.3	22.4	35.1	15.2	31.7	21.5	18.2	30.1
25.6	27.3	23.7	31.9	20.9	26.4	23.4	18.8

It is particularly interesting to calculate the various sample characteristics in this case, since we will be able to compare them with the known corresponding population values. First let us construct the sample histogram. To help with this, we have formed the frequency table using the same intervals as for the population:

Age of mother (years)	Frequency	Cumulative frequency	Percentage cumulative frequency
15–19	5	5	12.5
20–24	9	14	35.0
25–29	15	29	72.5
30–34	10	39	97.5
35–39	1	40	100
40–44	0	40	100
45–49	0	40	100

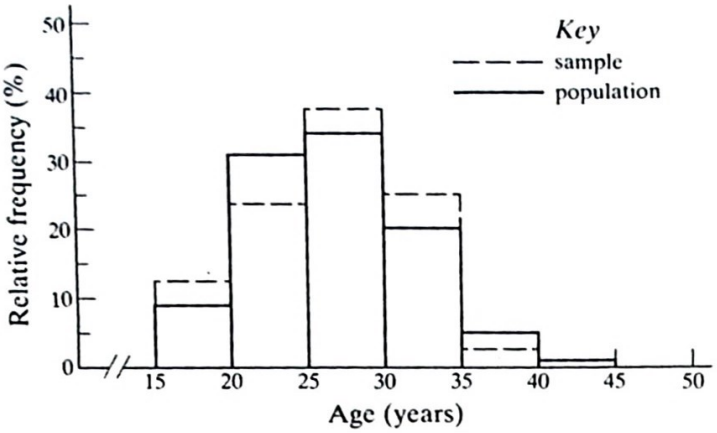


Figure 1.4

Figure 1.4 shows the sample histogram superimposed on the population histogram. Similarly, Figure 1.5 shows a graph of the sample cdf superimposed on a graph of the population cdf (note that in both cases we have drawn the *true* graphs of the cdfs rather than the smoothed versions such as we drew in Figure 1.2b in order to obtain a value for ϕ).

The notable feature of these figures, particularly when we remember how tiny is the sample compared to the size of the population (40 compared to 656 234), is how well the pictures for the sample reflect the corresponding population pictures. Let us see whether the numerical measures for the samples similarly bear any resemblance to those for the population.

The sample *mean*, which we will denote by $\bar{x}$, is calculated using the analogous formula to that for the population, namely

$$\text{sample mean} = \frac{\text{sum of all the values}}{\text{number of values}}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where we have denoted the n values in the sample by $x_1, x_2, x_3, \ldots, x_n$. Using this formula on our sample gives

$$\bar{x} = \frac{1058.3}{40} = 26.5 \qquad \text{(to one decimal place)}$$

The sample *median*, which we will denote by $\tilde{x}$, is the halfway point in the sample, which here, with a sample of 40, is halfway between the 20th and 21st members when the data have been ordered. A clever way to

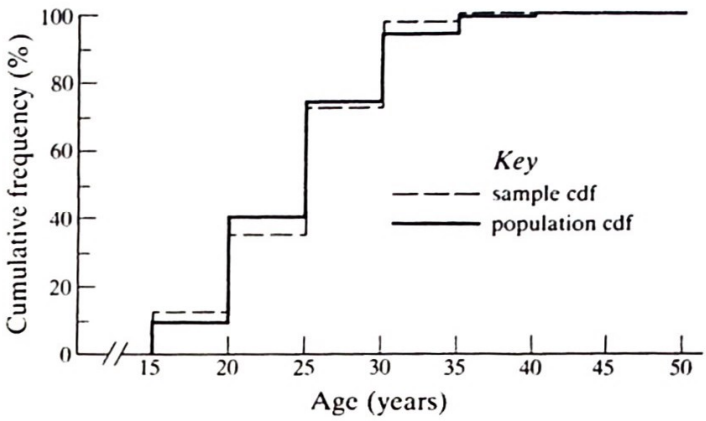


Figure 1.5

reduce the labour involved is to note from the cumulative frequency table that these two values must be in the age group '25 to 29', and so in fact we need only sort through the 15 values in this group. Doing this, we find that the 20th and 21st values are 26.8 and 27.0 respectively, giving

$$\tilde{x} = \tfrac{1}{2}(26.8 + 27.0) = 26.9$$

The standard deviation (sd) of the sample is also calculated in a similar fashion to its population counterpart. The formula for sd is

$$\text{sd} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

and for our sample this produces sd = 5.15. These sample statistics are remarkably close to their population counterparts as the following table shows:

	Mean	Median	Standard deviation
population	$\mu = 26.7$	$\phi = 26.5$	$\sigma = 5.4$
sample	$\bar{x} = 26.5$	$\tilde{x} = 26.9$	sd = 5.15

Of course, one cannot depend on the sample statistics always being as close as this to their population counterparts. However, with a sample as large as 40, it is quite usual for them to be fairly close. Indeed, classical statistical methods very largely depend on this being the case.

We have calculated the sample statistics in this instance for interest only, since we already know the population equivalents. In real-life situations, these population 'parameters' are unknown, and the sample statistics are used to provide information about them – this example

shows that they can be quite effective in so doing.

Although in classical statistics the sample mean $\bar{x}$ is used to estimate μ, the sample standard deviation sd is not generally used to estimate σ. The reason for this is that sd tends to under-estimate σ, particularly with small samples. The remedy for this problem is to use a modified version of the sample deviation, which is defined by

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

i.e. s is the same as sd with the n in the denominator replaced by $n-1$. So whenever we need to estimate σ, we shall invariably use s.

THEORETICAL DISTRIBUTIONS

Census data provide us with the relatively unusual situation of having comprehensive information about 'real-life' populations. In particular, this permits accurate evaluation of their important numerical characteristics. In most other cases, it is either too difficult or too expensive to obtain such complete data, and so then we need to sample the populations in order to find out something about these characteristics. To do this, we need knowledge of some common theoretical distributions which have proven good at modelling 'real-life' distributions. Fortunately, some of these distributions are also very useful in dealing with the theory of many of the test procedures we shall examine throughout the book.

Before looking at these theoretical distributions, we shall use the distribution of maternal ages to help introduce us to some of the concepts of theoretical distributions and also to some notions of probability which will be important later.

What information do probabilities give us? The probabilities of events are numerical assessments of the chance or likelihood of these events occurring. To begin with extreme cases, an impossible event is said to have the probability of 0 of occurring, while an event that is certain to occur is said to have a probability of 1. Usually, the probability of an event occurring lies somewhere between these extremes; a probability close to 0 (say, less than 0.05) indicates that there is only a small chance of the event happening and a probability close to 1 (say, greater than 0.95) indicates that there is a very strong chance of the event occurring. A probability of $\frac{1}{2}$ indicates an even (or '50–50') chance of the event happening; a probability below $\frac{1}{2}$ indicates that the odds are against the event happening, while a probability above $\frac{1}{2}$ indicates that the event is 'odds on'. A probability of $\frac{2}{3}$ indicates that, in the long run, the event happens '2 times out of 3', and a probability of 0.9 indicates that in the

long run the event happens '9 times out of 10'.

With these thoughts in mind, let us now look again at the distribution of maternal ages. Suppose that these maternal ages are just part of the details held on a computer filing system in a regional health authority. Then, when we speak of selecting a mother 'at random' in a particular age band, we are referring to selecting the appropriate computer reference rather than the actual mother!

For the moment, our ideas will be based on the histogram (Fig. 1.1) for this population, and we reproduce this in Figure 1.6 with the frequency scale replaced by a relative frequency scale (which effectively goes from 0 to 1 instead of from 0 to $N = 656\,234$), since this is the most helpful scale for our purposes. For convenience, we have also indicated the actual relative frequency on each block. From this figure we see that the proportion of ages over 40 is $0.009+0.001 = 0.010$, i.e. 1%. We can interpret this result as being the chance that when a mother is selected at random from the population, she was of age at least 40 when her child was born (obviously, quite a small chance). As a further example, consider the probability of selecting a mother whose age at the birth was between 20 and 29 years inclusive. From the histogram in Figure 1.6 we see that the proportions for the two blocks covering this age range are about 0.31 and 0.34, and so the probability of selecting a mother in the age band 20–29 years is about $0.31+0.34 = 0.65$, which indicates nearly a 2 in 3 chance of selecting someone from the population in this age range.

Now, of course, the sum of the proportions for all the individual classes is 1, and so the histogram in Figure 1.6 is showing us how the total probability of 1 is distributed over the various age groups, and this immediately gives us what is known as the *probability distribution* of the ages. Sometimes it is more convenient to display the probability distribution as a table:

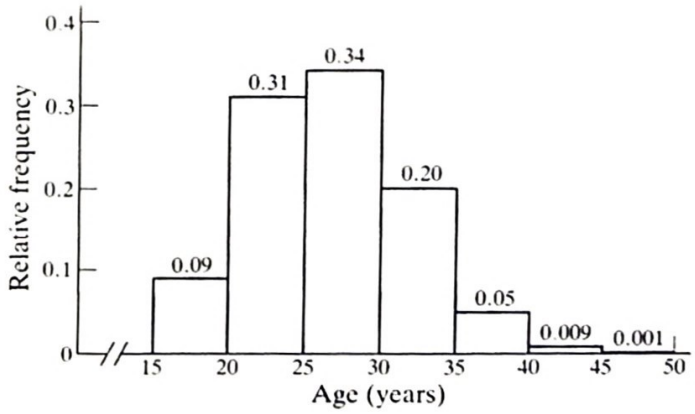


Figure 1.6

Age	Probability
15–19	0.09
20–24	0.31
25–29	0.34
30–34	0.20
35–39	0.05
40–44	0.009
45–49	0.001
Total	1

Actually we are combining ideas of both discrete and continuous distributions in this presentation. The age of a person is a continuous variable (it is a measurement of length of life), and so the above distribution is really a *discrete* approximation to what is essentially an underlying continuous distribution. In the case of a continuous variable, it is usual to have some kind of an algebraic description of the probability distribution, i.e. a formula, which can be used to draw a picture of the distribution. Clearly in this instance this is not possible. However, we can supply a graphical indication of the behaviour of the ages when they are viewed as a continuous variable. By a judicious use of free-hand on Figure 1.6, we obtain the sketch shown in Figure 1.7.

If we use such a curve to represent the probability distribution, then the probability of an age being 20 and 30 years will now correspond to the area indicated on the sketch in Figure 1.8. This is just a generalization of what happens in the histogram; if you think about it, you will realize that areas represent probabilities there as well.

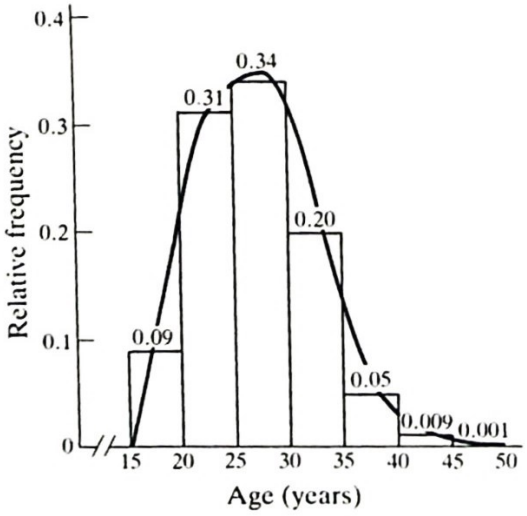


Figure 1.7

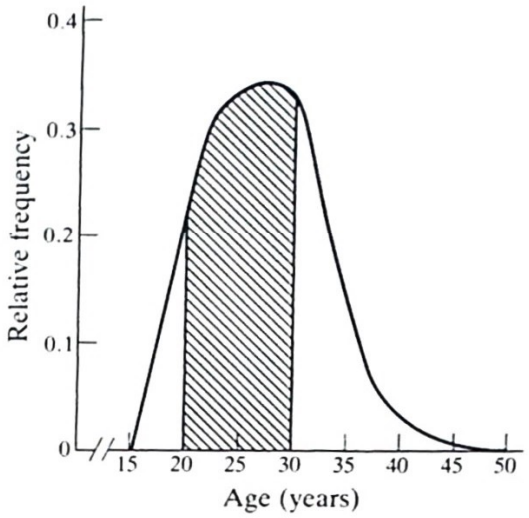


Figure 1.8

Now consider two age groups of equal lengths, say 20–24 years and 40–44 years. The probabilities associated with these groups are represented in Figure 1.9. Just by visually comparing the two areas, we see that the probability of selecting a mother between 20 and 24 years is, of course, far greater than for the other group. This feature is often expressed by saying that there is a greater 'density' of probability in the first group than in the second.

Because a curve such as that in Figure 1.9 indicates this relative density of probability, it is called a *probability density function* (or pdf for short). If the distribution is derived from theoretical arguments (instead of empirically, as here), then we are usually fortunate enough to find an algebraic representation of the pdf; this often allows us to obtain

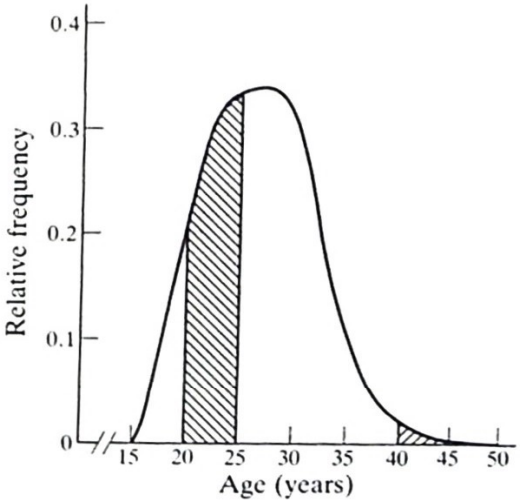


Figure 1.9

probabilities quite easily and more accurately.

Once we have available the probabilities of some events (either by empirical or theoretical means), there are a number of probability laws that can be used to find probabilities of other events in which we might be interested. We shall only mention here the few which we need for the purpose of this book.

First, consider again the two events, and their probabilities, highlighted in Figure 1.9. They are

$$\text{Prob(age of mother is between 20 and 24)} = 0.30$$
$$\text{Prob(age of mother is between 40 and 44)} = 0.01$$

Suppose now that we were interested in the probability of the mother's age being in either of these two groups. Since the area representing this probability is obviously obtained by adding the two individual shaded areas in Figure 1.9, we have

$$\begin{aligned}&\text{Prob((age is between 20 and 24) or (age is between 40 and 44))}\\ &= \text{Prob(age is between 20 and 24)} + \text{Prob(age is between 40 and 44)}\\ &= 0.30+0.01 = 0.31\end{aligned}$$

This is a direct application of the simple form of the so-called *addition law* of probability, which states that if two events, E_1 and E_2 say, are *mutually exclusive* (that is, they cannot occur simultaneously, which is indeed the case in our example), then

$$\text{Prob}(E_1 \textit{ or } E_2) = \text{Prob}(E_1) + \text{Prob}(E_2)$$

If E_1 and E_2 are *not* mutually exclusive, then a more general addition law is needed; this is

$$\text{Prob}(E_1 \textit{ or } E_2) = \text{Prob}(E_1) + \text{Prob}(E_2) - \text{Prob}(E_1 \textit{ and } E_2)$$

where $\text{Prob}(E_1 \textit{ and } E_2)$ is the probability of both E_1 and E_2 occurring.

The addition law is one of the two most-used probability laws. The other is the *multiplication law*. As the name implies, the *independence* of two events E_1 and E_2 states that the occurrence or non-occurrence of either event does not, in any way, affect the occurrence or non-occurrence of the other event; thus, the probability of E_1 (say) occurring does not depend at all on whether E_2 does or does not occur. When we have two events that are independent in this sense, the multiplication law provides us with a way of calculating the probability of both events occurring. The multiplication law is given by

$$\text{Prob}(E_1 \textit{ and } E_2) = \text{Prob}(E_1) \times \text{Prob}(E_2)$$

If we have more than two events, say, $E_1, E_2, \ldots, E_n$, which are all mutually independent, then similarly

$$\text{Prob}(E_1 \textit{ and } E_2 \textit{ and } \ldots \textit{ and } E_n) = \text{Prob}(E_1) \times \text{Prob}(E_2) \times \ldots \times \text{Prob}(E_n)$$

To illustrate this law, suppose that, in the population of mothers, we have information on the colour of the mothers' eyes. In particular, suppose that 40% of the mothers have blue eyes, so that the probability of selecting a mother with blue eyes is 0.40. Then, making the reasonable assumption that eye colour is *independent* of age, we can calculate the probability of, for example, selecting a mother who is both aged between 20 and 24 years *and* has blue eyes in the following manner:

$$\begin{aligned}&\text{Prob(aged between 20 and 24 } \textit{and} \text{ blue-eyed)}\\ &\quad = \text{Prob(aged between 20 and 24)} \times \text{Prob(blue-eyed)}\\ &\quad = 0.30 \times 0.40\\ &\quad = 0.12\end{aligned}$$

– quite a small chance! Having sown the seeds of a few of the main concepts of theoretical distributions, we now progress to a slightly more formal approach.

RANDOM VARIABLES

Undoubtedly, all of us at some time or other have thrown a die (or dice) in playing some game – Ludo, Monopoly, etc. Now before we throw a die, we have no idea as to which of the scores 1, 2, 3, 4, 5 or 6 will occur; all we know is that, *if* the die is fair, then each score has an equal probability (1 chance in 6, or $\frac{1}{6}$) of occurring. So we have a variable quantity (the score) whose values 1, 2, 3, 4, 5 or 6 occur according to a rule of chance: associated with each score is the probability of $\frac{1}{6}$. Other examples of variables whose outcomes depend on some rule of chance are: the number of defectives in a sample of, say, 20 electronic components; the number of shoppers arriving at a supermarket checkout over a prescribed period of time; the actual (as opposed to intended) time of arrival of buses at a particular stop; and the total score on two dice in a game of Monopoly. However, we note that the rules of chance for these variables are not so easily predetermined as with the score on a die!

In each case, we are dealing with variable quantities whose values (sometimes called *realizations*) occur according to a law of chance, i.e. its *probability distribution*. Such variables are called *random variables* to distinguish them from the sort of variables we get in algebra, which are

determined by definite formula (e.g. $y = 2x + 6$) rather than by chance behaviour.

Random variables can be classified as being either *discrete* or *continuous*. A discrete random variable can take on only certain values, usually (but not always) just integers, as with the scores 1, 2, 3, 4, 5 and 6 of a die. On the other hand, a continuous random variable can take any value over its possible range, as with the times of arrivals of buses at a stop. More often, the values of a discrete random variable are obtained by a counting process, while those of a continuous random variable are derived by a measurement process.

It is usual to denote random variables by upper-case (capital) letters and their realizations by lower-case letters. So, for example, if we are interested in the scores from throwing a die we would perhaps denote 'score on die' by D and then write its probability distribution as:

d	1	2	3	4	5	6
Prob($D = d$)	1/6	1/6	1/6	1/6	1/6	1/6

To illustrate these ideas further, let us consider the situation that arises in the game of Monopoly: a player throws two dice and then moves his marker the number of squares indicated by the total score of the dice. Let us use T to denote this random variable, i.e.

$$T = \text{total score on the two dice}$$

The possible values of T are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12, and the probabilities associated with these scores are easily obtained. Observe first that when two dice are thrown, 36 *equally likely* realizations are possible:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

where, say, the dice are of different colours (blue and red), and the first number in each pair represents the outcome on the blue die and the second number the outcome on the red die. We can then immediately see, for example, that a total score of 2 can be achieved in only one way, by (1,1), and so Prob(T=2) = 1/36. Similarly there are two ways out of 36 of scoring a total of 3, and so Prob(T=3) = 2/36. Continuing in this way, we obtain the distribution of T as:

t	2	3	4	5	6	7
Prob($T = t$)	1/36	2/36	3/36	4/36	5/36	6/36
t	8	9	10	11	12	
Prob($T = t$)	5/36	4/36	3/36	2/36	1/36	

These probabilities should be consistent with the experience of avid players of Monopoly; for example, the most likely score is 7 (but it still only occurs 1 time in 6 on average) while the scores of 2 and 12 each only occur 1 time in 36 in the long run.

Once the probability distribution of a random variable, which for generality we call X, has been obtained, we can derive its *cumulative* distribution function (cdf). This function, which is usually denoted by $F(x)$, gives probabilities of events of the type: $X \leqslant x$ (i.e. the event that X takes a value that is no larger than x) – and such events are often of considerable importance. The cdf can be obtained in a manner similar to that for the distribution of ages, i.e. we simply add successive probabilities. Doing this for the random variable T (the total score from two dice), we obtain

$$F(t) = \text{Prob}(T \leqslant t) = \begin{cases} 0 & t<2 \\ 1/36 & 2 \leqslant t<3 \\ 3/36 & 3 \leqslant t<4 \\ 6/36 & 4 \leqslant t<5 \\ 10/36 & 5 \leqslant t<6 \\ 15/36 & 6 \leqslant t<7 \\ 21/36 & 7 \leqslant t<8 \\ 26/36 & 8 \leqslant t<9 \\ 30/36 & 9 \leqslant t<10 \\ 33/36 & 10 \leqslant t<11 \\ 35/36 & 11 \leqslant t<12 \\ 1 & t \geqslant 12 \end{cases} \quad \text{for}$$

Figure 1.10 shows a graph of this cdf. Notice that the cdf is specified for all real numbers and not just the individual scores 2, 3, . . . , 12. Thus (if one wished!) we could say that Prob($T \leqslant 4.5$) = 6/36 and Prob($T \leqslant 20$) = 1; these results are undoubtedly true, though they are hardly likely to be very useful. However, the cdf does help us to find probabilities of more sensible events such as:

$$\text{Prob}(T \leqslant 7) = F(7) = 21/36$$

$$\text{Prob}(T > 10) = 1 - \text{Prob}(T \leqslant 10) = 1 - F(10) = 1 - 33/36 = 3/36$$