

Fundamentals of STATISTICS

VOLUME ONE

A. M. GOON, M.A., Ph.D.
M. K. GUPTA, M.Sc., Ph.D.
B. DASGUPTA, M.Sc.

CALCUTTA
THE WORLD PRESS PRIVATE LTD.
1975

*TO OUR TEACHERS
IN THE DEPARTMENTS OF STATISTICS OF
PRESIDENCY COLLEGE AND CALCUTTA UNIVERSITY*

PREFACE

Until recently, students of statistics had to depend very heavily on class lectures ; the more serious ones had to supplement the knowledge gathered from this source through reading of journals and books of specialised types. To undergraduate students (Pass and Honours) in Indian universities, in particular, this situation, arising out of a lack of texts that would deal with the subject in a comprehensive manner, had been a serious handicap.

It was in an attempt to fill this lacuna that we started out to write *Fundamentals of Statistics* some fourteen years ago. The book was well received by students and teachers alike, the first edition going out of print in less than a year. Heartened by the popularity of the first two editions, we decided to treat the subject more comprehensively, in two volumes, in subsequent editions. A number of topics were added in the third edition, while many of the chapters already included in the earlier editions were rewritten, many of the ideas were discussed in greater detail, and considerably more examples and exercises were covered, with a view to providing the serious reader with a firmer grip on the subject-matter.

Volume One of the fourth edition has been out of print for some time now, necessitating a newer edition, and this has enabled us to make a revision of the volume, besides giving it a new (and we suppose a more pleasant) garb. Some new topics have been added, together with some new examples and exercises. Among these, the sections on the theory of errors, that form Appendix A to the volume, deserve special mention. At the same time, mistakes detected in the fourth edition, by fellow-teachers or by ourselves, have been removed, and slight changes have been made here and there to improve the exposition.

Fundamentals of Statistics, we should point out, is a book of *statistical methods* (as opposed to *statistical theory* or mathematical statistics). But perhaps in quite a few respects it can claim some distinction among the numerous books of this kind that have since appeared in the market. For one thing, it is not intended to be a

cook-book of statistics. The theoretical principles underlying the methods are discussed in detail, and mathematical derivations of most of the results are supplied. Apart from the requirements of university curricula this is meant to enable the practitioner to go about his job with confidence, using the right or a reasonably good tool on every occasion.

Volume One of *Fundamentals* is concerned with the essential mathematical tools, including the numerical techniques of mathematics and the elements of probability theory, and with statistical methods of a general type. In Volume Two, on the other hand, are covered the important areas of analysis of variance and designs, both of experiments and of sample surveys, and statistical methods that are meant for some special fields of application, viz. demography, psychology and education, economics and industrial quality control.

Volume One starts in Chapter 1) with a brief survey of the concepts and results of algebra (including matrix algebra) and calculus that are essential for an understanding of the subsequent chapters. The mathematical bases of the many approximations that one has to make in the course of applying statistical methods to real problems—in terms of interpolation, numerical differentiation and integration and numerical solution of equations—are dealt with in Chapter 2. This chapter also includes a section on the common errors that arise in numerical work. In Chapter 3 are presented the basic ideas and results of probability theory. For the sake of simplicity the case of a finite sample space with equally likely elements only is adequately treated here, but the mode of extension to the general case (viz. Kolmogorov's axiomatic approach) is also indicated.

Part Two, that deals with statistics proper, constitutes the main body of this volume. Together with methods of tabulation and diagrammatic representation of data, the basic concepts of frequency distribution, central tendency, dispersion, etc., are presented in Chapters 4-8. The concepts are elaborated here by keeping in view sample data, but actually the distinction between a population and a sample from that population—not very important at this stage—is deliberately avoided. Chapter 9 gives numerous (univariate) theoretical distributions, including those of the Pearsonian system,

that are employed as simplifying models of distributions generally encountered in practice. Chapters 10-13 are concerned with the association of attributes and that of variables, together with such topics as rank correlation and intra-class correlation. A notable feature is Section 10.7, on the possible causal relationship between smoking and lung cancer, where the logical aspects of the so-called 'cancer controversy' are highlighted. Chapters 14-18 are devoted to random sampling and statistical inference. In Chapter 14, the idea of random sampling and methods of random sampling are presented, and some important sampling distributions are derived. The basic principles of point estimation, interval estimation and testing of hypotheses are the subject-matter of Chapter 15. In the next chapter, these principles are used to develop tests and estimates for some typical problems, the normal population distribution set-up being treated at length. While this chapter presents methods that are applicable to samples of any sizes whatsoever, those that are approximate and applicable only to large samples, but are simpler, are discussed in Chapter 17. Chapter 18 deals with non-parametric methods, that are playing a very important rôle in present-day statistical practice.

Some readers are likely to complain that multivariate analysis has not received in the book its due share. We have refrained from including this topic lest the volume should become too unwieldy and would just refer the interested reader to the books on this topic by M. G. Kendall, C. R. Rao and T. W. Anderson.

We have drawn upon our experience as teachers of statistics not only in regard to the exposition but also in formulating the numerous examples and exercises given in the book. These should help the reader towards a better understanding of the subject-matter and to readily adapt the method used in solving any problem to other problems of a similar type. Typical essay-type questions are there, too, mainly to help the student prepare for examinations.

We claim no great originality either in the choice or in the presentation of material. On the contrary, we shall deem our efforts amply rewarded if we have been able to give in this book an intelligible, systematic and reasonably accurate account of the principles and procedures of statistics. In carrying out the successive

revisions, too, our only aim has been to make the book increasingly useful for students as well as for research workers, who are relying more and more on statistical methods in their investigations.

The encouragement and suggestions that we have received from our teachers, Professors A. Bhattacharyya, P. K. Bose, B. N. Ghosh, H. K. Nandi and P. K. Banerji, and from numerous friends and colleagues have helped sustain our interest in this venture. Our friend, Professor R. Datta, of the Department of Economics, Calcutta University, continues to be a source of inspiration.

We take this opportunity to acknowledge once again our debt of gratitude to the publishers, The World Press, for the excellent co-operation that we have been receiving from them.

*Presidency College, Calcutta
February 1975*

THE AUTHORS

CONTENTS

CHAPTER		PAGES
Part One : MATHEMATICAL PRELIMINARIES		1—136
1	SOME USEFUL CONCEPTS AND RESULTS	3—34
1a	Algebra : Sum and product notations. Set. Real-number system. Sequence, series and their convergence. Binomial series. Exponential and logarithmic series. Some important algebraic inequalities. Matrices. Vectors. Determinants. Inverse matrix. Quadratic forms. Linear equations.	
1b	Calculus : Concept of a function. Limit of a function. Meaning of infinity. General results on limits. Some important limits. Continuity. Derivative of a function. Partial derivatives. Application of derivatives of functions. Definite integral. Infinite or improper integral. Indefinite integral. Double and multiple integrals. Some special integrals.	
1c	Stirling's approximation.	
2	NUMERICAL ANALYSIS	35—97
2a	Inaccuracies and approximations : Different types of inaccuracies. Rounding off. Significant figures. Absolute, relative and percentage errors.	
2b	Interpolation : The problem of interpolation. Finite differences. Error in a tabular value. Use of operators : Δ , E . Newton's forward interpolation formula. Newton's backward interpolation formula. Lagrange's interpolation formula. Divided differences. Newton's divided difference formula. Central difference formulae. Remainder terms in interpolation formulae. Bivariate interpolation.	
2c	Numerical differentiation.	
2d	Numerical integration : Trapezoidal rule. Simpson's one-third rule. Weddle's rule. Relative accuracy of quadrature formulae. Euler-Maclaurin formula.	
2e	Numerical solution of equations : Method of false position. Newton-Raphson method. Method of iteration. Convergence of the iteration method. Convergence of the Newton-Raphson method. Horner's method.	
3	ELEMENTS OF PROBABILITY THEORY	98—135
Meaning of probability. Notation and terminology. Classical definition of probability. Theorems of total probability. <u>Conditional probability</u> and statistical independence. Limitations of the classical definition. An axiomatic approach. <u>Random variable, and its expectation and variance.</u> Joint distribution of two random variables. <u>Law of large numbers.</u>		

CHAPTER		PAGES
	Part Two . GENERAL STATISTICAL METHODS	137—528
4 INTRODUCTION TO STATISTICAL METHODS		139—160
The nature of statistics Statistics and other disciplines. Collection of data Scrutiny of data Presentation of numerical data Diagrammatic representation of data		
5 FREQUENCY DISTRIBUTIONS		161—177
Summarisation of data Attribute and variable Frequency distribution of an attribute Discrete and continuous variables Frequency distribution of a variable Graphical representation of frequency distribution of a variable		
6 MEASURES OF CENTRAL TENDENCY		178—197
Descriptive measures of statistics. Central tendency Arithmetic mean Median Mode Comparison of mean, median and mode Other measures of central tendency		
7 MEASURES OF DISPERSION		198—214
Meaning of dispersion Range Mean deviation Standard deviation Comparison of range, mean deviation and standard deviation Measures based on mutual differences of observations Quartile deviation Measures of relative dispersion Curve of concentration		
8 MOMENTS AND MEASURES OF SKEWNESS AND KURTOSIS		215—231
Moments Central moments expressed in terms of moments about an arbitrary origin Moments about an arbitrary origin expressed in terms of central moments Sheppard's corrections for moments Skewness Kurtosis		
9 UNIVARIATE THEORETICAL DISTRIBUTIONS		232—274
<u>Population and sample</u> Theoretical distributions <u>Binomial distribution</u> Moments of the binomial distribution A recursion relation concerning moments of the binomial distribution Fitting a binomial distribution to an observed distribution <u>Poisson distribution</u> Moments of the Poisson distribution A recursion relation concerning moments of the Poisson distribution Fitting a Poisson distribution to an observed distribution <u>Hypergeometric distribution</u> Negative binomial distribution <u>Rectangular (or uniform) distribution</u> <u>Normal distribution</u> Properties of the normal distribution Limiting forms of binomial and Poisson distributions Fitting a <u>normal distribution</u> Importance of the normal distribution in statistics Log-normal distribution Generalised systems of frequency curves		

CHAPTER		PAGES
10 JOINT DISTRIBUTIONS OF ATTRIBUTES		275—296
Data on two or more attributes. Independence and association. Measures of association for the 2×2 case. Manifold two-way ($k \times l$) classification. Case of more than two attributes. Association and causal relationship. Smoking and lung cancer.		
11 BIVARIATE FREQUENCY DISTRIBUTIONS		297—332
Bivariate data. Scatter diagram. Correlation. Correlation coefficient. Properties of the correlation coefficient. Calculation of correlation coefficient from grouped data. Regression lines. Some important results relating to regression lines. Theoretical distribution of two variables. Limitations of the correlation coefficient. Correlation index and correlation ratio.		
12 MULTIVARIATE FREQUENCY DISTRIBUTIONS		333—358
Multivariate data. Multiple regression. Multiple correlation. Some results relating to multiple regression and multiple correlation. Partial correlation. Some relations connecting partial regression and partial correlation coefficients. Expression of a multiple correlation coefficient in terms of total and partial correlation coefficients. Expression of a higher-order coefficient in terms of coefficients of a lower order. Expression of a lower-order coefficient in terms of coefficients of a higher order. Multivariate normal distribution.		
13 SOME OTHER TYPES OF CORRELATION		359—376
Rank correlation coefficient. Spearman's rank correlation coefficient. Kendall's rank correlation coefficient. Grade correlation. Intra-class correlation. Population intra-class correlation.		
14 RANDOM SAMPLING AND SAMPLING DISTRIBUTIONS		377—400
<u>Random sampling.</u> Parameter, statistic and its sampling distribution. Expectation and standard error of sample mean. Expectation and standard error of sample proportion. Sampling distributions associated with discrete populations. Four fundamental distributions derived from the normal. Sampling distributions of mean and variance in sampling from a normal population.		
15 BASIC PRINCIPLES OF STATISTICAL INFERENCE		401—426
<u>Estimation and testing of hypotheses.</u> Point estimation of parameters. Maximum-likelihood estimation. Interval estimation of parameters. Test of significance. Neyman and Pearson's theory of testing of hypotheses. Likelihood-ratio tests.		

CHAPTER		PAGES
16 EXACT TESTS AND CONFIDENCE INTERVALS		427—469
Introduction Tests relating to binomial distributions Tests relating to Poisson distributions A test for independence of two attributes Problems regarding a univariate normal distribution Comparison of two univariate normal distributions Comparison of means of more than two normal populations Problems relating to a bivariate normal distribution Problems relating to simple regression Tests for multiple and partial correlation coefficients The normality assumption		
17 APPROXIMATE TESTS AND CONFIDENCE INTERVALS		470—506
Introduction Tests and confidence intervals for proportions Approximate tests and confidence limits for Poisson parameters Approximate standard error formulæ of some statistics z transformation of sample correlation and other transformations Frequency X^2 Test for goodness of fit hypothetical population completely specified Test for goodness of fit some parameters of hypothetical population unknown Test for homogeneity Test for independence Simplified formulæ Yates' correction for continuity		
18 NON PARAMETRIC METHODS		507—5.8
Introduction Non parametric estimation of location and dispersion Tolerance interval Non parametric tests for location Two sample non parametric tests for dispersion A general non parametric test for two independent samples One sample run test for randomness A non parametric measure and a test of association		
Appendices		529—552
A ELEMENTARY THEORY OF ERRORS		531—537
Introduction Normal law of errors Most probable value Measures of error		
B STATISTICAL TABLES		538—551
I	Ordinates and areas of the distribution of normal deviate	
II	Distribution of normal deviate Values of τ_α	
III	X^2 distribution Values of $X^2_{\alpha, v}$	
IV	t distribution Values of $t_{\alpha, v}$	
V	F distribution Values of F_{α, v_1, v_2}	
VI	Cumulative binomial probabilities of r or fewer successes in n independent trials with $p=0.5$	
VII	Critical values of T in the Wilcoxon signed rank test	
VIII	Critical values of U for the Mann Whitney test	
IX	Critical values of r in the run test	
X	Critical values of r_R , the Spearman rank correlation coefficient	
INDEX		553—558

“From time immemorial men must have been compiling information for peace and war. Statistics is in this sense as old as statecraft.”

“Statistics, like engineering, requires all the help it can receive from mathematics ; but... (statistics) can never become a branch of mathematics.”

P. C. Mahalanobis

PART ONE

MATHEMATICAL PRELIMINARIES

I

SOME USEFUL CONCEPTS AND RESULTS

In this chapter, we present some concepts and results of algebra and calculus which will be frequently used in subsequent chapters.

1a ALGEBRA

1a.1 Sum and product notations

To simplify the writing of the sum or product of n quantities we shall make use of two signs, which we explain below.

The Greek letter Σ (capital ‘sigma’) is used as a sign of summation. If we have n quantities x_1, x_2, \dots, x_n , their sum is denoted by

$$\sum_{i=1}^n x_i \quad \dots \quad (1.1)$$

or by $\sum_1^n x_i \quad \dots \quad (1.1a)$

When the range of values of i is evident from the context, one may write, simply,

$$\sum_i x_i. \quad \dots \quad (1.1b)$$

$\sum_{i=1}^n x_i$ is read as ‘sigma x_i from 1 to n ’.

Σ follows certain simple rules which can be easily verified :

- (i) $\sum_i (x_i \pm y_i) = \sum_i x_i \pm \sum_i y_i ;$
- (ii) $\sum_i kx_i = k \sum_i x_i$, k being a constant ;
- (iii) $\sum_i (k + x_i) = nk + \sum_i x_i.$

Suppose we have $m+n$ quantities arranged in m rows and n columns as follows :

$$\begin{array}{ccccccc}
 x_{11} & x_{12} & \dots & \dots & x_{1n} \\
 x_{21} & x_{22} & \dots & \dots & x_{2n} \\
 \dots & \dots & \dots & \dots & \dots \\
 x_{m1} & x_{m2} & \dots & \dots & x_{mn}
 \end{array}$$

Here x_{ij} is the quantity occurring both in the i th row and in the j th column. The use of two suffixes i and j provides a convenient way of representing the position of a quantity in the above arrangement.

The sum of all the $m \times n$ quantities may be obtained by first adding the quantities in each row and then adding the row totals:

$$\begin{aligned} & (x_{11} + x_{12} + \dots + x_{1n}) + (x_{21} + x_{22} + \dots + x_{2n}) + \dots \\ & \quad + (x_{m1} + x_{m2} + \dots + x_{mn}) \\ = & \sum_j x_{1j} + \sum_j x_{2j} + \dots + \sum_j x_{mj} \\ = & \sum_i (\sum_j x_{ij}). \end{aligned}$$

This grand total may also be obtained by first adding the quantities in each column and then adding the column totals:

$$\begin{aligned} & (x_{11} + x_{21} + \dots + x_{m1}) + (x_{12} + x_{22} + \dots + x_{m2}) + \dots \\ & \quad + (x_{1n} + x_{2n} + \dots + x_{mn}) \\ = & \sum_i x_{i1} + \sum_i x_{i2} + \dots + \sum_i x_{in} \\ = & \sum_j (\sum_i x_{ij}). \end{aligned}$$

Denoting the grand total by

$$\sum_{i,j} x_{ij},$$

we have thus

$$\begin{aligned} \sum_{i,j} x_{ij} &= \sum_i (\sum_j x_{ij}) \\ &= \sum_j (\sum_i x_{ij}), \quad \dots \quad (1.2) \end{aligned}$$

each of which is a double sum.

When we have more than two ways of classification, we shall have to use more than two suffixes in order to indicate the position of any quantity. The grand total will then be represented as

$$\sum_{i,j,k} x_{ijk}, \quad \dots \quad (1.3)$$

which may be obtained by summing the quantities with respect to the suffixes, taken one by one in any order we like.

The Greek letter Π (capital 'pi') is used as a sign of multiplication.

The product of the n quantities x_1, x_2, \dots, x_n is denoted by

$$\prod_{i=1}^n x_i \quad \dots \quad (1.4)$$

or $\prod_1^n x_i \quad \dots \quad (1.4a)$

or, simply, $\prod_i x_i. \quad \dots \quad (1.4b)$

$\prod_{i=1}^n x_i$ is read as 'product x_i from 1 to n '.

Like Σ , the product sign also obeys certain simple rules. For example,

$$\prod_i (x_i y_i) = (\prod_i x_i) (\prod_i y_i),$$

$$\prod_i (x_i / y_i) = \frac{\prod_i x_i}{\prod_i y_i},$$

etc.

1a.2 Set

In mathematics, we deal with collections of objects that have some common quality or feature, and we refer to such a collection as a *set*. The objects in a set are called the *elements* or *members* of the set. A State Legislative Assembly is a set whose elements are the MLAs. A set can be subdivided into smaller sets called *subsets*. Some examples of subsets are the set of Congress MLAs, the set of Communist MLAs and the set of female MLAs. If each element of a set B is also an element of a set A , then B is called a subset of A . If there is an element of A which is not in B , then B is called a *proper subset* of A . It is possible that a set may contain no elements ; then the set is called an *empty set* or a *null set*. For example, if there be no female MLAs in the State Legislature, then the subset of female MLAs is an empty set.

1a.3 Real-number system

We use the *real-number system* in elementary algebra. A short account of this system is given below.

The first step was the invention of the set of *counting numbers*. The symbols for the counting numbers in most countries are 1, 2, 3, 4, They answer the question "How many ?" Another name for a counting number is *positive integer*. We count

the elements of a finite set by corresponding one and only one element of the set with one and only one of the positive integers 1, 2, 3, ..., starting with 1 and continuing till each element of the set has been exhausted

From this simple beginning, one can expand the concept of number by introducing sophisticated ideas. Thus the need for a directed number to denote distance along either direction from a fixed point on a line suggests the number 0 and the negative integers $-1, -2, -3$,

If a system of numbers is to be useful, then it must have the property of *closure* under the four fundamental operations of addition, subtraction, multiplication and division. A set of elements is said to be closed under an operation if after the operation on the elements of the set, the resulting element is also a member of the set. It is easy to verify that the set of positive integers is closed under addition and multiplication, but not under subtraction. If a is a positive integer, then 0 and $-a$ may be defined as follows

$$0+a=a, a+(-a)=0$$

We now define the *set of integers* to be the set of numbers composed of the positive integers, zero and the negative integers. The introduction of 0 and the set of negative integers makes the set of integers closed under addition, subtraction and multiplication.

The operation of division requires an extension of our set of integers, and this is done by defining a set of *rational numbers*. A *quotient* or *fraction* is a ratio of integers where the denominator is different from zero. The set of rational numbers is the set of numbers each of which can be expressed as the quotient of two integers. The set of rational numbers is closed under addition, subtraction, multiplication and division.

If a rational number is converted into decimal form, one of two things will happen:

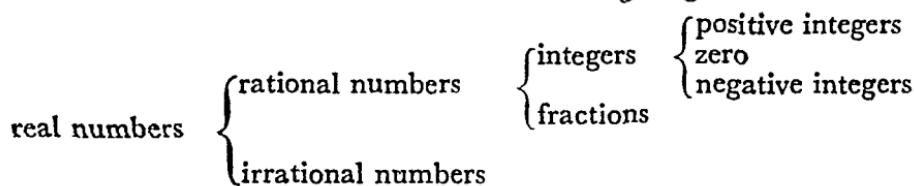
(i) The decimal terminates (e.g. 3, 0.24). This group contains all the integers and some rational numbers.

(ii) The decimal does not terminate but repeats (e.g. 4.12, 0.0413213). This group contains the remaining rational numbers.

An *irrational number* is a non-terminating, non-periodic decimal, e.g. $\sqrt{2}=1.4142 \dots, \sqrt{18}=4.24264 \dots$ Apparently, the digits can

be continued endlessly but no sequence of them will repeat. Not all irrational numbers can be obtained as the result of an algebraic operation, and an irrational number which is not algebraic is said to be *transcendental*. Typical examples are the base of natural logarithms (e) and the ratio of the diameter to the circumference of a circle (π).

We now define the set of *real numbers* to be the set of numbers consisting of the set of rational numbers and the set of irrational numbers. The set of real numbers is closed under the four fundamental operations. The relations between various subsets of the set of real numbers will be clear from the following diagram :



1a.4 Sequence, series and their convergence

A *sequence* is a set of quantities s_1, s_2, \dots , called elements, which can be arranged in an order, so that when n is given, the n th element of the sequence is completely specified. The elements are arranged by matching them one by one with the positive integers $1, 2, \dots, n, \dots$. A common symbol for a sequence is $\{s_n\}$.

A sequence $\{s_n\}$ is said to be *bounded* if there exists an arbitrary positive number M such that $|s_n| \leq M$, for all n . Otherwise, the sequence is said to be *unbounded*. If there exists an l such that for every choice of $\epsilon > 0$, there is an N such that $|s_n - l| < \epsilon$ for all $n > N$, then the sequence $\{s_n\}$ is *convergent* and has the *limit* l . It is customary to represent this by $\lim_{n \rightarrow \infty} s_n = l$. We shall see more of limit in

Section 1b.2. A sequence $\{s_n\}$ which is not convergent is called *divergent*. For example, the sequence of integers $\{n\}$ is divergent.

A *series* is an expression of the form $a_1 + a_2 + \dots + a_n + \dots$ which may have a finite or an infinite number of terms. Its partial sums, $s_1 = a_1, s_2 = a_1 + a_2, \dots, s_n = \sum_1^n a_i, \dots$, constitute a sequence $\{s_n\}$. A series is said to be *convergent* if its sequence of partial sums is convergent. If $\lim_{n \rightarrow \infty} s_n = s$, then s is the sum of the convergent

infinite series and $\sum_1^{\infty} a_n = s$. If $\lim_{n \rightarrow \infty} s_n = \pm\infty$, then the series is said to be *divergent* to $\pm\infty$. A series is said to *oscillate* if the sequence of partial sums does not tend to a definite limit but oscillates between two values, say, m and M . A simple example of oscillating series is $1 - 1 + 1 - 1 + \dots + (-1)^{n-1} \pm \dots$. This oscillates between 0 and 1. A series is *absolutely convergent* if the series of corresponding absolute values, $|a_1| + |a_2| + \dots + |a_n| + \dots$, is convergent. A series such as $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$, which is convergent but not absolutely convergent, is said to be *conditionally convergent*. Convergent series are usually the most useful type for practical applications, hence it is of importance to test a series for convergence. The comparison test and the ratio test (of d'Alembert) are most often used to test convergence of a series.

1a 5 Binomial series

A *binomial* is an expression that contains exactly two terms. In this section, we first give a formula which enables us to express any positive integral power of a binomial. By actual multiplication, we obtain the following expansions

$$\begin{aligned}(x+y)^1 &= x+y, \\ (x+y)^2 &= x^2 + 2xy + y^2, \\ (x+y)^3 &= x^3 + 3x^2y + 3xy^2 + y^3,\end{aligned}$$

etc

The general formula is

$$\begin{aligned}(x+y)^n &= \binom{n}{0} x^n + \binom{n}{1} x^{n-1}y + \binom{n}{2} x^{n-2}y^2 + \\ &\quad + \binom{n}{r} x^{n-r}y^r + \dots + \binom{n}{n} y^n,\end{aligned}\tag{15}$$

which holds for positive integral values of n *

When n is not a positive integer (i.e. is negative or fractional), some restriction has to be put on x and y . In this case, if $|y| < |x|$, where $|k|$, called 'modulus k ', denotes the numerical or absolute

* $\binom{n}{r}$ stands for the expression $n(n-1)\dots(n-r+1)/r!$, where $x' = x(r-1)$ 21
A similar symbol is $(n)_r = n(n-1)\dots(n-r+1)$

value of k , then one can write

$$(x+y)^n = \binom{n}{0} x^n + \binom{n}{1} x^{n-1}y + \binom{n}{2} x^{n-2}y^2 + \dots + \binom{n}{r} x^{n-r}y^r + \dots ad inf. \quad ... \quad (1.6)$$

(1.6) may be looked upon as the most general form of the binomial series, which includes (1.5) as a particular case.

Two important results, which follow from (1.6), are

$$(1+x)^{-1} = 1 - x + x^2 - \dots + (-1)^r x^r \pm \dots$$

$$\text{and } (1-x)^{-1} = 1 + x + x^2 + \dots + x^r + \dots$$

which are valid for $|x| < 1$.

1a.6 Exponential and logarithmic series

The series

$$1 + \frac{1}{1!} + \frac{1}{2!} + \dots + \frac{1}{n!} + \dots \quad \dots \quad (1.7)$$

is denoted by the letter e . It can be shown that the x th power of any positive number a is

$$a^x = 1 + \frac{x \log_e a}{1!} + \frac{x^2 (\log_e a)^2}{2!} + \dots + \frac{x^r (\log_e a)^r}{r!} + \dots \quad \dots \quad (1.8)$$

This is the most general form of the exponential series. As a particular case, we have, by putting $a=e$,

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots + \frac{x^r}{r!} + \dots \quad \dots \quad (1.9)$$

The logarithmic series, which gives the expansion of $\log_e(1+x)$ in ascending powers of x , is

$$\log_e(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots \quad \dots \quad (1.10)$$

From this we have also

$$\log_e(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \dots \quad \dots \quad (1.11)$$

Both (1.10) and (1.11) are valid for $|x| < 1$.

1a 7 Some important algebraic inequalities

(i) If a and b be any two numbers, then

$$|a+b| \leq |a| + |b|$$

Proof Whatever the signs of a and b , we shall have

$$-|a| \leq a \leq |a|$$

and

$$-|b| \leq b \leq |b|,$$

so that

$$-|a|-|b| \leq a+b \leq |a|+|b|,$$

i.e.

$$|a+b| \leq |a| + |b| \quad (1.12)$$

(ii) Let x_1, x_2, \dots, x_n be a set of positive quantities and suppose

$$A = \frac{x_1 + x_2 + \dots + x_n}{n},$$

$$G = (x_1 x_2 \dots x_n)^{1/n},$$

and

$$H = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n}$$

Then

$$A \geq G \geq H$$

Proof Since the square of a real quantity is necessarily non-negative,

$$(\sqrt{x_1} - \sqrt{x_2})^2 \geq 0,$$

so that

$$\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$$

From this again, we have

$$\frac{\frac{x_1 + x_2 + x_3 + x_4}{2} + \frac{x_1 + x_2 + x_3 + x_4}{2}}{2} \geq \sqrt{\frac{x_1 + x_2}{2} \times \frac{x_3 + x_4}{2}} \geq \sqrt{\sqrt{x_1 x_2} \sqrt{x_3 x_4}}$$

i.e.

$$\frac{x_1 + x_2 + x_3 + x_4}{4} \geq (x_1 x_2 x_3 x_4)^{1/4}$$

Proceeding in this way, we can prove the result $A \geq G$ for $n=2^m$, where m is a positive integer. It will now be shown that the inequality also holds when n is not of the form 2^m .

Let us take the smallest m such that $n < 2^m$. Then, along with x_1, x_2, \dots, x_n , we may consider $2^m - n$ other positive quantities,

$$x_{n+1}, x_{n+2}, \dots, x_{2^m},$$

all equal to A .

According to the result proved above,

$$\frac{x_1 + x_2 + \dots + x_n + x_{n+1} + \dots + x_{2^m}}{2^m} \geq \left(x_1 x_2 \dots x_n x_{n+1} \dots x_{2^m} \right)^{1/2^m},$$

or $\frac{x_1 + x_2 + \dots + x_n + (2^m - n)A}{2^m} \geq \left(x_1 x_2 \dots x_n A^{2^m - n} \right)^{1/2^m},$

or $\frac{nA + (2^m - n)A}{2^m} \geq \left(G^n A^{2^m - n} \right)^{1/2^m}$ or $A^{2^m} \geq \frac{G^n A^{2^m}}{A^n},$

or $A^n \geq G^n, \quad \text{or} \quad A \geq G. \quad \dots \quad (1.13)$

Since $\frac{1}{x_1}, \frac{1}{x_2}, \dots, \frac{1}{x_n}$ are themselves positive quantities, we have also

$$\frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n} \geq \left(\frac{1}{x_1} \cdot \frac{1}{x_2} \dots \frac{1}{x_n} \right)^{1/n},$$

i.e. $\frac{1}{H} \geq \frac{1}{G},$

so that $G \geq H. \quad \dots \quad (1.14)$

Combining (1.13) and (1.14), we can write

$$A \geq G \geq H. \quad \dots \quad (1.15)$$

The equality signs hold when all the n quantities are equal.

(iii) If x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n be two sets of real numbers, then

$$(x_1^2 + x_2^2 + \dots + x_n^2)(y_1^2 + y_2^2 + \dots + y_n^2) \geq (x_1 y_1 + x_2 y_2 + \dots + x_n y_n)^2.$$

Proof: We first note that if x_1, x_2, \dots, x_n are all equal to zero and/or y_1, y_2, \dots, y_n are all equal to zero, the inequality is trivially true, both sides being equal to zero. So we assume, without any loss of generality, that there is at least one non-zero value in the set of x 's as well as in the set of y 's.

Let us denote by A , B and C , respectively, the sums

$$x_1^2 + x_2^2 + \dots + x_n^2,$$

$$y_1^2 + y_2^2 + \dots + y_n^2$$

and $x_1y_1 + x_2y_2 + \dots + x_ny_n$

From the familiar inequality $a^2 + b^2 \geq 2|ab|$, we have

$$\frac{x_1^2}{A} + \frac{y_1^2}{B} \geq 2 \frac{|x_1y_1|}{\sqrt{AB}},$$

$$\frac{x_2^2}{A} + \frac{y_2^2}{B} \geq 2 \frac{|x_2y_2|}{\sqrt{AB}},$$

$$\frac{x_n^2}{A} + \frac{y_n^2}{B} \geq 2 \frac{|x_ny_n|}{\sqrt{AB}}$$

Adding the above n inequalities, we get

$$\frac{x_1^2 + x_2^2 + \dots + x_n^2}{A} + \frac{y_1^2 + y_2^2 + \dots + y_n^2}{B} \geq 2 \frac{|x_1y_1| + |x_2y_2| + \dots + |x_ny_n|}{\sqrt{AB}} \\ \geq 2 \frac{|x_1y_1 + x_2y_2 + \dots + x_ny_n|}{\sqrt{AB}},$$

i.e. $2 \geq 2 \frac{|x_1y_1 + x_2y_2 + \dots + x_ny_n|}{\sqrt{AB}}$

Hence

$$AB \geq (x_1y_1 + x_2y_2 + \dots + x_ny_n)^2 = C^2,$$

i.e. $(x_1^2 + x_2^2 + \dots + x_n^2)(y_1^2 + y_2^2 + \dots + y_n^2) \geq (x_1y_1 + x_2y_2 + \dots + x_ny_n)^2 \quad (116)$

This is known as the *Cauchy-Schwarz inequality*. The equality sign holds when $x_i = ky_i$ (or $y_i = kx_i$), k being a constant, for $i=1, 2, \dots, n$.

1a 8 Matrices

A *matrix A* of order $m \times n$ is a rectangular array of numbers or elements arranged in m rows and n columns as

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix},$$

where the elements are represented by a_{ij} . The subscripts i and j

refer, respectively, to the row number and the column number of the element in the array. To emphasize the order, we may write $A_{m \times n}$ instead of A . We shall assume that the elements a_{ij} are real numbers.

For example,

$$\begin{pmatrix} 2 & 0 \\ 4 & 5 \\ -3 & 1 \end{pmatrix}$$

is a 3×2 matrix. The matrix is nothing more than a set of elements arranged in a convenient manner. For example, the prices of m commodities for n periods of time may be represented by an $m \times n$ matrix, where the element a_{ij} is the price of the i th commodity for the j th period.

Two matrices $A = (a_{ij})$ and $B = (b_{ij})$ are called *equal*, and written as $A = B$, if and only if (1) they are of the same order and (2) $a_{ij} = b_{ij}$ for all i and j .

A matrix with the same number of rows as of columns is called a *square* matrix. The elements a_{ii} of a square matrix form the *main diagonal* of the matrix. A square matrix is *symmetric* if $a_{ij} = a_{ji}$ for all i and j . For example,

$$\begin{pmatrix} a & b & 0 \\ b & c & d \\ 0 & d & 0 \end{pmatrix}$$

is a symmetric matrix of order 3.

We shall now define the operations of addition, subtraction and multiplication with matrices.

1. The product of a matrix and a real number (scalar) c is defined as follows :

$$cA = c(a_{ij}) = (ca_{ij}),$$

i.e. a matrix each element of which is c times the corresponding element of A . We take, by convention, $cA = Ac$. $(-1)A$ is written as $-A$.

2. The sum of A and B is defined only when they are of the same order, and then

$$\begin{aligned} A + B &= (a_{ij}) + (b_{ij}) \\ &= (a_{ij} + b_{ij}). \end{aligned}$$

Thus the sum is again a matrix of the same order obtained by adding

corresponding elements of the two matrices. Similarly, we define the difference of the two matrices of the same order as

$$\begin{aligned} \mathbf{A}-\mathbf{B} &= (a_{ij}) - (b_{ij}) \\ &= (a_{ij} - b_{ij}), \end{aligned}$$

i.e. a matrix of the same order with elements $a_{ij} - b_{ij}$.

3. The product of two matrices is defined only when the number of columns of the first matrix is the same as the number of rows of the second. The product of $A_{m \times r}$ and $B_{r \times n}$ is a matrix $C_{m \times n}$, where the elements of C are given by

$$c_{ij} = \sum_{k=1}^r a_{ik} b_{kj}$$

Thus

$$\begin{aligned} A_{m \times r} B_{r \times n} &= (a_{ij})(b_{ij}) \\ &= \left(\sum_{k=1}^r a_{ik} b_{kj} \right) = C_{m \times n} \end{aligned}$$

It is to be noted that $B_{r \times n} A_{m \times r}$ is not defined unless $m=n$. Also, even for square matrices of the same order, AB need not be the same as BA .

These operations can be defined easily, step by step, for more than two matrices.

All the matrix operations, defined so far, are *associative* and *distributive*. Except for subtraction and the third operation, they are also *commutative*.

The transpose, A' , of a matrix A is A with its rows and columns interchanged. Thus if $A_{m \times n} = (a_{ij})$, then $A' = A'_{n \times m} = (a'_{ij})$, where $a_{ij} = a_{ji}$ for all i and j . It is easy to verify that

$$(A') = A, (A \pm B)' = A' \pm B', (AB)' = B' A'$$

Thus a square matrix is symmetric if it is equal to its transpose.

A *diagonal matrix* is a square matrix with all its non-diagonal elements equal to zero, and hence it is symmetric too.

A *unit matrix* (or *identity matrix*) I is a diagonal matrix with all diagonal elements equal to 1. Thus

$$I = (\delta_{ij}),$$

where $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$.

It is easy to verify that

$$I_{n \times n} A_{m \times n} = A_{m \times n} I_{n \times n} = A_{m \times n}$$

A *null matrix* is a matrix with all elements equal to zero and is denoted by \mathbf{O} . Obviously, we have

$$\mathbf{O}_{m \times r} \mathbf{A}_{r \times n} = \mathbf{O}_{m \times n}$$

and $\mathbf{A}_{n \times n} \pm \mathbf{O}_{n \times n} = \mathbf{A}_{n \times n}$.

A square matrix \mathbf{C} is called an *orthogonal matrix* if $\mathbf{C}'\mathbf{C} = \mathbf{I}$.

1a.9 Vectors

A *vector* is an ordered set of numbers arranged in one row or one column. A *row vector* of order n is a matrix of order $1 \times n$, while a *column vector* of order n is a matrix of order $n \times 1$. The transpose of a row vector is a column vector, and conversely.

We shall use a symbol like \mathbf{x} to denote a column vector and a symbol like \mathbf{x}' to denote a row vector.

Since a vector is a special type of matrix, the operations of addition, subtraction and multiplication are also defined for vectors with the same type of restrictions on the orders of the vectors. If \mathbf{x} and \mathbf{y} are two column vectors of order n , i.e. if

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

then $\mathbf{x}'\mathbf{x} = \sum_i x_i^2$, $\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x} = \sum_i x_i y_i$.

A set of non-null vectors, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r$, will be said to be *linearly dependent* if there exist numbers c_1, c_2, \dots, c_r , not all equal to zero, such that

$$c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \dots + c_r \mathbf{x}_r = \mathbf{0},$$

where $\mathbf{0}$ is the *null vector* (i.e. the vector with all elements equal to zero) of the same order.

The vectors will be said to be *linearly independent* if there do not exist c_i 's, not all equal to zero, such that the above relation is true.

1a.10 Determinants

To every square matrix $\mathbf{A}_{n \times n} = (a_{ij})$, there corresponds a number A known as the determinant of the matrix \mathbf{A} . It is also denoted by $|A|$ or $|a_{ij}|$ or

$$\left| \begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{array} \right|,$$

where a_{ij} is the element of the i th row and the j th column. It is defined as the sum of $n!$ terms of the form

$$\pm a_{1\alpha} a_{2\beta} \dots a_{n\nu} \quad (1.17)$$

there being one term corresponding to each permutation $(\alpha \beta \dots \nu)$ of the numbers 1, 2, ..., n . The sign of a term is determined according to the following rule

If the number of inversions in $(\alpha \beta \dots \nu)$ is even (including zero), then

$$a_{1\alpha} a_{2\beta} \dots a_{n\nu}$$

will have a plus sign. If, on the other hand, the number of inversions is odd, the term will have a minus sign.

(When two numbers in a permutation are not in natural order, the greater number preceding the smaller, like the 4 and 3 in [1243], or the 5 and 2 in [15342], they are said to give an *inversion*. The total number of inversions is 1 in the former permutation and 5 in the latter.)

The *minor* of an element a_{ij} in A is defined to be the determinant (of order $n-1$) formed by deleting from A the i th row and the j th column. The *co-factor* of a_{ij} is

$$(-1)^{i+j} \text{ times its minor}$$

and is generally denoted by A_{ij} . For instance, in the present case

$$A_{12} = - \begin{vmatrix} a_{21} & a_{23} & a_{2n} \\ a_{31} & a_{33} & a_{3n} \\ a_{n1} & a_{n3} & a_{nn} \end{vmatrix}$$

To evaluate a determinant one may use the relation

$$A = \sum_{i=1}^n a_{ii} A_{ii} \quad (\text{for any } i=1, 2, \dots, n), \quad (1.18)$$

which expresses A in terms of the elements of the i th row and the corresponding co-factors. The other type of relation, which expresses A in terms of the elements of the j th column and the corresponding co-factors, is

$$A = \sum_{i=1}^n a_{ij} A_{ij} \quad (\text{for any } j=1, 2, \dots, n) \quad (1.19)$$

We state below a useful property of a determinant :

$$\left. \begin{array}{l} \sum_i a_{ij} A_{kj} = 0 \text{ for } i \neq k; \\ \sum_i a_{ij} A_{ik} = 0 \text{ for } j \neq k. \end{array} \right\} \dots \quad (1.20)$$

It is easy to verify that $|A| = |A'|$. When A is not a square matrix, the determinant of any square sub-matrix of A is called a *minor* of A .

The *rank* of a matrix A is the greatest integer r such that A has at least one minor of order r which is different from zero.

1a.11 Inverse matrix

A square matrix $A_{n \times n}$ with $|A| \neq 0$ is called a *non-singular matrix*. $A_{n \times n}$ has then rank n . $A_{n \times n}$ is called a *singular matrix* if $|A_{n \times n}| = 0$.

Non-singular matrices have corresponding *inverse matrices*. When A is non-singular, the matrix A^{-1} , defined by

$$A^{-1} = \left(\frac{A_{ji}}{A} \right), \dots \quad (1.21)$$

is called the *inverse* or *reciprocal* of A , where $\frac{A_{ji}}{A}$ is the i th row, j th column element of A^{-1} . Usually $\frac{A_{ji}}{A}$ is denoted by a^{ij} and A^{-1} is then represented by (a^{ij}) . Two useful properties of inverse matrices are :

(1) $|A^{-1}| = 1/|A|$, and (2) if A is symmetric, then so is A^{-1} .

For an orthogonal matrix G , we have $G^{-1} = G'$.

1a.12 Quadratic forms

The expression

$$\sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j, \dots \quad (1.22)$$

where $a_{ij} = a_{ji}$, is called a *quadratic form* in the variables x_1, x_2, \dots, x_n . In matrix notation, it is $\mathbf{x}' \mathbf{A} \mathbf{x}$, and is denoted by $Q(\mathbf{x})$ or $Q(x_1, x_2, \dots, x_n)$. A is the matrix of the form Q . The *rank of a quadratic form* is by definition the same as the rank of the associated matrix A . Incidentally, A is a symmetric matrix. If $A = I$,

$$Q(\mathbf{x}) = \mathbf{x}' \mathbf{I} \mathbf{x} = \sum_{i=1}^n x_i^2.$$

If we apply a linear transformation

$$x_i = \sum_{j=1}^m b_{ij} y_j, \quad i=1, 2, \dots, n,$$

which in matrix notation is $\mathbf{x} = \mathbf{B}\mathbf{y}$ with $\mathbf{B} = \mathbf{B}_{n \times m}$, to $Q(\mathbf{x})$, we get a $Q(\mathbf{y})$

$$Q(x_1, x_2, \dots, x_n) = \mathbf{x}' \mathbf{A} \mathbf{x} = \mathbf{y}' \mathbf{B}' \mathbf{A} \mathbf{B} \mathbf{y} = Q(y_1, y_2, \dots, y_m),$$

where $\mathbf{B}' \mathbf{A} \mathbf{B}$ is the matrix of the transformed form and is symmetric

If for real values of x_1, x_2, \dots, x_n ,

$$Q(x_1, x_2, \dots, x_n) \geq 0,$$

then the form is said to be *non-negative*. If in the above relation equality sign holds if and only if x_i 's are all equal to zero, then the form is said to be *positive definite*. A *positive semi-definite form* is a non-negative form which is not positive definite.

A linear transformation $\mathbf{x} = \mathbf{C}\mathbf{y}$, using an orthogonal matrix \mathbf{C} , is called an *orthogonal transformation*. Orthogonal transformations are frequently used in statistical theory.

1a 13 Linear equations

A simultaneous system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= k_1 \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= k_m \end{aligned} \quad (1.23)$$

of m linear equations in n unknowns x_1, x_2, \dots, x_n has for a solution a set of n quantities h_1, h_2, \dots, h_n which, on substitution in (1.23) reduce the left-hand sides of (1.23) to k_1, k_2, \dots, k_m .

In matrix notation, (1.23) can be written as

$$\mathbf{Ax} = \mathbf{k}, \quad (1.24)$$

where

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix},$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad \mathbf{k} = \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_m \end{pmatrix}.$$

When \mathbf{k} is the null vector, (1.24) is said to be a *homogeneous system*, and it is called *non-homogeneous* when \mathbf{k} is non-null.

Let $\text{rank } A = r \leq n$. Consider the non-homogeneous system with A a non-singular matrix. Then we obtain the unique solution of the system by pre-multiplying both sides of (1.24) by A^{-1} , which gives

$$\mathbf{x} = A^{-1} \mathbf{k}$$

or, in explicit form,

$$x_i = \frac{1}{|A|} \sum_{j=1}^m k_j A_{ji} \quad \text{for } i=1, 2, \dots, n. \quad \dots \quad (1.25)$$

This is known as *Cramer's rule*.

In case A is singular or $m \neq n$, the system will be *consistent*, i.e. will have a solution, if and only if

$$\text{rank } A = \text{rank } (A | \mathbf{k}),$$

where $(A | \mathbf{k})$ is the augmented matrix obtained from A by attaching the column of constants as an extra column. A system of homogeneous equations is always consistent, for

$$\text{rank } A = \text{rank } (A | \mathbf{0}),$$

\mathbf{k} being now the null vector. The solution $x_i = 0, i=1, 2, \dots, n$, which always exists for a homogeneous system, is called the *trivial* solution. Thus the important question for a homogeneous system is the existence of a *non-trivial* solution. A homogeneous system of linear equations has a non-trivial solution if, and only if, $\text{rank } A < n$.

1b CALCULUS

1b.1 Concept of a function

If x and y are two variable quantities such that the value of y depends on the value of x , then y is said to be a function of x .

Consider the following relations between x and y :

$$(i) \quad y = 2 + 3x + 6x^2,$$

$$(ii) \quad y = 4 \log x + 5,$$

$$(iii) \quad y = \cos x.$$

In each of the above cases, when x is given y can be determined, though in none of these cases we have the individual values of x and y . When such is the case, the expressions on the right-hand side

of the relations are called functions of x . For instance, the volume of a cube depends on the length of its side given the length of side, we know the volume exactly. Thus the volume of a cube is a function of the length of its side.

The variable x , on whose value y depends, is called the independent variable and y the dependent variable. It is customary to indicate that y is a function of x by writing the independent variable within brackets and prefixing a letter, e.g.

$$y=f(x) \text{ or } \phi(x) \text{ or } F(x), \text{ etc}$$

$f(a)$ denotes the value of y at $x=a$

1b 2 Limit of a function

If there is a definite value A to which $f(x)$ approaches as x approaches the value a , then A is called the limit of $f(x)$ as x tends to a or approaches a . This is symbolically denoted by

$$\lim_{x \rightarrow a} f(x) = A \quad \text{or} \quad \text{Lt}_{x \rightarrow a} f(x) = A$$

When we say x tends to a , we mean that the difference between x and a becomes smaller and smaller and ultimately smaller than any given positive quantity, however small. The above definition of limit means that the difference between $f(x)$ and A can be made arbitrarily small by making the difference between x and a sufficiently small. With the help of the modulus notation this is stated as follows:

If given $\epsilon > 0$ however small, a δ can be found such that

$$|f(x) - A| < \epsilon$$

whenever

$$0 < |x - a| < \delta,$$

then A is said to be the limit of $f(x)$ as x tends to a .

Consider, for instance,

$$\lim_{x \rightarrow 1} (ax + b)$$

When x tends to 1, ax tends to a . Thus $ax + b$ tends to $a + b$. So

$$\lim_{x \rightarrow 1} (ax + b) = a + b$$

1b.3 Meaning of infinity (∞)

The function $\frac{1}{x}$ is not defined for $x=0$. But when $x(>0)$ is made smaller and smaller, $\frac{1}{x}$ becomes greater and greater. Thus as $x \rightarrow 0$ through positive values, $\frac{1}{x}$ does not tend to any number but increases without bound. This phenomenon is expressed as

$$\lim_{x \rightarrow +0} \frac{1}{x} = \infty.$$

Similarly, as $x \rightarrow 0$ taking negative values, $\frac{1}{x}$ remains a negative quantity whose magnitude becomes greater and greater. In symbols,

$$\lim_{x \rightarrow -0} \frac{1}{x} = -\infty.$$

If $f(x)$ approaches A as x increases without bound, A is said to be the limit of $f(x)$ for x tending to ∞ :

$$\lim_{x \rightarrow \infty} f(x) = A.$$

1b.4 General results on limits

1. The limit of the sum (or difference) of a finite number of functions is equal to the sum (or difference) of the limits of the individual functions. For instance,

$$\lim \{f_1(x) + f_2(x) - f_3(x)\} = \lim f_1(x) + \lim f_2(x) - \lim f_3(x).$$

2. The limit of the product of a finite number of functions is the product of their limits. For instance,

$$\lim \{f_1(x) \cdot f_2(x)\} = \lim f_1(x) \cdot \lim f_2(x).$$

3. The limit of the quotient of two functions is the quotient of their limits, provided the limit of the denominator does not vanish:

$$\lim \frac{f_1(x)}{f_2(x)} = \frac{\lim f_1(x)}{\lim f_2(x)}, \text{ provided } \lim f_2(x) \neq 0.$$

1b.5 Some important limits

$$(i) \quad \lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x = e; \quad (ii) \quad \lim_{x \rightarrow 0} (1+x)^{1/x} = e;$$

$$(iii) \quad \lim_{x \rightarrow a} \frac{x^r - a^r}{x - a} = r a^{r-1} \text{ (where } r \text{ is a rational number).}$$

1b.6 Continuity

The function $f(x)$ is said to be *continuous at a point $x=a$* if $\lim_{x \rightarrow a} f(x) = f(a)$. Otherwise, $f(x)$ is said to be *discontinuous at a* .

The function $f(x)$ is *continuous between a and b* if it is continuous at all points in that interval.

The function $f(x) = r^2$ is a continuous function, while

$$f(x) = \begin{cases} \frac{1}{x} & \text{for } x > 0 \\ 0 & \text{for } x = 0 \end{cases}$$

is discontinuous at $x=0$.

1b.7 Derivative of a function

Given a function

$$y = f(x),$$

the limiting value

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h},$$

when it exists, is called the *derivative* or the *differential coefficient* of $f(x)$ with respect to x at the point x . When the limit does not exist, $f(x)$ is not differentiable for that value of x .

Various notations are used to denote the derivative of $f(x)$:

$$\frac{dy}{dx}, \quad \frac{df(x)}{dx} \quad \text{or} \quad f'(x)$$

The derivative of $y = x^n$ is

$$\lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} = n x^{n-1}$$

Again, the derivative of $y = e^x$ is

$$\lim_{h \rightarrow 0} \frac{e^{x+h} - e^x}{h} = e^x \quad \lim_{h \rightarrow 0} \frac{(e^h - 1)}{h} = e^x$$

If $f'(x)$ happens to be a differentiable function, then we can differentiate $f'(x)$ again as we did with $f(x)$ and get the derivative of $f'(x)$. This is the *second derivative* or *second differential coefficient* of $f(x)$ and is denoted by $\frac{d^2 f(x)}{dx^2}$, i.e. $\frac{d}{dx} \left(\frac{d}{dx} f(x) \right)$, or $\frac{d^2 f(x)}{dx^2}$ or $f''(x)$.

Similarly, we can define higher order derivatives.

1b.8 Partial derivatives

If $u=f(x, y, z, \dots)$ is a function of the variables x, y, z, \dots , then the limit

$$\lim_{h \rightarrow 0} \frac{f(x+h, y, z, \dots) - f(x, y, z, \dots)}{h},$$

if it exists when y, z, \dots are regarded as constants, is called the first partial derivative of f with respect to x and is denoted by $\frac{\partial f}{\partial x}$ or f'_x . Similarly, the first partial derivative of f with respect to y , when the other variables are held constant, is denoted by $\frac{\partial f}{\partial y}$ or f'_y . Similarly for the other partial derivatives. The first partial derivatives f'_x, f'_y, \dots are again functions of x, y, z, \dots , and if these functions have further partial derivatives with respect to x, y, \dots , we may obtain the successive partial derivatives of higher orders. Thus

$$\frac{\partial}{\partial x} \left(\frac{\partial f}{\partial x} \right) = \frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x} f'_x = f''_{xx}, \text{ say,}$$

$$\frac{\partial}{\partial y} \left(\frac{\partial f}{\partial x} \right) = \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial y} \right) = f''_{xy} = f''_{yx}, \text{ say,}$$

$$\frac{\partial}{\partial y} \left(\frac{\partial f}{\partial y} \right) = \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y} f'_y = f''_{yy}, \text{ say.}$$

and so on.

1b.9 Application of derivatives of functions

Here we shall consider the behaviour of functions that depend on derivatives. In many cases it is possible to locate the maximum, minimum or the point of inflection of a function with the help of its derivatives.

We state below some definitions and results in this regard that will be useful in our later statistical work.

Definition A function $f(x)$ is said to have a *local* (or *relative*) maximum at $x=c$ if

$$f(c) > f(c+h)$$

for sufficiently small values of h (both positive and negative). For a *local* (or *relative*) minimum at $x=c$

$$f(c) < f(c+h)$$

for sufficiently small values of h (both positive and negative) The function $f(x)$ will have an *absolute maximum* at $x=c$ in an interval (a, b) if $f(c) > f(x)$, for all x in (a, b) and not only for x close to c

Similarly, there is an *absolute minimum* at $x=c$ if

$$f(c) < f(x), \text{ for all } x \text{ in } (a, b)$$

Definition The point of a curve where it changes from concavity to convexity or vice versa is called a *point of inflection*

We state below some rules that are generally applicable in locating a stationary value (i.e. a maximum or a minimum or a point of inflection) at a point $x=c$. Also, as we are considering application of derivatives we assume that $y=f(x)$ is differentiable upto the required order in the interval of interest

Rule 1 Test each value of x for which $f'(x)=0$ to determine the nature of stationary value. The usual tests are

(a) If

$$f'(x) > 0 \text{ for } x < c$$

$$= 0 \text{ for } x = c$$

$$< 0 \text{ for } x > c,$$

then a maximum occurs at $x=c$. If $f'(x)$ changes from negative to positive through zero as x advances through c , then there is a minimum at $x=c$. If $f'(x)$ does not change sign as x passes through c then a maximum or a minimum need not occur at $x=c$

(b) If $f''(c) < 0$ when $f'(c)=0$, then a maximum occurs at $x=c$

If $f''(c) > 0$ when $f'(c)=0$ then a minimum occurs at $x=c$

If $f''(c)=0$ when $f'(c)=0$, the above test fails. But if $f''(c)=0$ and $f'''(x)$ changes sign as it passes through $x=c$, then $x=c$ is a point of inflection

(c) If $f^{(n)}(c) \neq 0$ (for $n \geq 2$) and all the previous derivatives of $f(x)$ vanish at $x=c$ then

$f(x)$ has a maximum at $x=c$ if n be even and $f^{(n)}(c) < 0$,

$f(x)$ has a minimum at $x=c$ if n be even and $f^{(n)}(c) > 0$

and $f(x)$ has a point of inflection at $x=c$ if n be odd

(d) If the derivative does not exist at a point then examine the graph of $y=f(x)$ near that point for a possible stationary value

Lagrange multipliers

For the case where a stationary value is desired subject to certain auxiliary conditions, we use a technique known as *Lagrange's method of undetermined multipliers*.

Suppose $y=f(x_1, x_2, \dots, x_n)$ is a function of n variables x_1, x_2, \dots, x_n and we want to find a stationary value of f subject to $m < n$ auxiliary conditions : $\phi_j(x_1, x_2, \dots, x_n) = 0, j=1, 2, \dots, m$.

We use m constants $L_j, j=1, 2, \dots, m$, called Lagrange multipliers and form the equation

$$F = f + \sum_{j=1}^m L_j \phi_j.$$

If the n first partial derivatives of F are required to vanish, i.e. if $\frac{\partial F}{\partial x_i} = 0, i=1, 2, \dots, n$, then these n equations along with the m conditions $\phi_j = 0, j=1, 2, \dots, m$, make it possible to determine the $(m+n)$ quantities, viz. $x_1, x_2, \dots, x_n ; L_1, L_2, \dots, L_m$, so that f attains a stationary value for the values of x_1, x_2, \dots, x_n so determined.

Then to decide on the nature of the stationary value we apply appropriate tests.

1b.10 Definite integral

Let $\phi(x)$ be bounded in the finite interval (a, b) . Let the interval (a, b) be divided in any manner into n sub-intervals, equal or not, of width h_1, h_2, \dots, h_n . In these sub-intervals, choose perfectly arbitrary points, say r_1, r_2, \dots, r_n , respectively. If as $n \rightarrow \infty$, with the width of the intervals tending to zero, the sum

$$h_1\phi(r_1) + h_2\phi(r_2) + \dots + h_n\phi(r_n)$$

tends to a limit, then that limit is called the *definite integral* of $\phi(x)$ between a and b and is denoted by

$$\int_a^b \phi(x) dx.$$

This is read as 'the integral $\phi(x)dx$ from a to b '.

Geometrically, it can be seen that $\int_a^b \phi(x) dx$ is nothing but the area bounded by the curve $y=\phi(x)$, the x -axis and the two ordinates at a and b .

1b 11 Infinite or improper integral

When the limits a and/or b are infinite or the integrand $\phi(x)$ has infinite discontinuity i.e tends to infinity at one or more points in the finite range (a, b) the integral is called an *infinite* or *improper* integral.

When the range is infinite the integral can be evaluated in the following manner

$$(i) \int_a^{\infty} \phi(x) dx = \lim_{b \rightarrow \infty} \int_a^b \phi(x) dx \quad b \text{ is any number greater than } a$$

$$(ii) \int_{-\infty}^b \phi(x) dx = \lim_{a \rightarrow -\infty} \int_a^b \phi(x) dx \quad a \text{ is any number less than } b$$

provided the limits exist. When these limits are finite we say that the infinite integrals exist or are *convergent*. But if the limits are $+\infty$ or $-\infty$ we say the infinite integrals do not exist or are *divergent*.

(iii) When for an arbitrary finite point c the infinite integrals

$$\int_c^{\infty} \phi'(x) dx \quad \text{and} \quad \int_{-\infty}^c \phi(x) dx$$

both converge in the above sense then we say that the infinite integral $\int_{-\infty}^{\infty} \phi(x) dx$ is convergent and is given by

$$\int_{-\infty}^{\infty} \phi(x) dx = \int_{-\infty}^c \phi(x) dx + \int_c^{\infty} \phi(x) dx \quad c \text{ being any intermediate finite point}$$

Next let us consider the case where the integrand $\phi(x)$ tends to infinity at a finite number of points.

If a is the only point of infinite discontinuity then the infinite integral $\int_a^b \phi(x) dx$ is said to exist or converge if

$$\lim_{\epsilon \rightarrow 0} \int_{a+\epsilon}^b \phi(x) dx \text{ exists and is finite for arbitrary } \epsilon > 0$$

Similarly, if b is the only point of infinite discontinuity of $\phi(x)$, then $\int_a^b \phi(x)dx$ is convergent if $\lim_{\epsilon \rightarrow 0} \int_a^{b-\epsilon} \phi(x)dx$ exists and is finite for arbitrary $\epsilon > 0$. If a, b are both points of infinite discontinuity, then the infinite integral is said to converge if

$$\int_a^b \phi(x)dx = \int_a^c \phi(x)dx + \int_c^b \phi(x)dx, \text{ for arbitrary } c \text{ in } (a, b),$$

provided the infinite integrals on the right-hand side exist and are finite. If $\phi(x)$ has infinite discontinuity at an intermediate point c , $a < c < b$, then the infinite integral is defined by

$$\int_a^b \phi(x)dx = \lim_{\epsilon \rightarrow 0} \int_a^{c-\epsilon} \phi(x)dx + \lim_{\epsilon' \rightarrow 0} \int_{c+\epsilon'}^b \phi(x)dx,$$

provided the limits on the right-hand side exist for arbitrary $\epsilon, \epsilon' > 0$.

Sometimes definite limits exist only for $\epsilon = \epsilon'$ and then the limit is called the *principal value* of the infinite integral.

1b.12 Indefinite integral

The process of finding the integral of a function is called integration. It can be shown that integration is the inverse of differentiation. Thus if $y=f(x)$ be a function with derivative $f'(x)$, then the integral of $f'(x)$ with respect to x between a and b is $f(b)-f(a)$. So, for integrating a function $\phi(x)$, we look for an expression $f(x)$ whose derivative is equal to that function. Then the function $f(x)$ is called the indefinite integral of $\phi(x)$ and is denoted by $\int \phi(x)dx$.

Now, since $\frac{d}{dx}\{f(x)+c\}=f'(x)$, where c is a constant, we find that the integral of $f'(x)$ is $f(x)+c$. Hence $f(x)+c$ also is an indefinite integral. The constant c must not be forgotten; it is because of its presence that we call the integral an indefinite one.

Symbolically, if

$$\frac{d}{dx}\{f(x)+c\}=f'(x)=\phi(x), \text{ say},$$

then $\int \phi(x)dx=f(x)+c$,

where $\phi(x)$ is called the integrand and $\int \phi(x)dx$ is read as the 'indefinite integral of $\phi(x)$ with respect to x '. The method we have considered for integration is a matter of inspired guessing, but there are certain rules of integration which are helpful and at the same time simple.

Some important integrals are

$$\int x^a dx = \frac{x^{a+1}}{a+1} + c, \text{ provided } a \neq -1$$

$$\int e^{ax} dx = \frac{e^{ax}}{a} + c,$$

$$\int \frac{dx}{x} = \log x + c$$

1b 13 Double and multiple integral

Let $f(x, y)$ be a function of the variables x and y . Suppose $f(x, y)$ is bounded for a set of points in a rectangle of the (x, y) plane $R = \{(a \leq x \leq b, c \leq y \leq d)\}$. Let us subdivide R into sub-rectangles P_{rs} , by lines parallel to the x - and y axes. Let M and m be the upper and lower bounds of $f(x, y)$ in R , while M_{rs} and m_{rs} are the corresponding bounds in P_{rs} .

We form the upper (U) and lower (L) Riemann sums

$$U = \sum \sum M_{rs} p_{rs}, \quad L = \sum \sum m_{rs} p_{rs},$$

where p_{rs} denotes the area of the corresponding sub-rectangle P_{rs} .

If the lower bound of the upper sums tends to the upper bound of the lower sums, which bounds exist, when the number of sub-rectangles is increased indefinitely such that the area of each sub-rectangle tends to zero, then we say $f(x, y)$ is integrable over R . The common value of the two bounds is the *double integral* of $f(x, y)$ over R and is denoted by

$$\int_R \int f(x, y) dx dy$$

Double integral as a repeated integral

If $\int_c^d \int_a^b f(x, y) dx dy$ exists and if

either (1) $\int_a^b f(x, y) dx$ exists for all y in (c, d)

or (2) $\int_c^d f(x, y) dy$ exists for all x in (a, b) ,

then $\int_c^d \int_a^b f(x, y) dx dy = \int_c^d dy \int_a^b f(x, y) dx$
 $= \int_a^b dx \int_c^d f(x, y) dy.$

Transformation of integral :

Integrals may easily be evaluated by a change of variables. Let us evaluate the integral

$$\int_R \int f(x_1, x_2) dx_1 dx_2$$

by a change of variables defined by

$$x_1 = g_1(y_1, y_2) \text{ and } x_2 = g_2(y_1, y_2).$$

Here we assume that the transformation is one-to-one and that the first partial derivatives exist.

Let S be the image of R in the (y_1, y_2) plane. Let J be the Jacobian, defined by the determinant

$$J = \frac{\partial(g_1, g_2)}{\partial(y_1, y_2)} = \begin{vmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} \\ \frac{\partial g_2}{\partial y_1} & \frac{\partial g_2}{\partial y_2} \end{vmatrix}.$$

The transformed integral is then

$$\int_R \int f(x_1, x_2) dx_1 dx_2 = \int_S \int f(x_1, x_2) |J| dy_1 dy_2,$$

where on the right-hand side x_1 and x_2 are to be replaced by

$$g_1(y_1, y_2) \text{ and } g_2(y_1, y_2).$$

The above results can be easily extended to the case of triple and multiple integrals.

Let $f(x_1, x_2, \dots, x_n)$ be an integrable function of the n independent variables x_1, x_2, \dots, x_n ; then the n -ple integral is defined by

$$\int \int \dots \int_R f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n,$$

where R is an n -dimensional rectangle in the (x_1, x_2, \dots, x_n) space.

Let us apply the one-to-one transformation :

$$x_r = g_r(y_1, y_2, \dots, y_n), \quad r=1, 2, \dots, n,$$

where the first partial derivatives of g_r exist. Then the Jacobian is

$$J = \frac{\partial(x_1, \dots, x_n)}{\partial(y_1, \dots, y_n)} = \begin{vmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \dots & \dots & \frac{\partial g_1}{\partial y_n} \\ \vdots & \vdots & & & \vdots \\ \frac{\partial g_n}{\partial y_1} & \frac{\partial g_n}{\partial y_2} & \dots & \dots & \frac{\partial g_n}{\partial y_n} \end{vmatrix},$$

and

$$\begin{aligned} & \int \int \int_R f(x_1, \dots, x_n) dx_1 \dots dx_n \\ &= \int \int \int_S f(x_1, \dots, x_n) |J| dy_1 \dots dy_n, \end{aligned}$$

where S is the image of R in the (y_1, y_2, \dots, y_n) space and (x_1, x_2, \dots, x_n) on the right are to be replaced by (g_1, g_2, \dots, g_n) .

1b.14 Some special integrals

Beta function :

The definite integral

$$\int_0^1 x^{m-1} (1-x)^{n-1} dx$$

is convergent for $m > 0, n > 0$

We denote this, a function of m and n , by $B(m, n)$ and call it the *beta function* or *first Eulerian integral*. Thus

$$B(m, n) = \int_0^1 x^{m-1} (1-x)^{n-1} dx, \quad m > 0, n > 0. \quad \dots \quad (I 26)$$

Properties of beta function :

(1) If we write $y=1-x$, then

$$\begin{aligned} B(m, n) &= \int_0^1 x^{m-1}(1-x)^{n-1} dx \\ &= \int_0^1 y^{n-1}(1-y)^{m-1} dy \\ &= B(n, m). \end{aligned}$$

(2) Substituting $x=\sin^2\theta$, so that $dx=2\sin\theta\cos\theta d\theta$, $B(m, n)$ may be written in the following form :

$$B(m, n)=2 \int_0^{\pi/2} \sin^{2m-1}\theta \cos^{2n-1}\theta d\theta.$$

Gamma function :

The integral

$$\int_0^\infty \exp[-x] x^{n-1} dx$$

is convergent for $n>0$. This integral, a function of n , is denoted by $\Gamma(n)$ and is called the *gamma function* or *second Eulerian integral*. Thus

$$\Gamma(n)=\int_0^\infty \exp[-x] x^{n-1} dx, \quad n>0. \quad \dots \quad (1.27)$$

Properties of gamma function :

(1) If we integrate by parts, we get

$$\Gamma(n)=(n-1)\Gamma(n-1), \quad \dots \quad (1.28)$$

from which it follows that if n is a positive integer, then

$$\Gamma(n)=(n-1)!,$$

while if $n>0$ is not an integer, then

$$\Gamma(n)=(n-1)(n-2)\dots(p)\Gamma(p),$$

where $0 < p < 1$.

$$(2) \quad B(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \text{ for } m > 0, n > 0$$

We have

$$\begin{aligned} \Gamma(m)\Gamma(n) &= \left(\int_0^\infty \exp[-x] x^{m-1} dx \right) \left(\int_0^\infty \exp[-y] y^{n-1} dy \right) \\ &= \int_0^\infty \int_0^\infty \exp[-(x+y)] x^{m-1} y^{n-1} dx dy \end{aligned}$$

Applying the transformation

$$\begin{aligned} x &= uv \\ y &= v(1-u), \end{aligned} \quad \left\{ \begin{array}{l} \end{array} \right.$$

which has the Jacobian

$$\frac{\partial(x, y)}{\partial(u, v)} = v,$$

we obtain

$$\begin{aligned} \Gamma(m)\Gamma(n) &= \int_0^\infty \int_0^1 u^{m-1} (1-u)^{n-1} v^{m+n-1} \exp[-v] du dv \\ &\quad (\text{since } 0 < x < \infty, 0 < y < \infty \text{ if and only if } 0 < u < 1, 0 < v < \infty) \\ &= \left(\int_0^\infty \exp[-v] v^{m+n-1} dv \right) \left(\int_0^1 u^{m-1} (1-u)^{n-1} du \right) \\ &= \Gamma(m+n) B(m, n), \end{aligned}$$

which shows that

$$B(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \quad (1.29)$$

$$(3) \quad \int_0^\infty \exp[-ax] x^{n-1} dx = \frac{\Gamma(n)}{a^n}, \quad n > 0, a > 0$$

On substituting

we have

$$\int_0^\infty \exp[-ax] x^{n-1} dx = \frac{1}{a^n} \int_0^\infty \exp[-y] y^{n-1} dy$$

$$\frac{\Gamma(n)}{a^n}.$$

(4) We have

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \exp[-x] x^{-1/2} dx.$$

Hence

$$\{\Gamma\left(\frac{1}{2}\right)\}^2 = \int_0^\infty \int_0^\infty \exp[-(x+y)] x^{-1/2} y^{-1/2} dx dy.$$

On making the transformation

$$\begin{cases} x = z \cos^2 \theta \\ y = z \sin^2 \theta \end{cases} \quad (0 < z < \infty, 0 < \theta < \pi/2),$$

for which the Jacobian is

$$\frac{\partial(x, y)}{\partial(z, \theta)} = 2z \sin \theta \cos \theta,$$

we have

$$\begin{aligned} \{\Gamma\left(\frac{1}{2}\right)\}^2 &= 2 \int_0^\infty \int_0^{\pi/2} \exp[-z] d\theta dz \\ &= 2 \left[\int_0^\infty \exp[-z] dz \right] \left[\int_0^{\pi/2} d\theta \right] = \pi. \end{aligned}$$

Hence

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Dirichlet integral :

$$\int \dots \int \int dx_1 dx_2 \dots dx_n = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)}$$

Consider the transformation

$$\left. \begin{array}{l} x_1 = \cos \theta_1, \\ x_2 = \sin \theta_1 \cos \theta_2, \\ x_3 = \sin \theta_1 \sin \theta_2 \cos \theta_3, \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ x_n = \sin \theta_1 \sin \theta_2 \dots \sin \theta_{n-1} \cos \theta_n. \end{array} \right\}$$

The Jacobian of the transformation is $(-1)^n (\sin \theta_1)^n (\sin \theta_2)^{n-1} \dots \sin \theta_n$, and $0 < \theta_i < \pi$ ($i = 1, 2, \dots, n$) is the image of the n -dimensional unit sphere $\sum_i x_i^2 < 1$.

Using the result that $\int_0^\pi (\sin \theta)^n d\theta$ equals

$$2 \int_0^{\pi/2} (\sin \theta)^n d\theta = B\left(\frac{n+1}{2}, \frac{1}{2}\right) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} \sqrt{\pi},$$

we have

$$\begin{aligned} & \int_{\sum x_i^2 < 1} \int \int dx_1 dx_2 \dots dx_n \\ &= \int_0^\pi (\sin \theta_1)^n d\theta_1 \int_0^\pi (\sin \theta_2)^{n-1} d\theta_2 \dots \int_0^\pi \sin \theta_n d\theta_n \\ &= \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} \end{aligned}$$

1e STIRLING'S APPROXIMATION

We are often faced with the problem of evaluating factorial expressions which become laborious to compute from definition. However, for large n there exists a useful approximation to $n!$, which is due to Stirling. It is

$$n! \approx \sqrt{2\pi} \exp[-n] n^{n+1/2} \quad (1.30)$$

A better approximation is given by

$$n! \approx \sqrt{2\pi} \exp[-n + 1/12n] n^{n+1/2} \quad (1.30a)$$

In each case the percentage error decreases as n increases.

Many mathematical and scientific problems are such that they cannot be solved by existing analytical methods, or such solutions even when they exist are so complex that they will not lead to the desired numerical information. In such cases the desired result may be obtained by purely numerical methods. Thus the numerical methods are concerned with the practical methods of obtaining an approximate solution to the stated problem which is correct to a certain degree of accuracy.

The numerical data used in solving problems are usually not exact. They are usually correct to a certain number of decimal places. Not only are the data approximate, the methods and processes that are applied for getting a numerical solution are also approximate. So it is evident that our results will not be exact due to the approximate nature of the data and of the methods. Errors of data cannot be avoided, but the errors of calculations may be made small.

In the first section, we shall give some idea of the error and approximations in numerical calculations. And in subsequent sections, we shall consider different numerical methods for solving problems of various types.

2a INACCURACIES AND APPROXIMATIONS

2a.1 Different types of inaccuracies

We differentiate between different types of inaccuracies. A *blunder* will be usually committed by a person who is not familiar with the method. It is a gross inaccuracy. But even when we are familiar with the method and type of computation involved, we may make *mistakes*. Some of the common mistakes are the following :

- (i) *Copying mistakes*, e.g. writing 23,480 as 23,408.
- (ii) *Mistakes in decimal points*, e.g. writing $12\cdot3/30$ as 4·1.
- (iii) *Incorrectly reading a table*, e.g. reading from a wrong column.
- (iv) *Faulty memorisation of values of constants*, such as π , $\sqrt{2}$, etc.
- (v) *Mistakes that occur with an untidy and careless worker*.

Whenever a mistake is found, it should be carefully corrected. Whenever possible we should check the calculations at all stages. Neat work in appropriate tabular form reduces the chances of mistakes.

If we know our job and are careful, then we may avoid blunders and mistakes. But even then there will be a third type of inaccuracy, called *errors*, which will be sometimes impossible to avoid. Errors may arise from the following sources:

(i) *Errors in original data* The observations taken may have been rounded off and so are approximate.

(ii) *Errors due to the approximate nature of the formula* The mathematical formulation is mostly an idealised description of the problem. Sometimes we replace an infinite series by a finite number of terms.

(iii) *Rounding off errors* This is unavoidable in some calculations. For some of the quantities when expressed in decimals will be non terminating and also the capacity of the computing machines is limited. So in such cases we are to terminate after a number of places. These errors may accumulate as calculations proceed.

2a 2 Rounding off

Some quantities may not terminate and some may terminate after a large number of decimal places. Owing to the limited capacity of a computing machine, we can retain only a few of the digits. The process of cutting off superfluous digits and retaining only the desired number of digits is called rounding off. To round off a number to n digits we replace all digits to the right of the n th digit (counting from the left) by zero. If the discarded number contributes less than half a unit in the n th place, we leave the n th digit unchanged; if the discarded number contributes more than half a unit, we increase the n th digit by unity. In case the discarded number is exactly half a unit in the n th place, we leave the n th digit unaltered if it is even, but increase it by unity if it is odd. The numbers 27,598, 1 467205, 45 765, 2 0675 when rounded off to four digits become 27600, 1 467, 45 76, 2 068, respectively.

When the above rule is followed, the rounding off errors will be mostly eliminated by cancellations.

2a.3 Significant figures

The digits 1, 2, 3, , 9 are significant figures. 0 is also a significant figure if it is not used to fix the decimal point or to denote the discarded digits. The number of significant figures in a number expressed in the decimal form refers to the number of digits, starting from the left with the first non-zero digit and proceeding to the right that are assumed to be correct. Thus, in the number 0.0001408 the number of significant figures is only four, viz. 1, 4, 8 and 0 (between 4 and 8). The three 0's before the first significant figure 1 are not significant figures, for they are needed to fix the decimal point. In a number like 64,200 there is nothing to tell us whether the two 0's at the end are significant figures. To specify the number of significant figures in such cases it is customary to write it in the form 6.42×10^4 , 6.420×10^4 , 6.4200×10^4 to indicate that the number of significant figures is three, four and five, respectively.

2a.4 Absolute, relative and percentage errors

If m is the approximate value of a quantity whose true value (not necessarily known) is m^* , then the *absolute error modulus* of m is defined as

$$|e| = |m - m^*|. \quad \dots \quad (2.1)$$

It is the British practice to define the *absolute error* of m as

$$e = m - m^*, \quad \dots \quad (2.2)$$

while in other countries it is defined by

$$e = m^* - m. \quad \dots \quad (2.3)$$

Usually the magnitude of e is more important than its sign.

The absolute error has the dimension of the quantity, and a better measure of error is given by the dimensionless quantity r , called the *relative error*, defined by

$$r = \frac{|m - m^*|}{|m^*|}. \quad \dots \quad (2.4)$$

A somewhat similar measure is

$$\frac{|m - m^*|}{|m|}. \quad \dots \quad (2.4a)$$

The *percentage error* is $100r$.

The absolute error of a number correct to n significant figures cannot be more than half a unit in the last significant figure. For example, if 1 467 is correct to four significant figures, then its absolute error is not greater than 0.0005.

Let $m^* = f(m_1^*, m_2^*, \dots, m_k^*)$ denote a function of several independent quantities $m_1^*, m_2^*, \dots, m_k^*$. If m^* is evaluated by using approximate quantities m_1, m_2, \dots, m_k , which are subject to absolute errors $\epsilon_1, \epsilon_2, \dots, \epsilon_k$, respectively, then these errors will cause an error ϵ in the function m^* . By Taylor's theorem for a function of several variables, we get

$$m = f(m_1, m_2, \dots, m_k) = f(m_1^*, m_2^*, \dots, m_k^*) + \sum_{i=1}^k (m_i - m_i^*) f_i^0$$

(neglecting squares, products and higher powers of $m_i - m_i^*$'s) where f_i^0 is the value of the partial derivative $\partial f / \partial m_i$ evaluated at the point $(m_1^*, m_2^*, \dots, m_k^*)$. Then the absolute error ϵ in m is given approximately by

$$\epsilon = \sum_{i=1}^k \epsilon_i f_i^0 \quad (25)$$

The absolute error modulus of m is

$$|\epsilon| \leq \sum_{i=1}^k |\epsilon_i f_i^0| \quad (26)$$

This is the general result for determining the error of a function, and as corollaries we obtain the following

The absolute error modulus in a sum (difference) is less than or equal to the sum of the absolute error moduli of the separate terms.

The relative error in a product (quotient) is less than or equal to the sum of the relative errors in the various factors.

The following two fundamental theorems proved in [5] give the relation between the relative error and the number of significant figures in a number.

THEOREM 2.1 If the first significant figure of a number is k and the number is correct to n significant figures, then the relative error is less than $1/(k \times 10^{n-1})$.

The converse result is given in the other theorem.

THEOREM 2.2 If the relative error in an approximate number is less than $1/[(k+1) \times 10^{n-1}]$, then the number is correct to n significant figures or is at least in error by less than a unit in the n th significant figure.

Here also k is the first significant figure of the number.

In practice, the exact value of the absolute error modulus is not known, but it is known that for a number which has been rounded off to n decimals, this is not greater than $\frac{1}{2} \times 10^{-n}$. The errors as given by the above rules are the upper bounds of the errors that may occur in a computation. And these will be attained if the different errors reinforce each other. Sometimes the signs of errors will be such as to cancel each other and the actual error will be less than this upper bound.

We show below by examples how to calculate the sum, difference, product or quotient of numbers of various accuracies.

Ex. 2.1 Let us calculate the greatest value of the absolute error modulus of each of the expressions given below and then round off the results to the appropriate number of figures. Each one of the numbers is rounded off.

- (1) $4.12 + 0.768 - 2.71345$,
- (2) 2.145×0.45 ,
- (3) $0.468 / 4.3712$.

(1) In this case the greatest value of the absolute error modulus is $|e| \leq |e_1| + |e_2| + |e_3|$. The numbers 4.12 , 0.768 and -2.71345 are rounded off. So $|e_1| \leq \frac{1}{2} \times 10^{-2}$, $|e_2| \leq \frac{1}{2} \times 10^{-3}$ and $|e_3| \leq \frac{1}{2} \times 10^{-5}$. Note that the main contribution to $|e|$ is due to $|e_1|$, which is rounded off to two decimals only.

Now

$$|e| \leq \frac{1}{2} \times 10^{-2} + \frac{1}{2} \times 10^{-3} + \frac{1}{2} \times 10^{-5} = 0.5505 \times 10^{-2}.$$

As $4.12 + 0.768 - 2.71345 = 2.17455$, so the true value lies in the range 2.17455 ± 0.005505 , i.e. lies between 2.169045 and 2.180055 . So the result may be correctly rounded off to 2.2 or it may be rounded off to 2.17 with a possible error of one unit in the second decimal place.

Thus we find that in a sum (or difference) we do not get our result correct even up to the number of decimals of the number having smallest accuracy, so it will not pay to retain too many places in the

other numbers which may be more accurate. In such cases, we retain one more place in all the numbers than the number of places in the least accurate number and finally round off the result to the same place as that of the least accurate one and our result may be in error by one unit in that place.

(2) The product is 0.96525, and the maximum value of the relative error is

$$(\frac{1}{2} \times 10^{-3})/2.145 + (\frac{1}{2} \times 10^{-3})/0.45 = 0.0002 + 0.0111 = 0.0113$$

while the absolute error modulus is $\leq 0.965 \times 0.0113 = 0.0109$.

Thus the product lies between 0.9652 ± 0.0109 , i.e. between 0.9543 and 0.9761, and so the product may be correctly rounded off to 1.0 or may be rounded off to 0.97 with a possible error of two units in the second decimal place. Here the factor with the larger relative error is 0.45 and it determines the number of figures that will be trustworthy in the final result.

(3) The quotient is 0.10706, and the maximum value of the relative error is

$$(\frac{1}{2} \times 10^{-3})/0.468 + (\frac{1}{2} \times 10^{-4})/4.3712 = 0.00106 + 0.00001 = 0.00107$$

while the absolute error modulus will be $(0.107)(0.00107) = 0.00011$.

Thus the quotient lies in the range of 0.10706 ± 0.00011 , i.e. 0.10695 to 0.10717, so the quotient may be correctly rounded off to 0.107.

Errors sometimes occur from the loss of significant figures by subtraction of almost equal numbers unless we retain in the beginning more number of significant figures. We demonstrate this with the help of the following example.

Ex. 2.2 Find the difference $\sqrt{2.1} - \sqrt{2}$ correct to three significant figures.

We take

$$\sqrt{2.1} = 1.44914 \text{ and } \sqrt{2} = 1.41421.$$

Then

$$\sqrt{2.1} - \sqrt{2} = 1.44914 - 1.41421 = 0.03493.$$

The result is then 0.0349. Note that to start with we took more significant figures than are needed in the final result. If we took in the beginning only three significant figures, the first significant figure of the result will be in error and we would also not get the required number of significant figures.

2b INTERPOLATION**2b.1 The problem of interpolation**

The following table gives the values of $\log x$ (with base 10) corresponding to certain values of x :

x	$\log x$
95	1.9777236
96	1.9822712
97	1.9867717
98	1.9912261
99	1.9956352

Suppose in a certain computational work it is required to obtain the value of $\log 96.45$. It is not directly available from the above table, but can be obtained by a process known as *interpolation*.

The general problem can be stated as follows :

A function $y=f(x)$ is known for certain values of the argument* x_1, x_2, \dots, x_n , the corresponding values of $f(x)$ being $f(x_1), f(x_2), \dots, f(x_n)$. We are to find the value of $f(x)$ for some other value of the argument lying between x_1 and x_n .

The function will usually be of an unknown form or, even if known, of a complicated nature. Hence to solve the problem of interpolation we replace $f(x)$ by a simpler function, say $\phi(x)$, of known form. $\phi(x)$ is so chosen that

$$\phi(x_1) = f(x_1), \phi(x_2) = f(x_2), \dots, \phi(x_n) = f(x_n).$$

Then, for any other value of x , say x' , the value of $\phi(x')$ is taken to be an approximation to $f(x')$.

The function $\phi(x)$ is known as an interpolation formula.

We shall consider only the following particular form of $\phi(x)$:

$$\phi(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n,$$

where n is a positive integer and $a_n \neq 0$. $\phi(x)$ as defined here is known as a *polynomial* or a *rational integral function* in x of degree n .

The justification for the replacement of $f(x)$ by a polynomial $\phi(x)$ lies in an important theorem due to Weierstrass, which says that if $f(x)$ is continuous between x_1 and x_n , then it can be replaced by a

*In problems of interpolation, the function is generally referred to as the *entry* and the independent variable as the *argument*.

polynomial of suitable degree in that interval so that the difference between $f(x)$ and $\phi(x)$ will be as small as desired

2b 2 Finite differences

When the values of the argument are equidistant, the process of interpolation is facilitated by the use of what are called finite differences

Let h be some positive constant which we shall call the difference interval. Then, by definition,

$$\Delta f(x) = f(x+h) - f(x) \quad (27)$$

is the first difference of $f(x)$. This is also sometimes referred to as the first descending (or forward) difference of $f(x)$. Here Δ represents an operation and is not a quantity

Similarly,

$$\begin{aligned}\Delta^2 f(x) &= \Delta\{\Delta f(x)\} \\ &= \Delta f(x+h) - \Delta f(x) \\ &= f(x+2h) - 2f(x+h) + f(x)\end{aligned}$$

is the second difference of $f(x)$. The higher order differences—the third, fourth, etc.—are defined in a similar manner

$$\Delta^r f(x) = \{\Delta \Delta \dots \Delta \text{ } r \text{ times}\} f(x) \quad (28)$$

It must be always understood that Δ^r is not the r th power of a quantity, but simply the repetition of the operation represented by Δ r times, where r is a positive integer

The differences of various orders of the function $f(x)$ may be systematically obtained by constructing a diagonal difference table as follows

Argument	Entry	1st difference	2nd difference	3rd difference
x	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$
a	$f(a)$			
$a+h$	$f(a+h)$	$\Delta f(a)$		
$a+2h$	$f(a+2h)$	$\Delta f(a+h)$	$\Delta^2 f(a)$	
$a+3h$	$f(a+3h)$	$\Delta f(a+2h)$	$\Delta^2 f(a+h)$	$\Delta^3 f(a)$

In the above table, $f(a)$ is called the leading term and $\Delta f(a)$, $\Delta^2 f(a)$, $\Delta^3 f(a)$, the leading differences

It may be noted that to obtain $\Delta^r f(a)$ it is essential to know the values $f(a), f(a+h), \dots, f(a+rh)$. It is quite easy to form a difference table for a set of data, which is illustrated in the following example.

Ex. 2.3 Compute the difference table.

x	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$
2	17				
4	2	-15			
6	1	-1	14		
8	2	1	2	-12	
10	17	15	12	24	
12	82	65	36		

Let us consider a polynomial of degree n :

$$\phi(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n.$$

Then

$$\begin{aligned} \Delta\phi(x) &= \phi(x+h) - \phi(x) \\ &= a_1h + a_2\{(x+h)^2 - x^2\} + a_3\{(x+h)^3 - x^3\} + \dots \\ &\quad + a_n\{(x+h)^n - x^n\}, \end{aligned} \quad \dots \quad (2.9)$$

which is a polynomial of degree $(n-1)$.

The second difference $\Delta^2\phi(x)$, being the first difference of a polynomial of degree $(n-1)$, is a polynomial of degree $(n-2)$. Thus by taking difference once, the degree of the polynomial is reduced by unity. Hence $\Delta^n\phi(x)$ will be a polynomial of degree zero, i.e. a constant. $\Delta^{n+1}\phi(x)$ and higher order differences are all zeros.

This result, together with its converse, which also can be shown to be true, will be useful in solving interpolation problems.

2b.3 Error in a tabular value

Let y_0, y_1, \dots, y_{10} be the correct values of the entry $y=f(x)$ corresponding to the equidistant values x_0, x_1, \dots, x_{10} of the argument x . Suppose further that the central value y_5 is in error and the value recorded for it is $y_5 + \epsilon$. We construct the diagonal

difference table and exhibit the manner in which this error in y_5 is propagated through the different differences.

x	$y=f(x)$	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
x_0	y_0	Δy_0				
x_1	y_1	Δy_1	$\Delta^2 y_0$	$\Delta^3 y_0$		
x_2	y_2	Δy_2	$\Delta^2 y_1$	$\Delta^3 y_1$	$\Delta^4 y_0$	
x_3	y_3	Δy_3	$\Delta^2 y_2$	$\Delta^3 y_2 + \epsilon$	$\Delta^4 y_1 + \epsilon$	$\Delta^5 y_0 - 5\epsilon$
x_4	y_4	$\Delta y_4 + \epsilon$	$\Delta^2 y_3 + \epsilon$	$\Delta^3 y_3 - 3\epsilon$	$\Delta^4 y_2 - 4\epsilon$	$\Delta^5 y_1 + 10\epsilon$
x_5	$y_5 + \epsilon$	$\Delta y_5 - \epsilon$	$\Delta^2 y_4 - 2\epsilon$	$\Delta^3 y_4 + 3\epsilon$	$\Delta^4 y_3 + 6\epsilon$	$\Delta^5 y_2 - 10\epsilon$
x_6	y_6	Δy_6	$\Delta^2 y_5 + \epsilon$	$\Delta^3 y_5 - \epsilon$	$\Delta^4 y_4 - 4\epsilon$	$\Delta^5 y_3 + 5\epsilon$
x_7	y_7	Δy_7	$\Delta^2 y_6$	$\Delta^3 y_6$	$\Delta^4 y_5 + \epsilon$	$\Delta^5 y_4 - \epsilon$
x_8	y_8	Δy_8	$\Delta^2 y_7$	$\Delta^3 y_7$	$\Delta^4 y_6$	
x_9	y_9	Δy_9				
x_{10}	y_{11}					

We observe the following facts from the above table :

More and more differences are affected by the error ϵ as we take differences of higher orders. The maximum absolute error is along the line through y_5 or nearest this line on either side. The coefficients of ϵ 's are alternately positive and negative with binomial coefficients. The algebraic sum of the errors in each difference column is zero. If the function $y=f(x)$ be a polynomial of degree 4, then $\Delta^5 y$ values will be solely due to error and they will be symmetrical around the line through y_5 . Also sum of $\Delta^5 y$ values will be zero.

So we work out the differences till we come to a stage where the values are alternating in signs and symmetrical about a value, and their algebraic sum is zero. Then the line through the maximum absolute error or midway between the two largest absolute errors will locate the error in the entry. It is corrected by estimating ϵ from the last difference column which is solely due to error and subtracting this from $y_5 + \epsilon$, since $(y_5 + \epsilon) - \epsilon = y_5$, the true value.

We have considered above the simplest case, and it will be clear from the above difference table that if the error is not in the central part of the table and in each difference column we do not get all the differences affected by ϵ , then the observations we have made in the preceding paragraph will not be realised. There will be further complications if more than one value is in error. Bearing these in mind we illustrate the method with a simple case.

Ex. 2.4 Let us locate and correct an error in the values of a function, using the following difference table :

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
-1	-25	6			
1	1	2	-24	48	
3	3	26	24	54	6
5	29	104	78	30	-24
7	133	212	108	66	36
9	345	386	174	42	-24
11	731	602	216	48	6
13	1333	866	264		
15	2199				

We find that the $\Delta^4 y$ values are symmetrical about the line through $f(7)=133$, are alternating in sign and the algebraic sum of the $\Delta^4 y$ values is zero. Hence we conclude that the $\Delta^4 y$'s are solely due to error in $f(7)$. By referring to the earlier table, we find that the values in $\Delta^4 y$ column are $\epsilon, -4\epsilon, +6\epsilon, -4\epsilon$ and ϵ , respectively. Hence estimating ϵ from any one of these, say from $6\epsilon=36$, we get $\epsilon=6$, and the corrected value of $f(7)$ is $133-6=127$.

2b.4 Use of operators : Δ, E

We have already introduced the operator Δ in Section 2b.2. We now define the second operator E as

$$Ef(x) = f(x+h), \text{ where } h \text{ is the difference interval ;}$$

$$E^2 f(x) = E\{Ef(x)\} = f(x+2h) ;$$

and so on ;

$$\begin{aligned} E^n f(x) &= \{EE \dots n \text{ times}\} f(x) \\ &= f(x+nh). \end{aligned}$$

More generally, we shall denote $f(x+uh)$ by $E^u f(x)$, where u may be positive, negative or fractional.

Since

$$\begin{aligned} \Delta f(x) &= f(x+h) - f(x) \\ &= Ef(x) - f(x) = (E-1)f(x), \end{aligned}$$

one may take $\Delta \equiv E-1$ or $E \equiv 1+\Delta$. The relation, it should be noted, is one of equivalence and not of equality

From these, we have the general equivalence relations :

$$E^n \equiv (1+\Delta)^n \text{ and } \Delta^n \equiv (E-1)^n.$$

Further, it can be proved that for any $f(x)$ and n a positive integer

$$f(x+nh) = E^n f(x) = (1+\Delta)^n f(x)$$

$$= f(x) + \binom{n}{1} \Delta f(x) + \binom{n}{2} \Delta^2 f(x) + \dots + \binom{n}{n} \Delta^n f(x),$$

$$\text{and } \Delta^n f(x) = (E-1)^n f(x)$$

$$\begin{aligned} &= f(x+nh) - \binom{n}{1} f(x+n-1h) + \binom{n}{2} f(x+n-2h) + \dots \\ &\quad + (-1)^n \binom{n}{n} f(x). \end{aligned}$$

The above binomial expansions are also valid for any n , negative or fractional, if $f(x)$ is a polynomial in x .

Although Δ and E are not numbers, they satisfy the ordinary laws of algebra—the law of commutation, law of distribution and law of indices. Many problems may be solved by using these operators

Ex. 25 The values of a function $f(x)$ are given below for equidistant values of x

x	$f(x)$
4	3 11
5	2 96
6	2 85
7	
8	2 70

The value of $f(7)$ is unknown. Obtain as best an approximation as possible for $f(7)$.

Since only four values of $f(x)$ are given, we shall assume that $f(x)$ is a polynomial of third degree ; as a consequence, $\Delta^4 f(x)$ is to be regarded as zero. Now

$$\Delta^4 f(4) = 0,$$

i.e. $(E-1)^4 f(4) = 0,$

or $(E^4 - 4E^3 + 6E^2 - 4E + 1) f(4) = 0,$

or $f(8) - 4f(7) + 6f(6) - 4f(5) + f(4) = 0.$

Hence

$$f(7) = \frac{f(8) + 6f(6) - 4f(5) + f(4)}{4} = 2.77.$$

This is a problem of *missing term*. See also *Exercise 2.22*.

Ex. 2.6 The table below gives the average number of years of life (e_x^0) remaining to persons who survive to exact age x , for male African population of Belgian Congo :

x	e_x^0
0	37.64
5	44.04
10	41.40
15	37.78
20	34.41

Obtain e_x^0 for $x=1, 2, 3$ and 4 .

These can be obtained by applying Newton's forward interpolation formula. But when from the given table with equidistant values of the argument we want to form a new table with finer intervals, the new values of the argument also being equidistant, we would adopt the following simpler procedure.

Let $\Delta e_x^0, \Delta^2 e_x^0, \dots$ denote the differences for the given table with interval 5 and $\delta e_x^0, \delta^2 e_x^0, \dots$, etc., denote the differences for the new table to be formed with interval 1. Then using the properties of Δ and E , we have for the given table

$$e_{x+5}^0 = (1 + \Delta) e_x^0$$

and for the new table to be formed

$$e_{x+5}^0 = (1 + \delta)^5 e_x^0,$$

so that

$$(1 + \Delta) e_x^0 = (1 + \delta)^5 e_x^0.$$

That is,

$$1 + \Delta = (1 + \delta)^5,$$

or $\delta = (1 + \Delta)^{1/5} - 1,$

or $\delta = 2\Delta - 08\Delta^2 + 048\Delta^3 - 0336\Delta^4 +$

Again,

$$\delta^2 = (2\Delta - 08\Delta^2 + 048\Delta^3 -)^2,$$

or $\delta^2 = 04\Delta^2 - 032\Delta^3 + 02\ 6\Delta^4 -$

Similarly,

$$\delta^3 = 008\Delta^3 - 0096\Delta^4 + ,$$

and $\delta^4 = 0016\Delta^4 - ,$

and so on

The leading differences for the given table are

$$\Delta e_0^n = 6\ 40,$$

$$\Delta^2 e_0^n = -9\ 04,$$

$$\Delta^3 e_0^n = 8\ 06$$

and $\Delta^4 e_0^n = 6\ 83$

Here the fourth order differences $\Delta^4 e_0^n$ are assumed to be constant, which implies that $\delta^4 e_0^n$ are also constant. In that case,

$$\begin{aligned}\delta e_0^n &= (2\Delta - 08\Delta^2 + 048\Delta^3 - 0336\Delta^4)e_0^n \\ &= 2\Delta e_0^n - 08\Delta^2 e_0^n + 048\Delta^3 e_0^n - 0336\Delta^4 e_0^n \\ &= 2\ 619568,\end{aligned}$$

$$\begin{aligned}\delta^2 e_0^n &= (04\Delta^2 - 032\Delta^3 + 0256\Delta^4)e_0^n \\ &= 04\Delta^2 e_0^n - 032\Delta^3 e_0^n + 0256\Delta^4 e_0^n \\ &= -0\ 794368\end{aligned}$$

$$\begin{aligned}\delta^3 e_0^n &= (008\Delta^3 - 0096\Delta^4)e_0^n \\ &= 008\Delta^3 e_0^n - 0096\Delta^4 e_0^n \\ &= 0\ 130048,\end{aligned}$$

and $\delta^4 e_0^n = 0016\Delta^4 e_0^n = -0\ 010928$

Taking these as leading differences for the new difference table with interval of differencing 1, the table can be completed by using the relations :

$$\delta^3 e_1^0 = \delta^3 e_0^0 + \delta^4 e_0^0, \quad \delta^3 e_2^0 = \delta^3 e_1^0 + \delta^4 e_1^0,$$

$$\delta^2 e_1^0 = \delta^2 e_0^0 + \delta^3 e_0^0, \quad \delta^2 e_2^0 = \delta^2 e_1^0 + \delta^3 e_1^0,$$

etc. This is shown below :

x	e_x^0	δe_x^0	$\delta^2 e_x^0$	$\delta^3 e_x^0$	$\delta^4 e_x^0$
0	37.64				
1	40.259568	2.619568			
2	42.084768	1.825200	-0.794368	0.130048	
3	43.245648	1.160880	-0.664320	0.119120	-0.010928
4	43.861328	0.615680	-0.545200	0.108192	-0.010928
5	44.04	0.178672	-0.437008		

From the above table, the required values of e_x^0 are found to be

$$e_1^0 = 40.26, e_2^0 = 42.08, e_3^0 = 43.25 \text{ and } e_4^0 = 43.86.$$

The process illustrated in this example is known as *subtabulation*.

Ex. 2.7 Prove the identity

$$f(a) + f(a+h).x/1! + f(a+2h)x^2/2! + f(a+3h)x^3/3! + \dots$$

$$= \exp[x][f(a) + x\Delta f(a) + x^2 \Delta^2 f(a)/2! + x^3 \Delta^3 f(a)/3! + \dots].$$

This will be proved by using the equivalence relations $E \equiv 1 + \Delta$ and $E^n \equiv (1 + \Delta)^n$. Also the function $f(a)$ will be separated from the operators and will be introduced at the last stage. This is known as the method of *separation of symbols*.

In the present problem

$$\text{l.h.s.} = f(a) + Ef(a)x/1! + E^2 f(a)x^2/2! + E^3 f(a)x^3/3! + \dots$$

$$= \exp[Ex].f(a) = \exp[x + \Delta x].f(a)$$

$$= \exp[x].\exp[\Delta x]f(a)$$

$$= \exp[x].[1 + \Delta x/1! + \Delta^2 x^2/2! + \Delta^3 x^3/3! + \dots]f(a)$$

$$= \exp[x].[f(a) + x\Delta f(a) + x^2 \Delta^2 f(a)/2! + x^3 \Delta^3 f(a)/3! + \dots]$$

$$= \text{r.h.s.}$$

For similar problems see *Exercise 2.7*.

2b.5 Newton's forward interpolation formula

Let y_0, y_1, \dots, y_n be values of the function $y=f(x)$ corresponding to $(n+1)$ equidistant values of the argument x_0, x_1, \dots, x_n , the common difference being h . With these $(n+1)$ values, we can take as our interpolation formula a polynomial of degree n . A polynomial of a higher degree will contain more than $(n+1)$ constants, which cannot be determined when only $(n+1)$ values of y are given.

Let us take the n th degree polynomial in the form

$$\phi(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \\ + a_n(x-x_0)(x-x_1) \dots (x-x_{n-1}) \quad (2.10)$$

The constants a_0, a_1, \dots, a_n are determined by solving the equations

$$y_0 = \phi(x_0), \quad y_1 = \phi(x_1), \quad \dots, \quad y_n = \phi(x_n)$$

Now $\phi(x_0) = a_0$, so that

$$a_0 = y_0$$

$$\phi(x_1) = a_0 + a_1(x_1 - x_0)$$

Equating this to y_1 , we have

$$y_0 + a_1 h = y_1,$$

giving

$$a_1 = \frac{y_1 - y_0}{h} = \frac{\Delta y_0}{h}$$

$$\phi(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$$

Equating this to y_2 , we get

$$y_0 + \frac{\Delta y_0}{h} 2h + a_2 2! h^2 = y_2,$$

which leads to

$$a_2 = \frac{y_2 - y_0 - 2\Delta y_0}{2! h^2} = \frac{y_2 - 2y_1 + y_0}{2! h^2} = \frac{\Delta^2 y_0}{2! h^2},$$

and so on. Lastly, equating $\phi(x_n)$ to y_n , we find

$$a_n = \frac{\Delta^n y_0}{n! h^n}$$

Substituting these values for a_0, a_1, \dots, a_n in (2.10), one gets

$$\phi(x) = y_0 + (x-x_0) \frac{\Delta y_0}{h} + (x-x_0)(x-x_1) \frac{\Delta^2 y_0}{2! h^2} + \\ + (x-x_0)(x-x_1) \dots (x-x_{n-1}) \frac{\Delta^n y_0}{n! h^n} \quad (2.11)$$

This is *Newton's forward interpolation formula*. This may be put in a simpler form by substituting

$$u = \frac{x - x_0}{h}.$$

The formula then reduces to

$$\begin{aligned}\phi(x) &= \phi(x_0 + uh) \\ &= y_0 + u\Delta y_0 + \frac{u(u-1)}{2!} \Delta^2 y_0 + \dots \\ &\quad + \frac{u(u-1)(u-2)\dots(u-n+1)}{n!} \Delta^n y_0.\end{aligned}\dots \quad (2.12)$$

Since the formula contains values of the tabulated function beginning from y_0 forward (to the right) and none backward, it is called forward interpolation formula. It is used mainly for interpolation near the beginning of a set of tabulated values.

It is obvious that if $y=f(x)$ be a polynomial of degree $r (< n)$, either exactly or approximately, then one need take in the forward formula only the first $r+1$ terms of (2.12).

2b.6 Newton's backward interpolation formula

Newton's forward formula is not used for interpolation near the end of a set of tabulated values. For this purpose we have another formula, also due to Newton. As before, let there be $n+1$ pairs of values of the argument x and the entry $y=f(x)$:

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n),$$

the x_i 's being equidistant with common difference h . In this case the approximating polynomial $\phi(x)$ of degree n is taken as

$$\begin{aligned}\phi(x) &= b_0 + b_1(x-x_n) + b_2(x-x_n)(x-x_{n-1}) + \dots \\ &\quad + b_n(x-x_n)(x-x_{n-1})\dots(x-x_1).\end{aligned}\dots \quad (2.13)$$

The constants b_0, b_1, \dots, b_n are determined by solving the equations:

$$y_n = \phi(x_n), y_{n-1} = \phi(x_{n-1}), \dots, y_0 = \phi(x_0).$$

Now

$$\phi(x_n) = b_0, \text{ so that } b_0 = y_n.$$

$$\phi(x_{n-1}) = b_0 + b_1(x_{n-1} - x_n).$$

Equating this to y_{n-1} , we have

$$y_n - b_1 h = y_{n-1}, \text{ giving } b_1 = \frac{y_n - y_{n-1}}{h} = \frac{\Delta y_{n-1}}{h}.$$

Again,

$$\phi(x_{n-2}) = b_0 + b_1(x_{n-2} - x_n) + b_2(x_{n-2} - x_n)(x_{n-2} - x_{n-1})$$

Equating this to y_{n-2} , we get

$$y_n - \frac{4y_{n-1}}{h} 2h + b_2 2^1 h^2 = y_{n-2},$$

which leads to

$$b_2 = \frac{y_{n-2} - y_n + 2y_{n-1}}{2^1 h^2} = \frac{y_{n-2} - 2y_{n-1} + y_n}{2^1 h^2} = \frac{\Delta^2 y_{n-2}}{2^1 h^2},$$

and so on. Lastly, equating $\phi(x_0)$ to y_0 , we find

$$b_n = \frac{\Delta^n y_0}{n! h^n}$$

Substituting these values for b_0, b_1, \dots, b_n in (2 13), one gets

$$\begin{aligned} \phi(x) = & y_n + (x - x_n) \frac{\Delta y_{n-1}}{h} + (x - x_n)(x - x_{n-1}) \frac{\Delta^2 y_{n-2}}{2^1 h^2} + \\ & + (x - x_n)(x - x_{n-1}) \dots (x - x_1) \frac{\Delta^n y_0}{n! h^n} \end{aligned} \quad (2 14)$$

This is *Newton's backward interpolation formula*

If we put $u = \frac{x - x_n}{h}$, the formula reduces to a simpler form

$$\begin{aligned} \phi(x) = & \phi(x_n + uh) \\ = & y_n + u \Delta y_{n-1} + \frac{u(u-1)}{2^1} \Delta^2 y_{n-2} + \\ & + \frac{u(u+1)}{n!} \frac{(u+n-1)}{n!} \Delta^n y_0 \end{aligned} \quad (2 15)$$

This contains values of the function beginning from y_n all backward (to the left) and none forward. Hence the name 'backward interpolation formula'

Ex. 2 8 From the following table determine $e^{0.1285}$ and $e^{0.1694}$

x	e^x
0.12	1.127497
0.13	1.138828
0.14	1.150274
0.15	1.161834
0.16	1.173511
0.17	1.185305
0.18	1.197217
0.19	1.209250

First of all, we construct the diagonal difference table :

x	e^x	Δe^x	$\Delta^2 e^x$
0.12	1.127497		
0.13	1.138828	0.011331	
0.14	1.150274	0.011446	0.000115
0.15	1.161834	0.011560	0.000114
0.16	1.173511	0.011677	0.000117
0.17	1.185305	0.011794	0.000118
0.18	1.197217	0.011912	0.000121
0.19	1.209250	0.012033	

In the present case, the second differences are approximately constant, which indicates that the appropriate interpolation formula may be taken to be a polynomial of degree 2 [*vide* the concluding lines of Section 2b.2].

To determine $e^{0.1245}$ we have to interpolate at the beginning of the table, in which the values of the argument are equidistant. Hence we should use in this case the forward formula.

Here

$$h = 0.01, x_0 = 0.12 \text{ and } x = 0.1245,$$

so that

$$u = \frac{0.1245 - 0.12}{0.01} = 45.$$

Using formula (2.12), we have, therefore, as an approximate value of $e^{0.1245}$,

$$\phi(0.1245) = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0$$

$$= 1.127497 + 0.00509895 - 0.00001423$$

$$= 1.13258172, \text{ i.e. } 1.132582 \text{ (correct to 6 decimal places).}$$

For determining $e^{0.1895}$, we are to interpolate near the end of the table, for which we use the backward formula. In this case,

$$h = 0.01, x_n = 0.19 \text{ and } x = 0.1895,$$

so that

$$u = \frac{0.1895 - 0.19}{-0.01} = -0.05.$$

Using formula (2.15), we have, as an approximate value of $e^{0.1895}$,

$$\begin{aligned}\phi(0.1895) &= y_n + u \Delta y_{n-1} + \frac{u(u+1)}{2!} \Delta^2 y_{n-2} \\ &= 1.209250 - 0.00060165 - 0.00000287 \\ &= 1.20864548 \quad \text{i.e.} \quad 1.208645 \\ &\qquad\qquad\qquad (\text{correct to 6 decimal places})\end{aligned}$$

Ex 2.9 Determine $e^{0.1175}$ and $e^{0.1911}$ from the table of Ex 2.8

To determine $e^{0.1175}$ and $e^{0.1911}$ we are required to find values of the tabulated function e^x corresponding to values of x outside the range of given values from 0.12 to 0.19. This is not a problem in interpolation, but one in *extrapolation*. But here, too, one may use either Newton's forward formula or backward formula, depending on whether the required value is less than x_0 , the smallest tabulated value of the argument, or greater than x_n , the largest value. In the case of extrapolation, we assume that the function $y=f(x)$ is smooth near the ends of the range and that we are not extrapolating beyond a distance h on either side.

To extrapolate for $x=0.1175$, we take $x_0=0.12$, $x=0.1175$ and $h=0.01$, so that $u=\frac{0.1175-0.12}{0.01}=-0.25$

Using formula (2.12), we have as an approximate value of $e^{0.1175}$,

$$\begin{aligned}\phi(0.1175) &= y_0 + u \Delta y_0 + \frac{u(u-1)}{2!} \Delta^2 y_0 \\ &= 1.127497 - 0.25 \times 0.011331 + \frac{(-0.25)(-1.25)}{2} \times 0.000115 \\ &= 1.127497 - 0.00283275 + 0.00001797 \\ &= 1.12468222, \quad \text{i.e.} \quad 1.124682 \\ &\qquad\qquad\qquad (\text{correct to 6 decimal places})\end{aligned}$$

For determining $e^{0.1911}$, we take $x=0.1911$, $x_n=0.19$ and $h=0.01$, so that $u=\frac{0.1911-0.19}{0.01}=11$, and using formula (2.15) an approximate value of $e^{0.1911}$ is obtained as

$$\begin{aligned}\phi(0.1911) &= y_n + u \Delta y_{n-1} + \frac{u(u+1)}{2!} \Delta^2 y_{n-2} \\ &= 1.209250 + 0.11 \times 0.012033 + \frac{(0.11)(1.11)}{2} \times 0.000121\end{aligned}$$

$$= 1.209250 + 0.00132363 + 0.00000739$$

$$= 1.21058102, \text{ i.e. } 1.210581$$

(correct to 6 decimal places).

2b.7 Lagrange's interpolation formula

The above formulæ can be used only when the values of the argument are equidistant. But sometimes it may be difficult to obtain tabulated values of a function for equidistant values of the argument. To deal with such cases, we require a third formula which can be used even when the values of x are not equidistant.

Suppose we have $(n+1)$ pairs of values of x and $y=f(x)$, viz. $(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Here again we take as our interpolation formula a polynomial of degree n . Let us take the polynomial in the form :

$$\begin{aligned} \phi(x) = & c_0(x-x_1)(x-x_2)\dots(x-x_n) \\ & + c_1(x-x_0)(x-x_2)\dots(x-x_n) \\ & + \dots \dots \dots \dots \dots \\ & + c_r(x-x_0)(x-x_1)\dots(x-x_{r-1})(x-x_{r+1})\dots(x-x_n) \\ & + \dots \dots \dots \dots \dots \dots \\ & + c_n(x-x_0)(x-x_1)\dots(x-x_{n-1}). \quad \dots \quad (2.16) \end{aligned}$$

The constants c_0, c_1, \dots, c_n are determined, as in the previous cases, from the equations

$$y_0 = \phi(x_0), y_1 = \phi(x_1), \dots, y_n = \phi(x_n).$$

Now, $\phi(x_0) = c_0(x_0-x_1)(x_0-x_2)\dots(x_0-x_n)$, which on being equated to y_0 gives

$$c_0 = \frac{y_0}{(x_0-x_1)(x_0-x_2)\dots(x_0-x_n)}.$$

Again, $\phi(x_1) = c_1(x_1-x_0)(x_1-x_2)\dots(x_1-x_n)$. When it is equated to y_1 , we get

$$c_1 = \frac{y_1}{(x_1-x_0)(x_1-x_2)\dots(x_1-x_n)}.$$

Similarly for the other constants.

Substituting these values for c_0, c_1, \dots, c_n in (2.16), we get

$$\begin{aligned}\phi(x) = & \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} \dots \frac{(x-x_n)}{(x_0-x_n)} y_0 \\ & + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \dots \frac{(x-x_n)}{(x_1-x_n)} y_1 \\ & + \dots \dots \dots \dots \\ & + \frac{(x-x_0)(x-x_1)}{(x_r-x_0)(x_r-x_1)} \frac{(x-x_{r-1})(x-x_{r+1}) \dots (x-x_n)}{(x_r-x_{r-1})(x_r-x_{r+1}) \dots (x_r-x_n)}, \\ & + \dots \dots \dots \dots \dots \dots \dots \\ & + \frac{(x-x_0)(x-x_1)}{(x_n-x_0)(x_n-x_1)} \frac{(x-x_{n-1})}{(x_n-x_{n-1})} y_n \quad \dots \quad (2.17)\end{aligned}$$

This is *Lagrange's interpolation formula*, which is sometimes given in a different form that helps to minimize computational labour.

$$\begin{aligned}\frac{\phi(x)}{(x-x_0)(x-x_1)} &= \frac{y_0}{(x-x_0)(x_0-x_1)(x_0-x_2) \dots (x_0-x_n)} \\ & + \frac{y_1}{(x-x_1)(x_1-x_0)(x_1-x_2)} \frac{x_1}{(x_1-x_n)} \\ & + \dots \dots \dots \dots \\ & + \frac{y_n}{(x-x_n)(x_n-x_0)(x_n-x_1)} \frac{x_n}{(x_n-x_{n-1})}. \quad \dots \quad (2.18)\end{aligned}$$

If y is a function of x , in many cases it is also possible to look upon x as a function of y , $x=g(y)$, say. In such cases, by writing Lagrange's formula in the form

$$\begin{aligned}\phi(y) = & \frac{(y-y_1)(y-y_2)}{(y_0-y_1)(y_0-y_2)} \frac{(y-y_n)}{(y_0-y_n)} x_0 \\ & + \frac{(y-y_0)(y-y_2)}{(y_1-y_0)(y_1-y_2)} \frac{(y-y_n)}{(y_1-y_n)} x_1 \\ & + \dots \dots \dots \dots \\ & + \frac{(y-y_n)(y-y_1)}{(y_n-y_0)(y_n-y_1)} \frac{(y-y_{n-1})}{(y_n-y_{n-1})} x_n, \quad \dots \quad (2.19)\end{aligned}$$

one may use it to determine the value of x for a given value of y , e.g. to determine the value of x for which $\log x = 3.743$ (*vide* Ex. 2.12). This process is referred to as *inverse interpolation*.

It is seen that Lagrange's formula is of a more general nature than Newton's formulæ considered earlier. It is applicable in any part of the table, and for this the values of the argument need not be equidistant. The use of this formula, again, does not require the construction of a difference table. But, on the whole, it is of a more cumbrous form and its use will be comparatively laborious. Naturally, where the application of Newton's forward or backward formula is justified, this should be done in view of the resulting saving in labour, which in some cases may be considerable.

Ex. 2.10 Let us express Lagrange's formula in the following form :

$$\phi(x) = \sum_{r=0}^n \frac{F(x)}{(x-x_r)F'(x_r)} y_r,$$

where $F(x) = \prod_{r=0}^n (x-x_r)$ and $F'(x_r)$ is the value of $F'(x)$ at $x=x_r$.

$F(x)$ being the product of $(n+1)$ factors, the derivative of $F(x)$ will be the sum of $(n+1)$ terms :

$$\begin{aligned} F'(x) &= (x-x_1)(x-x_2)\dots(x-x_n) + (x-x_0)(x-x_2)\dots(x-x_n) \\ &\quad + \dots + (x-x_0)\dots(x-x_{r-1})(x-x_{r+1})\dots(x-x_n) \\ &\quad + (x-x_0)(x-x_1)\dots(x-x_{n-1}). \end{aligned}$$

Hence

$$F'(x_0) = (x_0-x_1)(x_0-x_2)\dots(x_0-x_n),$$

$$F'(x_1) = (x_1-x_0)(x_1-x_2)\dots(x_1-x_n),$$

⋮

⋮

$$F'(x_r) = (x_r-x_0)(x_r-x_1)\dots(x_r-x_{r-1})(x_r-x_{r+1})\dots(x_r-x_n),$$

⋮

⋮

$$F'(x_n) = (x_n-x_0)(x_n-x_1)\dots(x_n-x_{n-1}).$$

Thus it is now easy to verify that

$$\frac{F(x)}{(x-x_0)F'(x_0)}, \quad \frac{F(x)}{(x-x_1)F'(x_1)}, \quad \dots, \quad \frac{F(x)}{(x-x_n)F'(x_n)}$$

are the same as the coefficients of y_0, y_1, \dots, y_n , occurring on the right-hand side of (2.17). This establishes that Lagrange's formula has an alternative expression, given by

$$\phi(x) = \sum_{r=0}^n \frac{F(x)}{(x-x_r)F'(x_r)} y_r.$$

Ex. 2.11 Let us next show that the sum of the coefficients of y_0, y_1, \dots, y_n in Lagrange's formula is 1. Alternatively, we are to show that

$$\sum_{r=0}^n \frac{F(x)}{(x-x_r)F(x_r)} = 1$$

We deduce this result by decomposing $\frac{1}{F(x)}$ into partial fractions.

We assume

$$\frac{1}{F(x)} = \frac{A_0}{x-x_0} + \frac{A_1}{x-x_1} + \dots + \frac{A_n}{x-x_n}, \quad (2.20)$$

where A_0, A_1, \dots, A_n are independent of x , because to each linear and non repeated factor $(x-x_r)$ of $F(x)$ there corresponds a partial fraction $\frac{A_r}{x-x_r}$, where A_r is a constant, and $\frac{1}{F(x)}$ can be expressed as a sum of such fractions.

Since $F(x) = (x-x_0)(x-x_1) \dots (x-x_n)$, on simplification (2.20) takes the form

$$1 \equiv A_0(x-x_1)(x-x_n) + A_1(x-x_0)(x-x_2)(x-x_n) + \dots + A_n(x-x_0)(x-x_{n-1}).$$

In the above identity we put $x=x_0, x_1, \dots, x_n$ in succession and derive

$$\left. \begin{aligned} A_0 &= \frac{1}{(x_0-x_1)(x_0-x_n)} = \frac{1}{F(x_0)}, \\ A_1 &= \frac{1}{(x_1-x_0)(x_1-x_n)} = \frac{1}{F(x_1)}, \\ A_n &= \frac{1}{(x_n-x_0)(x_n-x_1)} = \frac{1}{F(x_n)} \end{aligned} \right\}$$

Thus determining the constants A_0, A_1, \dots, A_n , we now have, from (2.20),

$$\frac{1}{F(x)} = \frac{1}{(x-x_0)F(x_0)} + \frac{1}{(x-x_1)F(x_1)} + \dots + \frac{1}{(x-x_n)F(x_n)}.$$

Multiplying both sides by $F(x)$, we get the desired result

$$\sum_{r=0}^n \frac{F(x)}{(x-x_r)F(x_r)} = 1$$

Ex. 2.12

<u>x</u>	<u>log x</u>
5531	3.7428037
5532	3.7428822
5533	3.7429607
5534	3.7430392
5535	3.7431176

From the above table, determine antilog 1.743.

Consider first antilog 3.743. This can be evaluated by inverse interpolation, using Lagrange's formula in the form (2.19). Here y is to be taken as $\log x$. The given values of x may be denoted by x_0, x_1, \dots, x_4 and the corresponding values of $\log x$ by y_0, y_1, \dots, y_4 . To simplify calculations, let us form the following table :

	$y=3.743$	y_0	y_1	y_2	y_3	y_4
$y=3.743$	0					
y_0	0.0001963	0				
y_1	0.0001178	-0.0000785	0			
y_2	0.0000393	-0.0001570	-0.0000785	0		
y_3	-0.0000392	-0.0002355	-0.0001570	-0.0000785	0	
y_4	-0.0001176	-0.0003139	-0.0002354	-0.0001569	-0.0000784	0

In this table, each element is the value of y in the corresponding column *minus* the value of y in the corresponding row. The elements above the principal diagonal are not shown because each of them is equal in magnitude (but opposite in sign) to an element below this diagonal ; e.g., $(y_3 - y_2) = -(y_2 - y_3) = +0.0000785$.

Substituting these differences in (2.19), we have

$$\frac{(y-y_1)(y-y_2)(y-y_3)(y-y_4)}{(y_0-y_1)(y_0-y_2)(y_0-y_3)(y_0-y_4)} = +\frac{1178 \times 393 \times 392 \times 1176}{785 \times 1570 \times 2355 \times 3139} \\ = 0.0234250;$$

$$\frac{(y-y_0)(y-y_2)(y-y_3)(y-y_4)}{(y_1-y_0)(y_1-y_2)(y_1-y_3)(y_1-y_4)} = -\frac{1963 \times 393 \times 392 \times 1176}{785 \times 785 \times 1570 \times 2354} \\ = -0.156157;$$

$$\frac{(y-y_0)(y-y_1)(y-y_3)(y-y_4)}{(y_2-y_0)(y_2-y_1)(y_2-y_3)(y_2-y_4)} = +\frac{1963 \times 1178 \times 392 \times 1176}{1570 \times 785 \times 785 \times 1569} \\ = 0.702259;$$

$$\frac{(y-y_0)(y-y_1)(y-y_2)(y-y_4)}{(y_3-y_0)(y_3-y_1)(y_3-y_2)(y_3-y_4)} = + \frac{1963 \times 1178 \times 393 \times 1176}{2355 \times 1570 \times 785 \times 784} \\ = 0.469666,$$

$$\frac{(y-y_0)(y-y_1)(y-y_2)(y-y_3)}{(y_4-y_0)(y_4-y_1)(y_4-y_2)(y_4-y_3)} = - \frac{1963 \times 1178 \times 393 \times 392}{3139 \times 2354 \times 1569 \times 784} \\ = -0.0391929$$

Hence an approximate value of antilog 3.743 is

$$x = \phi(3.743) \\ = 0.234250 \times 5531 - 1.56157 \times 5532 + 7.02259 \times 5533 \\ + 4.69666 \times 5534 - 0.391929 \times 5535 \\ = 5533.501$$

Adjusting the decimal point, we get antilog 3.743 = 0.5533501

2b 8 Divided differences

When the values of the argument are not equidistant, we cannot obtain the finite differences and so cannot use interpolation formulæ based on those differences. We have discussed Lagrange's formula in Section 2b 7, which is useful in such cases.

When the values of the argument are not equidistant, we may, alternatively, form what are called *divided differences* and can use Newton's divided difference formula for interpolation.

Let the values of $f(x)$ be known for $x=x_0, x_1, \dots, x_n$

Then, by definition,

$$f(x_1, v_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \quad (2.21)$$

is the *divided difference of the first order* obtained for the values of the argument x_0, x_1 . Similarly,

$$f(x_2, v_1, x_0) = \frac{f(x_2, x_1) - f(x_1, x_0)}{x_2 - x_0} \quad (2.22)$$

is the *divided difference of the second order* obtained for the values of the argument x_0, x_1, x_2 . The higher order divided differences are defined in a similar way.

The divided differences of various orders can be obtained systematically by constructing a divided difference table.

The divided difference of a particular order can be expressed as a symmetric function of the values of the argument involved in it.

For example,

$$\begin{aligned} f(x_1, x_0) &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_1)}{x_1 - x_0} + \frac{f(x_0)}{x_0 - x_1} \\ &= \frac{f(x_0) - f(x_1)}{x_0 - x_1} = f(x_0, x_1). \end{aligned}$$

And, in general,

$$\begin{aligned} f(x_0, x_1, \dots, x_n) &= \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2) \dots (x_0 - x_n)} \\ &\quad + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2) \dots (x_1 - x_n)} + \dots \\ &\quad + \frac{f(x_n)}{(x_n - x_0)(x_n - x_1) \dots (x_n - x_{n-1})}. \quad \dots \quad (2.23) \end{aligned}$$

A result similar to one in the case of finite differences is the following :

The divided difference of order n of a polynomial of degree n is a constant. This follows from the simple results on divided differences given below, which are obtained from definition :

- (a) Divided difference of $[g(x) \pm h(x)]$ = divided difference of $g(x)$ \pm divided difference of $h(x)$.
- (b) Divided difference of $c.f(x) = c$ [divided difference of $f(x)$].
- (c) Divided difference of order n of x^n is a constant.

We may derive Lagrange's formula from the definition of divided difference as follows :

Let $f(x)$ be a polynomial of degree n and suppose $f(x)$ is known for $x=x_0, x_1, \dots, x_n$. Then, from the definition of divided difference,

$$\begin{aligned} f(x, x_0, \dots, x_n) &= \frac{f(x)}{(x - x_0) \dots (x - x_n)} + \frac{f(x_0)}{(x_0 - x)(x_0 - x_1) \dots (x_0 - x_n)} \\ &\quad + \dots + \frac{f(x_n)}{(x_n - x)(x_n - x_0) \dots (x_n - x_{n-1})} \end{aligned}$$

But $f(x, x_0, \dots, x_n) = 0$, being a divided difference of order $(n+1)$ of a polynomial of degree n . We have, therefore,

$$\begin{aligned} f(x) &= \frac{(x - x_1) \dots (x - x_n)}{(x_0 - x_1) \dots (x_0 - x_n)} f(x_0) + \frac{(x - x_0)(x - x_2) \dots (x - x_n)}{(x_1 - x_0)(x_1 - x_2) \dots (x_1 - x_n)} f(x_1) \\ &\quad + \dots + \frac{(x - x_0) \dots (x - x_{n-1})}{(x_n - x_0) \dots (x_n - x_{n-1})} f(x_n), \end{aligned}$$

and this is the Lagrange's formula.

When the values of the argument are equidistant, we may compute both finite and divided differences. We give below the relations between these two kinds of differences when the values of the argument are equidistant :

$$\begin{aligned}f(x, x+h) &= \frac{f(x+h) - f(x)}{h} = \frac{\Delta f(x)}{h}, \\f(\tau, x+h, x+2h) &= \frac{f(x+h, x+2h) - f(x, x+h)}{2h} \\&= \frac{\Delta f(x+h) - \Delta f(x)}{2h^2} \\&= \frac{\Delta^2 f(x)}{2! h^2},\end{aligned}$$

and, in general,

$$f(x, x-h, \dots, \tau + nh) = \frac{\Delta^n f(\tau)}{n! h^n}. \quad (2.24)$$

2b 9 Newton's divided difference formula

Let $f(x)$ be a function whose divided difference of order n is a constant. Suppose $f(x_0), f(x_1), \dots, f(x_n)$ are known and x is any other value of the argument for which we want the value of $f(x)$.

Now

$$f(x, x_0, \dots, x_{n-1}) = f(x_0, x_1, \dots, x_n)$$

since, by assumption, the n th order divided difference is a constant. So

$$f(x_0, x_1, \dots, x_n) = \frac{f(x, x_0, \dots, x_{n-1}) - f(x_0, x_1, \dots, x_{n-1})}{x - x_{n-1}},$$

$$\text{or } f(x, x_0, \dots, x_{n-2}) = f(x_0, x_1, \dots, x_{n-1}) + (x - x_{n-1}) f(x_0, x_1, \dots, x_n).$$

Again, writing

$$f(x, x_0, \dots, x_{n-2}) = \frac{f(x, x_0, \dots, x_{n-2}) - f(x_0, x_1, \dots, x_{n-2})}{x - x_{n-2}},$$

we have

$$\begin{aligned}f(x, x_0, \dots, x_{n-3}) &= f(x_0, x_1, \dots, x_{n-2}) + (x - x_{n-2}) f(x_0, x_1, \dots, x_{n-1}) \\&\quad + (x - x_{n-2})(x - x_{n-1}) f(x_0, x_1, \dots, x_n).\end{aligned}$$

Proceeding similarly, we get finally

$$\begin{aligned}f(x) &= f(x_0) + (x - x_0) f(x_0, x_1) + (x - x_0)(x - x_1) f(x_0, x_1, x_2) \\&\quad + \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1}) f(x_0, x_1, \dots, x_n). \quad \dots \quad (2.25)\end{aligned}$$

This is known as *Newton's divided difference formula*.

2b.10 Central difference formulæ

In this section, we shall consider interpolation formulæ that employ differences lying close to a horizontal line drawn through the central part of a diagonal difference table. The advantage of these formulæ over the formulæ considered in previous sections is that the former converge more rapidly than the latter when the value required is near the central part of a table. As a result, these central difference formulæ are useful for interpolating near the middle of a set of values. The two most important central difference formulæ are Stirling's and Bessel's formulæ. We next obtain Newton-Gauss forward and backward formulæ, and they in turn will give us Stirling's and Bessel's formulæ.

2b.10.1 Newton-Gauss forward formula (with $2n+1$ equidistant values of the argument)

Suppose $f(x)$ is known for

$$x = x_0 - nh, \dots, x_0 - 2h, x_0 - h, x_0, x_0 + h, x_0 + 2h, \dots, x_0 + nh.$$

Let us take

$$\begin{aligned} x_0 &= x_0, x_1 = x_0 + h, x_2 = x_0 - h, x_3 = x_0 + 2h, x_4 = x_0 - 2h, \dots, \\ x_{2n-1} &= x_0 + nh \text{ and } x_{2n} = x_0 - nh \end{aligned}$$

in Newton's divided difference formula. We have then

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0) f(x_0, x_0 + h) \\ &\quad + (x - x_0)(x - x_0 - h) f(x_0, x_0 + h, x_0 - h) \\ &\quad + (x - x_0)(x - x_0 - h)(x - x_0 + h) f(x_0, x_0 + h, x_0 - h, x_0 + 2h) + \\ &\quad \dots + (x - x_0)(x - x_0 - h)(x - x_0 + h) \dots (x - x_0 - nh) \times \\ &\quad \quad f(x_0, x_0 + h, x_0 - h, \dots, x_0 + nh, x_0 - nh). \end{aligned}$$

Replacing the divided differences by finite differences, we get

$$\begin{aligned} f(x) &= f(x_0) + \frac{(x - x_0)}{h} \Delta f(x_0) + \frac{(x - x_0)}{h} \cdot \frac{(x - x_0 - h)}{h} \cdot \frac{\Delta^2 f(x_0 - h)}{2!} \\ &\quad + \frac{(x - x_0)}{h} \cdot \frac{(x - x_0 - h)}{h} \cdot \frac{(x - x_0 + h)}{h} \cdot \frac{\Delta^3 f(x_0 - h)}{3!} + \dots \\ &\quad + \frac{(x - x_0)}{h} \cdot \frac{(x - x_0 - h)}{h} \dots \frac{(x - x_0 - nh)}{h} \cdot \frac{\Delta^{2n} f(x_0 - nh)}{(2n)!}. \end{aligned}$$

Let us put $u = \frac{x-x_0}{h}$, then

$$\begin{aligned} f(x) &= f(x_0 + uh) = f(x_0) + u\Delta f(x_0) + \binom{u}{2} \Delta^2 f(x_0 - h) + \binom{u+1}{3} \Delta^3 f(x_0 - 2h) \\ &\quad + \dots + \binom{u+n-1}{2n} \Delta^{2n} f(x_0 - nh). \quad \dots \quad (2.26) \end{aligned}$$

And this is the *Newton-Gauss forward formula*

2b.10.2 Newton-Gauss backward formula (with $2n+1$ equidistant values of the argument)

As in the previous formula, we assume that $f(x)$ is known for

$$x = x_0 - nh, \dots, x_0 - h, x_0, x_0 + h, \dots, x_0 + nh.$$

But now we take

$$x_0 = x_0, x_1 = x_0 - h, x_2 = x_0 + h, x_3 = x_0 - 2h, x_4 = x_0 + 2h, \dots \dots$$

$$x_{2n-1} = x_0 - nh \text{ and } x_{2n} = x_0 + nh$$

in Newton's divided difference formula. We have then

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0) f(x_0, x_0 - h) \\ &\quad + (x - x_0)(x_0 - x_0 + h) f(x_0, x_0 - h, x_0 + h) \\ &\quad + (x - x_0)(x - x_0 + h)(x - x_0 - h) f(x_0, x_0 - h, x_0 + h, x_0 - 2h) + \\ &\quad + (x - x_0)(x - x_0 + h)(x - x_0 - h) \dots (x - x_0 + nh) \times \\ &\quad f(x_0, x_0 - h, x_0 + h, \dots, x_0 - nh, x_0 + nh). \end{aligned}$$

Replacing the divided differences by finite differences, we get

$$\begin{aligned} f(x) &= f(x_0) + \frac{(x - x_0)}{h} \Delta f(x_0 - h) + \frac{(x - x_0)}{h} \frac{(x - x_0 + h)}{h} \frac{\Delta^2 f(x_0 - h)}{2!} \\ &\quad + \frac{(x - x_0)}{h} \frac{(x - x_0 - h)}{h} \frac{(x - x_0 - h)}{h} \frac{\Delta^3 f(x_0 - 2h)}{3!} + \dots \dots \\ &\quad + \frac{(x - x_0)}{h} \frac{(x - x_0 + h)}{h} \frac{(x - x_0 - h)}{h} \dots \frac{(x - x_0 + nh)}{h} \frac{\Delta^{2n} f(x_0 - nh)}{(2n)!}. \end{aligned}$$

Let us put $u = \frac{x-x_n}{h}$, then

$$\begin{aligned} f(x) &= f(x_0 + uh) = f(x_0) + u\Delta f(x_0 - h) + \binom{u}{2} \Delta^2 f(x_0 - h) \\ &\quad + \binom{u+1}{3} \Delta^3 f(x_0 - 2h) + \dots \dots + \binom{u+n}{2n} \Delta^{2n} f(x_0 - nh). \quad \dots \quad (2.27) \end{aligned}$$

This is the *Newton-Gauss backward formula*.

2b.10.3 Stirling's formula (with $2n+1$ equidistant values of the argument)

Let $f(x)$ be known for $x=x_0, x_0 \pm h, x_0 \pm 2h, \dots, x_0 \pm nh$, and suppose $f(x)$ is to be determined for $x_0 - h < x < x_0 + h$. Stirling's formula, which is meant for problems of this type, may be obtained by averaging Newton-Gauss forward and backward formulæ as follows :

$$\begin{aligned} f(x) = f(x_0 + uh) &= f(x_0) + u \cdot \frac{\Delta f(x_0) + \Delta f(x_0 - h)}{2} + \frac{\binom{u}{2} + \binom{u+1}{2}}{2} \Delta^2 f(x_0 - h) \\ &\quad + \binom{u+1}{3} \frac{\Delta^3 f(x_0 - h) + \Delta^3 f(x_0 - 2h)}{2} + \dots \\ &\quad + \frac{\binom{u+n}{2n} + \binom{u+n-1}{2n}}{2} \Delta^{2n} f(x_0 - nh). \end{aligned}$$

It contains alternately mean coefficients and mean differences of the formulæ from which it is obtained by averaging.

On simplification, this reduces to

$$\begin{aligned} f(x) = f(x_0 + uh) &= f(x_0) + u \cdot \frac{\Delta f(x_0) + \Delta f(x_0 - h)}{2} + \frac{u^2}{2!} \Delta^2 f(x_0 - h) \\ &\quad + \frac{u(u^2 - 1)}{3!} \frac{\Delta^3 f(x_0 - h) + \Delta^3 f(x_0 - 2h)}{2} + \dots \\ &\quad + \frac{u^2(u^2 - 1)(u^2 - 2^2) \dots (u^2 - n^2)}{(2n)!} \Delta^{2n} f(x_0 - nh). \quad \dots \quad (2.28) \end{aligned}$$

This is *Stirling's formula*.

2b.10.4 Bessel's formula (with $2n+2$ equidistant values of the argument)

Let $f(x)$ be known for $x=x_0, x_0 \pm h, x_0 \pm 2h, \dots, x_0 \pm nh, x_0 + (n+1)h$ and suppose $f(x)$ is to be determined for $x_0 - h < x < x_0 + h$.

Bessel's formula may be obtained by averaging the Newton-Gauss forward formula starting with $f(x_0)$ and the Newton-Gauss backward formula starting with $f(x_0 + h)$. These formulæ based on $(2n+2)$ equidistant values of the argument are

$$\begin{aligned} f(x) = f(x_0) + u \Delta f(x_0) + \binom{u}{2} \Delta^2 f(x_0 - h) + \dots \\ + \binom{u+n-1}{2n} \Delta^{2n} f(x_0 - nh) + \binom{u+n}{2n+1} \Delta^{2n+1} f(x_0 - nh) \end{aligned}$$

$$\text{and } f(x) = f(x_0 + h) + (u - 1) \Delta f(x_0) + \binom{u}{2} \Delta^2 f(x_0) + \\ + \binom{u+n-1}{2n} \Delta^{2n} f(x_0 - \overline{n-1}h) + \binom{u+n-1}{2n+1} \Delta^{2n+1} f(x_0 - nh)$$

Averaging them, we get

$$f(x) = \frac{f(x_0) + f(x_0 + h)}{2} + (u - \frac{1}{2}) \Delta f(x_0) + \binom{u}{2} \frac{\Delta^2 f(x_0) + \Delta^2 f(x_0 - h)}{2} + \\ + \binom{u+n-1}{2n} \frac{\Delta^{2n} f(x_0 - nh) + \Delta^{2n} f(x_0 - \overline{n-1}h)}{2} \\ + \frac{\binom{u+n}{2n+1} + \binom{u+n-1}{2n+1}}{2} \Delta^{2n+1} f(x_0 - nh)$$

This is *Bessel's formula*

We thus see that Bessel's formula also contains alternately mean coefficients and mean differences of the formulae from which it is obtained by averaging

By substituting $v = u - \frac{1}{2}$, we can put the above formula in a neat form as follows

$$f(x) = \frac{f(x_0) + f(x_0 + h)}{2} + v \Delta f(x_0) + \frac{\binom{v^2 - 1}{4}}{2!} \frac{\Delta^2 f(x_0 - h) + \Delta^2 f(x_0)}{2} \\ + \frac{v \binom{v^2 - 1}{4}}{3!} \Delta^3 f(x_0 - h) + \\ + \frac{v \binom{v^2 - 1}{4} \binom{v^2 - 9}{4}}{(2n+1)!} \frac{\left(v^2 - \frac{(2n-1)^2}{4}\right)}{2} \Delta^{2n+1} f(x_0 - nh) \quad (2.29)$$

An important special case of Bessel's formula is the *formula for interpolating to halves*, which occurs when $v=0$, i.e. when we are interested in $x=x_0 + \frac{h}{2}$. In that case all terms involving odd-order differences vanish and the formula reduces to

$$f(x) = \frac{f(x_0) + f(x_0 + h)}{2} - \frac{1}{8} \frac{\Delta^2 f(x_0 - h) + \Delta^2 f(x_0)}{2} \\ + \frac{3}{128} \frac{\Delta^4 f(x_0 - h) + \Delta^4 f(x_0 - 2h)}{2} + \quad (2.29a)$$

We have already mentioned the situations under which Newton's forward, backward and Lagrange's formulæ are appropriate for interpolation. In the beginning of Section 2b.10, we have also stated that the central difference formulæ are suitable for interpolating near the central part of a tabulated set of values. More specifically, it may be stated that Stirling's formula will give better results when $-0.25 \leq u \leq 0.25$, whereas Bessel's will be appropriate for $0.25 \leq u \leq 0.75$ (equivalently, for $-0.25 \leq v \leq 0.25$). Stated differently, it means that Stirling's formula will give more accurate results when interpolating near the beginning or end of a central interval, whereas near the middle of a central interval Bessel's formula is appropriate. The smaller the values of u and v the more rapidly will the formulæ (2.28) and (2.29) converge.

2b.10.5 Laplace-Everett formula

Another central difference formula that is helpful in interpolating within a central interval while subdividing an interval is known as the Laplace-Everett formula.

This formula may be derived by replacing all odd-order differences in the Newton-Gauss forward formula by lower even-order differences. Thus remembering that

$\Delta f(x_0) = f(x_0+h) - f(x_0)$, $\Delta^3 f(x_0-h) = \Delta^2 f(x_0) - \Delta^2 f(x_0-h)$, etc., we have, starting from Newton-Gauss forward formula,

$$\begin{aligned} f(x) &= f(x_0) + u\Delta f(x_0) + \binom{u}{2}\Delta^2 f(x_0-h) + \binom{u+1}{3}\Delta^3 f(x_0-h) \\ &\quad + \binom{u+1}{4}\Delta^4 f(x_0-2h) + \binom{u+2}{5}\Delta^5 f(x_0-2h) + \dots \\ &= f(x_0) + u[f(x_0+h) - f(x_0)] + \binom{u}{2}\Delta^2 f(x_0-h) \\ &\quad + \binom{u+1}{3}[\Delta^2 f(x_0) - \Delta^2 f(x_0-h)] + \binom{u+1}{4}\Delta^4 f(x_0-2h) \\ &\quad + \binom{u+2}{5}[\Delta^4 f(x_0-h) - \Delta^4 f(x_0-2h)] + \dots \\ &= uf(x_0+h) + \binom{u+1}{3}\Delta^2 f(x_0) + \binom{u+2}{5}\Delta^4 f(x_0-h) + \dots \end{aligned}$$

$$\begin{aligned}
 & + (1-u)f(x_0) + \left[\binom{u}{2} - \binom{u+1}{3} \right] \Delta^2 f(x_0-h) \\
 & + \left[\binom{u+1}{4} - \binom{u+2}{5} \right] \Delta^4 f(x_0-2h) + \\
 & = u f(x_0+h) + \binom{u+1}{3} \Delta^2 f(x_0) + \binom{u+2}{5} \Delta^4 f(x_0-h) + \\
 & + \zeta f(x_0) + \binom{\zeta+1}{3} \Delta^2 f(x_0-h) + \binom{\zeta+2}{5} \Delta^4 f(x_0-2h) + \dots
 \end{aligned} \tag{2.30}$$

where $\zeta = 1-u$

This is the Laplace Everett formula

Ex 2.13 From the table given in Ex 2.8 determine (a) $e^{0.1520}$ and (b) $e^{0.1662}$

Since the required values are near the central part of the table, we shall use appropriate central difference formulae

(a) Here $h=0.01$, $x=0.1520$ and we take $x_0=0.15$ so that

$$u = \frac{1520 - 15}{01} = 20$$

Since it is near the beginning of a central interval (also $-25 < u < 25$) we use Stirling's formula. Therefore as an approximate value of $e^{0.1520}$, we have

$$f(x_0) + u \frac{\Delta f(x_0) + \Delta f(x_0-h)}{2} + \frac{u^2}{2} \Delta^2 f(x_0-h)$$

since second differences are approximately constant

Thus

$$e^{0.1520} \approx 1.161834 + (20) \frac{0.11560 + 0.11677}{2} + \frac{(20)^2}{2} (0.000117)$$

$$= 1.161834 + 0.023237 + 0.00000234$$

$$= 1.16416004 \text{ i.e. } 1.164160 \text{ (correct to 6 decimal places)}$$

(b) In this case also, $h=0.01$ but $x=0.1662$, and we take $x_0=0.16$, so that

$$u = \frac{1662 - 16}{01} = 62$$

or

$$v = 12$$

As it is near the middle of a central interval (also $-0.25 < v < 0.25$), we use Bessel's formula. Therefore, as an approximate value of $e^{0.1662}$, we have

$$\frac{f(x_0) + f(x_0 + h)}{2} + v \Delta f(x_0) + \frac{v^2 - 0.25}{2} \cdot \frac{\Delta^2 f(x_0 - h) + \Delta^2 f(x_0)}{2}$$

since second differences are approximately constant. Thus

$$\begin{aligned} e^{0.1662} &\approx \frac{1.173511 + 1.185305}{2} + (0.12)(0.011794) \\ &\quad + \left(\frac{(0.12)^2 - 0.25}{2} \right) \left(\frac{0.000117 + 0.000118}{2} \right) \\ &= 1.179408 + 0.00141528 - 0.0000138415 \\ &= 1.1808094385, \text{ i.e. } 1.180809 \end{aligned}$$

(correct to 6 decimal places).

2b.11 Remainder terms in interpolation formulæ

The different interpolation formulæ discussed so far are polynomials of various orders, and these polynomials $\phi(x)$ coincide with the given function $f(x)$ at the given values of the argument, i.e. $\phi(x) = f(x)$ for $x = x_0, x_1, x_2, \dots, x_p$. But $\phi(x)$ is not necessarily the same as $f(x)$ for other values of x . In this section we shall study the remainder term, $f(x) - \phi(x)$, for the polynomial interpolation formulæ.

$\phi(x)$ is a polynomial of order p , where $p=n$ for Newton's forward, Newton's backward and Lagrange's formulæ and $p=2n$ for Stirling's and $p=2n+1$ for Bessel's formula. So the $(p+1)$ th derivative of $\phi(x)$ with respect to x is zero. Let us define an arbitrary function F as follows :

$$F(z) = f(z) - \phi(z) - [f(x) - \phi(x)] \frac{(z-x_0)(z-x_1)\dots(z-x_p)}{(x-x_0)(x-x_1)\dots(x-x_p)},$$

where x_0, x_1, \dots, x_p are the given values of the argument corresponding to which $f(x)$ is known. $(z-x_0)(z-x_1)\dots(z-x_p)$ is a polynomial of degree $(p+1)$ in z ; and the $(p+1)$ th derivative of it with respect to z is $(p+1)!$. Let us assume that $f(x)$ is continuous and has continuous derivatives of all orders in the interval from x_0 to x_p . Then the same is true for $F(z)$ also, and further $F(z)=0$ for $z=x, x_0, \dots, x_p$. By repeated application of Rolle's theorem we

find that there exists ζ , $x_0 < \zeta < x_p$, such that the $(p+1)$ th derivative of F at ζ is zero, i.e. $F^{(p+1)}(\zeta) = 0$

From the above defining equation of $F(z)$, we have

$$F^{(p+1)}(z) = f^{(p+1)}(z) - \{f(x) - \phi(x)\} \frac{(p+1)!}{(x-x_0)(x-x_1)\dots(x-x_p)},$$

and at $z = \zeta$,

$$f(x) - \phi(x) = (x-x_0)(x-x_1)\dots(x-x_p) \frac{f^{(p+1)}(\zeta)}{(p+1)!} \quad (231)$$

This is the remainder term in the interpolation formula $\phi(x)$

We state below the forms of the remainder term (231) for the different formulae

Remainder term in Lagrange's and Newton's forward and backward formulae is

$$(x-x_0)(x-x_1)\dots(x-x_n) \frac{f^{(n+1)}(\zeta)}{(n+1)!} \quad (232)$$

Remainder term in Stirling's formula is

$$(x-x_0)(x-x_1)(x-x_{-1})\dots(x-x_n)(x-x_{-n}) \frac{f^{(2n+1)}(\zeta)}{(2n+1)!} \quad (233)$$

Remainder term in Bessel's formula is

$$(x-x_0)(x-x_1)(x-x_{-1})\dots(x-x_n)(x-x_{-n})(x-x_{n+1}) \frac{f^{(2n+2)}(\zeta)}{(2n+2)!} \quad (234)$$

Here $x_r = x_0 + rh$ and $x_{-r} = x_0 - rh$

When the analytical form of $f(x)$ is unknown, $f^{(p)}(\zeta)$ may be replaced by an appropriate difference of $f(x)$, using the relations between differences and derivatives

2b 12 Bivariate interpolation

We have so far considered the problem of interpolation for a function of one argument. Sometimes the function may depend on two arguments, and we may have to interpolate for both the arguments. We can solve such a problem by first interpolating with respect to one argument and then interpolating with respect to the other. This is the same as applying the appropriate formulae twice and, as such, gives rise to no new problem.

A different method of solution can be obtained by using a formula similar to Newton's forward formula but modified to take account of two arguments instead of one.

Suppose the function is denoted by $z=f(x, y)$, where x and y are the two arguments. We shall consider the case of equidistant values of both the arguments with common differences h and k , respectively. We next define the two-way differences

$$\Delta^{1,0} f(x, y) = f(x+h, y) - f(x, y) = (E_x - 1) f(x, y)$$

$$\text{and } \Delta^{0,1} f(x, y) = f(x, y+k) - f(x, y) = (E_y - 1) f(x, y).$$

Similarly, we define

$$E_x^u E_y^v f(x, y) = f(x+hu, y+kv).$$

Then we have the following equivalence relation :

$$\Delta^{m,n} \equiv \Delta_x^m \Delta_y^n,$$

where $\Delta_x \equiv E_x - 1$, $\Delta_y \equiv E_y - 1$.

Now we give the general bivariate interpolation formula for the case of equidistant values of the arguments :

$$\begin{aligned} f(x, y) &= f(x_0 + hu, y_0 + kv) \\ &= E_x^u E_y^v f(x_0, y_0) = (1 + \Delta_x)^u (1 + \Delta_y)^v f(x_0, y_0) \\ &= \left[1 + u\Delta_x + \binom{u}{2} \Delta_x^2 + \binom{u}{3} \Delta_x^3 + \dots \dots \right] \left[1 + v\Delta_y + \binom{v}{2} \Delta_y^2 \right. \\ &\quad \left. + \binom{v}{3} \Delta_y^3 + \dots \dots \right] f(x_0, y_0) \\ &= \left[1 + u\Delta_x + v\Delta_y + \binom{u}{2} \Delta_x^2 + \binom{v}{2} \Delta_y^2 + uv\Delta_x\Delta_y + \dots \dots \right] f(x_0, y_0) \\ &= f(x_0, y_0) + u\Delta^{1,0} f(x_0, y_0) + v\Delta^{0,1} f(x_0, y_0) + \binom{u}{2} \Delta^{2,0} f(x_0, y_0) \\ &\quad + \binom{v}{2} \Delta^{0,2} f(x_0, y_0) + uv\Delta^{1,1} f(x_0, y_0) + \dots \dots \quad \dots \quad (2.35) \end{aligned}$$

This formula corresponds to Newton's forward formula. If we put either $u=0$ or $v=0$, we get Newton's forward formula for the univariate case. The formula for linear interpolation for both the arguments can be obtained from (2.35) by ignoring differences higher than $\Delta^{1,0}$ or $\Delta^{0,1}$, as

$$f(x, y) = f(x_0, y_0) + u\Delta^{1,0} f(x_0, y_0) + v\Delta^{0,1} f(x_0, y_0). \quad \dots \quad (2.36)$$

This linear formula can be easily extended to the case of functions of more than two arguments.

Thus

$$f(x, y, z) = f(x_0, y_0, z_0) + u \Delta_x f(x_0, y_0, z_0) + v \Delta_y f(x_0, y_0, z_0) + w \Delta_z f(x_0, y_0, z_0), \quad (2.37)$$

$$\text{where } u = \frac{x - x_0}{h}, \quad v = \frac{y - y_0}{k}, \quad w = \frac{z - z_0}{l}$$

2c NUMERICAL DIFFERENTIATION

The basic idea involved here is to approximate the given function $y = f(x)$ over a short interval of x by a suitable polynomial interpolation formula $\phi(x)$ and then to differentiate that formula rather than the actual function.

So we can obtain an estimate of the value of the derivative of a function $f(x)$ even though the algebraic form of $f(x)$ is not given. To do this we require a table of values of $f(x)$. This process of calculating the derivative of a function, with the help of the approximating interpolation formula and given a set of values of the function, is known as *numerical differentiation*. The problem is solved by selecting an appropriate interpolation formula and then differentiating it term by term as many times as is desired. If the given set of values of $f(x)$ are at equidistant values of x , then we choose an interpolation formula using differences. Again if the derivative required is at the beginning (end or central part) of the tabulated values, then we select Newton's forward (Newton's backward or a central difference) formula. Otherwise, we use Lagrange's formula or a divided difference formula.

We now obtain a general relation connecting the operators Δ and $D \equiv \frac{d}{dx}$. We assume that the Taylor expansion of $f(x)$ exists. Then $f(x+h)$ can be expressed as

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!} f''(x) +$$

$$\text{or } Ef(x) = \left[1 + hD + \frac{h^2 D^2}{2!} + \dots \right] f(x) = e^{hD} f(x)$$

$$\text{or } E \equiv e^{hD}$$

$$\text{or } \Delta \equiv [e^{hD} - 1]$$

$$(2.38)$$

So, knowing $\Delta f(x)$, $\Delta^2 f(x)$, etc., we can find $f'(x)$ using the relation

$$D \equiv \frac{1}{h} \log(1 + \Delta) \equiv \frac{1}{h} [\Delta - \Delta^2/2 + \Delta^3/3 - \dots], \dots \quad (2.39)$$

or an equivalent relation :

$$\begin{aligned} D &\equiv -\frac{1}{h} \log\left(\frac{1}{1+\Delta}\right) \equiv -\frac{1}{h} \log(1 - \Delta E^{-1}) \\ &\equiv \frac{1}{h} [\Delta E^{-1} + \Delta^2 E^{-2}/2 + \Delta^3 E^{-3}/3 + \dots]. \quad \dots \quad (2.39a) \end{aligned}$$

Ex. 2.14 Let us obtain the value of $d(e_x^0)/dx$ at $x=0.5$ using the table in Ex. 2.6.

Since the value required is at the beginning of a set of equidistant values of the tabulated function, we take Newton's forward formula and differentiate it. We take formula (2.12).

In the present case,

$$x_0 = 0, x = 0.5, h = 5,$$

$$\text{so } u = 0.1,$$

$$\text{and } \frac{de_x^0}{dx} = \frac{de_x^0}{du} \cdot \frac{du}{dx} = \frac{1}{h} \cdot \frac{de_x^0}{du} = \frac{1}{5} \cdot \frac{de_x^0}{du}.$$

Now we differentiate formula (2.12) up to terms involving $\Delta^4(e_0^0)$, since we are given five values of e_x^0 and so assume $\Delta^4 e_x^0$ is a constant.

Thus

$$\begin{aligned} d e_x^0 / dx &= \frac{1}{h} \cdot \frac{de_x^0}{du} = \frac{1}{h} \left[\Delta e_0^0 + \frac{2u-1}{2} \Delta^2 e_0^0 + \frac{3u^2-6u+2}{6} \Delta^3 e_0^0 \right. \\ &\quad \left. + \frac{4u^3-18u^2+22u-6}{24} \Delta^4 e_0^0 \right], \end{aligned}$$

so that

$$\begin{aligned} [d(e_x^0)/dx]_{x=0.5} &= \frac{1}{h} [a'(e_x^0)/du]_{u=0.1} \\ &= \frac{1}{5} \left[6.40 - 4(-9.04) + \frac{1.43}{6}(8.06) - \frac{3.976}{24} (-6.83) \right] \\ &= \frac{1}{5} [6.40 + 3.616 + 1.9209 + 1.1315] \\ &= \frac{13.0684}{5} = 2.6137. \end{aligned}$$

We can repeat this process and obtain the higher-order derivatives.

2d NUMERICAL INTEGRATION

In this section, we shall consider some simple approximate methods of finding the value of a definite integral from a given set of numerical values of the integrand. This is also known as *mechanical quadrature* when the integrand is a function of a single variable.

We replace the integrand by a suitable interpolation formula, usually one involving differences, and then integrate it term by term between the desired limits. We can get different quadrature formulae, as they are called, by replacing the integrand by different interpolation formulae or retaining terms up to different orders of difference. We shall obtain below some quadrature formulae by integrating Newton's forward formula.

In Newton's formula, $u = \frac{x-a_0}{h}$ and so $dx = h du$, and if the limits of integration for x are a_0 and $a_n = a_0 + nh$, the limits in terms of u will be 0 and n . Hence

$$\int_{a_0}^{a_0+nh} f(x) dx = h \int_0^n \left[f(a_0) + u \Delta f(a_0) + \binom{u}{2} \Delta^2 f(a_0) + \binom{u}{3} \Delta^3 f(a_0) + \dots \right] du$$

$$= h \left[n f(a_0) + \frac{n^2}{2} \Delta f(a_0) + \left(\frac{n^3}{3} - \frac{n^2}{2} \right) \frac{\Delta^2 f(a_0)}{2!} + \dots \right] \quad (240)$$

From this general form, we obtain the following particular formulae:

2d 1 Trapezoidal rule

Here we assume that the integrand is such that it can be well represented by a straight line in any interval of width h . That means $f(x)$ can be replaced by a first degree polynomial or, equivalently, $\Delta f(x)$ can be regarded as a constant. Accordingly, putting $n=1$ and neglecting differences of all orders higher than the first in (240), we get

$$\int_0^1 f(x) dx = h \left\{ f(a_0) + \frac{\Delta f(a_0)}{2} \right\} = \frac{h}{2} [f(a_0) + f(a_1)]$$

Similarly, we have for the other intervals

$$\int_{a_1}^{a_2} f(x) dx = \frac{h}{2} [f(a_1) + f(a_2)],$$

etc.

Adding all these, we get finally

$$\begin{aligned} \int_{a_0}^{a_n} f(x) dx &= \int_{a_0}^{a_1} f(x) dx + \int_{a_1}^{a_2} f(x) dx + \dots + \int_{a_{n-1}}^{a_n} f(x) dx \\ &= \frac{h}{2} [f(a_0) + f(a_1)] + \frac{h}{2} [f(a_1) + f(a_2)] + \dots \\ &\quad + \frac{h}{2} [f(a_{n-1}) + f(a_n)] = \frac{h}{2} [f(a_0) + 2f(a_1) \\ &\quad + 2f(a_2) + \dots + 2f(a_{n-1}) + f(a_n)]. \quad (2.41) \end{aligned}$$

This is known as the *trapezoidal rule*. This is useful where h is small, for any small segment of a smooth curve can be approximated by a straight line. Geometrically, this rule means that we are replacing the graph of $y=f(x)$ between a_0 and a_0+nh by n segments of straight lines and then approximating the area under the curve by that of a polygon.

2d.2 Simpson's one-third rule

Here we assume that the integrand is such that it can be replaced by a second-degree polynomial over any interval of width $2h$. Accordingly, we put $n=2$ and ignore differences of all orders above the second in (2.40). We have then

$$\begin{aligned} \int_{a_0}^{a_2} f(x) dx &= h \left[2f(a_0) + 4f(a_1) + \left(\frac{8}{3} - 2 \right) \frac{\Delta^2 f(a_0)}{2!} \right] \\ &= \frac{h}{3} [f(a_0) + 4f(a_1) + f(a_2)]. \end{aligned}$$

Similarly, we have for the next interval

$$\int_{a_2}^{a_4} f(x) dx = \frac{h}{3} [f(a_2) + 4f(a_3) + f(a_4)],$$

and so on for the other intervals.

Finally, we have (assuming n is even)

$$\begin{aligned} \int_{a_0}^{a_n} f(x) dx &= \int_{a_0}^{a_2} f(x) dx + \int_{a_2}^{a_4} f(x) dx + \dots + \int_{a_{n-2}}^{a_n} f(x) dx \\ &= \frac{h}{3}[f(a_0) + 4f(a_1) + f(a_2)] + \frac{h}{3}[f(a_2) + 4f(a_3) + f(a_4)] + \\ &\quad + \frac{h}{3}[f(a_{n-2}) + 4f(a_{n-1}) + f(a_n)] \\ &= \frac{h}{3}[f(a_0) - 4\{f(a_1) + f(a_3) + \dots + f(a_{n-1})\} + 2\{f(a_2) \\ &\quad + f(a_4) + \dots + f(a_{n-2})\}] + f(a_n) \end{aligned} \quad (242)$$

This is known as *Simpson's rule* or *Simpson's one third rule*

This is simple, accurate and the most useful of all the quadrature formulæ. In this case we have assumed that the interval is divided into an even number of sub intervals and, geometrically, we have replaced the graph of the given function by $n/2$ arcs of second-degree polynomials.

2d 3 Weddle's rule

Here we replace the integrand by a sixth degree polynomial over any interval of width $6h$. Accordingly, we put $n=6$ and ignore all differences of orders above the sixth in (240). We have then, after some simplifications,

$$\begin{aligned} \int_{a_0}^{a_6} f(x) dx &= h \left[6f(a_0) - 18\Delta f(a_0) + 27\Delta^2 f(a_0) - 24\Delta^3 f(a_0) \right. \\ &\quad \left. + \frac{123}{10}\Delta^4 f(a_0) + \frac{33}{10}\Delta^5 f(a_0) - \frac{41}{140}\Delta^6 f(a_0) \right] \end{aligned}$$

The coefficient of $\Delta^6 f(a_0)$ differs from $3/10$ by a small fraction, $1/140$. Making this change in the coefficient of $\Delta^6 f(a_0)$, which will be negligible if $\Delta^6 f(a_0)$ is small, we get

$$\begin{aligned} \int_{a_0}^{a_6} f(x) dx &= \frac{3h}{10} [f(a_0) + 5f(a_1) + f(a_2) - 6f(a_3) + f(a_4) \\ &\quad + 5f(a_5) + f(a_6)] \end{aligned}$$

For the next interval, we have

$$\int_{a_6}^{a_{12}} f(x) dx = \frac{3h}{10} [f(a_6) + 5f(a_7) + f(a_8) + 6f(a_9) + f(a_{10}) \\ + 5f(a_{11}) + f(a_{12})],$$

and so on for the other intervals.

Finally, we have (assuming n is a multiple of 6)

$$\int_{a_0}^{a_n} f(x) dx = \int_0^{a_6} f(x) dx + \int_{a_6}^{a_{12}} f(x) dx + \dots + \int_{a_{n-6}}^{a_n} f(x) dx \\ = \frac{3h}{10} [f(a_0) + 5f(a_1) + f(a_2) + 6f(a_3) + f(a_4) + 5f(a_5) \\ + 2f(a_6) + \dots + 2f(a_{n-6}) + 5f(a_{n-5}) + f(a_{n-4}) \\ + 6f(a_{n-3}) + f(a_{n-2}) + 5f(a_{n-1}) + f(a_n)]. \quad \dots \quad (2.43)$$

This is *Weddle's rule*.

It is the most accurate of the above formulæ. In usefulness it is second only to Simpson's rule. Geometrically, Weddle's rule means that we have replaced the graph of $y=f(x)$ in the interval a_0 to $a_n=a_0+nh$ by $n/6$ arcs of sixth-degree polynomials.

Similarly, by replacing $f(x)$ by higher-order polynomials, we can get other quadrature formulæ. Also, we can get central difference quadrature formulæ by replacing $f(x)$ by some central difference interpolation formula and integrating it between the desired limits.

2d.4 Relative accuracy of quadrature formulæ

In this section we shall obtain the error terms of the quadrature formulæ discussed in Sections 2d.1—2d.3 and in *Exercise 2.11*. In order that we may compare the error terms and make a comparative assessment of the formulæ, we must apply each to the same interval and divide it into the same number of subintervals. The minimum number of subintervals needed for this purpose is six. Let us, therefore, evaluate

$$I = \int_{a-3h}^{a+3h} f(x) dx$$

by all the four formulæ, after subdividing the interval into six subintervals of width h each. We also need the true value of I .

We assume that the integrand $f(x)$ is continuous and possesses continuous derivatives of all orders in the interval $(a-3h, a+3h)$. Let us also assume that the Taylor expansion of $f(x)$ exists in this interval. Let the indefinite integral of $f(x)$ be denoted by $F(x) + c$.

Then the true value of the integral is

$$\begin{aligned} I &= \int_{a-3h}^{a+3h} f(x) dx = F(a+3h) - F(a-3h) \\ &= 6hf(a) + 9h^3f''(a) + \frac{81}{20}h^5f'''(a) + \frac{243}{280}h^7f''''(a) + \dots, \end{aligned} \quad (244)$$

which is obtained by using Taylor's expansion of F and using the fact that $F'(a) = f(a)$, $F''(a) = f'(a)$, etc.

Now the approximate value of $\int_{a-3h}^{a+3h} f(x) dx$ by the trapezoidal rule is, say,

$$I_T = \frac{h}{2} [f(a-3h) + f(a+3h) + 2\{f(a-2h) + f(a-h) + f(a) + f(a+h) + f(a+2h)\}]$$

Next, we use Taylor's expansion of the functions and get

$$I_T = 6hf(a) + \frac{19}{2}h^3f''(a) + \frac{115}{24}h^5f'''(a) + \frac{859}{720}h^7f''''(a) + \dots \quad (245)$$

Then the error in the trapezoidal rule is

$$\begin{aligned} E_T &= I - I_T \\ &= -\frac{h^3}{2}f''(a) - \frac{89}{120}h^5f'''(a) - \frac{1639}{5040}h^7f''''(a) \end{aligned} \quad (246)$$

Now we obtain the approximate value of the integral by the one-third rule and proceed as before

$$\begin{aligned} I_{1/3} &= \frac{h}{3} [f(a-3h) + f(a+3h) + 4\{f(a-2h) + f(a) + f(a+2h)\} \\ &\quad + 2\{f(a-h) + f(a+h)\}] \\ &= 6hf(a) + 9h^3f''(a) + \frac{49}{12}h^5f'''(a) + \frac{329}{360}h^7f''''(a) \end{aligned} \quad (247)$$

Thus the error in the one-third rule is

$$\begin{aligned} E_{1/3} &= I - I_{1/3} \\ &= -\frac{h^5}{30} f^{IV}(a) - \frac{29}{630} h^7 f^{VI}(a) \dots \dots \\ &= -\frac{4h^5}{120} \left[f^{IV}(a) + \frac{58}{42} h^2 f^{VI}(a) \dots \dots \right]. \quad \dots \quad (2.48) \end{aligned}$$

Similarly, for the three-eighths rule (*vide Exercise 2.11*),

$$\begin{aligned} I_{3/8} &= \frac{3h}{8} [f(a-3h) + f(a+3h) + 3\{f(a-2h) + f(a+2h) \\ &\quad + f(a-h) + f(a+h)\} + 2f(a)] \\ &= 6hf(a) + 9h^3 f''(a) + \frac{33}{8} h^5 f^{IV}(a) + \frac{77}{80} h^7 f^{VI}(a) + \dots \dots \quad \dots \quad (2.49) \end{aligned}$$

Hence the error in the three-eighths rule is given by

$$\begin{aligned} E_{3/8} &= I - I_{3/8} \\ &= -\frac{3h^5}{40} f^{IV}(a) - \frac{53}{560} h^7 f^{VI}(a) \dots \dots \\ &= -\frac{9h^5}{120} \left[f^{IV}(a) + \frac{53}{42} h^2 f^{VI}(a) \dots \dots \right]. \quad \dots \quad (2.50) \end{aligned}$$

Treating the quantities in square brackets of (2.48) and (2.50) as nearly equal, we have the following approximate result :

$$E_{1/3}/E_{3/8} = 4/9. \quad \dots \quad (2.51)$$

The approximate value of the integral by Weddle's rule is

$$\begin{aligned} I_W &= \frac{3h}{10} [f(a-3h) + 5f(a-2h) + f(a-h) + 6f(a) + f(a+h) \\ &\quad + 5f(a+2h) + f(a+3h)] \\ &= 6hf(a) + 9h^3 f''(a) + \frac{81}{20} h^5 f^{IV}(a) + \frac{7}{8} h^7 f^{VI}(a) + \dots \dots, \quad \dots \quad (2.52) \end{aligned}$$

after replacing the functions by their Taylor expansions.

Thus the error in Weddle's rule is

$$\begin{aligned} E_W &= I - I_W \\ &= -\frac{h^7}{140} f^{VI}(a) \dots \dots \quad \dots \quad (2.53) \end{aligned}$$

2d 5 Euler-Maclaurin formula

Now we shall derive an important quadrature formula due to Euler and Maclaurin. This formula can also be used to find the sum of a series when the integral can be easily calculated.

Let it be required to find the sum $\sum_{r=0}^{n-1} f(a+rh)$. We define a new function $F(x)$ by the relation $\Delta F(x) = f(x)$. Thus we have

$$\sum_{r=0}^{n-1} f(a+rh) = F(a+nh) - F(a)$$

Now,

$$\begin{aligned} F(x) &= \Delta^{-1}f(x) \\ &= [e^{hD} - 1]^{-1}f(x), \text{ using relation (2.38)} \\ &= [hD + h^2 D^2/2 + h^3 D^3/6 + h^4 D^4/24 + \dots]^{-1}f(x) \\ &= (hD)^{-1}[1 + (hD/2 + h^2 D^2/6 + h^3 D^3/24 + \dots)]^{-1}f(x) \\ &= (hD)^{-1}[1 - hD/2 + h^2 D^2/12 - h^4 D^4/720 + \dots]f(x), \\ &\quad \text{using the expansion of } (1+x)^{-1} \\ &= \frac{1}{h} \left[D^{-1}f(x) - \frac{h}{2}f'(x) + \frac{h^2}{12}f''(x) - \frac{h^4}{720}f'''(x) + \dots \right]. \end{aligned}$$

As $Df(x) = f'(x)$, so $D^{-1}f(x) = \int f(r)dr$

Thus

$$\begin{aligned} \sum_{r=0}^{n-1} f(a+rh) &= F(a+nh) - F(a) \\ &= \frac{1}{h} \int_a^{a+nh} f(x)dx - \frac{1}{2}[f(a+nh) - f(a)] + \frac{h}{12}[f'(a+nh) - f'(a)] \\ &\quad - \frac{h^3}{720}[f'''(a+nh) - f'''(a)] \quad , \end{aligned} \tag{2.54}$$

$$\begin{aligned} \text{or } \frac{1}{h} \int_a^{a+nh} f(x)dx &= \left[\frac{1}{2}f(a) + f(a+h) + \dots + f(a+n-1h) + \frac{1}{2}f(a+nh) \right] \\ &\quad - \frac{h}{12}[f'(a+nh) - f'(a)] \\ &\quad + \frac{h^3}{720}[f'''(a+nh) - f'''(a)] \end{aligned} \tag{2.55}$$

This is known as the *Euler-Maclaurin formula*. Formula (2.54) is useful in finding the sum of a series. Using the Bernoullian numbers $B_1 = \frac{1}{2}$, $B_2 = \frac{1}{30}, \dots, (2.55)$ can be expressed as

$$\begin{aligned} \frac{1}{h} \int_a^{a+nh} f(x) dx &= \left[\frac{1}{2} f(a) + f(a+h) + \dots + f(a+\overline{n-1}h) + \frac{1}{2} f(a+nh) \right] \\ &\quad - \frac{B_1 h}{2!} [f'(a+nh) - f'(a)] \\ &\quad + \frac{B_2 h^3}{4!} [f'''(a+nh) - f'''(a)] \dots \quad (2.55a) \end{aligned}$$

Ex. 2.15 Calculate the value of the definite integral

$$\int_1^2 \frac{dx}{x}$$

correct to five places of decimals.

We divide the interval (1, 2) into six equal parts, each of width $h=1/6$. The values of the function $y=1/x$ are next tabulated for each of the seven points :

x	$1/x$
1	1.000000
7/6	0.857143
8/6	0.750000
9/6	0.666667
10/6	0.600000
11/6	0.545455
2	0.500000

(a) By trapezoidal rule, the integral is evaluated as

$$\begin{aligned} I_T &= \frac{1}{12} [1.500000 + 2(3.419265)] \\ &= 0.694877 \end{aligned}$$

= 0.69488, correct to five decimal places.

(b) Simpson's one-third rule gives

$$\begin{aligned} I_{1/3} &= \frac{1}{18} [1.500000 + 4(2.069265) + 2(1.350000)] \\ &= 0.69317, \text{ correct to five decimal places.} \end{aligned}$$

(c) Simpson's three-eighths rule gives

$$\begin{aligned} I_{3/8} &= \frac{1}{16} [1(500000) + 3(2752598) + 2(666667)] \\ &= 0.69320, \text{ correct to five decimal places} \end{aligned}$$

(d) By Weddle's rule, we have for the integral the value

$$\begin{aligned} I_W &= \frac{1}{10} [2(85) + 5(1402598) + 6(666667)] \\ &= 0.69315, \text{ correct to five decimal places} \end{aligned}$$

The true value of the integral is

$$I = \int_1^2 \frac{dx}{x} = \log_e 2 = 0.69315$$

Hence the absolute errors are

$$E_T = |I - I_T| = 0.00173,$$

$$E_{1/3} = |I - I_{1/3}| = 0.00002,$$

$$E_{3/8} = |I - I_{3/8}| = 0.00005$$

and $E_W = |I - I_W| = 0$

Ex 2.16 Find the sum $1^3 + 2^3 + \dots + n^3$

We use the Euler-Maclaurin formula in the form (2.54). In the present case we take $f(x) = x^3$, $a = 1$, $h = 1$, $n = n$ in (2.54). Thus

$$\begin{aligned} 1^3 + 2^3 + \dots + n^3 &= \int_1^{n+1} x^3 dx - \frac{1}{2} [(n+1)^3 - 1] + \frac{3}{12} [(n+1)^2 - 1] \\ &= [(n+1)^4 - 1]/4 - [(n+1)^3 - 1]/2 \\ &\quad + [(n+1)^2 - 1]/4 \\ &= [n(n+1)/2]^2 \end{aligned}$$

2e NUMERICAL SOLUTION OF EQUATIONS

In this section, we shall consider some methods of finding the roots of an equation, to any desired degree of accuracy, when the coefficients of the equation are pure numbers. Though most of the methods can be applied to simultaneous equations in more than one variable, we shall consider here only the case of a single variable and determine only the real roots.

In all the methods we are going to discuss, we need some approximation to the desired root. So we first consider two methods of finding approximate values of roots.

Let $f(x)=0$ be the equation whose roots have to be found. Then the graph of $y=f(x)$ will cross the x -axis at the points whose abscissæ are the roots. Approximate values of the real roots can, therefore, be obtained from the graph, and we need only that part of it where it crosses the x -axis.

Another method is based on the fact that if $f(x)$ is continuous in an interval containing the root and if in that interval we find that $f(a)$ and $f(b)$ are of opposite signs, then there will be at least one real root between a and b . For convergence of the approximate values of a root to the true value, it is necessary that a and b should be close to each other.

Next we consider the different methods of determining the real roots of an equation.

2e.1 Method of false position

The oldest method of determining the real roots of a numerical equation is the method of "false position" or *regula falsi*. Suppose the desired root lies between x_1 and x_2 , which are as close as possible, and $f(x_1)$, $f(x_2)$ have opposite signs. Assuming the part of the curve $y=f(x)$ between x_1 and x_2 to be smooth, we can approximate this part of the curve by a straight line. In other words, we perform linear interpolation to find the root of $f(x)=0$ and get

$$\frac{-f(x_1)}{x-x_1} = \frac{f(x_2)-f(x_1)}{x_2-x_1}$$

or
$$x = x_1 - \frac{(x_2-x_1)f(x_1)}{|f(x_2)|-|f(x_1)|}$$

or
$$x = x_1 + \frac{(x_2-x_1)|f(x_1)|}{|f(x_1)|+|f(x_2)|} \dots \quad (2.56)$$

This x is, however, not the true value of the root. This is only a better approximation to the true root than either x_1 or x_2 . We shall repeat this process a number of times till we get the root correct up to the desired number of decimal places.

2e 2 Newton-Raphson method

The real roots of $f(x)=0$ can be computed rapidly by this method if the derivative of $f(x)$ is a simple expression and is easily obtainable

Let x_0 be an approximate value of the root and h the correction to be applied, so that x_0+h is the correct value of the root

Then $f(x_0+h)=0$ and, expanding $f(x_0+h)$ by Taylor's theorem, we get

$$f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0 + \theta h) = 0, \quad |\theta| < 1$$

If x_0 is quite a good approximation to the unknown root h will be small and we can then neglect the term involving h^2 . Thus we have the relation

$$f(x_0) + h_1 f'(x_0) = 0,$$

which gives the approximation to h as

$$h_1 = -\frac{f(x_0)}{f'(x_0)} \quad (2.57)$$

The improved value of the root is then

$$x^{(1)} = x_0 + h_1,$$

and the other approximations are

$$x^{(2)} = x^{(1)} + h_2, \text{ where } h_2 = -\frac{f(x^{(1)})}{f'(x^{(1)})},$$

$$x^{(3)} = x^{(2)} + h_3, \text{ where } h_3 = -\frac{f(x^{(2)})}{f'(x^{(2)})}, \text{ etc}$$

This process is repeated till we get the root correct up to the desired number of decimal places

Equation (2.57) is the fundamental relation in this method. The larger the value of $f'(x)$ near the root, the more rapid will be the convergence of $x^{(1)}, x^{(2)}, \dots$ to the actual root. Should $f'(x)$ be small near the root, convergence would be slow, and the method will fail if $f'(x)=0$ in the neighbourhood of the root

Ex 2.17 Find the real root of

$$2x - \log_{10} x = 7$$

correct to five decimal places

Here $f(x) = 2x - \log_{10} x - 7$, and it is found that $f(3)$ and $f(4)$ are of opposite signs. So there is a real root of $f(x)=0$ between 3 and 4.

(a) *Method of false position*

	x	$f(x)$	
1st approx.	3	-1.47	$x = 3 + \frac{1 \times 1.47}{1.86} = 3 + .7903,$
	4	0.39	
	1	1.86	$x^{(1)} = 3.79.$
2nd approx.	3.7	-0.1682	$x = 3.7 + \frac{.1 \times .1682}{.1885}$
	3.8	0.0203	$= 3.7 + .08923,$
	.1	0.1885	$x^{(2)} = 3.78.$
3rd approx.	3.78	-0.01749	$x = 3.78 + \frac{.01 \times .01749}{.01885}$
	3.79	0.00136	$= 3.78 + .009278,$
	.01	0.01885	$x^{(3)} = 3.7893.$
4th approx.	3.7892	-0.0001475	$x = 3.7892 + \frac{.0001 \times .0001475}{.0001885}$
	3.7893	0.0000410	$= 3.7892 + .00007824,$
	.0001	0.0001885	$x^{(4)} = 3.789278.$

So the root is 3.78928, correct to five decimal places.

(b) *Newton-Raphson method*

$$f(x) = 2x - \log_{10} x - 7$$

and so

$$f'(x) = 2 - \log_{10} e/x.$$

It may be seen from graph that an approximate value of the root is 3.7. Thus

$$h_1 = \frac{7 + .5682017 - 7.4}{2 - 1.173767} = \frac{.1682017}{1.8826233} = .08934,$$

$$x^{(1)} = 3.7 + .089 = 3.789;$$

$$h_2 = \frac{7 + .5785246 - 7.578}{2 - 1.146196} = \frac{.0005246}{1.8853804} = .0002782,$$

$$x^{(2)} = 3.7892;$$

$$h_3 = \frac{7 + .5785475 - 7.5784}{2 - 1.146125} = \frac{.0001475}{1.8853875} = .000078,$$

$$x^{(3)} = 3.789278.$$

So the root, up to five decimal places, is 3.78928.

2e.3 Method of iteration

In those cases where the numerical equation $f(x)=0$ can be written as

$$x = \phi(x), \quad \dots \quad (2.58)$$

the real root can be determined easily by a process known as iteration or successive approximation. Here we start with an approximate value x_0 of the root and, substituting it on the right-hand side of (2.58), get an improved value of the root $x^{(1)}$, given by

$$x^{(1)} = \phi(x_0).$$

Again, we put $x^{(1)}$ for x on the right-hand side of (2.58) and get the second approximation as

$$x^{(2)} = \phi(x^{(1)}).$$

This process is repeated until we have the root correct to the desired number of decimal places.

This method is used only when $|\phi'(x)| < 1$ near the desired root, and the smaller the derivative the more rapid will be the convergence of the approximate roots to the correct value.

2e.4 Convergence of the iteration method

We now consider the condition under which the iteration method will converge, i.e. the condition under which the successive approximations $x_0, x^{(1)}, x^{(2)}, \dots$ will tend to the true value of the root. The true value of the root satisfies the equation

$$x = \phi(x),$$

and the first approximation is obtained as

$$x^{(1)} = \phi(x_0).$$

Subtracting, we get

$$x - x^{(1)} = \phi(x) - \phi(x_0).$$

By the mean value theorem,

$$\phi(x) - \phi(x_0) = (x - x_0)\phi'(\xi_0),$$

where ξ_0 is a point in the interval (x_0, x) or (x, x_0) . Thus we have

$$x - x^{(1)} = (x - x_0)\phi'(\xi_0).$$

Similar equations hold for the other approximations :

$$\begin{aligned}x - x^{(2)} &= (x - x^{(1)})\phi'(\xi_1), \\x - x^{(3)} &= (x - x^{(2)})\phi'(\xi_2), \\&\vdots \\x - x^{(n)} &= (x - x^{(n-1)})\phi'(\xi_{n-1}).\end{aligned}$$

Multiplying together the n equations, member for member, and dividing by the common factor $(x - x^{(1)})(x - x^{(2)}) \dots (x - x^{(n-1)})$, we get

$$x - x^{(n)} = (x - x_0) \prod_0^{n-1} \phi'(\xi_i).$$

Now, if the maximum value m of $|\phi'(\xi_i)|$ is less than 1 in the interval (x_0, x) or (x, x_0) , so that $|\phi'(\xi_i)| \leq m < 1$ for each i , we have

$$|x - x^{(n)}| \leq m^n |x - x_0|. \quad \dots \quad (2.59)$$

Thus the error after n repetitions of the process can be made as small as we please by increasing n suitably since the r.h.s of (2.59) depends on m^n , which approaches 0 as n increases. Thus the condition for convergence of the iteration method is that $|\phi'(x)| < 1$ in the neighbourhood of the desired root, where $\phi(x)$ is the function occurring in (2.58). The smaller the value of $|\phi'(x)|$ the more rapid is the convergence.

2e.5 Convergence of the Newton-Raphson method

The Newton-Raphson method can also be considered as an iteration method. The n th approximation of the method is given by

$$x^{(n)} = x^{(n-1)} - \frac{f(x^{(n-1)})}{f'(x^{(n-1)})},$$

which may be written in the form

$$x = \phi(x),$$

$$\text{with } \phi(x) = x - \frac{f(x)}{f'(x)}.$$

Then by the result concerning the convergence of the iteration method, we know that the Newton-Raphson method will converge if

$$\left| \frac{d}{dx} \left[x - \frac{f(x)}{f'(x)} \right] \right| < 1,$$

i.e. if

$$\left| \frac{f(x)f''(x)}{[f'(x)]^2} \right| < 1, \quad \dots \quad (2.60)$$

in the neighbourhood of the desired root.

Ex. 2.18 Find by iteration the positive root of the equation

$$e^x = 1 + 2x,$$

correct to four places of decimals

Here

$$e^x = 1 + 2x,$$

or

$$x = \log_e(1 + 2x)$$

$$= \log_{10}(1 + 2x) \log_{10} 10$$

Taking $f(x) = x - \log_{10}(1 + 2x) \log_{10} 10$ and forming a set of values of $f(x)$ for different x , it is found that $f(1.25)$ and $f(1.26)$ are of opposite signs. So a positive root lies between 1.25 and 1.26. Thus we begin the process of iteration with $x = 1.25$ as the starting value.

<u>x</u>	<u>$f(x) = \log_{10}(1 + 2x) \log_{10} 10$</u>
1.250000	1.252760
1.252760	1.254336
1.254336	1.255235
1.255235	1.255747
1.255747	1.256038
1.256038	1.256205
1.256205	1.256299
1.256299	1.256353
1.256353	1.256384
1.256384	1.256402
1.256402	1.256412
1.256412	1.256418
1.256418	1.256424
1.256424	1.256427

So the root is 1.2564, correct to four places of decimals

2e.6 Horner's method

Horner discovered a convenient method for obtaining the coefficients of an algebraic equation whose roots differ by a given constant from the roots of a given algebraic equation. By repeated application of this method, the real roots of an algebraic equation can be found up to a desired number of decimal places.

First, we shall discuss the method as applied in diminishing all the roots by a constant s . For the purpose of illustration, we take

$$f(x) = Ax^4 + Bx^3 + Cx^2 + Dx + E = 0 \quad \dots \quad (2.61)$$

as the given equation, but the method is general and can be applied to any polynomial. The equation whose roots are each diminished by s from the roots of (2.61) will be

$$f(x+s) = 0$$

$$\text{or} \quad f(s) + xf'(s) + \frac{x^2}{2}f''(s) + \frac{x^3}{6}f'''(s) + \frac{x^4}{24}f^{iv}(s) = 0. \quad \dots \quad (2.62)$$

Now, from (2.62) we find that

$$f'(s) = 4As^3 + 3Bs^2 + 2Cs + D,$$

$$f''(s) = 12As^2 + 6Bs + 2C,$$

$$f'''(s) = 24As + 6B,$$

$$f^{iv}(s) = 24A.$$

Thus the coefficients of (2.62) are given by

$$\left. \begin{aligned} f^{iv}(s)/24 &= A, & f'''(s)/6 &= 4As + B, \\ f''(s)/2 &= 6As^2 + 3Bs + C, \\ f'(s) &= 4As^3 + 3Bs^2 + 2Cs + D, \\ f(s) &= As^4 + Bs^3 + Cs^2 + Ds + E. \end{aligned} \right\} \dots \quad (2.63)$$

We next describe Horner's scheme of systematically obtaining the coefficients of (2.62). We write down the coefficients A, B, C, \dots of (2.61) in a row and form the following scheme. A letter below a line stands for the sum of two quantities immediately above the line ; e.g., $B + As = M$.

A	B	C	D	E
As	Ms	Ns	Os	Ps
M	N	O		Qs
As	P_s	Q_s		
P	Q		μ	
As	R_s			
R				
As				
		X		
				ϕ

It is easily verified that the quantities v , μ , x and ϕ obtained in the above scheme have the following values given in (2.63) :

$$\left. \begin{aligned} \phi &= 4As + B = f'''(s)/6, \\ x &= 6As^2 + 3Bs + C = f''(s)/2, \\ \mu &= 4As^3 + 3Bs^2 + 2Cs + D = f'(s), \\ v &= As^4 + Bs^3 + Cs^2 + Ds + E = f(s). \end{aligned} \right\} \quad \dots \quad (2.64)$$

Thus the equation (2.62), whose roots are the roots of (2.61) each diminished by s , is

$$Ax^4 + \phi x^3 + \chi x^2 + \mu x + v = 0 \quad \dots \quad (2.65)$$

We next illustrate with an example how this scheme is applied repeatedly in order to obtain successively the different digits of the desired root.

Ex. 2.19 Find the largest positive root of the equation

$$x^3 - 4.759x^2 + 1.759x + 7.518 = 0.$$

The desired root lies between 3 and 4. So we first diminish the roots of the above equation by 3.

$$\begin{array}{r} 1 \\ - 4.759 \\ \hline 3 \\ - 1.759 \\ \hline 3 \\ \hline 1.241 \\ 3 \\ \hline 4.241 \end{array} \qquad \begin{array}{r} 1.759 \\ - 5.277 \\ \hline - 3.518 \\ 3 \\ \hline 0.205 \\ \hline - 3.036 \end{array} \qquad \begin{array}{r} 7.518 \\ - 10.554 \\ \hline - 3.036 \end{array}$$

Thus the equation, whose roots are the roots of the given equation diminished by 3, is

$$x^3 + 4.241x^2 + 0.205x - 3.036 = 0.$$

The first decimal place of the desired root is now the same as the largest root of the above equation lying between 0 and 1, which is the same as the largest root lying between 0 and 10 of the following equation, obtained by multiplying all the roots by 10 :

$$x^3 + 42.41x^2 + 20.5x - 3036 = 0.$$

It is found to be between 7 and 8 ; so we diminish the roots by 7.

1	42·41	20·5	—3036
	7	345·87	2564·59
	49·41	366·37	— 471·41
	7	394·87	
	56·41	761·24	
	7		
	63·41		

The transformed equation is

$$x^3 + 63·41x^2 + 761·24x - 471·41 = 0,$$

and the desired root of the original equation up to the first two digits is 3·7. The third digit of the root is the largest root of

$$x^3 + 63·41x^2 + 761·24x - 471·41 = 0$$

lying between 0 and 1, or the largest root of

$$x^3 + 634·1x^2 + 76124x - 471410 = 0$$

lying between 0 and 10. It is found to be between 5 and 6. An approximate idea of the root can be obtained by solving the linear part of the equation, viz.

$$76124x - 471410 = 0.$$

We diminish the root of the equation by 5.

1	634·1	76124	—471410
	5	3195·5	396595·25
	639·1	79319·5	— 74815·75
	5	3220·5	
	644·1	82540·0	
	5		
	649·1		

The transformed equation is

$$x^3 + 649·1x^2 + 82540·0x - 74815·75 = 0,$$

and the desired root up to the first three digits is 3·75.

An approximate value for the fourth digit of the root is obtained by solving

$$82540x - 74815·75 = 0.$$

This suggests that the next digit is 9. In this way, we may repeat the process stated above and obtain as many digits of the desired root as needed.

Exercises

21 Show that the r th order divided difference is a symmetric function of the values of its arguments

22 Obtain Newton's forward and Newton's backward formulae from Newton's divided difference formula

23 Express the following divided differences in terms of finite differences

$$f(a, a-h, a+h) \text{ and } f(a-h, a, a+h, a+2h)$$

24 Define the first ascending (or backward) difference $\nabla f(x)$ of $f(x)$ by the relation

$$\nabla f(x) = f(x) - f(x-h)$$

and the higher-order ascending differences by

$$\nabla' f(x) = \nabla \underbrace{\nabla f(x)}_{r \text{ times}} = \nabla(\nabla^{r-1} f(x))$$

Then show that

$$\nabla^k f(x) = \Delta^k f(x-kh)$$

(The operator ∇ is called the *nabla* operator, from the Arabic word for harp)

25 Define the *central difference operator* δ by

$$\delta \equiv E^{1/2} - E^{-1/2}$$

or, equivalently, by

$$\delta \equiv \Delta E^{-1/2} \equiv \nabla E^{1/2}$$

The *averaging operator* μ is defined by

$$\mu \equiv \frac{1}{2}[E^{1/2} + E^{-1/2}]$$

Show that the central difference interpolation formulae can be elegantly expressed in terms of these operators δ and μ

26 Define the *factorial polynomials* $u^{(k)}$, for integral k , by

$$u^{(k)} = u(u-1)(u-2) \dots (u-k-1)$$

and the reciprocals of polynomials by

$$u^{(-k)} = \frac{1}{(u+1)(u+2) \dots (u+k)}$$

Then show that

$$\Delta u^{(k)} = k u^{(k-1)},$$

$$\Delta u^{(-k)} = -k u^{(-(k+1))}$$

2.7 Use the method of separation of symbols to prove the following identities :

$$(i) \quad xf(a) + x^2f(a+h) + x^3f(a+2h) + \dots$$

$$= \left(\frac{x}{1-x} \right) f(a) + \left(\frac{x}{1-x} \right)^2 \Delta f(a) + \left(\frac{x}{1-x} \right)^3 \Delta^2 f(a) + \dots$$

$$(ii) \quad (r+1)f(a) + \binom{r+1}{2} \Delta f(a) + \binom{r+1}{3} \Delta^2 f(a) + \dots + \Delta^r f(a)$$

$$= f(a) + f(a+h) + f(a+2h) + \dots + f(a+rh).$$

$$(iii) \quad f(a+nh) + x\Delta f(a+n-1h) + \binom{x+1}{2} \Delta^2 f(a+n-2h)$$

$$+ \binom{x+2}{3} \Delta^3 f(a+n-3h) + \dots$$

$$= f(a+\overline{x+nh}).$$

2.8 Show that the linear interpolation formula can be expressed in the form

$$\phi(x) = \frac{(x_2 - x) f(x_1) + (x - x_1) f(x_2)}{x_2 - x_1},$$

and the corresponding remainder term $R(x) = f(x) - \phi(x)$ has the following bound :

$$|R(x)| \leq \frac{(x_2 - x_1)^2}{8} M,$$

where $|f''(x)| \leq M$ and $x_1 < x < x_2$.

2.9 Suppose $f(x)$ is known for four equidistant values of x , viz. x_0 , $x_0 + h$, $x_0 + 2h$ and $x_0 + 3h$. Then show that the maximum error for interpolating between $x_0 + h$ and $x_0 + 2h$ is given by

$$\frac{3h^4}{128} M,$$

where $\max_{x_0 < x < x_0 + 3h} |f^{IV}(x)| = M$.

2.10 Obtain the following approximate results :

$$(i) \quad \frac{df(x)}{dx} = [\Delta f(x) - \Delta^2 f(x)/2 + \Delta^3 f(x)/3 - \dots]/h;$$

$$(ii) \quad \frac{df(x)}{dx} = [f(x+h) - f(x-h)]/2h.$$

[*Hints :* For (i) differentiate Newton's forward formula and for (ii) differentiate Stirling's formula.]

211 Obtain the following quadrature formula

$$\int_{a_0}^{a_0+3h} y dx = \frac{3h}{8} [f(a_0) + 3f(a_0+h) + 3f(a_0+2h) + f(a_0+3h)]$$

[*Hints* In (240) take $n=3$ and neglect differences of all orders above the third. This is known as *Simpson's three eighths rule* and is less accurate than *Simpson's one third rule*]

212 If u_x is quadratic in x find a quadrature formula for $\int_0^t u_x dx$ in terms of u_{-1} , u_0 and u_1

213 Show that Weddle's rule may be obtained by combining *Simpson's one third* and *three eighths rules* in the ratio 9 : 4

214 (*Gregory's formula*) Show that

$$\begin{aligned} \frac{1}{h} \int_a^{a+nh} f(x) dx &= \frac{1}{2} f(a) + f(a+h) + \dots + f(a+\overline{n-1}h) \\ &\quad - \frac{1}{2} f(a+nh) - \frac{1}{12} [4f(a+\overline{n-1}h) - 4f(a)] \\ &\quad - \frac{1}{24} [4^2 f(a+\overline{n-2}h) + 4^2 f(a)] - \frac{19}{720} [4^3 f(a+\overline{n-3}h) - 4^3 f(a)] \\ &\quad - \frac{3}{160} [4^4 f(a+\overline{n-4}h) + 4^4 f(a)] \end{aligned}$$

[*Hints* In formula (255) replace the derivatives at $f(a)$ by (239) and those at $f(a+nh)$ by (239a)]

This is a form of Euler Maclaurin formula using differences in place of derivatives and this modification is due to Gregory]

215 (*Stirling's approximation to the factorial*) Show that for large n the value of $n!$ is given approximately by

$$n! = \sqrt{2\pi n} n^n \exp\left[-n + \frac{1}{12n} - \frac{1}{360n^3}\right]$$

[*Hints* Evaluate $\log n! = \sum_{i=1}^n \log i$ by using formula (254) and express it as the sum of a function of n and a constant term c . Determine c by using Wallis formula]

$$\log(\pi/2) = \lim_{n \rightarrow \infty} \log[2^{4n} (n!)^4 / ((2n)^2 (2n+1))]$$

2.16 Show that the iteration formula for the square root of N is

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{N}{x_n} \right),$$

where x_n and x_{n+1} are respectively the n th and $(n+1)$ st approximations of \sqrt{N} .

[Hint : Take $f(x) = x^2 - N$ and apply Newton-Raphson method.]

2.17 (Continuation) Obtain the iteration formula for the reciprocal of N .

2.18 Given the following table, determine $\log 261.8$ and $\log 267.5$:

x	$\log x$
261	2.4166405
262	2.4183013
263	2.4199557
264	2.4216039
265	2.4232459
266	2.4248816
267	2.4265113
268	2.4281348

Ans. 2.4179697 ; 2.4273238.

2.19 Various compressional forces were applied to a spring. The following table shows the loads (y , in kg.) that were needed to produce different degrees of contraction (x , in mm.):

x	y
0	0
5	48.7
10	105.2
15	172.4
20	253.4
25	351.2
30	468.7

Using the above table, estimate the load needed to produce a contraction of 34 mm.

Ans. 578.9 kg.

2.20 The fourth powers of a number of integers are shown below :

n	n^4
80	40,960,000
83	47,458,321
86	54,700,816
89	62,742,241
92	71,639,296
95	81,450,625

Calculate the values of $(828)^4$, $(933)^4$, $(867)^4$ and $(884)^4$.

2.21 With the help of the following table, determine the value of θ for which $\sin \theta = 0.75$.

θ	$\sin \theta$
47°	0.73135
48°	0.74314
49°	0.75471
50°	0.76604
51°	0.77715

Ans. 48.59°.

2.22 The growth of population in India, according to the decennial censuses, is shown below

Census year	Population (in lakhs)
1901	2,383
1911	2,520
1921	2,512
1931	2,789
1941	
1951	3,610
1961	4,391

The census figure for 1941, which is not given here, is known to be highly unreliable. Give an estimate of the actual population for 1941.

Ans. 3,207 lakhs.

2.23 Given

$$f(0) = 0,$$

$$f(1) + f(2) = 10,$$

$$f(3) + f(4) + f(5) = 65,$$

find $f(4)$

[Hint. Take

$$f(x) = a + bx + cx^2$$

Estimate a , b , c from the three given equations.]

Ans. 21.

2.24 The following table gives the values of the function $F(0.05; \nu_1, \nu_2)$ for different values of ν_1 and ν_2 :

ν_1	10	20	30	40
ν_2				
5	4.74	4.56	4.50	4.46
10	2.98	2.77	2.70	2.66
15	2.54	2.33	2.25	2.20
20	2.35	2.12	2.04	1.99

Find the values of $F(0.05; 15, 8)$, $F(0.05; 24, 12)$ and $F(0.05; 38, 19)$, taking $1/\nu_1$ and $1/\nu_2$ as the arguments.

2.25 Find the smallest positive real root of $xe^x - 2 = 0$ to four significant figures by the method of false position. *Ans.* 0.8526.

2.26 Find by the Newton-Raphson method, to five significant figures, the root of $\sin x - \frac{x+1}{x-1} = 0$. (An approximate value of the root from graphs of $y = \sin x$ and $y = \frac{x+1}{x-1}$ is known to be -0.4 .)

Ans. -0.42036.

2.27 Find the positive root of

$$x^2 + 5x - 1000 = 0$$

correct to four significant digits.

SUGGESTED READING

- [1] Butler, R. & Kerr, E. *An Introduction to Numerical Analysis* (Chs. 1—5). Isaac Pitman, 1962.
- [2] Freeman, H. *Finite Differences for Actuarial Students* (Chs. 1—5, 9). Cambridge University Press, 1962.
- [3] Henrici, P. *Elements of Numerical Analysis*. John Wiley, 1964.
- [4] Kunz, K. S. *Numerical Analysis* (Chs. 1—7, 11). McGraw-Hill, 1957.
- [5] Scarborough, J. B. *Numerical Mathematical Analysis* (Chs. 1—7, 9.) Oxford University Press, 1958, and Oxford Book Co. (Indian Ed.), 1964.
- [6] Whittaker, E. & Robinson, G. *Calculus of Observations* (Chs. 1—4, 6, 7). Blackie, 1946.

3

3.1 Meaning of probability

The word *probability* may be used in two different contexts. First, it may be used in regard to some proposition. Take, e.g., the statement "It is very probable that India will adhere to the democratic system of government" or "It is very improbable that the country's 'brain drain' will stop in the near future". Probability here means the degree of belief in the proposition of the person making the statement. This may be called subjective probability.

Alternatively, the word may be used in regard to the result of an experiment that can conceivably be repeated an infinite number of times under *essentially similar* conditions. The results will be called *events*. (Note that the word 'event', as well as the word 'experiment', is being used in a perfectly general sense. Thus it is an event that a ball drawn from an urn is red, that in five tosses of a coin one gets two heads or that an article produced by a machine is defective.) The probability of an event here refers to the proportion of cases in which the event occurs in such repetitions of the experiment. This type of probability may be called objective, being a part of the real world, and it is with this sense of the word that we shall be concerned in the present discussion.

3.2 Notation and terminology

The events relating to an experiment will not be all of the same nature. Some events may be more or less complex than the others. E.g., when a six faced die is thrown, the appearance of an even number of points is a more complex event than, say, the appearance of five points. For the former can be looked upon as being composed of a number of events of the latter type—an even number of points means 2 points or 4 points or 6 points. The latter type of event cannot be decomposed further. The results of an experiment that cannot be decomposed further are called *elementary events*, and the whole set of elementary events is called the *sample space* of the experiment. The sample space must be kept clearly in mind for a proper understanding of probability theory.

An event, either elementary or composite, is denoted by one of the capitals, A, B, C , etc., in the beginning of the alphabet. Certain operations with events, defined in (a)-(d) below, will be important for a treatment of probability theory.

(a) Union : $A \cup B$ will denote the occurrence of *either A or B or both A and B*, and will be called the union of A and B ; similarly, the union $\bigcup_{i=1}^m A_i = A_1 \cup A_2 \cup \dots \cup A_m$ will denote the occurrence of at least one of the events A_1, A_2, \dots, A_m .

(b) Difference : $A - B$ will denote the occurrence of A together with the non-occurrence of B and will be called the difference of A from B .

(c) Intersection : $A \cap B$ will denote the occurrence of *both A and B*, and will be called the intersection of A and B ; similarly, the intersection $\bigcap_{i=1}^m A_i = A_1 \cap A_2 \cap \dots \cap A_m$ will denote the simultaneous occurrence of A_1, A_2, \dots, A_m .

(d) Complementation : A^C or \bar{A} or A' will denote the non-occurrence of A and will be called the complement of A .*

Note that the operations of union and intersection have the following properties :

$$\text{Commutativity : } A \cup B = B \cup A, \quad A \cap B = B \cap A. \quad \dots \quad (3.1)$$

$$\begin{aligned} \text{Associativity : } & A \cup (B \cup C) = (A \cup B) \cup C, \\ & A \cap (B \cap C) = (A \cap B) \cap C. \end{aligned} \quad \dots \quad (3.2)$$

$$\begin{aligned} \text{Distributivity : } & A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \\ & A \cup (B \cap C) = (A \cup B) \cap (A \cup C). \end{aligned} \quad \dots \quad (3.3)$$

$$\text{Idempotency : } A \cup A = A, \quad A \cap A = A. \quad \dots \quad (3.4)$$

Also, they obey de Morgan's rules :

$$(A^C)^C = A, \quad (A \cup B)^C = A^C \cap B^C$$

$$\text{and } (A \cap B)^C = A^C \cup B^C.$$

3.3 Classical definition of probability

Our treatment of probability theory will be restrictive in two respects. First, for the sake of simplicity we shall assume that the total number of elementary events in the sample space is *finite* (say r). Thus the treatment will not be applicable to cases where the number of elementary events is *infinite* (either countable or non-countable).

*It follows that $A - B = A \cap B^C$.

Secondly, it will be assumed that the experiment is such that the r elementary events are *equally likely*—in the sense that, when all relevant evidence is taken into account, no one of them can be expected to occur in preference to the others.

The probability $P(A)$ of any event A (not necessarily elementary) is then

$$P(A) = \frac{r(A)}{r}, \quad (35)$$

where $r(A)$ is the number of elementary events *favourable* to A (so that A happens when one of them happens and conversely).

Equation (35) gives what is called the *classical definition* of probability, associated with the names of, among others, Laplace and James Bernoulli.

Some elementary properties of the function P defined on the whole class of events follow immediately from (35).

(a) Since $0 \leq r(A) \leq r$, we have, on dividing by r ,

$$0 \leq P(A) \leq 1 \quad (36)$$

for any event A .

(b) If A is an *impossible* event, $r(A)=0$, implying that

$$P(A)=0$$

(c) If A is a *certain* event, $r(A)=r$, so that here

$$P(A)=1$$

(d) Let the occurrence of A *imply* the occurrence of B . Then every event that is favourable to A is necessarily favourable to B . Hence $r(A) \leq r(B)$, and so

$$P(A) \leq P(B)$$

In Ex 3.1—3.3 we shall consider some direct applications of the definition in computing the probabilities of given events.

Ex 3.1 Suppose a six faced die is thrown. What is the probability that the number appearing uppermost is even?

There are six possible cases as to the number appearing on the uppermost face of the die, viz 1, 2, 3, 4, 5 and 6. These give the six elementary events defining the sample space of the experiment. If the die is perfectly regular in shape and is homogeneous, and if the throw is made without giving any preference to any particular face, then these cases may also be considered equally likely. Of these six

cases, three are favourable to the appearance of an even number, viz. 2, 4 and 6. Hence the required probability is, under the above assumptions,

$$\frac{3}{6} = \frac{1}{2}.$$

Ex. 3·2 The digits 1, 2, 3, 4, 5, 6 and 7 are written down in random order to give a 7-digit number. What is the probability that the number is divisible by 4?

The digits can be arranged in a total of $7!$ ways, which define the sample space in the present case. That the digits are placed in *random order* means that the $7!$ ways are to be considered equally likely. A 7-digit number may be looked upon as the sum of two parts : (a) 100 times the number formed by the first 5 digits and (b) the number formed by the last 2 digits. The first part is always divisible by 4. Hence for the 7-digit number to be divisible by 4, it is only necessary that the last two digits form a multiple of 4. This will be the case if the last two digits are 12, 16, 24, 32, 36, 52, 56, 64, 72 or 76. In each case the first five places of the number can be filled in $5!$ ways, thus giving a total of

$$10 \times 5!$$

favourable cases. The required probability is, therefore,

$$\frac{10 \times 5!}{7!} = \frac{5}{21}.$$

Ex. 3·3 What is the probability of getting 25 points in 5 throws of a die?

There are six possible cases as to the number of points obtained in each throw. An elementary event may, therefore, be represented by a vector

$$(i_1, i_2, i_3, i_4, i_5),$$

where $i_1=1, 2, 3, 4, 5$ or 6 , and represents the number of points obtained in the 1st throw, and similarly for the other components.

The total number of elementary events in the sample space is then

$$6 \times 6 \times 6 \times 6 \times 6 = 6^5.$$

Provided the die is perfect and provided each throw is made without giving any conscious preference to any particular face to turn up, these 6^5 elementary events may be supposed to be equally likely.

As to the number of elementary events favourable to the occurrence of 25 points, we see that it is the same as the coefficient of x^{25} in the expansion of

$$(x^1+x^2+x^3+x^4+x^5+x^6)^5 = x^5 \left(\frac{1-x^6}{1-x} \right)^5$$

But this, again, is the same as the coefficient of x^{20} in the expansion of

$$(1-x^6)^5(1-x)^{-5} \quad (37)$$

Now,

$$(1-x^6)^5(1-x)^{-5} = (1-5x^6+10x^{12}-10x^{18}+\dots)(1+5x+\frac{5 \times 6}{2!}x^2+\frac{5 \times 6 \times 7}{3!}x^3+\dots)$$

Hence the coefficient of x^{20} in (37) is

$$\begin{aligned} & \frac{(24)_{20}}{20!} - 5 \times \frac{(18)_{14}}{14!} + 10 \times \frac{(12)_8}{8!} - 10 \times \frac{(6)_2}{2!} \\ &= \frac{24 \times 23 \times 22 \times 21}{4!} - 5 \times \frac{18 \times 17 \times 16 \times 15}{4!} \\ & \quad + 10 \times \frac{12 \times 11 \times 10 \times 9}{4!} - 10 \times \frac{6 \times 5}{2!} \\ &= 126 \end{aligned}$$

As such, the required probability is

$$= \frac{126}{6^5} = \frac{7}{2 \times 6^3} = \frac{7}{432}$$

3.4 Theorems of total probability

We shall now state and prove some theorems which will enable us to obtain the probability of the union (*i.e.* of the occurrence of *at least one*) of a set of events in terms of the probabilities of the component events and their intersections.

First, consider the case where the events in the set are *mutually exclusive*. Events A_1, A_2, \dots, A_m are called mutually exclusive if no two of them can occur simultaneously. Then $A_1 \cup A_2 \cup \dots \cup A_m$ means the occurrence of *one* of these events. For this case we have the following theorem.

Theorem 3.1 If A_1, A_2, \dots, A_m are mutually exclusive, then

$$P(A_1 \cup A_2 \cup \dots \cup A_m) = P(A_1) + P(A_2) + \dots + P(A_m)$$

. *Proof:* As in Section 3.3, let us suppose that the total number of elementary events is r , of which $r(A_i)$ are favourable to A_i . The number of elementary events that are favourable to either A_1 or A_2 is then $r(A_1) + r(A_2)$. Hence

$$\begin{aligned} P(A_1 \cup A_2) &= \frac{r(A_1) + r(A_2)}{r} \\ &= \frac{r(A_1)}{r} + \frac{r(A_2)}{r} \\ &= P(A_1) + P(A_2). \end{aligned} \quad \dots \quad (3.8)$$

Using result (3.8) repeatedly, we have, for any m ,

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_m) &= P([A_1 \cup A_2 \cup \dots \cup A_{m-1}] \cup A_m) \\ &= P(A_1 \cup A_2 \cup \dots \cup A_{m-1}) + P(A_m) \\ &= P(A_1 \cup A_2 \cup \dots \cup A_{m-2}) + P(A_{m-1}) + P(A_m) \\ &\quad \dots \quad \dots \\ &= P(A_1) + P(A_2) + \dots + P(A_m). \end{aligned}$$

Corollary 3.1.1 Suppose the events A_1, A_2, \dots, A_m are exhaustive, i.e. are such that one of them must occur (meaning that $A_1 \cup A_2 \cup \dots \cup A_m$ is a certain event), and also mutually exclusive. Then

$$P(A_1) + P(A_2) + \dots + P(A_m) = 1.$$

In particular, the events A and A^C , the complement of A , are exhaustive and mutually exclusive. Hence $P(A) + P(A^C) = 1$, implying that

$$P(A) = 1 - P(A^C).$$

This makes possible the computation of the probability of an event from that of its complement and *vice versa*.

Corollary 3.1.2 Let A_1, A_2, \dots, A_m be exhaustive forms of A (i.e., let $A_1 \cup A_2 \cup \dots \cup A_m = A$) and be mutually exclusive. Then

$$P(A) = P(A_1) + P(A_2) + \dots + P(A_m).$$

Corollary 3.1.3 Suppose the occurrence of A implies the occurrence of B . Then B occurs if, and only if, one of the mutually exclusive events A and $B - A$ occurs. Hence, from *Corollary 3.1.2*,

$$P(B) = P(A) + P(B - A),$$

so that

$$P(B - A) = P(B) - P(A).$$

Ex. 3.4 A lot of N objects contains Np objects of one kind (say, Np defectives) and Nq objects of another (say, Nq non-defectives). Out of the lot n objects are chosen at random. What is the probability that there will be k defectives among the chosen objects?

Of the n objects selected, the first one may be any one of the N in the lot, the second any one of the remaining $N-1$, and so on. Hence the total number of elementary events, i.e. the total number of ways in which the n objects may be chosen, regard being had to the order in which they appear, is

$$N(N-1)(N-2)\dots(N-n+1) = (N)_n$$

Since the selection is made at random, these are to be regarded as equally likely.

Now consider the number of ways in which k defectives and $n-k$ non-defectives may be chosen in a particular order, e.g. such that the first k are defective and the last $n-k$ non defective. This number is

$$\begin{aligned} Np(Np-1)(Np-2)\dots(Np-k+1)Nq(Nq-1)(Nq-2)\dots(Nq-n+k+1) \\ = (Np)_k(Nq)_{n-k} \end{aligned}$$

Hence the probability of having defective objects in the first k drawings and non-defective ones in the last $n-k$ is

$$(Np)_k(Nq)_{n-k}/(N)_n$$

But this obviously is also the probability of having k defectives and $n-k$ non-defectives in any other particular order. For our problem, the order is immaterial, and hence the required probability of having k defectives and $n-k$ non-defectives is, by the theorem of total probability for mutually exclusive events (rather, by Corollary 3.1.2),

$$c(Np)_k(Nq)_{n-k}/(N)_n,$$

where c is the total number of orders (permutations) in which k defectives and $n-k$ non defectives may appear. Since

$$c = \frac{n!}{k!(n-k)!},$$

the required probability is

$$\frac{(Np)_k(Nq)_{n-k}}{(N)_n}$$

Ex. 3·5 The conditions being the same as in Ex. 3·4, what is the probability that the sample will contain at least one defective object?

We shall first obtain the probability of the complementary event, viz. that there will be no defective object in the sample. By the same argument as in Ex. 3·4, this probability is

$$\binom{Nq}{n} / \binom{N}{n}.$$

The probability of getting at least one defective object is, by Corollary 3.1.1,

$$1 - \binom{Nq}{n} / \binom{N}{n} = 1 - (Nq)_n / (N)_n.$$

The next theorem deals with the probability of the union of events that are not necessarily mutually exclusive.

Theorem 3.2 Whatever be the events A_1, A_2, \dots, A_m ,

$$P(A_1 \cup A_2 \cup \dots \cup A_m) = \sum_{i=1}^m P(A_i) - \sum_{\substack{i,j=1 \\ i < j}}^m P(A_i \cap A_j) + \dots + (-1)^{m-1} P(A_1 \cap A_2 \cap \dots \cap A_m).$$

Proof (by mathematical induction) : Consider first $P(A_1 \cup A_2)$. The event $A_1 \cup A_2$ occurs if, and only if, one of the mutually exclusive events A_1 and $A_2 - [A_1 \cap A_2]$ occurs. Hence

$$P(A_1 \cup A_2) = P(A_1) + P(A_2 - [A_1 \cap A_2]).$$

But the occurrence of $A_1 \cap A_2$ implies the occurrence of A_2 , and so Corollary 3.1.3 gives

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2). \quad \dots \quad (3.9)$$

The theorem thus holds for $m=2$.

Let the theorem be true for $m=t$ (≥ 2). We shall show that in that case the theorem is necessarily true for $m=t+1$.

Now,

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_{t+1}) &= P([A_1 \cup A_2 \cup \dots \cup A_t] \cup A_{t+1}) \\ &= P(A_1 \cup A_2 \cup \dots \cup A_t) + P(A_{t+1}) \\ &\quad - P([A_1 \cap A_{t+1}] \cup [A_2 \cap A_{t+1}] \cup \dots \cup [A_t \cap A_{t+1}]), \quad \dots \quad (3.10) \end{aligned}$$

from (3.9), taken together with (3.3) Since the theorem holds for $m=t$, we have

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_t) &= \sum_{i=1}^t P(A_i) - \sum_{\substack{i,j \\ i < j}} P(A_i \cap A_j) + \\ &\quad + (-1)^{t-1} P(A_1 \cap A_2 \cap \dots \cap A_t) \end{aligned}$$

and

$$\begin{aligned} P([A_1 \cap A_{t+1}] \cup [A_2 \cap A_{t+1}] \cup \dots \cup [A_t \cap A_{t+1}]) &= \\ &= \sum_{i=1}^t P(A_i \cap A_{t+1}) - \sum_{\substack{i,j \\ i < j}} P(A_i \cap A_j \cap A_{t+1}) + \\ &\quad + (-1)^{t-1} P(A_1 \cap A_2 \cap \dots \cap A_t \cap A_{t+1}) \end{aligned}$$

Substituting these expressions in (3.10), we have, after a rearrangement of terms,

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_{t+1}) &= \sum_{i=1}^{t+1} P(A_i) - \sum_{\substack{i,j \\ i < j}}^{t+1} P(A_i \cap A_j) + \\ &\quad + (-1)^t P(A_1 \cap A_2 \cap \dots \cap A_{t+1}) \end{aligned}$$

Thus the theorem holds for $m=t+1$ if it holds for $m=t$. But it has already been proved to be true for $m=2$. Hence it must also be true for $m=3, 4, \dots$, i.e. for all positive integral values of m .

Ex 3.6 n objects marked 1, 2, ..., n are distributed over n places marked 1, 2, ..., n , one object being allotted to each place. What is the probability that none of the objects occupies the place corresponding to it?

Let us first obtain the probability of the complementary event, viz. that at least one of the objects will occupy the place corresponding to it.

Let A_i denote the event that the object numbered i occupies the place numbered i , for $i=1, 2, \dots, n$. Here A_1, A_2, \dots, A_n —it should be noted—are not mutually exclusive events.

Now, the n objects may be distributed over the n places in a total of $n!$ different ways, which may be assumed to be equally likely.

A_i occurs when the i th place is occupied by the i th object, the remaining $(n-1)$ places being occupied by the remaining objects in any arbitrary order. This can happen in $(n-1)!$ ways. Hence

$$P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n} \text{ for each } i$$

Again, both A_i and A_j occur when the i th and j th places are occupied by the corresponding objects, the other $(n-2)$ places being filled by the remaining objects in any order whatever. So

$$P(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)} \text{ for each } i, j \ (i < j).$$

Similarly for $P(A_i \cap A_j \cap A_k)$, $P(A_i \cap A_j \cap A_k \cap A_l)$, etc.

Lastly,

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = \frac{1}{n!}.$$

We have then, by virtue of the general theorem of total probability for events that are not necessarily mutually exclusive,

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= \binom{n}{1} \cdot \frac{1}{n} - \binom{n}{2} \cdot \frac{1}{n(n-1)} + \binom{n}{3} \cdot \frac{1}{n(n-1)(n-2)} \\ &\quad - \dots + (-1)^{n-1} \cdot \frac{1}{n!} \end{aligned}$$

(since there are n terms in $\sum_i P(A_i)$, $\binom{n}{2}$ terms in $\sum_{i < j} P(A_i \cap A_j)$, and so on),

$$\text{i.e. } = 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n-1} \cdot \frac{1}{n!}.$$

The required probability is

$$\begin{aligned} &1 - P(A_1 \cup A_2 \cup \dots \cup A_n) \\ &= 1 - \frac{1}{1!} + \frac{1}{2!} - \dots + (-1)^n \cdot \frac{1}{n!}. \end{aligned}$$

In case n is large, this may be approximately taken to be e^{-1} or 0.36788.

3.5 Conditional probability and statistical independence

Suppose the event A has non-zero probability, so that $r(A) > 0$. Consider, together with $r(A)$, the number of elementary events that are favourable to both A and B , i.e. the number $r(A \cap B)$. The ratio

$$\frac{r(A \cap B)}{r(A)}$$

is the proportion of elementary events that are favourable to B among elementary events that are favourable to A . Being analogous to the

ratio $\frac{r(B)}{r}$, which is the (unconditional) probability of B , it is called the *conditional* probability of B , given A (or under the condition that A has already occurred). In analogy with the symbol $P(B)$ for unconditional probability, the conditional probability of B given A is denoted by $P(B|A)$.

The following theorem, called the theorem of *compound probability*, is concerned with the joint occurrence (i.e. intersection) of events

Theorem 3.3 Whatever be the events A_1, A_2, \dots, A_m such that

$$P(A_1) > 0, P(A_2|A_1) > 0, P(A_3|A_1 \cap A_2) > 0, \dots,$$

$$P(A_{m-1}|A_1 \cap A_2 \cap \dots \cap A_{m-2}) > 0^*, \text{ we have}$$

$$P(A_1 \cap A_2 \cap \dots \cap A_m) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$$

$$\quad \quad \quad P(A_m|A_1 \cap A_2 \cap \dots \cap A_{m-1})$$

Proof First note that since $P(A_1) > 0$, i.e. since $r(A_1) > 0$,

$$P(A_1 \cap A_2) = \frac{r(A_1 \cap A_2)}{r}$$

can be written in the form

$$\frac{r(A_1)}{r} \cdot \frac{r(A_1 \cap A_2)}{r(A_1)}$$

But

$$\frac{r(A_1)}{r} = P(A_1)$$

and

$$\frac{r(A_1 \cap A_2)}{r(A_1)} = P(A_2|A_1)$$

Hence

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1) \quad (3.11)$$

Also, if $P(A_1) > 0$ and $P(A_2|A_1) > 0$, implying because of (3.11) that $P(A_1 \cap A_2) > 0$, we have

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1 \cap A_2)P(A_3|A_1 \cap A_2) \\ &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \end{aligned}$$

Proceeding in this way, we get the expression for the probability of the joint occurrence of A_1, A_2, \dots, A_m

* These conditions may be replaced by the equivalent single condition

$$P(A_1 \cap A_2 \cap \dots \cap A_{m-1}) > 0$$

Corollary 3.3.1 Suppose the events B_1, B_2, \dots, B_n are exhaustive and mutually exclusive. In that case, the events $A \cap B_1, A \cap B_2, \dots, A \cap B_n$ are mutually exclusive and $A = (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n)$. Hence

$$\begin{aligned} P(A) &= \sum_{i=1}^n P(A \cap B_i) \\ &= \sum_{i=1}^n P(B_i)P(A|B_i), \end{aligned}$$

provided none of B_1, B_2, \dots, B_n has zero probability.

Corollary 3.3.2 (Bayes' theorem) Let, as before, B_1, B_2, \dots, B_n be exhaustive and mutually exclusive events, none of which has zero probability. Further, let A be an event which, too, has non-zero probability. Then the equations

$$P(A \cap B_i) = P(B_i)P(A|B_i)$$

$$\text{and } P(A \cap B_i) = P(A)P(B_i|A)$$

lead to the result

$$\begin{aligned} P(A)P(B_i|A) &= P(B_i)P(A|B_i) \\ \Rightarrow P(B_i|A) &= \frac{P(B_i)P(A|B_i)}{P(A)} \end{aligned}$$

or, because of *Corollary 3.3.1*,

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)}, \text{ for } i=1, 2, \dots, n.$$

In many cases B_1, B_2, \dots, B_n may be looked upon as the possible *causes* of the *effect* A . *Bayes' theorem* then gives the probability of the cause B_i when the effect A has occurred or the probability of the *hypothesis* B_i in the light of the *datum* A . We may also say that it gives the *posterior probability* of B_i in terms of the *prior probabilities* $P(B_i)$, $i=1, 2, \dots, n$, and the conditional probabilities of A .

Suppose that A and B are such that

$$P(B|A) = P(B).$$

Then the probability of the occurrence of the event B is unaffected by the information that the event A has occurred, and we say that

B is (statistically) independent of A The following result then follows directly from (3.11)

$$P(A \cap B) = P(A)P(B) \quad (3.12)$$

Thus in this case the probability of the joint occurrence of the two events is given simply by the product of their individual unconditional probabilities Relation (3.12), being symmetrical in A and B , suggests that if B is independent of A , then A should also be considered independent of B —that, in fact, one should speak of the (mutual) independence of A and B . It is customary to take relation (3.12) as defining the independence of A and B . This definition is accepted irrespective of whether $P(A)$, say, is or is not equal to zero, although in the former case $P(B|A)$ is not defined.

In the same way, m events A_1, A_2, \dots, A_m are said to be *mutually independent* if the following $2^m - m - 1$ equations are satisfied

$$\left. \begin{aligned} P(A_i \cap A_j) &= P(A_i)P(A_j), \text{ for all } i, j (i < j), \\ P(A_i \cap A_j \cap A_k) &= P(A_i)P(A_j)P(A_k), \text{ for all } i, j, k (i < j < k), \\ P(A_1 \cap A_2 \cap \dots \cap A_m) &= P(A_1)P(A_2) \dots P(A_m) \end{aligned} \right\} \quad (3.13)$$

Ex. 3.7 Three urns contain, respectively, a_1 white, b_1 black balls, a_2 white, b_2 black balls, and a_3 white, b_3 black balls. One of the urns is chosen on the results of two throws of a coin—the first urn if head appears on each throw, the second urn if tail appears on each throw and the third urn in case head appears on one throw and tail on the other. Finally, a ball is drawn at random from the chosen urn. What is the probability for the ball being white?

We shall denote by B_1 , B_2 and B_3 the selection of the 1st, of the 2nd and of the 3rd urn, respectively, which are also equivalent to the appearance of two heads, of no head and of just one head in two throws of the coin. Also, we shall denote by A the drawing of a white ball. The events B_1 , B_2 and B_3 are exhaustive and mutually exclusive. Further, none of them is impossible. Hence, by Corollary 3.3.1, we have

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)$$

Now, if the coin is assumed to be perfect, then

$$P(B_1) = \frac{1}{4}, P(B_2) = \frac{1}{4}, P(B_3) = \frac{1}{2}$$

If the 1st urn is chosen, then the ball is taken at random from a set of $a_1 + b_1$ balls of which a_1 are white. Hence

$$P(A|B_1) = \frac{a_1}{a_1 + b_1}.$$

Similarly,

$$P(A|B_2) = \frac{a_2}{a_2 + b_2} \text{ and } P(A|B_3) = \frac{a_3}{a_3 + b_3}.$$

Thus

$$P(A) = \frac{1}{4} \left[\frac{a_1}{a_1 + b_1} + \frac{a_2}{a_2 + b_2} + \frac{2a_3}{a_3 + b_3} \right].$$

It is obvious that if the events A_1, A_2, \dots, A_m are mutually independent, then the theorem of compound probability takes the following form :

Theorem 3.4

$$P(A_1 \cap A_2 \cap \dots \cap A_m) = P(A_1)P(A_2)\dots P(A_m),$$

provided A_1, A_2, \dots, A_m are independent.

Ex. 3.8 A die is thrown 10 times. What is the probability of getting six points in each of 4 throws?

Under the usual assumptions, the probability of getting a six in a single throw is $\frac{1}{6}$. Since the throws may be supposed to be made independently of each other, the events A_1, A_2, \dots, A_{10} , where A_i stands either for the appearance of a six or for the non-appearance of a six in the i th throw, are to be taken as statistically independent. Hence, by the theorem of compound probability, the probability of getting a six in each of 4 particular throws, e.g. in each of the first 4 throws, (and a number other than six in each of the other 6 throws) is

$$\left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^6 = \frac{5^6}{6^{10}}.$$

But this is also the probability of getting 4 sixes in any other order. Since the total number of ways in which 4 sixes may appear is

$$\binom{10}{4},$$

the required probability is, by the theorem of total probability for mutually exclusive events,

$$\binom{10}{4} \frac{5^6}{6^{10}} = 0.05427, \text{ approximately.}$$

3.6 Limitations of the classical definition

The classical definition of probability has some obvious drawbacks

For one thing, direct application of this definition is not possible if the total number of elementary events of an experiment is infinite. It is obvious, for instance, that the definition will have to be modified and extended when we want to obtain the probability that a point selected in a given region will lie in a specified part of it. This type of probability has been called *geometrical probability* (or *probability in continuum*). The method generally used in this case is indicated in the following examples.

Ex 3.9 Suppose a line segment AB is bisected at the point C . What is the probability that a point taken on the line segment lies on AC ?

To evaluate the probability let us divide the whole line segment into n smaller segments, each of length

$$\alpha = \frac{\text{length of } AB}{n}$$

If the point is chosen *at random*, we may assume that the probability that the point lies on one of these small segments is the same for all segments. In that case the required probability will be equal to the ratio of the number of small segments within AC to the number of those within AB , for $\alpha \rightarrow 0$ (or $n \rightarrow \infty$). But, by taking smaller and smaller segments, we shall be making this ratio as near the ratio of the length of AC to that of AB as we please. Hence the required probability will be

$$\frac{\text{length of } AC}{\text{length of } AB} = \frac{1}{2}$$

The general expression for a probability of this kind is

$$\frac{\text{measure of the specified part of the region}}{\text{measure of the whole region}}, \quad (3.14)$$

where by measure we mean the length, the area or the volume of the region, according as it is in one, two or three dimensions.

Ex 3.10 A line segment AB is divided by a fixed point C , such that length $AC=a$ and length $CB=b$. Two points X and Y are chosen at random on AC and CB , respectively. What is the probability that the segments AX , XY and YB can form a triangle?

Let us denote the distance AX by x and the distance YB by y . Clearly, $0 < x < a$ and $0 < y < b$. Also, the distance XY is $a+b-x-y$. Now the segments AX , XY and YB can form a triangle if, and only if, the sum of the lengths of any two of them is greater than the length of the third segment, i.e. if, and only if,

$$(i) \quad x+y > a+b-x-y \text{ or } x+y > \frac{a+b}{2},$$

$$(ii) \quad x+(a+b-x-y) > y \text{ or } y < \frac{a+b}{2}$$

and (iii) $(a+b-x-y)+y > x$ or $x < \frac{a+b}{2}$.

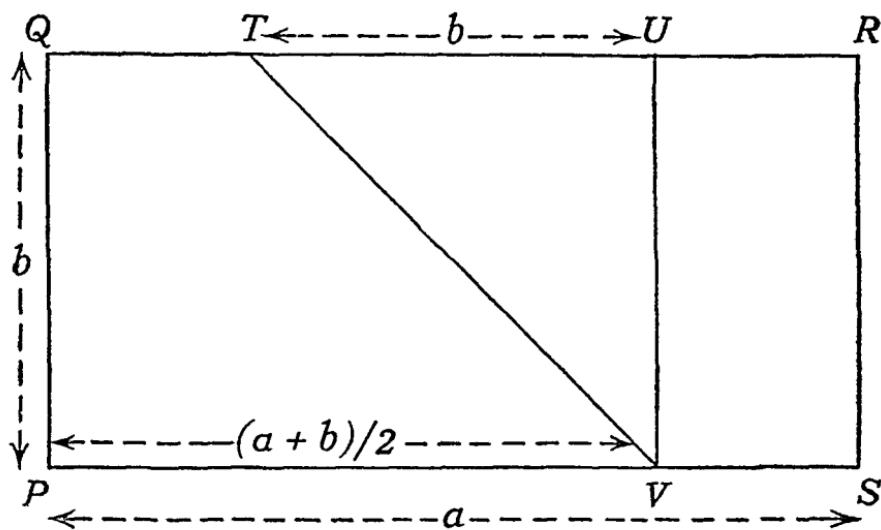


Fig. 3.1

Now, X and Y are chosen independently. Hence (x, y) may be regarded as a point selected at random on a rectangle, say $PQRS$, of sides a and b .

Case 1 : If $a > b$, (x, y) can satisfy conditions (i), (ii) and (iii) if, and only if, it lies within $\triangle TUV$ (Fig. 3.1). Again, $\triangle TUV$ is isosceles, the angle formed by the two equal sides being a right angle. Hence area $\triangle TUV = \frac{1}{2}b^2$, and the required probability is

$$\frac{\text{area } \triangle TUV}{\text{area } PQRS} = \frac{\frac{1}{2}b^2}{ab} = \frac{b}{2a}.$$

Case 2 : If $a < b$, the required probability can be similarly shown to be

$$\frac{\frac{1}{2}a^2}{ab} = \frac{a}{2b}.$$

Case 3 : If $a=b$, $PQRS$ is a square and (x, y) can satisfy conditions (i), (ii) and (iii) if, and only if, it lies within $\triangle QRS$, which again is isosceles and right-angled (Fig. 3.2). Also,

$$\text{area } \triangle QRS = \frac{1}{2}a^2,$$

$$\text{area } PQRS = a^2$$

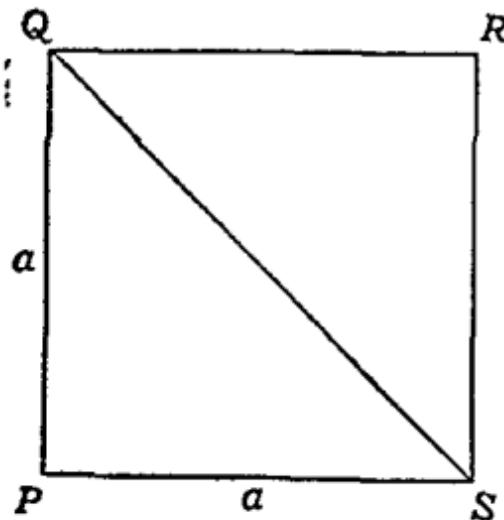


Fig. 3.2

Hence the required probability is now

$$\frac{\frac{1}{2}a^2}{a^2} = \frac{1}{2}.$$

Ex. 3.11 A board is covered with congruent rectangles, and a coin, the diameter of which is less than the smaller side of a rectangle, is thrown on the board. What is the probability that it will be partly in one rectangle and partly in another?

Let a and b be the lengths of the two sides of a rectangle and d the diameter of the coin. First consider the complementary event, i.e. the event that the coin will be completely in one rectangle. To obtain its probability we need consider only the rectangle in

which the centre of the coin lies. The coin will be wholly within this rectangle if the distance of the centre from each side is at least $d/2$ (i.e. at least equal to the radius of the coin). The probability of the complementary event, is therefore,

$$\frac{\text{area of rectangle with sides } (a-d) \text{ and } (b-d)}{\text{area of rectangle with sides } a \text{ and } b} = \frac{(a-d)(b-d)}{ab}.$$

The required probability is, then,

$$1 - \frac{(a-d)(b-d)}{ab} = \frac{d(a+b)-d^2}{ab}.$$

But the classical definition has a more serious drawback. Suppose we have a six-faced die known to be loaded in favour of 6. Here obviously the probability of 6 appearing uppermost in a throw is greater than $1/6$. But how to determine this probability? The classical definition leaves this question unanswered.

This definition of probability may also be criticised on the ground that it moves in a circle, being based on the idea of equally likely (i.e. equally *probable*) outcomes.

3.7 An axiomatic approach

The difficulties to which the classical definition leads may be obviated if we formulate a probability calculus on an axiomatic basis. For this we start from the notion of probability as relative frequency *in the long run*.

Suppose we perform a sequence of n repetitions of an experiment. Let f be the number of times A occurs among these n repetitions. f is called the *absolute frequency* (or, simply, frequency) of A , while the ratio $\frac{f}{n}$ is called the *relative frequency* of A .

If several sequences, each of n repetitions of the experiment, are now considered, several ratios will be obtained :

$$\frac{f_1}{n}, \frac{f_2}{n}, \frac{f_3}{n}, \dots$$

In many cases it is found that these relative frequencies differ from one another by small amounts, provided n is large. Thus in

such cases there is a tendency in the relative frequencies to accumulate in the neighbourhood of some fixed value. This limiting value as $n \rightarrow \infty$ is regarded as the probability of A in the experiment.

It is impossible here to practically determine the *exact* probability of an event. We cannot even prove the existence of a limit to the relative frequencies in any given case. On the other hand, we can regard any observed relative frequency as an *estimate* of the supposed limit. In this sense, one can even speak of the probability that a newborn child is a boy, that a man of sixty will not die within a year or that the height of a man is less than six feet.

Probability being interpreted in this way, the relations

(I) for any event A ,

$$P(A) \geq 0,$$

(II) if A is a sure event,

$$P(A)=1,$$

(III) for mutually exclusive events A_1 and A_2 ,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

(more generally, for mutually exclusive events A_1, A_2, \dots ,

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots,$$

and (IV) $P(A_1 \cap A_2) = P(A_1) P(A_2 | A_1)$, provided $P(A_1) > 0$

—are taken as axioms, formulated, of course, in the light of similar properties of relative frequencies. The other relations established in the previous sections will then follow from these axioms.

Thus, in whatever way we define probability, in each case it will be assumed to obey essentially the same laws.

Any function P , or $P(\)$, that obeys axioms (I)—(III) is called a *probability function* on the events in the sample space, and the function $P(\ | A_1)$ defined by axiom (IV) is called the *conditional probability function* corresponding to P and associated with the event A_1 .

Such a function P , for purposes of probability theory, need not be considered in the context of any experiment. For instance, in the case of a finite sample space, with elementary events e_α ($\alpha=1, 2, \dots, r$), we may associate with e_α a non-negative number p_α . Then, for any event A , we may take $P(A) = \sum_A p_\alpha$, the sum being taken over all e_α that belong to A . Actually, $p_\alpha = P(e_\alpha)$.

This axiomatic treatment of probability theory is due mainly to the Russian mathematician Kolmogorov.

In statistical theory, the axiomatic definition is of greater help than the classical definition, although the latter is simpler to understand and to deal with.

3.8 Random variable, and its expectation and variance

In most cases, with each elementary event in the sample space we may associate a real number. In tossing a coin, e.g., we may associate the number 1 with the appearance of a head and the number 0 with the appearance of a tail. In throwing a die, we have the numbers 1, 2, ..., 6 corresponding to the six possibilities regarding the face that appears uppermost. We thus define a function on the sample space. A (real-valued) function defined on the sample space is called a *random variable* or a stochastic variable. Obviously, to each value of a random variable x^* there corresponds a definite probability. Let x_1, x_2, \dots, x_k be the possible values of x , and let p_1, p_2, \dots, p_k be the corresponding probabilities. A statement of the possible values, together with the probabilities, gives the *probability distribution* of x . The probability p_i is, of course, to be interpreted as approximately the proportion of cases in which x takes the value x_i in a large series of repetitions of the experiment.

One important characteristic of a random variable is its *expectation*. Thus, let $x(e_\alpha)$ be the value of x corresponding to the elementary event e_α ($\alpha=1, 2, \dots, r$), and let $P(e_\alpha)$ be the probability associated with e_α . Then

$$E(x) = \sum_{\alpha=1}^r x(e_\alpha)P(e_\alpha) \quad \dots \quad (3.15)$$

is the expectation of x . $E(x)$ is often denoted by μ_x or, simply, μ .

Suppose now that the elementary events are numbered in such a way that $x(e_\alpha)$ for $\alpha=1, 2, \dots, r_1$ are all equal to x_1 , $x(e_\alpha)$ for $\alpha=r_1+1, r_1+2, \dots, r_2$ are all equal to x_2 , and so on. Then (3.15) may be written alternatively as

$$E(x) = x_1 \sum_{\alpha=1}^{r_1} P(e_\alpha) + x_2 \sum_{\alpha=r_1+1}^{r_2} P(e_\alpha) + \dots + x_k \sum_{\alpha=r_{k-1}+1}^{r_k} P(e_\alpha).$$

* Actually, here we have a special type of random variable, which is the only appropriate type for a finite sample space. For some other types, see Section 9.2.

But the coefficient of x_i here is nothing but the probability

$$P[x=x_i] = p_i$$

Hence we have also

$$E(x) = \sum_{i=1}^t x_i p_i \quad (3.16)$$

In this form the expectation of x is seen to be the sum of the products of the different possible values of x by their probabilities. Since p_i is the 'long run relative frequency' with which x assumes the value x_i , $E(x)$ may also be interpreted as the 'long-run average value' of x (*vide* Section 6.3).

There is yet a third formula for $E(x)$. Let A_1, A_2, \dots, A_t be a set of exhaustive and mutually exclusive events such that x takes the same value x_j for all elementary events that are favourable to A_j (for each $j=1, 2, \dots, t$). Then (3.15) may be expressed in the form

$$E(x) = \sum_{j=1}^t x_j P(A_j) \quad (3.17)$$

Indeed, this is the most general formula for $E(x)$. In actual application, we use one of these formulae—generally the one that best serves our purpose in the given context.

We have the following theorems on expectation.

Theorem 3.5 If $x=a$, a constant, then $E(x)=a$.

Proof Since $x=a$, we have $x(e_a)=a$ for all α . Hence (3.15) gives

$$E(x) = a \sum_{\alpha=1}^r P(e_\alpha) = a,$$

since $\sum_{\alpha=1}^r P(e_\alpha) = 1$

Theorem 3.6 If $y=bx$, then $E(y)=bE(x)$.

Proof Corresponding to the elementary event e_α , $x(e_\alpha)$ is the value of x and $y(e_\alpha)$ is the value of y . Then

$$\begin{aligned} E(y) &= \sum_{\alpha=1}^r y(e_\alpha) P(e_\alpha) \\ &= b \sum_{\alpha=1}^r x(e_\alpha) P(e_\alpha) \\ &= bE(x) \end{aligned}$$

Theorem 3.7 If x and y be two random variables and z a third random variable such that $z=x+y$, then $E(z)=E(x)+E(y)$.

Proof: Corresponding to the elementary event e_a , x , y and z have the values $x(e_a)$, $y(e_a)$ and $z(e_a)$, respectively.

Further, $z(e_a)=x(e_a)+y(e_a)$. Hence

$$\begin{aligned} E(z) &= \sum_{a=1}^r z(e_a)P(e_a) \\ &= \sum_{a=1}^r x(e_a)P(e_a) + \sum_{a=1}^r y(e_a)P(e_a) \\ &= E(x) + E(y). \end{aligned}$$

Theorem 3.8 If $y=a+bx$, then $E(y)=a+bE(x)$.

Proof: $E(y)=E(a)+E(bx)$ from *Theorem 3.7*

$$=a+bE(x) \quad \text{from } \textit{Theorems 3.5 and 3.6}.$$

Another characteristic of a random variable is its *variance*, which serves as a measure of the variation or dispersion of the random variable about its expectation. The variance of a random variable x is defined by

$$\text{var}(x)=E[x-E(x)]^2. \quad \dots \quad (3.18)$$

Often $\text{var}(x)$ is denoted by σ_x^2 or, simply, σ^2 . The positive square-root of the variance is called the *standard deviation* (denoted by σ_x or σ).

Since

$$[x-E(x)]^2=x^2-2xE(x)+[E(x)]^2,$$

we have, by virtue of *Theorems 3.5—3.8*,

$$\begin{aligned} \text{var}(x) &= E(x^2)-2E(x)\cdot E(x)+[E(x)]^2 \\ &= E(x^2)-[E(x)]^2. \quad \dots \quad (3.19) \end{aligned}$$

We have the following theorems on variance :

Theorem 3.9 If $x=a$, a constant, then $\text{var}(x)=0$.

Proof: From *Theorem 3.5*, $E(x)=a$. Hence $[x-E(x)]^2=0$ and so, from *Theorem 3.5* again,

$$\begin{aligned} \text{var}(x) &= E[x-E(x)]^2 \\ &= 0. \end{aligned}$$

Theorem 3.10 If $y = bx$, then $\text{var}(y) = b^2 \text{var}(x)$

Proof From *Theorem 3.6*, $E(y) = bE(x)$

$$\text{Hence } y - E(y) = b[x - E(x)]$$

$$\text{and } [y - E(y)]^2 = b^2[x - E(x)]^2$$

On applying *Theorem 3.6* again, we have

$$E[y - E(y)]^2 = b^2 E[x - E(x)]^2,$$

$$\text{i.e. } \text{var}(y) = b^2 \text{var}(x)$$

Theorem 3.11 If $y = a + bx$, then $\text{var}(y) = b^2 \text{var}(x)$

Proof *Theorem 3.8* gives $E(y) = a + bE(x)$. Hence $y - E(y) = b[x - E(x)]$. Next, proceeding as in the proof of *Theorem 3.10*, we have the stated result

Ex 3.12 Suppose two players, *A* and *B*, agree to play a game under the condition that *A* will get from *B* *a* rupees if he wins and will pay to *B* *b* rupees if he loses. Let the probability of *A*'s winning the game be *p* and that of *B*'s winning the game be $q = 1 - p$.

A's gain is then a random variable *x*, assuming two values, *a* (with probability *p*) and $-b$ (with probability *q*). Hence

$$E(x) = ap - bq$$

$$\begin{aligned} \text{and } \text{var}(x) &= [a - E(x)]^2 p + [-b - E(x)]^2 q \\ &= (a + b)^2 q^2 p + (a + b)^2 p^2 q \\ &= (a + b)^2 pq \end{aligned}$$

Ex 3.13 Let a die be thrown repeatedly till the first six appears. The number of throws needed to get the first six is then a random variable *x*, taking the values 1, 2, 3, ..., ad inf. Further, if the die is perfect, the probability of getting a six in a throw is $1/6$ and that of getting one of the other values (viz 1, 2, ..., 5) is $5/6$. Hence

$$P[x = k] = \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6},$$

for the throws are made independently of each other and $x = k$ if, and only if, a six is obtained in the *k*th throw but in none of the earlier throws.

The mathematical expectation of x is, therefore,

$$\begin{aligned} E(x) &= \sum_{k=1}^{\infty} k \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6} \\ &= \frac{1}{6} \left[1 + 2 \left(\frac{5}{6}\right) + 3 \left(\frac{5}{6}\right)^2 + \dots \dots \right] \\ &= \frac{1}{6} \left(1 - \frac{5}{6} \right)^{-2} = 6. \end{aligned}$$

Thus 'on the average' six throws will be needed to get the first six.

Again,

$$x^2 = x(x-1) + x,$$

so that

$$\begin{aligned} E(x^2) &= E[x(x-1)] + E(x) \\ &= \sum_{k=1}^{\infty} k(k-1) \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6} + 6 \\ &= \frac{1}{6} \left[2 \times 1 \left(\frac{5}{6}\right) + 3 \times 2 \left(\frac{5}{6}\right)^2 + 4 \times 3 \left(\frac{5}{6}\right)^3 + \dots \dots \right] \\ &\quad + 6 \\ &= 2 \times \frac{1}{6} \times \frac{5}{6} \left[1 + 3 \left(\frac{5}{6}\right) + \frac{3 \times 4}{2!} \left(\frac{5}{6}\right)^2 + \frac{3 \times 4 \times 5}{3!} \left(\frac{5}{6}\right)^3 + \dots \dots \right] \\ &\quad + 6 \\ &= 2 \times \frac{1}{6} \times \frac{5}{6} \left(1 - \frac{5}{6} \right)^{-3} + 6 = 60 + 6 = 66. \end{aligned}$$

Hence

$$\begin{aligned} \text{var}(x) &= E(x^2) - [E(x)]^2 \\ &= 66 - 36 = 30. \end{aligned}$$

3.9 Joint distribution of two random variables

In some investigations, we have to study more than one random variable at the same time. When a number of balls are taken from an urn containing red, white and black balls, e.g., the number of red balls obtained as well as the number of white balls may be a variable of interest. In studying the relationship of two variables, say, x and y , we have to consider the possible values of x , say, x_1, x_2, \dots, x_k , and the possible values of y , say, y_1, y_2, \dots, y_l , as also the probability for each pair of values (x_i, y_j) , where $i=1, 2, \dots, k$ and

$j=1, 2, \dots, l$. We shall denote this probability by p_{ij} . The p_{ij} 's give what is called the *joint probability distribution* of x and y . This may be represented by a two-way table like Table 3.1.

TABLE 3.1
JOINT DISTRIBUTION OF TWO RANDOM VARIABLES

$y \backslash x$	x_1	x_2	\dots	\dots	x_k	Marginal total
y_1	p_{11}	p_{12}	\dots	\dots	p_{1k}	p_{01}
y_2	p_{21}	p_{22}	\dots	\dots	p_{2k}	p_{02}
\vdots	\vdots	\vdots			\vdots	\vdots
y_l	p_{l1}	p_{l2}	\dots	\dots	p_{lk}	p_{0l}
Marginal total	p_{10}	p_{20}	\cdot	\dots	p_{k0}	1

Let

$$p_{10} = \sum_{j=1}^l p_{ij}, \quad \dots \quad (3.20)$$

Clearly,

$$p_{10} = P[x=x_i].$$

Also, let

$$p_{0j} = \sum_{i=1}^k p_{ij}, \quad \dots \quad (3.21)$$

so that

$$p_{0j} = P[y=y_j].$$

The p_{10} 's give the probability distribution (called the *marginal probability distribution*) of x . Likewise, the p_{0j} 's give the marginal distribution of y .

In studying the interdependence of x and y , we have to examine how $P[x=x_i, y=y_j]$ differs from the product $P[x=x_i]P[y=y_j]$, for each pair (i, j) . In case

$$P[x=x_i, y=y_j] = P[x=x_i]P[y=y_j], \quad \text{for all } i, j,$$

$$\text{i.e. } p_{ij} = p_{10} p_{0j}, \quad \text{for all } i, j, \quad \dots \quad (3.22)$$

we say that x and y are *statistically independent*. Otherwise, they are said to be *statistically associated*.

An important feature of the joint distribution of x and y is their covariance, which is used in measuring the association between x and y and is defined by

$$\text{cov}(x, y) = E[x - E(x)][y - E(y)]. \quad \dots \quad (3.23)$$

Since

$$[x - E(x)][y - E(y)] = xy - xE(y) - yE(x) + E(x)E(y),$$

we have, by virtue of Theorems 3.5—3.8,

$$\text{cov}(x, y) = E(xy) - E(x)E(y). \quad \dots \quad (3.24)$$

The ratio

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}, \quad \dots \quad (3.25)$$

called the *correlation coefficient* of x and y , is used as a measure of the association between the two random variables. The various properties of ρ are discussed in Chapter 11.

Ex. 3.14 An urn contains 2 red, 3 white and 3 black balls. If 3 balls are taken at random from the urn, what will be the covariance between the number of red balls and the number of white balls obtained?

TABLE 3.2
JOINT DISTRIBUTION OF THE NUMBER OF RED BALLS (x)
AND THE NUMBER OF WHITE BALLS (y)

$x \backslash y$	0	1	2	Marginal total
0	1/56	6/56	3/56	10/56
1	9/56	18/56	3/56	30/56
2	9/56	6/56	0	15/56
3	1/56	0	0	1/56
Marginal total	20/56	30/56	6/56	1

Let the number of red balls and that of white balls be denoted by x and y , respectively. The possible values of x are 0, 1 and 2, while the possible values of y are 0, 1, 2 and 3.

Also, $P_{ij} = P[x=x_i, y=y_j]$

$$= \frac{\binom{2}{x_i} \binom{3}{y_j} \binom{3}{3-x_i-y_j}}{\binom{8}{3}} \quad \text{for } \begin{cases} x_i = 0, 1, 2, \\ y_j = 0, 1, 2, 3. \end{cases}$$

The joint distribution and the marginal distributions of x and y are represented in Table 3.2.

Hence

$$E(x) = 1 \times \frac{30}{56} + 2 \times \frac{6}{56} = \frac{42}{56} = \frac{3}{4},$$

$$E(y) = 1 \times \frac{30}{56} + 2 \times \frac{15}{56} + 3 \times \frac{1}{56} = \frac{63}{56} = \frac{9}{8},$$

and $E(xy) = 1 \times \frac{18}{56} + 2 \times \frac{6}{56} + 2 \times \frac{3}{56} = \frac{36}{56} = \frac{9}{14}.$

Thus

$$\text{cov}(x, y) = \frac{9}{14} - \frac{3}{4} \times \frac{9}{8} = -\frac{45}{224}.$$

The following theorem on the probability distribution of two independent random variables is of frequent use in statistical theory :

Theorem 3.12 If x and y are independent random variables, then

$$E(xy) = E(x)E(y).$$

Proof: Using the notation already introduced, the expectation of the product xy may be written as

$$E(xy) = \sum_{i=1}^k \sum_{j=1}^l x_i y_j p_{ij},$$

where we are taking formula (3.17) for expectation. (The values x_i, y_j may not be all different, but they arise from an exhaustive set of mutually exclusive cases.) Now, since x and y are supposed to be independent, $p_{ij} = p_{i0} p_{0j}$ for all i, j . Hence

$$E(xy) = \sum_{i=1}^k \sum_{j=1}^l x_i y_j p_{i0} p_{0j}.$$

Since $x_i y_j p_{i0} p_{0j} = (x_i p_{i0})(y_j p_{0j})$, where the first factor depends on i alone and the second depends on j alone, we may write the above double sum as the product of two sums :

$$E(xy) = \left(\sum_{i=1}^k x_i p_{i0} \right) \left(\sum_{j=1}^l y_j p_{0j} \right).$$

But the first factor on the right-hand side is, by definition (3.16), the expectation of x and the second factor is similarly the expectation of y . As such,

$$E(xy) = E(x)E(y).$$

Corollary 3.12.1 If x and y are independent, we have, from (3.24) and (3.25),

$$\text{cov}(x, y) = 0$$

and

$$\rho_{xy} = 0.$$

In Section 3.8 we had an expression for the expectation of the sum of a number of random variables in terms of the expectations of the components. Here we give a corresponding expression for the variance of the sum of a number of random variables in terms of their variances and covariances.

Theorem 3.13 Let x_1, x_2, \dots, x_m be m random variables defined on the same sample space. Then

$$\text{var}(x_1 + x_2 + \dots + x_m) = \sum_{i=1}^m \text{var}(x_i) + 2 \sum_{\substack{i,j=1 \\ i < j}}^m \text{cov}(x_i, x_j).$$

Proof: Let us again start with two random variables, x and y . We have, from *Theorem 3.7*,

$$E(x+y) = E(x) + E(y).$$

Hence

$$\begin{aligned} \text{var}(x+y) &= E[(x+y) - E(x+y)]^2 \\ &= E[\{x-E(x)\} + \{y-E(y)\}]^2 \\ &= E[\{x-E(x)\}^2 + 2\{x-E(x)\}\{y-E(y)\} + \{y-E(y)\}^2] \\ &= \text{var}(x) + 2 \text{cov}(x, y) + \text{var}(y), \end{aligned}$$

by virtue of *Theorem 3.7*.

In the same way, for m random variables, x_1, x_2, \dots, x_m ,

$$\begin{aligned} \text{var}(x_1 + \dots + x_m) &= E[\sum_{i=1}^m \{x_i - E(x_i)\}]^2 \\ &= E[\sum_{i=1}^m \{x_i - E(x_i)\}^2 + 2 \sum_{\substack{i,j=1 \\ i < j}}^m \{x_i - E(x_i)\}\{x_j - E(x_j)\}] \\ &= \sum_{i=1}^m \text{var}(x_i) + 2 \sum_{\substack{i,j=1 \\ i < j}}^m \text{cov}(x_i, x_j). \end{aligned}$$

Corollary 3.13.1 If x_1, x_2, \dots, x_m are (at least pairwise) independent, then $\text{cov}(x_i, x_j) = 0$ for $i \neq j$. Hence in this case we have, simply,

$$\text{var}(x_1 + x_2 + \dots + x_m) = \sum_{i=1}^m \text{var}(x_i)$$

Ex 3.15 There is a lot of N objects, from which objects are taken at random one by one with replacement. What are the expected value and variance of the least number of drawings needed to get n different objects?

(Note that if the objects were taken without replacement, then just n drawings would be needed to get n different objects. Hence in that case the expected value would be n and the variance 0.)

Denoting the number of drawings required to obtain n different objects by x , we may write

$$x = 1 + x_1 + x_2 + \dots + x_{n-1},$$

where x_i = the least number of drawings required to obtain a new object when i different objects have already been obtained. From the nature of the problem, it is obvious that x_1, x_2, \dots, x_{n-1} are mutually (hence pairwise) independent, and for each i the possible values of x_i are 1, 2, ..., ad inf, with

$$P[x_i = k] = \left(\frac{i}{N}\right)^{k-1} \frac{N-i}{N}, \quad \text{for } k=1, 2,$$

Proceeding as in Ex 3.13, we then have

$$E(x_i) = \frac{N-i}{N} \left(1 - \frac{i}{N}\right)^{-1} = \frac{N}{N-i}$$

$$\begin{aligned} \text{and } E(x_i^2) &= \frac{N-i}{N} \left\{ 2 \cdot \frac{i}{N} \left(1 - \frac{i}{N}\right)^{-2} + \left(1 - \frac{i}{N}\right)^{-1} \right\} \\ &= 2 \cdot \frac{i}{N} \left(\frac{N}{N-i} \right)^2 + \frac{N}{N-i}, \end{aligned}$$

$$\text{so that } \text{var}(x_i) = \frac{N_i}{(N-i)^2}$$

Using Theorem 3.7 and Corollary 3.13.1, we get

$$\begin{aligned} E(x) &= 1 + E(x_1) + E(x_2) + \dots + E(x_{n-1}) \\ &= N \left\{ \frac{1}{N} + \frac{1}{N-1} + \dots + \frac{1}{N-n+1} \right\} \end{aligned}$$

and $\text{var}(x) = \text{var}(x_1) + \text{var}(x_2) + \dots + \text{var}(x_{n-1})$

$$= N \left\{ \frac{1}{(N-1)^2} + \frac{2}{(N-2)^2} + \dots + \frac{n-1}{(N-n+1)^2} \right\}.$$

(Use is made of the same technique in two other problems in Sections 14.3 and 14.4).

3.10 Law of large numbers

Before this chapter is brought to a close, we should state and prove certain theorems which have come to have an important bearing on statistical theory.

Theorem 3.14 (Chebyshev's lemma): Let x be a random variable that takes non-negative values only (more generally, let $P[x < 0] = 0$), and let μ be its mathematical expectation. Then, whatever be the non-zero quantity t , we must have

$$P[x \leq \mu t^2] > 1 - \frac{1}{t^2}.$$

Proof: In case $\mu = 0$, the result is trivially true. For the only possible value of x is then 0, so that

$$P[x \leq \mu t^2] = P[x = 0] = 1 > 1 - \frac{1}{t^2}, \text{ for all } t \neq 0.$$

We need, therefore, consider only the case when $\mu \neq 0$. Let the different possible values of x (or, more properly speaking, the different values of x with positive probabilities) be x_1, x_2, \dots, x_k and let p_1, p_2, \dots, p_k be the corresponding probabilities. Suppose further that

$$x_i \leq \mu t^2 \text{ for } i = 1, 2, \dots, a \quad \dots \quad (3.26)$$

$$\text{and } x_i > \mu t^2 \text{ for } i = a+1, a+2, \dots, k. \quad \dots \quad (3.27)$$

Now, by definition,

$$\mu = \sum_{i=1}^k x_i p_i,$$

and since x_i and p_i are all non-negative quantities,

$$\mu \geq \sum_{i=a+1}^k x_i p_i.$$

Again, by virtue of (3.27),

$$\mu \geq \sum_{i=a+1}^k x_i p_i > \mu t^2 \sum_{i=a+1}^k p_i.$$

But

$$\sum_{i=t+1}^k p_i = P[x > \mu t^2]$$

It follows that

$$\mu t^2 P[x > \mu t^2] < \mu$$

or

$$P[x > \mu t^2] < \frac{1}{t^2} \quad (\text{since } \mu > 0)$$

Because

$$P[x \leq \mu t^2] = 1 - P[x > \mu t^2],$$

we then have

$$P[x \leq \mu t^2] > 1 - \frac{1}{t^2}$$

Corollary 3.14.1 (Chebyshev's inequality) Since, for any random variable x with expectation μ , $(x-\mu)^2$ is itself a random variable assuming non negative values only and having expectation

$$E(x-\mu)^2 = \sigma^2,$$

which is the variance of x , we get, from Chebyshev's lemma,

$$P[(x-\mu)^2 \leq \sigma^2 t^2] > 1 - \frac{1}{t^2}$$

or, equivalently,

$$P[|x-\mu| \leq t\sigma] > 1 - \frac{1}{t^2},$$

where t is any positive quantity

This gives a fair idea of the sense in which σ may be looked upon as a measure of dispersion (*vide* Sections 7.4 and 9.2)

Theorem 3.15 (Weak law of large numbers) Let x_1, x_2, \dots be a sequence of random variables having expectations μ_1, μ_2, \dots

Further, let

$$V_n = \text{var}(x_1 + x_2 + \dots + x_n)$$

If

$$\frac{V_n}{n^2} \rightarrow 0 \text{ as } n \rightarrow \infty,$$

then, given any two positive quantities ϵ and η , however small, we can find an n_0 , depending on ϵ and η , such that

$$P \left[\left| \frac{x_1 + x_2 + \dots + x_n - \mu_1 - \mu_2 - \dots - \mu_n}{n} \right| \leq \epsilon \right] > 1 - \eta$$

for all $n \geq n_0$

Proof: Since

$$E\left(\frac{x_1+x_2+\dots+x_n}{n}\right) = \frac{\mu_1+\mu_2+\dots+\mu_n}{n}$$

and $\text{var}\left(\frac{x_1+x_2+\dots+x_n}{n}\right) = \frac{V_n}{n^2}$,

we have, from Chebyshev's inequality,

$$P\left[\left|\frac{x_1+x_2+\dots+x_n}{n} - \frac{\mu_1+\mu_2+\dots+\mu_n}{n}\right| \leq t\sqrt{\frac{V_n}{n}}\right] > 1 - \frac{1}{t^2}$$

for any positive t .

Choosing t in such a way that $t\sqrt{\frac{V_n}{n}} = \epsilon$, we therefore have, for any $\epsilon > 0$,

$$P\left[\left|\frac{x_1+x_2+\dots+x_n}{n} - \frac{\mu_1+\mu_2+\dots+\mu_n}{n}\right| \leq \epsilon\right] > 1 - \frac{V_n}{n^2\epsilon^2}.$$

Since $\frac{V_n}{n^2} \rightarrow 0$ as $n \rightarrow \infty$, given $\eta\epsilon^2 > 0$, however small, one can find an n_0 , depending on $\eta\epsilon^2$, such that

$$\frac{V_n}{n^2} < \eta\epsilon^2$$

for all $n \geq n_0$. For such an n_0 we have, therefore,

$$P\left[\left|\frac{x_1+x_2+\dots+x_n}{n} - \frac{\mu_1+\mu_2+\dots+\mu_n}{n}\right| \leq \epsilon\right] > 1 - \eta$$

whenever $n \geq n_0$. This proves the theorem.

Corollary 3.15.1 Let x_1, x_2, \dots be independently distributed with identical marginal distributions. Their common expectation will be denoted by μ and their common variance by σ^2 . Now $(x_1+x_2+\dots+x_n)/n = \bar{x}_n$ may be looked upon as a sample mean for a random sample (*vide* Section 14.5) of size n from a population with mean μ and variance σ^2 .

Also,

$$V_n = \sum_{i=1}^n \text{var}(x_i) = n\sigma^2,$$

so that $\frac{V_n}{n^2}$, which equals $\frac{\sigma^2}{n}$, $\rightarrow 0$ as $n \rightarrow \infty$ if σ^2 be finite. Hence given

$\epsilon > 0$ and $\eta > 0$, however small, we can find an n_0 , depending on ϵ and η , such that

$$P[|x - \mu| \leq \epsilon] > 1 - \eta$$

for all $n \geq n_0$, provided σ^2 is finite *

Corollary 3.15.2 Suppose we have a series of independent repetitions of an experiment for which the probability of an event A is $p_0 = P(A)$. Let x_i be a random variable associated with the i th repetition such that

$$x_i = \begin{cases} 1 & \text{if in the } i\text{th repetition } A \text{ occurs,} \\ 0 & \text{otherwise} \end{cases}$$

Then x_1, x_2, \dots are independent and identically distributed random variables, with

$$E(x_i) = 1 \times p_0 + 0(1 - p_0) = p_0$$

$$\text{and } \text{var}(x_i) = (1 - p_0)^2 p_0 + (0 - p_0)^2 (1 - p_0) = p_0(1 - p_0)$$

Also,

$$(x_1 + x_2 + \dots + x_n)/n = f_n/n \text{ (say)}$$

is the relative frequency of the event A in the first n repetitions of the experiment

Since $p_0(1 - p_0)$ is necessarily finite ($0 \leq p_0(1 - p_0) \leq 1/4$), it follows from the previous result that, given $\epsilon > 0$ and $\eta > 0$, however small, we can find an n_0 , depending on ϵ and η , such that

$$P\left[\left|\frac{f_n}{n} - p_0\right| \leq \epsilon\right] > 1 - \eta$$

for any $n \geq n_0$

This particular form of the law of large numbers has been called *Bernoulli's theorem* after James Bernoulli, who, however, arrived at the result by a much more elaborate argument

Bernoulli's theorem justifies and makes more precise the statement made in Section 3.7 that the probability of an event in an experiment is to be looked upon as the 'long-run relative frequency' of the event in repetitions of the experiment

* If the sample space be finite as we assumed in earlier sections, σ^2 must be finite too.

Questions and exercises

3.1 In what sense do we use the word ‘probability’ in statistical theory? Give the classical definition of probability and point out its limitations.

3.2 Give an outline of the axiomatic treatment of probability theory.

3.3 State and prove the theorem of total probability.

3.4 Supply an alternative proof for the theorem of total probability (*Theorem 3.2*) by showing that an elementary event ω that is favourable to *none* of the events A_1, A_2, \dots, A_m contributes 0 to the right-hand side of the equation, while an elementary event ω that is favourable to *exactly* t ($1 \leq t \leq m$) of the events contributes $P(\{\omega\})$.

3.5 Define conditional probability; state and prove the theorem of compound probability. What is meant by saying that a number of events are independent?

3.6 What is a random variable, and what are its expectation and variance? What is the covariance of two random variables?

3.7 State and prove the (weak) law of large numbers. Deduce, as a corollary, Bernoulli’s theorem and comment on its implications.

3.8 A_1 and A_2 are two events related to an experiment E . Given $P(A_1) = 1/2$, $P(A_2) = 1/3$ and $P(A_1 \cap A_2) = 1/4$, determine the following probabilities:

- (a) $P(A_1^c \cup A_2^c)$, (b) $P(A_1^c \cap A_2^c)$, (c) $P(A_1^c \cup A_2)$,
- (d) $P(A_1^c \cap A_2)$, (e) $P([A_1 \cup A_2]^c)$, (f) $P([A_1 \cap A_2]^c)$.

3.9 Arrange the following quantities in increasing order of magnitude with proper equality or inequality signs between them:

$$P(A_1), P(A_1) + P(A_2), P(A_1 \cup A_2), P(A_1 \cap A_2).$$

3.10 Assume that neither A nor B is an impossible event. (a) If now A and B are mutually exclusive, will they be independent? (b) If A and B are independent, will they be mutually exclusive?

3.11 Eight students are arranged at random (a) in a row and (b) in a ring. In each case, find the probability that two given students will be next to each other. *Ans.* (a) $\frac{1}{4}$; (b) $\frac{2}{7}$.

3.12 Three numbers are chosen at random from the first n natural numbers. What is the probability that the chosen numbers

will be in arithmetic progression? (Consider separately the cases when n is even and when n is odd)

$$\text{Ans } \frac{3n}{[2(n-1)(n-2)]} \text{ if } n \text{ is even,}$$

$$\frac{3(n-1)}{[2n(n-2)]} \text{ if } n \text{ is odd}$$

313 The nine digits 1, 2, ..., 9 are arranged in random order to form a nine digit number. Find the probability that 1, 2 and 3 appear as neighbours in the order mentioned. $\text{Ans } \frac{1}{4}$

314 Obtain the probability that the birth-days of seven people will fall on seven different days of the week, assuming equal probabilities for the seven days. $\text{Ans } 6!/7^6$

315 A box contains 40 envelopes, of which 25 are ordinary (not meant for air mail) and 16 are unstamped, while the number of unstamped ordinary envelopes is 10. What is the probability that an envelope chosen from the box is a stamped air mail envelope?

$$\text{Ans } \frac{3}{4}$$

316 Suppose A_1, A_2, \dots, A_k are independent events and $P(A_i) = p_i$. Find the probability (a) that none of the events occurs, (b) that at least one of the events occurs.

317 Two players, A and B , throw $n+1$ and n coins, respectively. Show that the probability that A will have more heads than B is $1/2$.

318 Two players A and B , throw a pair of dice. A , who starts the game, wins if he throws six before B throws seven and B wins if he throws seven before A throws six. Obtain the probability that A will win the game. $\text{Ans } \frac{3}{8}$

319 From a full pack of playing cards, 3 cards are taken at random. Evaluate each of the following probabilities and verify that their sum is unity.

- (a) that the cards are of the same denomination,
- (b) that 2 are of the same denomination and one different,
- (c) that all are of different denominations

$$\text{Ans } (a) \frac{4}{55}, (b) \frac{43}{55}, (c) \frac{11}{55}$$

320 What is the probability that in a game of poker, which requires the drawing of 5 cards from a full pack, all the cards drawn will be (a) of the same colour, (b) of the same suit?

$$\text{Ans } (a) \frac{1}{425}, (b) \frac{3}{425}$$

3.21. 15 balls are distributed at random among 5 boxes. What is the probability that exactly 2 boxes will remain empty?

$$\text{Ans. } 2/5^{14}[3^{15} - 3 \times 2^{15} + 3].$$

3.22 Obtain the probability that in k throws of a die each of the numbers 1, 2, ..., 6 will appear at least once.

$$\text{Ans. } 1 - 6\left(\frac{5}{6}\right)^k + 15\left(\frac{4}{6}\right)^k - 20\left(\frac{3}{6}\right)^k + 15\left(\frac{2}{6}\right)^k - 6\left(\frac{1}{6}\right)^k.$$

3.23 Five non-similar pairs of socks are in a closet. Four socks are selected at random. What is the probability that there will be at least one complete pair among the four socks chosen? $\text{Ans. } \frac{13}{21}.$

3.24 In the course of an experiment with a particular brand of D.D.T. on flies, it is found that 80% are killed in the initial application. Those which survive develop a resistance, so that the percentage of survivors killed in any later application is half that of the preceding application. Thus 40% of the survivors of the first application would succumb to the second, 20% of the survivors of the first two applications would succumb to the third, and so on. Find the probability

(a) that a fly will survive four applications;

(b) that it will survive four applications, given that it has survived the first one. $\text{Ans. (a) } 0.6864; \text{ (b) } 0.432.$

3.25 The probability that a family will have k children is αp^k , for $k=1, 2, \dots$, the probability that it will have no child being $1 - \alpha \sum_{k=1}^{\infty} p^k$. Assuming that the sex-ratio at birth (i.e. the ratio of male births to female births) is 1 : 1, obtain the probability that a family will have x sons. $\text{Partial ans. } 2\alpha p^x / (2-p)^{x+1}$ for $x \geq 1$.

3.26 It has been found from past experience that of the articles produced by a factory 20% come from Machine 1, 30% from Machine 2 and 50% from Machine 3. The percentages of satisfactory articles among those produced are 95% for Machine 1, 85% for Machine 2 and 90% for Machine 3. (a) An article is chosen at random from a lot. What is the probability that it is satisfactory? (b) Assuming that the article is satisfactory, what is the probability that it was produced by Machine 1? $\text{Ans. (a) } 0.895; \text{ (b) } 0.212.$

3.27 Three points, X , Y and Z , are taken at random on a line segment. What is the probability that Z lies between X and Y ?

$$\text{Ans. } \frac{1}{3}.$$

3.28 If a point is taken at random within a circle of radius r , what is the probability that its distance from the centre will exceed $r/2$? Ans $\frac{3}{4}$

3.29 Two points are taken at random on the circumference of a circle of radius r . What is the probability that the distance between them will not exceed $\pi r/4$? What will be the probability of this event if instead, the points are taken on a line segment of length $2\pi r$? Ans $\frac{1}{4} \cdot 1 - \frac{\pi^2}{16} = \frac{15}{16}$

3.30 A chord is drawn at random in a given circle. What is the probability that it is greater than the side of the equilateral triangle inscribed in the circle? Ans $\frac{1}{2}$

Ans $\frac{1}{2}$ if distance of chord from centre is chosen at random, $\frac{1}{3}$ if angle between chord and tangent at one end of it is chosen at random

3.31 Three points are taken at random on the circumference of a circle. What is the probability that they will all lie in a semi circle? Ans $\frac{3}{4}$

3.32 A dealer in electrical goods receives 20 new lamp bulbs. He tests them one by one until he finds one of satisfactory quality. Suppose x denotes the number of bulbs tested by the dealer. Give the possible values of x and the corresponding probabilities, assuming that there are 4 defective bulbs in the batch. Hence obtain the expected value and the variance of x .

3.33 10 000 tickets, each of Rs 2/-, are to be sold in a lottery in which there are 2 prizes of Rs 5,000/- each, 10 of Rs 200/- each and 100 of Rs 50/- each. Determine the expected gain (or loss) of a man who buys a ticket for the lottery. Ans Expected loss 30 P

3.34 In a lottery n tickets are drawn at a time out of N tickets numbered from 1 to N . Find the expectation and variance of the sum of the numbers on the tickets drawn.

Ans $n(N+1)/2, \frac{n(N^2-1)}{12} \left(1 - \frac{n-1}{N-1}\right)$

3.35 An urn has been filled with 25 balls, each of which is either black or white, by a chance mechanism. The expectation of the number of black balls is known to be 9. Show that the probability of drawing a black ball from the urn is 9/25.

3.36 Balls are taken one by one out of an urn containing a white and b black balls until the first white ball appears. Determine the expected number of black balls preceding the first white ball.

Ans. $b/(a+1)$.

3.37 (a) Give a counter-example to show that the converse of *Corollary 3.12.1* does not hold.

(b) Prove that if each of x and y takes two possible values, then zero covariance implies independence.

3.38 The random variables x_1, x_2, \dots are independent and x_i assumes just two values, $-\frac{1}{2^i}$ and $\frac{1}{2^i}$, with equal probabilities. Show that the law of large numbers holds for these variables.

3.39 Let the random variables x_1, x_2, \dots be such that x_i may depend on x_{i-1} and x_{i+1} but is independent of all other variables. Show that the law of large numbers holds, provided each variable has a finite variance.

3.40 Let the random variables x_1, x_2, \dots be such that the variances are bounded and the covariances are finite. Show that the law of large numbers applies.

SUGGESTED READING

- [1] Cramér, H. *The Elements of Probability Theory* (Chs. 1–4). John Wiley, 1955.
- [2] Feller, W. *An Introduction to Probability Theory and Its Applications*, Vol. I (Chs. 1–5, 9). John Wiley, 1957, and Wiley Eastern, 1960.
- [3] Goldberg, S. *Probability, an Introduction*. Prentice-Hall, 1960.
- [4] Parzen, E. *Modern Probability Theory and Its Applications* (Chs. 1–3, 5, 7, 8). John Wiley, 1960, and Toppian.
- [5] Uspensky, J. V. *Introduction to Mathematical Probability* (Chs. 1–3, 9, 10). McGraw-Hill, 1937.

PART TWO

GENERAL STATISTICAL METHODS

4.1 The nature of statistics

Although in modern times the word *statistics* is used on diverse occasions, there seems to be a lot of misconception regarding its connotation. It is, therefore, proper that we should begin our discussion by describing in a general way what ‘statistics’ really means.

‘Statistics’, as a plural noun, is used to mean numerical data arising in some sphere of human experience—to be precise, numerical data which result from a host of uncontrolled, and mostly unknown, causes acting together. It is in this sense that the term is used when our daily newspapers give vital statistics, crime statistics or soccer statistics of Calcutta, or when the Food Minister in the Lok Sabha quotes statistics of food production in India.

Used as singular, ‘statistics’ is a name for the body of scientific methods (the *statistical methods*) which are meant for the collection, analysis and interpretation of numerical data. One has this sense in mind when one says that this is a text-book of statistics or that Mr So-and-so is an outstanding statistician. Not that Mr So-and-so has got numerical data on different topics at his finger-tips ; he is just well versed in statistical methods and their applications.

The importance of statistical methods cannot be over-emphasised. The collection of facts and figures on different aspects of life and their analysis have become indispensable tasks of modern governments. Many a government has accepted planning as a matter of policy, and no planning is conceivable without the help of statistics. Business executives, again, are relying more and more on statistical techniques for controlling the quality of manufactured products and for studying the needs and desires of the consumers. To these people we may add politicians and social reformers who employ statistical facts as a basis for policy-making. To the layman a knowledge of statistical methods may come in handy in viewing in their proper perspective the data which are presented to him day in and day out by the Government

and the political parties This will make him a more responsible citizen, not to be easily taken in by unscrupulous propaganda

But statistical methods are to be regarded essentially as aids to scientific research Economics, biology, agriculture, meteorology,—to mention only a few of the sciences—have a common feature in that the phenomena with which they deal are caused by forces beyond the control of the investigator In fact, many of the tools of statistics were devised in the course of biological, agricultural and sociological investigations Even in the so called exact sciences, viz physics and chemistry, statistics is playing a useful rôle Indeed some recent theories in physics and chemistry have their origin in statistical ideas

4.2 Statistics and other disciplines

It is to be noted that statistics is not a science in the sense physics or chemistry or biology or even economics is Any science has for its objective the formulation of laws for explaining phenomena in some part of the real world Through observations or experiments a science builds up hypotheses regarding the phenomena, whose validity is then examined through further observations or experiments A hypothesis that stands repeated tests of this kind is raised to the level of a law Indeed, any science is but a body of such laws Statistics, however, is not a body of laws Among other functions, it formulates methods for the verification of hypotheses, for testing whether a hypothesis can claim to be a law It would be more proper to describe statistics as a quantitative method of scientific investigations

Also, statistics has to do with only the group characteristics of numerical data on the basis of which an investigation will be conducted The data may relate, e.g., to the scores of a number of students in a mathematics test If our interest lies in the score of each student separately, then the data will not be amenable to statistical treatment In statistics, the individual is important only in so far as it forms part of the group

Now in studying the characteristics of a group, the variation inherent in it has to be taken into account Statistics has, therefore been called the study of variation Indeed, a group without any variation, one whose members are all alike, is of no interest to

statistics. For here the numerical data would arise from a fixed system of causes—not a multiplicity of uncontrolled, and mostly unknown, causes. In this sense, statistics may be said to have ushered in a new era in scientific research.

Earlier scientists used to look at things from a deterministic viewpoint. They would ask : “Does B occur as a result of A ?” However, this meant an over-simplification of phenomena. For the real world is such that the conditions under which an experiment is conducted vary. Hence as a result of A , B may occur in some cases but may fail to occur in others. As such, it would be more pertinent to ask : “In what percentage of cases does B occur as a result of A ? Or, in other words, what is the probability that B will occur as a result of A ? ” The formulation of scientific laws has, therefore, to be made in probabilistic, rather than in deterministic, terms.

It has also to be remembered that a statistician is not interested solely in the (group) characteristics of the data on hand. He would rather like to have information about the characteristics of the bigger group of which the given individuals are but a sample. A principal task of statistics is, therefore, the inference of the numerical features of the whole group from those of a part. This problem is analogous to the problem of classical inductive logic. The only difference is that here induction has to be achieved within a probabilistic framework.

4.3 Collection of data

When one wants to study any problem by means of statistical methods, one will have first to collect the relevant numerical data. For instance, to know how the production of steel in India has been increasing since 1947, it is necessary to obtain the actual production figures for all years from 1947 to date. Sometimes the relevant data may exist in a published or an unpublished form, being collected by a private body or by the Government or by some research organisation, for its own use or for supplying popular information. In making use of such data (called *secondary data*), one has to be particularly careful about the definitions of terms and concepts used by the collecting authority and also about the method of collection and the reliability of the data.

More often, the person interested in a problem will have to collect data directly from the field of enquiry. The data are then said to be of the *primary* type. The collection of primary data may be done by interviewing a number of persons and filling in questionnaires relevant to the problem on the basis of information supplied by them. For instance, if one is interested in family income and expenditure on different items, one will generally interview the head of each family and collect the information sought from him. Sometimes it may be possible to avoid direct interviews by mailing out questionnaires to the persons or agencies concerned. The data obtained in this way are likely to be less reliable, however. There may be considerable non-response as well, in the sense that many questionnaires may not be returned by the informants.

The other method of obtaining primary data is by making *observations*. Thus, to determine the area under a particular crop in a region one has to visit each of the plots in the region and observe whether it contains the crop and, if so, in what proportion. The number of words in different pages of a book will be obtained by counting while the heights of a group of students will be found by direct measurements.

4.4 Scrutiny of data

The data collected should be examined carefully before they are subjected to statistical treatment. For, however useful statistical methods may be when properly applied, they cannot bring out any reliable information from faulty, unreliable data.

In some cases errors in the data can be readily detected. Thus, consider the following set of figures which contains two erroneous entries. sheer common sense enables the reader to detect them.

Stature (in cm) of 10 college students

140.9	161.2	153.9	172.2	162.9
159.1	147.2	773.5	181.5	1590.0

In a second type of situation, there may be figures which, although not impossible, are very unlikely to be true and should rouse suspicion. If 3 Kilograms of rice be stated to be the daily consumption of a family of 4, the matter calls for further investigation. Similarly, we should be hesitant in accepting 30 as the age of a son when the father's age is stated to be 45.

There is yet a third type of data, which may seem perfectly reliable at first sight but which, on thorough examination, may be found to be mutually inconsistent. Such internal checking should be made when we are given two or more series of figures between which some relationship is known to hold. For example, when the area, population and density of population are given for a number of regions, one can see if they are compatible with each other by verifying the relation :

$$\text{density} = \frac{\text{population}}{\text{area}}.$$

Again, if data regarding the price of a commodity during two different periods are given, together with the percentage increase in price in the later period over the earlier, the data may be checked by seeing if the relation

$$\text{p.c. increase} = \frac{\text{price in period II} - \text{price in period I}}{\text{price in period I}} \times 100$$

is satisfied.

Of course, no definite set of rules can be laid down for such scrutiny of data. One must use one's common sense, intelligence and whatever knowledge one may have about the problem under investigation.

4.5 Presentation of numerical data

The raw data collected by the investigator should be arranged in a neat, systematic form. This may be done (a) by using paragraphs of text or (b) by putting the data in a tabular form.

The following is an excerpt from the white paper on the general budget for 1957-58, appearing in the publication *Budget for 1957-58* of the Ministry of Finance, Govt. of India :

“.....The value of imports increased from Rs. 418 crores in April-November, 1955, to Rs. 535 crores in April-November, 1956. Of this increase of Rs. 117 crores, Rs. 96 crores was accounted for by the increase in the imports of machinery, iron & steel and other metals. Imports of machinery increased from Rs. 73.5 crores in April-November, 1955, to Rs. 105.8 crores in April-November, 1956 ; of iron and steel from Rs. 34.3 crores to Rs. 88 crores and of other metals from Rs. 16.4 crores to Rs. 26.2 crores.....”

The small decline in exports (from Rs 388.4 crores in April-November, 1955, to Rs 378 crores in April-November, 1956) is mostly accounted for by the decline in exports of oil (from Rs 26.2 crores to Rs 13.4 crores) and of cotton (from Rs 24.6 crores to Rs 10.9 crores) which is largely explained by the poor crop of oil-seeds and cotton in 1955-56. Exports of manganese ore also declined from Rs 10.6 crores in April-November, 1955, to Rs 6.8 crores in April-November, 1956, of cotton textiles from Rs 42.2 crores to Rs 40.1 crores, and of jute manufactures from Rs 83.3 crores to Rs 79.5 crores "

Such textual presentation of numerical data is not very effective. The reader has to go through the entire text, which is generally quite lengthy, before he can grasp the essential features of the data. It has an advantage in that the writer can draw special attention of the reader to certain points which he considers to be of special importance.

The same kind of information can be presented more effectively by means of a table. A table shows the data in a compact form, and a complete table with its title, headings and sub headings can bring all the essential features of the data into a clearer perspective. If necessary, one may add some foot notes or a short explanatory note to bring to the attention of the reader the source and the reliability of the data and other points of interest.

A table should always have a number attached to it, so that it may be easily referred to whenever necessary. Secondly, the table should possess a title which should be self-explanatory and should state briefly the nature and purpose of the data contained in the body of the table. The different columns of the table should possess headings and sub headings stating clearly what the data in different columns represent. The first column, called the 'stub' of the table, is used to give a description of the items on which data are available.

For the purpose of illustration, we give the information contained in the above excerpt in a tabular form (Table 4.1).

The columns have been numbered (1), (2), etc., so that these numbers may be quoted in any future reference.

For tabular representation no hard and fast rules can be laid down. One must exercise one's own judgement to determine in each individual case the most suitable form of tabular representation.

to be adopted. The most important criterion is clarity, and the ideal table is the one which brings the given data into the clearest perspective.

TABLE 4.1

VALUE OF IMPORTS INTO AND EXPORTS FROM INDIA DURING
APRIL-NOVEMBER, 1955, AND APRIL-NOVEMBER, 1956

Item (1)	Value (crores of rupees)		Increase (+) or Decrease (-) (4)
	April-November 1955 (2)	April-November 1956 (3)	
IMPORTS			
Machinery	73.5	105.8	+ 32.3
Iron and steel	34.3	88.0	+ 53.7
Other metals	16.4	26.2	+ 9.8
Other imports	293.8	315.0	+ 21.2
Total imports	418.0	535.0	+ 117.0
EXPORTS			
Oil	26.2	13.4	- 12.8
Cotton	24.6	10.9	- 13.7
Manganese ore	10.6	6.8	- 3.8
Cotton textiles	42.2	40.1	- 2.1
Jute manufactures	83.3	79.5	- 3.8
Other exports	201.5	227.3	+ 25.8
Total exports	388.4	378.0	- 10.4

Source : "White Paper on General Budget", *Budget for 1957-58*, Ministry of Finance, Govt. of India.

4.6 Diagrammatic representation of data

Representation of statistical data by diagrams—by graphs, charts or pictures—is more effective than tabular representation, being easily intelligible to a layman. Indeed, diagrams are almost essential whenever it is required to convey any statistical information to the general public. It must be stated, however, that information on a limited number of topics only can be presented in a single diagram so as to maintain its neatness. Moreover, a diagram can give only a rough idea about the magnitude of variations, whereas in a table the exact values may be quoted.

The more important types of diagram which are used in statistical work are being described below.

Line diagrams : Consider the data of Table 4.2, which show how the number of tourists coming to India has been changing over time. A very convenient method of representing such data is to use *line diagram*.

TABLE 4.2
NUMBER OF TOURISTS VISITING INDIA (EXCLUDING INDIAN NATIONALS ABROAD) DURING 1951—1957

Year	Number of tourists (000)
1951	19.9
1952	25.4
1953	28.1
1954	39.2
1955	43.6
1956	68.9
1957	80.5

Source : *Statistical Year Book, 1952*, United Nations.

We take the years along the horizontal axis and the number of tourists along the vertical. The numbers for the seven years give seven points on the graph, which are next joined by line segments. The resulting line diagram is shown in Fig. 4 I.

It may be noted that the line diagram would be exactly a straight line in case the values increased or decreased at a constant rate.

If we want to compare the volumes of tourist traffic for a number of countries, say India, the U.K. and the U.S.A., we may draw a line diagram for each country on the same graph paper. To distinguish the three diagrams, we may use, say, continuous lines for India, broken lines for the U.K. and dotted lines for the U.S.A.

A variant of the line diagram is the *semi-logarithmic chart* or *ratio chart*, where the vertical scale is logarithmic, but the horizontal scale is of the usual arithmetic type. Since the vertical scale

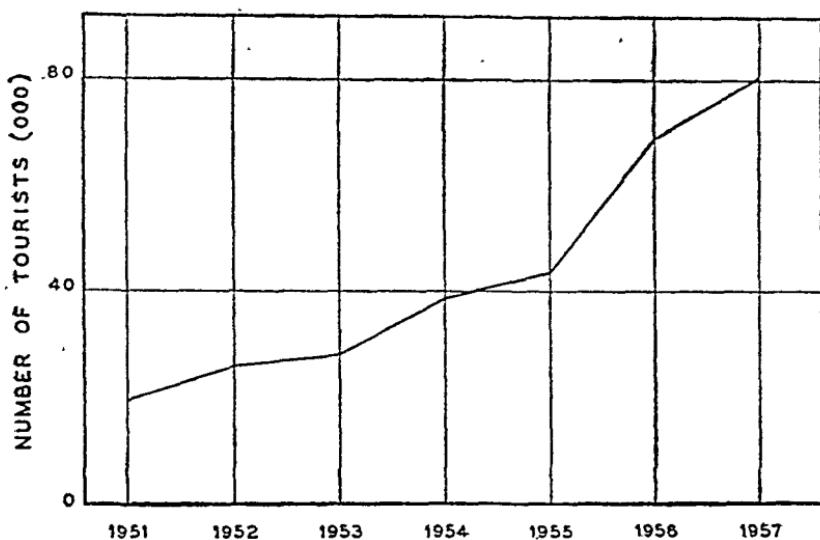


Fig. 4.1 Line diagram showing the number of tourists coming to India from abroad during 1951—1957.

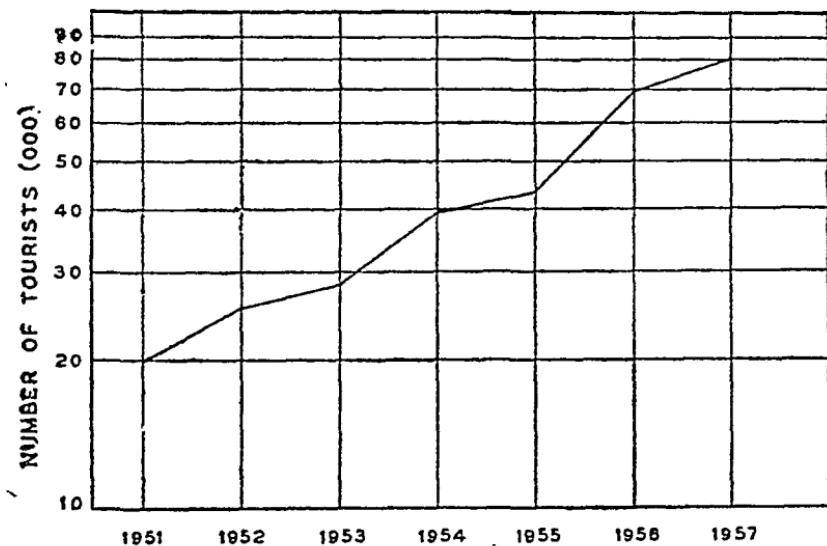


Fig. 4.2 Ratio chart showing the number of tourists coming to India from abroad during 1951—1957.

logarithmic, the distance between any two values on this scale is proportional to the difference of their logarithms. Obviously, if the variable under enquiry is increasing or decreasing at a constant ratio,

then the ratio chart will be exactly a straight line. The data of Table 4.2 are represented in a ratio chart in Fig. 4.2.

In drawing a line diagram, the following points should be borne in mind :

First, none of the axes should be too long or too short in comparison to the other, for fluctuations may seem over-emphasised in the first case and almost ironed out in the second.

Secondly, the zero of the vertical scale (but not necessarily that of the horizontal) should be included in the diagram ; otherwise, the diagram may give a false impression about the magnitude of fluctuations. However, if this is done in case the zero is too far below the actual range of variation, then the graph will lie at the top of the diagram. Here the actual variation may be brought into focus and the diagram made more agreeable to the eye by showing a definite break in the vertical axis, as in Fig. 4.4.

If the data on the variable under enquiry, varying over time, are given for each of a number of components, then we may use a type

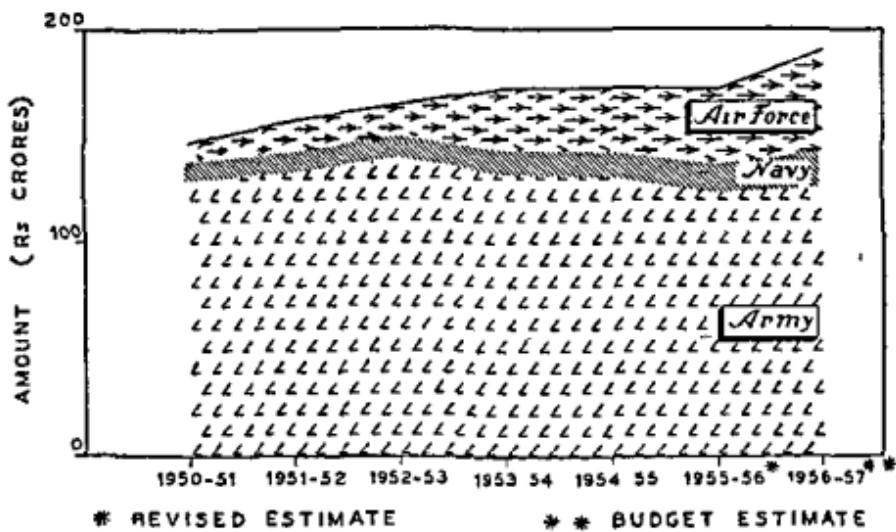


Fig. 4.3 Net expenditure on defence services, Govt. of India. (Source : *The Central Budget in Brief, 1956-57*, Govt of India.)

of line diagram called *component part chart* or *band chart*. It is actually a number of line diagrams, one for each component, superimposed one upon another. This is illustrated in Fig. 4.3, which represents

the expenditure on the defence services of India, separately for the Army, Navy and Air Force, for each year from 1950-51 to 1956-57. To distinguish the three parts of the diagram, three different shades have been used. Here the height of each band gives the expenditure on the corresponding defence service for different years, while the total height represents the total expenditure.

TABLE 4.3
AREA UNDER WHEAT WITH IRRIGATION FACILITIES AND YIELD
RATE OF WHEAT (INDIA, 1947-48 TO 1954-55)

Year	Area under wheat irrigated (000 acres)	Yield-rate of wheat (lb. per acre)
1947-48	6,882	599
1948-49	7,420	566
1949-50	7,618	584
1950-51	8,407	592
1951-52	8,506	582
1952-53	9,121	681
1953-54	9,601	670
1954-55	9,810	717

Source : *Statistical Abstract, India, 1955-57*, C.S.O., Govt. of India.

Another type of line diagram is the *multiple-axis chart*, which is meant to show the relationship between two or more series of data. In Table 4.3 the part of the area under wheat coming under irrigation and the yield-rate of wheat are given for India for a number of years. To show how the two series are related we may draw a chart of this kind. For this purpose we have to construct two line diagrams, one for each series, with a common horizontal axis but different vertical axes (Fig. 4.4).

Bar diagrams : Another mode of diagrammatic representation of data is the use of *bar diagrams*. These have more general applicability than line diagrams in the sense that they may be used for series varying either over time or over space. In this method, bars of equal width are taken for the different items of the series, the length of a bar

representing the value of the variable concerned. It is preferable to take the bars horizontally for data varying over space and vertically

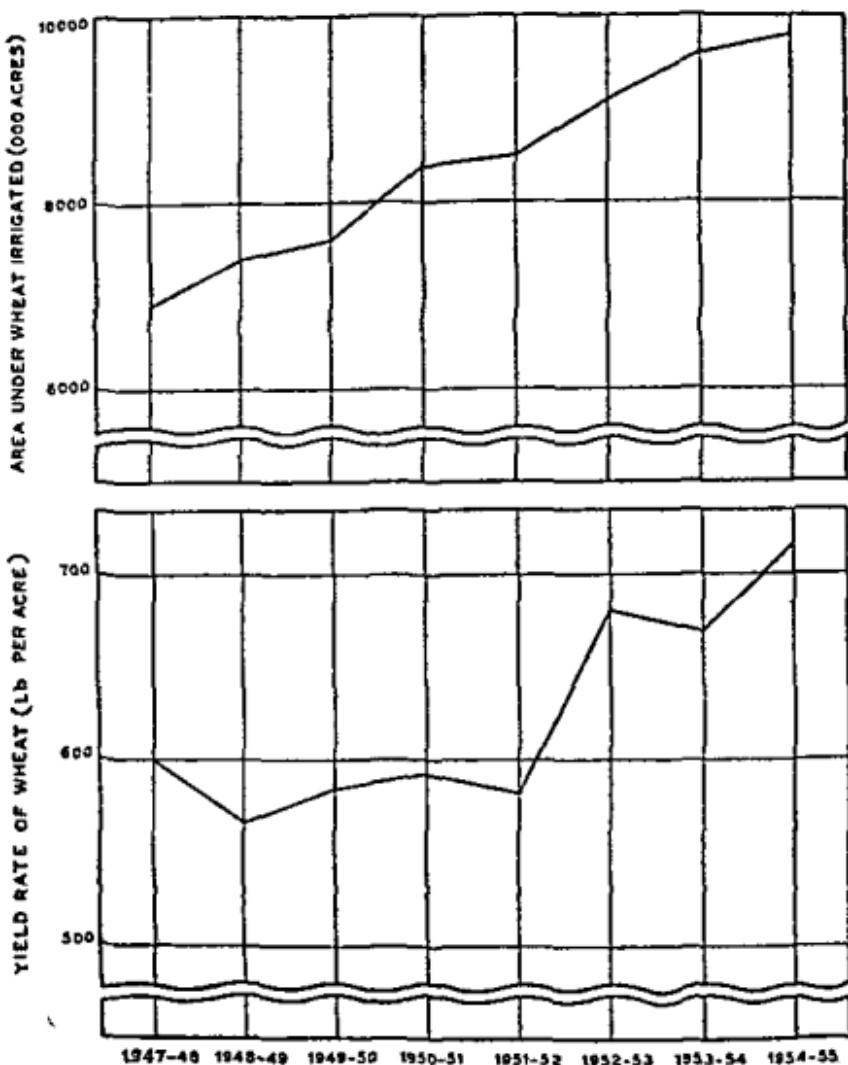


Fig. 4.4 Relationship between irrigation and yield-rate of wheat (data for India, 1947-48 to 1954-55).

in the case of a series varying over time. In Fig. 4.5 a bar diagram is shown, which represents the data of Table 4.2. Another bar diagram is given in Fig. 4.6.

A variant of the bar diagram is the *multiple bar diagram*, which is

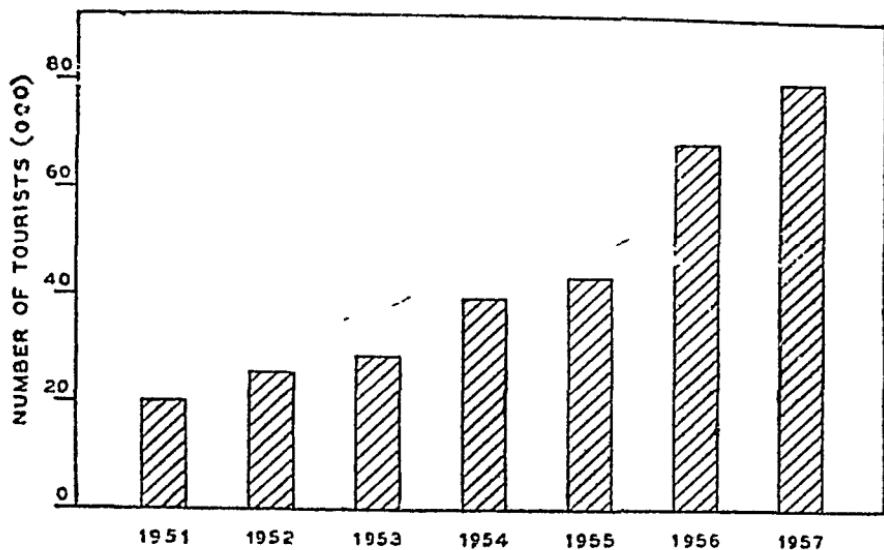


Fig. 4.5 Bar diagram indicating India's growing attractions to tourists from abroad (Table 4.2).

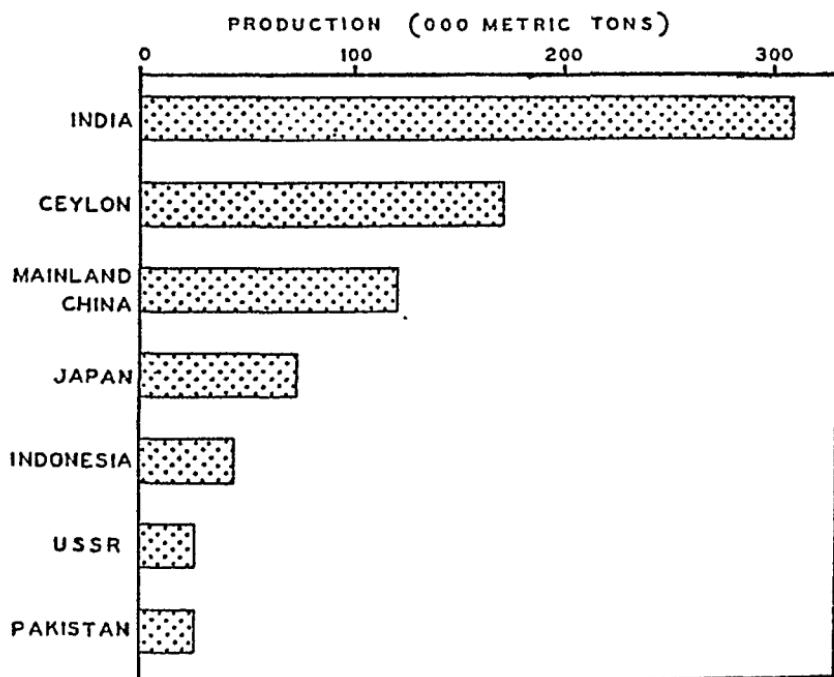


Fig. 4.6 Bar diagram showing production of tea in principal tea-growing countries in 1956.

TABLE 44
RURAL AND URBAN POPULATION OF WEST BENGAL
ACCORDING TO THE CENSUSES OF 1921—1951

Census year	Population (in millions)	
	Rural	Urban
1921	13 994	2 432
1931	14 792	2 897
1941	17 197	4 679
1951	20 025	6 282

Source *Statistical Abstract West Bengal 1957*, Govt of West Bengal

employed in comparing two or more series of data on the same variable. Thus we may want to compare the population figures, as recorded in a number of censuses, for two or more countries. Or we

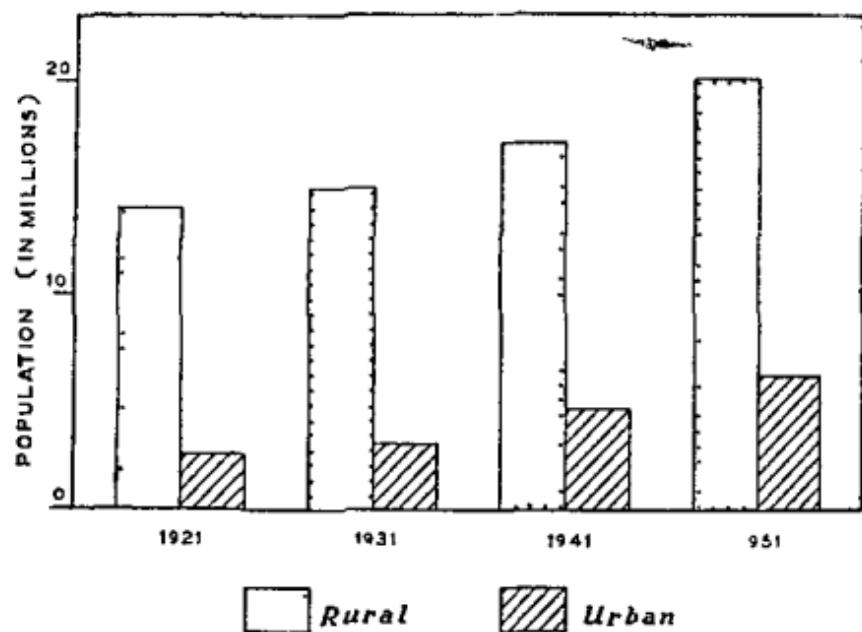


Fig. 4.7 Rural and urban population of West Bengal according to census figures

may like to compare the yield of paddy for a number of States for two or more time periods. Data of a similar type appear in Table 44.

Here we are primarily interested in knowing how the rural population of West Bengal compared with the urban population in each of the decennial censuses. To show this diagrammatically it is appropriate to use four sets of bars (for the four census years), each set containing two bars for the two sections of the population (Fig. 4.7).

Pictorial diagrams : The most vivid way of presenting numerical data is by using some pictorial device. Here a suitable symbol is first chosen to represent a certain number of units of the variable. Next, each value in the given series of data is represented either by

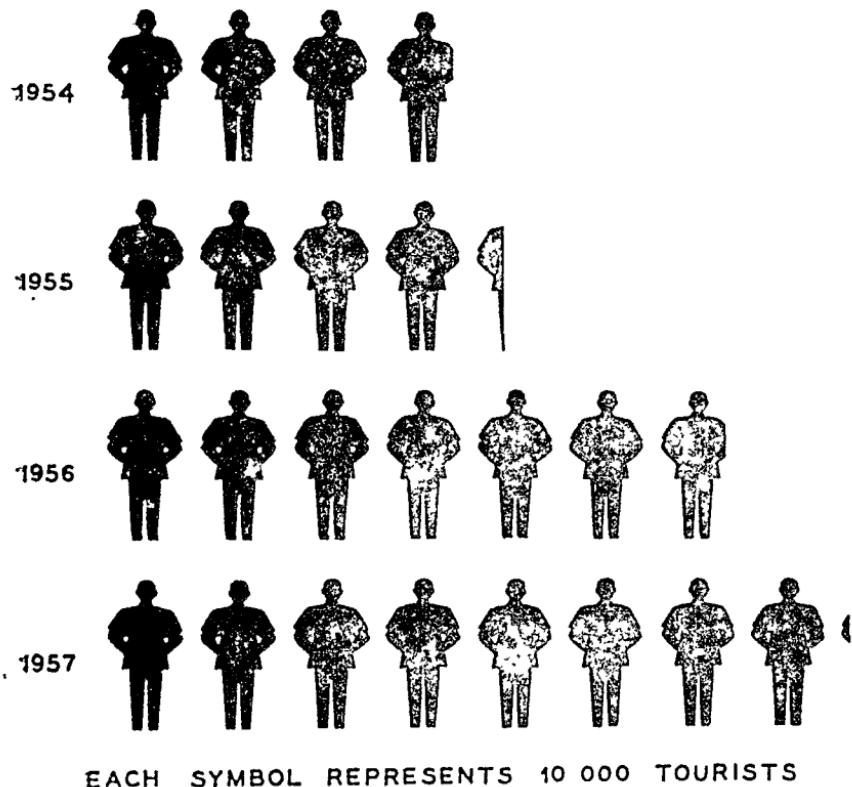


Fig. 4.8 Pictorial diagram showing number of tourists coming to India from abroad (Table 4.2).

taking a similar symbol, its size being proportional to the value, or by taking a number of symbols (including a fraction of a symbol) of the same size. The second method is preferable, because here an idea about the actual values is more readily obtained by looking at the diagram. Some of the data of Table 4.2, showing India's growing

TABLE 45
INVESTMENT ON DIFFERENT HEADS UNDER
THIRD FIVE YEAR PLAN

Item	Proposed investment (Rs Crores)
1 Agriculture, minor irrigation and community development	1,475
2 Major and medium irrigation	640
3 Power	975
4 Village and small industries	435
5 Industries and minerals	2,500
6 Transport and communication	1,650
7 Social services	1,725
8 Inventories	800
Total	10,200

Source *Third Five Year Plan—Draft Outline*, Planning Commission, Govt of India

TABLE 46
PERCENTAGES IN THE DIFFERENT CATEGORIES OF TABLE 45
AND EQUIVALENT ANGLES TO BE USED IN A PIE DIAGRAM

Item	Percentage investment	Angle to be used (degrees)
1 Agriculture, minor irrigation and community development	14.5	52.2
2 Major and medium irrigation	6.3	22.7
3 Power	9.5	34.2
4 Village and small industries	4.3	15.5
5 Industries and minerals	24.5	88.2
6 Transport and communication	16.2	58.3
7 Social services	16.9	60.8
8 Inventories	7.8	28.1
Total	100.0	360.0

attractions to tourists from abroad, are presented in this manner in Fig. 4.8.

Representation of percentages: When the values of a variable are given for a number of categories, we may be interested in the percentages for the different categories rather than in the absolute values,

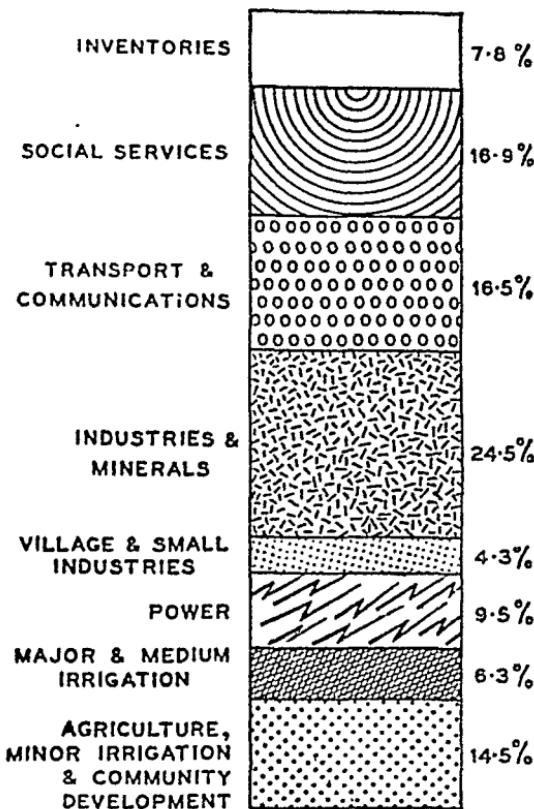


Fig. 4.9 Divided-bar diagram showing proposed investment on different heads under the Third Five Year Plan.

for the percentages are expected to give a better idea of the relative importance of each class. For this purpose one may draw either a *divided-bar diagram* or a *pie diagram*. In the first case, a bar of suitable length and width is taken, its total area being regarded as 100. If a vertical bar is chosen, then this area is divided into a number of sections by drawing lines parallel to the base, in such a way that the area of each section represents the percentage in the corresponding category. In the second case, a circle is used, the area enclosed by

it being taken as 100. It is then divided into a number of sectors by drawing angles at the centre, the area of each sector representing the corresponding percentage. Since the full angle at the centre is of 360° , it is clear that for any particular category the angle should be of $(360^\circ/100) \times$ corresponding percentage. The data on the total investment under the Draft Third Five Year Plan, given in Table 4.5, are represented by both these methods in Figs. 4.9 and 4.10.

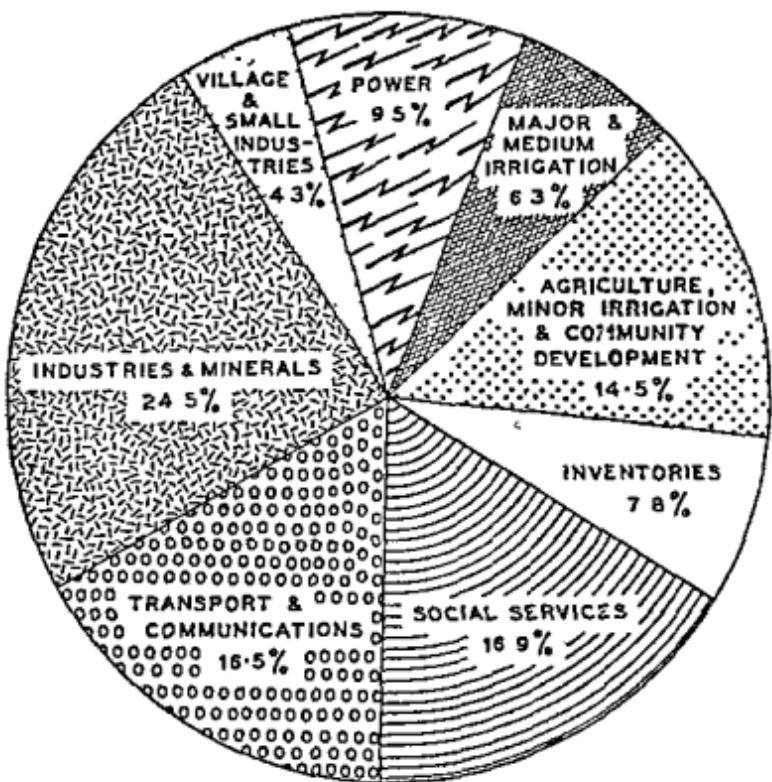


Fig. 4.10 Pie diagram showing proposed investment on different heads under the Third Five Year Plan

In order to draw a divided-bar diagram and a pie diagram, it is convenient to form beforehand a table of percentages (and for a pie diagram, also the corresponding angles to be drawn at the centre of the circle). These are shown in Table 4.6.

Statistical maps : Statistical information is sometimes presented in maps. This is appropriate when our concern is to show how a variable changes from one part of a region to another. Such a situation arises, for instance, when it is necessary to show diagrammatically

the variation in rainfall, in density of population or in yield-rate of a crop over the different parts of India. For this purpose, an outline map of the region is taken, and then the different portions of the map are shaded, a deeper shade indicating a larger value of the

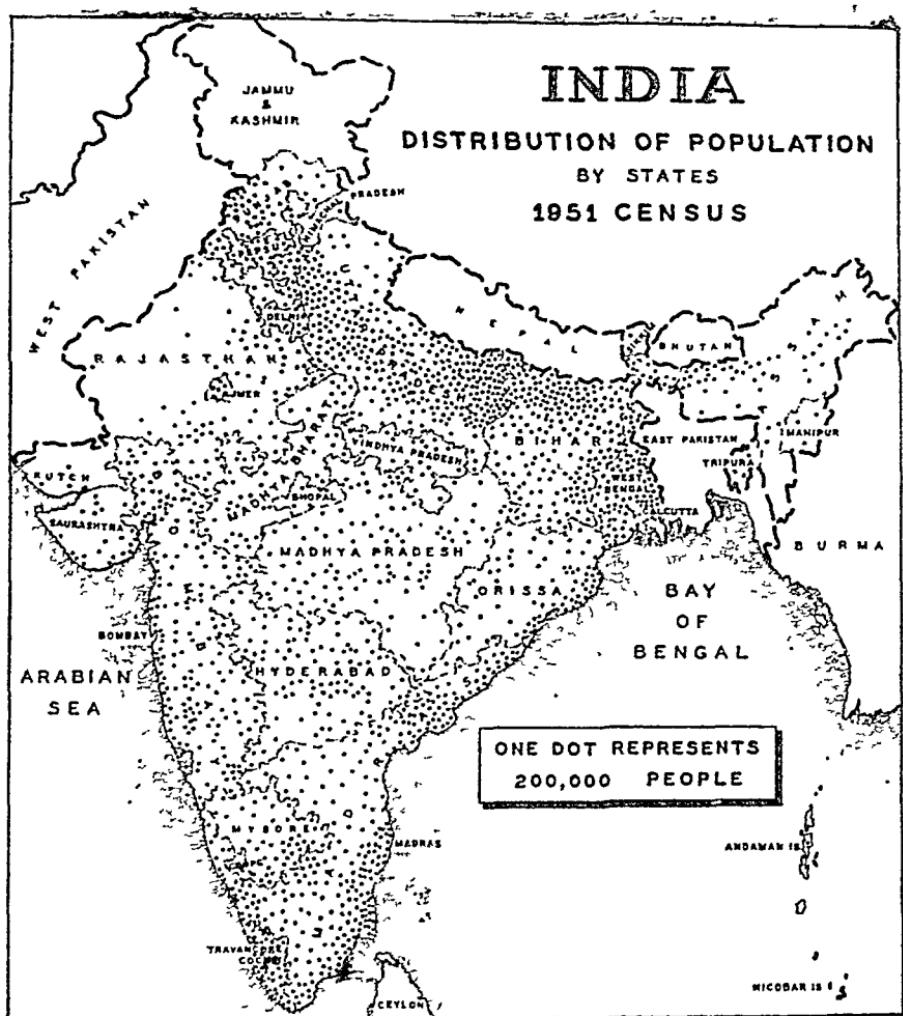


Fig. 4.11 A statistical map showing distribution of population by States in India, according to the 1951 census.

variable. Alternatively, points may be used to represent the value of the variable for each part of the region, the higher the value of the variable the greater being the density of points. A *statistical map* of this type is given in Fig. 4.11.

Questions and exercises

4.1 Write a note on the different senses in which the word 'statistics' may be used. Would it be correct to say that statistics is a science? Give reasons for your answer.

4.2 Comment on the general usefulness of statistical methods. What is the rôle of statistics in scientific research?

4.3 Give an account of the different modes of diagrammatic representation of statistical data.

4.4 Scrutinise the following data and state which of the figures, if any, you consider unreliable. Give reasons for your answer.

Census year	Population of country A (000)	Census year	Population of country A (000)
1900	212	1940	325
1910	235	1950	378
1920	274	1960	310
1930	510		

4.5 Examine the three series of figures given in the following table and state whether they may be regarded as mutually compatible.

Crop	Area under crop (000 acres)	Total yield (000 tons)	Yield-rate (lb. per acre)
Rice	74,424	24,209	728.63
Maize	9,325	2,944	707.79
Wheat	26,842	8,539	7125.9
Barley	7,999	2,786	780.17
Jowar	43,456	9,092	468.65
Bajra	27,350	3,555	491.16

1 ton = 2,240 lb.

4.6 Represent the information contained in the following passage in a suitable tabular form :

"The cropped area of vegetables (excluding potatoes) grown for human consumption in the United Kingdom rose in 1955-56 and was the highest since 1950-51. The cropped area increased to 509,000 acres, some 11,000 acres more than in 1954-55. The area of root vegetables increased by 8,100 acres to 62,400 acres, carrots alone increasing by 5,700 acres to 33,200 acres. The area of cabbage rose

slightly, thus halting the steady decline since 1947-48 ; the cropped area was 75,700 acres with 74,800 acres in 1954-55. The cropped area of cauliflower and broccoli was 33,400 acres, 2,400 acres less than in 1954-55. Peas harvested dry decreased by about 9,800 acres to 121,800 acres, but a larger area of beans, mainly broad beans and green peas, was grown. The area of broad beans increased by 2,600 acres to 7,300 acres and the area of green peas for canning and quick freezing rose by 7,000 acres to 50,400 acres.”

4.7 Use a suitable diagram to represent the following data relating to the Post and Telegraphs Department, Govt. of India (taken from *Statistical Abstract of the Indian Union, 1966* and *1968*).

Year	Net receipts (lakhs of rupees)	Year	Net receipts (lakhs of rupees)
1955-56	565.32	1961-62	497.56
1956-57	880.33	1962-63	567.12
1957-58	645.03	1963-64	963.80
1958-59	954.09	1964-65	871.33
1959-60	859.09	1965-66	516.17
1960-61	425.34	1966-67	936.09

4.8 The actual outlay on the public sector in the First and Third Five-Year Plans of India is shown below by heads of development ;

Head of development	First Plan Outlay Rs. Crores	Third Plan Outlay Rs. Crores
Agricultural and community development	290	1,096
Irrigation and power	583	1,927
Industries and mining	97	1,965
Transport and communications	518	2,113
Social services	412	1,422
Miscellaneous	60	85
Total	1,960	8,608

Draw suitable diagrams to show the relative importance attached to the various heads in each Plan. Hence make a comparison between the First and the Third Plans.

SUGGESTED READING

- [1] Croxton, F E and Cowden, D J *Applied General Statistics* (Chs 1—6) Prentice-Hall, 1964
- [2] Fisher, R A *Statistical Methods for Research Workers* (Chs 1—2). Oliver & Boyd, 1954
- [3] Jenkinson, B L *Bureau of the Census Manual of Tabular Presentation* U S Govt Printing Office, 1949
- [4] Mills, F C *Statistical Methods* (Chs 1—3) H Holt, 1955
- [5] Modley, R *How to Use Pictorial Statistics* Harper, 1937
- [6] Moroney, M J *Facts from Figures* Penguin, 1956
- [7] Myers, J W *Statistical Presentation* Littlefield, Adam & Co., 1956
- [8] Tippett, L H C *Statistics* Oxford University Press, 1943
- [9] Wallis, W A and Roberts, H V *Statistics a New Approach* (Chs 1—3, 5) Methuen, 1957

5

FREQUENCY DISTRIBUTIONS

5.1 Summarisation of data

In Chapter 4 we have discussed some elementary methods of dealing with numerical data in the detailed form in which they are collected. Often, however, the raw data will be so numerous that their significance will not be readily comprehended. In such cases it will become necessary to summarise the data to an easily manageable form. This kind of summarisation, of course, leads to some sacrifice of information. But this will not be a serious drawback unless we are interested in the individual (country, person or object) to which each figure refers. This is the case, for instance, when we are ultimately interested in information regarding such points as the minimum or the maximum height for a group of students, the percentage of students having height between 160 cm. and 166 cm., etc., and not in the height of each and every student of the group. And, as we have emphasized in Section 4.2, in statistics we are concerned only with such properties of aggregates.

For the present, it will be assumed that the order in which the data are obtained is not relevant to the problem under enquiry. Some cases where the order of the data is important will be discussed in some later chapters (in Volume 2).

5.2 Attribute and variable

It should be noted that although statistics always deals with numerical data, such data may arise in one of two ways. In some cases the data are numerical to start with, e.g. when we record the height for each of a group of men or the number of rooms in each house of a town. In other cases numbers arise only secondarily. When we record the sex of each newborn baby during a month or the language of each book in a library, the data are not numbers initially. We get numbers if, subsequently, we note the number of male babies and that of female babies, or the number of books written in English, in Hindi, in Bengali, and so forth.

We may, therefore, say that the first type of data arise if we are observing, for each individual of a group, a character which can be expressed in numbers. Such a character will be referred to as a quantitative character or a *variable* or a *ariate*. For the second type of data, the character observed (viz. the sex of a baby or the language in which a book is written) is not expressible in numerical terms. Such a character is, therefore, called a qualitative character or an *attribute*.

The distinction between a variable and an attribute should be clearly borne in mind, for they will generally require different methods of statistical treatment.

5.3 Frequency distribution of an attribute

In the course of a recent investigation conducted by the Indian Market Research Bureau, 1,674 inhabitants of Calcutta, Bombay and Madras were interviewed. Each was asked, among other questions, whether he/she knew about the Indian Airlines employees' agitation of 1973. On getting the data, the sponsors of the investigation put them into a systematic form. They just counted the number of those who knew about the agitation among the people interviewed and got the following table.

TABLE 5.1
RESULT OF SURVEY ON AIRLINES EMPLOYEES' AGITATION

State of knowledge	Number of people (frequencies)
Aware	619
Unaware	1,055
Total	1,674

The number 619 shows how many of the people interviewed were aware of the agitation. In statistical language, this is the *frequency* of the form 'aware' of the attribute 'state of knowledge', because it tells us how frequent this form was among the people interviewed. Similarly, the number 1,055 is the frequency of the form 'unaware'.

Perhaps a better picture is obtained if one uses, instead of the frequencies, the proportions or the *relative frequencies*, as they are called. These are shown in Table 5.2.

TABLE 5.2
PROPORTION OF PEOPLE AWARE OF
AIRLINES EMPLOYEES' AGITATION

State of knowledge	Relative frequency
Aware	0·370
Unaware	0·630
Total	1·000

Table 5.1 shows how the total frequency 1,674 is distributed over the two classes, 'aware' and 'unaware'. Such a table is, therefore, said to give a *frequency distribution*—in this case, the frequency distribution of the attribute 'state of knowledge'. Table 5.2 presents the same frequency distribution in a different form.

Tables 5.1 and 5.2 present a dichotomy, a classification of individuals into two classes. We may as well have frequency distributions of attributes with more than two classes. E.g., in the same survey, again, the people who knew of the agitation were asked whether they were sympathetic to the agitation or not. Their answers led to the following frequency distribution with three classes :

TABLE 5.3
ATTITUDE TOWARDS AIRLINES EMPLOYEES' AGITATION

Attitude	Number of people (frequency)
Sympathetic	161
Unsympathetic	291
Indifferent	167
Total	619

Data regarding attributes may be represented graphically on the basis of either the frequencies or the relative frequencies. Since the data shown in the form of frequencies, as in Table 5.1 or Table 5.3, are similar to those in Table 4.5, they may be represented by means of a bar diagram, preferably with horizontal bars. Similarly, the data given in the form of relative frequencies, as in Table 5.2, being comparable to those in Table 4.6, may be shown in a pie-diagram or a divided-bar diagram.

5.4 Discrete and continuous variables

When we pass on to the study of data regarding quantitative characters, it is immediately found that these may be of two principal types. In the first place, the character may take only *some isolated values*, like the number of letters in a word (word-length), number of petals in a flower, number of members in a family (family-size), and so forth. Alternatively, it may conceivably take *any value* within its range of variation. The height, weight or age of a man, the diameter of a bobbin, the temperature, rainfall or humidity in a region, etc., are variables of this type. Even in the second case the actual measurements will present a discreteness, e.g. when heights are given correct to the nearest inch or when weights are given correct to the nearest pound. But this discreteness, it should be noted, is completely artificial, being due to the limitations of the measuring instrument. Variables of the first type are called *discontinuous* or *discrete*, while those of the second type are called *continuous*.

The distinction between a discrete variable and a continuous variable is again to be borne in mind, because the statistical treatment of the one may differ in some cases from that of the other.

5.5 Frequency distribution of a variable

Going to market one winter morning, one of the authors bought, among other things, peas worth 25 P. Back home from the market, he found there were 198 pea-pods in his bag. He took each pod and counted the number of peas it contained. The figures thus obtained are given below.

TABLE 5.4
NUMBER OF PEAS IN EACH OF 198 PEA-PODS

4	3	5	3	5	2	4	5	2	4	4	4	5	3
5	3	6	3	2	2	3	4	3	2	3	3	4	3
4	6	4	3	3	3	1	3	2	4	3	3	3	3
3	2	4	5	3	4	3	2	4	3	3	2	2	6
1	3	5	2	4	4	3	3	5	4	2	3	3	3
7	6	4	4	3	3	2	3	4	4	3	3	2	3
6	3	4	2	4	4	3	3	2	2	3	5	3	4
4	2	3	2	3	4	5	3	4	5	2	5	3	3
4	3	5	5	6	4	5	4	3	5	4	3	3	3
5	5	4	4	4	3	3	6	4	4	4	1	4	4
3	2	2	4	3	2	3	5	3	4	3	2	6	3
5	4	4	3	2	2	5	3	3	4	3	2	2	3
3	3	3	4	3	5	4	3	4	5	2	3	3	3
5	3	3	4	3	2	2	3	4	4	1	5	5	3
2	2												

The significance of this mass of data cannot be easily comprehended. A need is immediately felt to summarise the data to a more comprehensible form. The first step in the process of summarisation is to count the number of pods with 1 pea, with 2 peas, with 3 peas, etc., and thus form a frequency distribution of the discrete variable 'number of peas per pod'. The labour involved in counting can be minimised if we adopt the following procedure :

On going through the whole set of figures, we find that the largest value is 7 and the smallest 1. We, therefore, form a table with seven classes for the seven values : 1, 2, , 7. Next, we take the given values of the variable one by one and for each value place a stroke (a tally mark) in the table against the appropriate class. To facilitate counting the tally marks are arranged in blocks of five, every fifth stroke being drawn across the preceding four. This is done in Table 5.5.

TABLE 55
TALLY MARKS FOR THE VALUES IN TABLE 54

NUMBER OF PEAS	TALLY MARKS
1	
2	tu tu tu tu tu tu tu
3	tu
4	tu tu tu tu tu tu tu tu
5	tu tu tu tu tu
6	tu tu
7	/

Finally, we count the number of tally marks in each class and get the frequency distribution of the variable, which is shown in the first two columns of Table 5.6

TABLE 56
FREQUENCY DISTRIBUTION OF NUMBER OF PEAS
PER POD FOR 198 POPS

Number of peas	Frequency	Relative frequency
1	4	0.0202
2	33	0.1667
3	76	0.3838
4	50	0.2525
5	26	0.1313
6	8	0.0404
7	1	0.0051
Total	198	1.0000

The same frequency distribution may be represented in a number of other ways. In the first place, one may give relative frequencies instead of frequencies, as in the case of an attribute. These will give

the proportion of pods with k peas, for different values of k . But suppose one asks : "What is the number of pods with k peas or less?" To answer such questions, we may form a table giving cumulative totals of the frequencies proceeding from the lowest class upwards, called *cumulative frequencies* of 'less-than type'. Similarly, the cumulative totals of the frequencies obtained by proceeding from the highest class of the table downwards are called cumulative frequencies of 'more-than type'. These cumulative frequencies give the number of pods with k peas or more, for different values of k .

TABLE 5.7
CUMULATIVE FREQUENCY TABLE FOR THE FREQUENCY
DISTRIBUTION OF NUMBER OF PEAS FOR 198 PODS

Number of peas	Cumulative frequency (less-than type)	Cumulative frequency (more-than type)
1	4	198
2	37	194
3	113	161
4	163	85
5	189	35
6	197	9
7	198	1

An alternative method of representing a frequency distribution is to give the *cumulative proportions*, which are the cumulative totals of relative frequencies. In the present case, they would give, for different values of k , the proportions of pods with k peas or less and with k peas or more.

When we are concerned with an attribute or a discrete variable, generally the nature of the data will itself dictate the mode of classification to be used. It would be natural to take one class for each form of an attribute or for each *different* value of a discrete variable. In the case of a continuous variable, on the other hand, if one takes a class for each different value of the variable, the number of classes may be unduly large, thus defeating the very purpose of classification,

viz. the summarisation of data. In fact, since a continuous variable can, by definition, assume an infinite number of possible values, the classification of such data is necessarily artificial. The statistician himself will have to decide upon the appropriate classification to be adopted in any given case. Some general observations can, however, be made in this connection. Let us take for illustration the data of Table 5 8

TABLE 5 8
STATURE (IN CM) OF 177 INDIAN ADULT MALES

169 0	164 5	154 2	163 0	171 6	157 5
166 7	166 7	161 8	163 1	167 0	160 5
159 9	165 0	156 5	157 6	169 5	170 3
157 8	159 7	161 7	160 9	163 7	167 0
169 9	158 9	145 6	152 5	162 5	171 3
158 4	168 9	162 2	167 9	166 8	162 2
171 7	163 9	162 0	164 2	160 2	169 4
160 4	162 0	165 5	167 5	163 9	170 0
167 5	165 2	167 2	164 2	171 0	166 6
161 0	167 4	159 8	171 7	156 4	160 5
168 8	172 8	169 0	167 7	170 0	160 6
167 8	168 2	165 3	168 0	161 4	168 0
164 0	166 8	158 0	168 1	160 5	155 9
167 4	163 5	169 5	164 4	173 2	162 0
167 8	159 3	169 2	165 2	174 2	161 4
165 2	163 1	161 5	163 5	161 0	172 2
163 5	168 9	166 7	176 4	161 4	156 0
170 4	166 3	162 6	160 9	165 4	163 2
159 0	164 5	171 3	164 2	160 0	172 0
158 1	162 1	166 7	161 0	156 9	152 6
157 6	182 0	168 0	158 4	164 9	167 1
159 2	158 5	160 8	171 4	167 3	161 3
167 7	183 5	168 0	159 5	159 5	170 1
170 2	163 5	156 0	162 7	165 2	158 7
169 0	170 1	169 0	160 5	160 8	167 0
157 5	167 7	172 5	171 7	170 1	178 4
161 5	157 4	171 1	163 7	166 4	165 5
165 8	164 9	168 2	162 3	168 1	
159 6	168 3	172 6	171 9	168 7	
160 3	164 0	169 3	169 7	165 4	

For one thing, the classes should be exhaustive, i.e. should be such that each of the given values is included in one of the classes.

Secondly, the classes should be mutually exclusive or non-overlapping. If the classes are mutually overlapping, e.g. 144·5—149·5, 149·5—154·5, etc., in the present case, difficulty will arise in classifying a value like 149·5.

Thirdly, the number of classes should not be too large; otherwise, the purpose of classification, viz. summarisation of data, will not be served. Moreover, by taking a large number of classes one will introduce an irregular pattern in the frequencies which may be completely absent in the actual distribution. (This applies to the case of an attribute or a discrete variable as well. If the number of different forms of an attribute or the number of different values of a discrete variable be too large, then each class should be constituted with a number of different forms or a number of different values.)

The number of classes should not be too small either, for this also may obscure the true nature of the distribution. Further, we shall see in later chapters that in computing various statistical measures, like mean, standard deviation, etc., from a frequency distribution, the assumption is made that each class-frequency corresponds to a single value, viz. the mid-point of the set of values defining a class. If the number of classes be too small, each class will be too wide, and this assumption will make the computed value of the measure extremely unreliable.

As a working rule, one should take ten to twenty classes, provided the total frequency is not small, say not less than 1,000. A still smaller number of classes may be taken if the total frequency be much smaller than 1,000.

Lastly, the classes should preferably be of equal width. Otherwise, the class-frequencies will not be comparable, and the computation of statistical measures will be laborious. The principle, however, cannot be rigidly followed, as will be apparent from the frequency distribution of Table 5.11. In forming Table 5.11, if we took classes of the same width, the number of classes would be exceedingly large (if the width were about 10 rupees, e.g.) or exceedingly small (if the width were about 250 rupees, e.g.). In the latter case certain important features of the data would be obscured.

It would not be apparent, for example, that more than 25% of the employees earn less than Rs 77.5 or that about 60% of the employees earn less than Rs 97.5

Bearing the above points in mind, we may take for classifying the data of Table 5.8 the 8 classes defined by the following values of the variable 144.6—149.5, 149.6—154.5, ..., 179.6—184.5. Here again it would be convenient to form a table like Table 5.5. As before, we go through the given values of the variable one by one and for each value put a stroke opposite the class to which it belongs (Table 5.9).

TABLE 5.9
TALLY MARKS FOR THE DATA OF TABLE 5.8

HEIGHT (cm)	TALLY MARKS
144.6—149.5	/
149.6—154.5	///
154.6—159.5	
159.6—164.5	
164.6—169.5	
169.6—174.5	
174.6—179.5	
179.6—184.5	

One more point is to be noted before we draw up the frequency distribution on the basis of Table 5.9. Consider the class 144.6—149.5. The values are recorded correct to $\frac{1}{10}$ of a cm. Hence 144.6 represents any value between 144.55 and 144.65. Similarly, 149.5 represents any value between 149.45 and 149.55. Thus the class 144.6—149.5 really stands for the *class interval* 144.55—149.55. Similar is the case for the other classes. 144.6 and 149.5 are called the lower and upper *class limits* for the first class, while 144.55 and 149.55 are the corresponding *class-boundaries*. One should state the

class-boundaries, rather than the class-limits, while drawing up the frequency distribution of a continuous variable. Table 5.10 shows the frequency distribution in terms of the absolute frequencies, relative frequencies and cumulative frequencies. It should be noted that here the cumulative frequencies of the less-than type correspond to the upper class-boundaries ; for instance, the third one, 28, is the number of persons with height 159.55 cm. or less. Similarly, the cumulative frequencies of the greater-than type correspond to the lower class-boundaries.

TABLE 5.10
FREQUENCY DISTRIBUTION OF HEIGHT FOR 177
INDIAN ADULT MALES

Height (cm.) class-interval	Frequency	Relative frequency	Cumulative frequency	
			'Less-than'	'Greater-than'
144.55—149.55	1	0.0057	1	177
149.55—154.55	3	0.0169	4	176
154.55—159.55	24	0.1356	28	173
159.55—164.55	58	0.3277	86	149
164.55—169.55	60	0.3390	146	91
169.55—174.55	27	0.1525	173	31
174.55—179.55	2	0.0113	175	4
179.55—184.55	2	0.0113	177	2
Total	177	1.0000	—	—

If the classes be of varying width, then the different class-frequencies will not be comparable. Comparable figures can be obtained if the frequencies are expressed per unit value of the variable by dividing them by the widths of the class-intervals. The ratios are called *frequency-densities*. Table 5.11 gives a frequency distribution where the classes are not equally wide. The frequency-densities appear in the third column of the table.

TABLE 5 11
FREQUENCY DISTRIBUTION OF INCOME FOR 1,870
EMPLOYEES OF A MANUFACTURING FIRM

Income (Rs) class-interval	Frequency	Frequency-density
42.5— 47.5	8	1 600
47.5— 52.5	12	2 400
52.5— 57.5	47	9 400
57.5— 62.5	61	12 200
62.5— 67.5	105	21 000
67.5— 72.5	338	33 600
72.5— 77.5	271	27 100
77.5— 82.5	260	26 000
82.5— 87.5	265	10 600
87.5— 122.5	117	4 680
122.5— 147.5	142	2 840
147.5— 197.5	105	2 100
197.5— 247.5	66	0 660
247.5— 347.5	57	0 570
347.5— 447.5	9	0 036
447.5— 697.5	5	0 010
697.5— 1,197.5	2	0 002
Total	1,870	—

5.6 Graphical representation of frequency distribution of a variable

Let us consider the frequency distribution of a discrete variable like that in Table 5 6 To represent such a distribution graphically, one may take two rectangular axes of co-ordinates—the horizontal for the variable, the vertical for frequency. The different values of the variable are then located as points on the horizontal axis.

Next, at each of these points a perpendicular is drawn to represent the corresponding frequency. Such a diagram may be called a *column diagram* or a *frequency bar diagram* (Fig. 5.1).

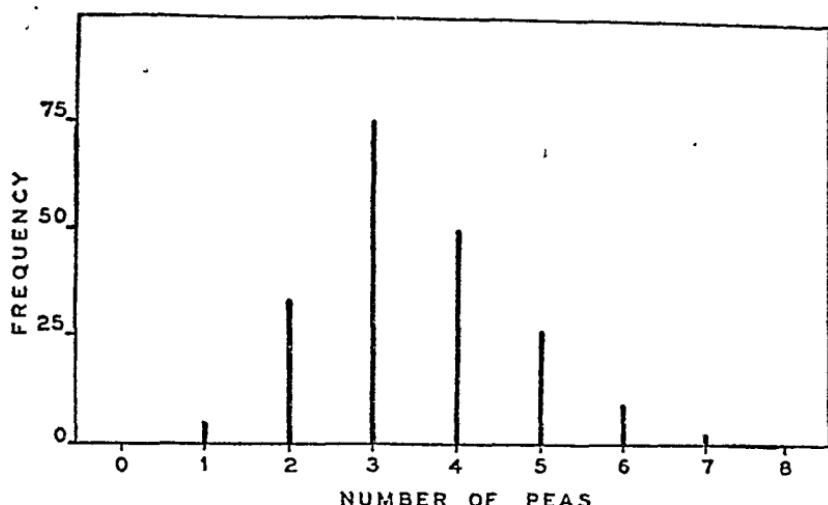


Fig. 5.1 Column diagram for the frequency distribution of number of peas for 198 pods (Table 5.6).

An alternative method of representing such a distribution is to use a *frequency polygon*, in which case the values and the corresponding frequencies are plotted as points with the help of rectangular co-ordinates, as in a frequency bar chart. For Table 5.6, let us first plot the points $(1, 4)$, $(2, 33)$, \dots , $(6, 8)$, $(7, 1)$ on graph paper. The value preceding 1, i.e. 0, has zero frequency ; so has the value following 7, i.e. 8. Hence let us take two more points, $(0, 0)$ and $(8, 0)$. Finally, we join the successive points by line segments to get a closed polygon (Fig. 5.2).

A frequency polygon may also be used to represent the frequency distribution of a continuous variable, provided the classes are of equal width. Here the frequencies are plotted against the mid-points of the corresponding class-intervals and joined successively by line segments, as in the case of a discrete variable.

But a better method for a continuous variable is to use a *histogram*. Here on the horizontal axis we locate the class-boundaries, and over each class-interval we erect a rectangle whose area represents the corresponding class-frequency. Obviously, the height of a rectangle is

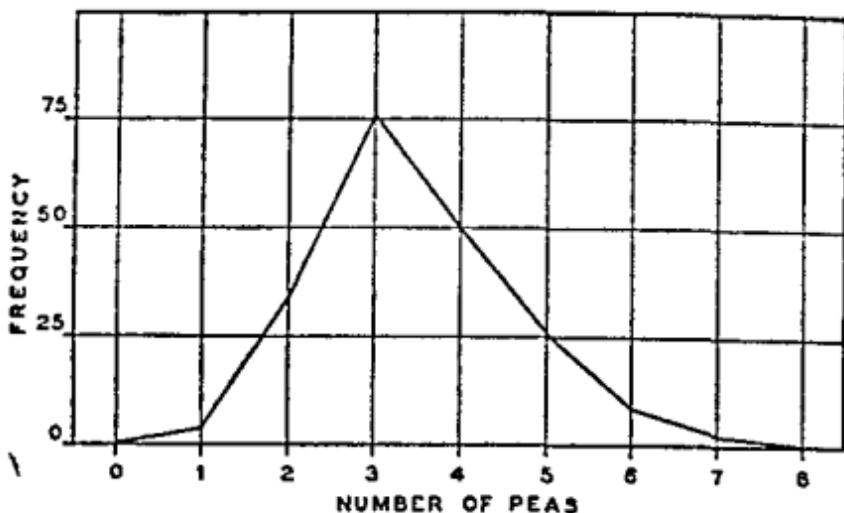


Fig. 5.2 Frequency polygon for the frequency distribution of number of peas for 198 pods (Table 5.6).

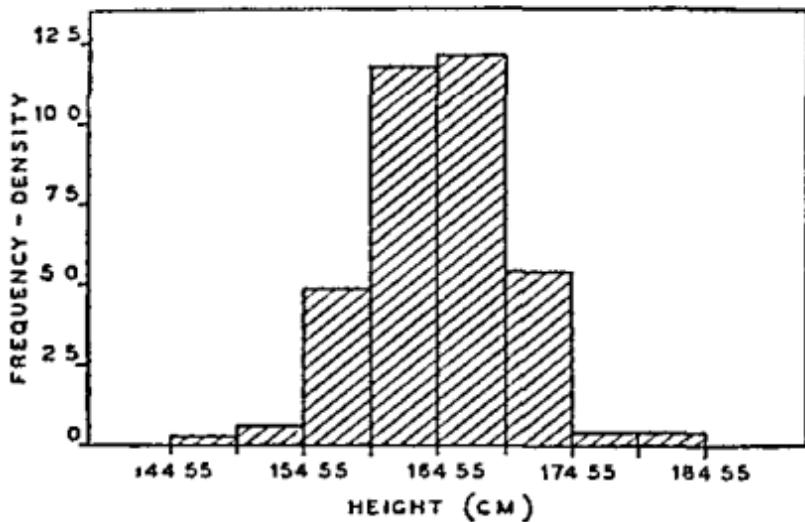


Fig. 5.3 Histogram representing the frequency distribution of height for 177 Indian adult males (Table 5.10).

to be taken equal to the corresponding frequency-density (Fig. 5.3). A histogram is sometimes used for a discrete variable as well, where each value is regarded as the mid-point of an interval. But its use here is not to be recommended, because in the discrete case each frequency corresponds to a single point and not to an interval.

However, the use of this method cannot be avoided in case each class includes more than one different value of the discrete variable.

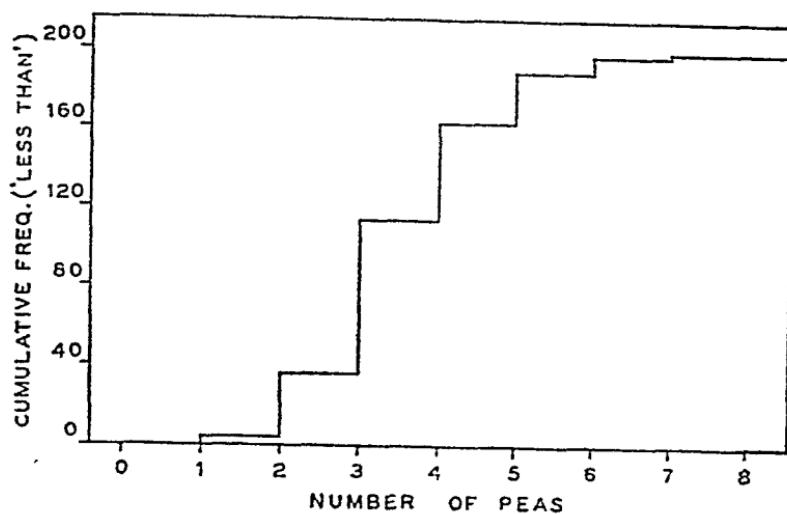


Fig. 5.4 Cumulative frequency diagram (less-than type) for the data on number of peas for 198 pods (Table 5.7).

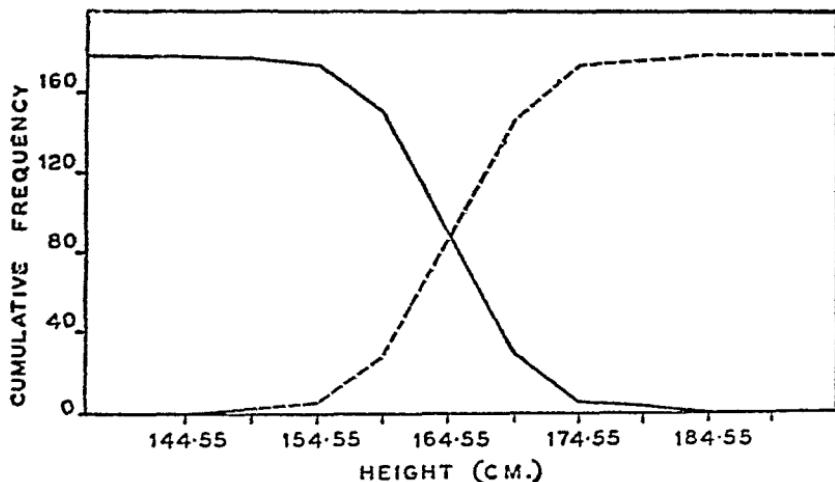


Fig. 5.5 Cumulative frequency diagrams for the frequency distribution of height for 177 Indian adult males (Table 5.10). (Continuous lines for 'greater-than' type, broken lines for 'less-than' type.)

A frequency distribution may be represented graphically on the basis of cumulative frequencies as well. To do this, the cumulative frequencies are plotted as points against the values to which they

correspond. By joining the points, a *cumulative frequency diagram* or *ogive* is obtained. Special attention should be given to the discrete case. In Table 5.7, the cumulative frequency is 0 for any value less than 1, is 4 for any value greater than or equal to 1 but less than 2, and so on. Hence the ogive will be a *step diagram*, as in Fig. 5.4. The ogives (of less than and greater than types) for the frequency distribution of Table 5.10 are shown in Fig. 5.5.

Questions and exercises

5.1 Explain with suitable examples the distinction (a) between an attribute and a variable and (b) between a discrete variable and a continuous variable.

5.2 Discuss the different considerations to be kept in view in drawing up a frequency distribution for data on a continuous variable.

5.3 Discuss the different modes of graphical representation of frequency distributions of different types.

5.4 The word length for each of 90 words in a poem by Tagore is shown below:

5	4	3	5	8	6	6	3	4
3	4	4	5	8	2	6	7	6
4	5	6	4	9	6	4	2	2
2	9	2	3	3	3	2	4	7
7	2	4	4	4	3	4	4	2
4	4	9	3	7	4	5	12	6
3	5	2	5	10	3	5	7	3
3	3	6	2	5	3	3	3	?
4	5	8	5	3	4	4	6	7
2	3	5	5	5	3	2	4	5

Construct a frequency table and also obtain the relative frequencies and the cumulative frequencies (of the 'less-than' type).

Represent the data in a column diagram, a frequency polygon and an ogive.

5.5 On the basis of the table constructed in Exercise 5.4, answer the following questions:

- What is the proportion of words with 9 letters?

(b) What is the number of words with 3 letters or less, and what is the number of words with 5 letters or more?

(c) What is the number of words with not less than 4 and not more than 6 letters? *Ans.* (a) 0.0333; (b) 32, 38; (c) 44.

5.6 With the data shown below, form a frequency distribution with six classes. Show the frequencies, the relative frequencies and the cumulative frequencies (of both the less-than and the greater-than types). Finally, represent the distribution by means of suitable diagrams.

Life (in hours) of 100 electric bulbs

511	991	1,177	1,016	600	777	895	749	1,067	980
923	1,314	1,108	1,137	906	1,230	1,099	1,242	803	1,131
918	1,240	1,057	980	992	763	759	1,394	1,111	1,117
1,143	808	948	857	962	922	817	1,057	665	1,171
936	1,068	750	873	1,139	1,127	1,163	934	515	907
1,061	1,198	1,027	1,081	991	1,155	1,199	806	950	1,262
848	1,293	956	1,140	885	1,330	1,166	1,333	1,146	933
820	880	982	912	1,100	1,293	1,192	1,371	1,023	1,298
1,059	1,092	1,091	1,182	699	803	1,069	922	1,245	706
1,053	1,001	939	1,248	850	985	1,219	945	1,012	846

SUGGESTED READING

- [1] Mills, F. C. *Statistical Methods* (Ch. 3). H. Holt, 1955.
- [2] Simpson, G. and Kafka, F. *Basic Statistics* (Chs. 8, 9). W. W. Norton, 1957, and Oxford & IBH, 1965.
- [3] Wallis, A. W. and Roberts, H. V. *Statistics : a New Approach* (Ch. 6). Methuen, 1957.
- [4] Yule, G. U. and Kendall, M. G. *An Introduction to the Theory of Statistics* (Ch. 4). Charles Griffin, 1953.

6

MEASURES OF
CENTRAL TENDENCY

6.1 Descriptive measures of statistics

It was noted in the previous chapter that the primary purpose of statistical methods is to summarise the information contained in the data. The purpose is served to some extent by classifying the data in the form of a frequency distribution and using various graphs. When the data relate to a variable, the process of summarisation can be taken a long step further by using certain descriptive measures. The aim is to determine certain features of the data which will describe their nature in a general way. The two most important features are *central tendency* and *dispersion*. Two other features which are also of some importance are *skewness* and *kurtosis*.

6.2 Central tendency

Quite often there will be found in the data a tendency, notwithstanding their variability, to cluster around a central value. This will be apparent from Table 6.1, where the figures seem to cluster around some point between 1,200 gm and 1,300 gm.

TABLE 6.1
YIELD PER PLANT FOR 12 TOMATO PLANTS OF A
PARTICULAR VARIETY

Plant No	Yield (gm.)	Plant No	Yield (gm.)
1	1,216	7	1,202
2	1,374	8	1,372
3	1,167	9	1,278
4	1,232	10	1,141
5	1,407	11	1,221
6	1,453	12	1,329

In such a case, it would be legitimate to use a single value, the central value, to represent the whole set of figures. Such a representative or typical value of a variable is called a measure of central tendency or an *average*.

The idea of average is a familiar one. One has this idea in mind, maybe rather vaguely, when one says that Germans live longer than Indians. By this one does not mean that the longevity of every German is higher than that of every Indian. All that is meant is that the longevity of a typical German is higher than the longevity of a typical Indian—in other words, the average longevity of Germans is higher than the average longevity of Indians. In connection with a frequency distribution, an average is also referred to as a measure of location, because it determines, as it were, the position of the distribution on the axis of the variable.

Three types of average are in general use, viz. *arithmetic mean*, *median* and *mode*.

6.3 Arithmetic mean

The arithmetic mean (or, simply, mean) of a variable is obtained by dividing the sum of its given values by their number. If the variable is denoted by x and if n values of x are given— x_1, x_2, \dots, x_n , then the arithmetic mean of x is

$$\bar{x} = \sum_{i=1}^n x_i/n. \quad \dots \quad (6.1)$$

Ex. 6.1 For the data of Table 6.1, the arithmetic mean is

$$(1,216 + 1,374 + \dots + 1,329)/12 = 15,392/12 = 1,282.67 \text{ gm.}$$

Ex. 6.2 Now consider Table 5.4. By using the above formula, the arithmetic mean of the number of peas per pod is found to be

$$(4+3+5+\dots+2+2)/198 = 683/198 = 3.45.$$

But the computation could be simplified if we took the help of Table 5.6. This table shows that in the sum the variate value 1 occurs 4 times, the value 2 occurs 33 times, and so on. Hence the sum of the given values would be

$$1 \times 4 + 2 \times 33 + \dots + 7 \times 1 = 683,$$

as before.

This shows that if the values of a discrete variable are arranged

in a frequency table, each class being formed by a single number, then formula (6 1) may be expressed in the alternative form

$$x = \sum_{i=1}^k x_i f_i / n, \quad (6 2)$$

where x_i is the value of x in the i th class, f_i is the corresponding frequency and $n = \sum_{i=1}^k f_i$.

When we have a frequency table with more than one variate value in a class formula (6 2) may still be used, x , now denoting the mid value of the i th class interval. But in this case (6 2) will give only an *approximate* value of the mean. The error will, however, be negligible provided the range of x is very large compared to the width of the class intervals.

It should be noted, however, that for a continuous variable formula (6 1) also will give only an approximate result because of the inevitable errors of observation contained in the data.

Formulae (6 1) and (6 2) with some simplifying modifications will be used for computing the mean and some other measures for a continuous variable in Ex 8 1 and Ex 8 2.

Some important properties of the arithmetic mean may be mentioned.

(a) By definition,

$$\sum_{i=1}^k x_i / n = x, \quad \text{or} \quad \sum_i x_i = nx$$

Subtracting x from each term on the left hand side and modifying the right hand side accordingly, we have

$$\sum_i (x_i - x) = nx - nx = 0 \quad (6 3)$$

This shows that the sum of the deviations of the given values of a variable from its mean is necessarily zero.

(b) Suppose the given values of x are all equal to a constant a . Then $\sum_{i=1}^k x_i = na$, and hence

$$x = a$$

Thus the mean of a variable whose given values are all equal must also be the same as their common value.

(c) Let $y = a + bx$. Corresponding to the value x_i of x , there will be the value $y_i = a + bx_i$ of y . Hence

$$\bar{y} = \sum_{i=1}^n y_i/n = \frac{\sum_{i=1}^n (a + bx_i)}{n} = \frac{na + b \sum_{i=1}^n x_i}{n} = a + b\bar{x}. \quad \dots \quad (6.4)$$

Thus, if $y = a + bx$, a linear function of x , then the arithmetic means of y and x are related in the same way as y and x themselves are.

(d) Let there be two sets of values of x , the number of values in the two sets being n_1 and n_2 and the means being \bar{x}_1 and \bar{x}_2 , respectively.

Let x_{1j} ($j=1, 2, \dots, n_1$)

denote the values in the first set and

x_{2j} ($j=1, 2, \dots, n_2$)

those in the second set. The sum of x values in the two sets taken together will then be equal to the sum of values in the first set plus the sum of values in the second set, i.e.

$$\sum_{j=1}^{n_1} x_{1j} + \sum_{j=1}^{n_2} x_{2j} = n_1 \bar{x}_1 + n_2 \bar{x}_2.$$

But this sum must be equal to $(n_1 + n_2)$ times the *grand mean* of x . Hence the grand mean is

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}.$$

Generally, if there be t sets of values of x , containing n_1, n_2, \dots, n_t values and having means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_t$, then the grand mean of x is

$$\bar{x} = \frac{\sum_{i=1}^t n_i \bar{x}_i}{\sum_{i=1}^t n_i}. \quad \dots \quad (6.5)$$

(e) Suppose the values of two variables, x and y , are given for each of n individuals. If a new variable is formed, viz.

$$z = ax + by,$$

then for the i th individual the value of the new variable is

$$z_i = ax_i + by_i.$$

Summing over all individuals, we have

$$\sum_{i=1}^n z_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n y_i,$$

so that

$$\bar{z} = a\bar{x} + b\bar{y}. \quad \dots \quad (6.6)$$

6.4 Median

If the given values of x are arranged in an increasing or decreasing order of magnitude, then the middle most value in this arrangement is called the median of x . (The median may alternatively be defined as a value of x such that half of the given values of x are smaller than or equal to it and half are greater than or equal to it.)

When the number of values, n , is odd, the middle-most value—i.e. the $\frac{(n+1)}{2}$ th value—in the arrangement will be the unique median of x .

On the other hand, when n is even, there will be no unique median. For any number between the $\frac{n}{2}$ th and $\left(\frac{n}{2}+1\right)$ th values of x in the arrangement, being regarded as middle most, is now to be taken as a median, according to the above definition. However, for the sake of definiteness, the arithmetic mean of the $\frac{n}{2}$ th and $\left(\frac{n}{2}+1\right)$ th values is here accepted as the median of x , by convention.

Ex 6.3 The yields (in gm) of barley from 7 plots, of size one sq. yd each were found to be

180, 191, 175, 111, 154, 141 and 176

To determine the median yield, these are first arranged in an increasing order of magnitude, i.e. as

111, 141, 154, 175, 176, 180, 191

The median is the 4th value in this arrangement, i.e. 175 gm.

Ex 6.4 The number of letters (word length) in each of six words taken from a dictionary is shown below

7, 9, 5, 10, 4 and 8

Arranged in increasing order, the values will be

4, 5, 7, 8, 9, 10

The number of values being six, any value between the 3rd and the 4th (i.e. between 7 and 8) may be considered a median of the variable. For the sake of definiteness, one may here take their arithmetic mean, 7.5, as the median.

In the above arrangement, if the 3rd and 4th values were both 7, then no difficulty would arise, because 7, being the middle most value, would then be the only median.

When the observations are grouped into a frequency distribution, the median can be obtained on the basis of the cumulative frequencies. For the cumulative frequency table itself provides an arrangement of the observations in an increasing or decreasing order of magnitude, according as the cumulative frequencies are of the 'less-than' or of the 'greater-than' type.

Ex. 6.5 Take the frequency distribution of number of peas per pod given in Table 5.6. The cumulative frequencies in Table 5.7 indicate that if the given values of the variable are arranged in an increasing order of magnitude, then the first four values will be all equal to 1, the 5th to the 37th will be equal to 2, the 38th to the 113th will be equal to 3, and so on. The total number of observations being 198, the median is any number between the 99th and 100th observations, which are both equal to 3. Hence the median for this distribution is 3.

The frequency distribution of a continuous variable needs special attention. The median here may be supposed to be the value for which the cumulative frequency is $n/2$, the reason for which will be apparent from the second form of the definition of median. By going through the cumulative frequency table, we can then determine in which interval the median lies. Suppose the cumulative frequencies are of the 'less-than' type. The following formula will then give an approximate value of the median.

Let us denote the lower and upper class-boundaries of the class containing the median by x_l and x_u and the corresponding cumulative frequencies by n_l and n_u , respectively. If we assume that cumulative frequency changes from n_l to n_u between x_l and x_u at a constant rate, i.e. if we assume that cumulative frequency is a linear function of x between x_l and x_u , then the median, which is the value with cumulative frequency $n/2$, will satisfy the relation

$$\frac{Mi - x_l}{x_u - x_l} = \frac{n/2 - n_l}{n_u - n_l}.$$

This gives

$$Mi = x_l + \frac{n/2 - n_l}{f_0} \times c, \quad \dots \quad (6.7)$$

where c and f_0 are the width and the frequency of the class-interval containing the median, Mi .

The same value may also be obtained geometrically, from the ogive of the frequency distribution. The median will be given by the abscissa of the point on the ogive for which the ordinate is $n/2$.

Ex 6.6 Let us consider the frequency distribution of Table 5.10. Here the total frequency is $n=177$, so that $n/2=88.5$.

Hence, on going through the cumulative frequencies of the 'less than' type, it is found that the median lies between $x_1=164.55$ cm and $x_2=169.55$ cm, for which the cumulative frequencies are $n_1=86$ and $n_2=146$. Here $c=5$ and $f_0=60$. Therefore, formula (6.7) gives

$$M_1 = 164.55 + \frac{88.5 - 86}{60} \times 5$$

$$= 164.55 + 0.208 = 164.758 \text{ cm}$$

Alternative method. The median for the same distribution is obtained graphically in Fig. 6.1. The ogive of the distribution is first drawn. Then through the point $n/2=88.5$ on the vertical axis a line parallel to the x -axis is taken, which intersects the ogive at P . From P a perpendicular is let fall on the x -axis. The point at which it meets the x -axis is the median, M_1 .

From Fig. 6.1 M_1 is found to be about 164.76 cm, as before.

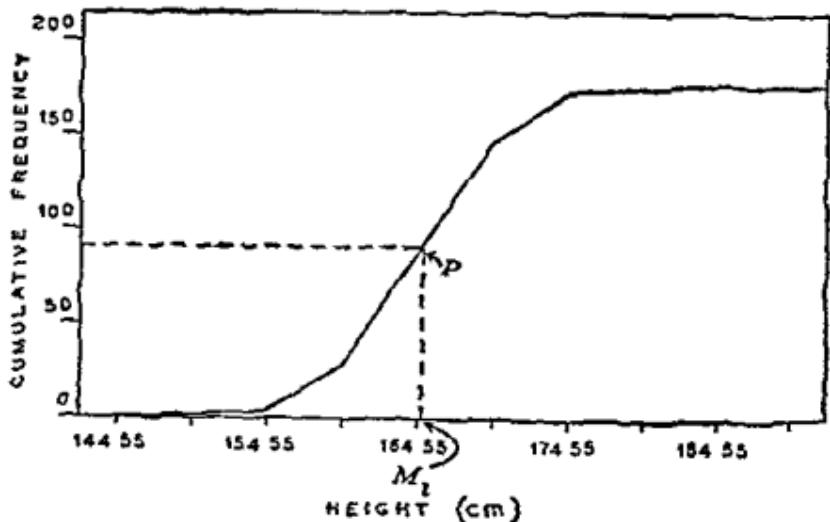


Fig. 6.1 Ogive showing the location of the median for the height distribution of Indian adult males (Table 5.10)

6.5 Mode

The mode of a variable is the value of the variable having the highest frequency*. This definition, properly speaking, applies to a discrete variable only.

Ex. 6.7 Consider the data of Table 5.4 again. It is found from Table 5.6 that the value 3 is the one with the highest frequency. Hence the mode of the number of peas per pod is 3.

For a continuous variable, the above definition needs to be modified. The mode here is the value of the variable with the highest frequency-density* corresponding to the ideal distribution which would be obtained if the total frequency were increased indefinitely and if, at the same time, the width of the class-intervals were decreased indefinitely. Graphically, it may be looked upon as the abscissa corresponding to the highest ordinate in the *frequency curve* (which is the limiting form of a histogram or a frequency polygon) of the ideal distribution. In Fig. 6.2 we have a frequency curve, where the mode is denoted by M_0 .

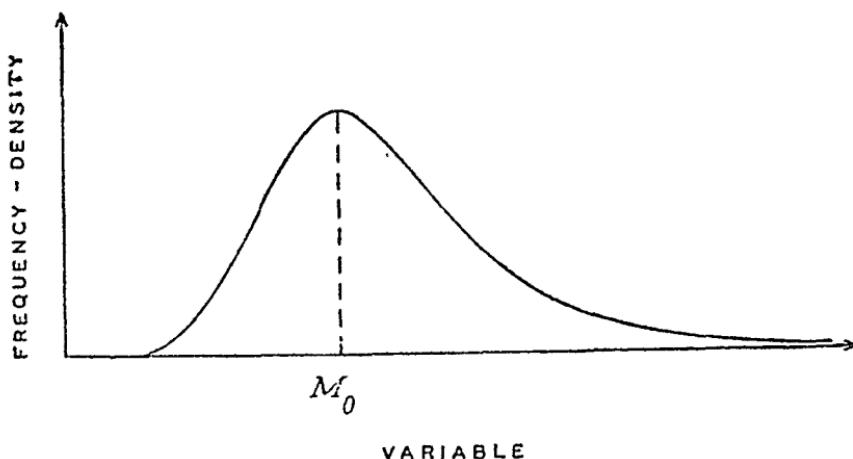


Fig. 6.2 A frequency curve.

For a frequency distribution obtained from a finite number of observations, like that in Table 5.10, the mid-value of the class-interval having the highest frequency may be approximately taken to be the mode. If this class-interval extends from x_l to x_u , then the

* If there are more than one value each with the highest frequency or frequency-density (as the case may be), then the mode is not defined.

mode is approximately given by

$$Mo = \frac{x_l + x_u}{2} = x_l + c/2,$$

c being the width of the interval.

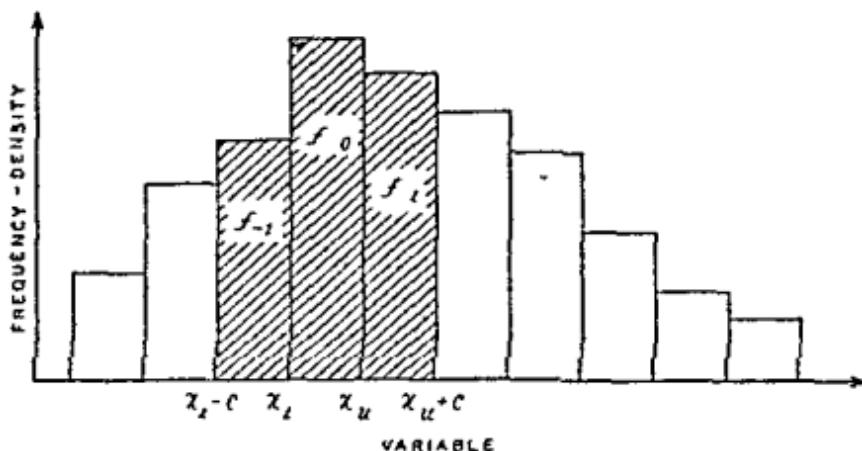


Fig. 6.3 Histogram of an observed distribution. The hatched portion shows the frequencies in the modal class and the two adjacent classes

This formula may generally be improved upon by considering, together with the class having the highest frequency, the class immediately preceding and the class immediately following it. Supposing these classes are of equal width c and have frequencies f_0 , f_{l-1} and f_l (vide Fig. 6.3), one would take $\frac{x_l + x_u}{2} = x_l + c/2$ as the mode, if $f_0 - f_{l-1}$ and $f_0 - f_l$ were equal in magnitude. If, on the other hand, $f_0 - f_{l-1}$ be smaller (greater) than $f_0 - f_l$, one would suppose that the mode is nearer (farther from) x_l than x_u .

Mathematically, this amounts to supposing that

$$\frac{Mo - x_l}{x_u - Mo} = \frac{f_0 - f_{l-1}}{f_0 - f_l}.$$

This leads to

$$Mo = x_l + \frac{f_0 - f_{l-1}}{2f_0 - f_{l-1} - f_l} \times c. \quad \dots \quad (6.8)$$

Another method of determining the mode is to make use of the empirical relation

$$\bar{x} - Mo = 3(\bar{x} - Mi), \quad \dots \quad (6.9)$$

which is found to be approximately valid for moderately skew distributions (*vide* Chapter 8). Relation (6.9) states that the amount by which the mean exceeds (is smaller than) the mode is three times the amount by which the mean exceeds (is smaller than) the median. Given the mean and the median, an approximate value of the mode may thus be obtained, viz.

$$3Mi - 2\bar{x}.$$

Ex. 6.8 Consider the height distribution of Indian adult males (Table 5.10). The class-interval with the highest frequency has boundaries $x_l = 164.55$ cm. and $x_u = 169.55$ cm. The frequencies in the two adjoining classes, which are also of the same width $c = 5$ cm., are $f_{-1} = 58$ and $f_1 = 27$. As to the modal class, it has frequency $f_0 = 60$. According to formula (6.8), the mode is approximately

$$\begin{aligned} Mo &= 164.55 + \frac{60 - 58}{2 \times 60 - 58 - 27} \times 5 \\ &= 164.55 + 0.286 = 164.836 \text{ cm.} \end{aligned}$$

Alternative method : For this distribution the mean is found to be $\bar{x} = 164.734$ cm. (*vide* Ex. 8.2) and the median is found to be $Mi = 164.758$ cm. (*vide* Ex. 6.6). Relation (6.9), therefore, gives

$$Mo = 3 \times 164.758 - 2 \times 164.734 = 164.806 \text{ cm.}$$

approximately.

6.6 Comparison of mean, median and mode

The mean is rigidly defined. So is the median except when there are an even number of observations, and so is the mode except when there are more than one value with the highest frequency or frequency-density.

The actual computation of all three measures involves almost the same amount of labour. But the determination of mode in the continuous case is impossible if only a few values of the variable are given. Even when a large number of observations grouped into a frequency distribution are available, the mode is difficult to determine. The method we have outlined above is not wholly satisfactory.

The general nature of the mean, like that of the median or mode, is easily comprehensible, the mean being the value that would be possessed by each of the given individuals if the total value ($\sum x_i$) were distributed equally among them.

Although in determining each of the three measures all observations have to be taken into account, in the actual computation only the mean *directly* uses all observations—so much so that its value changes when any one of the observations is changed. But this is not the case with the median or the mode. Some observations may be altered and yet the median or the mode may remain the same.

Moreover, the mean, as we have seen, has certain properties which enable it to be used readily in theoretical work. But the median and mode do not possess any such desirable property.

Besides, of the three measures the mean is generally the one that is the least affected by *sampling fluctuations*, although in some particular situations the median or the mode may be superior in this respect. (The term ‘sampling fluctuations’ will be explained in Section 14.2.)

It is, therefore, obvious that, by and large, the mean may be regarded as the best measure of central tendency.

In some special cases, however, the use of the mean is not to be recommended. Two such cases are considered below.

Suppose the values of the variable are given in the form of a frequency table of which one or both of the terminal classes are open (the table given in *Exercise 6.13* presents a case in point). Here the computation of the mean is impossible, because the class-marks of those terminal classes are indeterminate. But this will generally be no bar to the computation of the median or mode.

Again, let the weights of 8 persons be 138, 143, 141, 139, 152, 148, 160 and 267 lb. Here the mean is 161 lb., but this cannot be said to be a representative value, because seven out of the eight given values are smaller than 161. In cases of this sort, where the data contain a few extreme values widely different from the majority of the values, the mean should not be used. In the present example, the median would be the appropriate average.

In such a case the mean will also be subject to higher sampling fluctuations than the median. The following simple, though rather artificial, example illustrates this point. Let us consider drawing samples of size 3 from a set of 4 values

The different possible samples and the corresponding means and medians are shown below :

<i>Serial No.</i>	<i>Sample values</i>	<i>Mean</i>	<i>Median</i>
1	15, 30, 33	26	30
2	15, 30, 36	27	30
3	15, 33, 36	28	33
4	30, 33, 36	33	33

Clearly, the column of sample means shows greater variability than the column of medians.

The median has the additional property that the individual having the median value remains the same under any transformation that leaves the order of the values unchanged (*vide Exercise 6.9*). Hence whenever the order of the values is considered to be of importance, the median will be preferred to both the mean and the mode.

6.7 Other measures of central tendency

Besides the arithmetic mean, median and mode, there are two other averages which are relatively unimportant but may be appropriate to particular situations. These are the *geometric mean* and the *harmonic mean*.

If a variable x has n given values, x_1, x_2, \dots, x_n , then its geometric mean x_g is defined by

$$x_g = \left(\prod_i x_i \right)^{1/n}. \quad \dots \quad (6.10)$$

It is immediately seen that

$$\log x_g = \frac{1}{n} \sum_i \log x_i. \quad \dots \quad (6.11)$$

Thus the logarithm of the geometric mean of a variable is the arithmetic mean of its logarithm.

Consider two variables, x and y , whose values are given for each of n individuals. Let x_i and y_i ($i=1, 2, \dots, n$) be the values of x and y for the i th individual. Then

$$\left(\prod_i (x_i/y_i) \right)^{1/n} = \left(\prod_i x_i \right)^{1/n} / \left(\prod_i y_i \right)^{1/n};$$

that is, the geometric mean of the ratio of x and y is the ratio of their geometric means. Owing to this property of the geometric

mean, it is sometimes preferred for averaging ratios of two variables, for instance, in constructing price index numbers (*vide* Chapter 24 in Volume 2) Some consider the geometric mean to be the natural form of average for averaging price relatives since it gives equal emphasis to the same ratio of increase and of decrease in price Suppose the price-relative for one commodity is 5 (that is, its price has increased 5 times) and for another commodity it is $\frac{1}{5}$ (that is, its price now is $\frac{1}{5}$ of the original price) If equal importance is to be given to these two ratios, then their average should be 1 This criterion is satisfied by the geometric mean only

Ex 6.9 The ratios of the prices in 1964 to those in 1952 for four commodities are 0.92, 1.25, 1.75 and 0.85 To get the average price-ratio using the geometric mean, we see that

$$\begin{aligned}\log x_g &= (\log 0.92 + \log 1.25 + \log 1.75 + \log 0.85)/4 \\ &= (1.96379 + 0.09691 + 0.24304 + 1.92942)/4 \\ &= 0.05829 = \log 1.1436\end{aligned}$$

Thus

$$x_g = 1.144$$

The geometric mean also comes in if one wants to determine the value of a variable at the mid-point of a time interval when the variable changes over time exponentially Thus, if the values at two points of time, say 0 and t , be a and ar^t , then its value at the middle of the interval, i.e. at $t/2$, is

$$ar^{t/2} = \sqrt{a \cdot ar^t},$$

which is the geometric mean for the values at the terminal points of the interval

Formula (6.10) shows that the geometric mean is zero even if a single value of the variable happens to be zero, whatever the other values may be, and that it may be imaginary if some negative values are given This difficulty and the fact that it is of too abstract a nature are the reasons why it is not of common use in statistical work

The *harmonic mean* (x_h) of a variable x , with given values x_i , ($i=1, 2, \dots, n$), is defined by

$$x_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad (6.12)$$

or
$$\frac{1}{x_h} = \frac{1}{n} \sum_i \frac{1}{x_i}. \quad \dots \quad (6.13)$$

The second formula shows that the reciprocal of the harmonic mean of a variable is the arithmetic mean of its reciprocal.

Sometimes the variable may be in the form ' x per unit y ', e.g. miles per hour, rupees per maund, lb. per cubic foot, etc. In such cases, the harmonic mean would be the proper average if equal units of x are considered, while the arithmetic mean would be appropriate if equal units of y are considered. This may be illustrated with the following example :

Suppose a train moves n equal distances, each of s units, say, with speeds v_1, v_2, \dots, v_n miles per hour. The average speed is, of course, the total distance covered divided by the total time taken. Thus the average speed is

$$\frac{\frac{ns}{v_1} + \frac{ns}{v_2} + \dots + \frac{ns}{v_n}}{n} = \frac{1}{\frac{1}{v_1} + \frac{1}{v_2} + \dots + \frac{1}{v_n}},$$

the harmonic mean of the given speeds.

If, however, the train moves for n equal time intervals, each of length t hours, say, with the above speeds, then the average speed will be

$$\frac{v_1 t + v_2 t + \dots + v_n t}{nt} = \frac{v_1 + v_2 + \dots + v_n}{n},$$

the arithmetic mean of the given speeds.

Ex. 6.10 Suppose milk is sold at the rates of 0.80, 1.00, 1.20 and 1.50 rupees per litre in four different months. Assuming that equal amounts of money are spent on milk by a family in the four months, the average price in rupees per litre will be the harmonic mean of the given figures, viz.

$$x_h = \frac{4}{\frac{1}{0.80} + \frac{1}{1.00} + \frac{1}{1.20} + \frac{1}{1.50}} = \frac{4}{3.75} = 1.07.$$

In some cases, instead of simple means—simple arithmetic, geometric or harmonic means, *weighted means* are used. The values of the variable may not have the same importance. Hence one considers, along with the values x_1, x_2, \dots, x_n , a set of weights $w_1,$

w_1, \dots, w_n , where w_i indicates the importance of the value x_i in the given context. The weighted arithmetic mean is then

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}. \quad \dots \quad (6.14)$$

Obviously, formula (6.2) also gives in a way a mean of this type. It is the weighted arithmetic mean of the class-marks of τ with weights equal to the corresponding class-frequencies. Another frequently used weighted arithmetic mean is the cost of living index in its usual form (*vide* Section 24.4 in Volume 2).

In the same way, the weighted geometric mean of x for the above sets of values and weights will be

$$\left(\prod x_i^{w_i} \right)^{\frac{1}{\sum w_i}}, \quad (6.15)$$

and the weighted harmonic mean will be

$$\frac{\sum w_i}{\frac{\sum w_i}{\sum \frac{w_i}{x_i}}}. \quad (6.16)$$

Questions and exercises

6.1 What do you mean by the central tendency of a frequency distribution? Describe the common measures of central tendency.

6.2 What are the desiderata of a good measure of central tendency? Compare the mean, the median and the mode in the light of these desiderata.

6.3 How, in your opinion, should an average change when all values of the variable are increased or decreased?

- (1) by the same amount?
- (2) in the same proportion?

Judge in this light the different averages considered in this chapter.

6.4 State and prove the important properties of the arithmetic mean.

6.5 Give some examples where the geometric mean or the harmonic mean would be the appropriate type of average.

6.6 Show that the median of a variable is the abscissa of the point of intersection of its two ogives (of the 'less-than' and 'greater-than' types).

6.7(a) There are two sets of values of x . The first set with n_1 values has median M_1 and the second with n_2 values has median M_2 . Show that the median of all n_1+n_2 values taken together must lie between M_1 and M_2 .

(b) Show that if \bar{x}_1 and \bar{x}_2 are the means of the two sets, then the mean \bar{x} of the combined set also lies between \bar{x}_1 and \bar{x}_2 .

6.8 Let x be a variable assuming the values 1, 2, ..., k and let $F'_1=n, F'_2, \dots, F'_k$ be the corresponding cumulative frequencies of the 'greater-than' type. Show that

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k F'_i.$$

6.9(a) Suppose x is a variable (discrete or continuous) with median M_i . If $y=g(x)$ be a monotonically increasing or decreasing function of x , show that the median of y is $g(M_i)$.

(b) Can a similar statement be made with regard to the mean or the mode?

6.10 If $y=a+bx$ and M_o is the mode of x , then show that the mode of y must be $a+bM_o$.

6.11 Suppose the average of a number of temperature readings is to be determined. Show that it does not matter whether the readings are in the centigrade or in the Fahrenheit scale if one uses the arithmetic mean, but that it does matter if one uses the geometric or the harmonic mean.

6.12 Let x be a variable assuming positive values only. Show that (a) the arithmetic mean of the reciprocal of x cannot be smaller than the reciprocal of its arithmetic mean, and that (b) the arithmetic mean of the square-root of x cannot be greater than the square-root of its arithmetic mean.

6.13 The mean weight per student in a group of 6 students is 119 lb. The individual weights of 5 of them are 115 lb, 109 lb, 129 lb, 117 lb and 114 lb. What is the weight of the other student of the group ? *Ans* 130 lb

6.14 A factory has 5 sections employing 105, 184, 130, 93 and 124 workers. The mean earnings in a certain week per worker are Rs 33.84, 35.12, 35.27, 38.19 and 34.22 for the 5 sections. Determine the mean earnings per worker for the whole factory.

Ans Rs 35.21

6.15 The mean monthly income of a gentleman is Rs 819/- and his mean monthly expenditure comes out to be Rs 793/- What are his mean monthly savings ? *Ans* Rs 26/-

6.16 The following data give the length of ear-head (in cm) for 24 ears of a variety of wheat. Compute the mean and the median.

11.5	8.8	10.1
8.2	9.3	10.0
9.7	10.1	10.3
10.3	11.3	9.8
10.7	9.8	9.3
8.6	10.4	9.8
11.3	8.4	9.0
10.7	9.6	11.2

Ans 9.9 cm, 9.9 cm

6.17 For a certain frequency table with total frequency 150, the mean was found to be Rs 56.47. But while copying out the table, a typist left out two of the class frequencies, say f^* and f^{**} , so that the table is given to you in the following form.

Weekly wages in Rs (mid value)	45	50	55	60	65	70	75	Total
Frequency	5	48	f^*	30	f^{**}	8	6	150

Determine f^* and f^{**}

Ans $f^* = 41, f^{**} = 12$

6.18 The numbers of telephone calls received at an exchange in 245 successive one-minute intervals are shown in the following frequency distribution :

Number of calls	Frequency
0	14
1	21
2	25
3	43
4	51
5	40
6	39
7	12
Total	245

Evaluate the mean, median and mode.

Ans. 3.76 ; 4 ; 4.

6.19 Compute the mean, median and mode for the following frequency distribution :

Frequency distribution of I.Q. for 309 six-year-old children

I.Q.	Frequency
160—169	2
150—159	3
140—149	7
130—139	19
120—129	37
110—119	79
100—109	69
90— 99	65
80— 89	17
70— 79	5
60— 69	3
50— 59	2
40— 49	1
Total	309

Ans. 108.48 ; 108.41 ; 111.42, according to formula (6.8),
and 108.28, according to formula (6.9).

6.20 Determine the median and mode for the following distribution of monthly income for 580 middle class people

Monthly income (Rs.)	Frequency
—100	53
100—150	81
150—200	114
200—250	195
250—300	63
300—350	32
350—400	20
400—450	11
450—500	8
500—	3
Total	580

Ans Rs 210.77, Rs 219.01

6.21 The age-distribution of 4,488 Bengali males is given below

Age last birthday	Frequency
0	156
1	121
2	111
3	106
4	103
5—9	472
10—14	434
15—19	407
20—24	383
25—29	357
30—34	335
35—39	306
40—49	522
50—59	370
60—69	213
70—79	80
80—89	11
90—99	1
Total	4 488

Compute the mean age of Bengali males by means of formula (6.5)

Ans 27.40 years

SUGGESTED READING

- [1] Mills, F. C. *Statistical Methods* (Ch. 4). H. Holt, 1955.
- [2] Simpson, G. and Kafka, F. *Basic Statistics* (Chs. 10—12). W. W. Norton, 1957, and Oxford & IBH, 1965.
- [3] Wallis, W. A. and Roberts, H. V. *Statistics: a New Approach* (Ch. 7). Methuen, 1957.
- [4] Yule, G. U. and Kendall, M. G. *Introduction to the Theory of Statistics* (Ch. 5). Charles Griffin, 1953.

7.1 Meaning of dispersion

The average of a variable gives a general idea as to the whole set of its values. It is clear, however, that for a variable to be really *variable*, its given values will not be all equal to the average. In some cases they may lie very near the average, while in others they may be widely scattered about it. An example will make the point clearer.

Suppose two students, *A* and *B*, in a college received in eight monthly examinations the following marks in a particular subject :

Marks obtained by <i>A</i>	Marks obtained by <i>B</i>
63	61
47	54
56	56
44	57
66	60
65	59
80	55
43	62

If the arithmetic mean is taken to be the proper average to be used in this case, we find that the average score of each student was the same, viz. 58. Yet the overall nature of the scores of the two students was not at all the same. Thus *A* received as high a score as 80 and as low a score as 43. On the other hand, *B*'s score remained near about 58 throughout. In short, *B* gave a more consistent performance than *A*.

Thus, in order to give a proper idea about the overall nature of the given values of a variable, it is necessary, besides mentioning the average value, to state how scattered the given values are about the average. Mainly three different measures are used to determine this feature of a variable, which is called its *scatter* or *dispersion*. (It may be said that while the central tendency of a variable is the tendency

of its values to be *similar*, its dispersion represents the tendency of the values to be *different*.) These measures are (1) the *range*, (2) the *mean deviation* and (3) the *standard deviation*.

7.2 Range

The simplest measure of the dispersion of a variable is its range, which is defined as the difference between its highest and lowest given values. In the above example, the range of marks obtained is $80 - 43 = 37$ for *A* and $62 - 54 = 8$ for *B*.

7.3 Mean deviation

If *A* be the chosen average value of the variable *x*, then $x_i - A$ is the deviation of the *i*th given value of *x* from the average. Clearly, the higher the deviations

$$x_1 - A, x_2 - A, \dots, x_n - A$$

in magnitude, the higher is the dispersion of *x*. One may, therefore, consider some way of combining the deviations to get a measure of dispersion. It is readily seen that the simple arithmetic mean of the deviations, viz. $\frac{1}{n} \sum (x_i - A)$, cannot serve this purpose. For the sum of the deviations—and proportionately the arithmetic mean—may be quite small even when the individual deviations are large, positive and negative deviations almost cancelling each other. In fact, if *A* be the arithmetic mean of *x*, this sum vanishes, whatever the deviations are individually. This difficulty may be overcome by considering, instead of the deviations themselves, their absolute values, in which case their magnitude alone (and not their signs) will be taken into account. The arithmetic mean of these absolute deviations may be taken as the required measure of dispersion. It is referred to as the *mean deviation* of *x* about *A*. Denoting this mean deviation by MD_A , we have thus

$$MD_A = \frac{1}{n} \sum |x_i - A|. \quad \dots \quad (7.1)$$

It can be shown that the mean deviation is least when measured about the (a) median. A simple proof of the proposition is given below :

Let the given values of *x* be arranged in an increasing order of magnitude, $x_{(i)}$ being the *i*th value in this order. Obviously, we have

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Case 1 Let n be even and equal to $2m$. The sum $|x_{(1)} - A| + |x_{(2m)} - A|$ will be least, having the value $x_{(2m)} - x_{(1)}$, if A lies anywhere between $x_{(1)}$ and $x_{(2m)}$. Similarly, $|x_{(2)} - A| + |x_{(2m-1)} - A|$ is least when A lies between $x_{(2)}$ and $x_{(2m-1)}$, and so on. Lastly, $|x_{(m)} - A| + |x_{(m+1)} - A|$ is least when A lies between $x_{(m)}$ and $x_{(m+1)}$.

Each of these sums will be a minimum, and hence

$$MD_A = \frac{1}{n} \sum_i |x_i - A|$$

will be a minimum, when $x_{(m)} \leq A \leq x_{(m+1)}$.

Case 2 Let n be odd and equal to $2m+1$. As in the previous case, $|x_{(1)} - A| + |x_{(2m+1)} - A|$ is least if A is taken in between $x_{(1)}$ and $x_{(2m+1)}$. We take sums of pairs in this way, the last of this type being $|x_{(m)} - A| + |x_{(m+2)} - A|$, which is a minimum when A lies between $x_{(m)}$ and $x_{(m+2)}$. Finally, $|x_{(m+1)} - A|$ is smallest, i.e. equal to 0, if $A = x_{(m+1)}$.

Hence here MD_A will be a minimum when $A = x_{(m+1)}$.

In either case, it is seen that A should be the (a) median of x if MD_A is to be a minimum.

Owing to this result, it would seem proper to use the (a) median as origin in computing the mean deviation. In practice, however, the mean deviation is generally computed about the arithmetic mean.

Ex 7.1 Consider the data of Table 6.1. Here the median may be taken as the arithmetic mean of the 6th and 7th values in the arrangement of the data in an ascending order of magnitude, viz 1,232 gm and 1,278 gm. Thus the median is $M_1 = 1,255$ gm. The mean deviation about M_1 is obtained with the help of Table 7.1.

Thus the mean deviation about median is

$$MD_{M_1} = \frac{\sum_i |x_i - M_1|}{n}$$

$$= \frac{1,034}{12} = 86.17 \text{ gm}$$

If, instead, we want to calculate the mean deviation about the mean, some simplification can be made.

We may write

$$\sum_i |x_i - x| = \sum_1 (v - v_i) + \sum_2 (x_i - x),$$

Σ_1 containing the values of x that are less than or equal to \bar{x} and Σ_2 the values of x that are greater than \bar{x} .

But

$$\sum_i (x_i - \bar{x}) = 0,$$

that is,

$$\sum_1 (x_i - \bar{x}) + \sum_2 (x_i - \bar{x}) = 0,$$

so that

$$\sum_1 (\bar{x} - x_i) = \sum_2 (x_i - \bar{x}).$$

Hence

$$MD_{\bar{x}} = 2 \sum_1 (\bar{x} - x_i) / n = 2 \sum_2 (x_i - \bar{x}) / n.$$

Therefore, it is necessary to consider only one of the two sets of values of x in computing $MD_{\bar{x}}$.

TABLE 7.1

DETERMINATION OF MEAN DEVIATION ABOUT MEDIAN
FOR THE DATA OF TABLE 6.1

Yield (gm.) x_i	$ x_i - Mi $
1,216	39
1,374	119
1,167	88
1,232	23
1,407	152
1,453	198
1,202	53
1,372	117
1,278	23
1,141	114
1,221	34
1,329	74
Total	1,034

Ex. 7.2 For the data of Table 6.1, $\bar{x} = 1,282.67$ gm. The values of x which are smaller than or equal to \bar{x} are (in gm.) 1216, 1167, 1232, 1202, 1278, 1141 and 1221. The absolute deviations from \bar{x} are (in gm.) :

66.67, 115.67, 50.67, 80.67, 4.67, 141.67 and 61.67, and their sum is 521.69 gm.

Hence $MD_x = 2 \times 521.69/12 = 86.95$ gm.

7.4 Standard deviation

In considering the deviations $x_i - A$ for obtaining a measure of dispersion, we may get rid of their signs (thus taking their magnitude only into account) by taking, instead of their absolute values $|x_i - A|$, their squares $(x_i - A)^2$. Like the absolute values, these squares will also reflect the dispersion of the variable about A . The positive square root of the arithmetic mean of these quantities, i.e.

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - A)^2}, \quad \dots \quad (7.2)$$

which is called the *root-mean-square deviation* about A , may then be accepted as a measure of dispersion alternative to the mean deviation. The square root is taken in order to express the measure in the same units as those of x .

(7.2) is least when $A = \bar{x}$, as is evident from the fact that

$$\begin{aligned} \sum_i (x_i - A)^2 &= \sum_i ((x_i - \bar{x}) + (\bar{x} - A))^2 \\ &= \sum_i (x_i - \bar{x})^2 + 2(\bar{x} - A) \sum_i (x_i - \bar{x}) + n(\bar{x} - A)^2 \\ &= \sum_i (x_i - \bar{x})^2 + n(\bar{x} - A)^2 \\ &\quad [\text{since } \sum_i (x_i - \bar{x}) = 0, \text{ from (6.3)}] \\ &\geq \sum_i (x_i - \bar{x})^2, \end{aligned}$$

the equality sign being valid when and only when $A = \bar{x}$.

The measure of dispersion obtained by putting \bar{x} for A in (7.2) is called the *standard deviation* of x and is denoted by s . We have, therefore,

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad \dots \quad (7.3)$$

For computational purposes (7.3) may be expressed in a simpler form. We have

$$\begin{aligned} \sum_i (x_i - \bar{x})^2 &= \sum_i x_i^2 - 2\bar{x} \sum_i x_i + n\bar{x}^2 \\ &= \sum_i x_i^2 - n\bar{x}^2, \quad \text{since } \sum_i x_i = n\bar{x}. \end{aligned}$$

Hence

$$s = \sqrt{\frac{1}{n} \sum_i x_i^2 - \bar{x}^2}. \quad \dots \quad (7.4)$$

For grouped data, the standard deviation is given by

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 f_i}, \quad \dots \quad (7.5)$$

the symbols having the same significance as before. This becomes, on simplification,

$$s = \sqrt{\frac{1}{n} \sum_i x_i^2 f_i - \bar{x}^2}. \quad \dots \quad (7.6)$$

When the values of x are very large in magnitude, the computation of s by means of the above formula becomes laborious, as the squaring process yields still larger values. Short-cut methods of computing s will be discussed in Chapter 8. Some properties of the standard deviation are given below :

(a) Suppose that the given values of x are all equal : $x_i = a$, say, for $i = 1, 2, \dots, n$. Then $\bar{x} = \frac{\sum x_i}{n} = a$, so that for each i , $x_i - \bar{x} = 0$.

Hence

$$s = \sqrt{\frac{1}{n} \cdot n \cdot 0} = 0. \quad \dots \quad (7.7)$$

Thus the s.d. of a variable whose values are all equal must be zero. The converse also is true, as the reader can see for himself.

(b) Let y be equal to $a + bx$; then $\bar{y} = a + b\bar{x}$. Hence

$$y_i - \bar{y} = b(x_i - \bar{x}) \text{ for each } i.$$

It follows that

$$\sum_i (y_i - \bar{y})^2 = b^2 \sum_i (x_i - \bar{x})^2.$$

Dividing both sides by n and taking positive square roots, one has

$$s_y = |b| s_x, \quad \dots \quad (7.8)$$

where the symbols s_x and s_y denote the s.d.s of x and y , respectively.

(c) Let there be two sets of values of x with n_1 and n_2 values,

and let x_1, \bar{x}_1 be their means and s_1, s_2 their s.d.s. Then the s.d. of x for the two sets pooled together can be expressed in terms of $n_1, r_1, \bar{x}_1, \bar{x}_2$, and s_1, s_2 . By (6.5) the grand mean of x is

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Denoting by x_{ij} ($j=1, 2, \dots, n_1$) and by x_{2j} ($j=1, 2, \dots, n_2$) the values in the two sets, we may write the sum of squares of the deviations of the values from \bar{x} as

$$(n_1 + n_2)s^2 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x})^2 + \sum_{j=1}^{n_2} (x_{2j} - \bar{x})^2,$$

where s^2 is the total variance of x . But

$$\begin{aligned} \sum_j (x_{1j} - \bar{x})^2 &= \sum_j ((x_{1j} - \bar{x}_1) + (\bar{x}_1 - \bar{x}))^2 \\ &= \sum_j (x_{1j} - \bar{x}_1)^2 + n_1(\bar{x}_1 - \bar{x})^2 \\ &= n_1 s_1^2 + n_1(\bar{x}_1 - \bar{x})^2 \end{aligned}$$

Similarly,

$$\sum_j (x_{2j} - \bar{x})^2 = n_2 s_2^2 + n_2(\bar{x}_2 - \bar{x})^2$$

Thus

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2} \quad (7.9)$$

Generally, if there be t sets with n_1, n_2, \dots, n_t values, having means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_t$ and s.d.s s_1, s_2, \dots, s_t , then the s.d. of x for all the sets taken together is given by

$$s^2 = \frac{\sum_{i=1}^t n_i s_i^2}{\sum_{i=1}^t n_i} + \frac{\sum_{i=1}^t n_i (\bar{x}_i - \bar{x})^2}{\sum_{i=1}^t n_i}, \quad (7.10)$$

where \bar{x} is the grand mean of x .

Ex 7.3 The result in (b) above shows that the standard deviation remains unchanged under a change of origin. This often means a simplification in the computation of the standard deviation.

This is illustrated here for the data of Table 6.1

We are taking a new variable

$$u = x - 1,200,$$

the origin 1,200 being chosen to make the values of u small enough.

TABLE 7.2

DETERMINATION OF S.D. FOR THE DATA OF TABLE 6.1

Yield (gm.) x_i	$u_i = x_i - 1,200$	u_i^2
1,216	16	256
1,374	174	30,276
1,167	-33	1,089
1,232	32	1,024
1,407	207	42,849
1,453	253	64,009
1,202	2	4
1,372	172	29,584
1,278	78	6,084
1,141	-59	3,481
1,221	21	441
1,329	129	16,641
Total	992	195,738

Here

$$\bar{u} = \sum u_i / 12 = 992 / 12 = 82.67 \text{ gm.},$$

$$\begin{aligned}
 s_x^2 &= s_u^2 = \frac{1}{12} \sum u_i^2 - \bar{u}^2 \\
 &= \frac{1}{12} \times 195,738 - (82.67)^2 \\
 &= 16,311.5 - 6,834.33 \\
 &= 9,477.17 \text{ (gm.)}^2.
 \end{aligned}$$

Here

$$s_x = \sqrt{9,477.17} = 97.35 \text{ gm.}$$

7.5 Comparison of range, mean deviation and standard deviation

All three measures of dispersion, so far considered, are rigidly defined. The range is, however, inferior in one respect—it becomes meaningless when at least one of the two limits of the variable is infinite.

The range is the easiest to compute. The other two require almost the same amount of computational labour.

The significance of the range is easily comprehensible. The general nature of the mean deviation also is readily understood. The standard deviation, on the other hand, has a comparatively abstract nature.

Both the mean deviation and the standard deviation are based on all the given values of the variable. And hence, properly speaking, they characterise the whole set of values. The range is inferior in this respect, being based only on the two extreme values of the set. In fact, it often fails to give a proper idea of the dispersion. Suppose, for instance, that a variable assumes in one case the values

25, 25, 25, 25, 60, 60, 60, 60,

and in another the values

25, 31, 32, 44, 53, 56, 59, 60

The range is the same in the two cases, viz. $60 - 25 = 35$, but the dispersion cannot be said to be the same. For while in the first case four of the values are equal to 25 and four are equal to 60, in the second there is only one 25 and only one 60, the other values lying near about the average. The actual dispersion of the former set is, therefore, much larger than that of the latter.

The standard deviation has certain desirable properties which make it easily amenable to algebraical treatment. No such properties are, however, possessed by the range or the mean deviation.

It follows, therefore, that the standard deviation may be generally regarded as the best measure of dispersion, just as the arithmetic mean generally serves as the best measure of central tendency. In some particular cases the range is employed, instead of the standard deviation, because it is much easier to compute. This is often done, for example, in *statistical quality control*, where the analysis of the data must be done immediately after they are collected in order that effective action may be taken on the basis of the analysis.

7.6 Measures based on mutual differences of observations

Some statisticians suggest measures of dispersion that do not depend on any particular measure of central tendency, the choice of which in any case is bound to be more or less arbitrary. The notion of central tendency, they argue, is not relevant when we are looking for a measure of dispersion. Since dispersion really means the extent to which the given values of the variable differ from one another (rather than from any arbitrarily chosen average), any proper measure of dispersion, in their opinion, should be based solely on the mutual differences $x_i - x_j$ ($i, j = 1, 2, \dots, n$) of the values of x .

A measure of this type, suggested by C. Gini, is the mean of the absolute values of all n^2 mutual differences. Called Gini's *mean difference*, it is thus given by the formula

$$\Delta_1 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|. \quad \dots \quad (7.11)$$

In case of grouped data, we have, with the usual notation,

$$\Delta_1 = \frac{1}{n^2} \sum_{i=1}^k \sum_{j=1}^k f_i f_j |x_i - x_j|. \quad \dots \quad (7.11a)$$

However, this measure is, like a mean deviation, difficult to handle mathematically. The following measure, which is based on the squares of mutual differences of the observations, (or its positive square root) is better in this respect :

$$\Delta_2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2. \quad \dots \quad (7.12)$$

For grouped data, we have

$$\Delta_2 = \frac{1}{n^2} \sum_{i=1}^k \sum_{j=1}^k f_i f_j (x_i - x_j)^2. \quad \dots \quad (7.12a)$$

The merit of such measures is apparent from what we have said at the beginning of this section. Even so, it is doubtful whether a measure other than the standard deviation is at all necessary. Indeed, the standard deviation itself has this property of being based solely on the differences of the observations among themselves.

$$\text{For } \sum_i \sum_j (x_i - x_j)^2 = \sum_i \sum_j [(x_i - \bar{x}) - (x_j - \bar{x})]^2 \\ = n \sum_i (x_i - \bar{x})^2 - 2 \sum_i (x_i - \bar{x}) \sum_j (x_j - \bar{x}) + n \sum_i (x_i - \bar{x})^2 \\ = 2n s^2,$$

so that

$$s^2 = \frac{\Delta_2}{2} \quad (7.13)$$

We should rather say that here we have one more reason why the standard deviation should generally be regarded as the best measure of dispersion.

7.7 Quartile deviation

In the previous chapter, we defined the median as a value of the variable such that half of the given values are less than or equal to it and the remaining half are greater than or equal to it. The general name for a measure of this type is *quantile*. The quantile or *fractile* of order p (or the p -quantile) is a value of the variable such that a proportion p of the total number of given values are less than or equal to it and a proportion $(1-p)$ are greater than or equal to it. For a continuous variable this quantile (here denoted by z_p) may be approximately determined by the formula

$$z_p = x_l + \frac{np - n_l}{f_0} \times c, \quad (7.14)$$

where x_l = the lower boundary of the class-interval in which z_p lies,

c = width of this class interval,

n_l = cumulative frequency (of the 'less than' type) corresponding to x_l and

f_0 = frequency in this class

A frequency distribution may be briefly described by giving the values of some of its quantiles. Generally, it will be enough to give, together with the median which is $z_{\frac{1}{2}}$, the quantiles $z_{\frac{1}{4}}$ and $z_{\frac{3}{4}}$.

These three, $z_{\frac{1}{4}}$, $z_{\frac{1}{2}}$ and $z_{\frac{3}{4}}$, taken together divide, as it were, the frequency distribution of the variable into four equal parts. Hence they are also called the *quartiles*. $z_{\frac{1}{4}}$ is the first or lower quartile (denoted by Q_1), the median is the second quartile ($M_1 = Q_2$), and $z_{\frac{3}{4}}$ the third or upper quartile (Q_3).

The lower and upper quartiles provide us with another measure of dispersion. This measure is

$$Q = \frac{Q_3 - Q_1}{2}, \quad \dots \quad (7.15)$$

called the *quartile deviation* or *semi-interquartile range*.

The quartile deviation is rarely used, except when the computation of the standard deviation is extremely difficult or impossible—for instance, when the observations are given in a frequency table with class-intervals of varying width or with one or both of the terminal classes undefined.

Ex. 7.4 Consider the frequency distribution of Table 5.10. Here $n/4=44.25$. Hence we find, on going through the cumulative frequencies of the 'less-than' type, that the first quartile lies between $x_l=159.55$ cm. and $x_u=164.55$ cm., for which the cumulative frequencies are $n_l=28$ and $n_u=86$. Here $c=5$ and $f_0=58$. Hence, from formula (7.14),

$$\begin{aligned} Q_1 &= 159.55 + \frac{44.25 - 28}{58} \times 5 \\ &= 159.55 + 1.401 = 160.951 \text{ cm.} \end{aligned}$$

Similarly, since

$$\frac{3n}{4}=132.75,$$

$$\begin{aligned} Q_3 &= 164.55 + \frac{132.75 - 86}{60} \times 5 \\ &= 164.55 + 3.896 = 168.446 \text{ cm.} \end{aligned}$$

The quartile deviation is, therefore,

$$Q = \frac{168.446 - 160.951}{2} = 3.748 \text{ cm.}$$

7.8 Measures of relative dispersion

The measures of dispersion we have discussed above are all expressed in the same units as those of the variable. As such they cannot be used in comparing two distributions of different types with respect to their variability. A difficulty is encountered, for example, when we want to compare the dispersion of a set of heights (given in, say, cm.) with the dispersion of a set of weights (given in, say, kg.).

For purposes of such comparison, therefore, a measure of dispersion has to be made free from the units of the variable. The simplest procedure is to express a measure of dispersion as a percentage of a measure of central tendency. The most commonly used measure of this type is the *coefficient of variation*,

$$v = 100 \frac{s}{\bar{x}}, \quad (7.16)$$

where \bar{x} is supposed to be non-zero.

Besides the above use, such measures serve another purpose. Suppose repeated measurements are being taken of two rods, one of length 10 cm and another of length 100 cm. The means of the measurements will be 10 cm and 100 cm respectively, provided the measurements are free from bias. Let the standard deviation of each set of measurements be 2 cm. But a standard deviation of 2 cm in the first case does not mean the same thing as a standard deviation of 2 cm in the second. For the first set of measurements are then much less accurate than the second set of measurements. The coefficient of variation, on the other hand, will give a true picture of their relative accuracy. Thus it may be useful even when we want to compare sets of data expressed in the same units.

7.9 Curve of concentration

A special type of cumulative frequency graph, known as a *curve of concentration* or *Lorenz curve*, is useful in studying the concentration of wealth or income in relation to certain segments of the population and in similar other situations.

Let $F(x)$ denote the percent cumulative frequency for the variable up to the value x and let $\Phi(x)$ denote the percent cumulative total for the variable up to the value x . Naturally, both F and Φ vary from 0 to 100.

The curve obtained by plotting Φ against F for different fixed values of x is known as the *curve of concentration* or *Lorenz curve*. It can be shown that the curve is necessarily concave upwards (as in Fig. 7.1).

The line $\Phi=F$ is called the *line of equal distribution*. Such a curve would indicate that any specified proportion of persons would have precisely the same proportion of total value. In the case of an

income distribution, it would mean that 20% of persons would earn 20% of the income, 50% of persons would earn 50% of the income, 75% of persons would receive 75% of the income, and so on.

The more the departure of the Lorenz curve from the line of equal distribution, the more is the concentration of the total value (say, income) in a few individuals. Thus in a particular case, if we find that 50% of the persons receive only 20% of the total income, 75% of the persons receive only 30%, 90% receive 50%, and so on, it means that there is a lot of income concentration in a few individuals in the upper income groups. Thus the area between the line of equal distribution and the curve of concentration, called the *area of concentration*, is an indicator of the degree of concentration ; the larger the area the more is the concentration. Twice the area is Gini's *coefficient of concentration*.

The Lorenz curve for the data of *Exercise 6.20* has been drawn in Fig. 7.1. The necessary calculations are shown below :

TABLE 7.3

CALCULATIONS FOR DRAWING THE LORENZ CURVE FOR
THE DATA OF EXERCISE 6.20

Monthly income (Rs.)	Mid-point x	Frequency f	Cumulative frequency	Total income $x.f$	Cumulative total income	Percent cumulative frequency	Percent cumulative total income
—100	75	53	53	3,975	3,975	9·14	3·23
100—150	125	81	134	10,125	14,100	23·10	11·44
150—200	175	114	248	19,950	34,050	42·76	27·64
200—250	225	195	443	43,875	77,925	76·38	63·25
250—300	275	63	506	17,325	95,250	87·24	77·31
300—350	325	32	538	10,400	105,650	92·76	85·75
350—400	375	20	558	7,500	113,150	96·20	91·84
400—450	425	11	569	4,675	117,825	98·10	95·64
450—500	475	8	577	3,800	121,625	99·48	98·72
500—	525	3	580	1,575	123,200	100·00	100·00

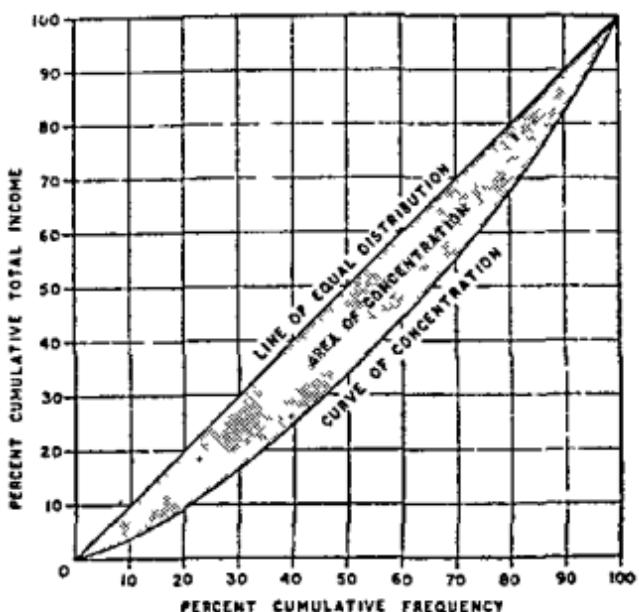


Fig. 7.1 The curve of concentration and the area of concentration for the data of *Exercise 6.20*.

Questions and exercises

7.1 What is dispersion? What are the common measures of dispersion?

7.2 How, in your opinion, should a measure of dispersion change when all values of the variable are increased or decreased

- (1) by the same amount?
- (2) in the same proportion?

Judge in this light the different measures of dispersion considered in this chapter.

7.3 Compare the range, the mean deviation and the standard deviation as measures of dispersion.

7.4 Define standard deviation. State and prove its important properties.

7.5 What is meant by relative dispersion? Define coefficient of variation and explain its usefulness.

7.6 Obtain the mean and standard deviation of the first n natural numbers.

$$Ans. \ (n+1)/2; \ \sqrt{(n^2-1)/12}.$$

7.7 Prove that for any set of values, x_1, x_2, \dots, x_n ,

$$x_1^2 + x_2^2 + \dots + x_n^2 \geq \frac{(x_1 + x_2 + \dots + x_n)^2}{n}.$$

7.8 Show that the mean deviation about mean cannot exceed the standard deviation. When are the two equal?

7.9 Suppose the given values of x (viz. x_i , for $i=1, 2, \dots, n$) are such that $a \leq x_i \leq b$ for each i .

Show that (i) $a \leq \bar{x} \leq b$; and (ii) $0 \leq s^2 \leq (b-a)^2/4$.

[Hint : For (ii) use the result

$$\sum_i (x_i - \bar{x})^2 \leq \sum_i \left(x_i - \frac{a+b}{2}\right)^2 = \sum_1 \left(x_i - \frac{a+b}{2}\right)^2 + \sum_2 \left(x_i - \frac{a+b}{2}\right)^2,$$

where for each i , x_i is included in \sum_1 or \sum_2 according as $x_i \leq (a+b)/2$ or $x_i > (a+b)/2$, so that in either case $\left(x_i - \frac{a+b}{2}\right)^2 \leq \frac{1}{4}(b-a)^2$.]

7.10 Let s and R be, respectively, the standard deviation and range of a set of n values of x . Show that

$$\frac{R^2}{2n} \leq s^2 \leq \frac{R^2}{4}.$$

When do the equalities hold?

7.11(a) Show that relation (7.9) can be expressed in the alternative form

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^2.$$

(b) In a batch of 10 children, the I.Q. of a dull boy is 36 below the average I.Q. of the other children. Show that the s.d. of I.Q. for all the children cannot be less than 10.8. If this standard deviation is actually 11.4, determine what the standard deviation will be when the dull boy is left out. *Ans.* 39.

7.12 Suppose that the variable x takes positive values only and that the deviations $x_i - \bar{x}$ are small compared to \bar{x} . Show that in such a case

$$(1) \quad x_g \simeq \bar{x} \left(1 - \frac{1}{2} \cdot \frac{s^2}{\bar{x}^2}\right) \text{ and } (2) \quad x_h \simeq \bar{x} \left(1 - \frac{s^2}{\bar{x}^2}\right).$$

7.13(a) Show that the mean deviation MD_A may be obtained by the formula :

$$nMD_A = S_2 - S_1 + A(n_1 - n_2),$$

where S_1 is the sum of the values that are less than A and n_1 is the

number of such values, while S_2 is the sum of the values that are greater than A and n_2 is the number of such values

(b) Hence supply an alternative proof for the result that MD_A is a minimum when A is a median

7.14 Determine the range, the mean deviation about mean and the standard deviation for the data of *Exercise 6.16*

Ans 33 cm, 0.71 cm, 0.91 cm

7.15 For the frequency distribution of *Exercise 6.18*, compute the mean deviation about median and the standard deviation

Ans 1.494, 1.858

7.16 Evaluate the three quartiles for the frequency distribution of *Exercise 6.19*. Next determine the mean deviation about median, the standard deviation and the quartile deviation

Ans 97.08, 108.41, 118.33, 13.36, 17.26, 10.63

7.17 Compute the standard deviation of the age-distribution of Bengali males given in *Exercise 6.21* by using formula (7.10)

Ans 20.39 yr

7.18 For a distribution of 250 observations on a variable x , the mean and standard deviation are, respectively, 65.7 and 4.4. However, on scrutinising the data it is found that two observations, which should correctly read as 71 and 83, had been wrongly recorded as 91 and 80. Obtain the correct values of the mean and the standard deviation

Ans $\bar{x}=65.6$, $s=4.2$

[*Hint* Use formulæ (6.5) and (7.10)]

7.19 The number of runs scored by cricketers A and B during a test series consisting of 5 test matches is shown below for each of the 10 innings

Cricketer A —5, 26, 97, 76, 112, 89, 6, 108, 24, 16

Cricketer B —51, 47, 36, 60, 58, 39, 44, 42, 71, 50

Make a comparative study of their batting performance

SUGGESTED READING

- [1] Mills, F C *Statistical Methods* (Ch 5) H Holt, 1955
- [2] Wallis, W A and Roberts, H V *Statistics a New Approach* (Ch 8) Methuen, 1957
- [3] Yule, G U and Kendall, M G *Introduction to the Theory of Statistics* (Ch 6) Charles Griffin, 1953

8

MOMENTS AND MEASURES OF
SKEWNESS AND KURTOSIS

8.1 Moments

In the two preceding chapters, we considered the mean,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{or} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^k x_i f_i,$$

and the variance,

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{or} \quad s^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 f_i,$$

as possible measures of the central tendency and dispersion of the variable x .

The most general measure of this type is

$$m'_r = \frac{1}{n} \sum_{i=1}^n (x_i - A)^r, \quad r = 0, 1, 2, \dots, \dots \quad (8.1)$$

which is called the r th moment of x about the origin A . If the given values are classified into a frequency table, the formula takes the form

$$m'_r = \frac{1}{n} \sum_{i=1}^k (x_i - A)^r f_i, \quad \dots \quad (8.1a)$$

x_i being the class-mark of the i th class and f_i its frequency.

When the origin of a moment is taken at the arithmetic mean of the variable, it is called a *central moment*. Thus the r th central moment, denoted by m_r , is

$$m_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r \quad \left. \begin{array}{l} \\ \end{array} \right\} \dots \quad (8.2)$$

or $m_r = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^r f_i, \quad \left. \begin{array}{l} \\ \end{array} \right\}$

according as the values are ungrouped or are grouped into a frequency table.

Evidently, we have always

$$m'_0 = m_0 = 1 \quad \text{and} \quad m_1 = 0.$$

The mean of a variable is its 1st moment about 0, while the variance is the 2nd central moment.

8.2 Central moments expressed in terms of moments about an arbitrary origin

When m'_1, m'_2, \dots, m_r —the moments about an arbitrary origin A —are given, the central moments up to that of the r th order can be obtained by using certain algebraic relations connecting the two sets of moments. These are deduced below.

We have

$$\begin{aligned} (x_i - x)^r &= \{(x_i - A) - (\bar{x} - A)\}^r \\ &= (x_i - A)^r - \binom{r}{1}(x_i - A)^{r-1}(x - A) \\ &\quad + \binom{r}{2}(x_i - A)^{r-2}(x - A)^2 - \\ &\quad + (-1)^{r-2}\binom{r}{r-2}(x_i - A)^2(x - A)^{r-2} \\ &\quad + (-1)^{r-1}\binom{r}{r-1}(x_i - A)(\bar{x} - A)^{r-1} + (-1)^r\binom{r}{r}(x - A)^r. \end{aligned}$$

Summing both sides for all i from 1 to n , and dividing by n , we get

$$\begin{aligned} m_r &= m_r - \binom{r}{1}m_{r-1}m'_1 + \binom{r}{2}m_{r-2}m_1^2 - \\ &\quad + (-1)^{r-2}\binom{r}{r-2}m_2m_1^{r-2} + (-1)^{r-1}\binom{r}{r-1}m'_1m_1^{r-1} \\ &\quad + (-1)^r\binom{r}{r}m_1^r, \end{aligned} \tag{8.3}$$

since

$$x - A = \frac{1}{n} \sum_{i=1}^n (x_i - A) = m_1 \tag{8.4}$$

It is easily seen that relation (8.3) holds for moments obtained from grouped data as well.

Some particular cases of (8.3) are

$$\left. \begin{aligned} m_1 &= m'_1 - m_1' = 0, \\ m_2 &= m'_2 - 2m'_1m_1' + m_1^2 = m'_2 - m_1^2, \\ m_3 &= m'_3 - 3m'_2m_1' + 3m'_1m_1^2 - m_1^3 = m'_3 - 3m'_2m_1' + 2m_1^3, \\ m_4 &= m'_4 - 4m'_3m_1' + 6m'_2m_1^2 - 4m'_1m_1^3 + m_1^4 \\ &= m'_4 - 4m'_3m_1' + 6m'_2m_1^2 - 3m_1^4 \end{aligned} \right\} \tag{8.5}$$

In most practical cases, it will be sufficient to calculate x, m_2, m_3

and m_4 . These computations are greatly facilitated by first computing moments about a suitably chosen origin and then by using relations (8.4) and (8.5). This will be evident from the following examples :

Ex. 8.1 Consider once again the data of Table 6.1, representing the yields in gm. of 12 tomato plants. Here the values are quite large, so that the direct computation of the first four moments will be extremely laborious, as this will require obtaining squares, cubes and fourth powers of the given quantities and finding their totals. The computational labour can be reduced a great deal by first taking deviations of the given values about a suitable origin and then computing moments about that origin.

In the present case, we may take the origin at 1,300 gm. The different steps in the calculation of moments about 1,300 gm. are indicated in the following table :

TABLE 8.1
CALCULATION OF MOMENTS FOR THE DATA OF TABLE 6.1

Yield (gm.) x_i	$u_i = x_i - 1,300$	u^2	u_i^3	u_i^4
1,216	-84	7,056	-592,704	49,787,136
1,374	74	5,476	405,224	29,986,576
1,167	-133	17,689	-2,352,637	312,900,721
1,232	-68	4,624	-314,432	21,381,376
1,407	107	11,449	1,225,043	131,079,601
1,453	153	23,409	3,581,577	547,981,281
1,202	-98	9,604	-941,192	92,236,816
1,372	72	5,184	373,284	26,873,856
1,278	-22	484	-10,648	234,256
1,141	-159	25,281	-4,019,679	639,128,961
1,221	-79	6,241	-493,039	38,950,081
1,329	29	841	24,387	707,281
Total	-208	117,338	-3,114,850	1,891,247,942

From this table, one gets

$$m_1 = \frac{1}{4} \times (-208) = -17,333 \text{ gm},$$

$$m_2 = \frac{1}{4} \times 117,338 = 9,778,1667 \text{ (gm)}^2,$$

$$m_3 = \frac{1}{4} \times (-3,114,850) = -259,570,8333 \text{ (gm)}^3$$

and $m_4 = \frac{1}{4} \times 1,891,247,942 = 157,603,995,1667 \text{ (gm)}^4$

Hence the mean, the variance (and the s.d.) and the third and fourth central moments will be as follows

$$\bar{x} = 1,300 + m_1 = 1,282.67 \text{ gm},$$

$$m_2 = m_2 - m_1^2 \\ = 9,778,1667 - 300,4433 = 9,477,7234 \text{ (gm)}^2,$$

so that $s = \sqrt{9,477,7234} = 97.35 \text{ gm},$

$$m_3 = m_3 - 3m_2m_1 + 2m_1^3 \\ = -259,570,8333 + 508,463,6906 - 1,041,5348 \\ = 247,851.32 \text{ (gm)}^3$$

and $m_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4 \\ = 157,603,995,1667 - 17,996,876,4994 \\ + 17,626,708,0278 - 270,798,5295 \\ - 156,963,028,17 \text{ (gm)}^4$

The values of \bar{x} and s computed by this short cut method may be compared with those obtained directly earlier (in Ex 6.1 and Ex 7.3). The two sets of values are, to all intents and purposes, identical. The slight difference between the two values of the standard deviation is to be attributed to errors of approximation.

In the case of grouped data, simplifications can be made in the calculation of moments if the deviations about the chosen origin are divided by a suitable number. The class width may be used for this purpose when the class intervals are equally wide.

To guard against computational mistakes, one may also use some checks which are due to Charlier. For a frequency distribution, these checks are based on the following relations:

We have $\sum_i (u_i + 1)f_i = \sum_i u_i f_i + n$. This provides a check on the value of $\sum_i u_i f_i$.

Again, $\sum_i (u_i + 1)^2 f_i = \sum_i u_i^2 f_i + 2 \sum_i u_i f_i + n$. This can be used to check the values of $\sum_i u_i f_i$ and $\sum_i u_i^2 f_i$.

And so on.

Ex. 8.2 Let us take for illustration the data on the heights of Indian adult males shown in Table 5.10.

TABLE 8.2
CALCULATION OF MOMENTS FOR THE DATA OF TABLE 5.10

Class-mark x_i	Frequency f_i	$u_i = \frac{x_i - 162.05}{5}$	$u_i f_i$	$u_i^2 f_i$	$u_i^3 f_i$	$u_i^4 f_i$	$(u_i + 1) f_i$
147.05	1	-3	-3	9	-27	81	16
152.05	3	-2	-6	12	-24	48	3
157.05	24	-1	-24	24	-24	24	0
162.05	58	0	0	0	0	0	58
167.05	60	1	60	60	60	60	960
172.05	27	2	54	108	216	432	2,187
177.05	2	3	6	18	54	162	512
182.05	2	4	8	32	128	512	1,250
Total	177	—	95	263	383	1,319	4,986

Here $\sum_i (u_i + 1)^4 f_i = 4,986$,

while $\sum_i u_i^4 f_i + 4 \sum_i u_i^3 f_i + 6 \sum_i u_i^2 f_i + 4 \sum_i u_i f_i + n$

$$= 1,319 + 1,532 + 1,578 + 380 + 177 = 4,986.$$

The two values being equal, the computations *may be* supposed to be free from errors.

Now the raw moments are

$$m'_1 = \left(\frac{1}{n} \sum_i u_i f_i \right) \times 5$$

$$= \left(\frac{1}{177} \times 95 \right) \times 5 = 0.53672 \times 5 \text{ cm.},$$

$$m_2' = \left(\frac{1}{n} \sum u_i^2 f_i \right) \times 5^2 \\ = \left(\frac{1}{177} \times 263 \right) \times 5^2 = 1\ 48588 \times 5^2 \text{ (cm)}^2,$$

$$m_3' = \left(\frac{1}{n} \sum u_i^3 f_i \right) \times 5^3 \\ = \left(\frac{1}{177} \times 383 \right) \times 5^3 = 2\ 16384 \times 5^3 \text{ (cm)}^3$$

and

$$m_4 = \left(\frac{1}{n} \sum u_i^4 f_i \right) \times 5^4 \\ = \left(\frac{1}{177} \times 1,319 \right) \times 5^4 = 7\ 45198 \times 5^4 \text{ (cm)}^4$$

We have, therefore,

$$x = 162\ 05 + m_1 = 162\ 05 + 0\ 53672 \times 5 \\ = 162\ 05 + 2\ 6836 = 164\ 734 \text{ cm},$$

$$m_2 = m_2 - m_1^2 \\ = (1\ 48588 - 0\ 28807) \times 5^2 \\ = 1\ 19781 \times 5^2 = 29\ 945 \text{ (cm)}^2,$$

so that $s = \sqrt{29\ 945} = 5\ 472 \text{ cm}$

$$m_3 = m_3 - 3m_1 m_2 + 2m_1^3 \\ = (2\ 16384 - 2\ 39250 + 0\ 30923) \times 5^3 \\ = 0\ 08057 \times 5^3 = 10\ 071 \text{ (cm)}^3,$$

and

$$m_4 = m_4 - 4m_1 m_2 + 6m_2 m_1^2 - 3m_1^4 \\ = (7\ 45198 - 4\ 64550 + 2\ 56822 - 0\ 24895) \times 5^4 \\ = 5\ 12575 \times 5^4 = 3,203\ 594 \text{ (cm)}^4$$

The values of x and s may also be obtained by using formulæ (6.2) and (7.6). The two sets of values will be identical for all practical purposes.

It is obvious from the above examples that for computational convenience the origin should be taken somewhere at the middle of the range of the given values of the variable. Secondly, it should be a round number (a class mark in the case of grouped data), so that the deviations about it can be readily obtained.

8.3 Moments about an arbitrary origin expressed in terms of central moments

Just as a central moment can be expressed in terms of moments about an arbitrary origin, so a moment about an arbitrary origin is expressible in terms of central moments.

We have

$$\begin{aligned}(x_i - A)^r &= \{(x_i - \bar{x}) + (\bar{x} - A)\}^r \\ &= \{(x_i - \bar{x}) + d\}^r,\end{aligned}$$

where

$$d = \bar{x} - A.$$

Thus

$$\begin{aligned}(x_i - A)^r &= (x_i - \bar{x})^r + \binom{r}{1} (x_i - \bar{x})^{r-1} d + \binom{r}{2} (x_i - \bar{x})^{r-2} d^2 + \dots \\ &\quad + \binom{r}{r-1} (x_i - \bar{x}) d^{r-1} + \binom{r}{r} d^r.\end{aligned}$$

Taking the sum of each side for all i from 1 to n , and dividing by n , we get

$$\begin{aligned}m'_r &= m_r + \binom{r}{1} m_{r-1} d + \binom{r}{2} m_{r-2} d^2 + \dots \\ &\quad + \binom{r}{r-1} m_1 d^{r-1} + \binom{r}{r} d^r. \quad \dots \quad (8.6)\end{aligned}$$

The last term but one is, of course, zero since $m_1 = 0$.

In particular,

$$\left. \begin{aligned}m'_1 &= d, \\ m'_2 &= m_2 + d^2, \\ m'_3 &= m_3 + 3m_2 d + d^3,\end{aligned}\right\} \quad \dots \quad (8.7)$$

and

$$m'_4 = m_4 + 4m_3 d + 6m_2 d^2 + d^4.$$

8.4 Sheppard's corrections for moments

In computing moments for data grouped into class-intervals by means of the formulæ

$$m'_r = \frac{1}{n} \sum_i (x_i - A)^r f_i \text{ and } m_r = \frac{1}{n} \sum_i (x_i - \bar{x})^r f_i,$$

we are acting as if the observations falling in a class (e.g. the f_i values falling in the i th class) were all equal to the class-mark, although the observations may be really unequal. The assumption naturally introduces some error, which is called the *error due to grouping*. To

correct for this grouping error the computed values of the moments have to be suitably adjusted. A method for adjusting the moments for grouped data where the classes are equally wide has been developed by Sheppard. Sheppard's corrections for moments about an arbitrary origin and for central moments of the first four orders are given below

$$\left. \begin{aligned} m_1 (\text{corrected}) &= m_1 \\ m_2 (\text{corrected}) &= m_2 - \frac{\epsilon^2}{12} \\ m_3 (\text{corrected}) &= m_3 - \frac{\epsilon^2}{4} m_1 \\ m_4 (\text{corrected}) &= m_4 - \frac{\epsilon^2}{2} m_2 + \frac{7}{240} \epsilon^4, \end{aligned} \right\} \quad (88)$$

and

$$\left. \begin{aligned} m_2 (\text{corrected}) &= m_2 - \frac{\epsilon^2}{12} \\ m_3 (\text{corrected}) &= m_3 \\ m_4 (\text{corrected}) &= m_4 - \frac{\epsilon^2}{2} m_2 + \frac{7}{240} \epsilon^4, \end{aligned} \right\} \quad (89)$$

where ϵ is the width of each class interval

These corrections will be valid only if certain conditions are fulfilled. First, it is necessary that the observations should relate to a continuous variable. Second, the frequency curve of the distribution should be continuous and should have high order contact at both ends of the range of the variable. The second condition cannot, of course, be verified when only a finite number of observations are available, which is usually the case. A fair indication may, however, be obtained from a frequency table of the data provided a sufficiently large number of observations are given. Thus if the class-frequencies decrease towards the two ends of the range smoothly and gradually (as in Table 5.10 or Table 8.3), the condition may be supposed to be fulfilled.

Two other conditions should be satisfied by the observed data in order that the use of Sheppard's corrections may be worth while. It is, in the first place, desirable that the total frequency should be sufficiently large, otherwise, the moments will be more affected by sampling errors (*vide* Chapter 14) than by grouping errors.

Furthermore, the width of the class-intervals should not be too small compared to the range of variation of the data—in other words, the number of classes should not be too large ; for otherwise Sheppard's corrections will make little difference to the uncorrected values of the moments. As a general rule, these corrections should not be applied unless the total frequency is higher than 1,000 and the number of classes is smaller than 20.

The use of Sheppard's corrections is illustrated in the following example :

Ex. 8.3 During a crop-cutting survey on rice in a State of India, 1,175 cuts (each of size $1/3,200$ acre) were taken. The yield of rice per cut for the 1,175 cuts is shown in Table 8.3 in the form of a frequency table.

TABLE 8.3
FREQUENCY DISTRIBUTION OF RICE YIELD FOR 1,175
CUTS OF $1/3,200$ ACRE EACH

Yield of rice (md.) class-mark x_i	Frequency f_i
1.25	5
3.75	37
6.25	68
8.75	131
11.25	142
13.75	158
16.25	186
18.75	171
21.25	99
23.75	75
26.25	35
28.75	27
31.25	18
33.75	12
36.25	5
38.75	4
41.25	2
Total	1,175

For this distribution, the uncorrected moments are found to be

$$\bar{x} = 15.8989 \text{ md},$$

$$m_2 = 45.9139 \text{ (md)}^2,$$

$$m_3 = 170.6747 \text{ (md)}^3$$

$$\text{and } m_4 = 7,272.9613 \text{ (md)}^4$$

Here we may apply Sheppard's corrections, because the given figures are such that the four conditions noted above may be taken to be fulfilled. We then have

$$\bar{x} (\text{corrected}) = 15.899 \text{ md},$$

$$m_2 (\text{corrected}) = 45.9139 - \frac{(2.5)^2}{12}$$

$$= 45.9139 - 0.5208 = 45.393 \text{ (md)}^2,$$

$$\text{so that } s (\text{corrected}) = \sqrt{45.393} = 6.737 \text{ md},$$

$$m_3 (\text{corrected}) = 170.675 \text{ (md)}^3$$

$$\text{and } m_4 (\text{corrected}) = 7,272.9613 - \frac{1}{2} \times (2.5)^2 \times 45.9139 + \frac{7}{240} \times (2.5)^4$$

$$= 7,272.9613 - 143.4809 + 1.1393$$

$$= 7,130.620 \text{ (md)}^4$$

8.5 Skewness

By skewness of a frequency distribution we mean the degree of its departure from symmetry. The frequency distribution of a discrete variable x is called *symmetrical* about the value x_0 , if the frequency of $x_0 - h$ is the same as the frequency of $x_0 + h$, whatever h

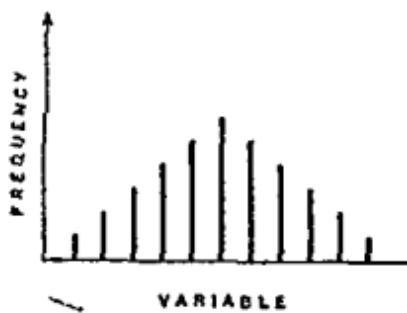


Fig. 8.1a A symmetrical distribution (discrete variable)

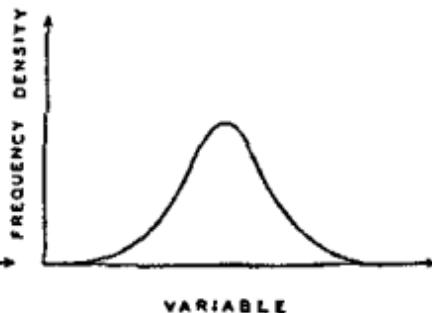


Fig. 8.1b A symmetrical distribution (continuous variable)

may be. In the case of a continuous variable, the term 'symmetry' should be used in relation to its frequency curve (*vide* Section 6.5). The frequency curve of a continuous variable is said to be symmetrical about x_0 if the frequency-density at x_0-h is the same as the frequency-density at x_0+h , whatever h may be. Figures 8.1a and 8.1b show two symmetrical distributions.

A distribution which is not symmetrical is called *asymmetrical* or *skew*. This skewness is said to be *positive* if the longer tail of the distribution is towards the higher values of the variable (Fig. 8.2a), and *negative* if the longer tail is towards the lower values of the variable (Fig. 8.2b).

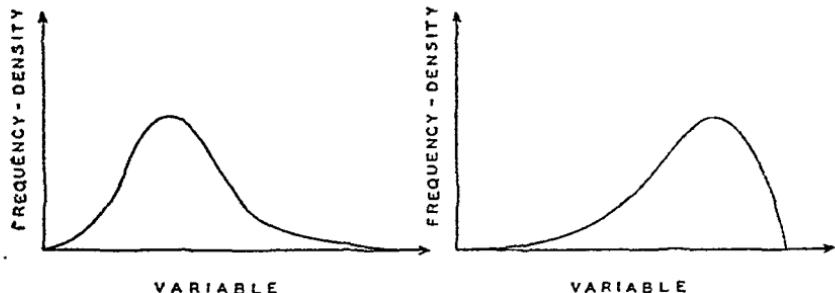


Fig. 8.2a A positively skew distribution.

Fig. 8.2b A negatively skew distribution.

An important point to be noted in this connection is that all odd-order central moments are zero for a symmetrical distribution, positive for a positively skew distribution and negative for a negatively skew distribution. Any such moment may, therefore, be considered a measure of the skewness of a distribution except, of course, m_1 which is necessarily zero for any distribution—symmetrical or skew. The simplest of these measures is m_3 . To make this measure free from the units of the variable, we divide it by s^3 and thus get an *absolute measure** :

$$g_1 = \frac{m_3}{s^3}. \quad \dots \quad (8.10)$$

An alternative measure of skewness is obtained from the relative positions of the mean and the mode in a distribution.

* This and the two subsequent measures involve the assumption that $s > 0$ (i.e. the assumption that x is not a constant in the given set of values.)

In a symmetrical distribution, the mean median and mode (assuming the distribution to be unimodal) coincide. If the distribution is positively skew, then

$$\text{mean} > \text{median} > \text{mode}$$

and if it is negatively skew, then

$$\text{mean} < \text{median} < \text{mode}$$

Hence the difference (mean - mode), divided by the s d , is taken as a measure of skewness

$$Sk = \frac{x - Mo}{s} \quad (8\ 11)$$

Since it is difficult to estimate the mode from a frequency distribution, the empirical relation (6 9) is used to get another measure of skewness, viz

$$Sk = \frac{3(x - M_t)}{s} \quad (8\ 12)$$

A fourth measure of skewness is obtained by considering the relative positions of the three quartiles of a frequency distribution. For a symmetrical distribution the lower and upper quartiles are equidistant from the median , for a positively skew distribution the lower quartile is nearer the median than the upper quartile is, while for a negatively skew distribution the upper quartile is nearer

Thus $(Q_3 - M_t) - (M_t - Q_1)$ may be taken as a measure of skewness It is expressed as a pure number on being divided by

$$(Q_3 - M_t) + (M_t - Q_1) = Q_3 - Q_1 = 2Q,$$

which is assumed to be non zero Thus the new measure is

$$Sk = \frac{(Q_3 - M_t) - (M_t - Q_1)}{2Q} \quad (8\ 13)$$

The measure given by equation (8 10) can theoretically assume any value between $-\infty$ and ∞ , but in practice its numerical value is rarely very high The measure (8 12) can vary between -3 and 3 The same may be said to be approximately the case with (8 11) because of the empirical relation (6 9), which is valid for moderately skew distributions As regards (8 13), it has the limits -1 and 1

Ex. 8.4 We may calculate the various measures of skewness for the frequency distribution of Table 5.10. It has been found already that for this distribution

$$\bar{x} = 164.734 \text{ cm.}, \quad Mo = 164.836 \text{ cm.},$$

$$Mi = 164.758 \text{ cm.}, \quad Q_1 = 160.951 \text{ cm..}$$

and $Q_3 = 168.446 \text{ cm.}, \quad \text{while} \quad s = 5.472 \text{ cm.}$

According to formula (8.11),

$$Sk = \frac{\bar{x} - Mo}{s} = -0.102/5.472 = -0.019.$$

Formula (8.12) gives

$$Sk = \frac{3(\bar{x} - Mi)}{s} = -0.072/5.472 = -0.013.$$

Again, from formula (8.13),

$$Sk = \frac{(Q_3 - Mi) - (Mi - Q_1)}{Q_3 - Q_1} \\ = -0.119/7.495 = -0.016.$$

Now consider the other measure of skewness, viz. (8.10). For the present distribution, $m_3 = 0.08057 \times 5^3$ (cm.)³ and $m_2 = 1.19781 \times 5^2$ (cm.)². Hence

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{0.08057}{(1.19781)^{3/2}}$$

and $\log g_1 = \log 0.08057 - \frac{3}{2} \times \log 1.19781$
 $= 2.7885916 - \log 0.06146.$ Thus $g_1 = 0.061.$

Thus all the measures are nearly equal to zero, indicating that the distribution is almost symmetrical.

Ex. 8.5 The frequency distribution of Table 8.3 may now be considered. Here

$$m_3 = 170.6747 \text{ (md.)}^3 \text{ and } m_2 = 45.3931 \text{ (md.)}^2,$$

so that $g_1 = \frac{m_3}{m_2^{3/2}} = \frac{170.6747}{(45.3931)^{3/2}}$

and $\log g_1 = \log 170.6747 - \frac{3}{2} \times \log 45.3931$
 $= 2.2321699 - \frac{3}{2} \times 1.6569899$
 $= 1.7466851 = \log 0.5581.$

Thus $g_1 = 0.558$, indicating that the distribution is positively skew. This is also apparent from the table itself.

8.6 Kurtosis

Another method of describing a frequency distribution is to specify its degree of peakedness or kurtosis. Two distributions may have the same mean and the same standard deviation and may be equally skew, but one of them may be more peaked than the other.

This feature of the frequency distribution is measured by*

$$g_2 = \frac{m_4}{s^4} - 3 \quad (8.14)$$

Obviously, it is a pure number. For a normal distribution, to be discussed in the next chapter, this measure has the value zero. A positive value of g_2 indicates that the distribution has high concentration of values near the central tendency and has high tails, in comparison with a normal distribution with the same standard deviation. In the same way, a negative value of g_2 means that the

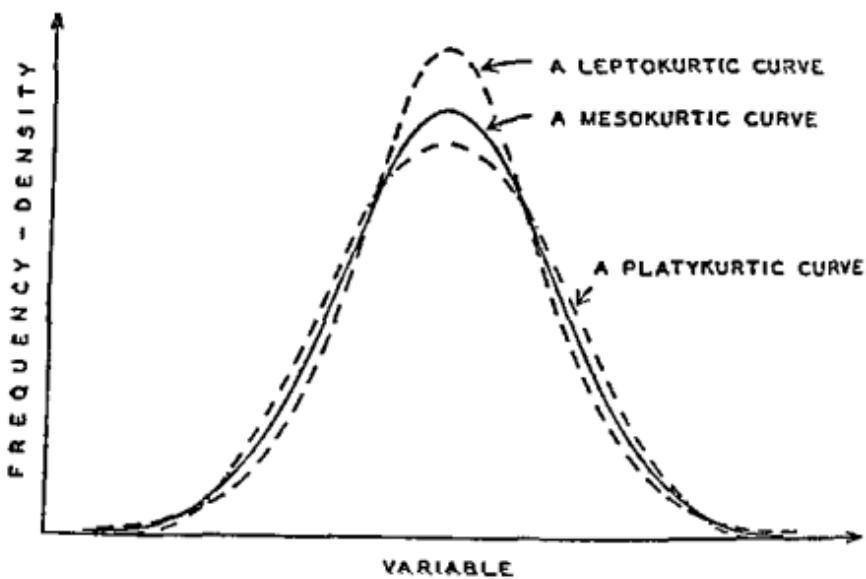


Fig. 8.3 Three symmetrical frequency curves with same mean and s.d. but with different degrees of kurtosis

distribution has low concentration of values in the neighbourhood of the central tendency and low tails, compared to a normal distribution with the same standard deviation. A normal curve is said to be

* This again involves the assumption that $s > 0$

mesokurtic (i.e. having medium kurtosis). A distribution with positive g_2 is called *leptokurtic*, and one with negative g_2 is known as *platykurtic*.

The quantities g_1^2 and $g_2 + 3$ themselves are sometimes used as measures of skewness and kurtosis, respectively. They are referred to as the b_1 and b_2 coefficients. Thus

$$b_1 = \frac{m_3^2}{m_2^3} \quad \dots \quad (8.15)$$

and

$$b_2 = \frac{m_4}{m_2^2}. \quad \dots \quad (8.16)$$

Ex. 8.6 Consider once again the frequency distribution of height for Indian adult males given in Table 5.10. Here

$$m_4 = 5.12575 \times 5^4 \text{ (cm.)}^4$$

$$\text{and } m_2 = 1.19781 \times 5^2 \text{ (cm.)}^2.$$

Therefore,

$$b_2 = \frac{m_4}{m_2^2} = 5.12575 / (1.19781)^2$$

$$\begin{aligned} \text{and } \log b_2 &= \log 5.12575 - 2 \times \log 1.19781 \\ &= 0.5529817 = \log 3.57257. \end{aligned}$$

Thus

$$b_2 = 3.573 \text{ and } g_2 = 0.573,$$

indicating that the distribution is slightly leptokurtic.

Ex. 8.7 For the frequency distribution of Table 8.3, $m_4 = 7,130.620$ (md.)⁴ and $m_2 = 45.393$ (md.)².

Hence

$$b_2 = \frac{m_4}{m_2^2} = \frac{7130.620}{(45.393)^2}$$

$$\begin{aligned} \text{and } \log b_2 &= \log 7130.620 - 2 \times \log 45.393 \\ &= 3.8531273 - 2 \times 1.6569889 \\ &= 0.5391495 = \log 3.4606. \end{aligned}$$

Thus

$$b_2 = 3.461 \text{ and } g_2 = 0.461,$$

which indicate that this distribution also is slightly leptokurtic.

Questions and exercises

8.1 Define the moments of a frequency distribution and explain their usefulness in describing the location and shape of the frequency distribution.

8.2 What are skewness and kurtosis? Give some suitable measures for skewness and kurtosis.

8.3 Using Cauchy-Schwarz inequality, or otherwise, prove that

$$(i) \quad b_2 \geq 1 \quad \text{and} \quad (ii) \quad b_2 - b_1 - 1 \geq 0$$

Discuss in detail the cases where $b_2 = 1$ and $b_2 - b_1 - 1 = 0$.

8.4 Suppose the values of a variable are grouped into a frequency table the width of each class being less than one-third of the standard deviation. Show that Sheppard's correction in such a case will make little difference (a difference of less than 0.5%) in the uncorrected value of the standard deviation.

8.5 Let $S = \sum_i^k |x_i - A| f_i$ be the sum of absolute deviations about the mid point A of the class in which our chosen average m lies. If $d = m - A$, show that the mean deviation about m will be

$$\frac{1}{n} \left[S + (n_1 - n_3)d + n_2 \left(\frac{h}{4} + \frac{d}{h} \right) \right]$$

where n_1 = frequency in the class containing m , h = width of the class-interval, n_1 = total frequency in all lower classes and n_3 = total frequency in all higher classes, provided we assume a uniform distribution of frequency in each class.

8.6 Show that the measure of skewness given by (8.12) must lie between -3 and 3 and that the measure given by (8.13) must lie between -1 and 1.

8.7 Prove by a geometrical argument, that for a J-shaped distribution with its longer tail towards the higher values of the variable, the median is nearer to the 1st quartile than to the 3rd. (A similar argument can be used to show that for the other type of J-shaped distribution, the median is nearer to the 3rd quartile than to the 1st.)

8.8 Consider any symmetrical frequency distribution for a discrete variable. Show that its central moments of odd orders must all be zero.

8.9 Compute \bar{x} , s , m_3 and m_4 for the data on length of earhead given in *Exercise 6.16*.

Ans. 9.9 ; 0.91 ; -0.061 ; 1.56 (in proper units).

8.10 For the frequency distribution of *Exercise 6.18*, compute the mean and the central moments up to that of the fourth order.

Ans. 3.763 ; 3.454 ; -1.736 ; 27.236.

8.11 Determine the mean and the central moments up to that of the fourth order for the frequency distribution of *Exercise 6.19*. For the same distribution compute the various measures of skewness and kurtosis.

Partial ans. $\bar{x}=108.481$; $m_2=297.748$;

$m_3=99.823$; $m_4=375,193.6$.

8.12 Consider the income-distribution of *Exercise 6.20* and measure its skewness by means of an appropriate formula.

Ans. $Sk=-0.202$, by formula (8.13).

8.13 The scores in English of 250 candidates appearing at an examination have

mean = 39.72, $m_2 = 97.80$, $m_3 = -114.18$ and $m_4 = 28,396.14$.

It is later found on scrutiny that the score 61 of a candidate has been wrongly recorded as 51. Make necessary corrections in the given values of the mean and the central moments.

Ans. Correct values are 39.76 ; 99.10 ; -93.27 ; 29,165.60.

8.14 Particulars relating to the wage-distributions of two manufacturing firms are given below :

	<u>Firm A</u>	<u>Firm B</u>
Mean wage	Rs. 277	Rs. 285
Median wage	Rs. 271	Rs. 262
Modal wage	Rs. 260	Rs. 251
Quartiles	Rs. 262 and 278	Rs. 258 and 290
Standard deviation	Rs. 32	Rs. 39

Compare the two distributions.

SUGGESTED READING

- [1] Kenney, J. F. and Keeping, E. S. *Mathematics of Statistics*, Part I (Ch. 7). Van Nostrand, 1954, and Affiliated East-West Press.
- [2] Mills, F. C. *Statistical Methods* (Ch. 5). H. Holt, 1955.
- [3] Yule, G. U. and Kendall, M. G. *Introduction to the Theory of Statistics* (Ch. 6). Charles Griffin, 1953.

9.1 Population and sample

By this time it should be apparent that in a statistical enquiry we are ultimately interested in some numerical characteristics of an aggregate of individuals—individual objects or beings—rather than in the characteristics of the individuals themselves. In statistical language such an aggregate is called a *population* or *universe*. In some cases it may be possible to study each and every member of the population for the purpose of the enquiry. More generally, there will be practical difficulties in studying the whole population. It may be too large for the enquirer, with his limited time and resources, to be able to examine each of its members. It may even be an infinite hypothetical population, e.g. the population of all possible throws with a coin or all possible yields of a crop on a given plot of land. In both cases the enquirer will have to remain content with the information gathered from a part of the population only. Such a part of a population, by means of which one seeks to represent the whole population, is called a *sample*. One may, therefore, say that, as a rule, the data collected in an enquiry relate to a sample, while the ultimate interest of the enquirer lies in a bigger aggregate of individuals—the population.

Among all types of sampling, *random sampling* is generally preferred on some theoretical considerations. Precisely what is meant by 'random sample', how to obtain such a sample—these questions will be dealt with in a later chapter (Chapter 14).

9.2 Theoretical distributions

The main problem with which we are generally concerned is then : how to infer the characteristics of the population from the known characteristics of the sample ? To take a concrete example, we may want to know what the average income of a family in Calcutta is, given the average income per family in a sample of, say, 1,000 families only. In a more general form, our problem is to make

inferences regarding the frequency distribution of some variable x in the population, given the frequency distribution of x in the sample.

To solve such problems it is generally necessary to make some assumptions regarding the form of the population distribution of x .

Our task becomes comparatively simple if the population distribution can be represented by a simple mathematical function of x . In the discrete case we search for a function $f(x)$, called the *probability-mass function* (p.m.f.), which gives for different values of x the corresponding probabilities. (These are, in fact, the relative frequencies of the values in the population, each being equal to the probability that a randomly chosen member of the population will have a particular value of x .) Such a function has to satisfy the conditions :

$$f(x) \geq 0 \text{ for any } x; \quad \dots \quad (9.1)$$

$$\text{and} \quad \sum_x f(x) = 1, \quad \dots \quad (9.2)$$

the sum being taken over all *possible* values of x .

In case x is a continuous variable, one cannot say that there are probabilities attached to particular values of x . One can, instead, assign a probability to an interval of values of x . Here we look for a function $f(x)$, called the *probability-density function* (p.d.f.) of x , which is such that

$$\int_a^b f(x) dx$$

gives $P[a \leq x \leq b]$, whatever a and b may be ($a \leq b$). This has obviously to satisfy the conditions :

$$f(x) \geq 0 \text{ for any } x; \quad \dots \quad (9.3)$$

$$\text{and} \quad \int_{\alpha}^{\beta} f(x) dx = 1, \quad \dots \quad (9.4)$$

(α, β) being the range of possible values of x .

Distributions defined in this way are called *theoretical distributions*, because they are ideal distributions which can hardly be observed in practice. The only purpose that such distributions are meant to serve is to give a fairly close approximation to the actual distribution of the variable in the population.

Some important theoretical distributions, which are useful even for ordinary statistical work, will be discussed in some detail in this chapter.

Like an observed distribution, a theoretical distribution also may be supposed to have measures of central tendency and dispersion, moments and similar other characteristics. If x be a discrete variable, its probabilities being given by the function $f(x)$, then its arithmetic mean will be defined, in analogy with (6.2), as

$$\mu = \sum_x xf(x),$$

the sum being taken over all possible values of x . This is obviously the mathematical expectation of x . We have therefore,

$$\mu = E(x) \quad (9.5)$$

If x be continuous the mean is defined as

$$\mu = \int_a^b xf(x)dx,$$

which is now the expectation of the variable.

The higher moments are defined in a similar manner. The r th moment about an arbitrary origin A is

$$\mu_r = E(x-A)^r, \quad (9.6)$$

while the r th central moment is

$$\mu_r = E(x-\mu)^r \quad (9.7)$$

We can define a function $M_A(t)$, called the moment-generating function (m.g.f.) about A , such that the coefficient of $\frac{t^r}{r!}$ in its expansion is μ_r , the r th moment about A . By definition,

$$\begin{aligned} M_A(t) &= E(e^{t(x-A)}) \\ &= E\{1 + t(x-A) + \frac{t^2}{2!}(x-A)^2 + \dots + \frac{t^r}{r!}(x-A)^r + \dots\} \\ &= 1 + tE(x-A) + \frac{t^2}{2!}E(x-A)^2 + \dots + \frac{t^r}{r!}E(x-A)^r + \dots \\ &= 1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \dots + \frac{t^r}{r!}\mu_r + \dots \end{aligned} \quad (9.6a)$$

Naturally, $M_\mu(t)$ will provide the central moments

The median (say, m_i) of x is taken as the value such that the probability for x to be less than or equal to it and the probability for x to be greater than or equal to it are both equal. In the continuous case, it is given by the equation

$$\int_{\alpha}^{m_i} f(x)dx = \int_{m_i}^{\beta} f(x)dx = \frac{1}{2}. \quad \dots \quad (9.8)$$

The mode for a discrete x is the most probable value. In the continuous case, it is the value with the highest probability density. In either case, we may say that the mode is the value of x for which $f(x)$ is the highest. Clearly, if x is a continuous variable such that the p.d.f. $f(x)$ has derivatives of the first and second orders at the modal value, m_0 , then m_0 can be determined from the relations

$$f'(m_0) = 0, \quad f''(m_0) < 0. \quad \dots \quad (9.9)$$

All these measures have the same properties as their sample counterparts.

9.3 Binomial distribution

Consider a set of trials (by ‘trial’ we mean an attempt to produce a particular event, which is neither certain nor impossible) each of which can result in an event E . The occurrence of E will be referred to as a ‘success’ and its non-occurrence as a ‘failure’. Suppose the trials are independent (i.e., suppose the probability of a success—as well as of a failure—in any trial is not affected by the outcome of any other trial), and the probability of a success in each trial is the same, say p , the probability of a failure being $q (=1-p)$. Such a set of trials is called a set of Bernoullian trials. Thus m tosses of a coin may be looked upon as a set of Bernoullian trials (where by success we may mean the appearance of a head), with $p=\frac{1}{2}$ if the coin is unbiased. Again, consider an urn containing two types of objects, a objects of the first type and b of the second. If m objects are drawn, one by one and with replacements, the objects in the urn being thoroughly mixed before each drawing, then one will again get a set of Bernoullian trials. The appearance of an object of the first type may be called a success, its probability being $p=\frac{a}{a+b}$.

Let us determine the probability of getting x successes in a set of m Bernoullian trials.

The probability of having successes in x trials, and hence failures in the remaining $m-x$ trials, in any preassigned order is

$$p^x q^{m-x},$$

since the trials are independent and in each trial the probability of success is p and the probability of failure is q . Now, x successes and $m-x$ failures may occur in any of $\binom{m}{x}$ orders, $\binom{m}{x}$ being the number of ways in which the x places to be occupied by successes can be chosen from the total number of m places in a sequence. The required probability is, therefore,

$$f(x) = \binom{m}{x} p^x q^{m-x}, \quad (9.10)$$

where x can assume the values

$$0, 1, 2, \dots, m$$

Evidently, $f(r)$ is positive for the given values of x and

$$\begin{aligned} \sum_{x=0}^m f(x) &= \sum_{x=0}^m \binom{m}{x} p^x q^{m-x} \\ &= (q+p)^m = 1 \end{aligned}$$

Thus (9.10) defines a probability distribution, which is known as Bernoulli's distribution or binomial distribution, because $f(r)$ is the $(r+1)$ st term in the expansion of the binomial $(q+p)^m$.

The constants m and p are such that if any of them is changed a new binomial distribution is obtained. Such constants in a probability distribution are called *parameters*. The binomial distribution is thus a biparametric distribution.

9.4 Moments of the binomial distribution

To start with, let us determine the moments about zero of the distribution defined by (9.10). It is seen that

$$\sum_{x=0}^m x(x-1) \dots (x-r+1) f(x) = \sum_{x=r}^m x(x-1) \dots (x-r+1) f(x)$$

(the first r terms being necessarily zero)

$$= \sum_{x=r}^m \frac{m!}{(x-r)!(m-x)!} p^x q^{m-x}$$

$$\begin{aligned}
 &= m(m-1) \dots (m-r+1) p^r \sum_{x'=0}^{n-r} \frac{(m-r)!}{x'!(m-r-x')!} p^{x'} q^{n-r-x'} \\
 &\quad \text{(putting } x' = x-r\text{)} \\
 &= m(m-1) \dots (m-r+1) p^r (q+p)^{n-r} \\
 &= m(m-1) \dots (m-r+1) p^r. \quad \dots \quad (9.11)
 \end{aligned}$$

Hence the 1st moment about zero is

$$\mu'_1 = \sum_{x=0}^m x f(x) = mp.$$

The 2nd moment about zero is

$$\begin{aligned}
 \mu'_2 &= \sum_{x=0}^m x^2 f(x) = \sum_x x(x-1)f(x) + \sum_x xf(x) \\
 &\quad [\text{since } x^2 = x(x-1) + x] \\
 &= m(m-1)p^2 + mp.
 \end{aligned}$$

In the same way, remembering that

$$x^3 = x(x-1)(x-2) + 3x(x-1) + x$$

$$\text{and } x^4 = x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x$$

and using relation (9.11), we have

$$\mu'_3 = m(m-1)(m-2)p^3 + 3m(m-1)p^2 + mp$$

$$\begin{aligned}
 \text{and } \mu'_4 &= m(m-1)(m-2)(m-3)p^4 + 6m(m-1)(m-2)p^3 \\
 &\quad + 7m(m-1)p^2 + mp.
 \end{aligned}$$

Thus the mean of the distribution is

$$\mu = mp. \quad \dots \quad (9.12)$$

By virtue of relations (8.5), which can be shown to be true for moments of a theoretical distribution as well, we have then

$$\begin{aligned}
 \mu_2 &= \mu'_2 - \mu'^2_1 \\
 &= \{m(m-1)p^2 + mp\} - m^2 p^2 \\
 &= mp(1-p) = mpq, \quad \dots \quad (9.13)
 \end{aligned}$$

so that the standard deviation is

$$\sigma = \sqrt{mpq}. \quad \dots \quad (9.14)$$

Similarly

$$\begin{aligned}\mu_3 &= \mu_2 - 3\mu_1\mu_1 + 2\mu_1^3 \\ &= mp(1-p)(1-2p) = npq(q-p)\end{aligned}\quad (9.15)$$

and $\mu_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$

$$\begin{aligned}&= 3m^2p^2(1-p)^2 + mp(1-p)(1-6p(1-p)) \\ &= 3m^2p^2q^2 - npq(1-6pq)\end{aligned}\quad (9.16)$$

The moments can be more easily obtained from the moment generating function. Here

$$\begin{aligned}M_n(t) &= E[e^{tx}] \\ &= \sum_{x=0}^m e^t \binom{m}{x} p^x q^{m-x} \\ &= (pe^t + q)^m \\ &= \{1 + (e^t - 1)p\}^m\end{aligned}$$

It follows that the measures of skewness and kurtosis corresponding to (8.10) and (8.14) for this distribution are

$$\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{mpq}} \quad (9.17)$$

$$\text{and } \gamma_2 = \beta_2 - 3 = \frac{1-6pq}{mpq} \quad (9.18)$$

The expression for γ_1 shows that the distribution is positively skewed, negatively skewed or symmetrical according as p is less than greater than or equal to q , i.e. according as p is less than greater than or equal to $\frac{1}{2}$.

9.5 A recursion relation concerning moments of the binomial distribution

The calculation of moments of the binomial distribution is facilitated by using a recursion relation.

From (9.12), we have

$$\mu = mp$$

By definition,

$$\begin{aligned}\mu_x &= F(x-\mu)^r \\ &= \sum_{x=0}^m (x-mp)^r \binom{m}{x} p^x q^{m-x}\end{aligned}$$

Differentiating this with respect to p , we get

$$\begin{aligned}\frac{d\mu_r}{dp} &= -rm \sum_{x=0}^m (x - mp)^{r-1} \binom{m}{x} p^x q^{m-x} \\ &\quad + \sum_{x=0}^m (x - mp)^r \binom{m}{x} x p^{x-1} q^{m-x} - \sum_{x=0}^m (x - mp)^r \binom{m}{x} (m-x) p^x q^{m-x-1} \\ &= -rm\mu_{r-1} + \sum_{x=0}^m (x - mp)^r \binom{m}{x} p^x q^{m-x} \left(\frac{x}{p} - \frac{m-x}{q} \right) \\ &= -rm\mu_{r-1} + \frac{1}{pq} \sum_{x=0}^m (x - mp)^{r+1} \binom{m}{x} p^x q^{m-x} \\ &= -rm\mu_{r-1} + \frac{\mu_{r+1}}{pq}\end{aligned}$$

or

$$\mu_{r+1} = pq \left(rm\mu_{r-1} + \frac{d\mu_r}{dp} \right). \quad \dots \quad (9.19)$$

Knowing that $\mu_0 = 1$, $\mu_1 = 0$ and using relation (9.19), one can quite easily calculate all the higher order moments of the binomial distribution.

9.6 Fitting a binomial distribution to an observed distribution

When we fit a theoretical distribution to a given observed distribution, our purpose is to examine whether the observed distribution may be regarded as the distribution in a random sample from the population characterised by the theoretical distribution. For this it is necessary to calculate, corresponding to the observed class-frequencies, the frequencies that are to be expected in case the given observed distribution is really a random sample from the assumed population. The first step in fitting a distribution is, therefore, to estimate the parameters of the theoretical distribution from the observed data, unless the values of the parameters are assumed or are known *a priori*.

A particularly simple method of estimating parameters is the *method of moments*. Supposing there are k parameters to be estimated, the method consists of the following steps :

- (1) To express the moments of the theoretical distribution in terms of the parameters.
- (2) To equate the first k theoretical moments, expressed in terms of the parameters, to the corresponding moments of the observed distribution.

(3) Finally, to solve the k resulting equations to determine the k parameters

The first k moments are taken because the errors in a moment due to sampling increase with the order of the moment

In a binomial distribution the first moment about zero is

$$\mu = mp,$$

while the first moment about zero of the observed distribution is x . Equating the two, we get

$$mp = x,$$

so that the estimate of p is

$$p = \frac{x}{m} \quad (9.20)$$

The estimated expected frequencies will, therefore, be

$$n \times f(x) = n \times \binom{m}{x} p^x (1-p)^{m-x}, \quad (9.21)$$

for $x = 0, 1, 2, \dots, m$,

where n is the total frequency, i.e. the total number of sets of m trials each

Ex 9.1 12 dice were thrown 2,630 times and each time the number of dice which had 5 or 6 on the uppermost faces was recorded. The results are given in the following table

Number of dice with 5 or 6 uppermost	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	18	115	326	548	611	519	307	133	40	11	2	0	0

Graduate the observed distribution (1) with a binomial distribution for which p is unknown (2) with a binomial distribution for which $p = \frac{1}{2}$.

Case 1 Here, to fit a binomial distribution, p has to be estimated from the observed distribution. The mean of the latter distribution is

$$x = \frac{\sum x f_x}{n} = \frac{10\ 662}{2,630} = 4.05399,$$

so the estimate of p is

$$p = \frac{4.05399}{12} = 0.33783$$

The probabilities $f(x)$ are calculated by using the relation

$$f(x) = \left\{ \frac{m-x+1}{x} \cdot \frac{\hat{p}}{\hat{q}} \right\} f(x-1).$$

Here $f(0) = \hat{q}^n$,

or $\log f(0) = m \log \hat{q} = 12 \log (0.66217)$

$$= 3.8516340 = \log (0.0071061),$$

so that $f(0) = 0.0071061$.

Also, $\frac{\hat{p}}{\hat{q}} = 0.51019$.

Subsequent calculations are shown in the following table :

TABLE 9.1

FITTING A BINOMIAL DISTRIBUTION TO THE FREQUENCY DISTRIBUTION OF NUMBER OF DICE WITH 5 OR 6 UPPERMOST IN 2,630 THROWS OF 12 DICE (p ESTIMATED FROM DATA)

x (1)	$\frac{m-x+1}{x}$ (2)	$\frac{m-x+1}{x}$ $\frac{\hat{p}}{\hat{q}}$ (3)	$f(x) = f(x-1)$ \times col. (3) (4)	Expected frequency $= n \times f(x)$ (5)	Observed frequency (6)
0	—	—	0.0071061	18.69	18
1	12	6.12228	0.0435055	114.42	115
2	5.5	2.80604	0.1420782	321.07	326
3	3.33333	1.70053	0.2076098	546.01	548
4	2.25	1.14793	0.2383215	626.79	611
5	1.6	0.81630	0.1945418	511.64	519
6	1.16667	0.59522	0.1157952	304.54	307
7	0.85714	0.43730	0.0506372	133.18	133
8	0.625	0.31887	0.0161467	42.47	40
9	0.44444	0.22675	0.0036613	9.63	11
10	0.3	0.15306	0.0005604	1.47	2
11, 12	—	—	0.0000363*	0.09	0
Total	—	—	1.0000000*	2,630.00	2,630

*Obtained from the identity : $f(11) + f(12) = 1 - \sum_{x=0}^{10} f(x)$.

Case 2 Here the procedure is the same as in *Case 1*, but for p we now use its given value, $1/3$. So

$$f(0) = q^m = (2/3)^{12},$$

or $\log f(0) = 12(\log 2 - \log 3)$

$$= 3.8869044 = \log (0.0077073),$$

giving $f(0) = 0.0077073$,

and $\frac{p}{q} = \frac{1}{2}$

TABLE 92

FITTING A BINOMIAL DISTRIBUTION TO THE FREQUENCY DISTRIBUTION OF NUMBER OF DICE WITH 5 OR 6 UPPERMOST IN 2,630 THROWS OF 12 DICE ($p = \frac{1}{3}$)

x (1)	$\frac{m-x+1}{x}$ (2)	$\frac{m-x+1}{x} p$ $\frac{1}{x} \cdot \frac{1}{3}$ (3)	$f(x) = f(x-1) \times \text{col (3)}$ (4)	Expected frequency $= n \times f(x)$ (5)	Observed frequency (6)
0	—	—	0.0077073	20.27	18
1	12	6.00000	0.0462438	121.62	115
2	5.5	2.75000	0.1271704	334.46	326
3	3.33333	1.66667	0.2119511	557.43	548
4	2.25	1.12500	0.2384450	627.11	611
5	1.6	0.80000	0.1907560	501.69	519
6	1.16667	0.58333	0.1112737	292.65	307
7	0.85714	0.42857	0.0476886	125.42	133
8	0.625	0.31250	0.0149027	39.19	40
9	0.44444	0.22222	0.0033117	8.71	11
10	0.3	0.15000	0.0004968	1.31	2
11, 12	—	—	0.0000529*	0.14	0
Total	—	—	1.0000000	2.630.00	2.630

*Obtained from the identity $f(11) + f(12) = 1 - \sum_{x=0}^{10} f(x)$

9.7 Poisson distribution

Let the variable x have the probability distribution defined by the p.m.f.

$$f(x) = \frac{\exp(-\lambda) \cdot \lambda^x}{x!}, \text{ where } x=0, 1, 2, \dots, \text{ ad inf.} \dots \quad (9.22)$$

and $\lambda > 0$.

It is readily seen that $f(x) > 0$ for all possible values of x and

$$\sum_{x=0}^{\infty} f(x) = \exp(-\lambda) \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \exp(-\lambda) \cdot \exp(\lambda) = 1.$$

The probability distribution defined by (9.22) is called a Poisson distribution. It may be looked upon as a limiting form of a binomial distribution. Suppose in a binomial distribution $m \rightarrow \infty$ and $p \rightarrow 0$, $mp = \lambda$ (say) remaining finite. In that case

$$\begin{aligned} & \lim \binom{m}{x} p^x q^{m-x} \\ &= \lim \frac{m(m-1)(m-2)\dots(m-x+1)}{x!} p^x (1-p)^{m-x} \\ &= \lim \frac{1 \left(1 - \frac{1}{m}\right) \left(1 - \frac{2}{m}\right) \dots \left(1 - \frac{x-1}{m}\right)}{x!} (mp)^x \left(1 - \frac{mp}{m}\right)^{m-x} \\ &= \frac{\lambda^x}{x!} \lim \left\{ 1 \left(1 - \frac{1}{m}\right) \left(1 - \frac{2}{m}\right) \dots \left(1 - \frac{x-1}{m}\right) \right\} \lim \left(1 - \frac{\lambda}{m}\right)^m \lim \frac{1}{\left(1 - \frac{\lambda}{m}\right)^x} \\ &= \frac{\exp(-\lambda) \cdot \lambda^x}{x!}, \end{aligned}$$

since

$$\lim_{m \rightarrow \infty} 1 \left(1 - \frac{1}{m}\right) \left(1 - \frac{2}{m}\right) \dots \left(1 - \frac{x-1}{m}\right) = 1,$$

$$\lim_{m \rightarrow \infty} \left(1 - \frac{\lambda}{m}\right)^m = \exp(-\lambda),$$

and

$$\lim_{m \rightarrow \infty} \left(1 - \frac{\lambda}{m}\right)^x = 1,$$

for fixed x . Thus we may say that as $m \rightarrow \infty$ and $p \rightarrow 0$, mp remaining finite ($=\lambda$), the binomial distribution defined by (9.10) tends to the Poisson distribution defined by (9.22).

The practical implication of this result is that the binomial probabilities $\binom{m}{x} p^x q^{m-x}$ may be well approximated with the much simpler quantities $\frac{\exp(-\lambda)}{x!} \lambda^x$ by putting $\lambda=mp$, provided m is sufficiently large and p is sufficiently small, mp being of moderate size.

Ex 9.2 Suppose 2 six-faced dice are thrown 100 times. What is the probability of getting a double six in at least 2 of the throws?

The probability of getting a double six in each of x throws is the binomial probability

$$f(x) = \binom{100}{x} \left(\frac{1}{36}\right)^x \left(\frac{35}{36}\right)^{100-x},$$

assuming both the dice are perfect.

The probability of getting a double six in at least 2 throws each is

$$1 - f(0) - f(1) = 1 - \left(\frac{35}{36}\right)^{100} - 100 \left(\frac{1}{36}\right) \left(\frac{35}{36}\right)^{99}$$

Now,

$$\log 35 = 1.5440680$$

and

$$\log 36 = 1.5563025$$

Hence

$$\begin{aligned} \log \left(\frac{35}{36}\right)^{100} &= 100(\log 35 - \log 36) \\ &= 2.776550 = \log(0.05978), \end{aligned}$$

so that

$$\left(\frac{35}{36}\right)^{100} = 0.05978$$

Again,

$$\begin{aligned} \log \frac{1}{36} \left(\frac{35}{36}\right)^{99} &= 99 \log 35 - 100 \log 36 \\ &= 3.2324820 = \log(0.0017080) \end{aligned}$$

Therefore,

$$\frac{1}{36} \left(\frac{35}{36}\right)^{99} = 0.0017080$$

The required probability is thus

$$1 - 0.05978 - 0.0017080 = 0.76942$$

Using the Poisson approximation with $\lambda = \frac{100}{36} = 2.7778$, we have

$$f(0) = \exp[-\lambda] \text{ and } f(1) = \exp[-\lambda] \lambda$$

Now,

$$\begin{aligned} \log \exp[-\lambda] &= -2.7778 \log e = -2.7778 \times 4.342945 \\ &= 2.7936167 = \log 0.062175 \end{aligned}$$

Hence

$$\exp[-\lambda] = 0.062175 \text{ and } \exp[-\lambda] \lambda = 0.172710$$

The required probability is then given as

$$1 - 0.062175 - 0.172710 = 0.76512.$$

This agrees with the exact probability up to two significant places.

We have considered the Poisson distribution as a limiting form of the binomial distribution. There are cases where the trials may be supposed to be of the Bernoullian type and the number of trials in each set, m , is known to be large, although its exact value is unknown. In such cases one cannot use the binomial distribution, but its approximation, the Poisson distribution, may still be employed :

1. Suppose records are kept of the number of calls made from a public telephone booth in a city for every one-hour interval during the busy hours of the day (say, from 10 a.m. to 5 p.m.). Here each interval may be looked upon as being composed of sub-intervals, each so small that within such a sub-interval, one call and no more may be made. Each interval then represents a set of trials, each trial giving a 'success' (when a call is made in the corresponding sub-interval) or a 'failure' (when no call is made within the sub-interval). Further, the calls may be supposed to be made independently of each other. Hence the number of calls made per one-hour interval may be supposed to be distributed binomially with parameters m (number of sub-intervals within an interval) and p (probability of a call being made in a sub-interval), p being supposed to be constant for the busy hours of the day. Here m is, of course, unknown but may be supposed to be very large. As regards p , it may be assumed to be quite small in this case. Hence the observed data are expected to agree fairly well with a Poisson distribution, whose parameter λ may be estimated with the observed mean number of calls per one-hour interval.

2. In the course of an experiment, 529 yeast cells were thoroughly mixed in a liquid. Next, the mixture was poured into a haemacytometer ruled into 400 small squares, which was then placed under a microscope, and the number of yeast cells falling in each square was counted. In this way a frequency distribution of the number of yeast cells per square of the haemacytometer was obtained.

Here, again, the mixture in each square might be conceived as being composed of a large number of groups of molecules, each of the

same size as an yeast cell. Each such group of molecules might be said to give a 'success' if it was an yeast cell and a 'failure' otherwise. Since the dilution was thoroughly mixed, the 'trials' represented by these groups might be supposed to be independent, the probability of a success being the same for each trial. The trials might thus be considered to be approximately of the Bernoullian type. Further, here the number of trials in each set (i.e. the number of groups in each square) might be supposed to be large, with p , the probability of a trial giving a success, being quite small. Hence it would be reasonable to expect that a Poisson distribution with λ estimated by $529/400=1.3225$ would give a very good fit to the observed distribution. This was, in fact, found to be the case.

Other variables for which the Poisson distribution is expected to be the appropriate theoretical distribution are the number of motor vehicles passing through a street-crossing per, say, $\frac{1}{2}$ minute or 1-minute interval during the busy hours of the day, the number of printing mistakes per page of one of the early proofs of a book, and so forth.

9.8 Moments of the Poisson distribution

For the Poisson distribution, we have

$$\begin{aligned} \sum_{x=0}^{\infty} x(x-1) & (x-r+1)f(x) = \sum_{x=r}^{\infty} x(x-1) & (x-r+1)f(x) \\ & = \sum_{x=r}^{\infty} \frac{\exp[-\lambda]\lambda^x}{(x-r)!} = \exp[-\lambda] \sum_{x=0}^{\infty} \frac{\lambda^{x+r}}{x!} & (\text{putting } x=x-r) \\ & = \lambda^r \exp[-\lambda] \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \lambda^r \exp[-\lambda] \exp[\lambda] = \lambda^r \end{aligned} \quad (9.23)$$

Hence the 1st moment about zero is

$$\mu_1 = \sum_{x=0}^{\infty} xf(x) = \lambda$$

The 2nd moment about zero is

$$\begin{aligned} \mu_2 & = \sum_{x=0}^{\infty} x^2 f(x) = \sum_x x(x-1)f(x) + \sum_x xf(x) \\ & = \lambda^2 + \lambda & [\text{since } x^2 = x(x-1) + x] \end{aligned}$$

In the same way, putting

$$x^3 = x(x-1)(x-2) + 3x(x-1) + x$$

and $x^4 = x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x$

and using equation (9.23), we find

$$\mu'_3 = \lambda^3 + 3\lambda^2 + \lambda$$

and $\mu'_4 = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda.$

Thus the mean of the distribution is

$$\mu = \lambda. \quad \dots \quad (9.24)$$

As to the central moments, we have

$$\mu_2 = \mu'_2 - \mu'^2_1 = (\lambda^2 + \lambda) - \lambda^2 = \lambda, \quad \dots \quad (9.25)$$

so that

$$\sigma = \sqrt{\lambda}. \quad \dots \quad (9.26)$$

Similarly,

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 = \lambda \quad \dots \quad (9.27)$$

and $\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 = 3\lambda^2 + \lambda. \quad \dots \quad (9.28)$

Thus it is seen that for the Poisson distribution with parameter λ , the mean, the variance and the third central moment are all equal to λ .

The moment-generating function of the distribution is

$$\begin{aligned} M_0(t) &= E(e^{tx}) \\ &= \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \exp[e^t \cdot \lambda] \\ &= \exp[\lambda(e^t - 1)]. \end{aligned} \quad \dots \quad (9.29)$$

Again, for this distribution

$$\gamma_1 = \sqrt{\bar{\beta}_1} = \frac{1}{\sqrt{\lambda}} \quad \dots \quad (9.29a)$$

and $\gamma_2 = \beta_2 - 3 = 1/\lambda. \quad \dots \quad (9.29b)$

The distribution is thus seen to be positively skew and leptokurtic.

9.9 A recursion relation concerning moments of the Poisson distribution

Like the moments of the binomial distribution, the moments of the Poisson distribution also satisfy a recursion relation.

Here we have

$$\mu_r = \sum_{x=0}^{\infty} (x - \lambda)^r \exp[-\lambda] \frac{\lambda^x}{x!}.$$

Differentiating with respect to λ , we have

$$\begin{aligned}
 \frac{d\mu_r}{d\lambda} &= -r \sum_{x=0}^{\infty} (x-\lambda)^{r-1} \exp[-\lambda] \frac{\lambda^x}{x!} - \sum_{x=0}^{\infty} (x-\lambda)^r \exp[-\lambda] \frac{\lambda^x}{x!} \\
 &\quad + \sum_{x=0}^{\infty} (x-\lambda)^r \exp[-\lambda] \frac{x\lambda^{x-1}}{x!} \\
 &= -r\mu_{r-1} + \sum_{x=0}^{\infty} (x-\lambda)^r \exp[-\lambda] \frac{\lambda^x}{x!} \left(\frac{x}{\lambda} - 1 \right) \\
 &= -r\mu_{r-1} + \frac{1}{\lambda} \sum_{x=0}^{\infty} (x-\lambda)^{r+1} \exp[-\lambda] \frac{\lambda^x}{x!} \\
 &= -r\mu_{r-1} + \frac{\mu_{r+1}}{\lambda} \\
 \text{or } \mu_{r+1} &= \lambda \left(r\mu_{r-1} + \frac{d\mu_r}{d\lambda} \right) \tag{9.30}
 \end{aligned}$$

Substituting $\mu_0 = 1$ and $\mu_1 = 0$ in (9.30) and putting $r = 1, 2, 3$, successively, one can obtain the higher order moments

9.10 Fitting a Poisson distribution to an observed distribution

The Poisson distribution has only one parameter λ , which can be estimated from the observed data by the method of moments. The first moment of the Poisson distribution about zero is

$$\mu = \lambda,$$

while the first moment about zero of the observed distribution is x . Equating these two, an estimate of λ is obtained, viz.

$$\hat{\lambda} = x$$

The expected frequencies corresponding to the observed frequencies f_x will then be obtained as

$$n \times f(x) = n \times \frac{\exp[-\hat{\lambda}] \hat{\lambda}^x}{x!}, \quad \text{for } x=0, 1, 2, \tag{9.31}$$

Ex. 9.3 In a textile factory, 100 pieces of cloth were inspected and the number of defects in each piece was recorded. In this way the following frequency distribution was obtained

Number of defects	0	1	2	3	4 or more
Frequency	79	18	2	1	0

In order to fit a Poisson distribution to the above data, we first have, as an estimate $\hat{\lambda}$ of the parameter λ , the observed mean

$$\bar{x} = \frac{\sum xf_x}{n} = \frac{0 \times 79 + 1 \times 18 + 2 \times 2 + 3 \times 1}{100} = 0.25.$$

The expected frequencies are then calculated from the formula :

$$n \times f(x) = 100 \times \frac{\exp[-0.25](0.25)^x}{x!}.$$

We first compute $f(0) = \exp[-\hat{\lambda}]$. Here

$$\begin{aligned}\log f(0) &= -0.25 \log e = -0.25 \times 0.4342945 \\ &= -1.8914264 = \log(0.77880).\end{aligned}$$

Hence

$$f(0) = 0.77880.$$

The other values are obtained successively from the relations

$$f(1) = f(0) \cdot \frac{\hat{\lambda}}{1}, \quad f(2) = f(1) \cdot \frac{\hat{\lambda}}{2}, \quad f(3) = f(2) \cdot \frac{\hat{\lambda}}{3},$$

and so forth.

These values, together with the expected frequencies, are shown in the table below :

TABLE 9.3
FITTING A POISSON DISTRIBUTION TO THE FREQUENCY
DISTRIBUTION OF NUMBER OF DEFECTS PER PIECE
FOR 100 PIECES OF CLOTH

Number of defects per piece x (1)	$\hat{\lambda}/x$ (2)	$f(x) = f(x-1) \times \hat{\lambda}/x$ (3)	Expected frequency $= n \times f(x)$ (4)	Observed frequency (5)
0	—	0.77880	77.88	79
1	0.25	0.19470	19.47	18
2	0.125	0.02434	2.44	2
3	0.08333	0.00203	0.20	1
4 or more	—	0.00013*	0.01	0
Total	—	1.00000	100.00	100

*Obtained from the identity : $\sum_{x=4}^{\infty} f(x) = 1 - \sum_{x=0}^3 f(x).$

9.11 Hypergeometric distribution

We know that we have a set of Bernoullian trials if individuals are drawn at random and *with* replacements from a population of N individuals, of which Np individuals possess a certain characteristic C and the remaining Nq ($q=1-p$) individuals do not possess C . If, however, the individuals are drawn at random and *without* replacements, then the trials are not of the Bernoullian type and the probability that, in a sample of size m , x individuals will possess the characteristic C and $m-x$ will not possess C is given by

$$\begin{aligned} f(x) &= \binom{Np}{x} \binom{Nq}{m-x} / \binom{N}{m} \\ &= \binom{m}{x} \frac{(Np)_x (Nq)_{m-x}}{(N)_m}, \quad x=0, 1, 2, \dots, m, \end{aligned} \quad (9.32)$$

since there are $\binom{N}{m}$ possible ways of selecting the sample and x C 's and $m-x$ not- C 's may be obtained in $\binom{Np}{x} \binom{Nq}{m-x}$ ways.

Obviously, $\sum_{x=0}^m f(x)=1$ since $\sum_{x=0}^m \binom{Np}{x} \binom{Nq}{m-x}$, being the coefficient of t^m in the expansion of $(1+t)^{Np}(1+t)^{Nq}=(1+t)^N$, equals $\binom{N}{m}$. Again, (9.32) may be written in the form

$$\frac{(Nq)_m}{(N)_m} \times \left(\frac{(Np)_x (m)_x}{(Nq-m+x)_x x!} \right) \quad (9.33)$$

The second factor of (9.33), it may be seen, is the coefficient of t^x in the expansion of the hypergeometric

$$F(\alpha, \beta, \gamma, t) = 1 + \frac{\alpha}{\gamma} \frac{t}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)} \frac{t^2}{2!} + \dots, \quad (9.34)$$

If we substitute

$$\alpha = -m, \beta = -Np \text{ and } \gamma = Nq - m + 1$$

Thus (9.33) is the coefficient of t^x in the expansion of

$$\frac{(Nq)_m}{(N)_m} F(-m, -Np, Nq-m+1, t) \quad (9.35)$$

Hence the distribution is called a *hypergeometric* distribution with parameters N , p and m .

The mean, μ , of this distribution is

$$\begin{aligned}
 \mu &= E(x) = \sum_{x=0}^m x \cdot \frac{m!}{x!(m-x)!} \cdot \frac{(\mathcal{N}p)_x (\mathcal{N}q)_{m-x}}{(\mathcal{N})_m} \\
 &= \sum_{x=1}^m x \cdot \frac{m!}{x!(m-x)!} \cdot \frac{(\mathcal{N}p)_x (\mathcal{N}q)_{m-x}}{(\mathcal{N})_m} \\
 &= \frac{m \cdot \mathcal{N}p}{\mathcal{N}} \sum_{x=1}^m \frac{(m-1)!}{(x-1)!(m-1-x-1)!} \cdot \frac{(\mathcal{N}p-1)_{x-1} (\mathcal{N}q)_{(m-1)-(x-1)}}{(\mathcal{N}-1)_{m-1}} \\
 &= mp \sum_{x'=0}^{m-1} \frac{(m-1)!}{x'!(m-1-x')!} \cdot \frac{(\mathcal{N}p-1)_{x'} (\mathcal{N}q)_{m-1-x'}}{(\mathcal{N}-1)_{m-1}} \\
 &\quad \text{[where } x' = x-1] \\
 &= mp. \tag{9.36}
 \end{aligned}$$

Proceeding similarly, we can show that

$$E\{x(x-1)\dots(x-r+1)\} = \frac{(m)_r (\mathcal{N}p)_r}{(\mathcal{N})_r}. \tag{9.37}$$

We can hence calculate the raw moments and central moments. It can be verified that

$$\mu_2 = mpq \frac{\mathcal{N}-m}{\mathcal{N}-1}, \tag{9.38}$$

$$\mu_3 = mpq(q-p) \frac{(\mathcal{N}-m)(\mathcal{N}-2m)}{(\mathcal{N}-1)(\mathcal{N}-2)}, \tag{9.39}$$

$$\begin{aligned}
 \text{and } \mu_4 &= \frac{mpq(\mathcal{N}-m)}{(\mathcal{N}-1)(\mathcal{N}-2)(\mathcal{N}-3)} [\mathcal{N}(\mathcal{N}+1)-6m(\mathcal{N}-m) \\
 &\quad + 3pq\{\mathcal{N}^2(m-2)-\mathcal{N}m^2+6m(\mathcal{N}-m)\}]. \tag{9.40}
 \end{aligned}$$

As expected, when $\mathcal{N} \rightarrow \infty$, the hypergeometric distribution tends to the binomial, since in case the number of individuals in the population is indefinitely large, it is immaterial whether individuals are drawn with or without replacements. Here we have

$$\begin{aligned}
 &\binom{m}{x} \frac{(\mathcal{N}p)_x (\mathcal{N}q)_{m-x}}{(\mathcal{N})_m} \\
 &= \frac{\binom{m}{x} p \left(p - \frac{1}{\mathcal{N}}\right) \dots \left(p - \frac{x-1}{\mathcal{N}}\right) q \left(q - \frac{1}{\mathcal{N}}\right) \dots \left(q - \frac{m-x-1}{\mathcal{N}}\right)}{\left(1 - \frac{1}{\mathcal{N}}\right) \left(1 - \frac{2}{\mathcal{N}}\right) \dots \left(1 - \frac{m-1}{\mathcal{N}}\right)} \\
 &\rightarrow \binom{m}{x} p^x q^{m-x} \quad \text{as } \mathcal{N} \rightarrow \infty.
 \end{aligned}$$

9.12 Negative binomial distribution

Consider an indefinite series of Bernoullian trials. Suppose p denotes the probability of the occurrence of an event E (called a 'success') in a trial and $q(=1-p)$ denotes the probability of its non-occurrence (or of a 'failure'). Let the trials be repeated until the event E occurs r times. The probability that at least m trials will be necessary to produce the event E r times is

$$\begin{aligned}
 & = \text{probability that } E \text{ occurs } r-1 \text{ times in the first } m-1 \text{ trials} \\
 & \quad \times \text{probability that } E \text{ occurs in the } m\text{th trial} \\
 & = \binom{m-1}{r-1} p^{r-1} q^{m-r} \times p \\
 & = \binom{m-1}{r-1} p^r q^{m-r}, \quad \text{for } m=r, r+1, \dots \quad \text{ad inf} \tag{9.41}
 \end{aligned}$$

If x denotes the number of failures preceding the r th success, naturally $m=x+r$. Thus the probability $f(x)$ that there are x failures preceding the r th success may be obtained by substituting $m=x+r$ in (9.41). Thus

$$\begin{aligned}
 f(x) & = p^r \binom{r+x-1}{r-1} q^x \\
 & = p^r \frac{(r+x-1)(r+x-2)}{x!} \dots r q^x, \quad \text{for } x=0, 1, 2, \dots \quad \text{ad inf}, \tag{9.42}
 \end{aligned}$$

which is the $(x+1)$ st term in the expansion of

$$p^r (1-q)^{-r}, \tag{9.43}$$

a binomial with a negative index. Hence the distribution of x is known as a *negative binomial distribution*. Obviously,

$$\begin{aligned}
 \sum_{x=0}^{\infty} f(x) & = p^r \sum_{x=0}^{\infty} \frac{(r+x-1)(r+x-2)}{x!} \dots r q^x \\
 & = p^r (1-q)^{-r} = 1
 \end{aligned}$$

The mean, μ , of the distribution is

$$\begin{aligned}
 \mu & = E(x) = p^r \sum_{x=0}^{\infty} x \binom{r+x-1}{r-1} q^x \\
 & = p^r \sum_{x=0}^{\infty} x \frac{(r+x-1)(r+x-2)}{x!} \dots r q^x \\
 & = rp^r q \sum_{x=1}^{\infty} \frac{(r+x-1)(r+x-2)}{(x-1)!} \frac{(r+1)}{x} q^{x-1}
 \end{aligned}$$

$$\begin{aligned}
 &= rp^r q \sum_{x'=0}^{\infty} \frac{(r+x')(r+x'-1)\dots(r+1)}{x'!} q^{x'}, \\
 &= rp^r q (1-q)^{-(r+1)} \quad \text{(putting } x'=x-1\text{)} \\
 &= r \frac{q}{p}. \quad \dots \quad (9.44)
 \end{aligned}$$

Similarly, $E\{x(x-1)\dots(x-k+1)\}=r(r+1)\dots(r+k-1)(q^k/p^k)$.

$$\begin{aligned}
 \text{Thus } \mu_2 &= E(x^2) - \{E(x)\}^2 = E\{x(x-1)\} + E(x) - \{E(x)\}^2 \\
 &= r(r+1) \frac{q^2}{p^2} + r \frac{q}{p} - r^2 \frac{q^2}{p^2} \\
 &= \frac{rq(q+p)}{p^2} = \frac{rq}{p^2}. \quad \dots \quad (9.45)
 \end{aligned}$$

Similarly, we can show that

$$\mu_3 = \frac{rq(1+q)}{p^3} \quad \dots \quad (9.46)$$

$$\text{and } \mu_4 = \frac{rq(1+4q+q^2)+3r^2q^2}{p^4}, \quad \dots \quad (9.47)$$

$$\text{so that } \gamma_1 = \frac{1+q}{\sqrt{rq}} \quad \dots \quad (9.48)$$

$$\text{and } \gamma_2 = \frac{1+4q+q^2}{rq}. \quad \dots \quad (9.49)$$

Thus the distribution is seen to be positively skew and leptokurtic.

9.13 Rectangular (or uniform) distribution

The theoretical distributions that we have considered so far are all meant for discrete variables. The rest of this chapter will be devoted to theoretical distributions of the continuous type. The simplest distribution of this group is the rectangular distribution, which has got equal probability-densities for all values throughout the range of the continuous variable x . The probability-density function, $f(x)$, is defined by

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b. \quad \dots \quad (9.50)$$

Clearly, $f(x) > 0$ for $a \leq x \leq b$,

$$\text{and } \int_a^b f(x) dx = \int_a^b \frac{dx}{b-a} = 1.$$

The mean, μ , of the distribution is

$$\mu = \int_a^b xf(x)dx = \int_a^b \frac{xdx}{b-a} = \frac{1}{2} \frac{b^2 - a^2}{b-a} = \frac{b+a}{2} \quad (9.51)$$

The r th moment about the start of the curve, a , is given by

$$\begin{aligned} \mu_r &= \int_a^b (x-a)^r f(x)dx = \int_a^b \frac{(x-a)^r dx}{b-a} = \frac{(b-a)^{r+1}}{(r+1)(b-a)} \\ &= (b-a)^r / (r+1) \end{aligned} \quad (9.52)$$

So the variance, σ^2 , is given by

$$\begin{aligned} \sigma^2 &= \mu_2 - \mu_1^2 \\ &= (b-a)^2/3 - (b-a)^2/4 = \frac{(b-a)^2}{12} \end{aligned} \quad (9.53)$$

Similarly, we have

$$\mu_3 = 0 \quad \text{and} \quad \mu_4 = (b-a)^4/80 \quad (9.54)$$

$$\text{Thus} \quad \gamma_1 = 0 \quad \text{and} \quad \gamma_2 = -1/2 \quad (9.55)$$

Obviously, the distribution is symmetrical and highly platykurtic

9.14 Normal distribution

Of all theoretical distributions for continuous variables, the most important is the so called normal or Gaussian distribution

The distribution is defined by the probability density function

$$f(x) = \frac{h}{\sqrt{\pi}} \exp[-h^2(x-a)^2], \quad -\infty < x < \infty, \quad (9.56)$$

where it is assumed that $h > 0$

Clearly, $f(x)$ is positive for all values of x . Further, using the improper integral

$$\int_0^\infty \exp[-y^2] dy = \frac{\sqrt{\pi}}{2}$$

we have

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \frac{2h}{\sqrt{\pi}} \int_a^{\infty} \exp[-h^2(x-a)^2] dx \quad (\text{since the distribution is symmetrical about } a) \\ &= \frac{2h}{h\sqrt{\pi}} \int_0^{\infty} \exp[-z^2] dz \quad [\text{putting } h(x-a)=z] \\ &= 1 \end{aligned}$$

9.15 Properties of the normal distribution

The more important properties of the distribution are as follows :

(1) From (9.56) it is seen that the distribution is symmetrical about the point $x=a$, since

$$f(a+u)=f(a-u)=\frac{h}{\sqrt{\pi}}\exp[-h^2u^2],$$

whatever u may be.

(2) Since the distribution is symmetrical about a , its mean and median coincide, both being equal to a . Its mode is also equal to a , as can be seen from the fact that

$$f'(a)=0 \text{ and } f''(a)<0.$$

This is also evident from the fact that $\exp[-h^2u^2]$ decreases monotonically as u^2 increases from zero, i.e. as u deviates from zero in either direction.

(3) Again, because the distribution is symmetrical, its odd-order central moments are identically equal to zero. As regards the central moments of even orders, we have

$$\begin{aligned} \mu_{2r} &= \frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} (x-a)^{2r} \exp[-h^2(x-a)^2] dx \\ &= \frac{2h}{\sqrt{\pi}} \int_a^{\infty} (x-a)^{2r} \exp[-h^2(x-a)^2] dx && \text{(since the integrand is symmetrical about } a) \\ &= \frac{2h}{\sqrt{\pi}} \int_0^{\infty} \left(\frac{z}{h^2}\right)^r \exp[-z] \frac{dz}{2h\sqrt{z}} && \text{[putting } z=h^2(x-a)^2, \\ &&& \text{so that } dx=\frac{1}{2h\sqrt{z}} dz] \\ &= \frac{1}{h^{2r}\sqrt{\pi}} \int_0^{\infty} \exp[-z] z^{r-1/2} dz \\ &= \frac{1}{h^{2r}} \cdot \frac{\Gamma(r+\frac{1}{2})}{\sqrt{\pi}} = \frac{(r-\frac{1}{2})(r-\frac{3}{2}) \dots \frac{3}{2} \cdot \frac{1}{2}}{h^{2r}} && \text{(since } \Gamma(\frac{1}{2})=\sqrt{\pi}). \\ &&& \dots \quad (9.57) \end{aligned}$$

In particular,

$$\sigma^2 = \mu_2 = \frac{1}{2h^2}. \quad \dots \quad (9.57a)$$

Hence (9.57) may also be written in terms of σ as

$$\mu_{2r} = (2r-1)(2r-3)\dots 3 \cdot 1 \cdot \sigma^{2r}. \quad \dots \quad (9.57b)$$

The moment-generating function about μ of the normal distribution is given by

$$M_p(t) = E\{\exp[t(x-\mu)]\}$$

$$\begin{aligned} &= \int_{-\infty}^{\infty} \exp[t(x-\mu)] \frac{1}{\sqrt{2\pi}\sigma} \exp[-(x-\mu)^2/2\sigma^2] dx \\ &= \exp[t^2\sigma^2/2] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\{(x-\mu)^2 - 2t(x-\mu)\sigma^2 + t^2\sigma^4\}/2\sigma^2] dx \\ &= \exp[t^2\sigma^2/2] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\{x-\mu-t\sigma^2\}^2/2\sigma^2] dx \\ &= \exp[t^2\sigma^2/2], \end{aligned}$$

so that $\mu_{2r+1}=0$

and $\mu_{2r} = \frac{(2r)!}{r!2^r} \sigma^{2r} = (2r-1)(2r-3)\dots 3 \cdot 1 \cdot \sigma^{2r}$,

as before.

For the normal distribution, we have

$$\mu_3=0 \quad \text{and} \quad \mu_4=3\sigma^4, \quad \dots \quad (9.58)$$

$$\text{giving} \quad \gamma_1=0 \quad \text{and} \quad \gamma_2=0. \quad \dots \quad (9.59)$$

Since $a=\mu$, the mean of the distribution, and $h=\frac{1}{\sigma\sqrt{2}}$, the probability-density function may be written in its more usual form :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2], \quad \dots \quad (9.60)$$

where $-\infty < x < \infty$.

(4) The curve has two points of inflection at a distance σ on either side of μ , since

$$\frac{df(x)}{dx} = -\frac{(x-\mu)}{\sigma^3\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2]$$

$$\text{and} \quad \frac{d^2f(x)}{dx^2} = \frac{(x-\mu)^2}{\sigma^5\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2] - \frac{1}{\sigma^3\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2]$$

$$= \frac{1}{\sigma^5\sqrt{2\pi}} \{(x-\mu)^2 - \sigma^2\} \exp[-(x-\mu)^2/2\sigma^2], \quad \dots \quad (9.61)$$

Thus the curve of the normal distribution is convex upwards within the interval $(\mu - \sigma, \mu + \sigma)$, and outside this interval the curve is concave upwards. Fig. 9.1 shows a number of normal curves with the same mean but with different standard deviations.

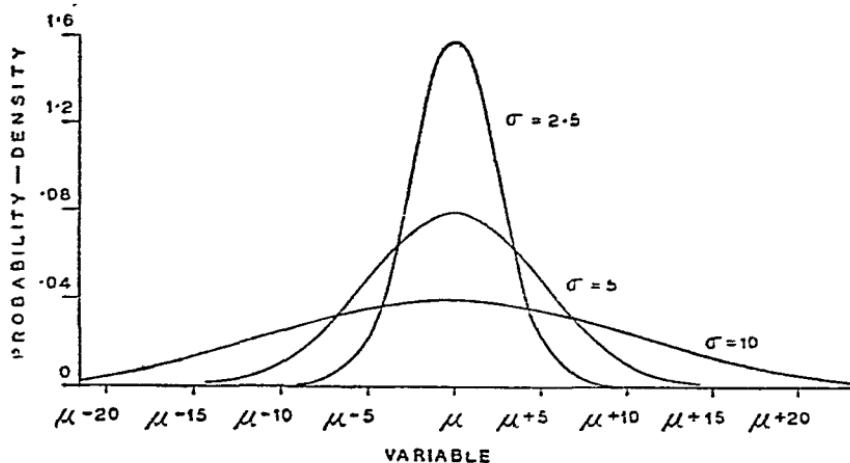


Fig. 9.1 Normal curves with the same mean but with different standard deviations.

(5) Let $\tau = \frac{x-\mu}{\sigma}$, where x is a normal variable with mean μ and standard deviation σ . Then obviously τ has mean zero and standard deviation unity. Further, it can be shown that τ is itself a normal variable (*vide* Section 14.6). Such a normal variable is called a *standard normal variable* or *normal deviate* (with unit standard deviation). It has the probability-density function

$$\phi(\tau) = \frac{1}{\sqrt{2\pi}} \exp[-\tau^2/2]. \quad \dots \quad (9.62)$$

The values of $\phi(k)$ and

$$\Phi(k) = \int_{-\infty}^k \phi(\tau) d\tau, \quad \dots \quad (9.63)$$

which is the cumulative probability $P[\tau \leq k]$, are given in statistical tables for different values of k . In the present book they appear in Table I in the Appendix.

The ordinates and the cumulative probabilities $P[x \leq a]$, for the

normal variable x with mean μ and standard deviation σ , are obtained from the tabulated values, using the following relations :

$$f(a) = \frac{1}{\sigma} \cdot \phi\left(\frac{a-\mu}{\sigma}\right) \quad \dots \quad (9.62a)$$

and $P[x \leq a] = \Phi\left(\frac{a-\mu}{\sigma}\right). \quad \dots \quad (9.63a)$

Further, since the distribution of τ is symmetrical about 0,

$$\phi(-k) = \phi(k) \quad \dots \quad (9.62b)$$

and $\Phi(-k) = 1 - \Phi(k). \quad \dots \quad (9.63b)$

Hence the values of $\phi(k)$ and $\Phi(k)$ are tabulated only for non-negative values of k .

(6) The probability for a normal variable x (with mean μ and standard deviation σ) to lie in any specified interval, say in the interval from a to b ($a < b$), can be obtained from the tabulated values of $\Phi(\tau)$. Thus

$$P[a < x \leq b] = P[x \leq b] - P[x \leq a] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \quad \dots \quad (9.62c)$$

Such probabilities are shown in the following table for some typical parts of the range :

	Approximate probability
Less than $\mu - 2\sigma$	0.02
Between $\mu - 2\sigma$ and $\mu - \sigma$	0.14
Between $\mu - \sigma$ and μ	0.34
Between μ and $\mu + \sigma$	0.34
Between $\mu + \sigma$ and $\mu + 2\sigma$	0.14
Above $\mu + 2\sigma$	0.02

(7) Another point to be noted is that although a normal variable can theoretically take any value between $-\infty$ and ∞ , for all practical purposes it may be assumed to lie between $\mu - 3\sigma$ and $\mu + 3\sigma$, the probability of its lying beyond these limits being very small, 0.0027 approximately. The interval $(\mu - 3\sigma, \mu + 3\sigma)$ is often referred to as the *effective range* of the normal variable.

9.16 Limiting forms of binomial and Poisson distribution

We have seen that a binomial distribution with parameters m and p has

$$\gamma_1 = \frac{q-p}{\sqrt{mpq}} \quad \text{and} \quad \gamma_2 = \frac{1-6pq}{mpq}.$$

When m is very large, and neither p nor q is very small,

$$\gamma_1 \rightarrow 0 \quad \text{and} \quad \gamma_2 \rightarrow 0.$$

This indicates—and this has, in fact, been rigorously proved—that under the above conditions, the binomial distribution can be approximated by a normal distribution. The normal distribution will have the same mean, mp , and the same standard deviation, \sqrt{mpq} , as the binomial distribution.

Again, a Poisson distribution with parameter λ has

$$\gamma_1 = 1/\sqrt{\lambda} \quad \text{and} \quad \gamma_2 = 1/\lambda,$$

both of which become negligibly small when λ is a very large number. This suggests that a Poisson distribution can be approximated by a normal distribution, provided λ is sufficiently large. (This result too can be rigorously proved.) The approximating normal distribution has the same mean, λ , and the same standard deviation, $\sqrt{\lambda}$, as the Poisson distribution.

One point is to be noted in this connection. The binomial and Poisson distributions are discrete distributions, whereas the normal distribution is continuous. The probability that x assumes the value r is

$$(i) \quad \binom{m}{r} p^r q^{m-r} \quad \text{or} \quad (ii) \quad \exp[-\lambda] \frac{\lambda^r}{r!},$$

according as x is a binomial or a Poisson variable. To approximate the above probability by means of a normal distribution, we have to integrate the appropriate normal density function from $r-1/2$ to $r+1/2$. This has to be done since we are replacing, while making the approximation, a discrete variable by a continuous variable. Hence the approximate values for the above probabilities will be :

$$(i) \quad \frac{1}{\sqrt{2\pi mpq}} \int_{r-1/2}^{r+1/2} \exp[-(x-mp)^2/2mpq] dx$$

$$\text{and} \quad (ii) \quad \frac{1}{\sqrt{2\pi \lambda}} \int_{r-1/2}^{r+1/2} \exp[-(x-\lambda)^2/2\lambda] dx.$$

Similarly, if we have to find the probability $P[a \leq x \leq b]$, a and b being two positive integers, and x being a binomial or a Poisson variable, then we have to integrate the corresponding normal density function from $a - 1/2$ to $b + 1/2$. However, it will almost be the same as the integral from a to b , provided $1/2$ is very small compared to \sqrt{mpq} or $\sqrt{\lambda}$ (as the case may be).

9.17 Fitting a normal distribution

As in other cases, the first step in fitting a normal distribution to observed data consists in estimating the parameters μ and σ by the method of moments. Since μ and σ are the mean and standard deviation of the theoretical distribution, the method of moments gives as their estimates \bar{x} and s , the mean and standard deviation of the observed distribution. With these estimates, we can then calculate the expected frequencies by using the tables of the normal deviate. Consider, for instance, the expected frequency for the interval from $x=a$ to $x=b$. This expected frequency is

$$\begin{aligned} n \int_a^b \frac{1}{s\sqrt{2\pi}} \exp[-(x-\bar{x})^2/2s^2] dx &= n \int_{(a-\bar{x})/s}^{(b-\bar{x})/s} \phi(\tau) d\tau \\ &\quad \left[\text{where } \tau = \frac{x-\bar{x}}{s} \right] \\ &= n \left[\int_{-\infty}^{(b-\bar{x})/s} \phi(\tau) d\tau - \int_{-\infty}^{(a-\bar{x})/s} \phi(\tau) d\tau \right] \\ &= n \left[\Phi\left(\frac{b-\bar{x}}{s}\right) - \Phi\left(\frac{a-\bar{x}}{s}\right) \right] \end{aligned}$$

The values of $\Phi(\tau)$ can be obtained from the tables of the normal deviate.

If one wants to draw the fitted curve over the histogram of the observed distribution, it will be necessary to compute the ordinates

$$n \frac{1}{s\sqrt{2\pi}} \exp[-(x-\bar{x})^2/2s^2]$$

for some appropriate values of x . Usually one takes the ordinates at the class-boundaries of the observed distribution. Multiplication by n is essential, for otherwise the ordinates will not be comparable to the frequency-densities of the observed distribution.

Now,

$$n \cdot \frac{1}{s\sqrt{2\pi}} \exp[-(x-\bar{x})^2/2s^2] = \frac{n}{s} \cdot \phi(\tau),$$

where $\tau = \frac{x-\bar{x}}{s}$. The values of $\phi(\tau)$ can be obtained from the tables of the normal deviate.

Ex. 9.4 Fit a normal distribution to the frequency distribution of height of Indian adult males given in Table 5.10. Also draw the fitted curve over the histogram of the observed distribution.

For the distribution of height of Indian adult males, the mean and standard deviation were found to be

$$\bar{x} = 164.734 \text{ mm. and } s = 5.472 \text{ mm.}$$

Here $n = 177$ and $\frac{n}{s} = 32.346$.

TABLE 9.4

FITTING A NORMAL DISTRIBUTION TO THE HEIGHT-DISTRIBUTION OF INDIAN ADULT MALES (TABLE 5.10)

Height (mm.) x (1)	$\tau = \frac{x-\bar{x}}{s}$ (2)	$\phi(\tau)$ (3)	Ordinate $= \frac{n}{s} \cdot \phi(\tau)$ (4)	$\Phi(\tau)$ (5)	$\Delta\Phi(\tau)$ (6)	Expected frequency $n \times \Delta\Phi(\tau)$ (7)	Observed frequency (1)
$-\infty$	$-\infty$	0	0	0	0.0001126*	0.020	0
144.55	-3.689	0.0004424	0.0143	0.0001126	0.0026478	0.469	1
149.55	-2.775	0.0084874	0.2745	0.0027604	0.0286123	5.064	3
154.55	-1.861	0.0706097	2.2839	0.0313727	0.1404492	24.860	24
159.55	-0.947	0.2547828	8.2412	0.1718219	0.3146168	55.687	58
164.55	-0.034	0.3987070	12.8966	0.4864387	0.3241316	57.371	60
169.55	0.880	0.2708640	8.7614	0.8105703	0.1530213	27.085	27
174.55	1.794	0.0798081	2.5815	0.9635916	0.0330236	5.845	2
179.55	2.708	0.0101984	0.3299	0.9966152	0.0032381	0.573	2
184.55	3.621	0.0005673	0.0183	0.9998533	0.0001467**	0.026	0
∞	∞	0	0	1			
Total	—	—	—	—	1.0000000	177.000	177

*It is the probability $P[x \leq 144.55]$.

**It is the probability $P[x \geq 184.55]$.

With these, we can now compute the expected frequencies for the different class-intervals and the ordinates at the class-boundaries in the manner explained above. In the tables $\phi(\tau)$ and $\Phi(\tau)$ are given for values of τ at intervals of 0.01, while in the present case we have taken $\tau = \frac{x - \bar{x}}{s}$ correct to 3 decimal places. For obtaining $\phi(\tau)$ and $\Phi(\tau)$ for these values, we have applied linear interpolation.

The agreement between the observed and the expected series of frequencies would seem to be fairly good. This agreement will also be apparent from Fig. 9.2, where we have the fitted normal curve, obtained on the basis of col. (4) of Table 9.4, superimposed on the histogram of the observed distribution.

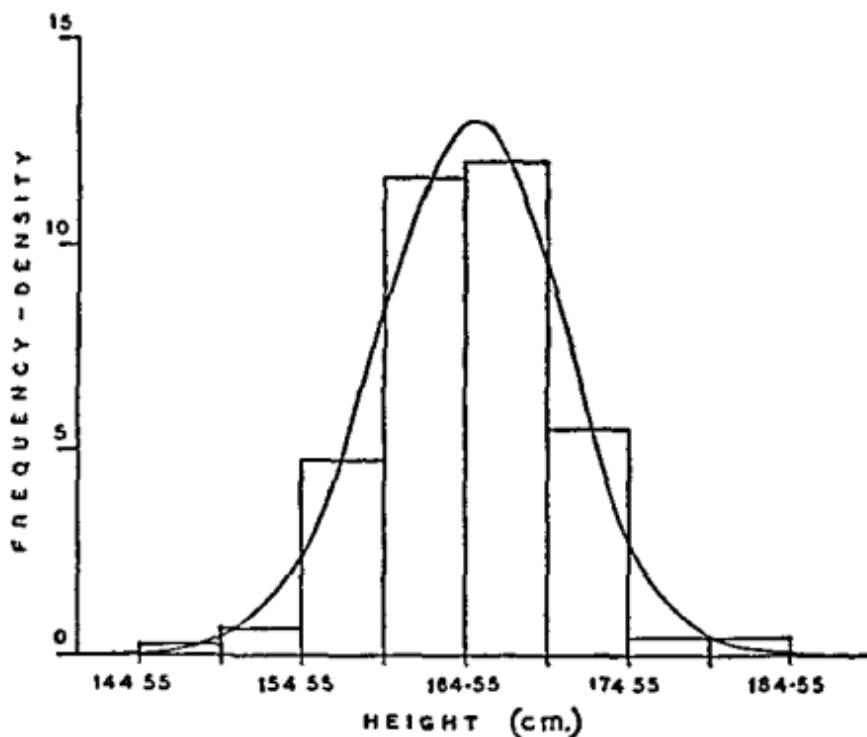


Fig. 9.2 Fitted normal curve together with the histogram of the height distribution of Indian adult males (Table 5.10).

9.18 Importance of the normal distribution in statistics

The normal distribution plays a very important rôle in statistical theory and its applications. As we have already seen, it has some

very simple properties which make it comparatively easy to deal with. Consequently, it will be a distinct advantage if in any case the population distribution of the variable under consideration may be assumed to be of the normal type. Generally, such an assumption is found legitimate in most cases of data arising from biological and psychological measurements. Under certain conditions, it can also be shown that the distribution of errors of observation in repeated measurements on a physical constant may be supposed to be normal. Such conditions being more or less valid in the field of manufacturing industry as well, most data arising there are also found to follow the normal law. Moreover, as we saw earlier, it serves as an approximation to the binomial and Poisson distributions, under certain conditions. Also, the sampling distributions of many statistics follow the normal form either exactly or approximately. (*Vide*, Chapters 14, 16 and 17.)

It should not, however, be supposed that the normal distribution is to be expected in all cases of continuous data. In fact, many distributions may be observed in practice, specially in the field of economics which deviate markedly from the normal law. Even in some of these cases, the normal distribution may be used as a first approximation, and conclusions arrived at in this way will be found virtually the same as those obtained by using the exact distribution. In some other cases, a transformation of the variable (like the logarithmic transformation used on economic data and discussed in Section 9.19) will make the distribution very nearly normal. In case none of the above procedures is feasible, one will, of course, have to make use of other theoretical distributions, e.g. the Pearsonian system of curves, Edgeworth's series, Gram-Charlier type *A* series, etc.

9.19 Log-normal distribution

The variable x is said to have a log-normal distribution if $\log x$ is normally distributed. As $\log x$ varies from $-\infty$ to ∞ , here x varies from 0 to ∞ . If $\log x$ has mean $\log \xi$ and standard deviation δ , then it has the p.d.f. $f(\log x)$ such that

$$f(\log x) d(\log x) = \frac{1}{\delta \sqrt{2\pi}} \exp[-(\log x - \log \xi)^2 / 2\delta^2] d(\log x),$$

$$-\infty < \log x < \infty, \dots \quad (9.64)$$

and hence the p.d.f. of x is

$$f(x) = \frac{1}{\delta x \sqrt{2\pi}} \exp[-(\log x - \log \xi)^2 / 2\delta^2], \quad 0 < x < \infty \quad (9.65)$$

Clearly, since $\log \xi$ is the median of the distribution of $\log x$ and $\log x$ is a monotonic function of x , the median of the distribution of x is ξ .

Also, the distribution is unimodal, since

$$\begin{aligned} \frac{df(x)}{dx} &= \frac{1}{\delta x \sqrt{2\pi}} \exp[-(\log x - \log \xi)^2 / 2\delta^2] \left\{ -\frac{1}{\delta^2 x} (\log x - \log \xi) \right\} \\ &\quad - \frac{1}{\delta^2 x^2 \sqrt{2\pi}} \exp[-(\log x - \log \xi)^2 / 2\delta^2] \\ &= -\frac{1}{\delta \sqrt{2\pi}} \exp[-(\log x - \log \xi)^2 / 2\delta^2] \left\{ \frac{1}{\delta^2 x^2} (\log x - \log \xi) + \frac{1}{x^2} \right\}, \end{aligned}$$

which vanishes, apart from doing so at $x=0$ and as $x \rightarrow \infty$, at

$$x = \xi \exp[-\delta^2], \quad (9.66)$$

the unique mode of the distribution.

The r th moment about 0 is

$$\begin{aligned} \mu_r' &= E(x^r) = \int_0^\infty x^r \frac{1}{\delta x \sqrt{2\pi}} \exp[-(\log x - \log \xi)^2 / 2\delta^2] dx \\ &= \int_{-\infty}^{\infty} \xi^r \exp[r\delta u] \frac{1}{\sqrt{2\pi}} \exp[-u^2/2] du \quad \left(\text{substituting } \frac{\log x - \log \xi}{\delta} = u \right) \\ &\quad \text{or } x = \xi \exp[\delta u], \text{ so that } \frac{dx}{\delta} = du \\ &= \xi^r \exp[r^2 \delta^2 / 2] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-(u - r\delta)^2 / 2] du \\ &= \xi^r \exp[r^2 \delta^2 / 2] \quad (9.67) \end{aligned}$$

In particular, the mean of the distribution is

$$\mu = E(x) = \xi \exp[\delta^2 / 2] = \xi w \text{ (say)}, \quad (9.68)$$

and

$$\mu_2 = \xi^2 \exp[2\delta^2], \quad (9.69)$$

so that

$$\begin{aligned} \mu_2 - \mu^2 &= \xi^2 \exp[2\delta^2] (\exp[\delta^2] - 1) \\ &= \xi^2 w^2 (w^2 - 1) \quad (9.70) \end{aligned}$$

Similarly,

$$\mu_3 = \xi^3 w^3 (w^2 - 1)^2 (w^2 + 2), \quad \dots \quad (9.71)$$

and $\mu_4 = \xi^4 w^4 (w^2 - 1)^2 (w^8 + 2w^6 + 3w^4 - 3).$... (9.72)

Hence

$$\gamma_1 = \sqrt{w^2 - 1} (w^2 + 2), \quad \dots \quad (9.73)$$

and $\gamma_2 = (w^2 - 1)(w^6 + 3w^4 + 6w^2 + 6).$... (9.74)

Thus the distribution is positively skew and leptokurtic.

9.20 Generalised systems of frequency curves

The failure of the normal distribution to fit many distributions observed in practice necessitated the development of generalised systems of frequency curves. The first approach, due to Karl Pearson, seeks to obtain a family of curves, known as Pearsonian curves, which would satisfactorily represent almost all practical distributions. The second approach, due to Bruns, Gram, Charlier and Edgeworth, seeks to represent a given density function as a linear combination of a simple density function and its derivatives. Thus if $\phi(x)$ is the normal density function and $\phi'(x)$ its r th derivative, the Gram-Charlier type A series has the form

$$f(x) = \phi(x) + c_1 \phi'(x) + c_2 \phi''(x) + \dots + c_r \phi^{(r)}(x). \quad \dots \quad (9.75)$$

The third approach, due to Edgeworth and others, seeks a transformation of the variable x so that the transformed variable has, at least approximately, a simple (say a normal) distribution. The last two approaches are beyond the scope of the present book. We shall, however, give a brief account of the Pearsonian system of frequency curves.

It is found that observed frequency distributions for homogeneous populations have (1) a single mode and (2) a high-order contact with the x -axis at the extremities. Thus if $f(x)$ be the probability-density function representing such a distribution, then

$$\frac{df}{dx} = 0 \text{ at } x = \alpha, \text{ the mode, and when } f = 0. \quad \dots \quad (9.76)$$

A differential equation satisfying these conditions is

$$\frac{df}{dx} = \frac{(x - \alpha)f}{b_0 + b_1 x + b_2 x^2}. \quad \dots \quad (9.77)$$

It is not necessary to take terms beyond $b_2 x^2$, because the

differential equation (9.77) is found adequate in providing curves of varying shapes and forms

For a general solution of the Pearsonian differential equation, it is written in the form

$$\frac{df}{f} = \frac{(x-\alpha)dx}{b_0 + b_1x + b_2x^2}$$

Integrating both sides, we have

$$\log f = \int \frac{(x-\alpha)dx}{b_0 + b_1x + b_2x^2} + \log C,$$

$\log C$ being the constant of integration, so that

$$f(x) = C \exp \left[\int \frac{(x-\alpha)dx}{b_0 + b_1x + b_2x^2} \right] \quad (9.78)$$

The explicit form of the function (9.78) depends upon the integral in the exponent, which again depends upon the roots of the quadratic $b_0 + b_1x + b_2x^2$ or, in other words, upon the values of the constants. A brief description of the important Pearsonian types is given below.

Type I This curve is obtained when the roots of the quadratic are real and of opposite signs. This happens when b_0 and b_2 are of opposite signs. Writing

$$\kappa = \frac{b_1^2}{4b_0b_2}, \quad (9.79)$$

we may say that we get the Type I curve when $\kappa < 0$.

The solution of the differential equation gives the Type I curve as

$$f(x) = J_0 \left(1 + \frac{x}{a_1} \right)^{m_1} \left(1 - \frac{x}{a_2} \right)^{m_2}, \quad -a_1 < x < a_2, \quad (9.80)$$

with origin at mode, where $m_1/a_1 = m_2/a_2$,

The curve reduces to the beta form under the substitution

$$z = \frac{x + a_1}{a_1 + a_2},$$

when the equation to the curve becomes

$$f(z) = \frac{1}{B(m_1+1, m_2+1)} z^{m_1} (1-z)^{m_2}, \quad 0 < z < 1 \quad (9.81)$$

The curve is bell shaped, J shaped or U shaped, according as the constants m_1 and m_2 both are positive, or only one is positive and the other negative, or both are negative.

Type VI. The curve is obtained when the roots are real and are of the same sign. This occurs when b_0 and b_2 are of the same sign and $b_1^2 > 4b_0b_2$ or, in other words, when $\kappa > 1$. The probability-density function is

$$f(x) = C(x-a)^{q_2}x^{-q_1}, \quad a < x < \infty, \quad \dots \quad (9.82)$$

the origin being a units before the start of the curve.

The distribution reduces to the beta form under the transformation $z=a/x$. The curve is bell-shaped if q_2 is positive and J-shaped if q_2 is negative.

Type IV. This curve is obtained when the roots are imaginary or when $b_1^2 < 4b_0b_2$, i.e. when $0 < \kappa < 1$. The equation to the curve is

$$f(x) = C \left(1 + \frac{x^2}{a^2}\right)^{-m} \exp\left[-\nu \tan^{-1} \frac{x}{a}\right], \quad -\infty < x < \infty, \quad \dots \quad (9.83)$$

with origin $\frac{\nu a}{2m-2}$ units above the mean.

The curve is always bell-shaped.

Type III. This is a transition type and is obtained when $b_2=0$ or, in other words, when $\kappa \rightarrow \pm\infty$. The curve is

$$f(x) = C \exp[-\gamma x] \left(1 + \frac{x}{a}\right)^{\gamma a}, \quad -a < x < \infty, \quad \dots \quad (9.84)$$

with origin at mode.

The distribution can be changed into the gamma form by using the transformation $z=\gamma(x+a)$, when the density function reduces to

$$f(z) = \frac{1}{\Gamma(p+1)} \exp[-z] z^p, \quad 0 < z < \infty, \quad \dots \quad (9.85)$$

where $p=\gamma a$.

The curve is bell-shaped if $\gamma a=p$ is positive and J-shaped if p is negative.

Type V. This transition type is obtained when the roots are equal, i.e. when $b_1^2=4b_0b_2$ or $\kappa=1$. The curve has the equation

$$f(x) = C x^{-p} \exp[-\gamma/x], \quad 0 < x < \infty, \quad \dots \quad (9.86)$$

with origin at the start of the curve.

It is transformed into the gamma form under the substitution $\gamma/x=z$ and is always bell-shaped.

Type II This is obtained when $b_1=0$ and b_0, b_2 are of opposite signs. The equation to the curve is

$$f(x) = C \left(1 - \frac{x^2}{a^2}\right)^m, \quad -a < x < a, \quad (9.87)$$

with origin at mean

Obviously, the curve is symmetrical about the origin and is bell-shaped or U-shaped, according as m is positive or negative. This reduces to the beta form under the transformation $z = 1 - \frac{x^2}{a^2}$

Type VII This occurs when $b_1=0$ and b_0, b_2 are of the same sign. The equation to the curve is

$$f(x) = C \left(1 + \frac{x^2}{a^2}\right)^{-m}, \quad -\infty < x < \infty, \quad (9.88)$$

the origin being at the mean

This is also symmetrical about the origin and is transformed into the beta form under the substitution $z = \left(1 + \frac{x^2}{a^2}\right)^{-1}$. This is always bell-shaped.

The *normal curve* is also a transition type of the Pearsonian family and is obtained when $b_1=b_2=0$.

It can be shown that b_0, b_1 and b_2 , and hence κ , can be expressed in terms of β_1 and β_2 . Thus the curves of the Pearsonian family can be specified by the β_1 and β_2 criteria.

Writing the differential equation in the form

$$\frac{df}{dx} = \frac{xf}{b_0 + b_1x + b_2x^2} \quad (\text{origin at mode, } \alpha),$$

we have

$$\frac{d^2f}{dx^2} = \frac{d}{dx} \left(\frac{xf}{b_0 + b_1x + b_2x^2} \right) = \frac{f}{(b_0 + b_1x + b_2x^2)^2} \{(1 - b_2)x^2 + b_0\}$$

Thus the curves of the Pearsonian family have two points of inflection, given by

$$x = \pm \sqrt{\frac{b_0}{b_2 - 1}}, \quad (9.89)$$

which are equidistant from the mode

Questions and exercises

9.1 Derive the binomial distribution from a suitable probability model. Obtain the mean and the s.d. of the distribution.

9.2 Derive the Poisson distribution as a limiting form of the binomial distribution. Give examples of data for which the Poisson distribution is expected to give a good fit.

9.3 Show that the normal distribution may be looked upon as a limiting form of the binomial and Poisson distributions. What are the important properties of this distribution? Account for the importance of the normal distribution in statistical theory and practice.

9.4 Determine the modes of the binomial and Poisson distributions. Show that the mode coincides with the mean when mp or λ (as the case may be) is an integer.

Partial ans. The modes are the highest integers contained in $(m+1)p$ and λ .

9.5 Let the intensity of accident-proneness, λ , of workmen follow a gamma distribution with p.d.f. $f(\lambda) = \frac{\gamma^\alpha}{\Gamma(\alpha)} \exp[-\gamma\lambda] \lambda^{\alpha-1}$, $0 < \lambda < \infty$, and let the number of accidents made by a workman whose intensity of accident-proneness is λ follow a Poisson distribution with p.m.f. $p(x|\lambda) = \exp[-\lambda] \frac{\lambda^x}{x!}$, $x=0, 1, 2, \dots$. Show that the number of accidents x , made by a workman of unknown accident-proneness, follows a negative binomial distribution.

9.6 Show that the cumulative probability of the binomial distribution may be expressed in the form

$$\sum_{x=0}^k \binom{m}{x} p^x q^{m-x} = \frac{1}{B(m-k, k+1)} \int_0^q z^{m-k-1} (1-z)^k dz$$

and that of the Poisson distribution in the form

$$\sum_{x=0}^k \exp[-\lambda] \frac{\lambda^x}{x!} = \frac{1}{\Gamma(k+1)} \int_\lambda^\infty \exp[-z] z^k dz.$$

9.7 Obtain the moment-generating function of the negative binomial distribution and hence determine its first four moments.

9.8 The Pascal distribution is defined by

$$f(x) = \frac{1}{1+\mu} \left(\frac{\mu}{1+\mu} \right)^x, \quad x=0, 1, 2, \dots,$$

where $\mu > 0$

Find the mean and variance of the distribution

9.9 Suppose 5% of the inhabitants of Calcutta are cricket fans. Determine approximately the probability that a sample of 100 inhabitants will contain at least 8 cricket fans? *Ans.* 0.88

9.10 The probability of getting no misprint in a page of a book is 0.14. What is the probability that a page contains more than 2 misprints? (State the assumption you make.)

Ans. 0.31 (under proper assumption)

9.11 A Poisson distribution has a double mode at $x=2$ and $x=3$. What is the probability that x will have one or the other of the two values? *Ans.* 0.224

9.12 Starting from an appropriate differential equation, obtain the curves of the Pearsonian system. Discuss their important properties.

9.13 Show that for a symmetrical probability distribution (either discrete or continuous) all odd-order moments are equal to zero.

9.14 A continuous random variable x having values only between 0 and 4 has the density function $f(x) = \frac{1}{2} - ax$. Evaluate a .

9.15 Find the mean and variance of each of the following continuous probability distributions

(i) $f(x) = \exp(-x)$, $x \geq 0$,

(ii) $f(x) = \frac{1}{2} \exp(-|x|)$, $-\infty < x < \infty$

9.16 The life (in hours) of electronic tubes of a certain type is supposed to be normally distributed with $\mu = 155$ hr and $\sigma = 19$ hr. What is the probability that the life of a tube will be

(1) between 136 hr and 174 hr?

(2) between 117 hr and 193 hr?

(3) less than 117 hr?

(4) more than 193 hr?

If a sample of 200 tubes is taken, how many are expected to be in each of the above groups?

Partial ans. The probabilities are

(1) 0.68, (2) 0.96, (3) 0.02, (4) 0.02

9.17 The results of a particular examination are shown below in summary form :

Result	Percentage of candidates
Passed with distinction	15
Passed without distinction	42
Failed	43
Total	100

It is known that a candidate gets plucked if he obtains less than 40 marks (out of 100), while he must obtain at least 75 marks in order to pass with distinction. Hence determine the mean and s.d. of the distribution of marks, assuming that it is of the normal type.

$$\text{Ans. } \mu = 45.09; \sigma = 28.86.$$

9.18 Show that the mean deviation about mean of normal distribution is $\sqrt{\frac{2}{\pi}} \cdot \sigma$, σ being the s.d. of the distribution.

9.19 If $\log x$ is normally distributed with $\mu=1$ and $\sigma^2=4$, find $P[\frac{1}{2} < x < 2]$.

$$\text{Ans. } 0.106.$$

9.20 There are 600 commerce students in the post-graduate classes of a university, and the probability for any student to need a copy of a particular text-book from the university library on any day is 0.05. How many copies of the book should be kept in the university library so that the probability may be greater than 0.90 that none of the students needing a copy from the library has to come back disappointed? (Use the normal approximation to the binomial probability law.)

$$\text{Ans. At least 37 copies.}$$

9.21 Suppose the life time (in hours) of a radio tube of a certain type obeys the exponential law $f(x) = \frac{1}{\lambda} \exp[-x/\lambda]$, $x > 0$, with $\lambda = 900$.

A company producing tubes wishes to guarantee for the articles a certain life time. For how many hours should the tube be guaranteed to function to achieve a probability of 0.90 that it will function (at least) for the number of hours guaranteed.

$$\text{Ans. 95 hours.}$$

9.22 For the continuous probability distribution

$$f(x) = \theta \exp(-\theta x), \quad 0 < x < \infty,$$

where $\theta > 0$, find the moment generating function. Obtain the mean, variance, β_1 and β_2 of the distribution.

9.23 In the course of an experiment, 15 mosquitoes were put in each of 120 jars and were next subjected to a dose of DDT. After $\frac{1}{2}$ hours the number alive in each jar was counted and the following frequency distribution was obtained

No. of mosquitoes alive	0	1	2	3	4	5	6	7	8
Frequency (no. of jars)	2	12	14	22	28	17	13	10	2

Find the frequencies that one would expect on the assumption that each mosquito has a common probability of survival.

9.24 When the first proof of a book containing 250 pages was read, the following distribution of misprints was found

No. of misprints per page	Frequency
0	139
1	76
2	28
3	4
4	?
5	1
Total	250

Fit a Poisson distribution to the above data.

9.25 A telephone switch board handles 720 calls on the average during a rush hour. The board can make 15 connections per minute. Estimate the probability that the board will be overtaxed during any minute in the rush hour.

Ans 0.156

9.26 The following distribution relates to the number of accidents to 647 women working on H.E. (high explosive) shells during a 5-week period (given by Greenwood and Yule in *J.R.S.S.*, 1920). Show that a negative binomial distribution, rather than a Poisson distribution, gives a very good fit to the data. How would you explain this?

Number of accidents	0	1	2	3	4	5
Frequency	447	132	42	21	3	2

Hint : Refer to the result of *Exercise 9.5*.

9.27 The following is the frequency distribution of right-hand grip for 345 European males:

Right-hand grip (in lb.)	Frequency
29·5—39·5	1
39·5—49·5	2
49·5—59·5	12
59·5—69·5	52
69·5—79·5	99
79·5—89·5	101
89·5—99·5	55
99·5—109·5	17
109·5—119·5	5
119·5—129·5	1
Total	345

Find the expected frequencies for the above classes, assuming that the population distribution of right-hand grip is normal. Draw the fitted curve and the histogram of the observed distribution on the same graph paper.

9.28 A car hire firm has two cars, which are hired out by the day. It has been found that the number of demands for cars of the firm on any day has a Poisson distribution with mean 1.5.

(a) Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused.

(b) If the two cars are used an equal number of times on the average, on what proportion of days is a given one of the cars not in use?

(c) What proportion of demand has to be refused?

Ans (a) 0.223, (b) 0.390, (c) 0.114

SUGGESTED READING

- [1] Elderton, W P and Johnson, N L *Systems of Frequency Curves* Cambridge University Press, 1969
- [2] Feller, W *An Introduction to Probability Theory and Its Applications*, Vol I (Chs 6—7) John Wiley, 1968, and Wiley Eastern, 1972
- [3] Hald, A *Statistical Theory with Engineering Applications* (Chs 5—7) John Wiley, 1952
- [4] Mood, A M and Graybill, F A *Introduction to the Theory of Statistics* (Chs 3, 6) McGraw Hill, 1963, and Kogakusha
- [5] Parzen, E *Modern Probability Theory and Its Applications* (Chs 3—6) John Wiley, 1960 and Wiley Eastern 1972
- [6] Weatherburn, C E *A First Course in Mathematical Statistics* (Ch 3) Cambridge University Press, 1947

10.1 Data on two or more attributes

In some investigations, it may be appropriate to collect data, for a given set of individuals, on more than one character at the same time. The object here would be to look for any relationship that may obtain among the characters. In this chapter we shall be concerned with data on several *attributes*. The case of several *variables* will be taken up in Chapters 11 and 12. (There may, of course, be a third case, where some of the characters are attributes and the others are variables. But this case can be dealt with by suitably adapting the methods appropriate for the first two cases and hence will not be discussed separately.)

Thus the data may relate to the proficiency in English and the proficiency in mathematics of a group of high-school students (for each attribute there being, say, five classes : very good/good/mediocre/bad/very bad) ; or to the subject-matter (fiction/non-fiction) and the readability (easy-reading/difficult-reading) of a number of books ; or to the sex, the economic status (rich/middle-class/poor) and the level of education(illiterate/primary/high-school/university) of a group of adults.

In Table 10.1 data on two attributes are presented in a summary form. The figure in each cell stands for the number of individuals (i.e. the frequency) corresponding to a pair of forms of the two attributes. Thus, e.g., 19 is the number of adults attacked with fever among those administered quinine during the period, 193 the number of adults attacked with fever among those not administered quinine, and so on. The cell-frequencies, together with their grand total, give the *joint (frequency) distribution* of the attributes, because they show how the two attributes vary jointly in the given group of individuals. From the joint distribution, we also obtain two other types of distribution. Thus the row-totals (marginal frequencies), together with the grand total, give the distribution of the attribute

'precautionary measure', to be called the *marginal frequency distribution* of precautionary measure in the present context. On the other hand the column totals, together with the grand total, give the marginal distribution of 'outcome'. The other type of distribution is given by each column or each row of frequencies of the table, together with the corresponding column or row total. Take, e.g., the frequencies in the first row, together with the row total 606. For these frequencies the form of the first attribute, 'precautionary measure', is the same but the form of the second varies. As such it is said to give a *conditional frequency distribution*—the conditional distribution of 'outcome' for the form 'quinine used' of precautionary measure. Similarly, the second row gives the conditional distribution of outcome for the form 'no quinine used' of precautionary measure. The frequencies in the first column and those in the second, in the same way, give the conditional distributions of the attribute 'precautionary measure' for the forms 'attacked with malaria' and 'not attacked with malaria', respectively, of the attribute 'outcome'.

TABLE 10.1
DATA ON THE USE OF QUININE AND INCIDENCE OF MALARIA
COLLECTED IN AN INVESTIGATION IN A STATE OF INDIA
(Each figure relates to the number of adults in each
category among a total of 3,540 adults)

Precautionary measure (B)	Outcome (A)		Total
	Attacked with malaria (A)	Not attacked with malaria (\bar{a})	
Quinine used (B)	19 (f_{AB})	587 ($f_{\bar{a}B}$)	606 (f_B)
No quinine used (\bar{B})	193 ($f_{A\bar{B}}$)	2 741 ($f_{\bar{a}\bar{B}}$)	2 934 ($f_{\bar{B}}$)
Total	212 (f_A)	3 328 ($f_{\bar{a}}$)	3 540 (n)

Clearly, the attributes need not have just two forms each, i.e., the table need not be a 2×2 table. Thus in Table 10.2 we have data on two attributes each of which has three forms.

In each case, we might consider the relative frequencies, instead of the frequencies, which would also give the distributions—joint marginal or conditional—of the attributes, although in a different form.

10.2 Independence and association*

Consider again two attributes, A and B . In the 2×2 case, the two forms of A may be denoted by A (the 'positive' form, indicating the *presence* of the character A) and α (the 'negative' form, indicating the *absence* of the character A) and, similarly, the two forms of B may be denoted by B and β . The four cell-frequencies may be denoted by f_{AB} , f_{AB} , $f_{\alpha B}$ and $f_{\alpha B}$ and the total by n . Also, the (marginal) frequencies for the A -classes may be denoted by f_A and f_α , and the (marginal) frequencies for the B -classes by f_B and f_β . Thus

$$\left. \begin{array}{l} f_A = f_{AB} + f_{AB}, f_\alpha = f_{\alpha B} + f_{\alpha B}, \\ f_B = f_{AB} + f_{\alpha B}, f_\beta = f_{AB} + f_{\alpha B}, \end{array} \right\} \dots \quad (10.1)$$

and

$$n = f_{AB} + f_{AB} + f_{\alpha B} + f_{\alpha B} \dots \quad (10.2a)$$

$$= f_A + f_\alpha \dots \quad (10.2b)$$

$$= f_B + f_\beta. \dots \quad (10.2c)$$

Suppose the individuals under consideration constitute the population itself and not just a sample from the population. Also suppose that none of the marginal frequencies is zero. Then the ratios f_{AB}/f_A and $f_{\alpha B}/f_\alpha$ give, respectively, the proportions of members of the population having B , among those having A and among those having α . If these proportions be equal, we may say that the presence or absence of the character A in an individual does not in any way determine whether B will be present. A and B may then be called *statistically unrelated* or *independent*. As opposed to the notion of independence, there is the notion of *association*. Thus A and B are said to be associated if they are not independent.

We have seen that, for A and B to be independent, we must have

$$\frac{f_{AB}}{f_A} = \frac{f_{\alpha B}}{f_\alpha}. \dots \quad (10.3)$$

This implies

$$\frac{f_{AB}}{f_A} = \frac{f_{AB} + f_{\alpha B}}{f_A + f_\alpha} = \frac{f_B}{n}$$

or

$$f_{AB} = \frac{f_A f_B}{n}. \dots \quad (10.4a)$$

*The ideas in this section are comparable to those in Section 3.5.

Actually, (10.3) also implies

$$f_{AB} = \frac{f_A f_B}{n}, \quad (10.4b)$$

$$f_{AB} = \frac{f_A f_B}{n} \quad (10.4c)$$

and

$$f_{AB} = \frac{f_A f_B}{n} \quad (10.4d)$$

Since equation (10.4a) itself leads to (10.4b), (10.4c), (10.4d) and to (10.3) it is taken as the defining equation for the independence of A and B . This is done irrespective of whether f_A and/or f_B is zero.

Suppose A and B are not independent, i.e. are associated. We may distinguish two cases (i) If

$$f_{AB} > \frac{f_A f_B}{n}, \quad (10.5)$$

A and B occur together more frequently than they would have if they had been independent. Hence in this case the attributes are said to be *positively associated* (or, simply, associated). (ii) On the other hand, if

$$f_{AB} < \frac{f_A f_B}{n}, \quad (10.6)$$

i.e. if A and B occur together less frequently than they would have if they had been independent, then they are said to be *negatively associated* (or *disassociated*).

As regards the definition of *perfect association*, we may adopt one of two alternatives (1) Thus we may say that there is perfect positive association between A and B if all A 's are B 's and/or all B 's are A 's, i.e. if $f_{AB}=0$ and/or $f_{AB}=0$. Likewise, there may be said to be perfect negative association if no A 's are B 's and/or no α 's are β 's, i.e. if $f_{AB}=0$ and/or $f_{AB}=0$. (2) Alternatively, we may say that there is perfect positive association if all A 's are B 's and all B 's are A 's, i.e. if $f_{AB}=0$ and $f_{AB}=0$, and that there is perfect negative association if no A 's are B 's and no α 's are β 's, i.e. if $f_{AB}=0$ and $f_{AB}=0$.

To keep these two cases distinct, the association will be said to be *complete* (positive or negative) in the first case and to be *absolute* in the second.

10.3 Measures of association for the 2×2 case

We shall consider measures of the extent to which A and B , each of which occurs in two possible forms, may be said to be associated. Clearly, there are certain desiderata that such a measure should fulfil. For one thing, it should be independent of the total frequency n , just as, say, the mean or the moments are, and should thus depend on the relative frequencies in the cells rather than on their frequencies. Secondly, it should be zero in the case of independence, negative in the case of negative association and positive in the case of positive association. Thirdly, it should increase from its lowest possible value through zero to its highest possible value as we proceed from perfect negative association through independence to perfect positive association. Lastly, it should preferably vary between two definite limits, like -1 and $+1$.

Obviously, the difference

$$\delta_{AB} = f_{AB} - \frac{f_A f_B}{n}, \quad \dots \quad (10.7)$$

between the actual frequency for the cell (A, B) and the value that it should assume if A and B are independent, may serve as the basis for such a measure. Keeping all the desiderata in mind, one may use

$$Q_{AB} = \frac{n\delta_{AB}}{f_{AB}f_{aB} + f_{AB}f_{aB}} \quad \dots \quad (10.8)$$

$$= \frac{f_{AB}f_{aB} - f_{AB}f_{aB}}{f_{AB}f_{aB} + f_{AB}f_{aB}} \quad \dots \quad (10.9)$$

as a measure of association. It has been called the *coefficient of association* between A and B and is due to Yule. It may be seen that Q statifies all the desiderata stated above. In particular, $Q_{AB}=0$ if and only if $\delta_{AB}=0$, i.e. if and only if A and B are independent. Its lowest possible value (-1) occurs when $f_{AB}f_{aB}=0$, i.e. when $f_{AB}=0$ and/or $f_{aB}=0$, i.e. when there is *complete* negative association between A and B . Likewise, its highest possible value ($+1$) occurs when there is *complete* positive association between A and B .

A measure with the same general properties as those of Q_{AB} is the *coefficient of colligation* Y_{AB} , also due to Yule and defined by

$$Y_{AB} = \frac{\sqrt{f_{AB}f_{aB}} - \sqrt{f_{AB}f_{aB}}}{\sqrt{f_{AB}f_{aB}} + \sqrt{f_{AB}f_{aB}}} \quad \dots \quad (10.10)$$

There is yet a third measure, viz.

$$V_{AB} = \frac{n\delta_{AB}}{\sqrt{f_A f_B f_{aB} f_{Bb}}} = \frac{f_{AB} f_{aB} - f_{AB} f_{aB}}{\sqrt{f_A f_a f_B f_B}}. \quad \dots \quad (10.11)$$

This has properties similar to those of Q and Y , but unlike Q and Y , $V = \mp 1$ when and only when there is *absolute association* between the two characters.

To prove this result, let us use the symbols a, b, c and d for f_{AB} , f_{aB} , f_{aB} and f_{aB} , respectively. Then

$$V_{AB} = \frac{ad - bc}{\{(a+b)(c+d)(a+c)(b+d)\}^{1/2}},$$

and this equals ∓ 1 if and only if

$$(ad - bc)^2 = (a+b)(c+d)(a+c)(b+d),$$

i.e. if and only if

$$\begin{aligned} a^2(bc + bd + cd) + b^2(ac + ad + cd) + c^2(ab + ad + bd) \\ + d^2(ac + ab + bc) + 4abcd = 0. \end{aligned} \quad \dots \quad (10.12)$$

But this expression can vanish only if at least two of the non-negative quantities, a, b, c and d , vanish. We assumed, however, that the marginal frequencies are all non-zero, precluding the cases $a=b=0, c=d=0, a=c=0$, and $b=d=0$. Hence $V_{AB} = \pm 1$ if and only if $b=c=0$ or $a=d=0$. In the former case there is absolute positive association between A and B and $V_{AB} = +1$, while in the latter there is absolute negative association and $V_{AB} = -1$.

Ex. 10.1 For the data of Table 10.1, let us denote by A the attribute 'outcome' and by B the attribute 'precautionary measure'. We then have for the data

$$\begin{aligned} Q_{AB} &= \frac{19 \times 2741 - 193 \times 587}{19 \times 2741 + 193 \times 587} = \frac{52079 - 113291}{52079 + 113291} \\ &= -61212/165370 = -0.37015, \end{aligned}$$

$$\begin{aligned} Y_{AB} &= \frac{\sqrt{52079} - \sqrt{113291}}{\sqrt{52079} + \sqrt{113291}} = \frac{228.208 - 336.587}{228.208 + 336.587} \\ &= -108.379/564.795 = -0.19189, \end{aligned}$$

while

$$\begin{aligned} V_{AB} &= \frac{19 \times 2741 - 193 \times 587}{\sqrt{212 \times 3328 \times 606 \times 2934}} = \frac{52079 - 113291}{\sqrt{125444583 \times 10^4}} \\ &= -61212/(11200 \times 10^2) = -0.05465. \end{aligned}$$

Each of the measures indicates only a slight negative association between the two attributes. In other words, there is only slight evidence in support of the belief that use of quinine is generally followed by exemption from attack of malaria.

One important point is to be noted in this connection. The notion of independence is, by its very nature, related to a population and so is the notion of association. However, it is perfectly legitimate to study the presence or absence of association in the population from sample data. We may thus compute a measure of association according to one of the formulæ given above, where n is now to be regarded as the sample size and the frequencies as the sample frequencies for the cells or the margins. As in many other cases, the sample measure is to be taken, at least for large n , as a good approximation to the corresponding population value. Indeed, the data of Table 10.1 are, more appropriately, to be considered to be sample data for a random sample of size 3,540 taken from the population of all adults in the given Indian State.

10.4 Manifold two-way ($k \times l$) classification

We shall now discuss cases where again there are two attributes, but at least one of which occurs in more than two forms. A two-way classification of this type is given in Table 10.2. The data relate to 830 professional workers living in Indian towns and cities, who were interviewed during a survey.

TABLE 10.2
CLASSIFICATION OF 830 PROFESSIONAL WORKERS ACCORDING
TO OCCUPATION GROUP AND ACTIVITY STATUS

		Activity status			Total
		Employees	Employers	Own-account workers	
Occupation group	Scientists and technicians	169	21	140	330
	Medical and health services	83	25	68	176
	Teachers	286	10	28	324
Total		538	56	236	830

The two attributes may again be denoted by A and B . Let A occur in one of k forms : A_1, A_2, \dots, A_k ; and let B occur in one of l forms : B_1, B_2, \dots, B_l . Suppose, of the n individuals under study, f_{ij} have the form A_i of A together with the form B_j of B . Then f_{ij} is the cell-frequency of the (i, j) cell or of the combination $A_i B_j$;

$$f_{i0} = \sum_{j=1}^l f_{ij} \quad (i=1, 2, \dots, k) \quad \dots \quad (10.13)$$

is the marginal frequency of A_i ; and

$$f_{0j} = \sum_{i=1}^k f_{ij} \quad (j=1, 2, \dots, l) \quad \dots \quad (10.14)$$

is the marginal frequency of B_j . Of course,

$$n = \sum_i f_{i0} \quad \dots \quad (10.15a)$$

$$= \sum_j f_{0j} \quad \dots \quad (10.15b)$$

$$= \sum_i \sum_j f_{ij}. \quad \dots \quad (10.15c)$$

The frequencies f_{ij} 's (together with n) define the joint distribution of A and B ; f_{i0} 's define the marginal distribution of A and f_{0j} 's the marginal distribution of B . The k conditional distributions of B for given forms of A are represented by the k columns of the two-way frequency table, while the l conditional distributions of A for given forms of B are represented by the l rows of the table.

As in the 2×2 case, here too we shall assume that

$$f_{i0} > 0 \text{ for each } i$$

and $f_{0j} > 0$ for each j .

In this case, we may consider A and B to be unrelated or *statistically independent* if

$$\frac{f_{10}}{f_{10}} = \frac{f_{20}}{f_{20}} = \dots = \frac{f_{k0}}{f_{k0}}$$

for each j or, equivalently, if

$$f_{ij} = \frac{f_{i0} f_{0j}}{n} \quad (\text{all } i, j). \quad \dots \quad (10.16)$$

The equations in (10.16) are used to verify that the two attributes are really independent. Note that of these kl only $(k-1)(l-1)$ are algebraically independent equations, there being a number of constraints, as implied by (10.13)—(10.15).

If, on the other hand,

$$f_{ij} \neq \frac{f_{i0} f_{0j}}{n} \quad \dots \quad (10.17)$$

for any pair (i, j) , then A and B will be said to be *associated*. It should be realised that in a manifold two-way classification generally it will not be meaningful to make a distinction between positive and negative association. However, this distinction will have a meaning in case the classification with respect to each attribute involves an implied ranking—when, e.g., a group of students is classified according to, say, proficiency in Subject I and in Subject II into five categories each : very good, good, mediocre, bad and poor. If students who are very good in Subject I are also generally found to be very good in Subject II, those who are good in Subject I are generally good in Subject II, and so on, then the two attributes may be said to be positively associated. If, on the other hand, those who are very good in Subject I are generally poor in Subject II, those who are good in Subject I are generally bad in Subject II, and so on, then the attributes may be considered to be negatively associated.

As in the 2×2 case, here too in constructing a measure of association we shall make use of the differences

$$\delta_{ij} = f_{ij} - \frac{f_{i0} f_{0j}}{n}, \quad \dots \quad (10.18)$$

between the actual cell-frequencies and the values they should assume if the characters A and B are, in fact, independent. The quantity

$$\chi^2_{AB} = \sum_i \sum_j \frac{\delta_{ij}^2}{(f_{i0} \times f_{0j})/n} \quad \dots \quad (10.19)$$

$$= n \sum_i \sum_j \delta_{ij}^2 / (f_{i0} \times f_{0j}) - n \quad \dots \quad (10.20)$$

may serve as a measure. This is zero if and only if A and B are independent (i.e. if and only if $\delta_{ij}=0$ for all i, j), and the higher the strength of association, the higher is the value of χ^2_{AB} . However, χ^2_{AB} depends too much on the total frequency n , and theoretically it can be infinitely large. A measure that does not suffer from this defect

is Karl Pearson's coefficient of contingency,

$$C_{AB} = \sqrt{\frac{\chi^2_{AB}}{n + \chi^2_{AB}}} \quad (10.21)$$

C_{AB} equals zero if and only if $\delta_{ij} = 0$ for each i, j (i.e. if and only if $\chi^2_{AB} = 0$). However, it has the defect that its least upper bound is less than unity (For $\chi^2_{AB} \geq 0$ and $n > 0$, so that $\chi^2_{AB} < n + \chi^2_{AB}$ and hence necessarily $C_{AB} < 1$) In other words, it does not attain the value unity even if A and B are perfectly associated. Consider, e.g., a table with k classes for each of the two attributes, where every diagonal frequency $f_{ii} > 0$ and $f_{ij} = 0$ for every non-diagonal cell. Surely, no greater degree of association than this can be imagined in the $k \times k$ case. Yet, since $f_{ij} = 0$ for $i \neq j$ and $f_{ii} = f_{i0} = f_{0i}$,

$$\begin{aligned}\chi^2_{AB} &= n \sum_i \sum_j \frac{f_{ij}^2}{f_{i0} \times f_{0i}} - n \\ &= n \sum_i \frac{f_{ii}^2}{f_{i0} \times f_{0i}} - n \\ &= n(k-1),\end{aligned}$$

and so

$$C_{AB} = \sqrt{\frac{n(k-1)}{n+n(k-1)}} = \sqrt{\frac{k-1}{k}}$$

To remove the stated defect, Tschuprow suggests an alternative coefficient

$$T_{AB} = \left\{ \frac{\chi^2_{AB}}{n \sqrt{(k-1)(l-1)}} \right\}^{1/2} \quad (10.22)$$

Like C_{AB} , T_{AB} vanishes if and only if A and B are independent. But unlike C_{AB} , if the attributes A and B are perfectly associated in a $k \times k$ table, then

$$T_{AB} = \left\{ \frac{n(k-1)}{n \sqrt{(k-1)(k-1)}} \right\}^{1/2} = 1.$$

However, not much is known about the behaviour of T_{AB} in the case of $k \times l$ tables with $k \neq l$.

Ex. 10.2 Consider the data of Table 10.2. Here for the two attributes, say A and B ,

$$\chi^2_{AB} = 830 \left[\frac{(169)^2}{538 \times 330} + \frac{(83)^2}{538 \times 176} + \frac{(286)^2}{538 \times 324} + \dots + \frac{(28)^2}{236 \times 324} - 1 \right]$$

TABLE 10.3
COMPUTATION OF MEASURES OF ASSOCIATION
FOR THE DATA OF TABLE 10.2

f_{ij}	$f_{i0} \times f_{0j}$	$(f_{ij})^2 / (f_{i0} \times f_{0j})$
169	177540	0.16087
83	94688	0.07275
286	174312	0.46925
21	18480	0.02386
25	9856	0.06341
10	18144	0.00551
140	77880	0.25167
68	41536	0.11133
28	76464	0.01025
Total	—	1.16890

From the above table, we get

$$\chi_{AB}^2 = 830 \times 0.16890 = 140.187,$$

so that $C_{AB} = \sqrt{\frac{140.187}{830 + 140.187}} = \sqrt{0.144495}$
 $= 0.38012,$

and $T_{AB} = \sqrt{\frac{140.187}{830 \times 2}} = \sqrt{0.08445}$
 $= 0.29060.$

Both indicate a moderate degree of association between the attributes.

10.5 Case of more than two attributes

When data are collected simultaneously on more than two attributes, one has to use suitable modifications and extensions of the principles set forth in the preceding sections. In *Exercise 10.12* we have a set of data of this type. We shall use a notation similar to that used in the case of two attributes. Thus, e.g., if

we have three attributes, A (having r forms : A_1, A_2, \dots, A_r), B (having s forms : B_1, B_2, \dots, B_s) and C (having t forms : C_1, C_2, \dots, C_t), we shall denote by f_{ijk} the frequency in the (i, j, k) cell and by n the total frequency. Also, we shall write

$$\left. \begin{aligned} f_{100} &= \sum_j \sum_k f_{1jk}, f_{0j0} = \sum_i \sum_k f_{ijk} \text{ and } f_{00k} = \sum_i \sum_j f_{ijk}; \\ f_{1j0} &= \sum_k f_{1jk}, f_{i0k} = \sum_j f_{ijk} \text{ and } f_{0jk} = \sum_i f_{ijk}. \end{aligned} \right\} \dots \quad (10.23)$$

Besides the joint distribution of A, B and C , defined by f_{ijk} 's, here we shall have two kinds of marginal distributions. In the first group will be the marginal distributions of A, B and C , defined respectively by f_{100} 's, f_{0j0} 's and f_{00k} 's; to the second group will belong the marginal distributions of A and B , of A and C and of B and C , given respectively by f_{1j0} 's, f_{i0k} 's and f_{0jk} 's. There will also be two kinds of conditional distributions. First, we shall have the conditional distributions of A for given forms of B and C , of B for given forms of A and C and of C for given forms of A and B . Secondly, we shall have the conditional distributions of A and B for given forms of C , of A and C for given forms of B and of B and C for given forms of A .

To discuss the types of problems that may arise when the data relate to more than two attributes, we need consider the case of three attributes only.

We may, first of all, want to investigate whether the attributes may be supposed to be *mutually independent*. The attributes, A, B and C , e.g., will be said to be mutually independent if the equations

$$f_{ijk} = \frac{f_{100} f_{0j0} f_{00k}}{n^2} \quad \dots \quad (10.24)$$

hold for all i, j and k . Not all these are algebraically independent equations. The number of independent equations is $(r-1)(s-1) + (r-1)(t-1) + (s-1)(t-1) + (r-1)(s-1)(t-1) = rst - r - s - t + 2$. E.g., if each attribute has just two forms— A and α , B and β , and C and γ , then the following 4 equations would be enough to ensure that A, B and C are independent attributes :

$$\left. \begin{aligned} f_{AB} &= f_A f_B / n, \\ f_{AC} &= f_A f_C / n, \\ f_{BC} &= f_B f_C / n \\ f_{ABC} &= f_A f_B f_C / n^2, \end{aligned} \right\} \quad \dots \quad (10.25)$$

and

where the symbols conform to the notation used in Section 10.2. As a measure of the *joint association* of the attributes, i.e. of the departure from mutual independence, we may use a variant of C or T , e.g.

$$C_{ABC} = \sqrt{\frac{\chi^2_{ABC}}{n + \chi^2_{ABC}}}, \quad \dots \quad (10.26)$$

where

$$\chi^2_{ABC} = \sum_i \sum_j \sum_k \left(f_{ijk} - \frac{f_{i00} f_{0j0} f_{00k}}{n^2} \right)^2 / \frac{f_{i00} f_{0j0} f_{00k}}{n^2}.$$

However, more commonly one will want to examine to what extent one attribute, considered to be of special importance, may be said to be associated with the others taken together. The attributes in the second group will then be lumped together, thus giving a two-way classification. E.g., if there be p attributes in all, each with 2 forms, then the two-way classification will have 2 classes in one direction and 2^{p-1} classes in the other. With the attributes A , B and C occurring, respectively, in r , s and t forms, as in the preceding paragraph. A will be said to be independent of or associated with B and C taken together, according as the following identities hold or do not hold simultaneously* :

$$f_{ijk} = \frac{f_{i00} f_{0jk}}{n}. \quad \dots \quad (10.27)$$

As a measure of the *multiple association* of A with B and C , we may then use a variant of Pearson's coefficient of contingency or of Tschuprow's T :

$$C_{A.BC} = \sqrt{\frac{\chi^2_{A.BC}}{n + \chi^2_{A.BC}}} \quad \dots \quad (10.28)$$

or

$$T_{A.BC} = \left\{ \frac{\chi^2_{A.BC}}{n \sqrt{(r-1)(st-1)}} \right\}^{1/2}, \quad \dots \quad (10.29)$$

where

$$\begin{aligned} \chi^2_{A.BC} &= \sum_i \sum_j \sum_k \left(f_{ijk} - \frac{f_{i00} f_{0jk}}{n} \right)^2 / \frac{f_{i00} f_{0jk}}{n} \\ &= n \sum_i \sum_j \sum_k \frac{f_{ijk}^2}{f_{i00} f_{0jk}} - n. \end{aligned}$$

Still more commonly, we shall have two attributes of primary importance, say A and B , considered along with some others which

*Of these only $(r-1)(st-1)$ are independent.

may have some influence on the former. In examining the association between A and B from a table like Table 10.1 or Table 10.2, all other characters are ignored, and hence this may be called *total association*. But we may like to take the influence of the other characters on both A and B into account. This will necessitate the measurement of the association between A and B for each combination of forms of the other attributes. Consider, for illustration, the case where A and B are studied together with one other attribute C , each of these attributes having two forms only. Then for each of the forms C and γ of the attribute C , we shall have to get a measure of association. These may be called the coefficient of *partial association* between A and B in the 'presence' of C and the coefficient of *partial association* between A and B in the presence of γ (i.e. in the 'absence' of C), and are given by the formulae

$$\begin{aligned} Q_{ABC} &= \frac{f_C \delta_{ABC}}{f_{ABC} f_{\alpha BC} + f_{ABC} f_{\alpha BC}} \\ &= \frac{f_{ABC} f_{\alpha BC} - f_{ABC} f_{\alpha BC}}{f_{ABC} f_{\alpha BC} + f_{ABC} f_{\alpha BC}} \end{aligned} \quad (10.30a)$$

and

$$\begin{aligned} Q_{AB\gamma} &= \frac{f_\gamma \delta_{AB\gamma}}{f_{AB\gamma} f_{\alpha B\gamma} + f_{AB\gamma} f_{\alpha B\gamma}} \\ &= \frac{f_{AB\gamma} f_{\alpha B\gamma} - f_{AB\gamma} f_{\alpha B\gamma}}{f_{AB\gamma} f_{\alpha B\gamma} + f_{AB\gamma} f_{\alpha B\gamma}}, \end{aligned} \quad (10.30b)$$

where

$$\delta_{ABC} = f_{ABC} - \frac{f_{AC} f_{BC}}{f_C}$$

and

$$\delta_{AB\gamma} = f_{AB\gamma} - \frac{f_{A\gamma} f_{B\gamma}}{f_\gamma}$$

To see how the influence of C on A and B may affect the association between them, let us write δ_{ABC} and $\delta_{AB\gamma}$ in terms of δ_{AB} , δ_{AC} and δ_{BC} . (Note that the signs of the Q 's are the same as those of the δ 's.) We have

$$\begin{aligned} \delta_{ABC} + \delta_{AB\gamma} &= f_{AB} - \frac{f_{AC} f_{BC}}{f_C} - \frac{f_{A\gamma} f_{B\gamma}}{f_\gamma} \\ &= \left(f_{AB} - \frac{f_A f_B}{n} \right) - \left(\frac{f_{AC} f_{BC}}{f_C} + \frac{f_{A\gamma} f_{B\gamma}}{f_\gamma} - \frac{f_A f_B}{n} \right) \\ &= \delta_{AB} - \frac{n f_\gamma f_{AC} f_{BC} + n f_C f_{A\gamma} f_{B\gamma} - f_A f_B f_C f_\gamma}{n f_C f_\gamma} \end{aligned}$$

$$\begin{aligned}
 &= \delta_{AB} - \frac{n^2}{n f_C f_\gamma} \left(f_{AC} - \frac{f_A f_C}{n} \right) \left(f_{BC} - \frac{f_B f_C}{n} \right) \\
 &= \delta_{AB} - \frac{n}{f_C f_\gamma} \delta_{AC} \delta_{BC}. \quad \dots \quad (10.31)
 \end{aligned}$$

Suppose now that

$$\delta_{AB,C} = \delta_{AB,\gamma} = 0.$$

Then

$$\delta_{AB} = \frac{n}{f_C f_\gamma} \delta_{AC} \delta_{BC},$$

which may be a non-zero quantity. This means that *A* and *B* may be independent when they are studied in the presence of *C* and also when they are studied in the absence of *C*. But, when *C* is ignored, they may appear to be associated simply because of the association of *C* with both *A* and *B* (i.e. because *C* may be such that neither δ_{AC} nor δ_{BC} vanishes). The association between *A* and *B*, as indicated by δ_{AB} , may thus be completely illusory. In the same way, (10.31) shows that δ_{AB} may be zero even when $\delta_{AB,C}$ or $\delta_{AB,\gamma}$ is non-zero. Thus the apparent independence of *A* and *B* may also be spurious, being due to the effect of a third attribute (*C*) on them.

10.6 Association and causal relationship

It should be obvious to the reader that an association between two attributes need not imply a causal relationship. For an association between two attributes, *A* and *B*, may be due to (a) *A* being a cause of *B* or (b) *B* being a cause of *A* or (c) both being caused by some other character or group of characters. Only in cases (a) and (b) is the relationship between *A* and *B* of the causal (i.e. cause-effect) type. To make sure that situation (c) does not obtain, we should study *A* and *B* together with other characters, *C*, *D*, etc., which are likely to have an influence on the former. That is to say, we should separately measure the association between *A* and *B* for each combination of forms of the other characters. The reason is that an apparent association between *A* and *B* may actually be due to the effect of those other characters on them, as has been seen in the last section.

For instance, we have seen in Ex 10.1 that the use of quinine as a precautionary measure (A) is negatively associated with attack of malaria (B). Even so it would be unwise to jump to the conclusion that the use of quinine provides protection against attack of malaria. It may be that the economic condition (C) of the people examined has something to do with the observed association. Consider a classification of the people according to C into two groups rich (C) and poor (y). It is well known that rich people are more health conscious than the poor and are more likely to afford the use of quinine as a precautionary measure. Hence A and C are likely to be positively associated. Again the rich live in more hygienic conditions than the poor and hence are less likely to be attacked with malaria. Thus B and C are likely to be negatively associated. And the non-zero association between A and B may actually be due to the non zero association of each of them with C .

As such while using a measure of association in looking for a causal relationship between two characters A and B we should consider them in conjunction with other characters C, D , etc., that are likely to have an effect on them. Only when A and B are found to be associated for fixed combinations of forms of these characters will it be proper to say that one of them is a cause of the other.

10.7 Smoking and lung cancer

The way in which the notion of association can be used in looking for a possible causal connection between two phenomena and the caution that one should exercise in interpreting the association between them are well illustrated by the so-called cancer controversy.

In the beginning of the 20th century statisticians noted an alarming rise in the incidence of lung cancer. While part of the rising trend might be attributed to improvements in diagnosis and the changing size and age composition of the population the evidence left little doubt that a true increase had taken place. With the increase in the incidence of lung cancer suspicions regarding the possible ill effects of smoking became deeper when medical men observed that lung cancer patients were predominantly heavy smokers of tobacco.

It was in 1952 that a remarkable piece of research was carried out on this issue by Bradford Hill and his colleagues. They picked from a number of hospitals in London about 1,500 patients who had been diagnosed as suffering from lung cancer and data were collected on their smoking habits. A similar number of non-cancer patients from the same hospitals were asked the same set of questions on their smoking habits. A comparison of the two groups showed that cigarette-smokers were more common among the lung cancer patients than among other patients. Further, as to the cigarette-smokers themselves, heavy smokers were found to be more common among the lung cancer patients than medium or light smokers. In the course of the next few years, more investigations on somewhat similar lines were conducted in different parts of the world, and they all confirmed Hill's findings. The British Medical Association thereupon started a crusade against smoking, demanding that the hazards of smoking "must be brought home to the public by all the modern devices of publicity".

However, R.A. Fisher, the celebrated British statistician, through a series of writings and lectures [1], deprecated this attempt on the part of the BMA "to plant fear in the minds of perhaps a hundred million smokers throughout the world—to plant it with the aid of all the means of modern publicity backed by public money". Hill's enquiry, Fisher said, had only made a good *prima facie* case against smoking. Further investigations subsequent to Hill's consisted very largely of the same type of observations and hence suffered from similar limitations. Fisher stressed that even if one admitted that the findings of Hill and others had established a significant positive association between cigarette-smoking and lung cancer, it might not indicate that the former causes the latter. For one thing, is it possible that lung cancer (rather the pre-cancerous condition involving slight chronic inflammation which exists for years in those who are going to show overt lung cancer) is a cause of smoking cigarettes? Maybe, this condition leads people to derive from smoking the same kind of satisfaction as one derives from it after a slight irritation or disappointment. For another (and Fisher thought this to be more likely), it may be that a common cause explains the observed association. The

obvious common cause is the genotype. Genotypic differences in men are expected to lead to differences in their susceptibility to lung cancer. Such differences are also likely to lead to differences in their smoking habits.

Although it was insinuated from some quarters that Fisher had entered the controversy at the behest of the cigarette-manufacturers, actually Fisher was perfectly logical in pointing out that the BMA was being too hasty in directing its guns towards smoking in its fight against lung cancer.

Anyway, it was important that the controversy should be set at rest and the world told the truth about the relationship between smoking and lung cancer. A thorough review of the investigations made till then came in the form of a report by the Royal College of Physicians of London. Their study, started in 1959 and continued till 1962, led to the conclusion that cigarette-smoking is indeed a cause of lung cancer. The same conclusion was reached in the course of a more thorough review made under the aegis of the U.S. Government. In early 1962, the U.S. Government set up a committee of experts—physicians, chemists, biochemists and statisticians—to make a thorough enquiry into the matter. Its report, briefly called the Surgeon-General's Report [5], came out in December, 1963.

The committee based its report principally on the findings of 29 retrospective and 7 prospective studies that had been conducted till 1963 in different parts of the world. In a retrospective study, the smoking histories of persons with a specified disease (e.g. lung cancer) would be compared with those of persons without the disease. In a prospective study, on the other hand, a comparison would be made between the death rates of smokers and non-smokers, both over-all and for specific causes of death. This is done by first recording the smoking habits of people and then obtaining death certificates for those who die after entering the study.

Now these studies indicated that a significant association does exist between smoking and lung cancer.

The committee noted that, while a direct experiment in man to test whether a causal relationship exists between smoking and lung cancer is not feasible, a considerable amount of experimental work

in many species of animals had shown that certain compounds identified in cigarette smoke can produce cancer. There are other substances in tobacco and smoke, which though not cancer-producing themselves, promote cancer production.

Second, the association was found to be consistent in the sense that none of the prospective studies and none but one of the retrospective studies showed results to the contrary. Thus, despite many variations in design and method, all the retrospective studies, except one which dealt with females, showed that there were proportionately more cigarette-smokers among lung cancer patients than among people without cancer.

Third, in the 9 retrospective studies in which relative risk ratios for smokers and non-smokers were calculated and in all 7 prospective studies, the relative risk ratios were uniformly high and fairly close to each other, thus attesting to the strength of the association. Moreover, a dose-effect phenomenon was apparent in the sense that the relative risk ratio was found to increase with the amount of tobacco or number of cigarettes consumed.

Fourth, the suggestion that genetic influences might underlie both the tendency to smoke and the tendency towards lung cancer was also examined by the committee and was found to be at variance with facts. For one thing, the great rise in the incidence of lung cancer among males that has occurred in recent decades points to the introduction of new factors without which the genotype would have little or no potency. Evidently, the genetic factors were not strong enough to cause lung cancer in large numbers of people under the environmental conditions that existed, say, half a century ago. And the possibility that the genetic constitution of man has changed simultaneously and identically in a large number of countries is also unlikely. For another, the risk of developing lung cancer has been found to diminish when smoking is discontinued, although the genetic constitution must have remained the same.

On the above considerations, the committee came to the conclusion that the high degree of association between smoking and lung cancer ought to be interpreted to mean that the former is a (major) cause of the latter.

Questions and exercises

10.1 Given n, f_A, f_B and f_{AB} , how would you find the other cell frequencies and marginal frequencies of the 2×2 table?

10.2 Consider three attributes, A , B and C , each occurring in two forms. Given $n, f_A, f_B, f_C, f_{AB}, f_{AC}, f_{BC}$ and f_{ABC} , how would you obtain the other frequencies?

10.3 On the basis of the performance of a group of students in a high school examination, the following statement was made

"Of the students concerned, 48% are good in English, 72% good in mathematics and 55% good in elementary science, 33% are good in both English and mathematics, 32% in English and elementary science and 38% in mathematics and elementary science, 30% are good in all three"

Show that the figures, as they stand, must be incorrect

10.4 Examine the following statement for possible contradictions

"Of 520 commuters interviewed at Howrah Station, 385 were found to be Government employees, 417 were smokers and 228 were in Western dress, 173 were both Government employees and smokers, 195 were Government employees in Western dress and 164 were smokers in Western dress"

10.5 For the case of two attributes, define independence and association (positive and negative). What are the different measures of association, and what are their properties?

10.6 Show that the two forms of Q , (10.8) and (10.9), are equivalent and that Q increases monotonically as ad increases (where $a=f_{AB}$ and $d=f_{AB}$), thus justifying the statement that Q changes from -1 through 0 to $+1$ as one goes from complete negative association through independence to complete positive association.

10.7 Show that

$$Q = \frac{2}{1+1^2}$$

and hence that Q is greater in absolute value than 1 , except when both are zero or ∓ 1 .

10.8 What is partial association? Explain the relevance of this concept to the investigation of a causal relationship between two attributes.

10.9 Examine the following statement : "One in every thousand smokers dies of lung cancer. Hence smoking must be a cause of lung cancer."

10.10 Compute a measure of association for the following data and comment.

DEATHS FROM TUBERCULOSIS IN ENGLAND & WALES IN 1956

	Males	Females	Total
Tuberculosis of respiratory system	3,534	1,319	4,853
Other forms of tuberculosis	270	252	322
Total	3,804	1,571	5,375

$$\text{Ans. } Q = 0.429.$$

10.11 For the following 3×3 classification, compute a measure of association. Would it be proper to attach a sign to this measure?

DATA ON 413 MALE COLLEGE STUDENTS (GIVING RESULTS OF A VISUAL ACUITY TEST AND A BALANCE TEST)

	Left-eyed	Ambiocular	Right-eyed	Total
Left-handed	48	25	52	125
Ambidextrous	32	13	25	70
Right-handed	94	33	91	218
Total	174	71	168	413

$$\text{Partial ans. } C = 0.076.$$

10.12 Would you say that the observed association between *A* and *B* for the following data is real? Or would you ascribe it to

the influence of C on A and B ? (A =type of high school education received, B =performance at university, C =income).

		English-medium schools	Others	Total
Successful at university	High income	38	97	135
	Low income	13	61	74
	Total	51	158	209
Unsuccessful at university	High income	16	141	157
	Low income	21	113	134
	Total	37	254	291
Total		88	412	500

Partial ans $Q_{AB}=0.378$, $Q_{ABC}=0.551$, $Q_{ABY}=0.068$.

SUGGESTED READING

- [1] Fisher, R. A *Smoking, the Cancer Controversy (Some Attempts to Assess the Evidence)* Oliver & Boyd, 1959.
- [2] Goodman, L. A and Kruskal, W. H. "Measures of association", *Jour. Am. Stat. Assocn.*, 49 (1954), pp 732-, and 54 (1959), pp. 123-.
- [3] Kendall, M. G. and Stuart, A *Advanced Theory of Statistics* (Ch. 33), Vol. II Charles Griffin, 1960
- [4] Kruskal, W. H. "Ordinal measures of association", *Jour. Am. Stat. Assocn.*, 53 (1958), pp. 814-.
- [5] U. S. Deptt. of Health, Education and Welfare. *Smoking and Health : Report of the Advisory Committee to the Surgeon-General of the Public Health Service*. Van Nostrand, 1964.
- [6] Yule, G. U. and Kendall, M. G. *An Introduction to the Theory of Statistics* (Ch. 1—3). Charles Griffin, 1950

11

BIVARIATE FREQUENCY DISTRIBUTIONS

11.1 Bivariate data

We have discussed in the last chapter methods of summarisation of data arising out of variation in two attributes. We shall now take up the case of two variables. The variables may be denoted by x and y . Thus x may be the height and y the weight of a person, or x may be the weight when it is green and y the weight of dry fibre for a jute plant. Our raw data will then consist of a number of pairs of values of x and y , each pair corresponding to a particular individual. As an example, we may consider the data of Table 11.1, which gives, for each of 20 undergraduate students, the marks obtained in statistics (Honours) in a college test (full marks 300) and in the subsequent university examination (full marks 600).

TABLE 11.1
MARKS OBTAINED BY 20 UNDERGRADUATE STUDENTS
IN STATISTICS HONOURS IN A COLLEGE TEST AND
IN THE SUBSEQUENT UNIVERSITY EXAMINATION

Serial No.	Marks obtained		Serial No.	Marks obtained	
	in college test	in university examination		in college test	in university examination
1	183	433	11	123	326
2	175	393	12	121	341
3	134	270	13	175	403
4	170	364	14	133	326
5	183	399	15	144	346
6	167	360	16	109	255
7	120	368	17	165	362
8	175	358	18	114	361
9	126	262	19	164	382
10	187	376	20	125	319

TABLE II-2
A BIVARIATE FREQUENCY TABLE, SHOWING INCOME AND PERCENTAGE OF
INCOME SPENT ON FOOD FOR 200 FAMILIES

		Family income (in Rs.)										Total
		100 < 150 5 150 5-200 5 200 5-250 5 250 5-300 5 300 5-350 5 350 5-400 5 400 5-450 5										
1	28.5-30.5											
	30.5-32.5											
	32.5-34.5											
	34.5-36.5											
	36.5-38.5											
	38.5-40.5	3										
	40.5-42.5											
	42.5-44.5											
	44.5-46.5											
	46.5-48.5											
	48.5-50.5											
Total		24	31	30	31	32	29	23	200			

When the data are considerably numerous, they may be summarised by using a two-way frequency table. For each variable a suitable number of classes are taken, keeping in view the same considerations as in the univariate case. If there are k classes for x and l classes for y , then there will be in all $k \times l$ cells in the two-way table. By going through the pairs of values of x and y , we can then find the number of individuals or frequency in each cell, perhaps using, for the sake of convenience in counting, the system of tally marks. The whole set of cell-frequencies will now define a frequency distribution—a *bivariate frequency distribution* (see Table 11.2).

In Table 11.2, the column-totals of frequencies show the numbers of individuals belonging to the corresponding x -classes, irrespective of their y -values (where x is taken to denote family income and y the percentage of family income spent on food). Thus they show the frequency distribution of x , called the *marginal distribution* of x in the present context. Similarly, the row-totals of frequencies give the marginal distribution of y . Again, if we consider a particular column of frequencies, we find the number of individuals falling in each y -class for the given values of x . It may be called a *conditional distribution* or an *array distribution* of y for given x . Similarly, any particular row of frequencies gives a conditional or an array distribution of x for given y .

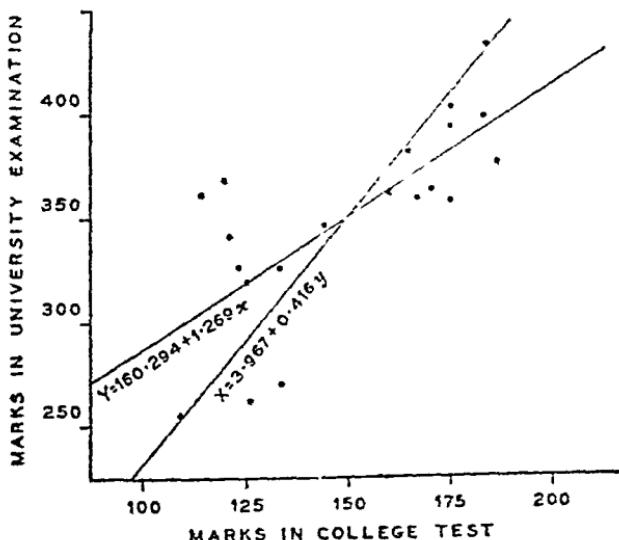


Fig. 11.1 Scatter diagram for the data of Table 11.1.

11.2 Scatter diagram

The simplest mode of diagrammatic representation of bivariate data is the use of a *scatter diagram* or *dot diagram*. Taking two perpendicular axes of co ordinates, one for x and the other for y , each pair of values is plotted as a point on graph paper. The whole set of points taken together constitutes a scatter diagram. The data of Table 11.1 have been represented in this way in Fig. 11.1. This method is of course, not suitable when the number of individuals is very large. In such a case, one may use some three-dimensional analogue of either the histogram or the frequency polygon, constructed on the basis of a bivariate frequency table.

11.3 Correlation

In the bivariate case, we may analyse the data relating to each variable separately by using the methods discussed in Chapters 5–9, but here we are primarily interested in the relationship between the two variables and for this some new methods have to be devised. The problems with which we are mainly concerned may be of two types. First, the data may reveal some relationship between the two variables and we may want to measure the extent to which they are related. Secondly, there may be one variable of particular interest and the other, regarded as an auxiliary variable, may be studied for its possible aid in throwing some light on the former. One is then interested in using any relationship that may be found from the observed data for making estimates or predictions of the principal variable in situations similar to the one under consideration.

Regarding the first type of problem, the simplest case occurs when, from the scatter diagram or otherwise, the variables are found to be linearly related, at least or approximately. Here, with the values of one variable lying in any assigned interval, however small, the corresponding values of the other variable may be found to differ considerably. Let us take the average of these values. If it is found that as one variable increases the other also increases, in general or on the average, there will be said to be *positive correlation* between them (Fig. 11.2a). On the other hand, as one variable increases, the other may decrease on the average. Here we say that there is *negative correlation* between them (Fig. 11.2b). There may still be

a third situation where as one variable increases, the other remains constant on the average. This is the case of *zero or no correlation* and the two variables are then said to be uncorrelated (Fig. 11.2c).

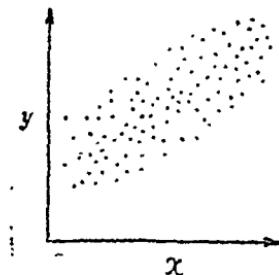


Fig. 11.2a
Positive correlation.

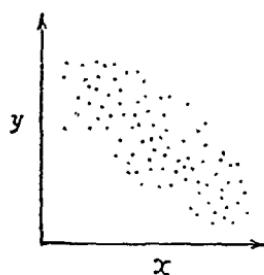


Fig. 11.2b
Negative correlation.

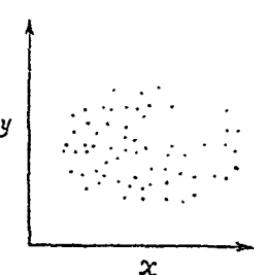


Fig. 11.2c
Zero correlation.

11.4 Correlation coefficient

In the scatter diagram, let us take two axes of co-ordinates for $x' = x - \bar{x}$ and $y' = y - \bar{y}$ (see Fig. 11.3). The origin of the new axes must be the point (\bar{x}, \bar{y}) , in terms of the original co-ordinates. The

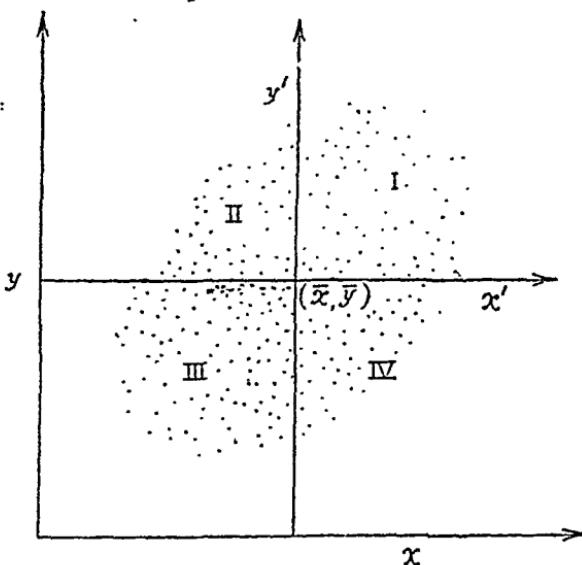


Fig. 11.3 A scatter diagram with the four quadrants of the (x', y') -plane.

points of the scatter diagram may now be seen as distributed over the four quadrants (I-IV) of the (x', y') -plane. Further, in quadrant I

x and y are both positive, in II x is negative and y positive, in III x and y are both negative, while in IV x is positive and y negative. Hence the product $x_i y_i$ is positive for all points occurring in quadrants I and III, while it is negative for all points in quadrants II and IV.

In the case of positive correlation, the general tendency of the points is to lie in quadrants I and III, so that in the sum $\sum_i x_i y_i$, the positive products outweigh the negative ones, and the sum thus becomes a positive quantity. In the case of negative correlation, the trend of the points is through quadrants II and IV in the sum $\sum_i x_i y_i$, the negative values now outweigh the positive ones, and the sum thus becomes a negative quantity. Lastly, when there is no correlation, the points are equally distributed over the four quadrants, and the sum $\sum_i x_i y_i$ becomes zero, as the set of positive products and that of negative products just balance each other.

Consequently a natural measure of correlation will seem to be the sum

$$\sum_i x_i y_i = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad (111)$$

But this sum is also dependent on some factors that have nothing to do with the correlation between the variables. For one thing it depends on the number of pairs of values of x and y which are taken into account. Secondly, it also depends on the units in which the variables x and y are measured and also on their variability. The first defect can be removed by dividing the sum by the number of pairs n . In order to eliminate the other defects, we divide the sum by the product of the standard deviations s_x and s_y (both assumed to be >0). The resulting measure of correlation is then

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}) / n}{s_x s_y} \quad (112)$$

r is called the (product moment) *correlation coefficient of the variables*. If one wants to specify the variables, one may use the symbol r_{xy} , as we shall have to do in a later chapter.

The quantity in the numerator of r is called the *covariance* of x and y , $\text{cov}(x, y)$, in analogy with the term *variance* that is used in the case of a single variable. Since the standard deviations, s_x and s_y , are the positive square-roots of the variances of x and y , say $\text{var}(x)$ and

$\text{var}(y)$, one may also write

$$r = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}. \quad \dots \quad (11.3)$$

Again,

$$\begin{aligned} n \text{cov}(x, y) &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i r_i y_i - n\bar{x}\bar{y} \\ &= \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n}, \end{aligned}$$

$$\begin{aligned} n \text{var}(x) &= \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2 \\ &= \sum_i x_i^2 - (\sum_i x_i)^2/n, \end{aligned}$$

and, similarly,

$$n \text{var}(y) = \sum_i y_i^2 - (\sum_i y_i)^2/n.$$

Hence r may be expressed in the alternative forms :

$$r = \frac{\frac{1}{n} \sum_i x_i y_i - \bar{x}\bar{y}}{\left(\frac{1}{n} \sum_i x_i^2 - \bar{x}^2 \right)^{1/2} \cdot \left(\frac{1}{n} \sum_i y_i^2 - \bar{y}^2 \right)^{1/2}} \quad \dots \quad (11.4)$$

$$= \frac{n \sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\left\{ n \sum_i x_i^2 - (\sum_i x_i)^2 \right\}^{1/2} \cdot \left\{ n \sum_i y_i^2 - (\sum_i y_i)^2 \right\}^{1/2}}. \quad \dots \quad (11.5)$$

The last form will be found to be the most convenient for computing r from raw data.

11.5 Properties of the correlation coefficient

(a) Obviously, the correlation coefficient of x and y is a pure number, that is, is independent of the units of measurement of x and y .

(b) Let $u = \frac{x-A}{c}$ and $v = \frac{y-B}{d}$, where A, B, c and d are four arbitrarily chosen constants. Corresponding to each pair of values (x_i, y_i) , we have a pair of values of the new variables, say (u_i, v_i) , where

$$u_i = \frac{x_i - A}{c}$$

$$v_i = \frac{y_i - B}{d}.$$

and

Since $x_i = A + cu_i$ and $\bar{x} = A + cu$,

$$x_i - \bar{x} = c(u_i - u)$$

Similarly,

$$y_i - \bar{y} = d(v_i - v)$$

Hence

$$\text{cov}(x, y) = cd \frac{1}{n} \sum_{i=1}^n (u_i - u)(v_i - v) = cd \text{cov}(u, v),$$

$$\text{var}(x) = c^2 \frac{1}{n} \sum_{i=1}^n (u_i - u)^2 = c^2 \text{var}(u),$$

$$\text{and } \text{var}(y) = d^2 \frac{1}{n} \sum_{i=1}^n (v_i - v)^2 = d^2 \text{var}(v)$$

Thus

$$\begin{aligned} r_{xy} &= \frac{cd \text{cov}(u, v)}{\sqrt{c^2 \text{var}(u)} \sqrt{d^2 \text{var}(v)}} \\ &= \frac{cd}{|c||d|} \frac{\text{cov}(u, v)}{\sqrt{\text{var}(u)} \sqrt{\text{var}(v)}} = \frac{cd}{|c||d|} r_{uv} \end{aligned} \quad (11.6)$$

If c and d have the same sign, $\frac{cd}{|c||d|}$ is 1. Thus in this case r_{xy} and r_{uv} are equal, both in magnitude and in sign. On the other hand, if c and d are of opposite signs, $\frac{cd}{|c||d|}$ is -1 , and here r_{xy} and r_{uv} will have the same magnitude but will be of opposite signs.

(c) In formula (11.2), let us put

$$x'_i = \frac{x_i - \bar{x}}{s_x}$$

and

$$y'_i = \frac{y_i - \bar{y}}{s_y}$$

Then

$$\sum_i x'^2_i = \sum_i y'^2_i = n$$

and

$$r = \frac{1}{n} \sum_i x'_i y'_i$$

Since

$$\frac{1}{n} \sum_i (x'_i + y'_i)^2 \geq 0,$$

or

$$\frac{1}{n} \sum_i x'^2_i + \frac{1}{n} \sum_i y'^2_i + 2 \frac{1}{n} \sum_i x'_i y'_i \geq 0,$$

or $2(1+r) \geq 0,$

we have

$$r \geq -1. \quad \dots \quad (11.7)$$

Again,

$$\frac{1}{n} \sum_i (x'_i - y'_i)^2 \geq 0,$$

or $\frac{1}{n} \sum_i x'^{2}_i + \frac{1}{n} \sum_i y'^{2}_i - \frac{2}{n} \sum_i x'_i y'_i \geq 0,$

or $2(1-r) \geq 0.$

Hence

$$r \leq 1. \quad \dots \quad (11.8)$$

Thus the correlation coefficient must necessarily lie between -1 and 1 .

r takes the lowest value -1 when, for each i ,

$$y'_i = -x'_i$$

or $y_i = \bar{y} - \frac{s_y}{s_x}(x_i - \bar{x}),$

and it takes the highest value 1 when, for each i ,

$$y'_i = x'_i$$

or $y_i = \bar{y} + \frac{s_y}{s_x}(x_i - \bar{x}).$

In these cases the variables are thus seen to be *perfectly correlated*; that is, each variable is an exact linear function of the other. The slope of the line is positive when $r=1$ and negative when $r=-1$.

Ex. 11.1 Let us consider the data of Table 11.1. The correlation coefficient between marks in college test and marks in the university examination is computed below.

We may denote by x the marks obtained in the college test and by y the marks in the university examination. For convenience in computations, we make a change of base for each of the variables and take

$$x-100 \text{ and } y-300$$

as u and v , respectively. The necessary calculations are indicated in the following table.

TABLE 11.3

**DETERMINATION OF CORRELATION COEFFICIENT BETWEEN
MARKS IN COLLEGE TEST AND MARKS IN UNIVERSITY
EXAMINATION (DATA OF TABLE 11.1)**

$v = x - 100$	$w = y - 300$	v^2	w^2	vw
83	133	6 889	16 689	11 039
75	93	5 625	8 649	6 975
34	-30	1 156	900	-1 020
70	64	4 900	4 096	4 480
83	99	6 889	9 801	8 217
67	60	4 489	3 600	4 020
20	68	400	4 624	1 360
75	58	5 625	3 364	4 350
26	-38	676	1 444	-988
87	76	7 569	5 776	6 612
23	26	529	676	598
21	41	441	1 681	861
75	103	5 625	10 609	7 725
33	26	1 089	676	858
44	46	1 936	2 116	2 024
9	-45	81	2 025	-405
65	0	4 225	3 844	4 030
14	61	196	3 721	854
64	82	4 096	6 724	5 248
25	19	625	361	475
993	1,004	63 061	92,376	67,313

The correlation coefficient between x and y is, by virtue of property (b),

$$\begin{aligned}
 r_{x,y} = r_{u,v} &= \frac{n \sum_i u_i v_i - (\sum_i u_i)(\sum_i v_i)}{\{n \sum_i u_i^2 - (\sum_i u_i)^2\}^{1/2} \{n \sum_i v_i^2 - (\sum_i v_i)^2\}^{1/2}} \\
 &= \frac{20 \times 67,313 - 993 \times 1,004}{\{20 \times 63,061 - (993)^2\}^{1/2} \{20 \times 92,376 - (1,004)^2\}^{1/2}} \\
 &= \frac{349,288}{(275,171)^{1/2} (839,504)^{1/2}} \\
 &= \frac{349,288}{524.6 \times 916.2} \\
 &= 0.727.
 \end{aligned}$$

11.6 Calculation of correlation coefficient from grouped data

Suppose the values of x and y are given in the form of a bivariate frequency table with k classes for x and l classes for y . Let us denote by x_i the mid-point of the i th class of x and by y_j the mid-point of the j th class of y . These two classes define the (i, j) cell of the bivariate frequency table. Let f_{ij} denote the frequency in this cell. Then the correlation coefficient is obtained from the formula

$$r = \frac{\sum_{i,j} (x_i - \bar{x})(y_j - \bar{y}) f_{ij}}{\{\sum_i (x_i - \bar{x})^2 f_{i0}\}^{1/2} \{\sum_j (y_j - \bar{y})^2 f_{0j}\}^{1/2}}, \quad \dots \quad (11.9)$$

where $f_{i0} = \sum_j f_{ij}$, $f_{0j} = \sum_i f_{ij}$

and $n = \sum_{i,j} f_{ij}$.

Obviously,

$$\bar{x} = \sum_i x_i f_{i0}/n, \quad \bar{y} = \sum_j y_j f_{0j}/n.$$

To reduce computational labour, we may make changes of base and scale for both x and y . It will be found advantageous (when the classes for each variable are equally wide) to take as bases two class-marks, say A and B , somewhere in the middle of the ranges of x and y , respectively, and as units the widths of the corresponding class-intervals, say c and d .

The new variables are then $u = \frac{x-A}{c}$ and $v = \frac{y-B}{d}$.

From property (b) stated in Section 11.5, it follows that

$$r = \frac{\sum_{ij} (u_i - \bar{u})(v_j - \bar{v}) f_{ij}}{\{\sum_i (u_i - \bar{u})^2 f_{i0}\}^{1/2} \{\sum_j (v_j - \bar{v})^2 f_{0j}\}^{1/2}} \quad (11.10)$$

$$\begin{aligned} &= \frac{\sum_{ij} u_i v_j f_{ij} - n \bar{u} \bar{v}}{\{\sum_i u_i^2 f_{i0} - n \bar{u}^2\}^{1/2} \{\sum_j v_j^2 f_{0j} - n \bar{v}^2\}^{1/2}} \\ &= \frac{n \sum_{ij} u_i v_j f_{ij} - (\sum_i u_i f_{i0})(\sum_j v_j f_{0j})}{\{n \sum_i u_i^2 f_{i0} - (\sum_i u_i f_{i0})^2\}^{1/2} \{n \sum_j v_j^2 f_{0j} - (\sum_j v_j f_{0j})^2\}^{1/2}}, \end{aligned} \quad (11.11)$$

since $\bar{u} = \sum_i u_i f_{i0} / n$ and $\bar{v} = \sum_j v_j f_{0j} / n$

It is easy to calculate $\sum_i u_i f_{i0}$, $\sum_i u_i^2 f_{i0}$, $\sum_j v_j f_{0j}$, $\sum_j v_j^2 f_{0j}$ from the bivariate frequency table. The calculation of $\sum_{ij} u_i v_j f_{ij}$ is performed in two stages. At first one may calculate, for different fixed values of j , $\sum_i u_i f_{ij} = U_j$, and then obtain the sum $\sum_j v_j U_j$, which gives

$$\sum_i \sum_j u_i v_j f_{ij} = \sum_j v_j U_j$$

Alternatively, one may calculate, for different fixed values of i , $\sum_j v_j f_{ij} = V_i$, and at the next stage $\sum_i u_i V_i$, which is also equal to $\sum_{ij} u_i v_j f_{ij}$. The relation

$$\sum_i u_i V_i = \sum_j v_j U_j \quad (11.12)$$

serves as a useful check on the calculations. Two other checks are

$$\sum_i V_i = \sum_j v_j f_{0j} = \sum_j v_j f_{0j}, \quad (11.13)$$

and $\sum_i U_i = \sum_i u_i f_{i0}$ (11.14)

Ex 11.2 We shall compute the correlation coefficient for the data of Table 11.2 by the above method. Let us denote the income of family by x and the percentage of income spent on food by y .

TABLE 11.4

DETERMINATION OF CORRELATION COEFFICIENT BETWEEN FAMILY INCOME AND PERCENTAGE OF FAMILY INCOME SPENT ON FOOD (FOR BIVARIATE FREQUENCY DISTRIBUTION OF TABLE 11.2)

U_i	125.5	175.5	225.5	275.5	325.5	375.5	425.5	U_j	$U_j U_i$
U_i	-3	-2	-1	0	1	2	3	$v_j f_{ij}$	$v_j^2 f_{ij}$
U_j	-5	-4	-3	-2	-1	0	1	$v_j f_{ij}$	$v_j^2 f_{ij}$
29.5	-	-	-	-	-	-	-	-40	-200
31.5	-	-	-	-	-	-	-	-40	-160
33.5	-	-	-	-	-	-	-	-84	-252
35.5	-	-	-	-	-	-	-	-64	-128
37.5	-1	-	-	-	-	-	-	-28	-28
39.5	0	-	-	-	-	-	-	-	-
41.5	1	-	-	-	-	-	-	-	-
43.5	2	-	-	-	-	-	-	-	-
45.5	3	-	-	-	-	-	-	-	-
47.5	4	-	-	-	-	-	-	-	-
49.5	5	-	-	-	-	-	-	-	-
								23	200
								58	-5
								32	725
								116	-103
								-71	-234
								-57	-118
								-57	-780
								-78	-5
								-118	-780
								-103	-1,229
								-5	-5
								-	-

check

Since the x -classes have width 50 each and the y classes have width 2 each, we take, as our new variables,

$$u = \frac{x - 275.5}{50} \text{ and } v = \frac{y - 39.5}{2},$$

275.5 and 39.5 being the arbitrarily chosen origins. The necessary calculations are shown in Table 11.4. On the basis of this table, we have

$$\begin{aligned} r &= \frac{n \sum u_i V_i - (\sum U_i)(\sum V_i)}{\sqrt{\{n \sum u_i^2 f_{10} - (\sum U_i)^2\}^{1/2} \{n \sum v_i^2 f_{01} - (\sum V_i)^2\}^{1/2}}} \\ &= \frac{200 \times (-780) - (-5) \times (-103)}{\{200 \times 725 - (-5)^2\}^{1/2} \{200 \times 1,229 - (103)^2\}^{1/2}} \\ &= \frac{156,515}{(144,975)^{1/2} (235,191)^{1/2}} = -\frac{156,515}{380.8 \times 485.0} \\ &= -0.847 \end{aligned}$$

11.7 Regression lines

Let us now consider the problem of predicting, for an individual, the value of one variable (say y) from the given value of another variable (say x). To solve this problem it would be necessary to express the relationship between y and x in a mathematical form. Suppose in a particular case the approximate relation may be represented by a line

$$Y = a + bx, \quad (11.15)$$

Y denoting the predicted value of y . To get an appropriate line, it is necessary to determine a and b from the observed data. Let us assume that there are n given pairs of values of x and y , the i th pair being denoted by (x_i, y_i) . The above line gives as an estimate of y , the value

$$Y_i = a + bx_i$$

The difference $y_i - Y_i$ is thus the error of estimate for the i th pair, $i = 1, 2, \dots, n$. Since the line is to be used for estimating purposes, it is reasonable to require that a and b should be such that these errors of estimate are as small as possible. However, it will not be enough to minimise the sum of these errors, because the errors, which may be positive or negative, may even add up to zero for a line for which

the individual errors are of high magnitude. In most cases a satisfactory method of determining a and b would be the *method of least squares*, which consists in minimising the sum of squares of the errors of estimation. Thus the problem is to choose a and b in such a way as to minimise

$$\begin{aligned} S^2 &= \sum_i (y_i - Y_i)^2 \\ &= \sum_i (y_i - a - bx_i)^2. \end{aligned}$$

The desired values are obtained by solving the simultaneous equations, called the *normal equations*,

$$\sum_i (y_i - a - bx_i) = 0 \quad | \quad \dots \quad (11.16)$$

and

$$\sum_i x_i (y_i - a - bx_i) = 0, \quad |$$

i.e.

$$\sum_i y_i = na + b \sum_i x_i \quad |$$

and

$$\sum_i x_i y_i = a \sum_i x_i + b \sum_i x_i^2. \quad |$$

The roots of the equations are

$$\begin{aligned} b &= \frac{n \sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n \sum_i x_i^2 - (\sum_i x_i)^2} \\ &= \frac{\frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y}}{\frac{1}{n} \sum_i x_i^2 - \bar{x}^2} \\ &= \frac{\frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_i (x_i - \bar{x})^2} \\ &= \frac{\text{cov}(x, y)}{\text{var}(x)} = r \frac{s_y}{s_x} \quad \dots \quad (11.17) \end{aligned}$$

$$\text{and} \quad a = \bar{y} - b \bar{x}. \quad \dots \quad (11.18)$$

Substituting these values in (11.15), we have the desired prediction formula :

$$Y = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x}). \quad \dots \quad (11.19)$$

The line (11.19) is called the *regression line of y on x* $a = \bar{y} - r \frac{s_y}{s_x} \bar{x}$
 s_y is the *y*-intercept of the line and $b = r \frac{s_y}{s_x}$ is its slope. The coefficient b

is the amount by which λ increases for a unit increment in the value of x . It is called the *regression coefficient of y on x*.

Similarly, if we are interested in predicting x from y , we use the regression line of x on y , which has the equation

$$Y = \tau + r \frac{s_x}{s_y} (y - \bar{y}) \quad (11.20)$$

$r \frac{s_x}{s_y}$, which is the amount by which λ increases for a unit increment in y , is the regression coefficient of x on y .

It may be noted that both the regression lines pass through the point (\bar{x}, \bar{y}) , which is in consequence, their point of intersection.

11.8 Some important results relating to regression lines

Consider any one of the regression lines, say that of y on x . It has the following properties:

(a) Let $u = \frac{x - A}{c}$ and $v = \frac{y - B}{d}$

Then the regression coefficient of y on x , denoted by b_{yx} for the sake of definiteness, is

$$b_{yx} = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{cd \text{cov}(u, v)}{c^2 \text{var}(u)} = \frac{d}{c} \frac{\text{cov}(u, v)}{\text{var}(u)} = \frac{d}{c} b_{xy}$$

or $b_{yx} = \frac{d}{c} \frac{n \sum u_i v_i - (\sum u_i)(\sum v_i)}{n \sum u_i^2 - (\sum u_i)^2} \quad (11.21)$

The other constants in the regression equation are, in terms of u and v ,

$$\bar{v} - B + dv$$

and $x = A + cu$

(b) Since $\bar{Y}_i = \bar{y} + r \frac{s_y}{s_x} (x_i - \bar{x})$

$$\sum_i Y_i = n\bar{y} + r \frac{s_y}{s_x} \sum_i (x_i - \bar{x})$$

Dividing both sides by n and remembering that $\sum_i (x_i - \bar{x}) = 0$, we have

$$\bar{Y} = \bar{y}. \quad \dots \quad (11.22)$$

In words, the mean of the observed values of y is equal to the mean of the corresponding predicted values.

From this it follows that the mean of the errors of estimates, $e_i = y_i - Y_i$, is zero.

(c) From (b) we have

$$Y_i - \bar{Y} = r \frac{s_y}{s_x} (x_i - \bar{x}).$$

$$\begin{aligned} \text{Hence } \text{var}(Y) &= \frac{1}{n} \sum_i (Y_i - \bar{Y})^2 \\ &= r^2 \frac{s_y^2}{s_x^2} \cdot \frac{1}{n} \sum_i (x_i - \bar{x})^2 \\ &= r^2 \frac{s_y^2}{s_x^2} \cdot s_x^2 \\ &= r^2 s_y^2. \end{aligned} \quad \dots \quad (11.23)$$

$$\text{Thus } |r| = \frac{s_y}{s_y}, \quad \dots \quad (11.24)$$

which may be interpreted as the proportion of the total variability of y which is accounted for by its linear regression on x .

(d) Again, the residual variance is

$$\begin{aligned} \text{var}(e) &= \frac{1}{n} \sum_i e_i^2 \\ &= \frac{1}{n} \sum_i (y_i - Y_i)^2 \\ &= \frac{1}{n} \sum_i \{(y_i - \bar{y}) - r \frac{s_y}{s_x} (x_i - \bar{x})\}^2 \\ &= \frac{1}{n} \sum_i (y_i - \bar{y})^2 - 2r \frac{s_y}{s_x} \cdot \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ &\quad + r^2 \frac{s_y^2}{s_x^2} \cdot \frac{1}{n} \sum_i (x_i - \bar{x})^2 \\ &= s_y^2 - 2r \frac{s_y}{s_x} \cdot r s_x s_y + r^2 \frac{s_y^2}{s_x^2} \cdot s_x^2. \end{aligned}$$

$$\text{Hence } \text{var}(e) = s_y^2(1 - r^2). \quad \dots \quad (11.25)$$

The standard deviation of ϵ , which is called the *standard error of estimate* of y from its linear regression on x , is denoted by $s_{y,x}$. We have then

$$s_{y,x} = s_y \sqrt{1 - r^2} \quad (11.26)$$

Since $\text{var}(\epsilon) \geq 0$, from (11.25) we have

$$r^2 \leq 1$$

$$\text{or} \quad -1 \leq r \leq 1,$$

a result which has already been proved in a different way.

If $r = \pm 1$, then $\text{var}(\epsilon) = 0$. That is, in this case $y_i = Y_i$ for each i , so that all points in the scatter diagram lie on the regression line. Here the linear regression equation will be the ideal predicting formula for y when x is given.

On the other hand, if $r = 0$, then $\text{var}(\epsilon) = s_y^2$. This means that the errors of estimation are as much variable as the original values of y , and hence the linear regression equation is of no help in predicting the value of y when the corresponding value of x is given. (This is also seen from the fact that for $r = 0$

$$Y = \beta_0 + \beta_1 x,$$

so that, so far as the linear regression equation is concerned, x throws no light whatever on the value of y .)

Similar results will, of course, hold for the regression of x on y .

From this discussion, it would be obvious that the numerical value of r also serves as a measure of the capability of the linear regression equation of one variable on the other as a predicting formula. The higher the numerical value of r , the more efficient is the regression equation.

(e) We have

$$\begin{aligned} \text{cov}(x, \epsilon) &= \frac{1}{n} \sum_i (x_i - \bar{x}) \epsilon_i \\ &= \frac{1}{n} \sum_i x_i \epsilon_i - \bar{x} \frac{1}{n} \sum_i \epsilon_i = 0 \end{aligned}$$

because of the normal equations (11.16). Hence

$$r_{x,y} = 0 \quad (11.27)$$

and ϵ may be looked upon as the part of y which is uncorrelated with x .

(f) Since $Y = a + bx$, we have

$$\begin{aligned}\text{cov}(Y, e) &= \frac{a}{n} \sum_i e_i + \frac{b}{n} \sum_i x_i e_i \\ &= 0,\end{aligned}$$

owing to the normal equations. Also,

$$y = Y + e,$$

so that

$$\text{cov}(y, Y) = \text{var}(Y) + \text{cov}(Y, e) = \text{var}(Y).$$

Hence

$$\begin{aligned}r_{yY} &= \frac{\text{cov}(y, Y)}{\sqrt{\text{var}(y)} \sqrt{\text{var}(Y)}} \\ &= \sqrt{\frac{\text{var}(Y)}{\text{var}(y)}} = |r|, \quad \dots \quad (11.28)\end{aligned}$$

which means that the correlation between y and its 'predicted' value Y must be non-negative and must be numerically the same as the correlation between y and x .

(g) We have seen that

$$b_{yx} = r \frac{s_y}{s_x}$$

and

$$b_{xy} = r \frac{s_x}{s_y}.$$

Hence

$$b_{yx} \times b_{xy} = r^2$$

or

$$r = \sqrt{b_{yx} \times b_{xy}}. \quad \dots \quad (11.29)$$

Thus the correlation coefficient is the geometric mean of the two regression coefficients. As regards the sign of r , it is the same as the common sign of the two regression coefficients.

Ex. 11.3 For the data of Table 11.1, the regression of y (marks in the university examination) on x (marks in the college test) is of considerable practical importance.

Here $\bar{x} = 100 + \frac{993}{20} = 149.65$

and $\bar{y} = 300 + \frac{1,004}{20} = 350.2.$

The regression coefficient of y on x is

$$b_{y,x} = b_{xy} = \frac{n \sum u_i v_i - (\sum u_i)(\sum v_i)}{n \sum u_i^2 - (\sum u_i)^2}$$

$$= \frac{349,288}{275,171} = 1.269$$

Hence the linear regression of y on x is represented by the equation

$$y = 350.2 + 1.269(x - 149.65)$$

or $y = 160.294 + 1.269x$

This equation may be used to predict how much a new student under similar conditions, is likely to score in the university examination from a knowledge of the marks he obtains in the college test. Since the correlation coefficient is quite high (0.727), the prediction is expected to be fairly precise.

The regression line of x on y is

$$x = 149.65 + 0.416(y - 350.2)$$

$$= 3967 + 0.416y$$

This regression equation is, however, mainly of theoretical interest. (This could be used to answer questions of the form "how much does a student getting 360 marks in the university examination score in the college test, on the average?")

The two regression lines are shown on the scatter diagram in Fig. 11.1

Lx 11.4 The bivariate frequency distribution given in Table 11.2 may now be considered. Here the variables are family income (x) and percentage of family income spent on food (y). Here, again, the regression of y on x is important from the practical point of view. On the basis of the calculations done in connection with Ex. 11.2, we have

$$x = 275.5 + 50 \times \frac{\sum U_i}{n} = 275.5 + 50 \times \frac{(-5)}{200} = 274.25$$

and $y = 39.5 + 2 \times \frac{\sum V_i}{n} = 39.5 + 2 \times \frac{(-103)}{200} = 38.47$

The regression coefficient of y on x is

$$\begin{aligned} b_{y,x} &= \frac{2}{50} \times b_{r,u} = \frac{2}{50} \times \frac{n \sum_i u_i V_i - (\sum_j U_j)(\sum_i V_i)}{n \sum_i u_i^2 f_{i0} - (\sum_j U_j)^2} \\ &= \frac{2}{50} \times \frac{(-156,515)}{144,975} = -0.0432. \end{aligned}$$

Hence the regression equation of y on x is

$$Y = 38.47 - 0.0432(x - 274.25)$$

or

$$Y = 50.32 - 0.0432x.$$

This, again, is expected to serve as a good prediction formula for y , given x , since the correlation coefficient between x and y is numerically quite high.

11.9 Theoretical distribution of two variables

As in the univariate case, here, too, we look for a simple mathematical function of x and y to represent the joint distribution of x and y in the population. If x and y are both continuous variables, then this distribution is defined by a probability-density function, $f(x, y)$, which is such that

$$\int_c^d \int_a^b f(x, y) dx dy$$

gives the compound probability that x lies in the interval (a, b) and y in the interval (c, d) , whatever the intervals may be. If the whole-range of x is from α to β and that of y is from γ to δ , then

$$\int_\gamma^\delta f(x, y) dy = g(x)$$

and

$$\int_\alpha^\beta f(x, y) dx = h(y)$$

define the marginal distributions of x and y , respectively.

Again,

$$\frac{f(x, y)}{g(x)} = f(y|x)$$

defines the conditional distribution or array distribution of y for

given x such that $g(x) > 0$ * Similarly,

$$\frac{f(x, y)}{h(y)} = f(x|y)$$

defines the conditional distribution of x for given y such that $h(y) > 0$

If the conditional distributions of one variable for given values of the other are all identical or, equivalently, if

$$f(x, y) = g(x) h(y)$$

for all x and y , then the two variables are said to be *independent*. Otherwise, they are said to be *associated*.

The means of these conditional distributions, say η_x and ξ , are of special importance. The term 'regression' actually refers to the relationship between η_x and x or between ξ , and y . The equation that expresses the conditional mean η_x as a function of x is the *regression equation of y on x* . If η_x is of the form

$$\eta_x = \alpha + \beta x, \quad (11.30a)$$

then the regression is said to be linear. In fitting a regression line to observed data, we thus attempt to estimate the true regression, assuming that it is linear. (11.30a) may be expressed in terms of the means (μ_x and μ_y), standard deviations (σ_x and σ_y) and correlation coefficient (ρ) as

$$\eta_x = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x) \quad (11.30b)$$

Obviously, $\rho \sigma_y / \sigma_x$ is the regression coefficient of y on x .

The regression of x on y similarly refers to the relationship between ξ , and y . In case this regression is linear,

$$\xi = \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y) \quad (11.31)$$

It will be sufficient for our purpose to consider one theoretical distribution of the bivariate type, viz. the *bivariate normal distribution*.

If x and y are two variables, then they are said to be distributed in the bivariate normal form if their probability density function is

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} \left\{ \frac{(x-\mu_x)^2}{\sigma_x^2} \right. \right. \\ \left. \left. - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right\} \right], \\ -\infty < x < \infty, -\infty < y < \infty \quad (11.32)$$

*If $g(x)=0$, then the conditional distribution is not defined (cf. Section 3.5).

The main properties of this distribution are stated below :

(1) If x and y are jointly normally distributed, then the marginal distribution of each variable is of the (univariate) normal form :

$$g(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \quad \dots \quad (11.33)$$

and $h(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right]. \quad \dots \quad (11.34)$

(2) The conditional distributions of each variable for given values of the other are also of the (univariate) normal form :

$$f(y|x) = \frac{1}{\sigma_y \sqrt{1-\rho^2} \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_y^2(1-\rho^2)} \left\{ y - \mu_y - \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x) \right\}^2\right], \quad \dots \quad (11.35)$$

$$f(x|y) = \frac{1}{\sigma_x \sqrt{1-\rho^2} \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_x^2(1-\rho^2)} \left\{ x - \mu_x - \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y) \right\}^2\right]. \quad \dots \quad (11.36)$$

(3) It is seen from (2) that the means of the conditional distributions of y for given values of x are given by

$$\eta_x = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x) \quad \dots \quad (11.37)$$

and those of the conditional distributions of x for given values of y by

$$\xi_y = \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y). \quad \dots \quad (11.38)$$

Thus the regression of y on x and that of x on y are both linear in case x and y are distributed in the bivariate normal form.

(4) From (2) it is also seen that the standard deviation of the conditional distribution of y for any given value of x is

$$\sigma_{y,x} = \sigma_y \sqrt{1 - \rho^2}. \quad \dots \quad (11.39)$$

These conditional distributions are, therefore, *homoscedastic* (i.e. have equal dispersion). Similarly, the conditional distributions of x for given values of y are homoscedastic, each having the standard deviation

$$\sigma_{x,y} = \sigma_x \sqrt{1 - \rho^2}.$$

(5) In the context of a bivariate normal distribution, the correlation coefficient ρ has a precise meaning.

Thus if $\rho=0$, then $f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left\{\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right\}\right]$,
 $=g(x) h(y)$. This means that if x and y are uncorrelated, then they are also independent.

Again, if $\rho=\pm 1$, then (considering the conditional distributions of y for fixed values of x) $\sigma_y\sqrt{1-\rho^2}=0$. So here the values of y in each array are exactly equal to the array-mean, and since the array-mean is a linear function of x , $\rho=\pm 1$ implies that there is an exact functional relationship (of the linear type) between x and y .

Furthermore, since $\sigma_y\sqrt{1-\rho^2}$ decreases as ρ increases numerically, it may be said that the higher the numerical value of ρ , the nearer are x and y to linear functional relationship.

11.10 Limitations of the correlation coefficient

Because of the close connection between correlation coefficient and linear regression, it is clear that the former can serve as a

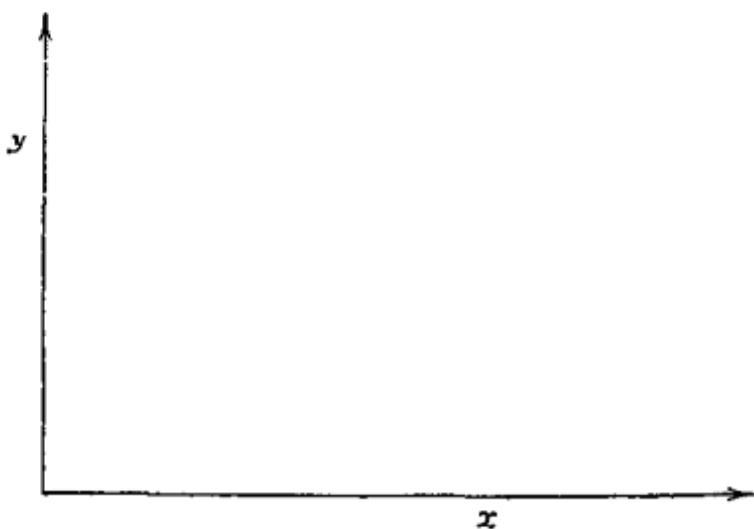


Fig 11.4 Scatter diagram, showing close non linear relationship between x and y

satisfactory measure of the relationship between two variables only when that relationship is of the linear type. Hence a low value of the correlation coefficient does not rule out the possibility that the variables are related in some other manner. If x and y have a non-linear relationship like that in Fig 11.4, the least-square regression lines will be approximately parallel to the axes of co ordinates and

hence r will be very small, although actually there may be a strong relationship between the variables.

Consider, e.g., the following data, which, although artificial, show that the correlation coefficient may be zero even when there is perfect (functional) dependence of one variable on another and may, therefore, totally fail to measure the relationship between the two variables :

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	5	6	7	8	9	10	9	8	7	6	5

Here

$$y = \begin{cases} 10+x & \text{for } x = -1, -2, \dots, -5 \\ 10-x & \text{for } x = 0, 1, 2, \dots, 5, \end{cases}$$

and hence y is an exact function of x .

But

$$\bar{x} = 0$$

and also

$$\sum_i x_i y_i = 0.$$

Hence

$$\text{cov}(x, y) = \frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y} = 0,$$

implying that the correlation coefficient also is zero,

It is advisable, therefore, that one should see whether the general relationship between x and y is linear or not (by drawing the scatter diagram, for instance) before using r as a measure of this relationship.

Again, the fact that two variables are correlated does not necessarily mean that they are causally related—that one variable is the cause of the other. Indeed, the two variables may appear to be correlated even when both are caused by some other variable or variables (*vide* Section 12.9).

A spurious correlation may also arise from the non-homogeneity of data. Thus suppose a number of groups of individuals, each by itself homogeneous but having varying means, are mixed up. Then

x and y may appear to be highly correlated in the combined group, even if in each of the constituent groups the variables have little or no correlation (see Example 11.14)

11.11 Correlation index and correlation ratio

When the general relationship between x and y is non linear, the correlation coefficient fails to measure the extent of their interdependence. This happens because of the close link between correlation coefficient and linear regression, as will be apparent from Section 11.8

In case the regression of any one variable (say, y) on another (say, x) is non linear, we may still like to devise a measure of the dependence of the first on the second.

We have seen in Section 11.8 that

$$\begin{aligned} r^2 &= \frac{\text{var}(Y)}{\text{var}(y)} \\ &= \frac{\sum_i (Y_i - \bar{Y})^2}{\sum_i (y_i - \bar{y})^2} \end{aligned}$$

where \bar{Y}_i is the predicted value of y from the linear regression equation when $x=x_i$.

In other words, r^2 may be interpreted as the proportion of the total variability of y which is accounted for by its linear regression on x .

This concept can be generalised. Suppose the appropriate regression equation is a polynomial of the p th degree ($p \leq n-1$). Then we can define a measure of association, similar to r^2 , called the *correlation index* of the p th order, r_p^2 , say, by

$$\begin{aligned} r_p^2 &= \frac{\text{var}(\bar{Y}_p)}{\text{var}(y)} \\ &= \frac{\sum_i (\bar{Y}_{pi} - \bar{Y}_p)^2}{\sum_i (y_i - \bar{y})^2}, \end{aligned} \quad (11.40)$$

where \bar{Y}_{pi} is the predicted value of y from the p th degree polynomial regression equation when $x=x_i$.

Now,

$$\begin{aligned} \sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - Y_{pi})^2 + \sum_i (Y_{pi} - \bar{y})^2, \\ &\quad \text{the product term vanishing by} \\ &\quad \text{virtue of the normal equations} \\ &\quad \text{determining } Y_{pi} \\ &= \sum_i (y_i - Y_{pi})^2 + \sum_i (Y_{pi} - \bar{Y}_p)^2 \quad \dots \quad (11.41) \\ &\quad (\text{since } \bar{y} = \bar{Y}_p). \end{aligned}$$

We shall show that as the degree of the polynomial increases, the value of the correlation index also increases.

Let the polynomial regression equations of degree p and degree $p-1$ be, respectively,

$$Y = a_0 + a_1 x + a_2 x^2 + \dots + a_p x^p$$

and

$$Y = a'_0 + a'_1 x + a'_2 x^2 + \dots + a'_{p-1} x^{p-1},$$

the constants in each case being determined by the least-square method.

By definition, then,

$$\sum_i (y_i - a_0 - a_1 x_i - \dots - a_p x_i^p)^2 \leq \sum_i (y_i - b_0 - b_1 x_i - \dots - b_p x_i^p)^2,$$

whatever the alternative set of constants b 's may be. Taking, in particular, $b_0 = a'_0$, $b_1 = a'_1$, \dots , $b_{p-1} = a'_{p-1}$, $b_p = 0$, we have thus

$$\begin{aligned} &\sum_i (y_i - a_0 - a_1 x_i - \dots - a_p x_i^p)^2 \\ &\leq \sum_i (y_i - a'_0 - a'_1 x_i - \dots - a'_{p-1} x_i^{p-1})^2, \end{aligned}$$

i.e.

$$\sum_i (y_i - Y_{pi})^2 \leq \sum_i (y_i - Y_{p-1,i})^2.$$

This implies, by virtue of (11.41), that

$$\sum_i (Y_{pi} - \bar{Y}_p)^2 \geq \sum_i (Y_{p-1,i} - \bar{Y}_{p-1})^2.$$

Hence for any p ,

$$r_p^2 \geq r_{p-1}^2.$$

Taking $p=2, 3, \dots, n$, successively, we thus have

$$r^2 \leq r_2^2 \leq r_3^2 \leq \dots \leq r_{n-1}^2.$$

We shall now introduce a more general measure of the degree of dependence of one variable on another. To this end, suppose the n pairs of values of x and y are arranged in arrays of y according to fixed values of x .

x_1	$y_{11}, y_{12}, \dots, y_{1n_1}$
x_2	$y_{21}, y_{22}, \dots, y_{2n_2}$
x_k	$y_{k1}, y_{k2}, \dots, y_{kn_k}$

This means that for n_i of the n individuals the values of x are the same, viz. x_i , while the values of y are $y_{i1}, y_{i2}, \dots, y_{in_i}$.

Suppose further that

$$\bar{y}_i = \sum_j y_{ij} / n_i, \quad (11.42)$$

the mean of the i th array, and

$$\bar{y} = \sum_i \sum_j y_{ij} / \sum_i n_i = \sum_i n_i \bar{y}_i / n, \quad (11.43)$$

the grand mean of y . Also, the mean of x is

$$\bar{x} = \sum_i n_i x_i / n \quad (11.44)$$

If the regression of y on x were linear, a measure of the dependence of y on x (or of the interdependence of x and y) would be $|r|$, given by

$$r = \frac{\sum_i n_i (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum_i n_i (\bar{x}_i - \bar{x})^2} \sqrt{\sum_i \sum_j (y_{ij} - \bar{y})^2}}, \quad (11.45)$$

and, in analogy with (11.28), we would have

$$r^2 = \frac{\sum_i n_i (\bar{x}_i - \bar{x})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2}, \quad (11.45a)$$

where \bar{x}_i is the value of x corresponding to x_i , as given by the regression line of y on x .

The values $Y_i = a + bx_i$ are obtained by minimising

$$\sum_i \sum_j (y_{ij} - a - bx_i)^2$$

or, equivalently, by minimising

$$\sum_i n_i (\bar{y}_i - a - bx_i)^2 \quad \dots \quad (11.46)$$

with respect to a and b , the normal equations being in each case

$$\begin{cases} \sum_i n_i \bar{y}_i = na + b \sum_i n_i x_i, \\ \sum_i n_i x_i \bar{y}_i = a \sum_i n_i x_i + b \sum_i n_i x_i^2. \end{cases} \quad \dots \quad (11.47)$$

Now it would be apparent from (11.46) that Y_i is just an approximation to \bar{y}_i . The expression (11.45) also shows that $|r|$ really purports to be a measure of the extent to which the array mean of y for given x depends on x . In case the regression is linear, i.e. in case $Y_i = \bar{y}_i$ for each i , $|r|$ achieves this purpose.

But, as we have already stated, in our case the regression is not linear and an alternative measure is called for. The discussion in the preceding paragraph suggests such a modification of $|r|$, through the replacement of Y_i in (11.45a) by \bar{y}_i . The new measure obtained, called the *correlation ratio of y on x* and denoted by e_{yx} , is thus the positive square-root of

$$e_{yx}^2 = \frac{\sum_i n_i (\bar{y}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2}. \quad \dots \quad (11.48)$$

In case the data are grouped into a $k \times l$ bivariate frequency table, the k classes of x may be supposed to give k arrays of y values. For the i th array, the total frequency is f_{i0} and the array mean is

$$\bar{y}_i = \sum_j f_{ij} y_j / f_{i0},$$

while the grand mean is

$$\bar{y} = \sum_j f_{0j} y_j / n.$$

The correlation ratio will then be given by the formula

$$e_{yx}^2 = \frac{\sum_i f_{i0} (\bar{y}_i - \bar{y})^2}{\sum_j f_{0j} (y_j - \bar{y})^2}. \quad \dots \quad (11.48a)$$

Note that

$$\begin{aligned}\sum_{i,j} (y_{ij} - \bar{y})^2 &= \sum_{i,j} \{(\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)\}^2 \\ &= \sum_i n_i (\bar{y}_i - \bar{y})^2 + \sum_{i,j} (y_{ij} - \bar{y}_i)^2,\end{aligned}$$

and, on dividing both sides by n , we may write this as

$$s^2 = e_{yx}^2 s^2 + (1 - e_{yx}^2) s^2. \quad \dots \quad (11.49)$$

Also,

$$\begin{aligned}\sum_i n_i (\bar{y}_i - \bar{y})^2 &= \sum_i n_i \{Y_i - \bar{y}\} + (\bar{y}_i - Y_i)^2 \\ &= \sum_i n_i (Y_i - \bar{y})^2 + \sum_i n_i (\bar{y}_i - Y_i)^2,\end{aligned}$$

the sum of the product-terms vanishing because of the normal equations yielding \bar{Y}_i . Again, dividing both sides by n , we may write this as

$$e_{yx}^2 s^2 = r^2 s^2 + (e_{yx}^2 - r^2) s^2. \quad \dots \quad (11.50)$$

Since

$$1 - e_{yx}^2 = \frac{1}{n} \sum_{i,j} (y_{ij} - \bar{y}_i)^2 / s^2 \geq 0$$

$$\text{and } e_{yx}^2 - r^2 = \frac{1}{n} \sum_i n_i (\bar{y}_i - Y_i)^2 / s^2 \geq 0,$$

we immediately have

$$r^2 \leq e_{yx}^2 \leq 1. \quad \dots \quad (11.51)$$

If follows from (11.50) that $e_{yx}^2 = r^2$ if, and only if, $\bar{Y}_i = \bar{y}_i$ for each i . Hence the difference $e_{yx}^2 - r^2$ measures the extent to which the true regression of y on x (i.e. the true relationship between the array mean of y for given x and y) departs from linearity. This is another, and the more common, use to which the correlation ratio is put.

Actually, the correlation ratio serves as the least upper bound to the numerical value of the correlation coefficient or a correlation index.

Just as r^2 may be interpreted as the proportion of the total variation of y that is explained by its linear regression on x , e_{yx}^2 may be interpreted, because of (11.49), as the proportion that is explained by the array means of y for given x values. But whereas the correlation coefficient is symmetrical in x and y , the correlation ratio is not, so that generally e_{yx} and e_{xy} will not be equal.

Questions and exercises

11.1 Distinguish between the correlation approach and the regression approach to the analysis of bivariate data.

11.2 Define correlation coefficient. State and prove its important properties. What are its limitations?

11.3 Define the term *regression*. State and prove all important results relating to regression lines.

11.4 Explain what is meant by a theoretical distribution in two variables. Define marginal distributions and array distributions.

11.5 Define the bivariate normal distribution and state its important properties.

11.6 Using Cauchy-Schwarz inequality, prove that r lies between -1 and $+1$. Hence interpret the cases $r = \pm 1$.

11.7 Show that the angles between the two regression lines (of y on x and of x on y) are $\tan^{-1} \left(\pm \frac{1-r^2}{r} \cdot \frac{s_x s_y}{s_x^2 + s_y^2} \right)$ and interpret the cases where $r=0$ and $r=\pm 1$.

11.8 What is correlation ratio? Show that $0 \leq r^2 \leq e_{yx}^2 \leq 1$ and discuss the following cases:

$$e_{xy}^2 = r^2, e_{yx}^2 = e_{xz}^2 = r^2, e_{yz}^2 = 1, e_{yz}^2 = e_{xz}^2 = 1.$$

11.9(a) Show that formula (11.48) for the correlation ratio may be reduced to the form

$$e_{yz}^2 = (\sum_i n_i \bar{y}_i^2 - n \bar{y}^2) / (\sum_{ij} y_{ij}^2 - n \bar{y}^2) = \frac{\sum_i (\sum_j y_{ij})^2 / n_i - (\sum_i \sum_j y_{ij})^2 / n}{\sum_{ij} y_{ij}^2 - (\sum_i \sum_j y_{ij})^2 / n}.$$

(b) Show that for a bivariate frequency table the correlation ratio may be computed by the formula

$$e_{yz}^2 = \frac{\sum_i V_i^2 / f_{i0} - (\sum_i V_i)^2 / n}{\sum_j v_j^2 f_{0j} - (\sum_i V_i)^2 / n}.$$

Hence determine the correlation ratio of y on x for Table 11.2 and comment.

Partial ans. $e_{yz} = 0.858$.

11.10 Show that the correlation ratio e_{yz} is the simple correlation coefficient between y and the array mean of y corresponding to x .

11.11 Let x and y be independent variables with standard deviations σ_x and σ_y . Show that the correlation coefficient between x and $x+y$ is

$$\sigma_x / \sqrt{\sigma_x^2 + \sigma_y^2}$$

11.12(a) Let x and y be jointly normally distributed with equal means and equal variances and with a positive correlation coefficient. Show that the conditional mean of y for a given $x > \mu$ will be less than x and *vice versa* (where μ is the common mean of x and y)

(b) Hence account for the paradox that for a tall father, on the average, the son (adult) though tall is shorter than the father himself, while for a tall son (adult), on the average, the father though tall is shorter than the son himself. (The feature has been termed 'regression to mediocrity' and is at the origin of the use of the term 'regression' in statistical literature.)

11.13 What would be the normal equations if a polynomial regression $Y = a_0 + a_1x + \dots + a_nx^n$ were fitted to the given values of x and y ? Suggest a measure of the usefulness of this regression equation as a predicting formula

11.14 Let there be k groups of data on x and y , with means \bar{x}_i and \bar{y}_i , variances $s_{x_i}^2$ and $s_{y_i}^2$, and correlations r_i ($i=1, 2, \dots, k$). Show that the correlation for the combined data is

$$r = \frac{\sum n_i r_i s_{x_i} s_{y_i} + \sum n_i (x_i - \bar{x})(y_i - \bar{y})}{[\sum_i n_i s_{x_i}^2 + \sum_i n_i (x_i - \bar{x})^2]^{1/2} [\sum_i n_i s_{y_i}^2 + \sum_i n_i (y_i - \bar{y})^2]^{1/2}}$$

where \bar{x} and \bar{y} are the grand means of x and y and n_i is the number of pairs in the i th group.

Hence explain the phenomenon that r_i may be zero for each i and yet r may be non zero.

11.15 Let x and y be subject to observational errors, so that what one observes are $x = x + \epsilon_x$ and $y = y + \epsilon_y$, instead of x and y . If ϵ_x and ϵ_y are independent of x and y , and if ϵ_x and ϵ_y are also mutually independent, show that the correlation coefficient between x and y becomes numerically smaller owing to the errors. (This is called the 'attenuation effect')

11.16 A gunner is aiming at the centre of a rectangular target, 40 ft. wide and 60 ft. high, at a given distance. Suppose the actual point of hit is (x, y) with origin at the centre, such that x and y are independently normally distributed with zero means and standard deviations $\sigma_x = 16$ ft. and $\sigma_y = 20$ ft. If four shots are fired, find the probability that the target will be hit at least once. *Ans.* 0.9899.

11.17 The following data relate to the stature (x) and sitting height (y), both in cm., for each of 30 people of a particular Indian caste :

x	y	x	y	x	y
172.8	83.9	157.7	77.7	170.2	83.4
166.0	83.6	146.7	76.4	153.3	79.1
164.1	81.3	153.2	77.2	173.7	86.1
164.4	85.4	155.3	80.1	155.8	78.6
168.8	83.9	151.5	76.9	158.0	80.1
165.2	81.1	161.1	81.5	157.2	81.6
170.0	84.9	156.3	80.9	156.2	78.6
163.5	81.1	169.4	83.1	168.2	82.5
169.4	84.9	159.9	84.2	164.4	84.1
159.1	79.6	161.7	80.3	165.5	87.1

(1) Represent the data by means of a scatter diagram.

(2) Compute the correlation coefficient of x and y .

Ans. $r=0.839$.

(3) Obtain the linear regression equations of y on x and of x on y . Hence determine what the sitting height of a man is expected to be if his stature is 175 cm.

11.18 During an investigation in an agricultural farm in Bengal, the length (in cm.) of green jute plant and the weight (in gm.) of dry jute fibre were observed for 350 plants. With these data the following bivariate frequency table was obtained :

	Length of green plant (cm) class mark							Total
	111.5	127.5	143.5	159.5	175.5	191.5	207.5	
Weight of dry jute fibre (gm) class-mark	3 175	32	25	15	1			53
	2 775	1	4	33	59	29	3	129
	4 375	1		4	28	35	14	84
	5 975				2	20	18	1
	7 575				1	1	14	5
	9 175					4	8	2
	10 775						3	2
	12 375							3
Total	14	29	52	91	89	60	15	350

From the above table, compute the coefficient of correlation of the two variables. Also, find the linear regression equation of weight of dry fibre on length of green plant.

Ans $r=0.755$, regression eqn $\bar{Y}=8.379+0.0756x$

II 19 The difference between upper face length (y) and nasal length (x), both measured in mm, is given for 15 Indian adult males

15	13	15
15	12	19
13	12	19
14	15	19
11	15	14

Calculate the correlation coefficient of x and y and the linear regression of y on x , given that for these people $\bar{x}=49.34$ mm, $\bar{y}=64.07$ mm, $s_x=3.53$ mm and $s_y=4.30$ mm

Partial ans $r=0.820$, $b_{xy}=0.999$

11.20 For 20 Army personnel, the regression of weight of kidneys (y) on weight of heart (x), both measured in oz., is

$$Y = 0.399x + 6.934,$$

and the regression of weight of heart on weight of kidneys is

$$X = 1.212y - 2.461.$$

Find the correlation between the two variables and also their means. Can you find their s.d.s as well?

Partial ans. $r = 0.695$, $\bar{x} = 11.509$, $\bar{y} = 11.526$.

11.21 Consider the following data :

x	-4	-3	-2	-1	0	1	2	3	4
y	0.1	2.5	3.4	3.9	4.1	3.8	3.5	2.8	0.3

Find r_{xy} and comment.

Partial ans. $r = 0.054$.

11.22 The figures of production of crude petroleum and production of wheat flour in India are given below for a number of years :

Year	1958	1959	1960	1961	1962	1963	1964	1965
Production of crude petroleum (000 tonnes)	439	450	454	514	1,077	1,653	2,212	2,176
Production of wheat flour (000 tonnes)	830	976	995	1,002	1,202	1,418	1,819	1,604

Compute the correlation coefficient and comment.

(Such a purely accidental correlation between two time series—one having no causal significance—is called *nonsense correlation*.)

SUGGESTED READING

- [1] Ezekiel, M. and Fox, K. A. *Methods of Correlation and Regression Analysis* (Chs. 5—9). John Wiley, 1959.
- [2] Goulden, C. H. *Methods of Statistical Analysis* (Chs. 6—7). John Wiley, 1952, and Asia Publishing House, 1959.

- [3] Kenney, J. F. and Keeping, E. S. *Mathematics of Statistics*, Part I (Ch. 15). Van Nostrand, 1954, and Affiliated East-West Press.
- [4] Moore, P. G. *Practical Statistical Techniques* (Ch. 14). Cambridge University Press, 1956.
- [5] Snedecor, G. W. *Statistical Methods* (Chs. 6, 7, 15). Iowa State College Press, 1956, and Allied Pacific, 1961.
- [6] Yule, G. U. "Why do we sometimes get nonsense correlation between time-series, etc?" *Jour. Roy. Stat. Soc.*, 82 (1926), pp. 1-.

12.1 Multivariate data

In some investigations, data may be collected, for the given set of individuals, on a number of (more than two) variables at the same time. Thus the data may relate to the scores obtained by each of a number of high school students in each of five subjects, say, English, major vernacular, mathematics, history and elementary science. In an agricultural experiment with a variety of wheat, the data may relate to the amount of fertiliser added, the number of man-hours spent in tilling the land, the yield of crop and the yield of straw.

Such data may also be arranged into a frequency distribution as in the univariate or the bivariate case. If there are p variables x_1, x_2, \dots, x_p , with k_1, k_2, \dots, k_p , classes, respectively, the distribution will be represented by $k_1 \times k_2 \times \dots \times k_p$ cells, the frequency in each cell being the number of individuals belonging simultaneously to the corresponding x_1 -class, x_2 -class, ..., x_p -class. From this joint distribution of the p variables, one may also obtain the *marginal distribution* of any p' of the variables ($1 \leq p' \leq p-1$) or the *conditional distribution* of any p' of the variables for given values of p'' of the other variables ($p', p'' \geq 1$ and $p'+p'' \leq p$). These are defined here in a way similar to that in the bivariate case.

12.2 Multiple regression

As with bivariate data, here too it may be that one of the p variables, say x_1 , is of primary interest to us and we consider x_2, x_3, \dots, x_p , together with x_1 in view of their possible influence on the latter. The object may be to build up a relationship between the 'dependent variable', x_1 , and the 'independent variables', x_2, x_3, \dots, x_p , with the idea of using this relationship for predicting the value of the dependent variable from a knowledge of the values of the independent variables. Thus, in estimating the rainfall at a place in a year, it is appropriate to consider the effect of the latitude, the longitude and the altitude of the place on rainfall. Similarly, in estimating the

yield of a crop in a year, it is proper to take into account the effect of, say, rainfall, average temperature and average humidity, during the period between the sowing and the harvesting of the crop. And common sense suggests that the higher the number of independent variables, the better is the prediction likely to be.

Let us assume that the relationship between x_1 and x_2, x_3, \dots, x_p , is, at least in an approximate sense, given by an equation of the form

$$X_{1 \text{ vs } p} = a + b_2 x_2 + b_3 x_3 + \dots + b_p x_p \quad (121)$$

Our data here will consist of p values, corresponding to the p variables, for each of n individuals. The values of the variables for the α th individual may be denoted by $x_{1\alpha}, x_{2\alpha}, \dots, x_{p\alpha}$ ($\alpha = 1, 2, \dots, n$).

In order to determine the constants a, b_2, b_3, \dots, b_p on the basis of the data, we again make use of the least square method.

If we denote by $x_{1 \text{ vs } p}$ the difference $x_{1\alpha} - X_{1 \text{ vs } p}$, then the error of estimate corresponding to the α th individual is $x_{1 \text{ vs } p, \alpha}$. The least-square method means that the constants a, b_2, b_3, \dots, b_p are to be so determined that

$$\sum_{\alpha} x_{1 \text{ vs } p, \alpha}^2 = \sum_{\alpha} (x_{1\alpha} - a - b_2 x_{2\alpha} - \dots - b_p x_{p\alpha})^2 \quad (122)$$

is a minimum.

The normal equations in this case (obtained by equating to zero the partial derivatives of (122) with respect to a, b_2, b_3, \dots, b_p) are

$$\left. \begin{aligned} \sum_{\alpha} x_{1 \text{ vs } p, \alpha} &= 0, \\ \sum_{\alpha} x_{2\alpha} x_{1 \text{ vs } p, \alpha} &= 0 \\ \sum_{\alpha} x_{3\alpha} x_{1 \text{ vs } p, \alpha} &= 0, \\ \sum_{\alpha} x_{p\alpha} x_{1 \text{ vs } p, \alpha} &= 0 \end{aligned} \right\} \quad (123a)$$

or

$$\left. \begin{aligned} \sum_{\alpha} x_{1\alpha} &= na + b_2 \sum_{\alpha} x_{2\alpha} + b_3 \sum_{\alpha} x_{3\alpha} + \dots + b_p \sum_{\alpha} x_{p\alpha}, \\ \sum_{\alpha} x_{2\alpha} x_{1\alpha} &= a \sum_{\alpha} x_{2\alpha} + b_2 \sum_{\alpha} x_{2\alpha}^2 + b_3 \sum_{\alpha} x_{2\alpha} x_{3\alpha} + \dots + b_p \sum_{\alpha} x_{2\alpha} x_{p\alpha}, \\ \sum_{\alpha} x_{3\alpha} x_{1\alpha} &= a \sum_{\alpha} x_{3\alpha} + b_2 \sum_{\alpha} x_{3\alpha} x_{2\alpha} + b_3 \sum_{\alpha} x_{3\alpha}^2 + \dots + b_p \sum_{\alpha} x_{3\alpha} x_{p\alpha}, \\ \sum_{\alpha} x_{p\alpha} x_{1\alpha} &= a \sum_{\alpha} x_{p\alpha} + b_2 \sum_{\alpha} x_{p\alpha} x_{2\alpha} + b_3 \sum_{\alpha} x_{p\alpha} x_{3\alpha} + \dots + b_p \sum_{\alpha} x_{p\alpha}^2 \end{aligned} \right\} \quad (123b)$$

The first equation gives, on being divided by n ,

$$\bar{x}_1 = a + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \dots + b_p \bar{x}_p, \quad \dots \quad (12.4)$$

which shows incidentally that the mean point $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ necessarily satisfies the prediction equation.

Multiplying (12.4) by $n\bar{x}_2, n\bar{x}_3, \dots, n\bar{x}_p$, and subtracting from the second, third, \dots, p th equation, respectively, of the system (12.3b), we have $(p-1)$ equations determining the b 's, viz.

$$\left. \begin{aligned} S_{21} &= b_2 S_{12} + b_3 S_{13} + \dots + b_p S_{1p}, \\ S_{31} &= b_2 S_{22} + b_3 S_{23} + \dots + b_p S_{2p}, \\ &\vdots \\ S_{p1} &= b_2 S_{p2} + b_3 S_{p3} + \dots + b_p S_{pp}, \end{aligned} \right\} \quad \dots \quad (12.5)$$

where

$$S_{ij} = \sum_a x_{ia} x_{ja} - n \bar{x}_i \bar{x}_j = \sum_a (x_{ia} - \bar{x}_i)(x_{ja} - \bar{x}_j). \quad \dots \quad (12.6)$$

It will be assumed that the matrix

$$\begin{pmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{pmatrix}$$

or, equivalently, that the *variance-covariance matrix*

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}, \quad \dots \quad (12.7)$$

where

$$s_{ij} = \frac{1}{n} S_{ij} = \begin{cases} \text{cov}(x_i, x_j) & \text{if } i \neq j \\ \text{var}(x_i) & \text{if } i = j, \end{cases} \quad \dots \quad (12.8)$$

is non-singular (i.e. is of rank p).*

This will mean, among other things, that

$$\begin{pmatrix} b_2 \\ b_3 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} s_{22} & s_{23} & \dots & s_{2p} \\ s_{32} & s_{33} & \dots & s_{3p} \\ \vdots & \vdots & & \vdots \\ s_{p2} & s_{p3} & \dots & s_{pp} \end{pmatrix}^{-1} \begin{pmatrix} S_{21} \\ S_{31} \\ \vdots \\ S_{p1} \end{pmatrix} \quad \dots \quad (12.9)$$

*This will mean that this symmetric matrix is positive definite.

or that

$$\begin{aligned}
 b_j &= s_{21}s^{j,1} + s_{31}s^{j,3} + \dots + s_{p1}s^{j,p} \\
 &= \frac{\left| \begin{array}{cccccc} s_{22} & s_{23} & s_{2(j-1)} & s_{21} & s_{2(j+1)} & s_{2p} \\ s_{32} & s_{33} & s_{3(j-1)} & s_{31} & s_{3(j+1)} & s_{3p} \\ s_{p2} & s_{p3} & s_{p(j-1)} & s_{p1} & s_{p(j+1)} & s_{pp} \end{array} \right|}{\left| \begin{array}{ccc} s_{22} & s_{23} & s_{2p} \\ s_{32} & s_{33} & s_{3p} \\ s_{p2} & s_{p3} & s_{pp} \end{array} \right|} . \quad (12.9a) \\
 &\qquad\qquad\qquad (\text{for } j=2, 3, \dots, p)
 \end{aligned}$$

Since $s_{ij} = r_{ij} s_i s_j$, where

s_i = standard deviation of x_i ,

$s_j = \dots, \dots, \dots, x_j$

and

r_{ij} = correlation coefficient of x_i and x_j ,

we may also write, on simplifying (12.9a),

$$b_j = (-1)^{j-2} \frac{s_1}{s_j} \left| \begin{array}{ccccc} r_{21} & r_{22} & r_{2(j-1)} & r_{2(j+1)} & r_{2p} \\ r_{31} & r_{32} & r_{3(j-1)} & r_{3(j+1)} & r_{3p} \\ r_{p1} & r_{p2} & r_{p(j-1)} & r_{p(j+1)} & r_{pp} \\ r_{22} & r_{23} & & & r_{2p} \\ r_{32} & r_{33} & & & r_{3p} \\ r_{p2} & r_{p3} & & & r_{pp} \end{array} \right| . \quad (12.10)$$

We shall write R for the matrix

$$\left(\begin{array}{cccc} r_{11} & r_{12} & r_{13} & r_{1p} \\ r_{21} & r_{22} & r_{23} & r_{2p} \\ r_{31} & r_{32} & r_{33} & r_{3p} \\ r_{p1} & r_{p2} & r_{p3} & r_{pp} \end{array} \right) , \quad (12.11)$$

which is the *correlation matrix* of x_1, x_2, \dots, x_p . R for the corresponding determinant and R_{ij} for the co factor of r_{ij} in R . We see that the determinant in the numerator of (12.10) is the minor of r_{jj} (in R) and hence $(-1)^{1+j} \times$ the co factor of r_{jj} , while the

determinant in the denominator is the minor (and also the co-factor) of r_{11} . Hence

$$\begin{aligned} b_j &= (-1)^{2j-1} \cdot \frac{s_1}{s_j} \cdot \frac{R_{1j}}{R_{11}} \\ &= -\frac{R_{1j}}{R_{11}} \cdot \frac{s_1}{s_j}, \quad \text{for } j=2, 3, \dots, p, \quad \dots \quad (12.12) \end{aligned}$$

while from (12.4)

$$a = \bar{x}_1 + \sum_{j=2}^p \frac{R_{1j}}{R_{11}} \cdot \frac{s_1}{s_j} \bar{x}_j. \quad \dots \quad (12.13)$$

Thus the prediction equation (called the *multiple regression equation of x_1 on x_2, x_3, \dots, x_p*) becomes

$$\begin{aligned} X_{1 \cdot 2 \cdot 3 \dots p} &= \bar{x}_1 - \frac{R_{12}}{R_{11}} \cdot \frac{s_1}{s_2} (x_2 - \bar{x}_2) - \frac{R_{13}}{R_{11}} \cdot \frac{s_1}{s_3} (x_3 - \bar{x}_3) - \dots \\ &\quad - \frac{R_{1p}}{R_{11}} \cdot \frac{s_1}{s_p} (x_p - \bar{x}_p). \quad \dots \quad (12.14) \end{aligned}$$

The coefficient $b_j = -\frac{R_{1j}}{R_{11}} \cdot \frac{s_1}{s_j}$ in (12.14) is called the *partial regression coefficient of x_1 on x_j* for fixed $x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_p$, and is often written in the more explicit form

$$b_{1j \cdot 2 \cdot 3 \dots (j-1)(j+1) \dots p}. \quad \dots \quad (12.15)$$

Evidently, it gives the amount by which the predicted value $X_{1 \cdot 2 \cdot 3 \dots p}$ increases when x_j increases by a unit amount, the other independent variables being kept fixed.

Ex. 12.1 The following table shows, for each of 18 cinchona plants, the yield of dry bark (in oz.), the height (in inches) and the girth (in inches) at a height of 6" from the ground.

Supposing we denote these variables by x_1 , x_2 and x_3 , respectively, it is of practical interest to study the dependence of x_1 on x_2 and x_3 . For this purpose, let us first determine the multiple regression equation of x_1 on x_2 and x_3 . The preliminary calculations are shown in Table 12.2.

TABLE 121
YIELD OF DRY BARK, HEIGHT AND GIRTH AT A LEVEL 6'
ABOVE GROUND FOR 18 CINCHOVA PLANTS

Plant No	Yield of dry bark (oz)	Height (in)	Girth at a height of 6' (in)	Plant No	Yield of dry bark (oz)	Height (in)	Girth at a height of 6' (in)
1	19	8	4	10	32	13	4
2	51	15	5	11	25	5	2
3	30	11	3	12	10	6	3
4	42	21	3	13	20	4	4
5	25	7	2	14	27	8	4
6	18	5	1	15	13	7	3
7	44	10	4	16	49	12	5
8	56	13	6	17	27	6	3
9	38	12	3	18	55	16	7

This gives

$$\bar{x}_1 = 581/18 = 32.28 \text{ oz},$$

$$\bar{x}_2 = 179/18 = 9.94 \text{ in},$$

$$\bar{x}_3 = 66/18 = 3.67 \text{ in},$$

$$s_1 = \frac{\sqrt{n \sum_{\alpha} x_{1\alpha}^2 - (\sum_{\alpha} x_{1\alpha})^2}}{n} = \frac{\sqrt{63,713}}{18} = \frac{252.41}{18} = 14.02 \text{ oz}.$$

$$s_2 = \frac{\sqrt{6,353}}{18} = \frac{79.706}{18} = 4.43 \text{ in},$$

$$s_3 = \frac{\sqrt{648}}{18} = \frac{25.456}{18} = 1.41 \text{ in},$$

$$r_{12} = \frac{n \sum_{\alpha} x_{1\alpha} x_{2\alpha} - (\sum_{\alpha} x_{1\alpha})(\sum_{\alpha} x_{2\alpha})}{\sqrt{n \sum_{\alpha} x_{1\alpha}^2 - (\sum_{\alpha} x_{1\alpha})^2} \sqrt{n \sum_{\alpha} x_{2\alpha}^2 - (\sum_{\alpha} x_{2\alpha})^2}}$$

$$= \frac{15,449}{\sqrt{63,713} \sqrt{6,353}} = \frac{15,449}{252.41 \times 79.706} = \frac{15,449}{20,118.59} = 0.768,$$

$$r_{18} = \frac{4,620}{\sqrt{63,713} \sqrt{6+8}} = \frac{4,620}{252.41 \times 25.456} = \frac{4,620}{6,425.35} = 0.719$$

and

$$r_{23} = \frac{1,056}{\sqrt{6.353} \sqrt{648}} = \frac{1,056}{79.706 \times 25.456} = \frac{1,056}{2,029.00} = 0.520.$$

TABLE 12.2

CALCULATION OF SUMS, SUMS OF SQUARES AND SUMS OF PRODUCTS FOR THE DATA OF TABLE 12.1

If the multiple regression equation is denoted by

$$\lambda_{1 \cdot 23} = a + b_{12 \cdot 3}x_2 + b_{13 \cdot 2}x_3,$$

then $b_{1 \cdot 23} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \frac{s_1}{s_2} = \frac{0.394}{0.730} \times \frac{14.02}{4.43} = \frac{5.52}{3.23} = 1.71$

and $b_{13 \cdot 2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \frac{s_1}{s_3} = \frac{0.320}{0.730} \times \frac{14.02}{1.41} = \frac{4.49}{1.03} = 4.36,$

while $a = \bar{x}_1 - b_{12 \cdot 3}x_2 - b_{13 \cdot 2}x_3 = -0.72$ (all in proper units)

Hence the multiple regression equation is obtained as

$$\lambda_{1 \cdot 23} = -0.72 + 1.71x_2 + 4.36x_3$$

12.3 Multiple correlation

In studying the dependence of x_1 on a set of independent variables, we may want to know to what extent x_1 is influenced by the independent variables. In the case of two variables x and y , we have seen that $r_{xy} = |r|$ serves as a measure of the strength of the interdependence of x and y or, if y may be looked upon as dependent on x , of the extent to which x influences y . Generalising this approach, we may take the simple correlation between x_1 and $\lambda_{1 \cdot 23 \cdot p}$, i.e. the value of x_1 given by the multiple regression equation of x_1 on x_2, x_3, \dots, x_p , as a measure of the joint influence of x_2, x_3, \dots, x_p on x_1 . It is called the *multiple correlation coefficient* of x_1 on x_2, x_3, \dots, x_p , and is denoted by $r_{1 \cdot 23 \cdot p}$. Obviously,

$$r_{1 \cdot 23 \cdot p} = \frac{\text{cov}(x_1, \lambda_{1 \cdot 23 \cdot p})}{\sqrt{\text{var}(x_1)} \sqrt{\text{var}(\lambda_{1 \cdot 23 \cdot p})}} \quad (12.16)$$

According to our notation

$$\text{var}(x_1) = s_1^2$$

Again, the mean of the predicted value $\lambda_{1 \cdot 23 \cdot p}$ is, from (12.14),

$$\begin{aligned} \frac{1}{n} \sum_{a=1}^n \lambda_{1 \cdot 23 \cdot p} &= x_1 - \frac{R_{12}}{nR_{11}} \frac{s_1}{s_2} \sum_{a=1}^n (x_{2a} - \bar{x}_2) - \frac{R_{13}}{nR_{11}} \frac{s_1}{s_3} \sum_{a=1}^n (x_{3a} - \bar{x}_3) \\ &\quad - \cdots - \frac{R_{1p}}{nR_{11}} \frac{s_1}{s_p} \sum_{a=1}^n (x_{pa} - \bar{x}_p) \\ &= x_1 \end{aligned} \quad (12.17)$$

so, since

$$x_1 = X_{1 \cdot 23 \dots p} + x_{1 \cdot 23 \dots p},$$

mean of $x_{1 \cdot 23 \dots p}$ is zero and

$$\begin{aligned}\text{cov}(x_1, X_{1 \cdot 23 \dots p}) &= \text{var}(X_{1 \cdot 23 \dots p}) + \text{cov}(X_{1 \cdot 23 \dots p}, x_{1 \cdot 23 \dots p}) \\ &= \text{var}(X_{1 \cdot 23 \dots p}),\end{aligned}\dots \quad (12.18)$$

so

$$\begin{aligned}n \text{cov}(X_{1 \cdot 23 \dots p}) &= \sum_{\alpha} X_{1 \cdot 23 \dots p, \alpha} x_{1 \cdot 23 \dots p, \alpha} \\ &= 0,\end{aligned}\dots \quad (12.19)$$

the normal equations (12.3a), $X_{1 \cdot 23 \dots p}$ being a linear function of \dots, x_p .

it

$$\begin{aligned}\text{av}(x_1, X_{1 \cdot 23 \dots p}) &= \frac{1}{n} \sum_{\alpha} (x_{1\alpha} - \bar{x}_1)(X_{1 \cdot 23 \dots p, \alpha} - \bar{x}_1) \\ &= \frac{1}{n} \sum_{\alpha} (x_{1\alpha} - \bar{x}_1) \left\{ -\frac{R_{12}}{R_{11}} \cdot \frac{s_1}{s_2} (x_{2\alpha} - \bar{x}_2) - \frac{R_{13}}{R_{11}} \cdot \frac{s_1}{s_3} (x_{3\alpha} - \bar{x}_3) \right. \\ &\quad \left. - \dots - \frac{R_{1p}}{R_{11}} \cdot \frac{s_1}{s_p} (x_{p\alpha} - \bar{x}_p) \right\} \\ &= -\frac{R_{12}}{R_{11}} \cdot \frac{s_1}{s_2} \cdot s_{12} - \frac{R_{13}}{R_{11}} \cdot \frac{s_1}{s_3} \cdot s_{13} - \dots - \frac{R_{1p}}{R_{11}} \cdot \frac{s_1}{s_p} \cdot s_{1p} \\ &= -\frac{s_1^2}{R_{11}} (r_{12} R_{12} + r_{13} R_{13} + \dots + r_{1p} R_{1p}) \\ &= -\frac{s_1^2}{R_{11}} (R - r_{11} R_{11}) = \left(1 - \frac{R}{R_{11}}\right) s_1^2.\end{aligned}\dots \quad (12.20)$$

ence

$$r_{1 \cdot 23 \dots p} = \frac{\left(1 - \frac{R}{R_{11}}\right) s_1^2}{\sqrt{s_1^2 \left(1 - \frac{R}{R_{11}}\right) s_1^2}} = \left(1 - \frac{R}{R_{11}}\right)^{1/2}.\dots \quad (12.20a)$$

The multiple correlation coefficient, being essentially a simple correlation coefficient, must lie between -1 and +1. But the variance between x_1 and $X_{1 \cdot 23 \dots p}$, being at the same time the variance of $X_{1 \cdot 23 \dots p}$, has to be a non-negative quantity. Hence we always

$$0 \leq r_{1 \cdot 23 \dots p} \leq 1.\dots \quad (12.21)$$

12.4 Some results relating to multiple regression and multiple correlation

Two results that we have already obtained in Section 12.3 are :

$$(1) \quad \bar{X}_{1 \cdot 23 \dots p} = \bar{x}_1,$$

implying that

$$\bar{x}_{1 \cdot 23 \dots p} = 0; \quad \dots \quad (12.17a)$$

and

$$(2) \quad \text{var}(X_{1 \cdot 23 \dots p}) = \left(1 - \frac{R}{R_{11}}\right) s_1^2 \quad \dots \quad (12.22)$$

$$= r_{1 \cdot 23 \dots p}^2 s_1^2. \quad \dots \quad (12.22a)$$

(3) We also find from the normal equations that, for $i=2, 3, \dots, p$,

$$\begin{aligned} \text{cov}(x_i, x_{1 \cdot 23 \dots p}) &= \frac{1}{n} \sum_a x_{ia} x_{1 \cdot 23 \dots p, a} \\ &= 0. \end{aligned} \quad \dots \quad (12.23)$$

Hence the residual part $x_{1 \cdot 23 \dots p}$ is uncorrelated with each of the independent variables (and hence with the regression part $X_{1 \cdot 23 \dots p}$).

$$(4) \quad \text{Since } x_1 = X_{1 \cdot 23 \dots p} + x_{1 \cdot 23 \dots p}$$

and

$$\text{cov}(X_{1 \cdot 23 \dots p}, x_{1 \cdot 23 \dots p}) = 0,$$

as we have seen in (12.19) or as may be seen from (12.17a) and (12.23), we get

$$\text{var}(x_1) = \text{var}(X_{1 \cdot 23 \dots p}) + \text{var}(x_{1 \cdot 23 \dots p}) \quad \dots \quad (12.24)$$

Hence

$$\text{var}(x_{1 \cdot 23 \dots p}) = \frac{R}{R_{11}} s_1^2 \quad (12.25)$$

or, denoting $\text{var}(x_{1 \cdot 23 \dots p})$ by $s_{1 \cdot 23 \dots p}^2$, $s_{1 \cdot 23 \dots p}$ being the *standard error of estimate*,

$$s_{1 \cdot 23 \dots p}^2 = (1 - r_{1 \cdot 23 \dots p}^2) s_1^2. \quad \dots \quad (12.25a)$$

Because of (12.22a) and (12.25a), we may write

$$r_{1 \cdot 23 \dots p}^2 = \frac{\text{var}(X_{1 \cdot 23 \dots p})}{\text{var}(x_1)} \quad (12.26)$$

$$= 1 - \frac{\text{var}(x_{1 \cdot 23 \dots p})}{\text{var}(x_1)}. \quad (12.27)$$

Now, $X_{1 \cdot 23 \dots p}$ and $x_{1 \cdot 23 \dots p}$ may be looked upon as, respectively, the part of x_1 that is accounted for and the part that is left unaccounted for by its multiple regression equation on x_2, x_3, \dots, x_p . Hence $r_{1 \cdot 23 \dots p}^2$ may be interpreted as the proportion of the total variance of x_1 that is explained by the multiple regression equation. Correspondingly, $1 - r_{1 \cdot 23 \dots p}^2$ is the proportion of the total variance of x_1 that is left unexplained by the multiple regression equation.

Also, equation (12.25a) shows that $s_{1 \cdot 23 \dots p}^2$ becomes smaller and smaller as $r_{1 \cdot 23 \dots p}$ increases from zero to unity. When $r_{1 \cdot 23 \dots p} = 1$, $s_{1 \cdot 23 \dots p}^2 = 0$, implying that $x_{1\alpha} = X_{1 \cdot 23 \dots p, \alpha}$, for each α , and here the multiple regression equation may be viewed as a perfect predicting formula. At the other extreme, if $r_{1 \cdot 23 \dots p} = 0$, then $\text{var}(X_{1 \cdot 23 \dots p}) = 0$, implying that $X_{1 \cdot 23 \dots p, \alpha} = \bar{x}_1$, i.e. is independent of x_2, x_3, \dots, x_p . Hence here the equation fails completely as a predicting formula for x_1 . That is why the multiple correlation coefficient, $r_{1 \cdot 23 \dots p}$, may also be regarded as a measure of the efficacy of the multiple regression equation as a formula for predicting x_1 when x_2, x_3, \dots, x_p are given.

Ex. 12.2 For the data of Ex. 12.1, the multiple correlation coefficient of weight of dry bark (x_1) on height (x_2) and girth at a height of 6" (x_3) may be computed. We have

$$r_{12} = 0.768, r_{13} = 0.719 \text{ and } r_{23} = 0.520.$$

Hence the required multiple correlation coefficient is

$$\begin{aligned} r_{1 \cdot 23} &= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \\ &= \sqrt{\frac{0.532505}{0.729600}} = \sqrt{0.729859} = 0.854. \end{aligned}$$

It indicates that x_2 and x_3 have considerable influence on x_1 . Viewed in a different way, it indicates that the multiple regression equation obtained in Ex. 12.1 serves as an excellent formula for predicting x_1 from given values of x_2 and x_3 .

12.5 Partial correlation

Sometimes the correlation between two variables, say x_1 and x_2 , may be partly (or wholly) due to the correlation of a group of variables, say x_3, x_4, \dots, x_p , with both x_1 and x_2 . In such situations, one may want to know what the correlation between x_1

and x_2 would be if the effect of x_3, x_4, \dots, x_p on each of them were eliminated. This correlation is called the *partial correlation* or *net correlation* between x_1 and x_2 , eliminating the effect of x_3, x_4, \dots, x_p , as opposed to their simple (or *total*) correlation.

Consider the least-square linear regression equations of x_1 on x_3, x_4, \dots, x_p and of x_2 on x_3, x_4, \dots, x_p . We may write

$$x_1 = X_{1 \cdot 34 \cdot p} + x_{1 \cdot 34 \cdot p}$$

and

$$x_2 = X_{2 \cdot 34 \cdot p} + x_{2 \cdot 34 \cdot p},$$

where $X_{1 \cdot 34 \cdot p}$ and $X_{2 \cdot 34 \cdot p}$ are the predicted values of x_1 and x_2 , $x_{1 \cdot 34 \cdot p}$ and $x_{2 \cdot 34 \cdot p}$ being the errors of estimation. Since $x_{1 \cdot 34 \cdot p}$ and $x_{2 \cdot 34 \cdot p}$ are uncorrelated with x_3, x_4, \dots, x_p [vide equation (12.23)], these may be looked upon as the parts of x_1 and x_2 , respectively, which are unaffected by this group of variables. Hence the simple correlation coefficient between $x_{1 \cdot 34 \cdot p}$ and $x_{2 \cdot 34 \cdot p}$ may be used to measure the partial correlation of x_1 and x_2 , eliminating the effect of x_3, x_4, \dots, x_p , in so far as this can be done with the help of linear regression equations. This is known as a *partial correlation coefficient* and is denoted by $r_{12 \cdot 34 \cdot p}$.

Thus

$$r_{12 \cdot 34 \cdot p} = \frac{\text{cov}(x_{1 \cdot 34 \cdot p}, x_{2 \cdot 34 \cdot p})}{\sqrt{\text{var}(x_{1 \cdot 34 \cdot p}) \text{var}(x_{2 \cdot 34 \cdot p})}}. \quad \dots \quad (12.28)$$

According to our notation,

$$\begin{aligned} x_{1 \cdot 34 \cdot p} &= x_1 - X_{1 \cdot 34 \cdot p} = (x_1 - \bar{x}_1) + \frac{R_{13}^{(2)}}{R_{11}^{(2)}} \frac{s_1}{s_3} (x_3 - \bar{x}_3) \\ &\quad + \frac{R_{14}^{(2)}}{R_{11}^{(2)}} \frac{s_1}{s_4} (x_4 - \bar{x}_4) + \dots + \frac{R_{1p}^{(2)}}{R_{11}^{(2)}} \frac{s_1}{s_p} (x_p - \bar{x}_p), \end{aligned}$$

where $R_{ij}^{(2)}$ is the co-factor of r_{ij} in $R^{(2)}$, the determinant obtained from R by deleting the 2nd row and the 2nd column. Or, putting

$$u_i = x_i - \bar{x}_i \quad \dots \quad (12.29)$$

and

$$u_{1 \cdot 34 \cdot p} = x_{1 \cdot 34 \cdot p} - \bar{x}_{1 \cdot 34 \cdot p} = x_{1 \cdot 34 \cdot p}, \quad \dots \quad (12.30)$$

$$u_{1 \cdot 34 \cdot p} = u_1 + \frac{R_{13}^{(2)}}{R_{11}^{(2)}} \frac{s_1}{s_3} u_3 + \frac{R_{14}^{(2)}}{R_{11}^{(2)}} \frac{s_1}{s_4} u_4 + \dots + \frac{R_{1p}^{(2)}}{R_{11}^{(2)}} \frac{s_1}{s_p} u_p.$$

Similarly, putting

$$u_{2 \cdot 34 \cdot p} = x_{2 \cdot 34 \cdot p} - \bar{x}_{2 \cdot 34 \cdot p} = x_{2 \cdot 34 \cdot p}, \quad \dots \quad (12.30a)$$

we have :

$$u_{2 \cdot 34 \dots p} = u_2 + \frac{R_{23}^{(1)}}{R_{22}^{(1)} s_3} \cdot \frac{s_2}{s_3} u_3 + \frac{R_{24}^{(1)}}{R_{22}^{(1)} s_4} \cdot \frac{s_2}{s_4} u_4 + \dots + \frac{R_{2p}^{(1)}}{R_{22}^{(1)} s_p} \cdot \frac{s_2}{s_p} u_p.$$

Analogously to (12.25), we have

$$\text{var}(x_{1 \cdot 34 \dots p}) = \frac{R_{12}^{(2)}}{R_{11}^{(2)}} s_1^2 \quad \dots \quad (12.31)$$

and

$$\text{var}(x_{2 \cdot 34 \dots p}) = \frac{R_{22}^{(1)}}{R_{22}^{(1)}} s_2^2. \quad \dots \quad (12.32)$$

Now, owing to the normal equations determining $X_{1 \cdot 34 \dots p}$ and $X_{2 \cdot 34 \dots p}$, we have

$$\sum_{\alpha} u_{1 \cdot 34 \dots p, \alpha} = 0, \quad \sum_{\alpha} u_{2 \cdot 34 \dots p, \alpha} = 0 \quad \dots \quad (12.33)$$

$$\text{and} \quad \sum_{\alpha} u_{i\alpha} u_{1 \cdot 34 \dots p, \alpha} = 0, \quad \sum_{\alpha} u_{i\alpha} u_{2 \cdot 34 \dots p, \alpha} = 0, \quad \dots \quad (12.34)$$

$$\text{for } i=3, 4, \dots, p.$$

Hence

$$\begin{aligned} n \text{cov}(x_{1 \cdot 34 \dots p}, x_{2 \cdot 34 \dots p}) &= \sum_{\alpha} u_{1 \cdot 34 \dots p, \alpha} u_{2 \cdot 34 \dots p, \alpha} = \sum_{\alpha} u_{1\alpha} u_{2\alpha} \\ &= \sum_{\alpha} u_{1\alpha} u_{2\alpha} + \frac{R_{23}^{(1)}}{R_{22}^{(1)} s_3} \sum_{\alpha} u_{1\alpha} u_{3\alpha} + \frac{R_{24}^{(1)}}{R_{22}^{(1)} s_4} \sum_{\alpha} u_{1\alpha} u_{4\alpha} + \dots + \frac{R_{2p}^{(1)}}{R_{22}^{(1)} s_p} \sum_{\alpha} u_{1\alpha} u_{p\alpha}, \\ \text{or} \quad \text{cov}(x_{1 \cdot 34 \dots p}, x_{2 \cdot 34 \dots p}) &= s_1 s_2 \left(r_{12} + r_{13} \frac{R_{23}^{(1)}}{R_{22}^{(1)}} + r_{14} \frac{R_{24}^{(1)}}{R_{22}^{(1)}} + \dots + r_{1p} \frac{R_{2p}^{(1)}}{R_{22}^{(1)}} \right) \\ &\quad (\text{since } \sum_{\alpha} u_{i\alpha} u_{j\alpha} = n \text{cov}(x_i, x_j) = nr_{ij} s_i s_j) \\ &= -\frac{R_{12}}{R_{22}^{(1)}} s_1 s_2. \end{aligned} \quad \dots \quad (12.35)$$

This is so because

$$r_{12} R_{22}^{(1)} + r_{13} R_{23}^{(1)} + r_{14} R_{24}^{(1)} + \dots + r_{1p} R_{2p}^{(1)}$$

= determinant obtainable from $R^{(1)}$ by replacing its first row, $(r_{21}, r_{22}, \dots, r_{2p})$, with $(r_{12}, r_{13}, \dots, r_{1p})$

$$\begin{aligned} &= \begin{vmatrix} r_{12} & r_{13} & \dots & r_{1p} \\ r_{22} & r_{23} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p2} & r_{p3} & \dots & r_{pp} \end{vmatrix} = \text{minor of } r_{21} \text{ in } R \\ &= \text{minor of } r_{12} \text{ in } R \\ &= -R_{12}. \end{aligned}$$

Thus, in terms of the simple (or total) correlation coefficients r_{ij} ,

$$\begin{aligned} r_{12 \cdot 34 \cdot p} &= \frac{(-R_{12}/R_{\frac{1}{2}\frac{3}{4}}^{(1)})s_1 s_2}{(R^{(1)} / R_{11}^{(1)})^{1/2} (R^{(1)} / R_{22}^{(1)})^{1/2} s_1 s_2} \\ &= -\frac{R_{12}}{\sqrt{R_{11} R_{22}}}, \end{aligned} \quad (12.36)$$

since $R^{(1)} = R_{11}$,

$$R^{(2)} = R_{22}$$

and $R_{11}^{(2)} = R_{\frac{1}{2}\frac{3}{4}}^{(1)}$,

both being obtainable from R by deleting its first two rows and first two columns.

In particular, with

$$R = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix},$$

we have

$$-R_{12} = \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & r_{33} \end{vmatrix} = r_{12} - r_{13} r_{23},$$

$$R_{11} = \begin{vmatrix} r_{22} & r_{23} \\ r_{32} & r_{33} \end{vmatrix} = 1 - r_{23}^2,$$

$$\text{and } R_{22} = \begin{vmatrix} r_{11} & r_{13} \\ r_{31} & r_{33} \end{vmatrix} = 1 - r_{13}^2,$$

and so

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}, \quad (12.37)$$

a result which could be obtained directly from the regression equation of x_1 on x_3 and that of x_2 on x_3 .

Unlike the multiple correlation coefficient,

$$-1 \leq r_{12 \cdot 34 \cdot p} \leq 1 \quad (12.38)$$

12.6 Some relations connecting partial regression and partial correlation coefficients

We have seen in (12.12) that

$$b_{12 \cdot 34 \cdot p} = -\frac{R_{12}}{R_{11}} \frac{s_1}{s_2}$$

Also,

$$s_{1 \cdot 23 \dots p}^2 = \frac{R}{R_{11}} s_1^2,$$

$$s_{2 \cdot 13 \dots p}^2 = \frac{R}{R_{22}} s_2^2$$

$$r_{12 \cdot 34 \dots p} = -\frac{R_{12}}{\sqrt{R_{11} R_{22}}}.$$

and

Hence

$$b_{12 \cdot 34 \dots p} = r_{12 \cdot 34 \dots p} \frac{s_{1 \cdot 23 \dots p}}{s_{2 \cdot 13 \dots p}}. \quad \dots \quad (12.39)$$

Again, since

$$s_{1 \cdot 34 \dots p}^2 = \frac{R^{(2)}}{R_{11}^{(2)}} s_1^2,$$

$$s_{2 \cdot 34 \dots p}^2 = \frac{R^{(1)}}{R_{22}^{(1)}} s_2^2,$$

and $R^{(1)} = R_{11}$, $R^{(2)} = R_{22}$ and $R_{11}^{(2)} = R_{22}^{(1)}$,

$$\begin{aligned} b_{12 \cdot 34 \dots p} &= -\frac{R_{12}}{\sqrt{R_{11} R_{22}}} \cdot \frac{\sqrt{R_{22} s_1^2}}{\sqrt{R_{11} s_2^2}} \\ &= r_{12 \cdot 34 \dots p} \frac{\sqrt{(R^{(2)}/R_{11}^{(2)}) s_1^2}}{\sqrt{(R^{(1)}/R_{22}^{(1)}) s_2^2}} \\ &= r_{12 \cdot 34 \dots p} \frac{s_{1 \cdot 34 \dots p}}{s_{2 \cdot 34 \dots p}} \quad \dots \quad (12.40) \end{aligned}$$

—a relation of the same form as

$$b_{12} = r_{12} \frac{s_1}{s_2},$$

with just the secondary suffixes 3, 4, ..., p added to each term.
We have also, analogously to the relation : $b_{12} = \text{cov}(x_1, x_2)/\text{var}(x_2)$,

$$b_{12 \cdot 34 \dots p} = \frac{\text{cov}(x_{1 \cdot 34 \dots p}, x_{2 \cdot 34 \dots p})}{\text{var}(x_{2 \cdot 34 \dots p})}. \quad \dots \quad (12.40a)$$

Interchanging the suffixes 1 and 2 in (12.40), we have an analogous expression for $b_{21 \cdot 34 \dots p}$, and the two lead to

$$r_{1 \cdot 34 \dots p}^2 = b_{12 \cdot 34 \dots p} b_{21 \cdot 34 \dots p}, \quad \dots \quad (12.41)$$

just as

$$r_{12}^2 = b_{12} b_{21}.$$

12.7 Expression of a multiple correlation coefficient in terms of total and partial correlation coefficients

We have, using the notation introduced in the last section,

$$\begin{aligned} \sum_{\alpha} u_{123 \dots p, \alpha}^2 &= \sum_{\alpha} u_{123 \dots (p-1), \alpha} u_{123 \dots p, \alpha}^* \\ &= \sum_{\alpha} u_{123 \dots (p-1), \alpha} [u_{1\alpha} - b_{1234 \dots p} u_{2\alpha} - b_{1324 \dots p} u_{3\alpha} - \\ &\quad - b_{1p23 \dots (p-1)} u_{p\alpha}] \\ &= \sum_{\alpha} u_{123 \dots (p-1), \alpha} u_{1\alpha} - b_{1p23 \dots (p-1)} \sum_{\alpha} u_{123 \dots (p-1), \alpha} u_{p\alpha} \\ &= \sum_{\alpha} u_{123 \dots (p-1), \alpha}^2 - b_{1p23 \dots (p-1)} \sum_{\alpha} u_{123 \dots (p-1), \alpha} u_{p23 \dots (p-1), \alpha} \end{aligned}$$

or, on dividing both sides by n and applying (12.41),

$$\begin{aligned} s_{123 \dots p}^2 &= [1 - b_{1p23 \dots (p-1)}] s_{123 \dots (p-1)}^2 \\ &= (1 - r_{1p23 \dots (p-1)}^2) s_{123 \dots (p-1)}^2 \end{aligned} \quad (12.42)$$

This shows incidentally that

$$s_{123 \dots p}^2 \leq s_{123 \dots (p-1)}^2 \quad (12.42a)$$

or, equivalently, that

$$r_{123 \dots p} \geq r_{123 \dots (p-1)} \quad (12.42b)$$

So by introducing an additional independent variable in the multiple regression equation, one can only improve its usefulness as a predicting formula.

Also, applying equation (12.42) successively to $s_{123 \dots (p-1)}^2$, $s_{123 \dots (p-2)}^2$, ..., s_{12}^2 , we have

$$s_{123 \dots p}^2 = (1 - r_{12}^2)(1 - r_{13 \dots 2}^2) \dots (1 - r_{1p23 \dots (p-1)}^2) s_1^2 \quad (12.43)$$

$$\text{or } 1 - r_{123 \dots p}^2 = (1 - r_{12}^2)(1 - r_{13 \dots 2}^2) \dots (1 - r_{1p23 \dots (p-1)}^2) \quad (12.43a)$$

Each of the factors on the right-hand side being less than or equal to unity, the multiple correlation coefficient, $r_{123 \dots p}$, must be numerically at least as high as any of the total or partial correlation coefficients of x_1 with the independent variables.

*This is because of the normal equations determining the residuals. In general, when one takes the sum of such products the secondary suffixes of one of the factors being common to those of the other one can drop any or all of the secondary suffixes of the former. Likewise one can add to the secondary suffixes of the former any or all of the secondary suffixes of the latter.

Suppose, without any loss of generality, that $p=3$. Since $s_1^2(1-r_{12}^2)$ and $s_1^2(1-r_{13}^2)$ are the residual variances if x_1 is estimated from x_2 and x_3 individually, while $s_1^2(1-r_{1\cdot23}^2)$ is the residual variance if x_1 is estimated from x_2 and x_3 taken together, it is obvious that the inclusion of an additional variable can only reduce the residual variance. Now, inclusion of x_3 , when x_2 has already been taken, for predicting x_1 , is worth while only when the resultant reduction in the residual variance is substantial. As (12.43a) indicates, this will be the case when the numerical value of $r_{13\cdot2}$ is sufficiently large. This illustrates the importance of the partial correlation coefficient in deciding whether to include or not an additional independent variable in regression analysis.

Ex. 12.3 Let us consider the data of Ex. 12.1 again. The partial correlation coefficient of x_1 (yield of dry bark) and x_2 (height of plant), the effect of x_3 (girth at a height of 6") being accounted for, is

$$\begin{aligned} r_{1\cdot23} &= \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \\ &= \frac{0.394}{\sqrt{0.483039}\sqrt{0.729600}} = \frac{0.394}{0.695 \times 0.854} = 0.663. \end{aligned}$$

The partial correlation coefficient of x_1 and x_3 , eliminating the effect of x_2 , is

$$\begin{aligned} r_{13\cdot2} &= \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} \\ &= \frac{0.320}{\sqrt{0.410176}\sqrt{0.729600}} = \frac{0.320}{0.640 \times 0.854} = 0.585. \end{aligned}$$

These values may be considered together with the total correlations

$$r_{12} = 0.768 \text{ and } r_{13} = 0.719.$$

Since r_{12} is quite large, one will naturally take x_2 as an independent variable for predicting x_1 . The partial correlation $r_{13\cdot2}$, being equal to 0.585, indicates that the inclusion of x_3 as an independent variable, in addition to x_2 , would be worth while as it would considerably increase the accuracy of prediction.

Ex 12.4 Ramakrishnan [*Sankhya*, 2 (1935-36), pp 43-54] considered annual data on the yield-rate of cotton, September rainfall, November rainfall and November maximum temperature for an Indian district with a view to building up a forecasting formula for the first variable. To make the data free from time dependent components the ratio of each figure to the 5 year moving average value was taken. The new variables will be called x_1 , x_2 , x_3 and x_4 respectively.

The total correlations were found to be

$$\begin{aligned}r_{12} &= 0.410 & r_{23} &= -0.287 \\r_{13} &= 0.307 & r_{24} &= -0.239 \\r_{14} &= -0.619 & r_{34} &= -0.517\end{aligned}$$

Since, of the total correlations r_{12} , r_{13} and r_{14} , the last one is numerically quite high and far higher than the other two, one should, of course take x_4 as an independent variable in the regression equation of x_1 .

Also

$$\begin{aligned}r_{12 \cdot 4} &= \frac{r_{12} - r_{14} r_{24}}{\sqrt{1 - r_{14}^2} \sqrt{1 - r_{24}^2}} \\&= 0.3437,\end{aligned}$$

$$\begin{aligned}r_{13 \cdot 4} &= \frac{r_{13} - r_{14} r_{34}}{\sqrt{1 - r_{14}^2} \sqrt{1 - r_{34}^2}} \\&= -0.01934,\end{aligned}$$

and

$$\begin{aligned}r_{23 \cdot 4} &= \frac{r_{23} - r_{24} r_{34}}{\sqrt{1 - r_{24}^2} \sqrt{1 - r_{34}^2}} \\&= -0.1966\end{aligned}$$

The partial correlation $r_{12 \cdot 4}$ is thus fairly high and much higher numerically than $r_{13 \cdot 4}$. Hence it would be advisable to include x_4 as an independent variable in addition to x_1 .

As to the other variable, x_3 , its inclusion will not be worth while since

$$\begin{aligned}r_{13 \cdot 24} &= \frac{r_{13 \cdot 4} - r_{12 \cdot 4} r_{23 \cdot 4}}{\sqrt{1 - r_{12 \cdot 4}^2} \sqrt{1 - r_{23 \cdot 4}^2}} \\&= 0.0524,\end{aligned}$$

which is a negligibly small quantity.

12.8 Expression of a higher-order coefficient in terms of coefficients of a lower order

We have

$$\begin{aligned}
 \sum_{\alpha} u_{1 \cdot 34 \dots p, \alpha} u_{2 \cdot 34 \dots p, \alpha} &= \sum_{\alpha} u_{1 \cdot 34 \dots (p-1), \alpha} u_{2 \cdot 34 \dots p, \alpha} \\
 &= \sum_{\alpha} u_{1 \cdot 34 \dots (p-1), \alpha} [u_{2 \alpha} - b_{23 \cdot 45 \dots p} u_{3 \alpha} - b_{24 \cdot 35 \dots p} u_{4 \alpha} \\
 &\quad - \dots - b_{2p \cdot 34 \dots (p-1)} u_{p \alpha}] \\
 &= \sum_{\alpha} u_{1 \cdot 34 \dots (p-1), \alpha} u_{2 \alpha} - b_{2p \cdot 34 \dots (p-1)} \sum_{\alpha} u_{1 \cdot 34 \dots (p-1), \alpha} u_{p \alpha} \\
 &= \sum_{\alpha} u_{1 \cdot 34 \dots (p-1), \alpha} u_{2 \cdot 34 \dots (p-1), \alpha} \\
 &\quad - b_{2p \cdot 34 \dots (p-1)} \sum_{\alpha} u_{1 \cdot 34 \dots (p-1), \alpha} u_{p \cdot 34 \dots (p-1), \alpha}.
 \end{aligned}$$

Dividing both sides by n , we get

$$\begin{aligned}
 \text{cov}(x_{1 \cdot 34 \dots p}, x_{2 \cdot 34 \dots p}) &= \text{cov}(x_{1 \cdot 34 \dots (p-1)}, x_{2 \cdot 34 \dots (p-1)}) \\
 &\quad - b_{2p \cdot 34 \dots (p-1)} \text{cov}(x_{1 \cdot 34 \dots (p-1)}, x_{p \cdot 34 \dots (p-1)}),
 \end{aligned}$$

or, using equation (12.40a),

$$b_{12 \cdot 34 \dots p} s_{2 \cdot 34 \dots p}^2 = [b_{12 \cdot 34 \dots (p-1)} - b_{1p \cdot 34 \dots (p-1)} b_{p2 \cdot 34 \dots (p-1)}] s_{2 \cdot 34 \dots (p-1)}^2$$

since

$$b_{2p \cdot 34 \dots (p-1)} = b_{p2 \cdot 34 \dots (p-1)} s_{2 \cdot 34 \dots (p-1)}^2 / s_{p \cdot 34 \dots (p-1)}^2$$

Also, from (12.41) and (12.42),

$$\begin{aligned}
 s_{2 \cdot 34 \dots p}^2 &= (1 - r_{2p \cdot 34 \dots (p-1)}^2) s_{2 \cdot 34 \dots (p-1)}^2 \\
 &= (1 - b_{2p \cdot 34 \dots (p-1)} b_{p2 \cdot 34 \dots (p-1)}) s_{2 \cdot 34 \dots (p-1)}^2
 \end{aligned}$$

Hence

$$b_{12 \cdot 34 \dots p} = \frac{b_{12 \cdot 34 \dots (p-1)} - b_{1p \cdot 34 \dots (p-1)} b_{p2 \cdot 34 \dots (p-1)}}{1 - b_{2p \cdot 34 \dots (p-1)} b_{p2 \cdot 34 \dots (p-1)}}. \quad \dots \quad (12.44)$$

Applying equation (12.40) and simplifying, we have also

$$r_{12 \cdot 34 \dots p} = \frac{r_{12 \cdot 34 \dots (p-1)} - r_{1p \cdot 34 \dots (p-1)} r_{p2 \cdot 34 \dots (p-1)}}{(1 - r_{1p \cdot 34 \dots (p-1)}^2)^{1/2} (1 - r_{p2 \cdot 34 \dots (p-1)}^2)^{1/2}}. \quad \dots \quad (12.44a)$$

Equations (12.44) and (12.44a) enable us to compute a regression coefficient or correlation coefficient of order $(p-2)$ from those of order $(p-3)$.

12.9 Expression of a lower-order coefficient in terms of coefficients of a higher order

Because of the normal equations, we also have

$$\sum_a u_{123} \dots p-a u_{234} \dots (p-1)a = 0$$

$$\text{or } \sum_a [u_{1a} - b_{1234} \dots p u_{a34} - b_{1324} \dots p u_{3a} - \\ - b_{1p23} \dots (p-1)u_{pa}] u_{234} \dots (p-1)a = 0$$

$$\text{or } \sum_a u_{1a} u_{234} \dots (p-1)a - b_{1321} \dots p \sum_a u_{2a} u_{234} \dots (p-1)a - \\ - b_{1p23} \dots (p-1) \sum_a u_{pa} u_{234} \dots (p-1)a = 0$$

$$\text{or } \sum_a u_{1321} \dots (p-1)a u_{234} \dots (p-1)a - b_{1231} \dots p \sum_a u_{234}^2 \dots (p-1)a - \\ - b_{1p23} \dots (p-1) \sum_a u_{pa} u_{234} \dots (p-1)a = 0$$

Hence, on dividing by n and using (12.40a) and simplifying,

$$b_{1231} \dots (p-1) + b_{1234} \dots p + b_{1p23} \dots (p-1) b_{p-31} \dots (p-1) \quad (12.45)$$

Also, if we interchange 1 and p in the suffixes, then

$$b_{p231} \dots (p-1) = b_{p-13} \dots (p-1) - b_{p123} \dots (p-1) b_{1234} \dots (p-1)$$

Substituting this value for $b_{p231} \dots (p-1)$ in (12.45), we have ultimately an expression for a regression coefficient of order $(p-3)$ in terms of those of order $(p-2)$, viz

$$b_{1231} \dots (p-1) = \frac{b_{1234} \dots p + b_{1p23} \dots (p-1) b_{p231} \dots (p-1)}{1 - b_{1p23} \dots (p-1) b_{p123} \dots (p-1)} \quad . \quad (12.45a)$$

Correspondingly, we have from (12.45) the following expression for a correlation coefficient of order $(p-3)$ in terms of those of order $(p-2)$

$$r_{1231} \dots (p-1) = \frac{r_{134} \dots p + r_{1p23} \dots (p-1) r_{2p13} \dots (p-1)}{(1 - r_{1p23}^2 \dots (p-1))^{1/2} (1 - r_{2p13}^2 \dots (p-1))^{1/2}} \quad (12.46)$$

In particular, with $p=3$, we get

$$b_{12} = \frac{b_{123} + b_{132} b_{321}}{1 - b_{132} b_{321}} \quad (12.47)$$

and

$$r_{12} = \frac{r_{123} + r_{132} r_{321}}{(1 - r_{132}^2)^{1/2} (1 - r_{321}^2)^{1/2}} \quad (12.48)$$

The attention of the reader is drawn particularly to equations (12.37) and (12.48)

Suppose for a set of data

$$r_{12 \cdot 3} = 0.$$

We have then

$$r_{12} = r_{13} r_{23},$$

which will not be zero if x_3 has non-zero correlation with x_1 and x_2 . Thus, although x_1 and x_2 may be uncorrelated when the effect of x_3 is eliminated from each, they may appear to be correlated when this is not done (i.e. when the influence of x_3 on them is ignored). This shows that one should be cautious in taking a non-zero total correlation between two variables as indicative of causal relationship. For the apparent relationship may really be due to the influence of another variable (or a group of variables) on both of them.

In the same way, the absence of correlation between two variables as shown by their total correlation coefficient may only be apparent. For r_{12} may be zero but $r_{12 \cdot 3}$, for instance, may not be, since in that case

$$r_{12 \cdot 3} = -r_{13 \cdot 2} r_{23 \cdot 1},$$

which may not vanish unless at least one of $r_{13 \cdot 2}$ and $r_{23 \cdot 1}$ does.

Besides, (12.37) and (12.48) also show that a total correlation coefficient may have a sign opposite to that of the corresponding partial coefficient.

12.10 Multivariate normal distribution

As in the univariate or the bivariate case, here too we have to consider for some purposes theoretical distributions that may serve as simplified models of observed multivariate distributions.

For continuous variables x_1, x_2, \dots, x_p , this distribution will be represented by a probability-density function, say $f(x_1, x_2, \dots, x_p)$, such that

$$P[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \dots, a_p < x_p < b_p]$$

$$= \int_{a_p}^{b_p} \dots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \dots, x_p) dx_1 dx_2 \dots dx_p.$$

Supposing that the range of possible values of x_i is from α_i to β_i ($i = 1, 2, \dots, p$), the p.d.f. has to satisfy the conditions :

$$(1) \quad f(x_1, x_2, \dots, x_p) \geq 0 \text{ for all values of } x_1, x_2, \dots, x_p, \quad (12.49)$$

$$(2) \quad \int_{a_p}^{b_p} \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2, \dots, x_p) dx_1 dx_2 \dots dx_p = 1 \quad (12.50)$$

We need consider here just one theoretical distribution of the continuous type, viz. the multivariate (p -variate) normal distribution whose p.d.f. is

$$f(x_1, x_2, \dots, x_p) = \frac{1}{(2\pi)^{p/2} |\sigma_{ij}|^{1/2}} \exp\left[-\frac{1}{2} \sum_i \sum_j \sigma_{ij}^{-1} (x_i - \mu_i)(x_j - \mu_j)\right] \quad -\infty < x_1, x_2, \dots, x_p < \infty, \quad (12.51)$$

where

$$\mu_i = E(x_i) \quad (12.52)$$

$$\sigma_{ij} = \begin{cases} \text{var}(x_i) = \sigma_i^2 & \text{if } i=j \\ \text{cov}(x_i, x_j) = \rho_{ij} \sigma_i \sigma_j & \text{if } i \neq j, \end{cases} \quad (12.53)$$

$$|\sigma_{ij}| = \text{determinant of } (\sigma_{ij}), \text{ the variance-covariance matrix (supposed to be positive definite)} \quad (12.54)$$

and

$$(\sigma^{ij}) = (\sigma_{ij})^{-1} \quad (12.55)$$

The more important properties of this distribution are as follows

(1) The marginal distribution of any p ($1 \leq p \leq p-1$) of the variables is p -variate normal

(2) The conditional distribution of any p variables for fixed values of any p'' other variables ($p', p'' \geq 1, p + p'' \leq p$) is also p -variate normal

(3) Consider the conditional distribution of any one variable, say x_1 , for fixed values of x_2, x_3, \dots, x_p . The mean and variance of the distribution are

$$E(x_1 | x_2, x_3, \dots, x_p) = \mu_1 - \sum_{i=2}^p \frac{R_{1i}}{R_{11}} \frac{\sigma_i}{\sigma_1} (x_i - \mu_i) \quad (12.56)$$

$$\text{and} \quad \text{var}(x_1 | x_2, x_3, \dots, x_p) = \frac{R_{11}}{R_{11}} \sigma_1^2, \quad (12.57)$$

where $R = |\rho_{ij}|$ = determinant of the correlation matrix (ρ_{ij})

and R_{ij} = co factor of ρ_{ij} in R

From (12.56), we see that the true regression of x_1 on x_2, x_3, \dots, x_p is linear

Writing, in accordance with (12.20),

$$\rho_{1 \cdot 23 \dots p} = \left(1 - \frac{R}{R_{11}}\right)^{1/2}$$

we have

$$\text{var}(x_1 | x_2, x_3, \dots, x_p) = (1 - \rho_{1 \cdot 23 \dots p}^2) \sigma_1^2, \quad \dots \quad (12.58)$$

which shows the rôle of the multiple correlation coefficient $\rho_{1 \cdot 23 \dots p}$ in the context of a multivariate normal distribution. The higher the multiple correlation, the smaller is the conditional variance.

(4) Again, consider the joint distribution of any two of the variables, say x_1 and x_2 , for fixed values of the others. The correlation coefficient between x_1 and x_2 in this distribution is

$$-\frac{R_{12}}{\sqrt{R_{11}R_{22}}}.$$

From (12.36), this is seen to be the partial correlation coefficient $\rho_{12 \cdot 34 \dots p}$. Thus in a multivariate normal set-up, the partial correlation $\rho_{12 \cdot 34 \dots p}$ is nothing but the correlation between x_1 and x_2 in the conditional distribution of x_1 and x_2 for each set of fixed values of x_3, x_4, \dots, x_p .

Questions and exercises

12.1 Obtain the multiple regression equation of x_1 on x_2, x_3, \dots, x_p , in terms of the means, the standard deviations and the inter-correlations of the variables.

12.2 Define multiple correlation and partial correlation. Deduce the formulæ for multiple and partial correlation coefficients in terms of total correlation coefficients.

12.3 Prove the relation

$$1 - r_{1 \cdot 23 \dots p}^2 = (1 - r_{12}^2)(1 - r_{13 \cdot 2}^2) \dots (1 - r_{1 \cdot p \cdot 23 \dots (p-1)}^2).$$

Use this relation to show that the multiple correlation coefficient is numerically greater than any of the total or partial correlation coefficients of x_1 with the other variables.

12.4 Find what the value of $r_{1 \cdot 23 \dots p}$ will be if the independent variables are pair-wise uncorrelated.

12.5 Show that if $r_{1i} = 0$ for each $i = 2, 3, \dots, p$, then $r_{1 \cdot 23 \dots p} = 0$, and conversely. What is the significance of this result in regard to the multiple regression equation of x_1 on x_2, x_3, \dots, x_p ?

12.6(a) Show that r_{12} , r_{13} and r_{23} must satisfy the inequality
 $r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} \leq 1$

[Hint Use the inequality $r_{12}^2 \leq 1$]

(b) Suppose a computer has found, for a given set of values of x_1 , x_2 and x_3 ,

$$r_{12} = 0.91, r_{13} = 0.33 \text{ and } r_{23} = 0.81$$

Examine whether his computations may be said to be free from errors

Ans No

12.7 Suppose x_1 , x_2 and x_3 satisfy the relation $a_1x_1 + a_2x_2 + a_3x_3 = k$

(a) Determine the three total correlation coefficients in terms of the standard deviations and the constants a_1 , a_2 and a_3

$$\text{Ans } r_{12} = \frac{a_3^2 s_3^2 - a_1^2 s_1^2 - a_2^2 s_2^2}{2a_1 a_2 s_1 s_2}, \text{ etc}$$

(b) State what the partial correlation coefficients will be

Partial ans All are equal to -1 , if a_1 , a_2 , a_3 are of the same sign

12.8 Suppose all the total correlation coefficients of x_1 , x_2 , \dots , x_p are equal to r (1) What is the value of $r_{123\dots p}$? (2) What are the values of the partial correlation coefficients of successive orders?

Show that if r be negative, then $r \geq -\frac{1}{p-1}$

Partial ans $r_{123\dots p}^2 = \frac{(p-1)r^2}{1+(p-2)r}$, each partial correlation

coefficient of order k is $\frac{r}{1+kr}$

12.9 Show that the multiple correlation coefficient $r_{123\dots p}$ is the highest possible value of the simple correlation coefficient between x_1 and a linear function of x_2 , x_3 , \dots , x_p

[Hint Take the correlation coefficient of x_1 and $Y = a_1 + b_2 x_2 + b_3 x_3 + \dots + b_p x_p$. Show that the values of b_2 , b_3 , \dots , b_p maximising this coefficient are proportional to the partial regression coefficients of x_1 on x_2 , x_3 , \dots , x_p , respectively.]

12.10 If $r_{ii} = r$ ($i=2, 3, \dots, p$) and $r_{ij} = r$ ($i, j=2, 3, \dots, p$, $i \neq j$), then what is $r_{123\dots p}$?

$$\text{Ans } r_{123\dots p}^2 = \frac{(p-1)r^2}{1+(p-2)r}$$

12.11 Let $x = u_1 + u_2 + \dots + u_r + v_1 + v_2 + \dots + v_s$, and $y = u_1 + u_2 + \dots + u_r + w_1 + w_2 + \dots + w_t$, the u 's, v 's and w 's being random

variables with unit variances and zero covariances. Show that the correlation coefficient between x and y is $\frac{r}{\sqrt{(r+s)(r+t)}}$.

12.12 What is a multivariate normal distribution? State its important properties.

12.13 On the basis of observations made on 35 cotton plants, the total correlations of yield of cotton (x_1), number of bolls, i.e. seed-vessels, (x_2) and height (x_3) are found to be

$$r_{12}=0.863, r_{13}=0.648 \text{ and } r_{23}=0.709.$$

Determine the multiple correlation coefficient $r_{1 \cdot 23}$ and the partial correlation coefficients $r_{1 \cdot 2 \cdot 3}$ and $r_{1 \cdot 3 \cdot 2}$, and interpret your results.

Partial ans. $r_{1 \cdot 23}=0.865; r_{1 \cdot 2 \cdot 3}=0.751, r_{1 \cdot 3 \cdot 2}=0.101$.

(12.14) The following constants are obtained from measurements on length in mm. (x_1), volume in c.c. (x_2) and weight in gm. (x_3) of 300 eggs :

$$\begin{array}{lll} \bar{x}_1=55.95 & s_1=2.26 & r_{12}=0.578 \\ \bar{x}_2=51.48 & s_2=4.39 & r_{13}=0.581 \\ \bar{x}_3=56.03 & s_3=4.41 & r_{23}=0.974 \end{array}$$

(a) Obtain the linear regression equation of egg-weight on egg-length and egg-volume. Hence estimate the weight of an egg whose length is 58.0 mm. and volume is 52.5 c.c.

$$\text{Ans. } X_3=3.49+0.053x_1+0.963x_2; 57.12 \text{ gm.}$$

(b) Give a measure of the usefulness of the above regression equation as a predicting formula.

(c) Compute the partial correlation coefficient of weight and volume, eliminating the effect of length. *Ans.* 0.961.

12.15 For a large group of students of statistics, x_1 =score in Theory, x_2 =score in Methods and x_3 =score in Lab. Work are approximately normally distributed. Also,

$$\begin{array}{lll} \bar{x}_1=50.4 & s_1=6.9 & r_{12}=0.69 \\ \bar{x}_2=45.1 & s_2=6.4 & r_{13}=0.45 \\ \bar{x}_3=53.3 & s_3=6.8 & r_{23}=0.58 \end{array}$$

Estimate the percentage of students whose total score exceeds 150.

[*Hint* : It can be proved that if x_1, x_2, \dots, x_p are distributed in the (multivariate) normal form, then $a+b_1x_1+b_2x_2+\dots+b_px_p$ is also normally distributed.]

$$\text{Ans. } 47.2\%$$

SUGGESTED READING

- [1] Ezekiel, M and Fox, K A *Methods of Correlation and Regression Analysis* (Chs 10—15) John Wiley, 1959
- [2] Goulden, C H *Methods of Statistical Analysis* (Ch 8) John Wiley, 1952, and Asia Publishing House, 1959
- [3] Kenney, J F and Keeping, E S *Mathematics of Statistics*, Part II (Ch 11) Van Nostrand, 1951, and Affiliated East-West Press.
- [4] Snedecor, G W *Statistical Methods* (Ch 14) Iowa State College Press, 1956, and Allied Pacific, 1961
- [5] Yule, G U and Kendall, M G *An Introduction to the Theory of Statistics* (Ch 12) Charles Griffin, 1950

13

SOME OTHER TYPES OF CORRELATION

13.1 Rank correlation coefficient

In calculating the product-moment correlation coefficient, it is essential that the two characters be definitely measurable. But in many cases the characters may not be measurable or, even if measurable, may not be measured for limitations of cost or time or for lack of appropriate measuring instruments. Sometimes, although measurements may be available for the calculation of the product-moment correlation, a rough and ready substitute may still be called for to reduce the arithmetical work involved. It may be possible to use a rank correlation coefficient in all these situations.

Suppose that it is possible to arrange the individuals according to the degree to which they possess the character under enquiry, although the character may not be directly measurable. Thus, for example, a number of operators may be arranged in order of efficiency by their supervisor, although it may not be easy to offer some numerical measure of efficiency. Such an ordered arrangement will be called a *ranking* and the ordinal number indicating the position of a given individual in the ranking is called its *rank*. A ranking where two or more individuals are allotted the same rank is called a *tie*. To be specific, a rank r means that with respect to the character under enquiry $r-1$ individuals have the character in a higher degree than the individual getting the rank r .

13.2 Spearman's rank correlation coefficient

First, let us consider the case where there is no tie.

Suppose we have n individuals ranked according to two characters, A and B , in the orders u_1, u_2, \dots, u_n and v_1, v_2, \dots, v_n , respectively, where the u 's and the v 's are permutations of the integers from 1 to n . Our problem is to have a suitable measure of the degree of relationship between u and v . Let $d_i = u_i - v_i$. The values of d give an indication of the closeness of the correspondence between A and B . We first observe that the relationship would be positively perfect

(i.e., there would be perfect agreement) if for each individual the ranks in the first and second series would coincide, and in that case

$$\sum_i d_i^2 = \sum_i (u_i - v_i)^2 = 0.$$

Again, the relationship would be negatively perfect (i.e., there would be perfect disagreement) if the ranking in the first case were completely reversed in the second (i.e. if $v_i = n - u_i + 1$ for all i), and in that case

$$\begin{aligned} \sum_i d_i^2 &= \sum_i (2u_i - n - 1)^2 = 4 \sum_i u_i^2 - 4(n+1) \sum_i u_i + n(n+1)^2 \\ &= \frac{4n(n+1)(2n+1)}{6} - 4(n+1) \frac{n(n+1)}{2} + n(n+1)^2 \\ &= \frac{n(n^2-1)}{3}, \end{aligned}$$

since $\sum_i u_i$ is just the sum of the first n natural numbers, and $\sum_i u_i^2$ is the sum of their squares.

Spearman suggested as his coefficient of rank correlation (say, r_R) the measure

$$r_R = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}. \quad \dots \quad (13.1)$$

Obviously, $r_R = 1$ for the case of perfect agreement and $r_R = -1$ for the case of perfect disagreement between the two series of ranks.

This coefficient can also be deduced from considerations of product-moment. Taking u 's and v 's as variate-values, we have

$$\bar{u} = \frac{\sum_i u_i}{n} = \frac{1}{n} \times \frac{n(n+1)}{2} = \frac{n+1}{2}.$$

$$\text{Similarly, } v = \frac{n+1}{2}.$$

$$\begin{aligned} \text{Also, } s_u^2 &= \frac{1}{n} \sum_i (u_i - \bar{u})^2 = \frac{1}{n} \sum_i u_i^2 - \bar{u}^2 \\ &= \frac{1}{n} \times \frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}, \end{aligned}$$

and, in the same way,

$$s_v^2 = \frac{n^2-1}{12}.$$

Now, we have

$$\begin{aligned}\frac{1}{n} \sum_i d_i^2 &= \frac{1}{n} \sum_i \{(u_i - \bar{u}) - (v_i - \bar{v})\}^2 \\ &= s_u^2 + s_v^2 - 2 \operatorname{cov}(u, v),\end{aligned}$$

so that $\operatorname{cov}(u, v) = \frac{n^2 - 1}{12} - \frac{1}{2n} \sum_i d_i^2.$

Hence

$$\begin{aligned}\operatorname{corr}(u, v) &= \frac{\operatorname{cov}(u, v)}{s_u s_v} = \frac{\frac{n^2 - 1}{12} - \frac{1}{2n} \sum_i d_i^2}{\frac{n^2 - 1}{12}} \\ &= 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}. \quad \dots \quad (13.2)\end{aligned}$$

Thus r_R may be regarded as the simple product-moment correlation coefficient between the two series of ranks.

Next, let us consider *the case of ties*. If the same rank is allotted to k individuals, then we have a tie of length k . By convention, if these k individuals follow r other individuals in the ranking, then each will be given the rank

$$\frac{(r+1) + (r+2) + \dots + (r+k)}{k} = r + \frac{k+1}{2},$$

i.e. the average of the ranks which these individuals would have received had there been no ties.

This tie does not affect the mean of the ranks. This will, however, affect the variance. The sum of squares of the untied ranks would be

$$(r+1)^2 + (r+2)^2 + \dots + (r+k)^2 = kr^2 + k(k+1)r + \frac{1}{6}k(k+1)(2k+1),$$

and the sum of squares of the tied ranks is

$$k \left\{ r + \frac{k+1}{2} \right\}^2 = kr^2 + k(k+1)r + \frac{1}{4}k(k+1)^2,$$

the difference being $\frac{1}{12}(k^3 - k)$. Consequently, the variance is lowered by $\frac{1}{12n}(k^3 - k)$ in the case of tied ranks. Also, it is obvious that the effect of tying different sets is additive.

Now, suppose that in the ranking with respect to the first character, there are s ties of length k_1, k_2, \dots, k_s , and in the ranking

with respect to the second character, there are t ties of length k_1, k_2, \dots, k_t . The variances would then be

$$s_u^2 = \frac{n^2 - 1}{12} - T_u, \text{ where } T_u = \frac{1}{12n} \sum_{j=1}^t (k_j^3 - k_j),$$

$$\text{and } s_v^2 = \frac{n^2 - 1}{12} - T_v, \text{ where } T_v = \frac{1}{12n} \sum_{j=1}^t (k_j^3 - k'_j)$$

Similarly, since

$$2 \operatorname{cov}(u, v) = s_u^2 + s_v^2 - \frac{1}{n} \sum_i d_i^2,$$

the covariance for the case of tied ranks would be

$$\operatorname{cov}(u, v) = \frac{n^2 - 1}{12} - \frac{T_u + T_v}{2} - \frac{1}{2n} \sum_i d_i^2,$$

so that Spearman's rank correlation coefficient in the case of tied ranks becomes

$$r_P = \frac{\frac{n^2 - 1}{12} - \frac{T_u + T_v}{2} - \frac{1}{2n} \sum_i d_i^2}{\left(\frac{n^2 - 1}{12} - T_u \right)^{1/2} \left(\frac{n^2 - 1}{12} - T_v \right)^{1/2}}. \quad (13.3)$$

In case there is perfect agreement between the two sets of ranks, we shall have $u_i = v_i$, and hence

$$r_P = \frac{\frac{n^2 - 1}{12} - T_u}{\frac{n^2 - 1}{12} - T_u} = 1$$

Again, if there is perfect disagreement between the two sets of ranks, we shall have $v_i = n - u_i + 1$ and $r_P = -1$ (the relationship between u and v being represented exactly by a straight line with a negative slope)

Ex. 13.1 Ten hand-writings were ranked by two judges in a competition. The rankings are given below. Calculate Spearman's coefficient to measure the closeness of the two rankings.

	Hand writing									
	A	B	C	D	E	F	G	H	I	J
Judge 1	3	8	5	4	7	10	1	2	6	9
Judge 2	6	4	7	5	10	3	2	1	9	8

The differences d between the two series of ranks for the 10 hand-writings are :

$$-3, 4, -2, -1, -3, 7, -1, 1, -3, 1.$$

$$\text{Hence } \sum_i d_i^2 = 9 + 16 + 4 + 1 + 9 + 49 + 1 + 1 + 9 + 1 = 100.$$

$$\text{Also, } n(n^2 - 1) = 10^3 - 10 = 990.$$

Thus Spearman's r_R is

$$\begin{aligned} r_R &= 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \\ &= 1 - \frac{6 \times 100}{990} \\ &= 1 - 0.606 = 0.394. \end{aligned}$$

Thus there is no remarkable closeness between the two sets of ranks.

Ex. 13.2 Two supervisors ranked 12 workers working under them in order of efficiency. Calculate Spearman's rank correlation coefficient between the two rankings.

Worker	A	B	C	D	E	F	G	H	I	J	K	L
Supervisor 1	5	6	1	2	3	8½	8½	4	7	11	10	12
Supervisor 2	5½	5½	2	2	2	9	7	4	8	10½	12	10½

In the first ranking, there is only one tie of length 2. Thus

$$T_u = \frac{1}{12} \times \left(\frac{2^3 - 2}{12} \right)$$

$$= \frac{1}{24} = 0.0417.$$

In the second ranking, there are three ties of length 2, 3, and 2. Here

$$T_v = \frac{1}{12} \times \left\{ \frac{2^3 - 2}{12} + \frac{3^3 - 3}{12} + \frac{2^3 - 2}{12} \right\}$$

$$= \frac{1}{12} \times \left\{ \frac{1}{2} + 2 + \frac{1}{2} \right\} = 0.25.$$

Also,

$$\frac{n^2 - 1}{12} = \frac{12^2 - 1}{12} = \frac{143}{12} = 11.9167,$$

$$\text{and } \sum_{i=1}^n d_i^2 = \frac{1}{12} \times \left\{ \frac{1}{4} + \frac{1}{4} + 1 + 0 + 1 + \frac{1}{4} + \frac{9}{4} + 0 + 1 + \frac{1}{4} + 4 + \frac{9}{4} \right\} \\ = \frac{12.5}{12} = 1.0417$$

Thus

$$\tau = \frac{11.9167 - 1.0417 + 2500}{2} - \frac{1.0417}{2} = \frac{11.2500}{\sqrt{11.9167 - 1.0417} \sqrt{11.9167 - 2500}} = \frac{11.2500}{\sqrt{11.8750 \times 11.6667}} \\ = \frac{11.2500}{11.7715} = 0.956,$$

which indicates that the supervisors agree closely in their judgement

13.3 Kendall's rank correlation coefficient

Kendall's rank correlation coefficient τ may be obtained as follows. First, let us suppose that there is no tie. Consider each possible pair of individuals (i, j) and the order of this pair in the two rankings. If the pair appears in the same order in both rankings, we allot it a score of +1, and if it appears in reverse orders, a score of -1. The score is thus obtained for each of the $\binom{n}{2} = \frac{n(n-1)}{2}$ possible pairs. We then define a rank correlation coefficient τ as

$$\begin{aligned} \tau &= \frac{\text{total score}}{\text{maximum possible total score}} \\ &= \frac{\text{total score}}{n(n-1)/2} \end{aligned} \quad (13.4)$$

Obviously, $\tau = +1$ for perfect agreement, because the score for each pair, each being in the same order in both rankings, is +1, and $\tau = -1$ for perfect disagreement, because the score for each pair, the pair being in reverse orders in the two rankings, is now -1.

Suppose the ranking in one series is in the natural order, viz. the order 1, 2, ..., n . Let us consider the corresponding ranking in the other series. Suppose out of the $\binom{n}{2}$ pairs for the second series, P pairs have ranks in the natural order and Q pairs have ranks in the

reverse order. Obviously, the P pairs will receive a score of +1 each, while the Q pairs will receive a score of -1 each. Thus, according to (13.4),

$$\begin{aligned}\tau &= \frac{P-Q}{\binom{n}{2}} \\ &= 1 - \frac{2Q}{\binom{n}{2}} \quad \dots \quad (13.5a)\end{aligned}$$

$$= \frac{2P}{\binom{n}{2}} - 1, \quad \dots \quad (13.5b)$$

since

$$\begin{aligned}P+Q &= \text{total number of pairs} \\ &= \binom{n}{2}.\end{aligned}$$

This indicates that in case neither of the two series of ranks has ties, in computing τ one need determine P only (or Q only).

τ may also be regarded as a product-moment coefficient. For the ranking with respect to the first character, let u_i be the rank of the i th individual. For the pair (i, j) , with $i < j$, we define a_{ij} such that

$$a_{ij} = \begin{cases} +1 & \text{if } u_i < u_j, \\ -1 & \text{if } u_i > u_j. \end{cases}$$

We similarly define b_{ij} for the ranking with respect to the second character. It can then be easily verified that

$$\tau = \frac{\sum(a_{ij}b_{ij})}{\sqrt{\sum a_{ij}^2 \sum b_{ij}^2}}, \quad \dots \quad (13.6)$$

each summation being over all the possible pairs (i, j) , with $i < j$.

In the case of tied ranks, we proceed almost in the same way, but now we take

$$a_{ij} = 0 \text{ if } u_i = u_j,$$

and, similarly, we take

$$b_{ij} = 0 \text{ if } v_i = v_j.$$

Thus if there is a tie of length k , the score is reduced by $\frac{k(k-1)}{2}$, since $a_{ij} \times b_{ij} = 0$ if either a_{ij} or b_{ij} or both are zero. Therefore, if there are s ties of length k_1, k_2, \dots, k_s , in the ranking with respect

to the first character, $\sum a_{ij}^2$ would be reduced to

$$\frac{n(n-1)}{2} - T'_* \text{, where } T'_* = \frac{1}{2} \sum_{j=1}^t k_j(k_j-1).$$

Similarly, if there are t ties of length k'_1, k'_2, \dots, k'_t in the ranking with respect to the second character, then $\sum b_{ij}^2$ would be reduced to

$$\frac{n(n-1)}{2} - T'_{**} \text{, where } T'_{**} = \frac{1}{2} \sum_{j=1}^t k'_j(k'_j-1).$$

(The total score also will, naturally, get reduced.)

Thus the formula for τ in the case of tied ranks becomes

$$\tau = \frac{\text{total score}}{\left\{ \frac{n(n-1)}{2} - T'_* \right\}^{1/2} \left\{ \frac{n(n-1)}{2} - T'_{**} \right\}^{1/2}}. \quad \dots \quad (13.7)$$

It is apparent that both the coefficients, r_R and τ , are easy to calculate, and these have been advocated as rough and ready substitutes for the product-moment correlation between two measurable characters. In case the characters are not measurable in practice or are very difficult to measure, the rank correlation coefficients may be used as measures of the correspondence between the two characters. Kendall's τ has an advantage over Spearman's r_R in that it may be adapted more easily to the theory of sampling.

Ex. 13.3 Let us calculate Kendall's τ coefficient for the data of Ex. 13.1.

To calculate τ , it is convenient to rearrange one ranking so as to put it in the natural order 1, 2, ..., n . If we do so for the ranking by Judge 1, the corresponding ranking by Judge 2 becomes :

2, 1, 6, 5, 7, 9, 10, 4, 8 and 3.

The score obtained by considering the first member, 2, in conjunction with the others is $8-1=7$, because only 1 is smaller than 2. Similarly, the score involving the member 1 is 8, the score involving the member 6 is $4-3=1$, and so on. The total score is

$$7+8+1+2+1-2-3+0-1=19-6=13.$$

On the other hand, the maximum possible score is $(10 \times 9)/2=45$.

Thus

$$\tau = 13/45 = 0.289.$$

Ex. 13·4 To calculate τ for the data of Ex. 13·2, we rearrange the ranking of Supervisor 1 in the natural order, and then we have the two sets of ranks as follows :

Supervisor 1	1	2	3	4	5	6	7	$8\frac{1}{2}$	$8\frac{1}{2}$	10	11	12
Supervisor 2	2	2	2	4	$5\frac{1}{2}$	$5\frac{1}{2}$	8	9	7	12	$10\frac{1}{2}$	$10\frac{1}{2}$

The total score in this case is

$$9+9+9+8+6+6+3+3+3-2-0=54.$$

Here $T'_u = \frac{1}{2}\{2 \times 1\} = 1,$

and $T'_v = \frac{1}{2}\{2 \times 1 + 3 \times 2 + 2 \times 1\} = 5.$

Hence, from formula (13.7),

$$\begin{aligned}\tau &= \frac{54}{\sqrt{66-1}\sqrt{66-5}} \\ &= \frac{54}{\sqrt{65 \times 61}} \\ &= \frac{54}{62.97} = 0.858.\end{aligned}$$

Thus both Spearman's r_R and Kendall's τ indicate a very high degree of agreement between the two series of ranks.

13.4 Grade correlation

The ranking of an individual, as the r th in a group, may be regarded as a numerical statement to the effect that there are $(r-1)$ members who are given precedence over that individual. We will then define the *grade* of an individual as the proportion of individuals in the whole group with a lower variate value than the value possessed by that individual. If we have a discontinuous population of size N , the grade of an individual ranked, according to the variate values, as the r th (assuming that the ranking proceeds from the higher to the lower variate values) will be $\frac{N-r}{N}$. In case the population is continuous, its members cannot be ranked ; but if a sample of size n is selected, the grade of an individual with rank r can be estimated if

it is assumed that one half of that member is to be assigned to each of the two parts into which the variate value divides the variate range, and we have g_r , the grade corresponding to rank r , as

$$g_r = \frac{n-r+\frac{1}{2}}{n} = 1 - \frac{r-\frac{1}{2}}{n} \quad (13.8)$$

For a continuous bivariate population, there will be no rank correlation, but one can nevertheless think of a grade correlation. To each individual of a bivariate population, there will be attached two grades corresponding to the two variate values, and the product moment correlation of these grades is the grade correlation. For a bivariate normal population with correlation coefficient ρ , it can be shown that

$$\rho = 2 \sin(\pi \rho_g / 6), \quad (13.9)$$

where ρ_g is the grade correlation. This relation, however, should not be used to transform a rank correlation coefficient obtained from a sample into the product moment correlation coefficient in that sample, or in the population from which the sample was taken, unless the population is known to be bivariate normal and the sample size is sufficiently large.

Ex 13.5 Find an estimate of the correlation coefficient in a bivariate normal population if the rank correlation from a sample is 0.45

Here ρ_g (estimated) = 0.45

so that $\frac{\pi \rho_g}{6}$ (estimated) = 13.5°

Thus ρ (estimated) = $2 \sin \frac{\pi \rho_g}{6}$ (estimated)

$$= 2 \sin 13.5^\circ = 0.47$$

13.5 Intra-class correlation

We have considered in Chapter 11 the correlation between two clearly defined variates, such as the marks in college test and marks in university examination for a group of students or the family income and percentage of family income spent on food for a group of families. There sometimes arise cases in which we require the correlation with respect to a particular variate between members of

the same class, e.g. with respect to marks in a university examination of students belonging to the same tutorial group or with respect to height of brothers in the same family. By correlation here we mean the extent to which the members of the same class (group or 'family') resemble each other with respect to the given variable. Such a correlation we shall call *intra-class correlation*, to distinguish it from ordinary correlation, which may be called *inter-class correlation*.

Suppose we are investigating the correlation between heights of brothers, and suppose there are two brothers in each family. If we take the height of the elder brother or the taller brother as the first variate and the height of the younger or the shorter as the second variate and find the correlation, we would get the correlation between the height of elder brother and the height of younger brother or between the height of taller brother and the height of shorter brother and not the relationship of heights of brothers in general. To get the relationship of heights of brothers in general, we have to take in turn the height of each brother as the first variate and that of the other as the second variate, and thus get two pairs of heights for each family. Similarly, if there are k brothers in the family, there will be $k(k-1)$ pairs. Thus if we have p families of k brothers each, there will be $pk(k-1)$ pairs of values in the correlation table. Let x_{ij} denote the variate value for the j th member of the i th family ($i=1, 2, \dots, p$; $j=1, 2, \dots, k$). If we arbitrarily regard the first column of the correlation table as corresponding to a variate U and the second column as corresponding to a variate V , then the mean of each variate is given by

$$\bar{U} = \bar{V} = \frac{1}{pk(k-1)} \sum_{i=1}^p (k-1) \sum_{j=1}^k x_{ij} = \frac{1}{pk} \sum_i \sum_j x_{ij}$$

$$= \bar{x}, \text{ the grand mean of } x, \quad \dots \quad (13.10)$$

since each of the x_{ij} 's occurs $(k-1)$ times in each column of the correlation table, along with the values for the other $(k-1)$ members of the family occurring in the other column.

Similarly, the variance of each variate is given by

$$s_U^2 = s_V^2 = \frac{1}{pk(k-1)} \sum_{i=1}^p (k-1) \sum_{j=1}^k (x_{ij} - \bar{x})^2$$

$$= \frac{1}{pk} \sum_i \sum_j (x_{ij} - \bar{x})^2 = s^2, \text{ the total variance of } x. \quad \dots \quad (13.11)$$

The covariance between U and V is given by

$$\text{cov}(U, V) = \frac{1}{pk(k-1)} \sum_{i=1}^p \sum_{\substack{j, j' = 1 \\ j \neq j'}}^k (x_{ij} - \bar{x})(x_{ij'} - \bar{x}),$$

where the second summation extends over all pairs of members (j, j') in the i th family, with $j \neq j'$. But

$$\begin{aligned} & \sum_{\substack{j, j' = 1 \\ j \neq j'}}^k (x_{ij} - \bar{x})(x_{ij'} - \bar{x}) \\ &= \sum_{i=1}^p \sum_{j=1}^k (x_{ij} - \bar{x})(x_{ij'} - \bar{x}) - \sum_{i=1}^p (x_{ii} - \bar{x})^2 \\ &= \sum_{i=1}^p (x_{ii} - \bar{x}) \sum_{j=1}^k (x_{ij} - \bar{x}) - \sum_{i=1}^p (x_{ii} - \bar{x})^2 \\ &= k^2 (\bar{x}_i - \bar{x})^2 - \sum_{i=1}^p (x_{ii} - \bar{x})^2, \end{aligned}$$

where $\bar{x}_i = \frac{1}{k} \sum_{j=1}^k x_{ij}$, the mean of x in the i th family.

Thus

$$\text{cov}(U, V) = \frac{k^2}{pk(k-1)} \sum_{i=1}^p (\bar{x}_i - \bar{x})^2 - \frac{1}{pk(k-1)} \sum_{i=1}^p \sum_{j=1}^k (x_{ij} - \bar{x})^2. \quad \dots \quad (13.12)$$

Writing $s_m^2 = \frac{1}{p} \sum_{i=1}^p (\bar{x}_i - \bar{x})^2$, the variance of the means of the p families we have

$$\text{cov}(U, V) = \frac{k}{k-1} s_m^2 - \frac{s^2}{k-1}$$

Thus the coefficient of intra-class correlation r_I is given by

$$\begin{aligned} r_I &= \frac{\text{cov}(U, V)}{\sqrt{\text{var}(U)\text{var}(V)}} = \frac{\frac{k}{k-1} s_m^2 - \frac{s^2}{k-1}}{s^2} \\ &= \frac{1}{k-1} \left\{ \frac{s_m^2}{s^2} - 1 \right\}. \quad \dots \quad (13.13) \end{aligned}$$

Note that $0 \leq s_m^2 \leq s^2$. Hence this coefficient is a maximum, viz. equal to +1, when $s_m^2 = s^2$, i.e. when the variance between the mean is equal to the total variance, which happens when the variance within families is zero. In this case the variate-values for member within each family are all equal.

Again, this coefficient is a minimum, viz. equal to $-\frac{1}{k-1}$, when s_m^2 , the variance between family means, is zero, which happens when the variance within families is the maximum possible. Thus the coefficient of intra-class correlation may be looked upon as a measure of the extent to which the total variance may be explained away by the variance between means.

Some caution is necessary in the interpretation of the intra-class correlation. It is clearly seen that here r_I varies from $-\frac{1}{k-1}$ to $+1$.

The lower limit is larger than -1 unless $k=2$. It is thus a skew coefficient in the sense that a negative value has not the same significance (as a departure from zero) as the corresponding positive value.

In the general case, when there are k_i members in the i th class, in the correlation table each member of the i th class will appear $(k_i - 1)$ times in each column in association with the other members of the class. Here

$$\bar{U} = \bar{V} = \frac{1}{N} \sum_{i=1}^p \left\{ (k_i - 1) \sum_{j=1}^{k_i} x_{ij} \right\} = \bar{x}_0, \text{ say}$$

(which is not the grand mean of x),

where $N = \sum_{i=1}^p k_i(k_i - 1)$,

the first summation is over the p classes, while the second is over all members of the i th class. Similarly,

$$s_U^2 = s_V^2 = \frac{1}{N} \sum_{i=1}^p \left\{ (k_i - 1) \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_0)^2 \right\}.$$

Again,

$$\begin{aligned} \text{cov}(U, V) &= \frac{1}{N} \sum_{i=1}^p \sum_{j,j'=1}^{k_i} (x_{ij} - \bar{x}_0)(x_{ij'} - \bar{x}_0), \quad \text{with } j \neq j' \\ &= \frac{1}{N} \sum_{i=1}^p \sum_{j,j'=1}^{k_i} (x_{ij} - \bar{x}_i)(x_{ij'} - \bar{x}_i) - \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_i)^2 \end{aligned}$$

(where \sum' extends over all possible pairs, including the case $j=j'$)

$$= \frac{1}{N} \sum_i k_i (\bar{x}_i - \bar{x}_0)^2 - \frac{1}{N} \sum_i \sum_j (x_{ij} - \bar{x}_i)^2,$$

where \bar{x}_i is the mean of the i th class, being equal to $\frac{1}{k_i} \sum_{j=1}^{k_i} x_{ij}$.

Hence

$$r_I = \frac{\text{cov}(U, V)}{\sqrt{\text{var}(U)\text{var}(V)}} = \frac{\sum k_i^2(x_i - \bar{x}_0)^2 - \sum \sum (x_{ij} - \bar{x}_0)^2}{\sum \{(k_i - 1) \sum (x_{ij} - \bar{x}_0)^2\}} \quad (13.14)$$

Ex 13.6 The weights in gm. of a number of copper wires, each of length 1 metre, were obtained. These are shown below classified according to the die from which they come. Determine the intra-class correlation.

Die No				
I	II	III	IV	V
1 33	1 30	1 32	1 31	1 30
1 32	1 35	1 29	1 29	1 32
1 36	1 33	1 31	1 33	1 33
1 35	1 34	1 28	1 31	1 33

The intra-class correlation coefficient is invariant under a change of base and scale. Let our new variable be

$$u = 100(x - 1.28),$$

where x is the original variable

This means that

$$u_i = 100(x_{ij} - 1.28)$$

for $i=1, 2, 3, 4, 5$ and $j=1, 2, 3, 4$. These values are shown in Table 13.1

TABLE 13.1
WEIGHTS OF COPPER WIRES AFTER CHANGE OF BASE AND SCALE

Die No					
	I	II	III	IV	V
	5	2	4	3	2
	4	7	1	1	4
	8	5	3	5	5
	7	6	0	3	5
Total	24	20	8	12	16

Here the grand mean is $\bar{u} = \frac{80}{20} = 4.0$.

$$\begin{aligned}\text{Also, } s^2 &= \frac{1}{20} \sum_{i=1}^5 \sum_{j=1}^4 (u_{ij} - \bar{u})^2 \\ &= \frac{1}{20} \{1+0+16+9+4+9+1+4+0+9+1+16 \\ &\quad + 1+9+1+1+4+0+1+1\} \\ &= \frac{88}{20} = 4.4,\end{aligned}$$

$$\text{and } s_m^2 = \frac{1}{5} \sum_i (\bar{u}_i - \bar{u})^2 = \frac{1}{5} \{4+1+4+1+0\} = 2.$$

Hence the intra-class correlation coefficient r_I is given by

$$\begin{aligned}r_I &= \frac{1}{k-1} \left\{ \frac{ks_m^2}{s^2} - 1 \right\} \\ &= \frac{1}{4} \left\{ \frac{5 \times 2}{4.4} - 1 \right\} \\ &= \frac{1}{4} \{2.273 - 1\} \\ &= \frac{1.273}{4} = 0.318.\end{aligned}$$

This indicates that the copper wires coming from the same die resemble each other, in respect of their weights, only to a moderate degree.

13.6 Population intra-class correlation

In many situations, the value x_{ij} may be supposed to have arisen from sampling a population in two stages : first a family is chosen at random from a whole set of families, and next a member is selected at random from all members of the chosen family (and the value of x for this member being noted). In such a case, one may write

$$x_{ij} = \mu + m_i + e_{ij}, \quad \dots \quad (13.15)$$

where μ (a constant) is the mean of x for all individuals in all families taken together, m_i (a random variable) is the amount by which the mean of the chosen family differs from μ and e_{ij} (a random variable) is the amount by which x_{ij} differs from the family mean.

We shall denote the variance of all m_i 's by σ_m^2 , while the variance

of e_{ij} 's (assumed to be the same for each i) will be denoted by σ_e^2 . Further, it will be assumed that the m_i 's are mutually independent, the e_{ij} 's are mutually independent, while the m_i 's are independent of the e_{ij} 's.*

Now, if we consider two members of the same family, for which the values of x are x_{ij} and x_{il} , then

$$\text{var}(x_{ij}) = \text{var}(x_{il}) = \sigma_m^2 + \sigma_e^2 = \sigma^2 \text{ (say),}$$

while

$$\text{cov}(x_{ij}, x_{il}) = \sigma_m^2$$

Hence the correlation coefficient between x_{ij} and x_{il} , which is nothing but the intra class correlation coefficient under the above model, is

$$\rho_I = \frac{\sigma_m^2}{\sigma^2} \quad (13.16)$$

This is seen to be the proportion of the total variance of x that is represented by the variance of family means. Hence the higher the value of σ_m^2 (i.e. the smaller the value of σ_e^2) relative to σ^2 , the higher is the intra-class correlation.

Formula (13.16) may be seen to have the same form as formula (13.13) if we remember that under the present model k is supposed to be practically infinite.

Questions and exercises

13.1 What is a rank correlation coefficient? Deduce Spearman's formula for rank correlation coefficient. How should the formula be modified for tied ranks?

13.2 Discuss the rationale behind Kendall's τ coefficient for rank correlation. Also indicate how the formula can be adapted to the case of tied ranks.

13.3 Show that both Spearman's r_s and Kendall's τ will lie between -1 and $+1$. Interpret the marginal cases.

13.4 Define intra-class correlation and distinguish it from inter-class correlation. Derive the formula for intra class correlation when a variate x is observed for p families, each consisting of k members.

*For a further discussion of this model, see Section 19.5, Vol. 2

13.5 Show that the coefficient derived in *Exercise 13.4* lies between the limits $-\frac{1}{k-1}$ and +1. Interpret the marginal cases.

13.6 Show that the intra-class correlation coefficient defined by (13.14) also takes its highest possible value +1 when the members of each family have the same value of x .

Verify that formula (13.14) reduces to formula (13.13) in case $k_i=k$, i.e. in case the families are all of the same size.

13.7 Two judges rank a number of competitors in a certain art competition as follows :

	Competitor											
	1	2	3	4	5	6	7	8	9	10	11	12
Judge A	5	1	4	2	7	3	6	8	10	9	11	12
Judge B	10	5	1	2	3	4	7	6	8	11	9	12

Measure the association between the judgements of the two judges by using (1) Spearman's r_R and (2) Kendall's τ .

$$\text{Ans. } r_R = 0.706, \tau = 0.515.$$

13.8 Six boys and six girls are ranked below on the basis of their performance in a mathematics test. Do you find any association between sex and performance? Use (1) Spearman's r_R and (2) Kendall's τ formulæ for tied ranks, assuming that, in ranking with respect to sex, boys are considered superior to girls.

Sex	B	B	G	B	B	G	G	G	B	G	B	G
Rank	2	2	2	4	5	6.5	6.5	8	9.5	11	9.5	12

B=boy, G=girl.

$$\text{Ans. } r_R = 0.407, \tau = 0.299.$$

13.9 For each of six families, the heights in inches of three brothers belonging to it are recorded below. Compute the coefficient of intra-class correlation.

<i>Family</i>	<i>Heights of brothers</i>		
1	69 5	70 6	72 3
2	71 2	70 8	72 0
3	65 6	67 2	66 7
4	62 2	63 6	63 5
5	68 0	70 5	70 5
6	64 4	64 3	64 6

$$Ans \quad r_I = 0.910$$

13.10 The birth weights of babies born to 5 mothers in a Calcutta nursing home are given below. Compute the intra class correlation

<i>Mother</i>				
1	2	3	4	5
6 lb 3 oz	7 lb 9 oz	5 lb 10 oz	8 lb 6 oz	6 lb 9 oz
6 lb 8 oz	8 lb 2 oz	6 lb 2 oz	7 lb 10 oz	6 lb 6 oz
6 lb 0 oz		6 lb 4 oz	7 lb 5 oz	
		6 lb 4 oz		

$$Ans \quad r_I = 0.826$$

SUGGESTED READING

- [1] Fisher, R A *Statistical Methods for Research Workers* (Ch 7) Oliver & Boyd, 1948
- [2] Kendall, M G *Rank Correlation Methods* (Chs 1—3) Charles Griffin, 1948
- [3] Yule, G U and Kendall, M G *Introduction to the Theory of Statistics* (Ch 11) Charles Griffin, 1953

14

RANDOM SAMPLING AND SAMPLING DISTRIBUTIONS

14.1 Random sampling

The selection of a sample (consisting of, say, n members) from a population (having, say, N members) may be done in a number of ways ; that is to say, we may have different types of sampling. Suppose, for instance, that a social scientist wants to determine the average income per family for families residing in Calcutta. To know this figure exactly, he will have to study the income of each of some ten lakhs of families living in the city. This will require a lot of time and money, which the enquirer may not be able to afford. To manage within his prescribed time and limited resources, he will in such a case study a sample of families only—some 100 or 1,000 of them—and will base his conclusions on the characteristics of the sample. In this investigation, the n families in the sample may be the first n or the last n appearing in the National Register for Calcutta ; alternatively, they may be chosen by considering, say, every 100th family in the register beginning with the first entry ; and so on. All these methods, however, are defective in that they frequently lead to unrepresentative samples. A more serious drawback is that in such sampling procedures no idea can be obtained from the sample regarding the possible deviations of the characteristics of the sample from the characteristics of the population.

To avoid the second drawback, one may use what is called *probability sampling*. It also takes care of the first defect in the long run (in a sense to be explained later).* In this case, the sampling procedure is such that each member of the population gets a definite probability of being included in the sample. The simplest and the most commonly used type of probability sampling is *simple random sampling* (or *random sampling*, for short). In this kind of sampling, each member of the population has the same probability of being included in the sample.

*E.g., if \bar{x} be the sample mean of x for random samples drawn from a population, the population mean of x being μ , then \bar{x} is a *consistent estimator* of μ .

We may again have two distinct types of simple random sampling. In one case, the n units of the sample are drawn from the population one by one, after each drawing the individual selected being returned to the population, in such a way that at each drawing each of the N members of the population gets the same probability $\frac{1}{N}$ of being selected. This is simple random sampling *with replacements*.

Clearly, here the same unit of the population may occur more than once in the sample, there are N^n possible samples, regard being had to the order in which the n sample units appear, and each has the probability $\frac{1}{N^n}$ to materialise.

If, on the other hand, the n members of the sample are drawn one by one but the member obtained at any drawing is not returned to the population and if at each stage every remaining unit of the population (at the r th drawing each of the remaining $N-r+1$ units) is given the same probability $\frac{1}{N-r+1}$ of being included in the sample, then we have simple random sampling *without replacements*.

Here no member of the population can occur more than once in the sample. There are $\binom{N}{n}$ conceivable samples, provided the order in which the sample units are obtained is ignored, and each such sample has the probability

$$\frac{n}{N} \times \frac{n-1}{N-1} \times \dots \times \frac{1}{N-n+1} = \frac{1}{\binom{N}{n}}$$

to materialise. This is so because at the r th stage one is to choose from $N-r+1$ individuals one of the $n-r+1$ individuals to be included in the sample, which have not yet been chosen in earlier drawings. It may be seen that in this case, too, the probability that any specified individual, say the i th, is selected at any drawing, say the k th drawing, is

$$\frac{N-1}{N} \times \frac{N-2}{N-1} \times \dots \times \frac{N-k+1}{N-k+2} \times \frac{1}{N-k+1} = \frac{1}{N},$$

as in simple random sampling with replacements.

It is obvious that if one takes n individuals all at a time from the

population, giving equal probability to each of the $\binom{N}{n}$ combinations of n members out of the N members in the population, one will still have simple random sampling without replacements.

Some practical methods of obtaining random samples will be discussed in Volume 2.

14.2 Parameter, statistic and its sampling distribution

Generally in statistical investigations, our ultimate interest lies in one or more characters possessed by the members of the population. On taking a sample, we shall then observe the forms or the values of the characters for the individuals included in the sample. Supposing there is only one character of importance, it can be assumed, without any loss of generality, to be a variable x ; for the case of an attribute can also be tackled by methods meant for a variable.* If x_i be the value of x for the i th member of the sample, then x_1, x_2, \dots, x_n are the sample observations.

Generally, again, our primary interest will be in knowing the values of different measures of the variable x , like its mean, standard deviation, etc. A measure of this type, calculated on the basis of population values of x , is called a *parameter*. (The word ‘parameter’ is being used here in a general sense. In a narrower sense, a parameter is a measure that occurs in the probability distribution of the variable; e.g., λ is the parameter of a Poisson variable, μ and σ are the parameters of a normal variable.) A corresponding measure computed on the basis of sample values is called a *statistic*.

Since the sets of population members included in different samples from the same population may be different, the value of the statistic itself is liable to vary from one sample to another. These differences in the values of a statistic are called *sampling fluctuations*. Thus if a number of samples, each of size n , are taken from the same population and if for each sample the value of the statistic is calculated, a series of values of the statistic will be obtained. If the number of samples is large, these may be arranged into a frequency table. The frequency distribution of the statistic that would be obtained if the number of samples, each of the same size (say n),

*In the case of an attribute, one will deal with the sample frequencies for the different classes, which are variables assuming non-negative integral values.

were infinite is called the *sampling distribution* of the statistic. In the case of random sampling, the nature of the sampling distribution of a statistic can be deduced theoretically, provided the nature of the population is given, from considerations of probability theory.

Like any other distribution, a sampling distribution may have its mean, standard deviation and moments of higher orders. Of particular importance is the standard deviation, which is designated as the *standard error* of the statistic. As illustrations, in the next section we derive for the case of random sampling the means (expectations) and standard errors of a sample mean and a sample proportion.

Some people prefer to use 0.6745 times the standard error, which is called the *probable error* of the statistic. The relevance of the probable error stems from the fact that for a normally distributed variable x with mean μ and s.d. σ ,

$$P[\mu - 0.6745\sigma \leq x \leq \mu + 0.6745\sigma] = 0.50,$$

approximately

14.3 Expectation and standard error of sample mean

Suppose a random sample of size n is drawn from a population of size N .

Let X_α ($\alpha=1, 2, \dots, N$) (14.1)

be the value of the variable x for the α th member of the population. Then the population mean of x is

$$\mu = \frac{1}{N} \sum_{\alpha} X_\alpha, \quad (14.2)$$

and the population variance is

$$\sigma^2 = \frac{1}{N} \sum_{\alpha} (X_\alpha - \mu)^2 \quad (14.3)$$

Again, let us denote by

$$x_i (i=1, 2, \dots, n) \quad (14.4)$$

the value of x for the i th member (i.e. the member selected at the i th drawing) of the sample. The sample mean of x is then

$$\bar{x} = \frac{1}{n} \sum_i x_i \quad (14.5)$$

For deriving the expectation and standard error of \bar{x} , we may consider two distinct cases.

Case 1. Random sampling with replacements

It is immediately seen from *Theorems 3.6 and 3.7* that

$$E(\bar{x}) = \frac{1}{n} \sum_i E(x_i)$$

and

$$\text{var}(\bar{x}) = E\{\bar{x} - E(\bar{x})\}^2$$

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

$$= \frac{1}{n^2} \sum_i E\{x_i - E(x_i)\}^2 + \frac{1}{n^2} \sum_{i \neq j} E\{x_i - E(x_i)\} \{x_j - E(x_j)\}$$

$$= \frac{1}{n^2} \sum_i \text{var}(x_i) + \frac{1}{n^2} \sum_{i \neq j} \text{cov}(x_i, x_j).$$

To obtain $E(x_i)$ and $\text{var}(x_i)$, we note that x_i can assume the values X_1, X_2, \dots, X_N , each with probability $\frac{1}{N}$.

$$\text{Hence } E(x_i) = \sum_a X_a \cdot P[x_i = X_a] = \sum_a X_a \cdot \frac{1}{N} = \mu,$$

$$\begin{aligned} \text{and } \text{var}(x_i) &= E(x_i - \mu)^2 = \sum_a (X_a - \mu)^2 \cdot P[x_i = X_a] \\ &= \sum_a (X_a - \mu)^2 \cdot \frac{1}{N} = \sigma^2, \end{aligned}$$

for each i .

Again,

$$\begin{aligned} \text{cov}(x_i, x_j) &= E(x_i - \mu)(x_j - \mu) \\ &= \sum_{a,a'} (X_a - \mu)(X_{a'} - \mu) \cdot P[x_i = X_a, x_j = X_{a'}]. \end{aligned}$$

Since in sampling with replacements the composition of the population remains the same throughout the sampling process, x_j can take any one of the values X_1, X_2, \dots, X_N , with probability $1/N$, irrespective of the value taken by x_i . In other words, for $i \neq j$, x_i and x_j are independent, so that

$$P[x_i = X_a, x_j = X_{a'}] = P[x_i = X_a] \cdot P[x_j = X_{a'}] = \frac{1}{N^2}.$$

$$\begin{aligned} \text{Hence } \text{cov}(x_i, x_j) &= \frac{1}{N^2} \sum_{a,a'} (X_a - \mu)(X_{a'} - \mu) \\ &= \frac{1}{N^2} \sum_a (X_a - \mu) \sum_{a'} (X_{a'} - \mu) = 0, \end{aligned}$$

for each i, j ($i \neq j$), since $\sum_a (X_a - \mu) = \sum_{a'} (X_{a'} - \mu)$, being the sum of the deviations of X_1, X_2, \dots, X_N from their mean, is zero.

Hence we have, finally,

$$E(\bar{x}) = \frac{1}{n} \times n\mu = \mu \quad (14.6)$$

and

$$\begin{aligned} \text{var}(\bar{x}) &= \frac{1}{n^2} \times n\sigma^2 + \frac{1}{n^2} \times n(n-1) \times 0 \\ &= \frac{\sigma^2}{n} \end{aligned} \quad . \quad (14.7)$$

The standard error of \bar{x} is, therefore,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (14.8)$$

Case 2 Random sampling without replacements

As before, for each i ,

$$E(x_i) = \mu \text{ and } \text{var}(x_i) = \sigma^2,$$

since here, too, x_i can take one of the values X_1, X_2, \dots, X_N with the same probability $\frac{1}{N}$. The covariance terms, however, need special attention.

Here, for $i \neq j$

$$\begin{aligned} P[x_i = X_\alpha, x_j = X_{\alpha'}] &= P[x_i = X_\alpha] P[x_j = X_{\alpha'} | x_i = X_\alpha] \\ &= \frac{1}{N} \times \frac{1}{N-1} \text{ if } \alpha \neq \alpha' \end{aligned}$$

(since x_j can take any value except X_α , the value which is known to have been already assumed by x_i , with equal probability $\frac{1}{N-1}$),

and

$$= 0 \text{ if } \alpha = \alpha'$$

Hence

$$\begin{aligned} \text{cov}(x_i, x_j) &= \sum_{\substack{\alpha, \alpha' \\ \alpha \neq \alpha'}} (X_\alpha - \mu)(X_{\alpha'} - \mu) \frac{1}{N(N-1)} \\ &= \frac{1}{N(N-1)} \sum_{\alpha} (\lambda_{\alpha} - \mu) \{ \sum_{\alpha'} (X_{\alpha'} - \mu) - (X_{\alpha} - \mu) \} \\ &= \frac{1}{N(N-1)} \{ \sum_{\alpha} (X_{\alpha} - \mu) \sum_{\alpha'} (X_{\alpha'} - \mu) - \sum_{\alpha} (X_{\alpha} - \mu)^2 \} \\ &= -\frac{1}{N(N-1)} N\sigma^2 = -\frac{\sigma^2}{N-1} \end{aligned}$$

Thus, in this case we have

$$E(\bar{x}) = \frac{1}{n} \times n\mu = \mu \quad \dots \quad (14.9)$$

and

$$\begin{aligned} \text{var}(\bar{x}) &= \frac{1}{n^2} \times n\sigma^2 + \frac{1}{n^2} \times n(n-1) \times \left(-\frac{\sigma^2}{N-1}\right) \\ &= \frac{\sigma^2}{n} \left(1 - \frac{n-1}{N-1}\right). \end{aligned} \quad \dots \quad (14.10)$$

Hence the standard error of \bar{x} is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}. \quad \dots \quad (14.11)$$

In both cases, the standard error decreases with increasing n . The standard error of the mean in sampling without replacements is, however, smaller than that in sampling with replacements. But the difference becomes negligible if N is very large compared to n . Also, in sampling without replacements, the standard error of the sample mean vanishes if $n=N$, which is to be expected because the sample mean now becomes a constant, i.e. the same as the population mean. However, this is not the case with sampling with replacements.

14.4 Expectation and standard error of sample proportion

Suppose in a population of N members, there are Np members with a particular character A and Nq members with the character *not-A*. Then p is the proportion of members in the population having the character A . Let a sample of size n be drawn from the population, and let f be the number of members in the sample having the character A . To find the expectation and the standard error of the sample proportion f/n , we adopt the following procedure.

We assign to the a th member of the population the value X_a , which is equal to 1 if this member possesses the character A and equal to 0 otherwise. Similarly, to the i th member of the sample we assign the value x_i , which is equal to 1 if this member possesses A and equal to 0 otherwise.

In this way, we get a variable x , which has population mean

$$\frac{1}{N} \sum_a X_a = p$$

and population variance

$$\frac{1}{N} \sum_{\alpha} X_{\alpha}^2 - p^2 = p - p^2 = pq$$

The sample mean of the variable x , on the other hand, is

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

Hence we find, on replacing x by \bar{x} , p by p and σ^2 by pq in the expressions (14.6), (14.8), (14.9) and (14.11), that

$$E(\bar{x}) = p \quad (14.12)$$

and $\sigma_{\bar{x}} = \sqrt{\frac{pq}{n}}$ (14.13)

in the case of random sampling with replacements, and

$$E(\bar{x}) = p \quad (14.14)$$

and $\sigma_{\bar{x}} = \sqrt{\frac{pq}{n} \left(1 - \frac{n-1}{N-1}\right)}$ (14.15)

in the case of random sampling without replacements

The comments made in connection with the standard error of the mean apply here also

14.5 Sampling distributions associated with discrete populations

We shall here and in the next section derive some common sampling distributions

Note that the essential feature of random sampling from a *finite population* is that each of the sample observations x_1, x_2, \dots, x_n has the same (marginal) distribution and this common distribution is identical with the distribution of x in the population. Indeed, this feature is taken to *define* random sampling from a population, finite or infinite. In what follows, we shall assume that x_1, x_2, \dots, x_n are *not only random but also independent*. Hence (a) in the discrete case, if $f(x)$ is the probability mass function of x , then the joint probability-mass function of x_1, x_2, \dots, x_n will be supposed to be

$$\prod_{i=1}^n f(x_i)$$

Similarly, (b) in the continuous case, if $f(x)$ be the probability-density function of x , then the joint probability density function of

x_1, x_2, \dots, x_n will be supposed to be

$$\prod_{i=1}^n f(x_i).$$

(a) *Sampling distribution of sample total : binomial parent*

Suppose x_1 and x_2 are distributed independently in the binomial form with parameters m_1, p and m_2, p , respectively. Consider then the distribution of the sum x_1+x_2 . Obviously, the values this sum can take are $0, 1, 2, \dots, m_1+m_2$.

Also,

$$\begin{aligned} P[x_1+x_2=k] &= \sum_{k_1=0}^k P[x_1=k_1] \cdot P[x_2=k-k_1] \\ &= \sum_{k_1=0}^k \binom{m_1}{k_1} \binom{m_2}{k-k_1} p^{k_1} (1-p)^{m_1+m_2-k} \\ &= p^k (1-p)^{m_1+m_2-k} \sum_{k_1=0}^k \binom{m_1}{k_1} \binom{m_2}{k-k_1}. \end{aligned}$$

Now, this sum is nothing but the sum of products of the coefficients of t^{k_1} in $(1+t)^{m_1}$ and of t^{k-k_1} in $(1+t)^{m_2}$, for varying k_1 , and hence equals the coefficient of t^k in $(1+t)^{m_1+m_2}$, which is $\binom{m_1+m_2}{k}$. Thus

$$P[x_1+x_2=k] = \binom{m_1+m_2}{k} p^k (1-p)^{m_1+m_2-k}.$$

This shows that x_1+x_2 is itself binomially distributed with parameters m_1+m_2 and p . We also get from this the general result that if x_1, x_2, \dots, x_n are independently distributed binomial variables with parameters $m_1, p ; m_2, p ; \dots ; m_n, p$, then the sum $x_1+x_2+\dots+x_n$ is also a binomial variable with parameters $m_1+m_2+\dots+m_n$ and p .

This implies that if x_1, x_2, \dots, x_n are randomly and independently taken from a binomial distribution with parameters m and p , then the sampling distribution of the statistic $x_1+x_2+\dots+x_n$ is also binomial with parameters nm and p .

(b) *Sampling distribution of sample total : Poisson parent*

Suppose x_1 and x_2 are distributed independently in the Poisson form with parameters λ_1 and λ_2 , respectively. The sum x_1+x_2

can then take the values 0, 1, 2, ... Also,

$$\begin{aligned}
 P[x_1 + x_2 = k] &= \sum_{k_1=0}^k P[x_1 = k_1] P[x_2 = k - k_1] \\
 &= \sum_{k_1=0}^k \frac{\exp(-\lambda_1)\lambda_1^{k_1}}{k_1!} \frac{\exp(-\lambda_2)\lambda_2^{k-k_1}}{(k-k_1)!} \\
 &= \frac{\exp(-\lambda_1-\lambda_2)}{k!} \sum_{k_1=0}^k \binom{k}{k_1} \lambda_1^{k_1} \lambda_2^{k-k_1} \\
 &= \exp(-\lambda_1-\lambda_2) \frac{(\lambda_1+\lambda_2)^k}{k!},
 \end{aligned}$$

which shows that $x_1 + x_2$ is itself a Poisson variable with parameter $\lambda_1 + \lambda_2$. It immediately follows that if x_1, x_2, \dots, x_n are independently distributed Poisson variables with parameters $\lambda_1, \lambda_2, \dots, \lambda_n$, then the sum $x_1 + x_2 + \dots + x_n$ is also a Poisson variable with parameter $\lambda_1 + \lambda_2 + \dots + \lambda_n$.

The above result gives, in particular, the sampling distribution of the statistic $x_1 + x_2 + \dots + x_n$ when x_1, x_2, \dots, x_n are independent random observations from a Poisson population with parameter λ . This sampling distribution is also of the Poisson form with parameter $n\lambda$.

14.6 Four fundamental distributions derived from the normal

(a) Distribution of normal deviate

The definition of the normal deviate has already been given in Chapter 9 – it is a normal variable with mean zero and standard deviation unity. Thus the probability-density function of the distribution is

$$f(\tau) = \frac{1}{\sqrt{2\pi}} \exp[-\tau^2/2], \quad (14.16)$$

where $-\infty < \tau < \infty$

The properties of this distribution may be deduced from those of a simple normal distribution.

We shall denote by τ_α the value of τ such that

$$P[\tau > \tau_\alpha] = \alpha \quad (14.17)$$

It is called the upper α -point (or $100\alpha\%$ -point) of the normal

deviate. Because of the symmetry of the distribution about zero, we have

$$\tau_{1-\alpha} = -\tau_\alpha.$$

Thus the lower α -point of the normal deviate, $\tau_{1-\alpha}$ —which is the value of τ such that

$$P[\tau < \tau_{1-\alpha}] = \alpha \quad \dots \quad (14.18)$$

—is the same as the upper α -point in magnitude but has the opposite sign.

Fig. 14.1 shows the curve of this distribution.

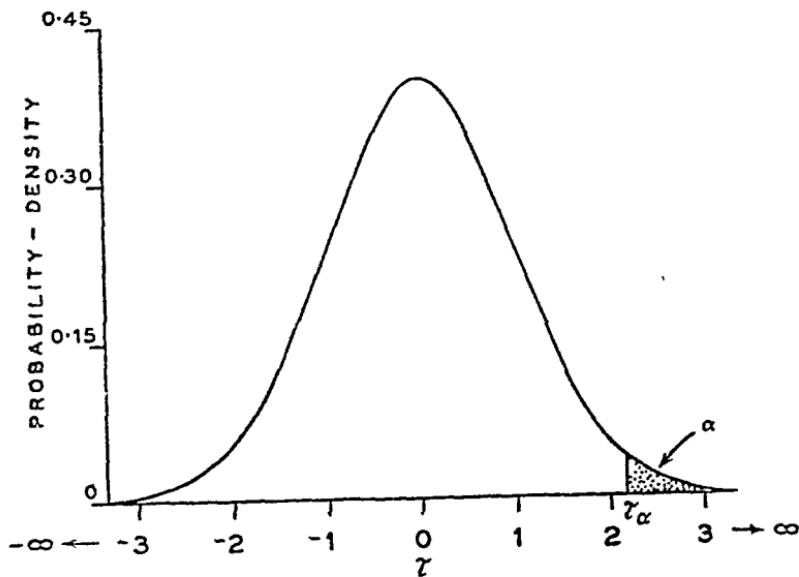


Fig. 14.1 Distribution of normal deviate.

It follows from the following theorem that if x is normally distributed with mean μ and variance σ^2 , then $(x-\mu)/\sigma$ is a normal deviate. Conversely, if $(x-\mu)/\sigma$ is a normal deviate, then x is a normal variable with mean μ and variance σ^2 .

Theorem 14.1 If x is normally distributed with mean μ and variance σ^2 , then $y = a + bx$, where $b \neq 0$, is also normally distributed with mean $a + b\mu$ and variance $b^2\sigma^2$.

Proof: Let us denote the p.d.f.s of x and y by $f(x)$ and $g(y)$, respectively.

Assuming $b > 0$, from the result

$$P[c < y < d] = P\left[\frac{c-a}{b} < x < \frac{d-a}{b}\right],$$

we have

$$\int_c^d g(y) dy = \int_{\frac{c-a}{b}}^{\frac{d-a}{b}} f(x) dx = \int_{\frac{c-a}{b}}^{\frac{d-a}{b}} f\left(\frac{y-a}{b}\right) \frac{dx}{dy} dy$$

(on making the transformation $y = a + bx$) If $b < 0$, we similarly have, from

$$P[c < y < d] = P\left[\frac{d-a}{b} < x < \frac{c-a}{b}\right],$$

$$\int_c^d g(y) dy = - \int_{\frac{c-a}{b}}^{\frac{d-a}{b}} f\left(\frac{y-a}{b}\right) \frac{dx}{dy} dy$$

Combining the two results, we get*

$$g(y) = f\left(\frac{y-a}{b}\right) \left| \frac{dx}{dy} \right|$$

But $\frac{dx}{dy} = 1 / \left(\frac{dy}{dx} \right) = \frac{1}{b}$ and $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2]$

$$\begin{aligned} \text{Hence } g(y) &= \frac{1}{|b|\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{y-a}{b}-\mu\right)^2/2\sigma^2\right] \\ &= \frac{1}{|b|\sigma\sqrt{2\pi}} \exp[-(y-a-b\mu)^2/2b^2\sigma^2], \end{aligned}$$

which proves the theorem

(b) χ^2 distribution

Let y_1, y_2, \dots, y_v be v mutually independent normal deviates. Then the sum of their squares

$$\sum_{i=1}^v y_i^2$$

is called a χ^2 (chi-square) with v degrees of freedom (d.f.) It has the probability-density function

$$f(\chi^2) = \frac{1}{2^{v/2} \Gamma(v/2)} \exp(-\chi^2/2) (\chi^2)^{v/2-1}, \quad (14.19)$$

where $0 < \chi^2 < \infty$

*This indicates why in each case one is to take in the p.d.f. of the new variables $|J|$ rather than J itself.

The p.d.f. of χ^2 , the positive square-root of χ^2 , is immediately found to be

$$\frac{1}{2^{(\nu-2)/2} \Gamma(\nu/2)} \exp(-\chi^2/2) \chi^{\nu-1}, \quad \dots \quad (14.19a)$$

where $0 < \chi^2 < \infty$.

For $\nu \leq 2$ the density (14.19) steadily decreases as χ^2 increases, while for $\nu > 2$ there is a unique maximum at $\chi^2 = \nu - 2$. The distribution is thus always positively skew. The curve of the distribution of χ^2 with 7 d.f. is shown in Fig. 14.2

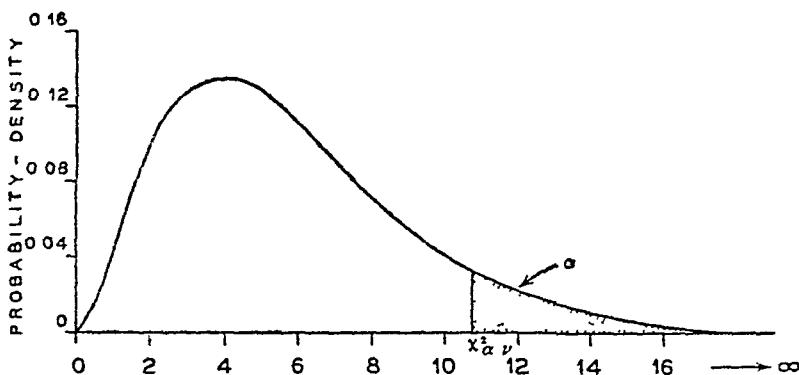


Fig. 14.2 χ^2 distribution with 7 degrees of freedom ($\nu=7$).

Consider, to begin with, the distribution of the square of a single normal deviate, say

$$z = y^2.$$

z varies from 0 to ∞ and for $0 < a < b < \infty$, we have, noting that the transformation from y to z is two-to-one,

$$P[a < z < b] = P[\sqrt{a} < y < \sqrt{b}] + P[-\sqrt{b} < y < -\sqrt{a}].$$

Thus if $g(z)$ be the p.d.f. of z and $f(y)$ that of y , then

$$\begin{aligned} \int_a^b g(z) dz &= \int_{\sqrt{a}}^{\sqrt{b}} f(y) dy + \int_{-\sqrt{b}}^{-\sqrt{a}} f(y) dy \\ &= \int_0^b \left[f(\sqrt{z}) \frac{d}{dz} (\sqrt{z}) - f(-\sqrt{z}) \frac{d}{dz} (-\sqrt{z}) \right] dz \end{aligned}$$

(on putting $y = \sqrt{z}$ and $y = -\sqrt{z}$ in the first and second integrals, respectively).

Hence

$$\begin{aligned}g(z) &= [f(\sqrt{z}) + f(-\sqrt{z})] \left| \frac{df}{dz} \right| \\&= 2 \frac{1}{\sqrt{2\pi}} \exp(-z/2) \frac{1}{2\sqrt{z}} \\&= \frac{1}{2^{1/2}\Gamma(1/2)} z^{-1/2} \exp(-z/2)\end{aligned}$$

Thus the result (14.19) is seen to be true for $v=1$. If it is then assumed to be true for $v=t$, the p.d.f. of $u = \sqrt{\sum_{i=1}^t y_i^2}$ is, from (14.19a),

$$\frac{1}{2^{(t-2)/2}\Gamma(t/2)} \exp(-u^2/2) u^{t-1}, \quad 0 < u < \infty,$$

and that of $v = \sqrt{y_{t+1}^2}$ is

$$\frac{2^{1/2}}{\Gamma(1/2)} \exp(-v^2/2), \quad 0 < v < \infty$$

The joint p.d.f. of u and v is then

$$\frac{1}{2^{(t-3)/2}\Gamma(1/2)\Gamma(t/2)} \exp[-(u^2+v^2)/2] u^{t-1}$$

Now make the one-to-one polar transformation

$$\left. \begin{array}{l} u = u' \cos \theta, \\ v = u' \sin \theta \end{array} \right\} \quad (0 < u' < \infty, \quad 0 < \theta < \pi/2)$$

$$\text{Then } u' = \sqrt{u^2 + v^2} \quad (= \sqrt{\sum_{i=1}^{t+1} y_i^2})$$

Also, the Jacobian of the transformation is

$$\left| \begin{array}{cc} \frac{\partial u}{\partial u'} & \frac{\partial u}{\partial \theta} \\ \frac{\partial v}{\partial u'} & \frac{\partial v}{\partial \theta} \end{array} \right| = u'$$

Hence the joint p.d.f. of u' and θ is

$$\frac{1}{2^{(t-3)/2}\Gamma(1/2)\Gamma(t/2)} \exp(-u'^2/2) (u')^{t-1} \cos^{t-1} \theta, \quad 0 < u' < \infty, \quad 0 < \theta < \pi/2$$

Since $2 \int_0^{\pi/2} \cos^{t-1}\theta d\theta = B(1/2, t/2)$, the p.d.f. of u' is

$$\frac{1}{2^{(t-1)/2} \Gamma\left(\frac{t+1}{2}\right)} \exp(-u'^2/2) (u')^t,$$

whence the p.d.f. of u'^2 comes out to be

$$\frac{1}{2^{(t+1)/2} \Gamma\left(\frac{t+1}{2}\right)} \exp(-u'^2/2) (u'^2)^{(t-1)/2}.$$

The result (14.19) thus holds for $\nu=t+1$ if it is assumed to hold for $\nu=t$. Since it has been already shown to be valid for $\nu=1$, by mathematical induction, it is found to hold for all positive integral values of ν .

An important result regarding the χ^2 distribution is to be noted. Let Y_1 and Y_2 be two independent variables distributed as χ^2 's with ν_1 and ν_2 d.f., respectively. Then the sum $Y_1 + Y_2$ may be shown to be distributed in the same form with $\nu_1 + \nu_2$ d.f. This may be regarded as a consequence of the definition of χ^2 , since $Y_1 + Y_2$ is the sum of squares of $\nu_1 + \nu_2$ mutually independent normal deviates. A direct proof may be given by considering the joint distribution of Y_1 and Y_2 and deriving from it the joint distribution of $Y = Y_1 + Y_2$ and θ , which are such that

$$\sqrt{Y_1} = \sqrt{Y} \cos \theta,$$

$$\sqrt{Y_2} = \sqrt{Y} \sin \theta$$

$$(0 < Y < \infty, 0 < \theta < \pi/2).$$

This property is designated as the additive property of χ^2 's.

For large ν , $\sqrt{2\chi^2}$ can be shown to be approximately normally distributed with mean $\sqrt{2\nu-1}$ and standard deviation 1. This approximation is generally used to calculate values of χ^2 at different probability levels for $\nu > 30$.

We shall denote by $\chi_{\alpha, \nu}^2$ the value of χ^2 (with ν d.f.) for which

$$P[\chi^2 > \chi_{\alpha, \nu}^2] = \alpha. \quad \dots \quad (14.20)$$

$\chi_{\alpha, \nu}^2$ is then the upper α -point of the χ^2 distribution with ν d.f., while the lower α -point is $\chi_{1-\alpha, \nu}^2$.

(c) *t distribution*

If y be a normal deviate and Y a chi-square with v d.f. distributed independently of y , then the new variable

$$\frac{y}{\sqrt{Y/v}}$$

is called a *t* with v degrees of freedom. It has the distribution

$$f(t) = \frac{1}{v^{1/2} B(1/2, v/2)} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad (14.21)$$

where $-\infty < t < \infty$

Let us start from the joint density function of y and Y , which is

$$\frac{1}{\sqrt{2\pi} 2^{v/2} \Gamma(v/2)} Y^{(v-2)/2} \exp(-y^2/2) \exp(-Y/2),$$

$-\infty < y < \infty, 0 < Y < \infty$

Making the one-to-one transformation

$$\begin{aligned} t &= \frac{y}{\sqrt{Y/v}}, \\ u &= Y \end{aligned} \quad \left. \right\} \quad (-\infty < t < \infty, 0 < u < \infty),$$

so that $y = t\sqrt{u/v}$, $Y = u$, and noting that

$$J = \begin{vmatrix} \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \\ \frac{\partial Y}{\partial t} & \frac{\partial Y}{\partial u} \end{vmatrix} = \begin{vmatrix} \sqrt{u/v} & \frac{t}{2\sqrt{uv}} \\ 0 & 1 \end{vmatrix} = \sqrt{u/v},$$

we have the joint p.d.f. of t and u as

$$\frac{1}{v^{1/2} 2^{(v+1)/2} \sqrt{\pi} \Gamma(v/2)} u^{(v-1)/2} \exp\left[-\frac{u}{2}\left(1 + \frac{t^2}{v}\right)\right],$$

$-\infty < t < \infty, 0 < u < \infty$

The p.d.f. of t is, therefore,

$$\begin{aligned} &\frac{1}{v^{1/2} 2^{(v+1)/2} \sqrt{\pi} \Gamma(v/2)} \int_0^\infty u^{(v-1)/2} \exp\left[-\frac{u}{2}\left(1 + \frac{t^2}{v}\right)\right] du \\ &= \frac{\Gamma\left(\frac{v+1}{2}\right) \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}}{v^{1/2} \sqrt{\pi} \Gamma(1/2)} \\ &= \frac{1}{v^{1/2} B(1/2, v/2)} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} \end{aligned}$$

'Like the distribution of the normal deviate, the t distribution is symmetrical about $t=0$, but unlike the normal distribution, it has $\gamma_2 > 0$, i.e. it is more peaked than a normal distribution with the same standard deviation.

The symbol $t_{\alpha, v}$ will be used to denote the value of t (with v d.f.) such that

$$P[t > t_{\alpha, v}] = \alpha. \quad \dots \quad (14.22)$$

Owing to the symmetry of the distribution,

$$t_{1-\alpha, v} = -t_{\alpha, v}. \quad \dots \quad (14.23)$$

For small v the t distribution differs considerably from the distribution of the normal deviate, $t_{\alpha, v}$ being always greater than t_{α} if $0 < \alpha < 1/2$. For large values of v , however, the t distribution tends to the distribution of the normal deviate and $t_{\alpha, v}$ may then be well approximated by t_{α} .

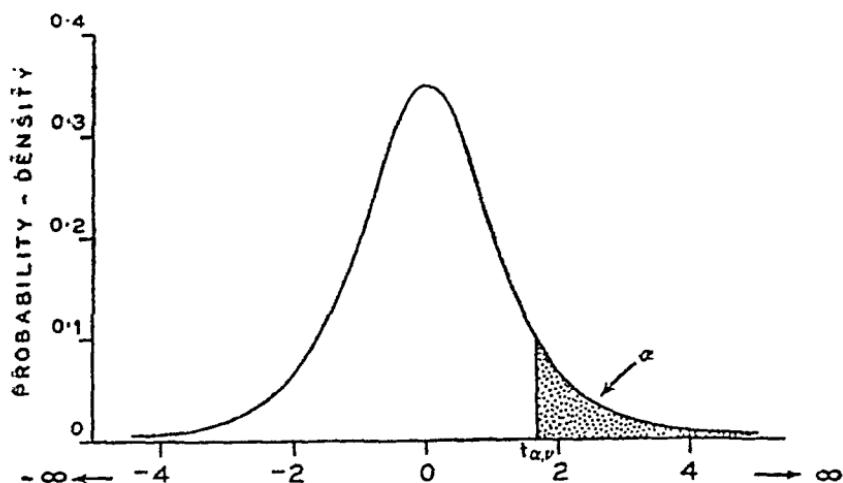


Fig. 14.3 t distribution with 5 degrees of freedom ($v=5$).

(d) F distribution

Let Y_1 and Y_2 be independently distributed as χ^2 s with v_1 and v_2 degrees of freedom, respectively. The random variable

$$\frac{Y_1/v_1}{Y_2/v_2}$$

is then called an F with v_1, v_2 degrees of freedom. This has the

distribution

$$f(F) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2, \nu_2/2)} F^{(\nu_1-2)/2} \left(1 + \frac{\nu_1}{\nu_2} F\right)^{-(\nu_1+\nu_2)/2}, \quad \dots \quad (14.24)$$

where $0 < F < \infty$.

To derive this result, note that the joint p.d.f. of Y_1 and Y_2 is, from (14.19),

$$g(Y_1, Y_2) = \frac{1}{2^{(\nu_1+\nu_2)/2} \Gamma(\nu_1/2) \Gamma(\nu_2/2)} Y_1^{(\nu_1-2)/2} Y_2^{(\nu_2-2)/2} \times \exp\left(-\frac{Y_1 + Y_2}{2}\right),$$

$0 < Y_1 < \infty, 0 < Y_2 < \infty.$

Let us make the one-to-one transformation

$$\begin{aligned} F &= \frac{Y_1/\nu_1}{Y_2/\nu_2}, \\ u &= Y_2 \end{aligned} \quad \left\{ \quad (0 < F < \infty, 0 < u < \infty),$$

so that

$$Y_1 = \frac{\nu_1}{\nu_2} F u$$

and

$$Y_2 = u$$

The Jacobian of the transformation is

$$J = \begin{vmatrix} \frac{\partial Y_1}{\partial F} & \frac{\partial Y_1}{\partial u} \\ \frac{\partial Y_2}{\partial F} & \frac{\partial Y_2}{\partial u} \end{vmatrix} = \begin{vmatrix} \frac{\nu_1 u}{\nu_2} & \frac{\nu_1 F}{\nu_2} \\ 0 & 1 \end{vmatrix} = \frac{\nu_1 u}{\nu_2}$$

Hence the joint p.d.f. of F and u is

$$\begin{aligned} h(F, u) &= \frac{1}{2^{(\nu_1+\nu_2)/2} \Gamma(\nu_1/2) \Gamma(\nu_2/2)} \left(\frac{\nu_1 F u}{\nu_2}\right)^{(\nu_1-2)/2} u^{(\nu_2-2)/2} \\ &\quad \times \exp\left[-\frac{u}{2}\left(1 + \frac{\nu_1}{\nu_2} F\right)\right] \left(\frac{\nu_1 u}{\nu_2}\right) \\ &= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{2^{(\nu_1+\nu_2)/2} \Gamma(\nu_1/2) \Gamma(\nu_2/2)} F^{(\nu_1-2)/2} u^{(\nu_2-2)/2} \\ &\quad \times \exp\left[-\frac{u}{2}\left(1 + \frac{\nu_1}{\nu_2} F\right)\right], \\ &\quad 0 < F < \infty, 0 < u < \infty. \end{aligned}$$

The p.d.f. of F is, therefore,

$$\begin{aligned} f(F) &= \int_0^{\infty} h(F, u) du \\ &= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \cdot F^{(\nu_1-2)/2} \cdot \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}} \\ &= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2, \nu_2/2)} \cdot \frac{F^{(\nu_1-2)/2}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, \quad 0 < F < \infty. \end{aligned}$$

This distribution is highly positively skew. It is easily seen from the definitions of t and F that an F with $\nu_1=1$ is a t^2 , t having ν_2 d.f.

As in the previous cases, we shall denote by $F_{\alpha; \nu_1, \nu_2}$ the upper α -point of the F distribution with (ν_1, ν_2) d.f.; i.e.,

$$P[F > F_{\alpha; \nu_1, \nu_2}] = \alpha. \quad \dots \quad (14.25)$$

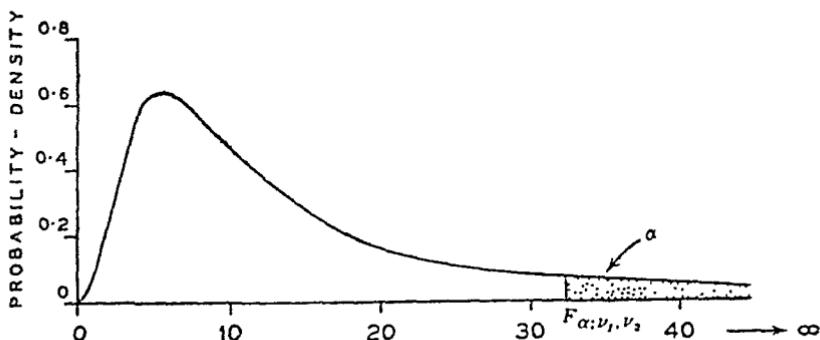


Fig. 14.4 F distribution with 10, 4 degrees of freedom
($\nu_1=10, \nu_2=4$).

As regards the lower α -point $F_{1-\alpha; \nu_1, \nu_2}$, we see that

$$P[F < F_{1-\alpha; \nu_1, \nu_2}] = \alpha$$

or $P\left[\frac{1}{F} > \frac{1}{F_{1-\alpha; \nu_1, \nu_2}}\right] = \alpha.$

Now $\frac{1}{F}$, which is of the form $\frac{Y_2/v_2}{Y_1/v_1}$, is itself distributed as an F with (v_2, v_1) d.f. It follows that

$$F_{1-\alpha} \cdot \frac{1}{v_1 \cdot v_2} = F_{\alpha} \cdot v_2 \cdot v_1$$

or $F_{1-\alpha} \cdot v_1 \cdot v_2 = F_{\alpha} \cdot \frac{1}{v_2 \cdot v_1}$ (14.26)

It is, therefore, unnecessary to tabulate the lower α -points of F distributions with various d.f.s, once the upper α points are tabulated.

14.7 Sampling distributions of mean and variance in sampling from a normal population

Let x_1, x_2, \dots, x_n be independent random observations from a normal population whose p.d.f. is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2] \quad -\infty < x < \infty$$

We shall denote the sample mean and the sample variance of x by \bar{x} and s^2 , respectively. Thus

$$\bar{x} = \sum_i x_i / n$$

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

In order to obtain the sampling distributions of \bar{x} and s^2 , we start from the joint p.d.f. of x_1, x_2, \dots, x_n , which is

$$\prod_i f(x_i) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp[-\sum_i (x_i - \mu)^2/2\sigma^2]$$

We make the following one-to-one transformation from x_i ($i=1, 2, \dots, n$) to y_i ($i=1, 2, \dots, n$)

$$\left. \begin{aligned} y_1 &= \frac{(x_1 - \mu)/\sigma + (x_2 - \mu)/\sigma + \dots + (x_n - \mu)/\sigma}{\sqrt{n}}, \\ y_i &= a_{i1} \frac{(x_1 - \mu)}{\sigma} + a_{i2} \frac{(x_2 - \mu)}{\sigma} + \dots + a_{in} \frac{(x_n - \mu)}{\sigma}, \end{aligned} \right\} \quad (14.27)$$

for $i=2, 3, \dots, n$,

where the $(n-1)$ vectors $(a_{i1}, a_{i2}, \dots, a_{in})$ are of unit length, mutually orthogonal and each orthogonal to the vector

$$\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right)$$

One such set of vectors is :

$$\begin{aligned} & \frac{1}{\sqrt{2}}(1, -1, 0, 0, \dots, 0, 0), \\ & \frac{1}{\sqrt{6}}(1, 1, -2, 0, \dots, 0, 0), \\ & \quad \vdots \\ & \frac{1}{\sqrt{n(n-1)}}(1, 1, 1, \dots, 1, -(n-1)). \end{aligned}$$

The Jacobian of the transformation is then J , such that

$$\frac{1}{J} = \left| \begin{array}{cccc} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \dots & \frac{\partial y_n}{\partial x_n} \end{array} \right| = \frac{1}{\sigma^n}, \quad \dots \quad (14.28)$$

implying that $J = \sigma^n$.

Further,

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2.$$

Hence the joint p.d.f. of y_1, y_2, \dots, y_n is

$$\frac{1}{(\sqrt{2\pi})^n} \exp[-\sum_i y_i^2/2].$$

This shows that y_1, y_2, \dots, y_n are independently and identically distributed, each being a normal deviate.

But

$$y_1 = \frac{1}{\sigma \sqrt{n}} \sum_i (x_i - \mu) = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma},$$

a linear function of \bar{x} . Since y_1 is a normal deviate, \bar{x} must be a normal variable with mean μ and variance $\frac{\sigma^2}{n}$ (from Theorem 14.1). Thus the p.d.f. of \bar{x} is

$$g(\bar{x}) = \frac{\sqrt{n}}{\sigma \sqrt{2\pi}} \exp[-n(\bar{x} - \mu)^2/2\sigma^2],$$

$$-\infty < \bar{x} < \infty. \quad \dots \quad (14.29)$$

Again,

$$\begin{aligned}\sum_{i=2}^n y_i^2 &= \sum_{i=1}^n y_i^2 - y_1^2 \\ &= \frac{1}{\sigma^2} \left[\sum_{i=1}^n (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \right] \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{(n-1)s'^2}{\sigma^2}. \quad \dots \quad (14.30)\end{aligned}$$

Now $\sum_{i=2}^n y_i^2$, being the sum of squares of $(n-1)$ independent normal deviates, is a χ^2 with $(n-1)$ d.f., and this is distributed independently of y_1 . It follows that $(n-1)s'^2/\sigma^2$ is distributed as a χ^2 with $(n-1)$ d.f. and is independent of \bar{x} . From (14.19) the p.d.f. of s'^2 is, therefore, obtained as

$$h(s'^2) = \frac{(n-1)^{(n-1)/2}}{(2\sigma^2)^{(n-1)/2} \Gamma\left(\frac{n-1}{2}\right)} \exp[-(n-1)s'^2/2\sigma^2] (s'^2)^{(n-3)/2}, \quad 0 < s'^2 < \infty. \quad \dots \quad (14.31)$$

Questions and exercises

14.1 Define simple random sampling. Describe some practical methods of drawing a random sample from a finite population.

14.2 Explain the terms 'parameter', 'statistic', 'sampling distribution' and 'standard error of a statistic'.

14.3 Obtain the expectation and standard error of sample mean for a random sample of size n drawn from a population of size N (a) with replacements, (b) without replacements.

14.4 Show that for a random sample of size 100, drawn with replacements, the standard error of sample proportion cannot exceed 0.05.

14.5 Suppose that a statistic T is normally distributed with mean θ . Show that it is very unlikely that the percentage error in estimating θ by T will be greater than 3 times the coefficient of variation of T . [Note that in this case the coefficient of variation is $100\sigma_T/\theta$.]

14.6 Identify the distributions of χ^2 , t and F as members of the Pearsonian family.

14.7 Starting from the density function of a normal deviate, obtain that of a χ^2 with v d.f.

14.8 Starting from the definition of the t -statistic with v d.f., obtain its density function.

14.9 Define an F with v_1, v_2 d.f. Hence obtain its density function.

14.10 Show that \bar{x} and s'^2 for random samples of size n from a normal population are distributed independently of each other. Also, show that \bar{x} is distributed normally with mean μ and variance σ^2/n , while $\frac{(n-1)s'^2}{\sigma^2}$ is distributed as a χ^2 with $(n-1)$ d.f.

14.11 Proceeding as in Section 14.7, show that if x_i ($i=1, \dots, n$) are distributed independently and normally with means μ_i ($i=1, 2, \dots, n$) and variances σ_i^2 ($i=1, 2, \dots, n$), then the linear function $a + \sum_i b_i x_i$ (where at least one b is non-zero) is normally distributed with mean $a + \sum_i b_i \mu_i$ and variance $\sum_i b_i^2 \sigma_i^2$.

14.12 If x has the exponential distribution with p.d.f.

$$f(x) = \theta \exp(-\theta x), \quad 0 < x < \infty,$$

where $\theta > 0$, then what is the distribution of $\sum_i x_i$, x_i ($i=1, \dots, n$) being random and independent observations from this distribution?

14.13 If x is distributed in the rectangular form with p.d.f.

$$f(x) = \frac{1}{\theta}, \quad 0 < x < \theta,$$

show that $-2 \log_e(x/\theta)$ is a χ^2 with 2 d.f.

14.14 If Y_1 and Y_2 are independent χ^2 's with v_1 and v_2 d.f., find the joint distribution of $Y_1 + Y_2$ and Y_1/Y_2 . Show that $Y_1 + Y_2$ is itself a χ^2 , while Y_1/Y_2 is of the form $\frac{v_1}{v_2} F$, and that these two are also mutually independent.

[Hint : Make the polar transformation :

$$\sqrt{Y_1} = \sqrt{Y} \cos \theta,$$

$$\sqrt{Y_2} = \sqrt{Y} \sin \theta.$$

Note that $Y = Y_1 + Y_2$ and $\cot^2 \theta = Y_1/Y_2$.]

14.15 Denoting by λ_λ a Poisson variable with parameter λ and by χ^2_k a χ^2 variable with k d.f., establish the following identity, for all positive integers k

$$P[\lambda_\lambda \leq k-1] = P[\chi^2_k > 2\lambda]$$

14.16 Denoting by $\Lambda_{n,p}$ a binomial variable with parameters (n, p) and by F_{v_1, v_2} an F statistic with v_1 and v_2 degrees of freedom, establish the following identity

$$P[\Lambda_{n,p} \leq k-1] = P[F_{2k-n-(k-1)} > \frac{n-k+1}{k} \frac{p}{1-p}]$$

14.17 Let x_i ($i=1, 2, \dots, n$) be a random sample from a continuous distribution with p.d.f. $f(x)$. If y and z be the smallest and the largest of the observations, show that y has the p.d.f.

$$n \left[\int_y^\infty f(x) dx \right]^{n-1} f(y),$$

while z has the p.d.f.

$$n \left[\int_{-\infty}^z f(x) dx \right]^{n-1} f(z)$$

14.18 (Continuation) Show that in random sampling from the exponential distribution $f(x)$ of Exercise 14.12, y and z also have exponential marginal distributions.

SUGGESTED READING

- [1] Goon, A M, Gupta, M K and Dasgupta, B *An Outline of Statistical Theory* (Ch 10) World Press, 1970
- [2] Hogg, R V and Craig, A T *Introduction to Mathematical Statistics* (Chs 3, 4) Macmillan, 1965, and Amerind
- [3] Keeping, E S *Introduction to Statistical Inference* (Ch 8) Van Nostrand, 1962, and Affiliated East West Press
- [4] Mood, A M and Graybill, F A *Introduction to the Theory of Statistics* (Ch 10) McGraw Hill, 1963, and Kogakusha
- [5] Yule, G U and Kendall, M G *An Introduction to the Theory of Statistics* (Ch 14) Charles Griffin, 1953

15

BASIC PRINCIPLES OF STATISTICAL INFERENCE

15.1 Estimation and testing of hypotheses

Since a sample is but a part of a population, the features of the former will generally differ from those of the latter. The question that naturally arises is then : what can be said about the properties of the population from a knowledge of the properties of the sample ? Although an answer to this question may not be found in all cases, in the case of random sampling this can be answered with the help of probability theory. In sampling theory, we are primarily concerned with this very question. The process of going from the known sample to the unknown population has been called *statistical inference*.

The basic problem of sampling theory usually presents itself in one of two forms : (1) Some feature of the population in which an enquirer is interested may be completely unknown to him, and he may want to make a guess about this feature completely on the basis of a random sample from the population. (2) Some information as to the feature of the population may be available to the enquirer, and he may want to see whether the information is tenable in the light of the random sample taken from the population. The first type of problem is called the *problem of estimation* and the second the *problem of testing of hypotheses*.

We shall assume in this chapter and in Chapter 16 that the form of the population (binomial, normal, etc.) is either known or is not of importance to the enquirer in the particular context, in which case he will be interested in some unknown parameter or parameters of the population. The usual problem is then to estimate the unknown parameters or to test some hypotheses regarding these parameters on the basis of the given sample. In a later chapter, we shall also consider the problem of testing hypotheses regarding the form of a population.

15.2 Point estimation of parameters

Let θ be an unknown parameter of the distribution of a variable x . For estimating θ on the basis of a random sample, x_1, x_2, \dots, x_n ,

we may use the statistic T . Then T is the *estimator* of θ , and the value of T obtained from a *given* sample is its *estimate*. Clearly, for T to be a good estimator, the difference $|T - \theta|$ should be as small as possible. However, since T is itself a random variable, all that we can hope to ensure is that the difference be small with a high probability.

Unbiasedness and minimum variance

One way of achieving this would be to see that the sampling distribution of T has a central tendency towards θ and a small dispersion. If we agree to accept the mean as the proper measure of central tendency and the variance as the proper measure of dispersion, then we would want that T should be *unbiased**¹, i.e.,

$$E(T) = \theta, \quad \text{whatever the true value of} \\ \theta \text{ may be,} \quad (15.1)$$

and that, among all unbiased estimators, T should have the smallest variance, i.e.

$$\text{var}(T) \leq \text{var}(T'), \quad \text{whatever the true value of} \\ \theta \text{ may be} \quad (15.2)$$

where T' is any other unbiased estimator.

A statistic T of this type is called a *minimum-variance unbiased estimator* of θ .

Ex 15.1 It has been shown (in Section 14.4) that if we consider Bernoullian trials with probability of success p for each trial and if f be the number of successes in n such trials, then

$$E(f/n) = p,$$

whatever the true value of p . Hence f/n is an unbiased estimator of p . It can also be shown that, among all unbiased estimators, f/n has the smallest variance. Hence f/n is a *minimum variance unbiased estimator of p* .

Ex 15.2 If x_1, x_2, \dots, x_n be a random sample from a population with mean μ and if x be the sample mean, then

$$E(x) = \mu,$$

whatever the true value of μ , so that x is an unbiased estimator of μ (*vide* Section 14.3).

*If $E(T) = \theta + b(\theta)$, then $b(\theta)$ is the bias of T .

Suppose further that the observations are not only random but also independent and that the population is *normal* with mean μ and variance σ^2 . Here it can be shown that \bar{x} has the least variance among all unbiased estimators of μ ; i.e., \bar{x} is a minimum-variance unbiased estimator of μ .

Consistency and efficiency

An alternative approach would be to demand that the estimator should behave more and more satisfactorily as the sample size n becomes larger and larger. In particular, it may be required that the values of T , which purports to be a good estimator, should be more and more clustered around θ with increasing sample size. To put it in probabilistic terms, it may be required that the statistic T should *converge stochastically* (or *in probability*) to θ as $n \rightarrow \infty$. In other words, given two positive quantities, ϵ and η , however small, it should be possible to find an n_0 , depending on ϵ and η , such that

$$P[|T - \theta| \leq \epsilon] > 1 - \eta \quad \dots \quad (15.3)$$

whenever $n \geq n_0$. A statistic T with this property is called a *consistent estimator* of θ .

It can be shown that a set of sufficient conditions for T to be consistent are that

$$E(T) \rightarrow \theta \quad \dots \quad (15.4a)$$

$$\text{and} \quad \text{var}(T) \rightarrow 0 \quad \dots \quad (15.4b)$$

as $n \rightarrow \infty$.

There may be found a large number of consistent estimators for θ . Indeed, if T be consistent, so are, e.g., $T + \frac{a}{\psi(n)}$ and $T\left\{1 + \frac{a}{\psi(n)}\right\}$, where a is any constant independent of n and $\psi(n)$ is any increasing function of n .

To choose among these rival estimators, some additional criterion would be needed. Thus we may consider, together with stochastic convergence, the rate of stochastic convergence; i.e., we may demand not only that T should converge stochastically to θ but that it should do so sufficiently rapidly. We shall confine our attention to consistent estimators that are asymptotically normally distributed (*vide Chapter 17*). In that case, the rapidity of convergence will be indicated by the inverse of the variance of the asymptotic distribution.

bution Denoting the asymptotic variance by 'avar' we may then say that T is the best estimator of θ if it is consistent and normally distributed and if

$$\text{avar}(T) \leq \text{avar}(T')$$

whatever the other consistent and asymptotically normal estimator T' may be

A consistent, asymptotically normal statistic T having this property is called *efficient*

Ex. 15.3 Consider the proportion of successes, f/n , for a set of n Bernoullian trials with probability of success p . From Corollary 3.15.2, it follows that f/n is a consistent estimator of p .

Further, the fact that f has the binomial distribution, with parameters n and p , means that f is asymptotically normally distributed with mean np and variance $np(1-p)$. Hence f/n is seen to be asymptotically normally distributed with expectation p and variance $p(1-p)/n$. Since this can also be shown to be the smallest asymptotic variance for an asymptotically normally distributed consistent estimator of p , f/n is also efficient.

Ex 15.4 Let x_1, x_2, \dots, x_n be independent random observations from a normal population with mean μ and variance σ^2 . If σ^2 is finite, it follows from Corollary 3.15.1 that the sample mean \bar{x} is a consistent estimator of μ .

Now the sampling distribution of \bar{x} is exactly normal with mean μ and (exact) variance σ^2/n . Also, σ^2/n can be shown to be the smallest asymptotic variance for an asymptotically normally distributed estimator of μ . Hence \bar{x} is also an efficient estimator of μ .

On the contrary, the sample median x_{m1} has*

$$E(x_{m1}) \underset{\sim}{\approx} \mu \quad (15.5a)$$

and

$$\text{var}(x_{m1}) \underset{\sim}{\approx} \frac{\pi\sigma^2}{2n} \quad (15.5b)$$

Since $E(x_{m1}) \rightarrow \mu$ and $\text{var}(x_{m1}) \rightarrow 0$ as $n \rightarrow \infty$, the sample median is a consistent estimator of μ . Like \bar{x} , x_{m1} is also asymptotically normal. But since x_{m1} has asymptotic variance $\pi\sigma^2/2n$, which is greater than σ^2/n , it is an inefficient estimator.

*The symbol ' $\underset{\sim}{\approx}$ ' denotes asymptotic equality

Sufficiency

The criteria of consistency and efficiency for a good estimator have been suggested by R. A. Fisher.

Now, a preliminary choice among statistics for the purpose of estimating θ , before looking for a minimum-variance unbiased or an efficient consistent estimator, can be made on the basis of another criterion suggested by Fisher. This is the criterion of *sufficiency*. A statistic T is called sufficient for θ (or, rather, for the family of distributions characterised by θ) if the conditional distribution of any other statistic for given T is independent of θ . Obviously, if T is of this type, then any inference regarding θ can be made on the basis of T alone, instead of starting with all n observations. In other words, T provides a method of summarising the information regarding θ contained in the whole sample into a single statistic.

A necessary and sufficient condition for T to be sufficient for θ is that the joint probability-density function or the joint probability-mass function of x_1, x_2, \dots, x_n should be of the form :

$$f(x_1, x_2, \dots, x_n | \theta) = g(T | \theta) \cdot h(x_1, x_2, \dots, x_n), \quad \dots \quad (15.6)$$

where the first part of the right-hand side depends on T and θ , while the second part is independent of θ . This provides a simple method of judging whether T is really a sufficient statistic.

Ex. 15.5 Consider a set of n Bernoullian trials with probability of success p . With the i th trial we may associate a variable x_i having the probability-mass function

$$f(x_i | p) = p^{x_i} (1-p)^{1-x_i}, \text{ for } x_i = 0, 1.$$

The joint probability-mass function of x_1, x_2, \dots, x_n is

$$\begin{aligned} f(x_1, x_2, \dots, x_n | p) &= p^{\sum x_i} (1-p)^{n-\sum x_i}, \\ &= g(\sum x_i | p) \cdot h(x_1, x_2, \dots, x_n), \end{aligned}$$

$$\text{where } g(\sum x_i | p) = p^{\sum x_i} (1-p)^{n-\sum x_i}.$$

$$\text{and } h(x_1, x_2, \dots, x_n) = 1.$$

Hence $\sum x_i$, the number of successes in the n trials taken together, is a sufficient statistic for p . So is $p = \sum_{i=1}^n x_i / n$.

Ex 15.6 Let x_1, x_2, \dots, x_n be as in Ex 15.4. Suppose further that μ is unknown but σ^2 is known. Then the joint density function of x_1, x_2, \dots, x_n is

$$\begin{aligned}f(x_1, x_2, \dots, x_n | \mu) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2\right] \\&= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{n}{2\sigma^2}(x - \mu)^2\right] \exp\left[-\frac{1}{2\sigma^2} \sum_i (x_i - \bar{x})^2\right] \\&= g(x | \mu) h(x_1, x_2, \dots, x_n) \text{ say}\end{aligned}$$

Hence x , as well as $nx = \sum_i x_i$, is a sufficient statistic for μ .

15.3 Maximum-likelihood estimation

There is a simple method of obtaining estimators with desirable properties. This is the method of maximum likelihood.

Consider $f(x_1, x_2, \dots, x_n | \theta)$, the joint probability-density or probability-mass of the sample observations. For fixed θ , it may be looked upon as a function of the sample observations and then it gives their probability density function or probability mass function. But, when x_1, x_2, \dots, x_n are given, it may also be looked upon as a function of θ , called the *likelihood function of θ* and denoted by $L(\theta)$. The principle of maximum likelihood consists in taking that value as the estimator of θ for which $L(\theta)$ is a maximum. Thus if $\hat{\theta}$ be the *maximum-likelihood estimator of θ* , then by definition

$$L(\hat{\theta}) = \max_{\theta} L(\theta) \quad (15.7)$$

In many cases, it will be convenient to deal with $\log L(\theta)$, rather than $L(\theta)$, and since $\log L(\theta)$ attains its highest value for the same value of θ as $L(\theta)$ does, θ is such that

$$\log L(\hat{\theta}) = \max_{\theta} \log L(\theta) \quad (15.8)$$

This $\hat{\theta}$, again, will in many cases be obtainable by differentiating $\log L(\theta)$, i.e. will be that value of θ for which

$$\frac{d \log L(\theta)}{d \theta} = 0 \quad (15.9)$$

But one must make sure that the value obtained by solving (15.9), which gives a local maximum of $L(\theta)$, also gives the absolute (global) maximum. Indeed, the derivative may not exist at $\theta = \hat{\theta}$, and then this method will fail.

Ex. 15.7 With x_1, x_2, \dots, x_n the same as in Ex. 15.5, let us obtain the maximum-likelihood estimate of p . Here the likelihood function is

$$L(p) = p^{\sum x_i} (1-p)^{n-\sum x_i}$$

and $\log_e L(p) = (\sum x_i) \log_e p + (n - \sum x_i) \log_e (1-p).$

Hence

$$\frac{d \log_e L(p)}{dp} = \frac{(1-p) \sum x_i - p(n - \sum x_i)}{p(1-p)} = \frac{\sum x_i - np}{p(1-p)},$$

which equals zero for $p = \sum x_i/n$, provided $0 < \sum x_i < n$. When $\sum x_i = 0$, it is directly found that $\log_e L(p)$ or $L(p)$ is highest for the smallest value of p , i.e. for $p=0$. Similarly, when $\sum x_i = n$, $L(p)$ is highest for the highest value of p , i.e. for $p=1$. Hence, whatever the value of $\sum x_i$, the maximum-likelihood estimate is

$$\hat{p} = \frac{\sum x_i}{n} = f/n,$$

the sample proportion of successes.

Ex. 15.8 Let x_1, x_2, \dots, x_n be a random sample of independent observations from a Poisson population with parameter λ . Then the likelihood function is

$$L(\lambda) = \frac{\exp[-n\lambda] \cdot \lambda^{\sum x_i}}{\prod_i (x_i!)}$$

and $\log_e L(\lambda) = -n\lambda + (\sum x_i) \log_e \lambda - \sum_i \log_e (x_i!).$

Hence

$$\frac{d \log_e L(\lambda)}{d\lambda} = -n + (\sum x_i) \frac{1}{\lambda},$$

which equals zero if, and only if, $\lambda = \sum x_i/n$, provided $\sum x_i > 0$. In case $\sum x_i = 0$, we find directly that $\log_e L(\lambda)$ or $L(\lambda)$ becomes a maximum when λ takes its least possible value, i.e. when $\lambda=0$. Thus, whatever the value of $\sum x_i$, the maximum-likelihood estimate of λ is

$$\hat{\lambda} = \sum x_i/n,$$

the sample mean

Ex 15.9 Suppose x_1, x_2, \dots, x_n are independent random observations from a normal distribution with mean μ and variance σ^2 .

Case 1 μ unknown, σ known ($=\sigma_0$)

Here

$$L(\mu) = \frac{1}{(\sigma_0 \sqrt{2\pi})^n} \exp\left[-\sum_i (x_i - \mu)^2 / 2\sigma_0^2\right]$$

Proceeding as in Ex 15.7 and Ex 15.8, we have

$$\hat{\mu} = x$$

Case 2 μ known ($=\mu_0$), σ unknown

Here

$$L(\sigma) = \frac{1}{(\sigma \sqrt{2\pi})^n} \exp\left[-\sum_i (x_i - \mu_0)^2 / 2\sigma^2\right],$$

and, proceeding as before, we get

$$\sigma = \sqrt{\sum_i (x_i - \mu_0)^2 / n}$$

Case 3 Both μ and σ unknown

Since both parameters are unknown, the likelihood function here is

$$L(\mu, \sigma) = \frac{1}{(\sigma \sqrt{2\pi})^n} \exp\left[-\sum_i (x_i - \mu)^2 / 2\sigma^2\right]$$

Now,

$$\log_e L(\mu, \sigma) = -\frac{n}{2} \log_e (2\pi) - n \log_e \sigma - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2},$$

and

$$\frac{\partial \log_e L(\mu, \sigma)}{\partial \mu} = \frac{\sum_i (x_i - \mu)}{\sigma^2},$$

$$\frac{\partial \log_e L(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_i (x_i - \mu)^2}{\sigma^3}$$

The maximum likelihood estimates of μ and σ will be obtained by solving the simultaneous equations

$$\frac{\partial \log_e L(\mu, \sigma)}{\partial \mu} = 0, \quad \frac{\partial \log_e L(\mu, \sigma)}{\partial \sigma} = 0$$

We thus have

$$\mu = x$$

and

$$\sigma = \sqrt{\sum_i (x_i - x)^2 / n}$$

Apart from their intuitive appeal, maximum-likelihood estimators possess several nice properties :

(1) Consistency : The maximum-likelihood estimator $\hat{\theta}$ of a parameter θ is, under very general conditions, a consistent estimator.

(2) Asymptotic normality : It has also been found that, under general conditions, $\hat{\theta}$ is asymptotically normally distributed with mean θ .

(3) Efficiency : Among all asymptotically normal consistent estimators of θ , $\hat{\theta}$ has generally the smallest asymptotic variance. Hence $\hat{\theta}$ is generally efficient.

(4) Sufficiency : If there at all exists a sufficient statistic for θ , then $\hat{\theta}$ is also sufficient or is a function of a sufficient statistic.

(5) Unbiasedness : Generally, $\hat{\theta}$ will *not* be an unbiased estimator, but a simple modification will in most cases make it unbiased. In Ex. 15.9 (Case 3), for instance,

$$\hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n} = s^2$$

is not an unbiased estimator of σ^2 , but $\frac{n\hat{\sigma}^2}{n-1} = s'^2$ is.

(6) Invariance : A maximum-likelihood estimator of θ possesses the very desirable property of invariance. Thus if $\hat{\theta}$ is the maximum-likelihood estimator of θ , then $\psi(\hat{\theta})$ is also the maximum-likelihood estimator of $\psi(\theta)$, ψ being a single-valued function of θ with a unique inverse.

15.4 Interval estimation of parameters

Estimation of a parameter by a single value, as in the above section, is referred to as *point estimation*. An alternative procedure is to give an interval within which the parameter may be supposed to lie. This is called *interval estimation*. This may also be illustrated with the help of a variable x which has a normal distribution in the population with mean μ (unknown) and standard deviation σ (known). Let x_1, x_2, \dots, x_n be the values of x in a random sample of size n from this population, the observations being mutually independent.

Now, it is known that *any linear function of normal variables is itself normally distributed*. The sample mean \bar{x} , being a linear function of normal variables x_1, x_2, \dots, x_n , is normally distributed and, as has already been shown, it has mean μ and variance σ^2/n .

Hence $\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}$

is a normal deviate It follows that

$$P\left[-2576 \leq \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \leq 2576\right] = 0.99$$

or $P\left[x - 2576 \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + 2576 \frac{\sigma}{\sqrt{n}}\right] = 0.99$

The latter relation shows that in repeated sampling it is very likely, the probability being 0.99, that the interval $\left(x - 2576 \frac{\sigma}{\sqrt{n}}, x + 2576 \frac{\sigma}{\sqrt{n}}\right)$ will include μ . In other words, if a very large number of samples, each of size n , are taken from the population and if for each such sample the above interval is determined, then in about 99% of the cases the interval will include μ , while in the remaining 1% it will fail to do so (*vide* Section 3.7) One will, therefore, be justified in saying, on the basis of a given sample, that μ lies between $x - 2576 \frac{\sigma}{\sqrt{n}}$ and $x + 2576 \frac{\sigma}{\sqrt{n}}$, the limits being computed from the observations in hand These are called 99% *confidence limits* to μ , 0.99 being the *confidence coefficient*—a sort of measure of the trust or confidence that one may place in these limits for actually including μ .

The choice of the confidence coefficient in any particular case will depend on the discretion of the experimenter himself Naturally, a value close to unity is selected The general symbol for denoting a confidence coefficient is $1-\alpha$

We shall now deal with a more general set-up Let θ be a parameter and T a statistic based on a random sample of size n from the corresponding population We shall suppose that T is a sufficient statistic

Now, in many cases it will be possible to find a function, say $\psi(T, \theta)$, whose distribution is independent of θ The statement

$\psi_{1-\alpha/2} \leq \psi(T, \theta) \leq \psi_{\alpha/2}$, where $\psi_{1-\alpha/2}$ and $\psi_{\alpha/2}$ are the lower and upper $\alpha/2$ -points of the distribution of $\psi(T, \theta)$,

can often be written in an equivalent form as, say,

$$\theta_1(T) \leq \theta \leq \theta_2(T)$$

Hence $P[\theta_1(T) \leq \theta \leq \theta_2(T)]$

$$= P[\psi_{1-\alpha/2} \leq \psi(T, \theta) \leq \psi_{\alpha/2}] = 1 - \alpha, \quad \dots \quad (15.10)$$

whatever the true value θ may be. $\theta_1(T)$ and $\theta_2(T)$ will then be confidence limits to θ with confidence coefficient $1 - \alpha$.

In the above example, $\psi(\bar{x}, \mu) = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$, which is distributed as a normal deviate and hence independently of μ . The confidence limits discussed in Chapter 16 also belong to this category.

15.5 Test of significance

Suppose a variable x is known to be normally distributed in a given population, with a *known* variance σ^2 but with an *unknown* mean μ . Also, suppose it is suggested to us that the mean may be equal to a specified value, say μ_0 , and we want to see how acceptable this suggestion is. We have then the hypothesis

$$H_0 : \mu = \mu_0,$$

which needs to be verified. Such a hypothesis is called a *null hypothesis*, because it states that there is *no* difference between μ and μ_0 . The verification (or test) of H_0 has to be done on the basis of a random sample from this population. Let x_1, x_2, \dots, x_n be the values of x for a random sample of size n , the observations being independent.

In order to test H_0 , let us assume, to begin with, that it is true—that, in fact, the population mean of x is μ_0 . From this assumption a number of results will follow. The most important result for our purpose is that \bar{x} is, according to the assumption, normally distributed with mean μ_0 and variance σ^2/n —in other words, $\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$ is a normal deviate, τ . As such,

$$P\left[\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma} > 2.576\right] = 0.01.$$

To put it in a different way, in repeated sampling from this population, in only one in hundred samples is the value of $\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$ expected to exceed 2.576 numerically. This fact then provides a test for the hypothesis. If in a given sample $\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma}$ exceeds 2.576.

then it means that a value has been obtained which is very improbable under the hypothesis. In such a case the hypothesis itself will be held in suspicion. We say H_0 is *rejected*. On the other hand, if in the given sample $\frac{\sqrt{n}|x - \mu_0|}{\sigma}$ does not exceed 2.576, i.e. if it takes a value which is not improbable under the hypothesis, one would find no reason to suspect the hypothesis. It would then be said to be *accepted*.

By acceptance of a hypothesis we do not mean that it is proved to be true. All that is implied is that, so far as the given sample is concerned, we find no reason to question the validity of the hypothesis. Nor does rejection of H_0 mean a disproof of H_0 . It means simply that, in the light of the given sample, H_0 does not seem to be a plausible hypothesis.

The mode of argument may be restated as follows. Some difference between the sample mean \bar{x} and the hypothetical population mean μ_0 is to be expected because of the inevitable sampling fluctuations. However, if this difference be too large, say greater than $2.576\sigma/\sqrt{n}$ —in other words, if $\frac{\sqrt{n}|x - \mu_0|}{\sigma} > 2.576$ —then one would say

that it may not be due to sampling fluctuations alone but arises because the true population mean is not μ_0 . One would thus take it as significant or indicative of the falsity of the hypothesis.

Hence a test of this kind is also called a *test of significance*. The probability 0.01, on the basis of which the differences are being regarded as significant of the falsity of the hypothesis or not, is called the *level of significance*. The choice of the level of significance, of course, depends on the experimenter himself. If he thinks that rejection of the hypothesis when actually it is true will be a serious error, he will choose a rather small value, say 0.01 or 0.001. On the other hand, if he thinks that this error is not so serious, he will not mind taking a value as high as, say, 0.05 or 0.1. The general symbol for the level of significance is α *.

The above test procedure will be appropriate when we are interested in knowing whether μ is or is *not equal* to μ_0 , i.e. when we

*It is customary in common statistical work to take $\alpha=0.05$ or 0.01

want to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypotheses $H : \mu \neq \mu_0$.

In some cases, however, we may want to know whether μ is equal to μ_0 or greater. H_0 is then to be tested against the alternative hypotheses $H : \mu > \mu_0$.

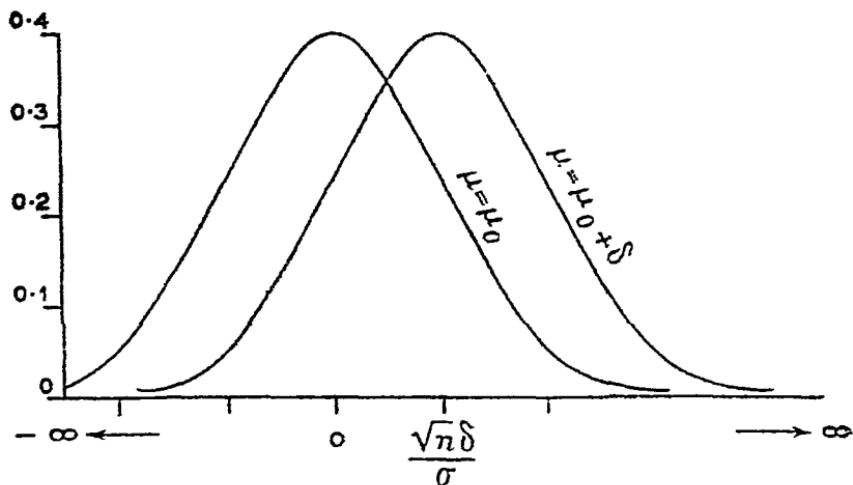


Fig. 15.1 Distribution of $\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$ for $\mu = \mu_0$ and for $\mu = \mu_0 + \delta$.

Here, too, very small values of the statistic $\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$, as well as very large values, are to be regarded as unlikely when $\mu = \mu_0$. But then very small values of the statistic are still more unlikely for a value of μ greater than μ_0 (as will be apparent from Fig. 15.1). And since here we are concerned with a choice between μ_0 and values of μ greater than μ_0 , very small values of the statistic should lead to the acceptance of H_0 rather than to its rejection. On the other hand, very large values of the statistic, being unlikely when $\mu = \mu_0$ but not so unlikely for $\mu > \mu_0$, should lead to the rejection of H_0 . If the level of significance be 0.01, H_0 is, therefore, to be rejected in such a situation if for a given sample it is found that

$$\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} > 2.326,$$

since $P[\tau > 2.326] = 0.01$, and is to be accepted otherwise.

Similarly, when one wants to examine if μ is equal to μ_0 or

smaller, i.e. when one wants to test $H_0: \mu = \mu_0$ against the alternative hypotheses $H: \mu < \mu_0$, one should take only very small values of $\frac{\sqrt{n}(x-\mu_0)}{\sigma}$ as indicative of the falsity of H_0 . As regards very large

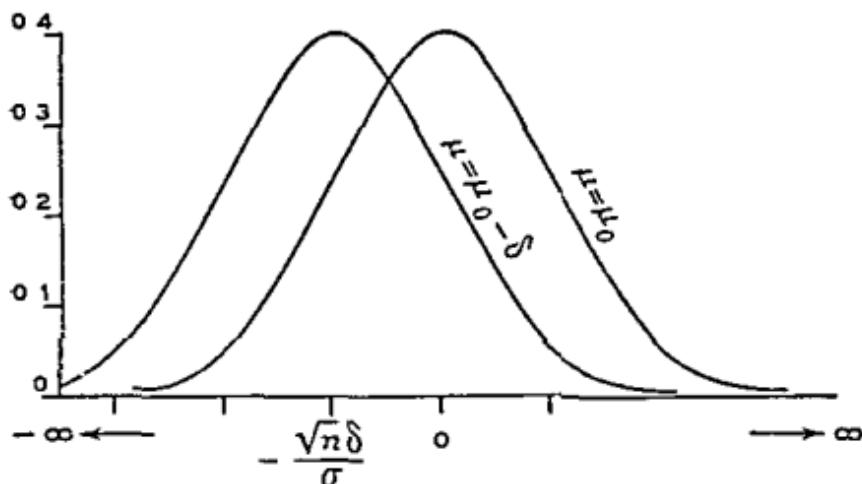


Fig. 15.2 Distribution of $\frac{\sqrt{n}(x-\mu_0)}{\sigma}$ for $\mu=\mu_0$ and for $\mu=\mu_0-\delta$

values of the statistic they are of course extremely improbable when $\mu=\mu_0$. But then they are still more improbable for any value of μ less than μ_0 (see Fig. 15.2). If the level of significance be 0.01, H_0 is, therefore, to be rejected when for the given sample

$$\frac{\sqrt{n}(x-\mu_0)}{\sigma} < -2.326$$

and is to be accepted otherwise.

From the above discussion, it will also be apparent why, in testing H_0 against the alternatives $H: \mu \neq \mu_0$ at the level of significance 0.01, we reject H_0 when $\frac{\sqrt{n}(x-\mu_0)}{\sigma}$ is less than -2.576 as well as when it is greater than 2.576.

The nature of the alternative hypotheses thus determines which type of test (one sided or two-sided) is to be used in any given case—whether the left tail of the curve of the distribution of the relevant statistic or its right tail or both are to be taken for defining values that lead to the rejection of the null hypothesis.

15.6 Neyman and Pearson's theory of testing of hypotheses

For a solution to the problem of testing of hypotheses, we have used here an intuitive approach. In order to give a more rational treatment of the problem, it is necessary to consider the probabilities of the two types of error that one may commit in rejecting or accepting a hypothesis on the basis of sample observations. These are the probability of the error committed in rejecting H_0 when, in fact, it is true (error I)*, and the probability of the error committed in accepting H_0 when actually an alternative hypothesis is true (error II). This approach has been adopted by J. Neyman and E. S. Pearson in formulating a theory of testing of hypotheses.

To present the basic principles of this theory, we shall suppose that the population, from which the variables x_1, x_2, \dots, x_n are a random sample, depends on a single unknown parameter θ , the form of the distribution being known. We shall suppose that the population distribution is continuous and that

$$p(\mathbf{x}|\theta) \quad \dots \quad (15.11)$$

is the joint density function of the variables $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

We shall denote by W the whole set of possible values of \mathbf{x} . W is called the sample space and may be looked upon as a region in n dimensions.

Consider the hypothesis

$$H_0 : \theta = \theta_0.$$

Any test for H_0 is nothing but a rule for rejecting or accepting H_0 , depending on the nature of the sample observations \mathbf{x} . A test would thus specify a set of points in W , called the *critical region* or *region of rejection* and denoted by w , and would require the rejection of H_0 if \mathbf{x} lies in w and the acceptance of H_0 if \mathbf{x} lies outside w (or in $W-w$). A test is thus determined by its critical region, and conversely.

The probability of error I associated with the test is then the same as the probability that \mathbf{x} lies in w when θ_0 is the true value of θ , or, in symbols,

$$P[\mathbf{x} \in w | \theta_0] = \int_{w} \dots \int p(\mathbf{x} | \theta_0) d\mathbf{x}, \quad \dots \quad (15.12)$$

where $d\mathbf{x} = dx_1 dx_2 \dots dx_n$.

*The probability is the same as the level of significance α (in case H_0 is simple).

As regards the probability of error II, suppose θ_1 , and not θ_0 , is the true value of θ . Then the probability of error II with respect to θ_1 is the same as the probability that x lies in $W - w$ when θ_1 is true or, in symbols,

$$\begin{aligned} P[x \in W - w | \theta_1] &= 1 - P[x \in w | \theta_1] \\ &= 1 - \int_{-\infty}^{\infty} \int p(x | \theta_1) dx \end{aligned} \quad (15.13)$$

$P[x \in w | \theta_1]$ which is the probability of rejecting H_0 when, in fact, θ_1 is the true value of θ , is called the *power* of the test with respect to θ_1 , because it measures in a sense the capacity of the test to detect the falsity of H_0 . Of course, $P[x \in w | \theta_1]$ may vary with θ_1 , and thus we may talk about the *power function* of the test.

Now, it would be an ideal situation if the test could minimise the probabilities of both types of error at the same time. However, with a fixed sample size, this is not possible—as one probability decreases the other increases. A reasonable procedure would then be to fix the probability of error I at a desirable level, i.e. to make

$$P[x \in w | \theta_0] = \alpha \text{ (say)}, \quad (15.14)$$

and to choose from all critical regions w , satisfying (15.14), one which has the maximum power (the minimum probability of error II). In case we are interested in the alternative hypotheses

$$H: \theta \neq \theta_0,$$

a critical region w_a will be the best if it be such that

$$P[x \in w_a | \theta_0] = \alpha \quad (15.15a)$$

and $P[x \in w_a | \theta] \geq P[x \in w | \theta]$ for all $\theta \neq \theta_0$, $(15.15b)$

whatever the other region w , satisfying (15.15a), may be. Such a region w_a is called a *uniformly most powerful critical region* (and the corresponding test a uniformly most powerful test) of size (or of level) α for

$$H_0: \theta = \theta_0$$

against the alternative hypotheses

$$H: \theta \neq \theta_0$$

In the same way, if we are interested in the alternative hypotheses $H: \theta > \theta_0$ ($H: \theta < \theta_0$), then w_a will be an ideal critical region, called

a uniformly most powerful critical region of size α for $H_0 : \theta = \theta_0$ against the alternatives $H : \theta > \theta_0$ ($H : \theta < \theta_0$), if it be such that

$$P[x \in w_0 | \theta_0] = \alpha \quad \dots \quad (15.16a)$$

and $P[x \in w_0 | \theta] \geq P[x \in w | \theta]$ for all $\theta > \theta_0$ (all $\theta < \theta_0$), $\dots \quad (15.16b)$

whatever the other region w , satisfying (15.16a), may be.

Unfortunately, for two-sided alternatives ($H : \theta \neq \theta_0$) in most situations no uniformly most powerful test exists, although for one-sided alternatives ($H : \theta > \theta_0$ or $H : \theta < \theta_0$) a uniformly most powerful test exists in most cases.

For two-sided alternatives, therefore, some additional criterion has to be found in making a choice among rival critical regions of size α . One such criterion is *unbiasedness*. A region w is called biased if

$$P[x \in w | \theta] < P[x \in w | \theta_0]$$

for some $\theta \neq \theta_0$. A biased region would thus reject H_0 with a smaller probability when H_0 is false than it would when H_0 is true. This is an undesirable feature, and we should, therefore, look for unbiased regions alone and restrict our choice of a desirable region of size α among those that are unbiased. In this situation, a w_0 such that

$$P[x \in w_0 | \theta_0] = \alpha, \quad \dots \quad (15.17a)$$

$$P[x \in w_0 | \theta] \geq \alpha \quad \text{for all } \theta \neq \theta_0 \quad \dots \quad (15.17b)$$

and $P[x \in w_0 | \theta] \geq P[x \in w | \theta] \quad \text{for all } \theta \neq \theta_0, \quad \dots \quad (15.17c)$

whatever the other region w satisfying (15.17a) and (15.17b), may be regarded as best. It would be called a *uniformly most powerful unbiased region* of size α for testing $H_0 : \theta = \theta_0$.

The intuitive method of test construction has been followed in this book because of its simplicity. But it may be stated that the tests obtained are, as a rule, the best available even from the point of view of Neyman and Pearson.

Thus, when the alternative is one-sided our suggested test is uniformly most powerful, while in the case of a two-sided alternative the suggested test is, as a rule, uniformly most powerful among unbiased tests.

Ex. 15.10 The breaking strength of pieces of a type of string has mean 18.2 lb. and standard deviation 2.1 lb. A new method of

manufacture of strings is supposed to give a higher mean breaking strength, but for this the standard deviation is known to be practically the same 15 pieces manufactured by the new method have breaking strength (in lb) as follows

17.9	21.0	20.5
16.8	19.2	17.4
18.4	15.9	16.9
18.0	19.8	17.6
21.8	19.3	21.9

Do these results indicate that the new method is really a better one?

Let us denote the breaking strength of a string by x . The new method of manufacture may be said to be better than the existing method if in the population of strings produced by the new method x has mean (μ) greater than 18.2 lb. We have thus to test a null hypothesis in this case, viz.

$$H_0 \quad \mu = 18.2 \text{ lb},$$

against the alternatives

$$H \quad \mu > 18.2 \text{ lb}$$

For this purpose, we assume (1) that in the population x is distributed in the normal form and (2) that the given values of x , say

$$x_i \quad (i=1, 2, \dots, 15),$$

are random and independent observations

The population standard deviation of x is given $\sigma = 2.1$ lb. Again, for the given sample $v = 18.83$ lb, so that

$$\tau = \frac{\sqrt{15}(v - 18.2)}{2.1} = \frac{3.873(18.83 - 18.2)}{2.1} = 1.162$$

Since it is smaller than 2.326 as well as 1.645, the hypothesis H_0 is to be accepted at both 1% and 5% levels of significance. In other words, the new method of manufacture does not seem to be superior to the old method.

Simple and composite hypotheses

The hypothesis to be tested may or may not specify the population distribution completely. When it does, it is called a *simple hypothesis*, otherwise, it is called *composite*. Thus the hypotheses considered above

are all simple hypotheses. On the other hand, suppose for a normal distribution the mean (μ) and variance (σ^2) are both unknown, and let the hypothesis to be tested be $H_0 : \mu = \mu_0$. It specifies the mean, but the variance is left unspecified. Hence it is a composite hypothesis with one *degree of freedom* (one parameter being unspecified).

In the case of a composite hypothesis too, the general principles to be followed remain the same as described previously. But here in order to make the level of the test equal to α , we must have

$$P[E \in w | H_0] = \text{constant, whatever the unspecified parameters, say, } \theta', \theta'', \dots, \text{ may be.} \dots \quad (15.18)$$

Not all regions of W may have this property. Those that do are called *similar regions* (or regions similar to the sample space). Our choice then has to be made among similar regions. We first select all *similar regions of size α* . Next, from these we choose one that is uniformly most powerful, and if no uniformly most powerful similar region exists, then one that is uniformly most powerful among unbiased similar regions of size α .

15.7 Likelihood-ratio tests

Closely allied to the maximum-likelihood method of estimation, there is a simple procedure for obtaining tests which has an intuitive appeal and which generally gives tests with desirable properties.

Suppose we are to test a simple hypothesis

$$H_0 : \theta = \theta_0 \quad \dots \quad (15.19)$$

against all alternatives. Given the sample observations, x_1, x_2, \dots, x_n , a natural way of judging the acceptability or otherwise of the hypothesis would be to compare the likelihood $L(\theta_0)$ with the maximum possible value of $L(\theta)$. If

$$\lambda = \frac{L(\theta_0)}{\max L(\theta)} \quad \dots \quad (15.20)$$

were near to unity, then in the light of the given sample H_0 would seem highly plausible; on the other hand, if this were near to zero, H_0 would seem to have little validity. A test for H_0 is thus provided by a critical region defined by $\lambda < \lambda_0$, where λ_0 is such that $P[\lambda < \lambda_0 | H_0] = \alpha$. This relation fixes the probability of error I at α . As regards power, it can be shown that the likelihood-ratio test

provides a sort of compromise among tests that would be most powerful for individual alternatives.

When more than one parameter are unknown, a hypothesis like (15.19) leaves some parameter or parameters (say, θ) unspecified, so $L(\theta, \theta)$ in that case is not a constant. Hence the comparison is then made between $\max_{H_0} L(\theta, \theta)$, the highest possible value of $L(\theta, \theta)$ under H_0 , the condition that the hypothesis is true and the absolute maximum, $\max L(\theta, \theta)$. The critical region here is defined by $\lambda < \lambda_0$, where

$$\lambda = \frac{\max_{H_0} L(\theta, \theta)}{\max L(\theta, \theta)} \quad (15.21)$$

and λ_0 is such that

$$P[\lambda < \lambda_0 | H_0] = \alpha$$

A result that will simplify the test in the case of large samples is that $-2\log \lambda$ is, under H_0 , distributed approximately as a χ^2 with $v d.f.$, where v denotes the number of parameters specified by H_0 .

Ex 15.11 Consider a sample of size n from a normal distribution such that the sample observations

$$x_1, x_2, \dots, x_n$$

are random and independent,

(a) Suppose the mean of the population (μ) is unknown but the variance is known ($\sigma^2 = \sigma_0^2$). To obtain the likelihood ratio test for the simple hypothesis

$$H_0 \quad \mu = \mu_0,$$

note that here the likelihood function is

$$L(\mu) = \frac{1}{(\sigma_0 \sqrt{2\pi})^n} \exp[-\sum_i (x_i - \mu)^2 / 2\sigma_0^2]$$

The maximum likelihood estimate of μ has been found to be \bar{x} in Section 15.13. Hence

$$\max L(\mu) = \frac{1}{(\sigma_0 \sqrt{2\pi})^n} \exp[-\sum_i (x_i - \bar{x})^2 / 2\sigma_0^2]$$

and

$$\lambda = \frac{L(\mu_0)}{\max L(\mu)} = \exp[-n(\bar{x} - \mu_0)^2 / 2\sigma_0^2]$$

The critical region is thus given by

$$\lambda < \lambda_0$$

or $\frac{n(\bar{x} - \mu_0)^2}{\sigma_0^2} > C$

or $\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma_0} > k,$

where k is such that

$$P\left[\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma_0} > k \mid H_0\right] = \alpha.$$

Since under H_0 $\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0}$ is a normal deviate, we must have $k = \tau_{\alpha/2}$. Hence the critical region is given by

$$\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma_0} > \tau_{\alpha/2}.$$

(b) If both μ and σ^2 are unknown, we may like to test the composite hypothesis

$$H_0 : \mu = \mu_0.$$

The likelihood function is now

$$L(\mu, \sigma) = \frac{1}{(\sigma \sqrt{2\pi})^n} \exp\left[-\sum_i (x_i - \mu)^2 / 2\sigma^2\right].$$

The maximum-likelihood estimates of μ and σ are \bar{x} and s , respectively. Hence

$$\max L(\mu, \sigma) = \frac{1}{(s \sqrt{2\pi})^n} \exp[-n/2].$$

Under H_0 , the likelihood function is

$$L(\mu_0, \sigma) = \frac{1}{(\sigma \sqrt{2\pi})^n} \exp\left[-\sum_i (x_i - \mu_0)^2 / 2\sigma^2\right].$$

Here the maximum-likelihood estimate of σ is

$$s_0 = \sqrt{\frac{\sum_i (x_i - \mu_0)^2}{n}}.$$

As such, $\max_{H_0} L(\mu, \sigma) = \max_{H_0} L(\mu_0, \sigma) = \frac{1}{(s_0 \sqrt{2\pi})^n} \exp[-n/2]$.

The critical region for the likelihood-ratio test is then given by

$$\lambda < \lambda_0,$$

or by

$$\frac{s^2}{ns_0^2} < k,$$

or, since $ns_0^2 = n(x - \mu_0)^2 + ns^2$ and $ns^2 = (n-1)s'^2$, by

$$\frac{\sqrt{n|x - \mu_0|}}{s} > C,$$

where C is such that

$$P\left[\frac{\sqrt{n|x - \mu_0|}}{s} > C \mid H_0\right] = \alpha$$

But $\frac{\sqrt{n(x - \mu_0)}}{s'}$ is, under H_0 , distributed as a t with $n-1$ df

Hence the test is defined by the critical region

$$\frac{\sqrt{n|x - \mu_0|}}{s} > t_{\alpha/2, n-1}$$

Ex 15.12 Consider n Bernoullian trials with probability of success p . We may associate with the i th trial a variable x_i with probability mass function

$$f(x_i) = p^{x_i} (1-p)^{1-x_i}, \text{ for } x_i = 0, 1$$

For a given set of values of x_1, x_2, \dots, x_n , the likelihood function is

$$L(p) = p^x (1-p)^{n-x},$$

where x = number of successes in the n trials taken together

In order to test the hypothesis

$$H_0: p = \frac{1}{2},$$

we shall make use of the likelihood-ratio

$$\frac{L(1/2)}{\max L(P)} = \frac{(1/2)^x}{\binom{n}{x} \left(\frac{1-x}{n}\right)^{n-x}} = \frac{(n/2)^x}{x^x (n-x)^{n-x}},$$

since the maximum-likelihood estimate of p is $\frac{x}{n}$

The critical region is then given by

$$x^* (n-x)^{n-x} > C,$$

or by

$$x < \frac{n}{2} - C_1 \text{ or } x > \frac{n}{2} + C_1,$$

the function $x^* (n-x)^{n-x}$ being symmetrical about $x = \frac{n}{2}$ and increasing as x deviates in either direction from $\frac{n}{2}$. The constant C_1 is such that

$$P\left[x < \frac{n}{2} - C_1 \mid H_0\right] + P\left[x > \frac{n}{2} + C_1 \mid H_0\right] = \alpha.$$

Actually, the level, in most cases, will only be approximately equal to α , since here we are dealing with a discrete variable.

Questions and exercises

15.1 Explain the problems of estimation and testing of hypotheses. Distinguish between point estimation and interval estimation.

15.2 Discuss, with examples, the notion of a minimum-variance unbiased estimator.

15.3 Explain, with suitable illustrations, the criteria of consistency, efficiency and sufficiency, as used in the theory of estimation.

15.4 Give an outline of the Neyman-Pearson theory of testing of hypotheses, explaining the concepts of the errors of type I and type II, power and unbiasedness.

15.5(a) Describe the maximum-likelihood method of estimation. What are the properties of a maximum-likelihood estimator?

(b) What is a likelihood-ratio test? What can be said of the large-sample behaviour of the associated test criterion?

15.6 A variable x is normally distributed in the population with mean 100 and standard deviation 5. Determine how large a sample is to be taken from this population in order that the sample mean will not differ from 100 by more than 1 with probability 0.95.

Ans. 97.

15.7 From a large lot of freshly-minted coins a random sample of size 50 is taken. The mean weight of coins in the sample is found to be 28.57 gm. Assuming that the population standard deviation

of weight is 1.25 gm will it be reasonable to suppose that the population mean is 28 gm ? *Partial ans* $\tau=3.224$

15.8 For the data of *Exercise 15.7* obtain the 99% confidence limits to the mean weight of all coins in the lot

Ans 28.11 gm and 29.03 gm

15.9 Find the maximum-likelihood estimators of θ for random and independent observations x_1, x_2, \dots, x_n from the populations

$$(a) f(x) = \frac{1}{\theta} \exp\left[-\frac{x}{\theta}\right], \quad 0 < x < \infty,$$

$$(b) f(x) = \exp[-(x-\theta)] \quad \theta \leq x < \infty$$

$$(c) f(x) = \frac{1}{2} \exp[-|x-\theta|], \quad -\infty < x < \infty$$

Ans (a) $\hat{\theta}$ =sample mean, (b) $\hat{\theta}=x_{(1)}$, the smallest observation, (c) $\hat{\theta}$ =sample median

15.10(a) An urn contains white and black balls in unknown proportions, the total number of balls being 8. Three balls are taken at random, of which 2 are found to be white and 1 black. Find the maximum-likelihood estimate of the number of white balls in the urn.

[Hint] The likelihood function is

$$L(N_1) = \frac{\binom{N_1}{2} \binom{8-N_1}{1}}{\binom{8}{3}}$$

Ans Both 5 and 6 are m.l. estimates

(b) A random sample of size n has been taken (without replacements) from a population of size N . N is unknown, but the number of individuals in the population with the character A is known to be N_1 . Let x among the n members of the sample have the character A . Show that the maximum-likelihood estimate of N is, approximately, $\frac{nN_1}{x}$.

[Hint] The likelihood function is

$$L(N) = \frac{\binom{N_1}{x} \binom{N-N_1}{n-x}}{\binom{N}{n}}$$

Discuss how this method may be used in estimating the number of fish in a pond or the number of birds in an aviary.

5.11 For the set-up of Ex. 15.1, obtain the expected value of $\frac{f}{n}(1-\frac{f}{n})$. Hence suggest an unbiased estimator of $p(1-p)$.

15.12 (a) Find the maximum-likelihood estimator of $\frac{1}{p}$ for the observation x from the discrete distribution

$$f(x) = (1-p)^{x-1} p, \text{ for } x=1, 2, \dots$$

(b) Show that the estimator is unbiased. What is the variance of the estimator ? *Partial ans.* $1/p=x$.

15.13 Let x be normally distributed with known mean μ_0 but unknown variance σ^2 . Evaluate

$$E(|x-\mu_0|).$$

Hence suggest an unbiased estimator of σ based on a set of random and independent observations, x_1, x_2, \dots, x_n . Derive the expectation of $s_0 = \sqrt{\frac{1}{n} \sum_i (x_i - \mu_0)^2}$ and suggest an alternative unbiased estimator of σ . Compare the variances of the two estimators and comment.

15.14 Starting from the equation

$$\sigma^2 = E(x^2) - \mu^2,$$

obtain an unbiased estimator of μ^2 . What is its principal defect ?

15.15 x_1, x_2, \dots, x_n are independent, random observations from the rectangular population with density

$$f(x) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta.$$

Consider the critical region $x_{(n)} > 0.8$ for testing the hypothesis $H_0 : \theta = 1$, where $x_{(n)}$ is the largest of x_1, x_2, \dots, x_n . What is the associated probability of error I and what is the power function ?

Partial ans. Power function is

$$P[E \in w | \theta] = \begin{cases} 0 & \text{for } \theta \leq 0.8, \\ 1 - \left(\frac{0.8}{\theta}\right)^n & \text{for } \theta > 0.8. \end{cases}$$

15.16 Obtain the likelihood-ratio tests, based on independent random observations x_1, x_2, \dots, x_n , for the following hypotheses

$$(a) H_0: \theta = \theta_0 \text{ when } f(x) = \frac{1}{\theta} \exp\left[-\frac{x}{\theta}\right], \quad 0 < x < \infty,$$

$$(b) H_0: \theta = \theta_0 \text{ when } f(x) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta$$

15.17 Let t_1, t_2, \dots, t_k be mutually independent and unbiased estimators of μ with variances V_1, V_2, \dots, V_k , respectively. Consider a linear function

$$T = a + \sum_i b_i t_i$$

Choose the constants a, b_1, b_2, \dots, b_k in such a way that T is unbiased and has the smallest variance among all unbiased linear estimators

$$Ans \quad T = \sum_i (t_i/V_i) / \sum_i (1/V_i)$$

SUGGESTED READING

- [1] Anderson, R. L. and Bancroft, T. A. *Statistical Theory in Research* (Chs 8—11) McGraw-Hill, 1952
- [2] Hogg, R. V. and Craig, A. T. *Introduction to Mathematical Statistics* (Chs 5, 9—11) Macmillan, 1965, and Amerind
- [3] Keeping, E. S. *Introduction to Statistical Inference* (Chs 5, 6) Van Nostrand, 1962, and Affiliated East-West Press
- [4] Mood, A. M. and Gravbill, F. A. *Introduction to the Theory of Statistics* (Chs 7, 8, 11, 12) McGraw-Hill, 1963, and Kōgakusha
- [5] Rao, C. R. *Advanced Statistical Methods in Biometric Research* (Chs 4, 8a) John Wiley, 1952
- [6] Wald, A. *Principles of Statistical Inference* Notre Dame, 1942

16

EXACT TESTS AND CONFIDENCE INTERVALS

16.1 Introduction

The general procedure followed in testing a hypothesis regarding a parameter or in obtaining confidence limits for a parameter has been explained in the previous chapter. We shall now consider the application of the general procedure to particular problems.

16.2 Tests relating to binomial distributions

Suppose a random sample of size n is drawn from a population for which the proportion of individuals having a character A , say p , is unknown. In order to test the hypothesis

$$H_0 : p = p_0,$$

we make use of the statistic x , the number of members of the sample having the character A .

Now, under the hypothesis H_0 , x is distributed binomially with parameters n and p_0 .

Let the observed value of x be x_0 .

(1) In case we are required to test H_0 against the alternatives $H : p > p_0$, we shall compute the probability

$$P[x \geq x_0 | p_0] = \sum_{x \geq x_0} \binom{n}{x} p_0^x (1-p_0)^{n-x}.$$

If this is smaller than the specified level of significance, α , we shall consider x_0 to be an unlikely value under the hypothesis and shall, therefore, reject H_0 . Otherwise, H_0 will be accepted.

(2) If, on the other hand, the alternative hypotheses are $H : p < p_0$, we shall compute

$$P[x \leq x_0 | p_0] = \sum_{x \leq x_0} \binom{n}{x} p_0^x (1-p_0)^{n-x}$$

and shall reject or accept H_0 according as this probability is or is not smaller than α .

(3) The two sided alternatives $H: p \neq p_0$ may be of interest in case $p_0 = \frac{1}{2}$ (e.g. when our problem is to test whether a coin is unbiased) Here we compute

$$\begin{aligned} & P\left[x - \frac{n}{2} \leq -|x_0 - \frac{n}{2}| \middle| \frac{1}{2}\right] + P\left[x - \frac{n}{2} \geq |x_0 - \frac{n}{2}| \middle| \frac{1}{2}\right] \\ &= P\left[x \leq \frac{n}{2} - d_0 \middle| \frac{1}{2}\right] + P\left[x \geq \frac{n}{2} + d_0 \middle| \frac{1}{2}\right] \\ &= \frac{1}{2^n} \sum_{x < \frac{n}{2} - d_0} \binom{n}{x} + \frac{1}{2^n} \sum_{x > \frac{n}{2} + d_0} \binom{n}{x}, \end{aligned}$$

where $d_0 = |x_0 - \frac{n}{2}|$, and compare it with α for rejection or acceptance of $H_0: p = \frac{1}{2}$

For some selected n and p , these probabilities may be had from Table 37 of the *Biometrika Tables*, Vol I. Much more extensive are the *Tables of the Binomial Probability Distribution*, prepared by the (U.S.) National Bureau of Standards.

Consider next two populations for which the proportions of individuals having a character A are p_1 and p_2 . Again, we shall denote by x_1 and x_2 respectively, the numbers of members having the character A in random samples of sizes n_1 and n_2 drawn independently from the two populations. We may be interested in the hypothesis

$$H_0: p_1 = p_2$$

We shall again make use of the statistics x_1 and x_2 , but shall concentrate our attention on samples for which $x = x_1 + x_2$ is a constant (the same as the observed sum of x_1 and x_2). Under H_0 , if we denote the common value of the two proportions by p , the pmf's of x_1 , x_2 and $x = x_1 + x_2$ are

$$f(x_1) = \binom{n_1}{x_1} p^{x_1} (1-p)^{n_1 - x_1},$$

$$f(x_2) = \binom{n_2}{x_2} p^{x_2} (1-p)^{n_2 - x_2}$$

and

$$f(x) = \binom{n_1 + n_2}{x} p^x (1-p)^{n_1 + n_2 - x}$$

The conditional p.m.f. of x_1 for given x is, therefore,

$$\begin{aligned} f(x_1|x) &= \frac{\binom{n_1}{x_1} \binom{n_2}{x-x_1} p^{x_1+x_2} (1-p)^{n_1+n_2-x_1-x_2}}{\binom{n_1+n_2}{x} p^x (1-p)^{n_1+n_2-x}} \\ &= \frac{\binom{n_1}{x_1} \binom{n_2}{x-x_1}}{\binom{n_1+n_2}{x}} \quad \dots \quad (16.1) \end{aligned}$$

(i.e. is of the hypergeometric form).

If the observed value of x_1 is x_{10} and that of x is x_0 , we consider the conditional p.m.f. $f(x_1|x_0)$ for testing H_0 .

(1) If we are interested in the alternatives $H: p_1 > p_2$, we compute

$$P[x_1 \geq x_{10} | x = x_0] = \sum_{x_1 \geq x_{10}} \frac{\binom{n_1}{x_1} \binom{n_2}{x_0 - x_1}}{\binom{n_1+n_2}{x_0}}$$

and reject or accept H_0 according as this probability is or is not smaller than α .

(2) On the other hand, if we are interested in the alternatives $H: p_1 < p_2$, we compute

$$P[x_1 \leq x_{10} | x = x_0] = \sum_{x_1 \leq x_{10}} \frac{\binom{n_1}{x_1} \binom{n_2}{x_0 - x_1}}{\binom{n_1+n_2}{x_0}}$$

and compare it with α for rejection or acceptance of H_0 .

These probabilities may be obtained directly from the *Tables of the Hypergeometric Distribution* (Stanford University Press).

Ex. 16.1 A manufacturer of fluorescent tubes claims that no more than 6% of his products are defective. A sample of 20 tubes is found to contain 4 defective tubes. Does the manufacturer's claim seem justified in the light of these data?

The number of defectives in a sample of size 20 may be supposed to be distributed in the binomial form with parameters $n=20$ and p (unknown). Under the hypothesis $H_0: p=0.06$, which is to be tested against the alternatives $H: p>0.06$, the probability that a sample

will have 4 defectives or more is

$$\sum_{x=4}^{20} \binom{20}{x} (0.06)^x (0.94)^{20-x} = 1 - \sum_{x=0}^3 \binom{20}{x} (0.06)^x (0.94)^{20-x}$$

$$= 1 - 0.97104 = 0.02896$$

This probability being less than 0.05, the hypothesis is to be rejected at the 5% level. The data thus seem to contradict the claim made by the manufacturer. (If one uses the 1% level, then, of course, the hypothesis will have to be accepted, i.e. one will not consider the data to be incompatible with the manufacturer's assertion.)

Ex 16.2 It is required to compare two methods of treating a type of allergy. Method I was used on 15 patients and Method II on 14. The results are shown below

	Method I	Method II
Cured	6	11
Not cured	9	3
Total	15	14

Is Method II better than Method I?

Here $n_1 = 15$, $n_2 = 14$, $x_0 = 17$ and $x_{10} = 6$

As the hypothesis $H_0: p_1 = p_2$ is to be tested against $H: p_1 < p_2$ we compute

$$P[x_1 \leq x_{10} | x=x_0] = \frac{\left[\binom{15}{3} + \binom{15}{4} \binom{14}{13} + \binom{15}{5} \binom{14}{12} + \binom{15}{6} \binom{14}{11} \right]}{\binom{29}{17}}$$

$$= 0.0407$$

Since this is less than 0.05, H_0 is to be rejected at the 5% level. Thus at the 5% level, Method II is to be considered better than Method I.

16.3 Tests relating to Poisson distributions

Suppose x_1, x_2, \dots, x_n are independent random observations from a Poisson population with unknown parameter λ . Here we may be required to test the hypothesis

$$H_0: \lambda = \lambda_0$$

To develop a test we make use of the sufficient statistic

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

which, as has been shown in Section 14.5, is itself distributed in the Poisson form with parameter $n\lambda$. The p.m.f. of y under H_0 is, therefore,

$$f(y) = \frac{\exp(-n\lambda_0)(n\lambda_0)^y}{y!} \quad (y=0, 1, 2, \dots).$$

Let y_0 be the observed value of y .

(1) If the alternatives of interest are $H: \lambda > \lambda_0$, we evaluate

$$P[y \geq y_0 | \lambda_0] = \sum_{y=y_0}^{\infty} \frac{\exp(-n\lambda_0)(n\lambda_0)^y}{y!}$$

and proceed as in Section 16.2.

(2) On the other hand, if the alternatives of interest are $H: \lambda < \lambda_0$, then we have to compute

$$P[y \leq y_0 | \lambda_0] = \sum_{y=0}^{y_0} \frac{\exp(-n\lambda_0)(n\lambda_0)^y}{y!}.$$

These probabilities may be obtained from Table 7 of *Biometrika Tables*, Vol. I.

We may, again, be interested in a comparison of the parameters λ_1 and λ_2 of two Poisson populations. Let x_{1i} ($i=1, 2, \dots, n_1$) be independent random observations from the first population and x_{2i} ($i=1, 2, \dots, n_2$) be independent random observations from the second. A test for

$$H_0: \lambda_1 = \lambda_2$$

should be based on the sufficient statistics

$$y_1 = \sum_{i=1}^{n_1} x_{1i};$$

$$\text{and} \quad y_2 = \sum_{i=1}^{n_2} x_{2i}.$$

We consider only those samples of sizes n_1 and n_2 for which $y = y_1 + y_2$ is a constant (the same as the observed sum of y_1 and y_2). Under H_0 , if we denote the common value of the two parameters by

λ , then the p m fs of y_1, y_2 and $y = y_1 + y_2$ are

$$f(y_1) = \exp(-n_1\lambda)(n_1\lambda)^{y_1}/y_1!$$

$$f(y_2) = \exp(-n_2\lambda)(n_2\lambda)^{y_2}/y_2!$$

and

$$f(y) = \exp[-(n_1+n_2)\lambda][(n_1+n_2)\lambda]^y/y!$$

The conditional p m f of y_1 for given y is, therefore,

$$\begin{aligned} f(y_1|y) &= \frac{\exp[-(n_1+n_2)\lambda](n_1\lambda)^{y_1}(n_2\lambda)^{y-y_1}/(y_1!(y-y_1)!)}{\exp[-(n_1+n_2)\lambda][(n_1+n_2)\lambda]^y/y!} \\ &= \binom{y}{y_1} \left(\frac{n_1}{n_1+n_2}\right)^{y_1} \left(\frac{n_2}{n_1+n_2}\right)^{y-y_1} \end{aligned} \quad (16.2)$$

(i.e. is binomial with parameters y and $\frac{n_1}{n_1+n_2}$)

Denoting the observed values of y_1 and y by y_{10} and y_0 , respectively, we consider the conditional p m f $f(y_1|y_0)$ for testing H_0

(1) In case our interest lies in the alternatives $H: \lambda_1 > \lambda_2$, we compute

$$P[y_1 \geq y_{10} | y = y_0] = \sum_{y_1 \geq y_{10}} \binom{y_0}{y_1} \left(\frac{n_1}{n_1+n_2}\right)^{y_1} \left(\frac{n_2}{n_1+n_2}\right)^{y_0-y_1}$$

and compare it with α for rejection or acceptance of H_0

(2) If, instead, we are interested in the alternatives $H: \lambda_1 < \lambda_2$, we have to compute

$$P[y_1 \leq y_{10} | y = y_0] = \sum_{y_1 \leq y_{10}} \binom{y_0}{y_1} \left(\frac{n_1}{n_1+n_2}\right)^{y_1} \left(\frac{n_2}{n_1+n_2}\right)^{y_0-y_1}$$

and compare it with α for rejection or acceptance of H_0

16.4 A test for independence of two attributes

In many investigations one is faced with the problem of judging whether two qualitative characters, say A and B , may be said to be independent

Let us denote the forms of A by A_i ($i=1, 2, \dots, k$), the forms of B by B_j ($j=1, 2, \dots, l$), and the probability associated with the cell $A_i B_j$ in the two-way classification of the population by p_{ij} . The probability associated with A_i is then

$$p_i = \sum_{j=1}^l p_{ij}$$

and the probability associated with B_j is

$$p_{0j} = \sum_{i=1}^k p_{ij}.$$

The hypothesis to be tested here is

$$H_0 : p_{ij} = p_{i0} p_{0j}, \text{ all } i, j.$$

Suppose now that for a random sample of size n drawn with replacements, n_{ij} is the observed frequency for the cell $A_i B_j$. The marginal frequency of A_i is

$$n_{i0} = \sum_{j=1}^l n_{ij},$$

and the marginal frequency of B_j is

$$n_{0j} = \sum_{i=1}^k n_{ij}.$$

Note that the joint p.m.f. of n_{ij} is multinomial :

$$\begin{aligned} & f(n_{11}, n_{12}, \dots, n_{kl} | p_{11}, p_{12}, \dots, p_{kl}) \\ &= \frac{n!}{\prod_{i=1}^k \prod_{j=1}^l (n_{ij}!)^i} \prod_{i=1}^k \prod_{j=1}^l (p_{ij})^{n_{ij}}. \end{aligned}$$

Under H_0 , this is

$$\frac{n!}{\prod_{i=1}^k \prod_{j=1}^l (n_{ij}!)^i} \prod_{i=1}^k (p_{i0})^{n_{i0}} \prod_{j=1}^l (p_{0j})^{n_{0j}}.$$

This could be used for testing H_0 if p_{i0} ($i=1, 2, \dots, k$) and p_{0j} ($j=1, 2, \dots, l$) were known quantities.

When these are unknown, we use, instead of the unconditional distribution of n_{ij} 's, their conditional distribution for fixed marginals, n_{i0} 's and n_{0j} 's. Under H_0 , the joint p.m.f. of n_{i0} ($i=1, 2, \dots, k$) is

$$\frac{n!}{\prod_{i=1}^k (n_{i0}!)^i} \prod_{i=1}^k (p_{i0})^{n_{i0}}$$

and the joint p.m.f. of n_{0j} ($j=1, 2, \dots, l$) is

$$\frac{n!}{\prod_{j=1}^l (n_{0j}!)^j} \prod_{j=1}^l (p_{0j})^{n_{0j}}.$$

Therefore, under H_0 , the conditional distribution of n_{ij} 's for fixed marginals has the pmf

$$\frac{\frac{n!}{\prod_i \prod_j (n_{ij}!)}}{\frac{\prod_i (p_{i0})^{n_{i0}} \prod_j (p_{0j})^{n_{0j}}}{\prod_i (n_{i0}!) \prod_j (n_{0j}!)}} = \frac{\prod_i (n_{i0}!) \prod_j (n_{0j}!)}{n! \prod_i \prod_j (n_{ij}!)}, \quad (16)$$

This conditional distribution now provides us with a test for H_0 .

The use of this technique is illustrated in the following example with a 2×2 table.

Ex. 16.3 The following table is based on a random sample of persons attending the preview of a motion picture

	Age below 40	Age 40 or above	Total
Liked the picture	32	8	40
Did not like the picture	4	6	10
Total	36	14	50

Judge whether the picture has equal appeal to the young and the old or whether it is more liked by the young.

The problem here is to test the hypothesis (H_0) that the two attributes, youth and liking for the picture, are independent, against the alternative that they are positively associated. We have then added up the probabilities, under H_0 , of the given table and of those indicating more extreme positive association (and having the same marginals). These tables are

32	8
4	6

Probability=0.0172

33	7
3	7

Probability=0.0024

34	6
2	8

Probability=0.0002

35	5
1	9
Probability=0.0000	

36	4
0	10
Probability=0.0000	

Since the sum of these probabilities, viz. 0.0198, is smaller than 0.05, H_0 is to be rejected at the 5% level. In other words, here the data do indicate that the picture is more popular with the young than with the old.

In the following sections, we shall consider different problems in testing of hypotheses and interval estimation which can be solved by using the four fundamental distributions discussed in Section 14.6. It will be assumed in all cases that the population distribution is normal.

16.5 Problems regarding a univariate normal distribution

Consider a population where x is normally distributed with mean μ and standard deviation σ . Let x_1, x_2, \dots, x_n be n random and independent observations on x obtained from this population. We shall denote by \bar{x} the sample mean of x :

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

and by s'^2 the sample variance of x :

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

The distinction between s'^2 and s^2 is to be noted. In s'^2 the divisor is $n-1$, which makes it an unbiased estimator of σ^2 . For

$$\sum_i (x_i - \bar{x})^2 = \sum_i (x_i - \mu)^2 - n(\bar{x} - \mu)^2,$$

so that

$$\begin{aligned} E(s'^2) &= \frac{1}{n-1} E\left\{ \sum_i (x_i - \bar{x})^2 \right\} \\ &= \frac{1}{n-1} E\left\{ \sum_i (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \right\} \\ &= \frac{1}{n-1} \left\{ \sum_i \text{var}(x_i) - n \text{var}(\bar{x}) \right\} \\ &= \frac{1}{n-1} \left\{ n\sigma^2 - n \cdot \frac{\sigma^2}{n} \right\} = \sigma^2. \end{aligned} \quad \dots \quad (16.4)$$

Case 1 μ unknown, σ known

Here we may be required to test the null hypothesis $H_0: \mu = \mu_0$. It has been shown in the previous chapter that the test procedure for H_0 in this case is based on the statistic $\frac{\sqrt{n}(x - \mu_0)}{\sigma}$, which is distributed as a normal deviate (τ) under this hypothesis.

1 For alternatives $H: \mu > \mu_0$, H_0 is rejected if for the given sample $\tau > \tau_{\alpha}$ (and is accepted otherwise).

2 For alternatives $H: \mu < \mu_0$, H_0 is rejected if for the given sample $\tau < \tau_{1-\alpha} (= -\tau_\alpha)$.

3 For alternatives $H: \mu \neq \mu_0$, H_0 is rejected if for the given sample $|\tau| > \tau_{\alpha/2}$.

In each case α denotes the chosen level of significance.

As regards the problem of interval estimation of μ , it has been shown that the limits $x - \tau_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ and $x + \tau_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, computed for the given sample, are the confidence limits to μ with confidence coefficient $1 - \alpha$.

Case 2 μ known, σ unknown

Here one may be interested in testing a hypothesis regarding σ or in estimating σ .

It is seen that x_i is a normal variable with mean μ and standard deviation σ . Hence

$$\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2, \quad (16.5)$$

being the sum of squares of n independent normal deviates, is distributed as a χ^2 with n d.f.

For testing $H_0: \sigma = \sigma_0$, we make use of the fact that

$$\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma_0} \right)^2 = \frac{\sum (x_i - \mu)^2}{\sigma_0^2}$$

is a χ^2 with n d.f. under this hypothesis.

1 For alternatives $H: \sigma > \sigma_0$, H_0 is rejected in case for the given sample $\chi^2 > \chi^2_{\alpha, n}$.

2 For alternatives $H: \sigma < \sigma_0$, H_0 is rejected if for the given sample $\chi^2 < \chi^2_{1-\alpha, n}$.

3 For alternatives $H: \sigma \neq \sigma_0$, H_0 is rejected if for the given sample $\chi^2 < \chi^2_{1-\alpha/2, n}$ or $\chi^2 > \chi^2_{\alpha/2, n}$.

As a point estimate of σ , we have

$$\sqrt{\frac{1}{n} \sum_i (x_i - \mu)^2}.$$

To get a confidence interval for σ , we note that

$$P\left[\chi^2_{1-\alpha/2, n} \leq \frac{\sum_i (x_i - \mu)^2}{\sigma^2} \leq \chi^2_{\alpha/2, n}\right] = 1 - \alpha$$

$$\text{or } P\left[\frac{\sum_i (x_i - \mu)^2}{\chi^2_{\alpha/2, n}} \leq \sigma^2 \leq \frac{\sum_i (x_i - \mu)^2}{\chi^2_{1-\alpha/2, n}}\right] = 1 - \alpha.$$

The confidence limits to σ^2 are, therefore,

$$\frac{\sum_i (x_i - \mu)^2}{\chi^2_{\alpha/2, n}} \text{ and } \frac{\sum_i (x_i - \mu)^2}{\chi^2_{1-\alpha/2, n}}.$$

The confidence limits to σ are just the positive square-roots of these quantities, with the same confidence coefficient, $1 - \alpha$.

Case 3. μ and σ both unknown

Here to test $H_0 : \mu = \mu_0$ or to have confidence limits to μ , one cannot use the statistic $\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$ since σ is unknown. σ is in this case replaced by its sample estimate, $s' = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$. The resulting expression will be

$$\frac{\sqrt{n}(\bar{x} - \mu)}{s'}.$$

Now, it has been shown in Section 14.7 that

$$\frac{(n-1)s'^2}{\sigma^2} = \frac{\sum_i (x_i - \bar{x})^2}{\sigma^2} \quad \dots \quad (16.6)$$

is a χ^2 with $n-1$ d.f. and is distributed independently of \bar{x} . Thus

$$\frac{\sqrt{n}(\bar{x} - \mu)}{s'} = \frac{\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}}{\sqrt{\frac{(n-1)s'^2}{\sigma^2} / (n-1)}} \quad \dots \quad (16.7)$$

being of the form $\frac{\tau}{\sqrt{\chi^2/(n-1)}}$, where χ^2 has $n-1$ d.f. and is independent of τ , is distributed as a t with $n-1$ d.f.

To test $H_0: \mu = \mu_0$ we may, therefore, use the statistic

$$t = \frac{\sqrt{n}(x - \mu_0)}{s}$$

with $df = n - 1$. We shall have to compare t (computed from the given sample) with $t_{\alpha/2, n-1}$, or t with $-t_{\alpha/2, n-1}$, or $|t|$ with $t_{\alpha/2, n-1}$ according as the alternatives of interest are $H: \mu > \mu_0$, $H: \mu < \mu_0$ or $H: \mu \neq \mu_0$.

In order to obtain confidence limits to μ , we see that

$$P\left[-t_{\alpha/2, n-1} \leq \frac{\sqrt{n}(x - \mu)}{s} \leq t_{\alpha/2, n-1}\right] = 1 - \alpha,$$

$$\text{i.e. } P\left[\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}\right] = 1 - \alpha$$

The $100(1 - \alpha)\%$ confidence limits to μ will, therefore, be $\bar{x} - t_{\alpha/2, n-1} \frac{s'}{\sqrt{n}}$

and $\bar{x} + t_{\alpha/2, n-1} \frac{s'}{\sqrt{n}}$, these being computed from the given sample.

In this case we may have also the problem of testing $H_0: \sigma = \sigma_0$ or the problem of obtaining confidence limits to σ .

From what has been said above, it is clear that

$$\frac{\sum (x_i - \bar{x})^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2}$$

is, under the hypothesis H_0 , a χ^2 with $n-1$ d.f. This provides us with tests for H_0 . The value of this χ^2 , computed from the given sample, is compared with $\chi^2_{\alpha, n-1}$ or $\chi^2_{1-\alpha, n-1}$, according as the alternatives are $H: \sigma > \sigma_0$ or $H: \sigma < \sigma_0$. For alternatives $H: \sigma \neq \sigma_0$, on the other hand, the computed value is to be compared with both $\chi^2_{1-\alpha/2, n-1}$ and $\chi^2_{\alpha/2, n-1}$. H_0 being rejected if the computed value is smaller than the former or exceeds the latter value.

Since

$$P\left[\chi^2_{1-\alpha/2, n-1} \leq \frac{(n-1)s'^2}{\sigma^2} \leq \chi^2_{\alpha/2, n-1}\right] = 1 - \alpha,$$

$$\text{i.e. } P\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \leq \sigma^2 \leq \frac{(n-1)s'^2}{\chi^2_{1-\alpha/2, n-1}}\right] = 1 - \alpha,$$

the confidence limits to σ^2 are

$$\frac{(n-1)s'^2}{\chi^2_{\alpha/2, n-1}} \text{ and } \frac{(n-1)s'^2}{\chi^2_{1-\alpha/2, n-1}}.$$

The confidence limits (with the same confidence coefficient, $1-\alpha$) to σ are, of course, the positive square-roots of these quantities.

Ex. 16.4 The following are 12 determinations of the melting point of a compound (in degrees centigrade) made by an analyst, the true melting point being 165°C.

Would you conclude from these data that his determinations are free from bias?

164.4	161.4
169.7	162.2
163.9	168.5
162.1	163.4
160.9	162.9
160.8	167.7

The determinations made by the analyst may be said to be unbiased if the mean determination in the population, that could be obtained if he took an infinite number of readings, can be supposed to be 165 degrees. We have, therefore, to test the null hypothesis $H_0 : \mu = 165$ against all alternatives $H : \mu \neq 165$.

It will be assumed (a) that the population distribution of determinations is of the normal type and (b) that the sample observations are random and mutually independent.

Under these assumptions, a test for H_0 is provided by the statistic

$$t = \frac{\sqrt{n}(\bar{x} - 165)}{s'},$$

which has $(n-1)$ d.f.

For the given observations,

$$n=12,$$

$$\sum_i u_i = 47.9$$

and

$$\sum_i u_i^2 = 292.83$$

(where $u_i = x_i - 160$).

Hence

$$x = 160 + \frac{47.9}{12} = 160 + 3.992 = 163.992 \text{ degrees}$$

and $s' = \sqrt{\frac{292.83 - 12 \times (3.992)^2}{11}} = \sqrt{\frac{292.83 - 191.2328}{11}}$
 $= \sqrt{9.2361} = 3.039 \text{ degrees,}$

so that

$$t = \frac{\sqrt{12}(163.992 - 165)}{3.039} = -\frac{3.464 \times 1.008}{3.039} = -1.149$$

From Table IV in the Appendix,

$$t_{0.025, 11} = 2.201$$

and $t_{0.005, 11} = 3.106$

Since for the given sample $|t|$ is smaller than both these tabulated values, H_0 is to be accepted at both the 1% and the 5% levels of significance. In other words, we find no reason to suppose that the analyst's determinations are not free from bias.

Ex 16.5 The weights at birth for 15 babies born in a Calcutta hospital are given below. Each figure is correct to the nearest tenth of a pound.

6.2	5.7	8.1
6.7	4.8	5.0
7.1	6.8	5.8
6.9	7.6	7.9
7.5	7.8	8.5

Give two limits between which the mean weight at birth for all such babies is likely to lie.

Let us denote by x the variable weight at birth per baby. Our problem here is then to find, on the basis of the given sample of 15 babies, confidence limits for the population mean of x . We shall assume (a) that in the population x is normally distributed (with a mean μ and a standard deviation σ , both of which are unknown) and (b) that the given sample observations are random and mutually independent.

Under these assumptions, the $100(1-\alpha)\%$ confidence limits to μ will be

$$\bar{x} - t_{\alpha/2, n-1} \cdot \frac{s'}{\sqrt{n}} \text{ and } \bar{x} + t_{\alpha/2, n-1} \cdot \frac{s'}{\sqrt{n}}.$$

In the present case,

$$n=15,$$

$$\sum_i x_i = 102.4$$

and

$$\sum_i x_i^2 = 716.88.$$

Hence

$$\bar{x} = 102.4/15 = 6.827 \text{ lb.}$$

$$\text{and } s' = \sqrt{\frac{716.88 - 15 \times (6.827)^2}{14}} = \sqrt{\frac{716.88 - 699.12}{14}} \\ = 1.126 \text{ lb.}$$

Again, consulting Table IV in the Appendix, we find that

$$t_{.005, 14} = 2.977.$$

Hence the 99% confidence limits to μ are

$$6.827 - 2.977 \times \frac{1.126}{\sqrt{15}} = 6.827 - \frac{3.352}{3.873} = 6.827 - 0.865 = 5.962 \text{ lb.}$$

and

$$6.827 + 0.865 = 7.692 \text{ lb.}$$

The confidence coefficient being as high as 0.99, one may well assert that the population mean lies between 5.962 lb. and 7.692 lb.

Ex. 16.6 A firm manufacturing rivets wants to limit variations in their length as much as possible. The lengths (in cm.) of 10 rivets manufactured by a new process are :

2.15	1.99	2.05	2.12	2.17
2.01	1.98	2.03	2.25	1.93

In the past, the standard deviation of length of rivets manufactured by the firm has been 0.145 cm. Examine whether the new process seems to be superior to the old.

If σ be the standard deviation of length for all rivets manufactured by the new process, then this may be considered superior if $\sigma < 0.145$. The null hypothesis is then $H_0 : \sigma = 0.145$, which is to be tested against the alternatives $H : \sigma < 0.145$.

Under the usual assumptions, the test will be given by the statistic

$$\frac{\sum(x_i - \bar{x})^2}{(0.145)^2}$$

which is, under H_0 , a χ^2 with $n-1$ d.f. For the present data,

$$\begin{aligned} \sum_i(x_i - \bar{x})^2 &= \sum_i(u_i - u)^2 = \sum_i u_i^2 - \frac{(\sum u_i)^2}{n} \quad [\text{putting } u = x - 2.00] \\ &= 0.1372 - \frac{(0.68)^2}{10} = 0.1372 - 0.04624 = 0.09096 \end{aligned}$$

Hence $\chi^2 = \frac{0.09096}{0.021025} = 4.326$

Now $\chi^2_{0.99, 1} = 2.088$, and $\chi^2_{0.95, 1} = 3.325$. The observed value is thus insignificant, and the null hypothesis is to be accepted, i.e., the new process does not seem to be superior to the old.

16.6 Comparison of two univariate normal distributions

Let the distribution of x in each of two populations be normal. Suppose the mean and the standard deviation of x for one population are μ_1 and σ_1 , while for the other they are μ_2 and σ_2 , respectively. Suppose further that $x_{11}, x_{12}, \dots, x_{1n_1}$ are random and independent observations obtained from the first population, and $x_{21}, x_{22}, \dots, x_{2n_2}$ are random and independent observations obtained from the second. The first set of observations is also supposed to be independent of the second set.

$$\bar{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j}/n_1$$

and $s_1' = \sqrt{\frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2 / (n_1 - 1)}$

are the mean and the standard deviation of x in the first sample.

$$\bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}/n_2$$

and $s_2' = \sqrt{\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2 / (n_2 - 1)}$

are the corresponding statistics for the second sample.

Case 1. μ_1, μ_2 unknown but σ_1, σ_2 known

In this case one may be concerned with a comparison between the population means. One may have to test the hypothesis that μ_1 and μ_2 differ by a specified quantity, say

$$H_0 : \mu_1 - \mu_2 = \xi_0,$$

or one may like to obtain confidence limits for the difference $\mu_1 - \mu_2$.

It may be seen that $\bar{x}_1 - \bar{x}_2$, being a linear function of normal variables, is itself normally distributed. It has mean

$$\begin{aligned} E(\bar{x}_1 - \bar{x}_2) &= E(\bar{x}_1) - E(\bar{x}_2) \\ &= \mu_1 - \mu_2 \end{aligned} \quad \dots \quad (16.8)$$

and variance

$$\begin{aligned} \text{var}(\bar{x}_1 - \bar{x}_2) &= \text{var}(\bar{x}_1) + \text{var}(\bar{x}_2) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \end{aligned} \quad \dots \quad (16.9)$$

the covariance term being zero since \bar{x}_1 and \bar{x}_2 are independent.

As such,

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)^{1/2}}$$

is distributed as a normal deviate.

To test

$$H_0 : \mu_1 - \mu_2 = \xi_0$$

we make use of the statistic

$$\frac{(\bar{x}_1 - \bar{x}_2) - \xi_0}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)^{1/2}}, \quad \dots \quad (16.10)$$

which is distributed as a normal deviate (τ) under H_0 . H_0 is to be rejected on the basis of the given samples if $\tau > \tau_\alpha$ or if $\tau < -\tau_\alpha$, according as the alternative hypotheses in which the experimenter is interested are $H : \mu_1 - \mu_2 > \xi_0$ or $H : \mu_1 - \mu_2 < \xi_0$. On the other hand, if the alternatives are $H : \mu_1 - \mu_2 \neq \xi_0$, H_0 is to be rejected when $|\tau| > \tau_{\alpha/2}$. In the commonest case, the null hypothesis will be $H_0 : \mu_1 = \mu_2$, for which $\xi_0 = 0$.

If the problem is one of interval estimation, then it will be found,

following the usual mode of argument, that the confidence limits to $\mu_1 - \mu_2$ (with confidence coefficient $1-\alpha$) are

$$(x_1 - x_2) - \tau_{\alpha/2} \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)^{1/2} \text{ and } (x_1 - x_2) + \tau_{\alpha/2} \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)^{1/2}$$

Case 2 μ_1, μ_2 known but σ_1, σ_2 unknown

Here it may be necessary to test the hypothesis that the ratio of the two unknown standard deviations has a specified value, say $H_0: \sigma_1/\sigma_2 = \xi_0$, or to set confidence limits to this ratio

Since

$$\sum_{j=1}^{n_1} (x_{1j} - \mu_1)^2 / \sigma_1^2 \text{ and } \sum_{j=1}^{n_2} (x_{2j} - \mu_2)^2 / \sigma_2^2$$

are independent χ^2 's with n_1 d.f. and n_2 d.f., respectively,

$$\frac{\sum_{j=1}^{n_1} (x_{1j} - \mu_1)^2 / n_1 \sigma_1^2}{\sum_{j=1}^{n_2} (x_{2j} - \mu_2)^2 / n_2 \sigma_2^2} \quad (16.11)$$

is distributed as an F with n_1, n_2 d.f.

Under the hypothesis $H_0: \sigma_1/\sigma_2 = \xi_0$ therefore,

$$\frac{\sum_{j=1}^{n_1} (x_{1j} - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2} (x_{2j} - \mu_2)^2 / n_2} \sim F_{1, \xi_0^2}$$

is an F with n_1, n_2 d.f. This provides a test for H_0 . When the alternatives are $H: \frac{\sigma_1}{\sigma_2} > \xi_0$, H_0 is to be rejected if for the given samples

$$F > F_{\alpha, n_1, n_2}$$

If the alternatives are $H: \frac{\sigma_1}{\sigma_2} < \xi_0$, H_0 is to be rejected if for the given samples

$$F < F_{1-\alpha, n_1, n_2},$$

i.e. if

$$\frac{1}{F} > F_{\alpha, n_2, n_1}$$

Lastly, when the alternatives are $H: \frac{\sigma_1}{\sigma_2} \neq \xi_0$, H_0 is to be rejected if the samples in hand give either

$$F < F_{1-\alpha/2, n_1, n_2}, \text{ i.e. } \frac{1}{F} > F_{\alpha/2, n_2, n_1},$$

or

$$F > F_{\alpha/2, n_1, n_2}$$

The commonest form of the null hypothesis will be $H_0: \sigma_1 = \sigma_2$, for which $\xi_0 = 1$, and here

$$F = \frac{\sum_j (x_{1j} - \mu_1)^2 / n_1}{\sum_j (x_{2j} - \mu_2)^2 / n_2}$$

simply.

For the purpose of getting confidence limits to σ_1/σ_2 , we see that

$$P\left[\frac{1}{F_{\alpha/2; n_1, n_2}} \leq \frac{\sum_j (x_{1j} - \mu_1)^2 / n_1}{\sum_j (x_{2j} - \mu_2)^2 / n_2} \leq F_{\alpha/2; n_1, n_2}\right] = 1 - \alpha,$$

$$\text{i.e. } P\left[\frac{1}{F_{\alpha/2; n_1, n_2}} \cdot \frac{\sum_j (x_{1j} - \mu_1)^2 / n_1}{\sum_j (x_{2j} - \mu_2)^2 / n_2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq F_{\alpha/2; n_1, n_2} \cdot \frac{\sum_j (x_{1j} - \mu_1)^2 / n_1}{\sum_j (x_{2j} - \mu_2)^2 / n_2}\right] = 1 - \alpha.$$

The confidence limits to $\frac{\sigma_1^2}{\sigma_2^2}$ (with confidence coefficient $1 - \alpha$) will, therefore, be

$$\frac{1}{F_{\alpha/2; n_1, n_2}} \cdot \frac{\sum_j (x_{1j} - \mu_1)^2 / n_1}{\sum_j (x_{2j} - \mu_2)^2 / n_2} \text{ and } F_{\alpha/2; n_1, n_2} \cdot \frac{\sum_j (x_{1j} - \mu_1)^2 / n_1}{\sum_j (x_{2j} - \mu_2)^2 / n_2}.$$

The corresponding limits to σ_1/σ_2 will naturally be the positive square-roots of these quantities.

Case 3. Means and standard deviations all unknown

We shall first consider methods of testing for the difference of the two means and of setting confidence limits to this difference.

(a) For the sake of simplicity, we shall assume that the two unknown standard deviations are equal. Now, if σ denotes the common standard deviation, then

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{1/2}}$$

is a normal deviate, while

$$\frac{(n_1 - 1)s'^2_1 + (n_2 - 1)s'^2_2}{\sigma^2} = \frac{\sum_j (x_{1j} - \bar{x}_1)^2}{\sigma^2} + \frac{\sum_j (x_{2j} - \bar{x}_2)^2}{\sigma^2},$$

which is the sum of two independent χ^2 s, one with $n_1 - 1$ d.f. and the

The problems that generally arise in relation to one univariate normal distribution and two univariate normal distributions have been discussed above. The more general case of k univariate normal distributions will be partly discussed in the next section.

Ex. 16.7 The following data give the lives in hours of 2 batches of electric lamps. Test whether there is a significant difference between the batches in respect of average length of life.

<u>Batch 1</u>	<u>Batch 2</u>
1,505	1,799
1,556	1,618
1,801	1,604
1,629	1,655
1,644	1,708
1,607	1,675
1,825	1,728
1,748	

Let us denote by μ_1 and μ_2 the average lives for lamps in the populations corresponding to Batch 1 and Batch 2, respectively. We have to test the null hypothesis $H_0 : \mu_1 = \mu_2$ against all alternatives $H : \mu_1 \neq \mu_2$.

We assume (1) that in each population the life of bulb is normally distributed, (2) that the unknown standard deviations of the two distributions are equal and (3) that the sample observations in each set are random and mutually independent, one set being independent of the other. The test for H_0 is then provided by the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s' \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which is a t with $n_1 + n_2 - 2$ d.f.

For the given samples,

$$\sum_j u_{1j} = 515, \quad \sum_j u_{2j} = 587,$$

$$\sum_j u_{1j}^2 = 126,717 \text{ and } \sum_j u_{2j}^2 = 76,639$$

(putting $u = x - 1,600$), and

$$\begin{aligned} \bar{x}_1 &= 1,600 + 515/8 \\ &= 1,664.375, \end{aligned}$$

$$\begin{aligned} \bar{x}_2 &= 1,600 + 587/7 \\ &= 1,683.857, \end{aligned}$$

$$\begin{aligned}
 s' &= \left\{ \frac{\left(\sum_j u_{1j}^2 - n_1 \bar{u}_1^2 \right) + \left(\sum_j u_{2j}^2 - n_2 \bar{u}_2^2 \right)}{n_1 + n_2 - 2} \right\}^{1/2} \\
 &= \left\{ \frac{(126,717 - 8 \times (64.375)^2) + (76,639 - 7 \times (83.857)^2)}{13} \right\}^{1/2} \\
 &= \left(\frac{120,978.900}{13} \right)^{1/2} = \sqrt{9,306.069} = 96.468,
 \end{aligned}$$

so that

$$\begin{aligned}
 t &= \frac{1664.375 - 1683.857}{96.468 \sqrt{\left(\frac{1}{8} + \frac{1}{7}\right)}} = -\frac{19.482}{96.468 \times 0.5175} \\
 &= -0.390.
 \end{aligned}$$

Since $t_{.025, 13} = 2.160$ and $t_{.005, 13} = 3.012$, the observed t leads to the acceptance of H_0 . In other words, observed difference in mean life is found to be insignificant.

Ex. 16.8 Two experimenters, A and B , take repeated measurements on the length of a copper wire. On the basis of the data obtained by them, which are given below, test whether B 's measurements are more accurate than A 's. (It may be supposed that the readings taken by both are unbiased.)

A 's measurements (in mm.)		B 's measurements (in mm.)	
12.47	12.44	12.06	12.34
11.90	12.13	12.23	12.46
12.77	11.86	12.46	12.39
11.96	12.25	11.98	
12.78	12.29	12.22	

Since the readings of both the experimenters are unbiased, B 's measurements may be considered more accurate if they have a smaller population standard deviation than A 's measurements. The null hypothesis is then

$$H_0 : \sigma_1 = \sigma_2,$$

to be tested against the alternatives $H : \sigma_1 > \sigma_2$.

Under the usual assumptions, the test is given by

$$F = \frac{s'_1{}^2}{s'_2{}^2} \text{ with } n_1 - 1 \text{ and } n_2 - 1 \text{ d.f.}$$

Here

$$\begin{aligned}s_1^2 &= \frac{1}{n_1-1} \left\{ \sum_j x_{1j}^2 - n_1 \bar{x}_1^2 \right\} \\&= \frac{1}{n_1-1} \left\{ \sum_j u_{1j}^2 - n_1 \bar{u}_{1j}^2 \right\} = \frac{1}{n_1-1} \left\{ \sum_j u_{1j}^2 - \frac{(\sum u_{1j})^2}{n_1} \right\}\end{aligned}$$

and $s_2^2 = \frac{1}{n_2-1} \left\{ \sum_j u_{2j}^2 - \frac{(\sum u_{2j})^2}{n_2} \right\}$

where we take $u=x-12.00$

For the given samples,

$$\sum_j u_{1j} = 2.85 \quad \sum_j u_{2j} = 2.14,$$

$$\sum_j u_{1j}^2 = 1.8105, \quad \sum_j u_{2j}^2 = 0.7962,$$

$$s_1^2 = \frac{1}{9} \left\{ 1.8105 - \frac{(2.85)^2}{10} \right\} = \frac{1}{9} \{ 1.8105 - 0.8122 \} = 0.1109$$

and $s_2^2 = \frac{1}{7} \left\{ 0.7962 - \frac{(2.14)^2}{8} \right\} = \frac{1}{7} \{ 0.7962 - 0.5724 \} = 0.03197$

Hence

$$F = \frac{0.1109}{0.03197} = 3.469 \text{ (with 9 and 7 d.f.)}$$

The tabulated values are

$$F_{0.05, 9, 7} = 3.68 \text{ and } F_{0.01, 9, 7} = 6.72$$

The observed F is thus insignificant at both the levels and H_0 therefore, should be accepted. Thus we find no reason to suppose that B 's measurements are more accurate than A 's.

16.7 Comparison of means of more than two normal populations

Suppose there are k populations in each of which the variable x is normally distributed. Let μ_i ($i=1, 2, \dots, k$) be the unknown mean of x in the i th population. We want to test whether the k means may be supposed to be equal. Thus our null hypothesis is

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k,$$

which is to be tested against all alternatives

The standard deviations, assumed unknown, will be supposed to be equal, the common value being denoted by σ .

Let a random sample of independent observations be taken from each population, the size of the sample from the i th population being n_i (≥ 2 for at least one i). The observations from the i th population may be denoted by

$$x_{i1}, x_{i2}, \dots, x_{in_i}.$$

Let $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$,

the i th sample mean, and let

$$\bar{x} = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}/n,$$

the grand mean, where

$$n = \sum_{i=1}^k n_i.$$

Now, consider the sum of squares of the deviations of the observations from the grand mean, to be called the 'total SS' :

$$\text{total } SS = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2. \quad \dots \quad (16.14)$$

We can write

$$\begin{aligned} \text{total } SS &= \sum_{i=1}^k \sum_{j=1}^{n_i} \{(\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)\}^2 \\ &= \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2. \quad \dots \quad (16.15) \end{aligned}$$

The first component represents the sum of (weighted) squares of deviations of the sample means from the grand mean, which may also be looked upon as representing the extent to which the sample means differ among themselves. This is called the 'SS between groups' (SSB).

The second component, on the other hand, uses deviations of values within sample from the sample mean and is called the 'SS within groups' (SSW).

It can be shown that

$$E(SSW) = (n - k)\sigma^2, \quad \dots \quad (16.16)$$

while $E(SSB) = (k - 1)\sigma^2 + \sum_{i=1}^k n_i (\mu_i - \bar{\mu})^2, \quad \dots \quad (16.17)$

where $\bar{\mu} = \sum_{i=1}^k n_i \mu_i / n.$

If we put

$$MSB = SSB/(k-1) \quad (161)$$

and

$$MSW = SSW/(n-k), \quad (162)$$

to be called the 'mean square between groups' and the 'mean-square within groups', respectively, then

$$E(MSB) = \sigma^2 + \frac{1}{k-1} \sum_{i=1}^k n_i (\mu_i - \mu)^2 = \sigma_1^2, \text{ say}, \quad (163)$$

and $E(MSW) = \sigma^2$

$$\text{Also, } \sum_{i=1}^k n_i (\mu_i - \mu)^2 = 0$$

if, and only if, all the means μ_i are equal. Otherwise, it is positive. Hence our problem reduces to the problem of testing

$$H_0: \sigma_1^2 = \sigma^2$$

against the alternatives

$$H: \sigma_1^2 > \sigma^2,$$

which is similar to the problem posed in Section 16.6 regarding the equality of two variances or standard deviations.

The test is given by

$$F = \frac{MSB}{MSW}, \quad (164)$$

which is, under H_0 , distributed as an F statistic with $(k-1)$ and $(n-k)$ d.f. We reject H_0 if

$$F > F_{\alpha, (k-1), (n-k)}$$

and accept it otherwise, α being the chosen level of significance.

The process of splitting the total SS into independent components like SSB and SSW , which can be attributed to different sources of variation, is called an *analysis of variance*, and is generally put in tabular form.

Computational procedure for the analysis of variance

(1) Calculate the total for each group $T_{10}, T_{20}, \dots, T_{10}$,

where

$$T_{10} = \sum_{j=1}^{n_i} x_{ij}$$

(2) Calculate the grand total $T_{00} = \sum_i \sum_j x_{ij} = \sum_i T_{10}$

(3) Calculate the raw total SS $\sum_i \sum_j x_{ij}^2$

- (4) Calculate $\sum_i T_{i0}^2/n_i$.
- (5) Calculate correction factor : T_{00}^2/n .
- (6) Total SS = $\sum_i \sum_j x_{ij}^2/n - T_{00}^2/n$ = value obtained in step (3)—that in step (5).
- (7) SSB = $\sum_i T_{i0}^2/n_i - T_{00}^2/n$ = value obtained in step (4)—that in step (5).
- (8) SSW = total SS - SSB = value in step (6)—that in step (7).

It may be noted that sometimes calculations may be simplified by making a change of base and scale of the observations. This will not affect the test, for the F statistic defined by (16.20) remains unaltered under such transformations.

Ex. 16.9 The weights in gm. of a number of copper wires, each of length 1 metre, are obtained. These are shown below classified according to the dies from which the wires come :

Die No.				
I	II	III	IV	V
1.30	1.28	1.32	1.31	1.30
1.32	1.35	1.29	1.29	1.32
1.36	1.33	1.31	1.33	1.30
1.35	1.34	1.28	1.31	1.33
1.32		1.33	1.32	
1.37		1.30		

Test the hypothesis that there is no difference between the mean weights of wires coming from the different dies.

To test the hypothesis we shall assume that the distribution of the weight of wire (x) is normal for each die and that the variances for different dies are equal.

Let x_{ij} be the weight of the j th copper wire coming from the i th die. Taking the origin at 1.28 gm. and the unit as 0.01 gm., our new variable is

$$u = 100(x - 1.28),$$

and so its value for the j th wire from the i th die is

$$u_{ij} = 100(x_{ij} - 1.28).$$

TABLE 16 I
WEIGHTS OF COPPER WIRES AFTER CHANGE OF
BASE AND SCALE

Die No				
I	II	III	IV	V
2	0	4	3	2
4	7	1	1	4
8	5	3	5	2
7	6	0	3	5
4		5	4	
9		2		
Total	34	18	15	13

Here $T_{00} = \sum \sum u_{ij} = 96,$

$$\sum \sum u_{ij}^2 = 504,$$

$$\sum \frac{T_{ij}^2}{n_i} = \frac{34^2}{6} + \frac{18^2}{4} + \frac{15^2}{6} + \frac{16^2}{5} + \frac{13^2}{4}$$

$$= 192.6667 + 81 + 37.5 + 51.2 + 42.25 \\ = 404.6167$$

and correction factor $= \frac{T_{00}^2}{n} = \frac{96^2}{25} = 368.64$

Hence

$$\text{total } SS = \sum \sum u_{ij}^2 - T_{00}^2/n$$

$$= 504 - 368.64 = 135.36,$$

$$SSB (\text{e.g. } SS \text{ due to dies}) = \sum T_{ij}^2/n_i - T_{00}^2/n$$

$$= 404.6167 - 368.64$$

$$= 35.9767$$

and

$$SSW = \text{total } SS - SSB$$

$$= 135.36 - 35.9767 = 99.3833$$

TABLE 16.2
ANALYSIS OF VARIANCE FOR THE DATA ON WEIGHTS OF
COPPER WIRES

Source of variation	d.f.	SS	MS	F	F at level 1% 5%
Between groups	4	35.9767	8.9942	1.81	4.43 2.87
Within groups	20	99.3833	4.9692		
Total	24	135.3600	—	—	—

The observed F , being less than $F_{0.05; 4, 20}$ and $F_{0.01; 4, 20}$, is insignificant at both levels of significance. Thus the hypothesis under test may be accepted ; i.e., there seems to be no reason to suppose that (population) mean weights of copper wires for different dies are unequal.

16.8 Problems relating to a bivariate normal distribution

Suppose in a given population the variables x and y are distributed in the bivariate normal form with means μ_x and μ_y , standard deviations σ_x and σ_y , and correlation coefficient ρ . Let (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) be the values of x and y observed in a sample of size n drawn from this population. We shall suppose that the n pairs of sample observations are random and independent. We shall also assume that all the parameters are unknown.

(a) Test for correlation coefficient

Here the sample correlation coefficient is

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\left\{ \sum_i (x_i - \bar{x})^2 \right\}^{1/2} \left\{ \sum_i (y_i - \bar{y})^2 \right\}^{1/2}}$$

where \bar{x} and \bar{y} are the sample means. When $\rho=0$, the sampling distribution of r assumes a simple form (*vide Exercise 16.5*), and in that case

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \dots \quad (16.21)$$

can be shown to be distributed as a t with $n-2$ d.f. This fact provides us with a test for $H_0 : \rho=0$. As to the general hypothesis

$H_0: \rho = \rho_0$, an exact test becomes difficult, because for $\rho \neq 0$ the sample correlation has a complicated sampling distribution. An approximate test, which may be used for even moderately large n will be given in Chapter 17.

Ex 16.10 The correlation coefficient between nasal length and stature for a group of 20 Indian adult males was found to be 0.203. Test whether there is any correlation between the characters in the population.

The null hypothesis here is $H_0: \rho = 0$, to be tested against all alternatives. As we have seen, under certain assumptions the test is given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which has $n-2$ d.f.

Here

$$t = \frac{0.203\sqrt{18}}{\sqrt{1-(0.203)^2}} = \frac{0.203 \times 4.243}{\sqrt{1-0.0409}} = \frac{0.8613}{0.9792} = 0.880$$

The tabulated values are

$$t_{0.025, 18} = 2.101 \text{ and } t_{0.05, 19} = 2.878$$

The observed value is, therefore, insignificant at both the levels, i.e., the population correlation may be supposed to be zero.

(b) *Problems regarding the difference between μ_x and μ_y .*

Information regarding the difference between the means μ_x and μ_y may be of some importance when x and y are variables measured in the same units.

To begin with, we note that if we take a new variable

$$z = x - y,$$

then this z , being a linear function of normal variables, is itself normally distributed with mean

$$\mu_z = \mu_x - \mu_y,$$

and variance

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y,$$

It will follow, from what we have said in Section 16.5, that if we put $z_i = x_i - y_i$, $z = \sum_i z_i/n$ and $s_z^2 = \frac{1}{n-1} \sum_i (z_i - z)^2$, then

$$\frac{\sqrt{n}(z - \mu_z)}{s_z} \quad (16.22)$$

will be distributed as a t with $n-1$ d.f. This will provide us with a test for $H_0 : \mu_x - \mu_y = \zeta_0$, which is equivalent to $H_0 : \mu_z = \zeta_0$, and with confidence limits to the difference $\mu_z = \mu_x - \mu_y$. The statistic (16.22) is often referred to as a *paired t*.

We may, instead, be interested in the ratio $\mu_x/\mu_y = \eta$ (say). In this case we shall take

$$z = x - \eta y,$$

which again is normally distributed with mean

$$\mu_z = \mu_x - \eta \mu_y = 0.$$

Hence the statistic

$$t = \frac{\sqrt{n} \bar{z}}{s_z'}$$

is distributed as a t (i.e. paired t) with $n-1$ d.f. This can be used for testing the hypothesis $H_0 : \mu_x/\mu_y = \eta_0$ or for setting confidence limits to the ratio μ_x/μ_y .

Ex. 16.11 The weights of ten boys before they are subjected to a change of diet and after a lapse of six months are recorded below :

Serial No.	Weight (in lb.)	
	Before	After
1	109	115
2	112	120
3	98	99
4	114	117
5	102	105
6	97	98
7	88	91
8	101	99
9	89	93
10	91	89

Test whether there has been any significant gain in weight as a result of the change of diet.

If we denote by y and x the weight of a boy before and after the change of diet, then the hypothesis to be tested is $H_0 : \mu_x = \mu_y$, the alternatives being $H : \mu_x > \mu_y$.

Under the assumptions (1) that x and y are jointly normally distributed and (2) that the pairs of values of x and y in the sample are random and independent, the test for H_0 is given by

$$t = \frac{\sqrt{n}\bar{z}}{s_z}$$

with $n-1$ d.f., where $z=x-y$

For the given sample,

the values of z are

$$6 \quad \text{Hence } z=25/10=2.5,$$

8

$$1 \quad s_z = \sqrt{\frac{153 - 10 \times (2.5)^2}{9}}$$

3

$$3 \quad = \sqrt{\frac{90.5}{9}} = \sqrt{10.0556} = 3.171$$

1

$$-2 \quad \text{and } t = \frac{\sqrt{10} \times 2.5}{3.171} = \frac{3.162 \times 2.5}{3.171}$$

4

$$-2 \quad = 2.493$$

Now, $t_{0.05, 9}=1.833$ and $t_{0.01, 9}=2.821$. The observed value is thus significant at the 5% but insignificant at the 1% level of significance. If we choose the 5% level, then the null hypothesis should be rejected and we should say that the change of diet results in a gain in average weight.

(c) *Problems regarding the ratio σ_x/σ_y ,*

When x and y are variables measured in identical units, one may also be interested in the ratio σ_x/σ_y . Let us denote this ratio by ξ . If we consider the new variables

$$u=v+\xi y$$

$$\text{and} \quad v=x-\xi y,$$

then u and v are jointly normally distributed, like x and y , and

$$\text{cov}(u, v)=\sigma_u^2-\xi^2\sigma_y^2=0$$

Thus u and v are uncorrelated normal variables

In going to test for the hypothesis

$$H_0 : \sigma_x/\sigma_y = \xi_0,$$

we shall, therefore, take two new variables

$$u = x + \xi_0 y$$

and

$$v = x - \xi_0 y$$

and shall instead test for the equivalent hypothesis $H_0 : \rho_{uv} = 0$. This test will be given by the statistic

$$t = \frac{r_{uv}\sqrt{n-2}}{\sqrt{1-r_{uv}^2}} \quad \dots \quad (16.23)$$

with $n-2$ d.f., r_{uv} being the sample correlation between u and v .

To have confidence limits for ξ , we utilise the fact that, with

$$u = x + \xi y$$

and

$$v = x - \xi y,$$

$$P\left[\frac{|r_{uv}| \sqrt{n-2}}{\sqrt{1-r_{uv}^2}} \leq t_{\alpha/2, n-2}\right] = 1-\alpha$$

$$\text{or } P\left[\frac{r_{uv}^2(n-2)}{1-r_{uv}^2} \leq t_{\alpha/2, n-2}^2\right] = 1-\alpha.$$

By solving the equation

$$r_{uv}^2(n-2) = t_{\alpha/2, n-2}^2(1-r_{uv}^2)$$

or, say, $\psi(\xi) = 0$

for the unknown ratio $\xi = \sigma_x/\sigma_y$, two roots will be obtained. In case the roots, say ξ_1 and ξ_2 , are real ($\xi_1 < \xi_2$), these will be the required confidence limits for ξ with confidence coefficient $1-\alpha$.

Again, $\psi(\xi)$ may be either a convex or a concave function. In the former case we shall say $\xi_1 \leq \xi \leq \xi_2$, while in the latter we shall say $0 \leq \xi \leq \xi_1$ or $\xi_2 \leq \xi < \infty$.

But the roots may as well be imaginary, in which case we shall say that for the given sample the $100(1-\alpha)\%$ confidence limits do not exist.

16.9 Problems relating to simple regression

Let x and y be two variables such that the conditional distribution of y for each value of x is normal, with mean

$$\eta_x = \alpha + \beta x \text{ (say)}$$

—which means that the population regression of y on x is linear—and a constant variance σ^2 .

Let x_i and y_i , ($i=1, 2, \dots, n$) denote the sample values of x and y for a sample of size n . The pairs of values are supposed to be random and independent.*

The sample (least-square) regression line will be given by

$$Y = a + bx,$$

where $b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2}$

and $a = \bar{y} - b\bar{x}$

In order to study the sampling distributions of a , b and Y , we shall consider only those samples of size n in which the values of x_1, x_2, \dots, x_n are the same as those observed in the given sample. This restriction implies that x_i is not treated as a variable, once the first sample of size n has been selected; only y_i is supposed to vary from sample to sample. In such samples, b will be looked upon as a linear function of y 's:

$$b = \sum_i w_i y_i,$$

where $w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}.$

Being a linear function of y 's, which are normal variables, b is itself a normal variable; and

$$\begin{aligned} E(b) &= \sum_i w_i E(y_i) = \sum_i w_i (\alpha + \beta x_i) \\ &= \alpha \sum_i w_i + \beta \sum_i w_i x_i = \beta \end{aligned} \quad \dots \quad (16.24)$$

[since $\sum_i w_i = 0$

$$\text{and } \sum_i w_i x_i = \frac{\sum (x_i - \bar{x}) x_i}{\sum (x_i - \bar{x})^2} = 1],$$

and

$$\text{var}(b) = \sum_i w_i^2 \text{var}(y_i) \quad [\text{the } y \text{'s being independent}]$$

$$= \sigma^2 \sum_i w_i^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}. \quad \dots \quad (16.25)$$

* It is also assumed that the values of x are not all equal, so that $\sum_i (x_i - \bar{x})^2 > 0$

Similarly, a , which is a linear function of the normal variables y_i , is itself normally distributed with

$$E(a) = E(\bar{y}) - E(b) \cdot \bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha$$

and $\text{var}(a) = \text{var}(\bar{y}) - 2\bar{x}\text{cov}(\bar{y}, b) + \bar{x}^2\text{var}(b)$ (16.26)

Now,

$$\text{var}(\bar{y}) = \frac{\sigma^2}{n},$$

$$\text{var}(b) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2},$$

while

$$\begin{aligned} \text{cov}(\bar{y}, b) &= \frac{1}{n} \text{cov}\left(\sum_i y_i, \sum_i w_i y_i\right) \\ &= \frac{1}{n} \sum_i w_i \text{var}(y_i) \\ &= \frac{\sigma^2}{n} \sum_i w_i = 0, \end{aligned}$$

so that

$$\text{var}(a) = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right\}. \quad \dots \quad (16.27)$$

Again, for any given x , the estimated value (\hat{Y}) of y , being a linear function of a and b , is normal with

$$E(\hat{Y}) = E(a) + E(b) \cdot x = \alpha + \beta x = \eta_x \quad \dots \quad (16.28)$$

and $\text{var}(\hat{Y}) = \text{var}(\bar{y} + b(x - \bar{x})) = \text{var}(\bar{y}) + (x - \bar{x})^2\text{var}(b)$

$$= \frac{\sigma^2}{n} + \frac{(x - \bar{x})^2 \sigma^2}{\sum_i (x_i - \bar{x})^2} = \sigma^2 \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right\}. \quad \dots \quad (16.29)$$

Hence if σ^2 be known, tests and confidence limits for α and β can be obtained considering

$$\frac{a - \alpha}{\sigma \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right\}^{1/2}} \quad \dots \quad (16.30)$$

and $\frac{(b - \beta) \sqrt{\sum_i (x_i - \bar{x})^2}}{\sigma} \quad \dots \quad (16.31)$

to be normal deviates. Similarly, to test whether, corresponding to a given value of x , η_x has a specified value or to set confidence limits

to η_x , we can use the statistic

$$\frac{Y - \eta_x}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} \quad (16.32)$$

as a normal deviate

In case σ^2 is unknown, one would use its unbiased estimate*

$$s_{\eta_x}^2 = \frac{\sum (y_i - Y_i)^2}{n-2} = \frac{\sum y_i^2 - a \sum y_i - b \sum x_i y_i}{n-2} \quad (16.33)$$

For tests and confidence limits for α , β and η_x , the statistics

$$\frac{\alpha - \alpha}{s_{\eta_x} \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum (x_i - \bar{x})^2}}} \quad (16.34)$$

$$\frac{(b - \beta) \sqrt{\sum (x_i - \bar{x})^2}}{s_{\eta_x}} \quad (16.35)$$

and

$$\frac{Y - \eta_x}{s_{\eta_x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} \quad (16.36)$$

each of which is a t with $n-2$ d.f., will be used.

By using the above results, we can also compare two regression equations, say, $\eta_x = \alpha + \beta x$ and $\eta_x = \alpha' + \beta' x$, i.e. we can compare α and α' or β and β' . The method will be similar to that used in comparing the means of two univariate normal distributions.

In this connection, one may also consider the problem of predicting, by giving suitable limits based on the regression line, the value of y corresponding to a given value of x . For given x , $y - Y$ is a normal variable with

$$E(y - Y) = \eta_x - \eta_x = 0$$

and

$$\text{var}(y - Y) = \text{var}(y) + \text{var}(Y)$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right],$$

so that

$$P \left[Y - \tau_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \leq y \leq Y + \tau_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right] = 1 - \alpha \quad (16.37)$$

*Here it is assumed that $n \geq 3$

If σ be known, then the prediction limits to y (with "confidence coefficient" $1-\alpha$) will be :

$$Y \mp \tau_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}.$$

In case σ is unknown, it will be replaced by $s'_{y,x}$ and $\tau_{\alpha/2}$ will be replaced by $t_{\alpha/2, n-2}$.

Ex. 16.12 For 20 pairs of fathers and sons, the regression equation of height of son (y) on height of father (x), both measured in inches, was found to be

$$Y = 3.66 + 0.932x.$$

Test whether a differs significantly from zero and b differs significantly from unity. For the given data, $\bar{x} = 66.21$, $\sum_i (x_i - \bar{x})^2 = 120.56$ and $\sum_i (y_i - \bar{y})^2 = 145.61$.

For making the test, we assume that the population regression of y on x is linear :

$$\eta_x = \alpha + \beta x \text{ (say).}$$

We have then to test $H_0 : \alpha = 0$ and $H_0 : \beta = 1$. We make the usual assumptions regarding the conditional distribution of y for given x .

Here

$$\begin{aligned} s'_{y,x}^2 &= \frac{\sum_i (y_i - Y_i)^2}{n-2} = \frac{\sum_i (y_i - \bar{y})^2 - b^2 \sum_i (x_i - \bar{x})^2}{n-2} \\ &= \frac{145.61 - (0.932)^2 \times 120.56}{18} \\ &= \frac{145.61 - 104.72}{18} = 2.272. \end{aligned}$$

Hence for testing $H_0 : \alpha = 0$, we have

$$\begin{aligned} t &= \frac{\alpha - 0}{s'_{y,x} \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right\}^{1/2}} \quad (\text{with } 18 \text{ d.f.}) \\ &= \frac{3.66}{(2.272)^{1/2} \left\{ \frac{1}{20} + \frac{4,383.76}{120.56} \right\}^{1/2}} \\ &= \frac{3.66}{(2.272)^{1/2} (0.05 + 36.36)^{1/2}} = \frac{3.66}{1.507 \times 6.034} = 0.403. \end{aligned}$$

For testing $H_0: \beta=1$, we have

$$t = \frac{(b-1)\sqrt{\sum_i(x_i-\bar{x})^2}}{s_{y-x}} \quad (\text{with } 18 \text{ d.f.})$$

$$= \frac{(0.932-1)\sqrt{120.56}}{1.507} = -\frac{0.068 \times 10.98}{1.507}$$

$$= -0.495$$

The table of the t distribution gives

$$t_{0.025, 18} = 2.101 \text{ and } t_{0.005, 18} = 2.878$$

The observed value of each t is thus insignificant at both 1% and 5% significance levels. Thus the observed a does not differ significantly from 0, nor does the observed b differ significantly from 1.

16.10 Tests for multiple and partial correlation coefficients

Many problems regarding the joint distribution of more than two variables may also be solved by using the four basic distributions considered at the beginning of this chapter.

We shall consider below the problems of testing for a multiple correlation coefficient and for a partial correlation coefficient. Let x_1, x_2, \dots, x_p be p variables whose joint distribution is of the p -variate normal form. Let $\rho_{123\dots p}$ and $\rho_{123\dots p}$ be the population multiple correlation of x_1 on x_2, x_3, \dots, x_p , and the population partial correlation of x_1 and x_2 , eliminating the effects of x_3, x_4, \dots, x_p . Further, suppose $r_{123\dots p}$ and $r_{123\dots p}$ are the corresponding sample coefficients based on a sample of size $n (\geq p+1)$.

It can be shown that if $\rho_{123\dots p}=0$, then the statistic

$$\frac{r_{123\dots p}^2/(p-1)}{(1-r_{123\dots p}^2)/(n-p)} \quad (16.38)$$

is distributed as an F with $p-1$ and $n-p$ d.f. This statistic, therefore, supplies a test for

$$H_0: \rho_{123\dots p}=0$$

Similarly,

$$\frac{r_{1234\dots p}\sqrt{n-p}}{\sqrt{1-r_{1234\dots p}^2}}, \quad (16.39)$$

which is a t with $n-p$ d.f. when $\rho_{1234\dots p}=0$, supplies a test for $H_0: \rho_{1234\dots p}=0$.

Ex. 16.13 60 students are examined in statistics, physics and mathematics. The total correlations between the scores obtained are

$$r_{12}=0.64, r_{13}=0.75 \text{ and } r_{23}=0.82$$

(where x_1 , x_2 and x_3 are taken to denote scores in statistics, physics and mathematics, respectively). It is conjectured that the correlation between x_1 and x_2 is due to the influence of x_3 on both x_1 and x_2 . Test whether this conjecture seems valid in the light of the above data.

The conjecture may be said to be valid if the population partial correlation coefficient of x_1 and x_2 , eliminating the effect of x_3 from both, is zero. Hence we have to test the hypothesis $H_0: \rho_{12 \cdot 3} = 0$ against all alternatives.

Under the assumptions of normality of x_1 , x_2 , x_3 and of randomness and independence of the sample observations, the statistic to be used for making the test is

$$t = \frac{r_{12 \cdot 3} \sqrt{n-3}}{\sqrt{1-r_{12 \cdot 3}^2}} \text{ with } n-3=57 \text{ d.f.}$$

Here

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{0.0250}{\sqrt{0.4375}\sqrt{0.3276}} = 0.0661.$$

Hence

$$\begin{aligned} t &= \frac{0.0661 \sqrt{57}}{\sqrt{0.9956}} \\ &= \frac{0.4990}{0.9978} = 0.500. \end{aligned}$$

On comparing this with the tabulated values

$$t_{.025, 57} = 2.003$$

and

$$t_{.005, 57} = 2.657,$$

we find that H_0 is to be accepted. In other words, the conjecture that the correlation between x_1 and x_2 is due to the effect of x_3 upon them seems to be borne out by the data.

16.11 The normality assumption

The tests and confidence intervals for different parameters discussed in Sections 16.2—16.10 have been derived under the assumption that the underlying distributions are of the normal type. By assuming that the underlying distributions are of a different type, naturally a different set of tests and confidence limits would be obtained.

The above results are, therefore, strictly valid in sampling from normally distributed populations. Some work has been done to have an idea as to whether they hold for other types of distribution as well. It has been found that provided the population distribution diverges only slightly from normality, the results given remain valid to a large extent. However, if in any case the population distribution departs markedly from normality, the above methods should not be used. Some methods for dealing with such situations will be discussed in Chapter 18.

Questions and exercises

16.1 Suggest some exact test procedures for hypotheses concerning the parameter p of a binomial distribution and the parameter λ of a Poisson distribution.

16.2 Describe how one may test for association between two attributes in case the sample size is not large.

16.3 Why is it that, for data of somewhat similar types, we sometimes use Fisher's t test and at some other times the paired t test? What type of test would you use in case the assumption of homoscedasticity underlying Fisher's t test is untenable?

16.4 How do you test for, or set confidence limits to, the ratio of two variances? Consider separately the case of two univariate normal distributions and that of a bivariate normal distribution.

16.5 The sample correlation coefficient r , for sampling from a bivariate normal population with $\rho=0$, has the p.d.f.

$$f(r) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} (1-r^2)^{(n-4)/2}, \quad -1 < r < 1$$

Show that the statistic $r\sqrt{n-2}/\sqrt{1-r^2}$ has the t distribution with $n-2$ d.f.

Explain how this statistic can be used to test the hypothesis $H_0: \rho=0$.

16.6 Write a note on tests of significance for correlation coefficients--total, multiple and partial, especially touching upon the question of degrees of freedom of the test statistics.

16.7 For bivariate data, how does one test for, or set confidence limits to, the array mean of one variable (y) for a given value of the other (x) ? Describe also the method of predicting the value of y for a given value of x .

16.8 24 guinea-pigs are suffering from a disease. In order to study the therapeutic value of a serum, it is administered to a randomly selected group of 15 guinea-pigs, the remaining 10 being left untreated. The results are as follows :

	<i>Treated</i>	<i>Untreated</i>
<i>Recovered</i>	9	3
<i>Died</i>	6	7

Would you consider the serum beneficial ?

16.9 A 5-foot specimen of a new type of fibre is found to have 13 defects, while the manufacturer claims that there are no more than 150 defects per 100 feet. Do the above data support this claim ?

16.10 The mean I. Q. for a group of 25 children is 108.481, the standard deviation being 17.255. Test whether the observed mean is significantly greater than 100. *Partial ans.* $t=2.458$.

16.11 The mean yield per plant for 11 tomato plants of a particular variety was found to be 1,284.73 gm. with a standard deviation of 96.41 gm. Set up 99% confidence limits to the mean yield of all plants of this variety. *Ans.* 1,192.61 and 1,376.85 gm.

16.12 For the data of *Exercise 16.11*, obtain 99% confidence limits to the population standard deviation of yield of plants.

Ans. 60.75 and 207.67 gm.

16.13 The marks obtained by 20 students of College *A* and 15 students of College *B* in a mathematics test are given below :

<i>College A</i>				<i>College B</i>			
89	71	47	29	79	12	22	
76	84	81	49	61	55	90	
63	97	32	73	36	81	76	
69	88	43	80	42	35	67	
55	52	86	44	50	73	62	

Do you think students of College *A* are more proficient in mathematics than students of College *B* ? *Partial ans.* $t=1.275$.

16.14 15 bars of steel produced by Process I have mean breaking strength 46.2 with $\pm d$ 8.7, while 12 bars produced by Process II have mean breaking strength 57.5 with $\pm d$ 10.6. There is not enough ground to suppose that the population $\pm d$ s are equal. Test whether the population means may be supposed to be equal.

16.15 It is known that the mean diameters of rivets produced by two firms, I and II, are practically the same, but the standard deviations may differ. For 22 rivets produced by Firm I the standard deviation is 2.9 mm, while for 16 rivets manufactured by Firm II the standard deviation is 3.8 mm. Do you think that products of Firm I are of a better quality than those of Firm II? Partial ans. $t=1.72$

16.16 The additional hours of sleep gained after using each of 2 drugs by the same group of 12 patients are given below.

<i>Patient</i>	<i>Additional hours of sleep</i>	
	<i>Drug 1</i>	<i>Drug 2</i>
1	2.1	3.6
2	0.2	4.7
3	0.9	1.8
4	3.8	5.5
5	3.5	4.6
6	0.2	0.3
7	-1.3	-0.4
8	-0.3	1.9
9	-1.7	2.0
10	0.7	1.7
11	0.8	2.1
12	1.3	1.1

Test whether the second drug gives on the average, at least an hour more of sleep than the first drug. Partial ans. $t=1.633$

16.17 For the data of Exercise 16.16, judge whether the standard deviations of the two series are significantly different.

Partial ans. $|t|=0.320$

16.18 The correlation coefficient between head length and stature for a sample of 36 members of an Indian tribe has been found to be 0.4339. Is it reasonable to assume that in the population the characters are uncorrelated? Partial ans. $t=2.807$

16.19 The age in years (x) and chest-girth in inches (y) were recorded for two groups of school-boys consisting of 15 and 18 boys, respectively. On the basis of these data, the following values were obtained :

	<i>Group 1</i>	<i>Group 2</i>
$\sum x_i$	202.7	244.1
$\sum y_i$	23.3	53.8
$\sum x_i^2$	2,742.56	3,314.01
$\sum y_i^2$	44.77	174.40
$\sum x_i y_i$	315.07	729.82

Determine for each group the linear regression equation of y on x . Hence examine if the corresponding population regression equations (assumed linear) may be supposed to be identical or parallel.

16.20 For the data of Ex. 12.1, the multiple correlation of weight of dry bark (x_1) on height (x_2) and girth at a height of 6" (x_3) was found to be 0.854. Test whether this value may be supposed to have arisen in sampling from a population where the multiple correlation is zero.

Partial ans. $F=20.22$.

16.21 Under assumptions similar to those made in Section 16.9, show that $E(b_{12\cdot34\dots p})=\beta_{12\cdot34\dots p}$ and $\text{var}(b_{12\cdot34\dots p})=S^{22}\sigma^2$, where $(S^{ij})=(S_{ij})^{-1}$ with $i, j=2, 3, \dots, p$. Hence suggest a test for a hypothesis concerning $\beta_{12\cdot34\dots p}$.

SUGGESTED READING

- [1] Anderson, R. L. and Bancroft, T. A. *Statistical Theory in Research* (Chs. 7, 13). McGraw-Hill, 1952.
- [2] Goulden, C. H. *Methods of Statistical Analysis* (Chs. 4, 6—8). Asia Publishing House, 1959.
- [3] Hald, A. *Statistical Theory with Engineering Applications* (Chs. 9—11, 18). John Wiley, 1952.
- [4] Johnson, N. L. and Leone, F. C. *Statistics and Experimental Design*, Vol. I (Chs. 8, 12). John Wiley, 1964.
- [5] Keeping, E. S. *Introduction to Statistical Inference* (Chs. 8, 11). Van Nostrand, 1962, and Affiliated East-West Press.
- [6] Mood, A. M. and Graybill, F. A. *Introduction to the Theory of Statistics* (Chs. 11, 12). McGraw-Hill, 1963.

17.1 Introduction

The results obtained in the two preceding chapters are exact in the sense that the probability connected with any test of significance or any confidence interval is exact, provided, of course, the underlying assumption regarding the form of the population is in each case satisfied. These results are valid irrespective of the sample size.

In the present chapter, on the other hand, we shall consider some approximate results which are valid only for sufficiently large samples. This is, no doubt, a limitation, but otherwise these have wider applicability because they hold for all populations which satisfy certain general conditions, rather than being valid for some particular type of population (like the normal). Furthermore, the results are comparatively easy to apply in actual hypothesis-testing or in setting confidence limits to a parameter, requiring, as a rule, the use of the distribution of the normal deviate only. These approximate results are based on the following facts:

In the first place, it has been found that the random sampling distributions of many statistic tend, for large samples, to a form either exactly or very nearly normal, except for some rare types of population.

Secondly, for large n the expectation of a statistic will generally be approximately equal to the corresponding parameter, while its variance will be approximately of the form δ^2/n , δ^2 being a finite quantity. Hence the larger the sample size, the more concentrated will be the distribution about the parameter and the small will be, on the average, the deviations between the values of the statistic and the corresponding parameter. It follows that any unknown parameter can, in general, be well estimated by the value of the statistic found from a given sample, provided the sample size is large enough.

Lastly, the characteristics of the sampling distribution of a statistic, like its mean, moments, etc., may also be estimated from the given sample.

Take, for instance, the sample mean \bar{x} for n random and independent observations on the variable x . It can be shown that if the population variance σ^2 is finite, then the sampling distribution of \bar{x} tends to normality for large n —this is the so-called Central Limit Theorem. Again, \bar{x} has expectation μ , the population mean of x , and variance σ^2/n , so that in case σ^2 is finite, the sample mean will serve as a good estimate of the population mean, provided the sample size is sufficiently large. Consider next the other characteristics (besides the mean) of the sampling distribution of \bar{x} . The standard error σ/\sqrt{n} may be taken for illustration. This will be well estimated by s/\sqrt{n} , s being the sample standard deviation.

For a statistic T of this type $\frac{T-\theta}{\sigma_T}$, where θ is the corresponding parameter and σ_T is the standard error of T , is, thererfore, approximately a normal deviate. If one has to test a hypothesis regarding θ , say,

$$H_0 : \theta = \theta_0,$$

one will utilise the fact that under H_0

$$\frac{T-\theta_0}{\sigma_T} \dots \quad (17.1)$$

is approximately a normal deviate. If σ_T be *known*, either *a priori* or from the hypothesis, one would, therefore, compute T from the sample, use it to calculate $\frac{T-\theta_0}{\sigma_T}$ and compare the resulting value with $\tau_{\alpha/2}$,

τ_α or $-\tau_\alpha$, as the case may be (the level of significance being approximately α). If σ_T be known *a priori*, the confidence limits for θ will be

$$T - \tau_{\alpha/2} \sigma_T \text{ and } T + \tau_{\alpha/2} \sigma_T. \dots \quad (17.2)$$

The confidence coefficient will be $1 - \alpha$ approximately.

In case σ_T is *unknown*, one can substitute its sample value, say, $\hat{\sigma}_T$ and still make the test for H_0 by taking

$$\frac{T-\theta_0}{\hat{\sigma}_T} \dots \quad (17.3)$$

as approximately a normal deviate.

The confidence limits to θ will now be

$$T - \tau_{\alpha/2} \hat{\sigma}_T \text{ and } T + \tau_{\alpha/2} \hat{\sigma}_T. \dots \quad (17.4)$$

A question that naturally arises is Precisely how large should n be for such approximations to be valid? An answer to this question will depend on the nature of the population from which the samples are being taken, on the nature of the statistic and, of course, on the degree of accuracy aimed at. Some practical rules may, however, be suggested. In case one is dealing with sample means or sample proportions, such approximations will usually be good if $n > 30$. In the case of sample medians, variances, coefficients of skewness and kurtosis, correlation coefficients (population correlation being in the neighbourhood of zero), it is necessary that n should be at least about 100. For sample correlation coefficients, when population correlation is considerably different from zero, it is found that even samples of 300 do not give satisfactory approximation.

Ex 17.1 For 150 beans of a particular variety, the mean and standard deviation of breadth of bean were found to be $\bar{x} = 8.512$ mm and $s = 0.616$ mm. Test if the observed mean differs significantly from 8 mm.

The null hypothesis here is $H_0: \mu = 8$, where μ is the mean breadth per bean in the population. Since the sample size in this case is quite large, on the assumption of random and independent sample observations, an approximate test for H_0 will be provided by the statistic $\frac{\sqrt{n}(\bar{x}-8)}{s}$, which is, under H_0 , approximately a normal deviate τ . For the given sample,

$$\begin{aligned}\tau &= \frac{\sqrt{150}(8.512 - 8)}{0.616} \\ &= \frac{12.247 \times 0.512}{0.616} \\ &= 10.179\end{aligned}$$

In the present case we are to use a two sided test since the alternative hypotheses are $H: \mu \neq 8$. Since 10.179 is greater than $\tau_{0.025} = 1.960$ as well as $\tau_{0.005} = 2.576$, the null hypothesis is to be rejected. The observed mean thus differs significantly from given.

Ex. 17·2 With the data of Ex. 17·1, we may also have confidence limits for the unknown population mean. If the chosen confidence coefficient is, say, 0·95, then the corresponding limits will be approximately

$$\bar{x} - 1.960 \frac{s}{\sqrt{n}} = 8.512 - 1.960 \times \frac{0.616}{\sqrt{150}} = 8.512 - 1.960 \times \frac{0.616}{12.247}$$

$$= 8.512 - 0.099 = 8.413 \text{ mm.}$$

and $\bar{x} + 1.960 \frac{s}{\sqrt{n}} = 8.512 + 0.099 = 8.611 \text{ mm.}$

17.2 Tests and confidence intervals for proportions

Suppose in a population p is the proportion of members with a character A . If random samples of size n be drawn from this population, the n drawings being mutually independent, and if we denote by f the number of members of the sample who possess the character A , then the sampling distribution of f will be of the binomial form :

$$\binom{n}{f} p^f (1-p)^{n-f}.$$

The exact treatment of such samples will, therefore, necessitate the use of the binomial distribution. However, if n is sufficiently large, we can use instead the normal distribution. For, as we have stated earlier (in Chapter 9), a binomial distribution tends to the normal form for large n , provided p is not very nearly equal to zero or unity. As a working rule, one may insist that p should lie between $\frac{9}{n}$ and $\frac{n-9}{n}$.

(The reason will be apparent from what we say in Section 17.5).

Since f has mean np and variance $np(1-p)$, one may, therefore, use $\frac{f-np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}}$, where $\hat{p} = \frac{f}{n}$ is the sample proportion of A , approximately as a normal deviate.

Suppose it is required to test the hypothesis $H_0 : p=p_0$ at the level of significance α . Under this hypothesis,

$$\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}} \dots (17.5)$$

is approximately a normal deviate. To test the hypothesis one will, therefore, compute the above quantity from the given sample. (a) If the question is whether p is equal to p_0 or greater, H_0 will be rejected

if the computed value exceeds τ_α (and it will be accepted otherwise).—
 (b) Secondly, if one is interested to know whether p is equal to p_0 , or smaller, H_0 will be rejected when the computed value is found to be smaller than $-\tau_\alpha$. (c) When the question is whether p is or is not equal to p_0 , the hypothesis will be rejected if the computed value exceeds $\tau_{\alpha/2}$ or is smaller than $-\tau_{\alpha/2}$.

Next, in order to get a pair of confidence limits for p with confidence coefficient $1-\alpha$, we see that although the exact variance of \hat{p} is $\frac{p(1-p)}{n}$, if n is large enough, \hat{p} in this expression may be replaced by the sample value \hat{p} . Hence we have approximately

$$P\left[\left|\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}}\right| \leq \tau_{\alpha/2}\right] = 1-\alpha$$

$$\text{or } P\left[\hat{p}-\tau_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p}+\tau_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = 1-\alpha$$

Hence for the given sample

$$\hat{p}-\tau_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+\tau_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (17.6)$$

will be confidence limits to p with confidence coefficient approximately equal to $1-\alpha$.

Suppose now that we have two populations, the proportion of A being p_1 in the one and p_2 in the other. Let random samples of sizes n_1 and n_2 , respectively, be obtained from the first and the second populations through independent drawings, and let $\hat{p}_1=f_1/n_1$ and $\hat{p}_2=f_2/n_2$ be the two sample proportions of A . We have then

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$$

$$\text{and } \text{var}(\hat{p}_1 - \hat{p}_2) = \text{var}(\hat{p}_1) + \text{var}(\hat{p}_2)$$

$$= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Further, as is obvious from the preceding discussion, $\hat{p}_1 - \hat{p}_2$ will also be approximately normal when n_1 and n_2 are sufficiently large. Hence in that case

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

will be distributed approximately as a normal deviate.

Consider then the hypothesis

$$H_0 : p_1 = p_2.$$

According to this hypothesis,

$$E(\hat{p}_1 - \hat{p}_2) = 0$$

and $\text{var}(\hat{p}_1 - \hat{p}_2) = p(1-p)\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}$,

where p is the common value of p_1 and p_2 . If p were given by the hypothesis, one would, therefore, use

$$\tau = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} \quad \dots \quad (17.7)$$

for testing H_0 .

But here, as is usually the case, p is unknown and has to be estimated from the data. The proper estimate will be the proportion of A in the two samples taken together, i.e.

$$\hat{p} = \frac{f_1 + f_2}{n_1 + n_2}.$$

To test H_0 one would, therefore, compute from the given samples

$$\tau = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} \quad \dots \quad (17.8)$$

and compare it with the appropriate tabulated value of the normal deviate for the acceptance or rejection of the hypothesis.

Ex. 17.3 An antibiotic is claimed to cure at least 90% of cases of tuberculosis. 80 T. B. patients are treated with the antibiotic and out of them 59 get cured. Do you consider the claim to be justified?

The null hypothesis in this case is $H_0 : p = 0.9$ (where p is the proportion of patients whom the antibiotic is expected to cure), which is to be tested against the alternatives $H : P < 0.9$. Under the usual assumptions, the test is given by the statistic

$$\frac{\sqrt{n}(\hat{p} - 0.9)}{\sqrt{0.9 \times 0.1}},$$

which may be supposed to be approximately a normal deviate under H_0 since here n is fairly large.

For the given sample,

$$\hat{p} = \frac{59}{80} = 0.7375$$

Hence

$$\begin{aligned}\tau &= \frac{\sqrt{80(0.7375 - 0.9)}}{\sqrt{0.9 \times 0.1}} = \frac{-8.944 \times 0.1625}{0.3} \\ &= -14.53/0.3 = -4.843\end{aligned}$$

This is smaller than $-\tau_{0.05} = -1.645$ as well as $-\tau_{0.01} = -2.326$. As such, the null hypothesis is to be rejected, i.e., the claim that the antibiotic cures at least 90% of cases of T.B. does not seem to be justified in the light of the data.

Ex 17.4 An investigation of the performance of two machines in a factory manufacturing large numbers of bobbins, gives the following results

	No of bobbins examined	No of bobbins found defective
Machine 1	375	17
Machine 2	450	22

Test whether there is any significant difference in the performance of the two machines.

The two machines may be said to be significantly different in their performance if the proportion of defective bobbins for Machine 1 is different from the proportion of defective bobbins for Machine 2. Thus we have to test the null hypothesis $H_0: p_1 = p_2$ against the alternatives $H: p_1 \neq p_2$.

Here the sample proportions are

$$\hat{p}_1 = \frac{17}{375} = 0.04533$$

and $\hat{p}_2 = \frac{22}{450} = 0.04889$,

while $\hat{p} = \frac{\hat{p}_1 + \hat{p}_2}{n_1 + n_2} = \frac{17 + 22}{375 + 450} = \frac{39}{825} = 0.04727$

Hence

$$\tau = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

which is approximately a normal deviate under H_0 , has the value

$$\begin{aligned} & \frac{0.04533 - 0.04889}{\sqrt{0.04727 \times 0.95273 \left(\frac{1}{375} + \frac{1}{450} \right)}} \\ & = -\frac{0.00356}{\sqrt{0.00022018}} = -\frac{0.00356}{0.01484} = -0.240. \end{aligned}$$

Since the observed value of the normal deviate is numerically smaller than both $\tau_{.005} = 2.576$ and $\tau_{.025} = 1.960$, the null hypothesis ought to be accepted. In other words, the given figures do not indicate any significant difference in the performance of the two machines.

17.3 Approximate tests and confidence limits for Poisson parameters

Let x_1, x_2, \dots, x_n be a set of random and independent observations from a Poisson distribution with unknown parameter λ . An approximate test or confidence interval for λ can be obtained from the fact that the sufficient statistic

$$y = \sum_i x_i$$

is approximately normally distributed with mean and variance both equal to $n\lambda$, provided $n\lambda$ is sufficiently large.

A test for

$$H_0 : \lambda = \lambda_0$$

will then be given by the statistic

$$\tau = \frac{y - n\lambda_0}{\sqrt{n\lambda_0}},$$

which is approximately normally distributed under H_0 .

Similarly, the fact that approximately

$$P\left[-\tau_{\alpha/2} \leq \frac{y - n\lambda}{\sqrt{n\lambda}} \leq \tau_{\alpha/2}\right] = 1 - \alpha$$

will provide us with confidence limits to λ , the associated confidence coefficient being approximately $1 - \alpha$.

Again, we may be interested in a comparison among k Poisson distributions with unknown parameters λ_i ($i = 1, 2, \dots, k$). If y_i ($i = 1, 2, \dots, k$) be the totals, for the k populations, of independent

random observations taken from them, then

$$\sum_i \frac{(y_i - n_i \lambda_i)^2}{n_i \lambda_i}$$

is approximately a χ^2 with k d.f. Hence to test for the hypothesis

$$H_0: \lambda_1 = \lambda_2 = \dots = \lambda_k,$$

we could use the statistic

$$\chi^2 = \sum_i \frac{(y_i - n_i \lambda)^2}{n_i \lambda} \quad (\text{with } k \text{ d.f.}) \quad (17.9)$$

if the common value λ were given. But λ will in most cases be unspecified and will have to be estimated from the data. The maximum likelihood estimate is

$$\hat{\lambda} = \frac{\sum_{i=1}^k y_i}{\sum_{i=1}^k n_i}$$

If we replace λ with $\hat{\lambda}$ in (17.9), we may still use

$$\chi^2 = \sum_i \frac{(y_i - n_i \hat{\lambda})^2}{n_i \hat{\lambda}} \quad (\text{with } k-1 \text{ d.f.}) \quad (17.10)$$

for an approximate test for H_0 . This is distributed approximately as a χ^2 with $k-1$ d.f., the loss of one d.f. resulting from the estimation of λ by $\hat{\lambda}$.

Ex 17.5 In a study relating to the traffic conditions in a city, the average daily numbers of motor car accidents during April, 1962 were found to be as follows

Zone	Average daily number
North	17
East	13
South	10
West	12
Central	14

Do you think that the traffic problem is equally acute in all five zones?

Denoting the *average* daily number of accidents in zone i by x_i , we see that $30x_i$, which is the number of accidents for the whole month, may be supposed to be a Poisson variable, say with parameter λ_i . The hypothesis to be tested is

$$H_0 : \lambda_1 = \lambda_2 = \dots = \lambda_5.$$

Under H_0 ,

$$\begin{aligned} \sum_i (30x_i - 30\bar{x})^2 / 30\bar{x} &= 30 \sum_i (x_i - \bar{x})^2 / \bar{x} \\ &= \frac{30}{\bar{x}} [\sum_i x_i^2 - 5\bar{x}^2] \end{aligned}$$

is approximately a χ^2 with 4 d.f.

For the given data, this statistic has the value

$$\frac{30}{13.2} [898 - 5 \times (13.2)^2] = \frac{30 \times 26.8}{13.2} = 60.909.$$

Since this exceeds both $\chi^2_{.05,4} = 9.488$ and $\chi^2_{.01,4} = 13.277$, H_0 is to be rejected. Hence the traffic problem does not seem to be equally acute in all five zones.

17.4 Approximate standard error formulæ of some statistics

The formulæ for standard error as well as those for expectation for sample mean and sample proportion, which we have already come across, are exact and of a simple nature. When we pass on to other statistics, we find that the exact formulæ are of a complicated nature and in some cases are difficult to derive. In the large-sample case, however, we need not know these exact expressions ; approximate formulæ serve our purpose.

We give below such approximate formulæ for the standard errors—rather for the variances—of some of the more important statistics. As regards this expectation, it may be noted that the expectation of every statistic considered here may be taken to be approximately equal to the corresponding parameter. (In the derivation of these formulæ, it has been assumed that the sample observations are random and independent.)

These results stem from the fact that if T_1, T_2, \dots, T_k are k statistics and $\theta_1, \theta_2, \dots, \theta_k$ are the corresponding parameters, then for a sufficiently well-behaved function $\psi(T_1, T_2, \dots, T_k)$, one has

$$\psi(T_1, T_2, \dots, T_k) \approx \psi(\theta_1, \theta_2, \dots, \theta_k) + \sum_i (T_i - \theta_i) \left(\frac{\partial \psi}{\partial T_i} \right).$$

in the neighbourhood of the point $T_1 = \theta_1, T_2 = \theta_2, \dots, T_k = \theta_k$, $\left(\frac{\partial\psi}{\partial T_i}\right)_s$ being the value of the partial derivative of ψ with respect to T_i at this point.

Hence if $E(T_i) \approx \theta_i$ for each i and $\text{var}(T_i)$, $\text{cov}(T_i, T_j)$ are $O(1/n)^*$, then

$$E[\psi(T_1, T_2, \dots, T_k)] \approx \psi(\theta_1, \theta_2, \dots, \theta_k) \quad (17.11a)$$

and $\text{var}[\psi(T_1, T_2, \dots, T_k)] \approx \sum_i \left(\frac{\partial\psi}{\partial T_i}\right)_s^2 \text{var}(T_i)$

$$+ \sum_{i \neq j} \left(\frac{\partial\psi}{\partial T_i}\right)_s \left(\frac{\partial\psi}{\partial T_j}\right)_s \text{cov}(T_i, T_j) \quad (17.11b)$$

Central moments If m_r be the r th central moment of the sample and μ_r the corresponding population moment, then

$$\text{var}(m_r) \approx \frac{1}{n} (\mu_2 - \mu_1^2 - 2r\mu_{r-1}\mu_{r+1} + r^2\mu_{r-1}^2) \quad (17.12)$$

In particular, for the sample variance s^2 we have

$$\text{var}(s^2) \approx \frac{1}{n} (\mu_4 - \mu_2^2) \quad (17.13a)$$

If the population is *normal*, then $\mu_4 = 3\mu_2^2$, so that

$$\text{var}(s^2) \approx 2\sigma^4/n \quad (17.13b)$$

If we consider, instead of the sample variance, the sample standard deviation s , it may be proved that in the *normal* population case,

$$\text{var}(s) \approx \sigma^2/2n \quad (17.14)$$

g₁ and g₂ coefficients In sampling from a *normal* population,

$$\text{var}(g_1) \approx 6/n \quad (17.15)$$

and $\text{var}(g_2) \approx 24/n \quad (17.16)$

Coefficient of variation Let us denote by v and V the sample and population coefficient of variation, respectively, of x . If x is normally distributed in the population, then

$$\text{var}(v) \approx \frac{V^2}{2n} \left[1 + \frac{2V^2}{10^4} \right] \quad (17.17)$$

* $O(1/n)$ is a quantity such that $\lim_{n \rightarrow \infty} n O(1/n)$ is finite e.g. $5/n + 2/n^2$

Sample correlation coefficient : If r be the sample correlation of the variables x and y which are distributed in the bivariate normal form in the population with correlation coefficient ρ , then

$$\text{var}(r) \approx \frac{(1-\rho^2)}{n}. \quad \dots \quad (17.18)$$

Sample quantiles : Let z_p be the sample quantile of order p of the variable x and let ζ_p be the corresponding population quantile, then

$$\text{var}(z_p) \approx \frac{p(1-p)}{n} [f(\zeta_p)]^{-2}. \quad \dots \quad (17.19)$$

Here x is supposed to be continuous with probability-density function $f(x)$, so that $f(\zeta_p)$ is the probability-density at $x = \zeta_p$.

For the median, $p = \frac{1}{2}$. Hence the variance of the sample median is, approximately,

$$\frac{1}{4n} [f(\zeta_{1/2})]^{-2}.$$

In case x has a normal distribution with variance σ^2 ,

$$f(\zeta_{1/2}) = \frac{1}{\sigma \sqrt{2\pi}}.$$

Therefore, in sampling from a *normal* parent, the variance of the sample median is approximately

$$\frac{\pi \sigma^2}{2n}. \quad \dots \quad (17.20)$$

We find incidentally that although both the sample mean and sample median in the case of a normal population have expectation μ , the population mean, exactly or approximately, the variance of the latter is about 1.57 times the variance of the former. Hence it is considered advisable to take the sample mean, rather than the sample median, as the proper estimate of μ (*vide* Ex. 15.4).

The use of some of these formulæ is illustrated in the following examples.

Ex. 17.6 The standard deviation of life in hours per bulb for a sample of 150 electric bulbs, taken from those produced by a factory in a particular year, was found to be 464 hours. The standard deviation of the same variable for 175 electric bulbs, taken from those produced in the following year, was 653 hours. Test if there has been a significant change in the variability of life of bulbs.

In the present case we have to test (against all alternatives) $H_0: \sigma_1 = \sigma_2$, where σ_1 and σ_2 are the standard deviations of life of bulb for bulbs produced in the first and second years, respectively.

Assuming that the variable is normally distributed in each population and that the sample observations are random and mutually independent, the test may be performed by means of

either

$$\frac{s_1 - s_2}{\sigma \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}}$$

or

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

each of which is approximately a normal deviate under H_0 *.

Method I The common value σ^2 of the two population variances being unknown we estimate it by

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

If we compare the two sample standard deviations, then the test is given by

$$\tau = \frac{s_1 - s_2}{s \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}}$$

For the given samples,

$$s_1 = 464, s_2 = 653,$$

$$\text{so that } s^2 = \frac{150 \times (464)^2 + 175 \times (653)^2}{150 + 175} = \frac{106,915,975}{325} = 328,979$$

$$\text{and } s = 574$$

$$\begin{aligned} \text{Hence } \tau &= \frac{464 - 653}{574 \sqrt{\frac{1}{300} + \frac{1}{350}}} = -\frac{189}{574 \sqrt{0.006190}} \\ &= -\frac{189}{574 \times 0.0787} = -4.184 \end{aligned}$$

Since this is numerically greater than both $\tau_{0.025} = 1.960$ and $\tau_{0.005} = 2.576$, H_0 is to be rejected.

*The F ratio approaches normal as n_1 and n_2 increase.

Method 2. H_0 may be tested by means of the statistic

$$\tau = \frac{s_1^2 - s_2^2}{\sqrt{\frac{2}{n_1} + \frac{2}{n_2}}},$$

which is also approximately distributed as a normal deviate under H_0 . For the given samples,

$$\begin{aligned}\tau &= \frac{215,296 - 426,409}{328,972 \sqrt{0.024762}} = -\frac{211,113}{328,972 \times 0.1574} \\ &= -4.077,\end{aligned}$$

which is almost equal to the value of τ obtained by using *Method I*. Thus both the tests lead to the rejection of the null hypothesis. In other words, each of the tests indicates that there has been a significant change in the variability of life of bulbs during the period.

Ex. 17.7 For 600 beans of a particular variety, the frequency distribution of breadth (in mm.) has

$$g_1 = -0.128 \text{ and } g_2 = 0.195.$$

Examine if the population distribution may be supposed to be normal.

If the population distribution be really of the normal type, then $\gamma_1 = 0$, $\gamma_2 = 0$. We should, therefore, test the hypotheses $H_0 : \gamma_1 = 0$ and $H_0 : \gamma_2 = 0$. On the assumption of random and independent observations, the tests are given by

$$g_1 \sqrt{\frac{n}{6}} \text{ and } g_2 \sqrt{\frac{n}{24}},$$

respectively, which are distributed approximately as normal deviates under the null hypotheses. For the given sample,

$$g_1 \sqrt{\frac{n}{6}} = -0.128 \times \sqrt{100} = -1.28$$

and $g_2 \sqrt{\frac{n}{24}} = 0.195 \times \sqrt{25} = 0.975.$

On comparing their absolute values with

$$\tau_{.025} = 1.960 \text{ and } \tau_{.005} = 2.576,$$

we find that both the hypotheses are acceptable. The population distribution, therefore, may be supposed to be of the normal form.

17.5 z -transformation of sample correlation and other transformations

It has been noted in the previous section that in random sampling from a bivariate normal population, the sample correlation r is approximately normally distributed about the population correlation ρ with approximate variance $(1-\rho^2)^2/n$

The sampling distribution of r tends to normality fairly rapidly when ρ is not very different from zero. However, when ρ differs widely from zero, e.g. when $\rho = \pm 0.7$, this sampling distribution tends to normality so slowly that the use of the normal approximation will not be advisable even if n is as large as 100.

For such values of ρ , it is advisable to use the transformation

$$z = \frac{1}{2} \log_e \frac{1+r}{1-r}, \quad (17.21)$$

introduced by Fisher. The new statistic z may be assumed to be normally distributed even when n is as small as 10, although ρ may be widely different from zero. It has been shown that z has approximate mean

$$\zeta = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} \quad (17.21a)$$

and approximate variance

$$\frac{1}{n-3} \quad (17.21b)$$

One can, therefore, test any hypothesis regarding ρ or get confidence limits for ρ by using the statistic

$$\sqrt{n-3}(z - \zeta)$$

as approximately a normal deviate, even for moderately large values of n .

Other transformations of this type are given by the following functions

(1) $\sin^{-1} \sqrt{f/n}$, where f is the observed number of successes in a series of n Bernoullian trials with probability of success p , with

$$E(\sin^{-1} \sqrt{f/n}) \approx \sin^{-1} \sqrt{p}, \quad (17.22a)$$

$$\text{var}(\sin^{-1} \sqrt{f/n}) \approx \frac{1}{4n}, \quad (17.22b)$$

(2) \sqrt{x} , where x is a Poisson variable with parameter λ (assumed large), with

$$E(\sqrt{x}) \approx \sqrt{\lambda}, \quad \dots \quad (17.23a)$$

$$\text{var}(\sqrt{x}) \approx \frac{1}{4}; \quad \dots \quad (17.23b)$$

(3) $\log_e s^2$, where s^2 is the sample variance in sampling from a normal population with variance σ^2 , with

$$E(\log_e s^2) \approx \log_e \sigma^2, \quad \dots \quad (17.24a)$$

$$\text{var}(\log_e s^2) \approx \frac{2}{n}. \quad \dots \quad (17.24b)$$

These transformations have a two-fold merit. First, the transformed statistic tends to normality much more rapidly than the original statistic. Secondly, the transformed statistic has an asymptotic variance which is independent of population parameters, thus providing a better test or confidence interval than the original statistic.

Ex. 17.8 In Ex. 11.1 the correlation coefficient between marks in statistics Hons. in a college test and those in the subsequent university examination, for 20 students, was found to be 0.727. What can be said about the population correlation coefficient?

Let us assume (1) that in the population the two variables are jointly normally distributed and (2) that the 20 sample observations, on which the observed correlation coefficient r is based, are random and mutually independent. We can then have confidence limits for the population correlation ρ .

We have, approximately,

$$P[-r_{025} \leq \sqrt{n-3}(z-\zeta) \leq r_{025}] = 0.95$$

or $P\left[z - \frac{r_{025}}{\sqrt{n-3}} \leq \zeta \leq z + \frac{r_{025}}{\sqrt{n-3}}\right] = 0.95.$

Hence the 95% confidence limits for ζ are

$$z - \frac{r_{025}}{\sqrt{n-3}} \text{ and } z + \frac{r_{025}}{\sqrt{n-3}}.$$

For the given sample,

$$\begin{aligned} z &= \frac{1}{2}(\log 1.727 - \log 0.273) \log_e 10 \\ &= \frac{1}{2}(0.23729 - 1.43616) 2.30259 = 0.92234, \end{aligned}$$

so that the confidence limits for ζ are

$$0.92234 - \frac{1.960}{\sqrt{17}} = 0.92234 - \frac{1.960}{4.1234} \\ = 0.92234 - 0.47537 = 0.44697$$

and $0.92234 + 0.47537 = 1.39771$

Now $\frac{1+\rho}{1-\rho} = \text{antilog}(2\zeta \log e) = \lambda$, say, or $\rho = \frac{\lambda-1}{\lambda+1}$

When $\zeta = 0.44697$, $\lambda = \text{antilog}(2 \times 0.44697 \times 0.43429)$
 $= \text{antilog}(0.38823) = 2.445$

and $\rho = 1.445/3.445 = 0.419$

When $\zeta = 1.39771$, $\lambda = \text{antilog}(2 \times 1.39771 \times 0.43429)$
 $= \text{antilog}(1.21402) = 16.369$

and $\rho = 15.369/17.369 = 0.885$

The 95% confidence limits for ρ are, therefore, 0.419 and 0.885

Ex 17.9 The correlation coefficient between head-length and head-breadth is 0.324 for 90 Brahmins and is 0.278 for 130 Chattris. Test whether the two coefficients differ significantly.

Denoting the correlation coefficient between the two characters in the population of all Brahmins by ρ_1 and the corresponding coefficient in the population of all Chattris by ρ_2 , we have here as our null hypothesis $H_0: \rho_1 = \rho_2$, which is to be tested against all alternatives.

We shall assume that in each population the two characters are distributed in the bivariate normal form. Further, the n_1+n_2 pairs of sample observations taken from the two populations will be supposed to be random and independent.

If we put $z_1 = \frac{1}{2} \log \frac{1+r_1}{1-r_1}$ and $z_2 = \frac{1}{2} \log \frac{1+r_2}{1-r_2}$, r_1 and r_2 being the sample correlations for samples taken from the first and second populations, then an approximate test for H_0 is provided by the statistic

$$\frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}},$$

which is approximately a normal deviate under H_0 .

For the given samples,

$$z_1 = \frac{1}{2}(\log 1.324 - \log 0.676) 2.30259$$

$$= \frac{1}{2}(0.12189 - 1.82995) 2.30259 = 0.33611$$

and

$$z_2 = \frac{1}{2}(\log 1.278 - \log 0.722) 2.30259$$

$$= \frac{1}{2}(0.10653 - 1.85854) 2.30259 = 0.28551,$$

so that

$$\tau = \frac{0.33611 - 0.28551}{\sqrt{\frac{1}{87} + \frac{1}{127}}} = \frac{0.05060}{\sqrt{0.019368}} = \frac{0.05060}{0.13917} = 0.364.$$

Since the value exceeds neither $\tau_{.005} = 2.576$ nor $\tau_{.025} = 1.960$, the hypothesis is to be accepted. In other words, the difference between the two sample correlation coefficients is to be regarded as insignificant.

17.6 Frequency χ^2

Suppose a population consists of k mutually exclusive classes, the proportion of members falling in the i th class being p_i , $i = 1, 2, \dots, k$. This classification may be with respect to either an attribute or a variable. (In the case of a continuous variable, the classification will necessarily be artificial, being achieved by dividing the whole range of the variable into k arbitrarily defined intervals.) Obviously,

$$\sum_{i=1}^k p_i = 1.$$

If a random sample of size n be drawn from this population, the drawings being mutually independent, then the probability that f_1 of the members of the sample will belong to the first class, f_2 to the second, \dots , f_k to the last is

$$\frac{n!}{f_1! f_2! \dots f_k!} p_1^{f_1} p_2^{f_2} \dots p_k^{f_k}. \quad \dots \quad (17.25)$$

For $k=2$, it defines a binomial distribution. In the general case the distribution is called a *multinomial distribution*. Note that (17.25) may be expressed in the form

$$\frac{\frac{e^{-np_1} (np_1)^{f_1}}{f_1!} \times \frac{e^{-np_2} (np_2)^{f_2}}{f_2!} \times \dots \times \frac{e^{-np_k} (np_k)^{f_k}}{f_k!}}{\frac{e^{-(np_1+np_2+\dots+np_k)} (np_1+np_2+\dots+np_k)^n}{n!}}. \quad \dots \quad (17.25a)$$

It is known that the sum of k mutually independent Poisson variables, say, x_i ($i=1, 2, \dots, k$) with parameters λ_i , is itself a Poisson variable with parameter $\sum_i \lambda_i$. Hence in the form (17.25a) the multinomial distribution appears as the conditional distribution of k independent Poisson variables f_1, f_2, \dots, f_k , subject to the condition

$$f_1 + f_2 + \dots + f_k = n \quad (17.26)$$

Now, it is also known that a Poisson variable with parameter λ tends to normality if $\lambda \rightarrow \infty$. In the present case, the variables are f_i , with parameters np_i . Hence for each i ,

$$\frac{f_i - np_i}{\sqrt{np_i}}$$

is approximately a normal deviate if np_i is sufficiently large. Thus, approximately,

$$\sum_{i=1}^k \left(\frac{f_i - np_i}{\sqrt{np_i}} \right)^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i} \quad (17.27)$$

comes out to be the sum of squares of k normal deviates. The approximate normal deviates are, however, subject to the linear constraint (17.26), which may be written in the form

$$\sum_i \sqrt{np_i} \left(\frac{f_i - np_i}{\sqrt{np_i}} \right) = 0 \quad (17.26a)$$

In consequence, (17.27) will have approximately the χ^2 distribution with $k-1$ d.f., provided the theoretical frequencies np_i are large enough. In statistical literature, (17.27) is referred to as a *Pearsonian χ^2* (after Karl Pearson) or a *frequency χ^2* .

For computational purposes, one may use the following simplified form of (17.27)

$$\sum_{i=1}^k \frac{f_i^2}{np_i} - n \quad (17.27a)$$

For $k=2$, (17.27) becomes

$$\begin{aligned} & \frac{(f_1 - np_1)^2}{np_1} + \frac{(f_2 - np_2)^2}{np_2} \\ &= (f_1 - np_1)^2 \left\{ \frac{1}{np_1} + \frac{1}{n(1-p_1)} \right\} \quad [\text{since } f_1 + f_2 = np_1 + np_2 = n, \\ & \quad \text{so that } f_1 - np_1 = -(f_2 - np_2)] \\ &= \frac{(f_1 - np_1)^2}{np_1(1-p_1)} - \left\{ \frac{\sqrt{n}(f_1/n - p_1)}{\sqrt{p_1(1-p_1)}} \right\}^2, \end{aligned}$$

f_1/n being the sample proportion for the first class. This is approximately a χ^2 with 1 d.f.—a fact of which we are already aware, since

$$\frac{\sqrt{n}(f_1/n - p_1)}{\sqrt{p_1(1-p_1)}}$$

is known to be approximately a normal deviate.

It has been stated above that for this χ^2 approximation to be valid, the theoretical frequencies np_i should be sufficiently large. As a working lower limit, we may take 5, since both practical and theoretical investigations show that the approximation is usually satisfactory if $np_i \geq 5$ for each i , provided the number of classes is also greater than or equal to 5. If the number of classes is smaller than 5, it is advisable to have each of the expected frequencies somewhat greater than 5. When it is found that for some class np_i is less than 5, one should amalgamate or coalesce this class with one or more of the adjacent classes so as to make the theoretical frequency in the combined class greater than or equal to 5. The number of degrees of freedom will then be :

(number of classes after coalescing) — 1.

In fact, some recent studies made by Cochran [1] suggest that if relatively few expected frequencies are less than 5 (say, just one out of five or more, or two out of ten or more), then even as low a value as 1 is allowable for an expected frequency in using the χ^2 approximation.

We shall see presently how this statistic may be used to solve various problems in hypothesis-testing.

17.7 Test for goodness of fit : hypothetical population completely specified

In earlier chapters, we considered tests for parametric hypotheses. In developing such tests, we made the assumption that the parent population is of a specified nature. (In most cases it was assumed that the population is of the normal type). We shall now consider hypotheses of a more fundamental nature, where these assumptions themselves are questioned and one seeks to verify them on the basis of sample observations.

Let us first take up the case where the hypothetical population is completely specified, there being no unknown parameter in its

distribution Let us visualise the population as being composed of k mutually exclusive classes, and let us suppose that, according to the hypothesis, the population proportion in the i th class is p_i^0 . If the frequency in the i th class in random samples of size n from this population be denoted by f_i , we find from Section 17.6 that, under the hypothesis,

$$\sum_i \frac{(f_i - np_i^0)^2}{np_i^0} = \sum_i \frac{f_i^2}{np_i^0} - n \quad (17.28)$$

is approximately a χ^2 with $(k-1) d f$, provided np_i^0 is large enough for each i . The χ^2 statistic, therefore, provides an approximate test for the hypothesis. The greater the differences between the observed frequencies f_i and the expected frequencies (under the hypothesis) np_i^0 , the greater will be the value of (17.28). Hence it would appear that a very high value of (17.28) should indicate falsity of the given hypothesis. If α be the chosen level of significance, then our test procedure consists in the rejection of the hypothesis if in a given sample $\sum_i f_i^2 / np_i^0 - n$ exceeds $\chi_{\alpha, (k-1)}^2$ and in its acceptance otherwise.

Since our task here is to see how well the expected frequencies np_i^0 are in agreement with (or how well they fit) the observed frequencies f_i , such a test is also called a test for goodness of fit.

Ex 17.10 In the course of an experiment on the breeding of peas, a botanist obtained 556 peas, of which 315 were round and yellow, 108 were round and green, 101 were angular and yellow and 32 were angular and green. According to a genetic theory, such peas should be obtained in the ratio of 9 3 3 1. Are the experimental results compatible with this theory?

If we denote by p_1, p_2, p_3 and p_4 the proportions of peas in the four classes in the whole population of peas that may be obtained in experiments of this type, then the null hypothesis to be tested is

$$H_0: p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$$

Let us assume that the given 556 peas have been taken randomly and independently from this population. A test for the hypothesis H_0 is then supplied by the statistic $\sum_i \frac{f_i^2}{np_i^0} - n$, which is, under H_0 , distributed approximately as a χ^2 with 3 d.f. The calculation of this χ^2 for the given sample is shown in the following table.

(1) Class	(2) Observed frequency	(3) Expected frequency	(4) (2) ² /(3)
Round and yellow	315	312.75	317.266
Round and green	108	104.25	111.885
Angular and yellow	101	104.25	97.851
Angular and green	32	34.75	29.468
Total	556	556.00	556.470

Hence

$$\chi^2 = 556.470 - 556 = 0.470.$$

The tabulated values are

$$\chi^2_{.05,3} = 7.815 \text{ and } \chi^2_{.01,3} = 11.345.$$

The observed χ^2 being insignificant at both the levels of significance, the experimental results seem compatible with the genetic theory.

17.8 Test for goodness of fit : some parameters of hypothetical population unknown

This is the more usual form of the problem of testing for goodness of fit. It differs from the preceding problem in that now the proportions p_i^0 are not completely specified by the hypothesis but are dependent on some unknown parameter or parameters. Such is the case, for instance, when the hypothesis says that the population is of the Poisson or of the normal type, without specifying the value of λ or those of μ and σ .

In the general case, let us suppose that the hypothetical population depends on r unknown parameters ($r < k-1$), which may be estimated from the given sample itself. If the appropriate estimates of the parameters (e.g. those obtained by the method of moments) are considered and if the corresponding estimates of the population proportions are denoted by \hat{p}_i^0 , then

$$\sum_i \frac{(f_i - n\hat{p}_i^0)^2}{n\hat{p}_i^0} = \sum_i \frac{f_i^2}{n\hat{p}_i^0} - n \quad \dots \quad (17.29)$$

will still be approximately a χ^2 if $n\hat{p}_i^0$ are large enough. However, the number of degrees of freedom will get reduced, for the estimation of each parameter imposes a homogeneous linear constraint on the

(approximate) normal deviates

$$\frac{f_i - n\hat{p}_i^0}{\sqrt{n\hat{p}_i^0}}$$

The number of degrees of freedom of the above χ^2 statistic will, therefore, be

$$(k-1) - (\text{number of parameters estimated}) = k - r - 1 \quad (17.30)$$

Ex 17.11 In Ex 9.4 a normal distribution was fitted to the observed distribution given in Table 5.10. We shall assume that the given 177 persons have been taken randomly and independently from the population of all Indian adult males. We may then test for the goodness of fit of the normal distribution (i.e., we may judge whether or not the population distribution of height may be supposed to be of the normal type) by means of the χ^2 statistic.

The computation of the χ^2 is shown in the table below. It may be noted that here the first three and the last three class intervals of the variable (*vide* Table 9.4) have been amalgamated to form only two intervals. This has been done in order to make the expected frequency in each class greater than 5.

(1) Height (mm.)	(2) Observed frequency	(3) Expected frequency	(4) $(3)^2/(2)$
-154.55	4	5.553	2.881
154.55-159.55	24	24.860	23.170
159.55-164.55	58	55.687	60.409
164.55-169.55	60	57.371	62.749
169.55-174.55	27	27.085	26.915
174.55-	4	6.444	2.483
Total	177	177.000	178.607

From the above table,

$$\chi^2 = 178.607 - 177 = 1.607$$

Remembering that in fitting the normal distribution, two parameters had to be estimated from the sample, we see that the χ^2 statistic now has $5-2=3$ d.f. The relevant tabulated values are $\chi^2_{0.05,3}=7.815$ and $\chi^2_{0.1,3}=11.345$. Since the observed χ^2 is much smaller than the tabulated values, the normal distribution seems to have given a very good fit.

17.9 Test for homogeneity

Suppose there are l similarly classified populations. Let k be the number of classes in each population and p_{ij} the proportion of the j th population in the i th class ($i=1, 2, \dots, k$ and $j=1, 2, \dots, l$). The populations may be represented as follows :

Class	Population					
	1	2	3	l
1	p_{11}	p_{12}	p_{13}	p_{1l}
2	p_{21}	p_{22}	p_{23}	p_{2l}
3	p_{31}	p_{32}	p_{33}	p_{3l}
\vdots	\vdots	\vdots	\vdots	\vdots
k	p_{k1}	p_{k2}	p_{k3}	p_{kl}
Total	1	1	1			1

When p_{ij} are unknown, one may want to know if the l population distributions may be supposed to be identical (or homogeneous). One has then to test the hypothesis

$$H_0 : p_{i1} = p_{i2} = \dots = p_{il} \quad \text{for each } i.$$

Let a random sample of size n_j be drawn from the j th population ($j=1, 2, \dots, l$), the drawings being mutually independent, and let the number of members of this sample which belong to the i th class be f_{ij} . We have then

$$\sum_{i=1}^k f_{ij} = n_j.$$

The position may be visualised from the following table :

Class	Sample						Total
	1	2	3	l	
1	f_{11}	f_{12}	f_{13}	f_{1l}	f_{10}
2	f_{21}	f_{22}	f_{23}	f_{2l}	f_{20}
3	f_{31}	f_{32}	f_{33}	f_{3l}	f_{30}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
k	f_{k1}	f_{k2}	f_{k3}	f_{kl}	f_{k0}
Total	n_1	n_2	n_3	n_l	n

Under the present set-up, for each j

$$\sum_{i=1}^k \frac{(f_{ij} - n_j p_{ij})^2}{n_j p_{ij}}$$

is distributed approximately as a χ^2 with $k-1$ d.f.

Hence

$$\sum_{j=1}^l \sum_{i=1}^k \frac{(f_{ij} - n_j p_{ij})^2}{n_j p_{ij}} = \sum_{i=1}^k \sum_{j=1}^l \frac{(f_{ij} - n_j p_{ij})^2}{n_j p_{ij}},$$

being the sum of l independent χ^2 's each with $k-1$ d.f., is itself a χ^2 , with $(k-1)l$ d.f. According to the hypothesis, therefore,

$$\sum_{i,j} \frac{(f_{ij} - n_j p_i^0)^2}{n_j p_i^0}, \quad \dots \quad (17.31)$$

where p_i^0 is the common value of p_{ij} for all j , is approximately a χ^2 with $(k-1)l$ d.f. This statistic could be used to test H_0 if p_i^0 's were known quantities. Suppose we replace each p_i^0 by its estimate—the proper estimate is the sample proportion obtained by combining all samples, viz

$$\hat{p}_i^0 = \frac{\sum f_{ij}}{\sum_j n_j} = \frac{f_{i0}}{n},$$

where $f_{i0} = \sum_j f_{ij}$ and $n = \sum_j n_j$. Then the frequency χ^2 takes the form

$$\sum_{i,j} \frac{\left(f_{ij} - \frac{n_j f_{i0}}{n} \right)^2}{\frac{n_j f_{i0}}{n}} = n \sum_{i,j} \frac{f_{ij}^2}{f_{i0} n_j} - n. \quad \dots \quad (17.32)$$

Because of this estimation, the number of degrees of freedom will get reduced by $(k-1)$ —and not by k , since when $k-1$ proportions are estimated, the remaining one is automatically determined by virtue of the property that the sum of all proportions is unity.

The hypothesis H_0 will, therefore, be rejected or accepted according as

$$n \left(\sum_{i,j} \frac{f_{ij}^2}{f_{i0} n_j} - 1 \right)$$

exceeds $\chi^2_{\alpha, (k-1)(l-1)}$ or not, $\chi^2_{\alpha, (k-1)(l-1)}$ being the upper α -point of the χ^2 distribution with

$$(k-1)l - (k-1) = (k-1)(l-1) \text{ d.f.} \quad \dots \quad (17.33)$$

17.10 Test for independence

Let a population be classified according to two attributes, A and B , into k and l classes respectively, say,

$$A_1, A_2, \dots, A_k$$

and

$$B_1, B_2, \dots, B_l.$$

Let p_{ij} be the proportion of members of the population belonging simultaneously to the i th class of A (i.e. A_i) and the j th class of B (i.e. B_j). The structure of the population will then be as follows :

	B_1	B_2	B_3	...	B_l	Total
A_1	p_{11}	p_{12}	p_{13}	...	p_{1l}	p_{i0}
A_2	p_{21}	p_{22}	p_{23}	...	p_{2l}	p_{i0}
A_3	p_{31}	p_{32}	p_{33}	...	p_{3l}	p_{30}
\vdots	\vdots	\vdots	\vdots		\vdots	\vdots
A_k	p_{k1}	p_{k2}	p_{k3}	...	p_{kl}	p_{k0}
Total	p_{01}	p_{02}	p_{03}	...	p_{0l}	1

The proportions p_{ij} define the *joint distribution* of A and B . The marginal totals

$$p_{i0} = \sum_{j=1}^l p_{ij}$$

give the *marginal distribution* of A ; while the other marginal totals

$$p_{0j} = \sum_{i=1}^k p_{ij}$$

give the *marginal distribution* of B .

When p_{ij} are unknown, we may enquire whether A and B are independent. We have then to test the hypothesis

$$H_0 : p_{ij} = p_{i0} \times p_{0j} \quad (\text{for all } i, j)$$

Let a random sample of size n be drawn from the population, the drawings being mutually independent. If we denote by f_{ij} the number of members of the sample that belong both to the i th class of A and to the j th class of B , then, under the hypothesis H_0 ,

$$\sum_{i=1}^k \sum_{j=1}^l \frac{(f_{ij} - n p_{i0} p_{0j})^2}{n p_{i0} p_{0j}} \quad \dots \quad (17.34)$$

will be distributed approximately as a χ^2 with $kl - 1$ d.f. This statistic could be used to test H_0 if p_{10} and p_{01} were known quantities. In the present case, however, they are unknown and have to be estimated from the sample itself.

The proper estimate of p_{10} , which is the population proportion in the class A_1 , is the corresponding sample proportion

$$f_{10}/n,$$

where $f_{10} = \sum_i f_{1ij}$. Similarly, the proper estimate of p_{01} is

$$f_{01}/n,$$

where $f_{01} = \sum_i f_{0ij}$. The structure of the sample is shown below

	B_1	B_2	B_3	B_l	Total
A_1	f_{11}	f_{12}	f_{13}	f_{1l}	f_{10}
A_2	f_{21}	f_{22}	f_{23}	f_{2l}	f_{21}
A_3	f_{31}	f_{32}	f_{33}	f_{3l}	f_{31}
A_k	f_{k1}	f_{k2}	f_{k3}	f_{kl}	f_{k1}
Total	f_{01}	f_{02}	f_{03}	f_{0l}	n

Substituting these estimates for p_{10} and p_{01} in (17.34), we get the new statistic

$$\sum_i \sum_j \frac{\left(f_{ij} - \frac{f_{10} f_{01}}{n} \right)^2}{\frac{f_{10} f_{01}}{n}} = n \sum_i \sum_j \frac{f_{ij}^2}{f_{10} f_{01}} - n \quad (17.35)$$

We are using $k+l$ estimates, of which $(k-1)+(l-1)$ are independent. For, given $(k-1)$ of the estimates for p_{10} , the other automatically follows, and similarly, given $(l-1)$ of the estimates for p_{01} , the other is automatically determined. Hence (17.35) is distributed approximately as a χ^2 with

$$(kl-1)-(k-1)-(l-1)=(k-1)(l-1) \quad (17.36)$$

degrees of freedom.

Thus we see that, although the problem discussed here is different from that in the preceding section, the solution of each is formally

the same, the χ^2 statistic used in each case being of the form

$$\text{(grand total)} \sum_{i=1}^k \sum_{j=1}^l \frac{(\text{class frequency})^2}{(\text{row total}) \times (\text{column total})} - \text{(grand total)}$$

with $d.f. = (\text{no. of rows} - 1) \times (\text{no. of columns} - 1)$.

7.11 Simplified formulæ

When either k or l is equal to 2, the expressions (17.32) and 17.35) reduce to much simpler forms.

In the first place, suppose $l=2$ and $k>2$. Here the sample frequencies and their totals may be represented as follows :

Row	Column 1	Column 2	Total
1	a_1	b_1	T_1
2	a_2	b_2	T_2
3	a_3	b_3	T_3
\vdots	\vdots	\vdots	\vdots
k	a_k	b_k	T_k
Total	T_a	T_b	n

In terms of these symbols, the expression corresponding to (17.32) or (17.35) is now

$$\begin{aligned}
 & n \sum_{i=1}^k \frac{\left(a_i - \frac{T_i T_a}{n}\right)^2}{T_i T_a} + n \sum_{i=1}^k \frac{\left(b_i - \frac{T_i T_b}{n}\right)^2}{T_i T_b} \\
 & = n \left(\frac{1}{T_a} + \frac{1}{T_b} \right) \sum_{i=1}^k \frac{\left(a_i - \frac{T_i T_a}{n}\right)^2}{T_i}, \text{ since } \left(a_i - \frac{T_i T_a}{n}\right) = -\left(b_i - \frac{T_i T_b}{n}\right) \\
 & = \frac{n^2}{T_a T_b} \left(\sum_i \frac{a_i^2}{T_i} - 2 \frac{T_a}{n} \sum_i a_i + \frac{T_a^2}{n^2} \sum_i T_i \right) \\
 & = \frac{n^2}{T_a T_b} \left(\sum_i \frac{a_i^2}{T_i} - \frac{T_a^2}{n} \right). \quad \dots \quad (17.37a)
 \end{aligned}$$

This formula or its equivalent,

$$\frac{n^2}{T_a T_b} \left(\sum_i \frac{b_i^2}{T_i} - \frac{T_b^2}{n} \right), \quad \dots \quad (17.37b)$$

will be found more convenient for computational purposes.

In case k and l are both equal to 2, i.e. for a 2×2 table, the sample frequencies and the totals may be written as follows:

Row	Column		Total
	1	2	
1	a	b	$a+b$
2	c	d	$c+d$
Total	$a+c$	$b+d$	n

The difference between any observed frequency, such as a or c , and the corresponding expected frequency, such as $\frac{(a+b)(a+c)}{n}$ or $\frac{(a+b)(b+d)}{n}$, is numerically the same for all four cells of the table.

Hence the approximate χ^2 statistic for testing homogeneity or independence (as the case may be) will now be

$$\begin{aligned}
 & n \left\{ a - \frac{(a+b)(a+c)}{n} \right\}^2 \left\{ \frac{1}{(a+b)(a+c)} + \frac{1}{(a+b)(b+d)} + \frac{1}{(c+d)(a+c)} \right. \\
 & \quad \left. + \frac{1}{(c+d)(b+d)} \right\} \\
 & = \frac{1}{n} \left\{ na - (a+b)(a+c) \right\}^2 \frac{n^2}{(a+b)(c+d)(a+c)(b+d)} \\
 & = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \quad \dots \quad (17.3)
 \end{aligned}$$

This again will be found much easier to apply than the original formula.

17.12 Yates' correction for continuity

We have stated in Section 17.5 that, for the validity of the approximation, it is necessary that the expected frequency in each class should be sufficiently large, say greater than 5. When some expected frequency is too small, we coalesce some of the classes in order to satisfy this condition. However, it should be apparent that this procedure is ruled out in the case of testing for homogeneity or independence in a 2×2 table.

Yates has suggested a correction to be applied to the observed frequencies in a 2×2 table in case any expected frequency is found to be too small. This consists in increasing or decreasing the observed frequencies by $\frac{1}{2}$, in such a way that the marginal totals remain unaltered. If we consider the 2×2 table of Section 17.11, then it is necessary to increase a and d by $\frac{1}{2}$ each (while each of b and c is to be decreased by $\frac{1}{2}$) in case $ad < bc$. On the other hand, if $ad > bc$, then a and d have to be decreased by $\frac{1}{2}$ each, while each of b and c has to be increased by the same amount. Since the marginal totals remain unaltered, obviously the expected frequencies are not to be changed.

If Yates' correction is applied, the formula for χ^2 , corresponding to (17.38), will be

$$\frac{\{|ad - bc| - n/2\}^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)}.$$

Ex. 17.12 The breakdowns occurring during a year for each of 4 machines in a factory were classified as follows according to shift, there being 3 shifts daily. Judge whether the differences among the four distributions may be attributed to sampling fluctuations alone.

Shift	Machine			
	1	2	3	4
1	15	9	18	20
2	16	18	29	31
3	19	15	19	27
Total	50	42	66	78

Here we are to test for the homogeneity of the four frequency distributions corresponding to the four machines. Assuming that the required conditions are satisfied, we may apply the χ^2 test, with

$$\chi^2 = n \left(\sum_i \sum_j \frac{f_{ij}^2}{f_{i0} f_{0j}} - 1 \right), \text{ d.f.} = 6.$$

For the given samples, f_{ij} and the products $f_{i0} f_{0j}$ are shown in the table below together with

$$\frac{f_{ij}^2}{f_{i0} f_{0j}}.$$

(1) Observed frequency	(2) Row total \times col. total	(3) $(1)^2/2$
15	3,100	0.07258
9	2,604	0.03111
18	4,092	0.07918
20	4,836	0.08271
16	4,700	0.05447
18	3,948	0.08207
29	6,204	0.13556
31	7,332	0.19107
19	4,000	0.09025
15	3,360	0.06696
19	5,280	0.06837
27	6,240	0.11683
Total 236	—	1.01116

Hence

$$\begin{aligned} \chi^2 &= 236(1.01116 - 1) \\ &= 2.632 \end{aligned}$$

On comparing this with $\chi^2_{0.05, 6} = 12.592$ and $\chi^2_{0.01, 6} = 16.812$, it is seen that the hypothesis of homogeneity is to be accepted. That is to say, the observed differences among the four distributions may be attributed to sampling fluctuations alone.

Ex. 17.13 88 residents of an Indian city, who were interviewed during a simple survey, are classified below according to sex and according to whether they drink tea or not. Do these data reveal any association between sex and drinking of tea?

	Male	Female	Total
Drink tea	40	33	73
Do not drink tea	3	12	15
Total	43	45	88

In order to test for the null hypothesis that the two attributes are

independent in the population of residents, we shall use the χ^2 test. For the present data,

$$\begin{aligned}\chi^2 &= \frac{(ad - bc)^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)} \\ &= \frac{(381)^2 \times 88}{73 \times 15 \times 43 \times 45} = \frac{12,774,168}{2,118,825} = 6.029.\end{aligned}$$

Since two of the expected cell-frequencies are rather small, it would, however, be proper to use Yates' correction for continuity. On applying that correction, we have

$$\begin{aligned}\chi^2 &= \frac{\{|ad - bc| - n/2\}^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)} \\ &= \frac{(337)^2 \times 88}{73 \times 15 \times 43 \times 45} = \frac{9,994,072}{2,118,825} = 4.717.\end{aligned}$$

The tabulated values are

$$\chi^2_{.05,1} = 3.841$$

and $\chi^2_{.01,1} = 6.635$.

If we choose the 5% level of significance, then the null hypothesis should be rejected ; i.e., the two attributes should be taken to be associated in the population. It may be noted that although both of the computed χ^2 values are significant at the 5% level, the non-application of Yates' correction grossly exaggerates this significance.

Questions and exercises

17.1 Discuss the large-sample approximations of sampling theory and comment on their merits.

17.2 State the important uses of the χ^2 distribution in sampling theory.

17.3 (a) Supposing that r_i ($i=1, 2, \dots, k$) are sample correlations for independent random samples of sizes n_i ($i=1, 2, \dots, k$) from a bivariate normal population with unknown correlations ρ , indicate how you would combine the k sample correlations to get an estimate of ρ . *Partial ans.* Estimate of ρ to be obtained from

$$z_0 = \sum_i (n_i - 3) z_i / \sum_i (n_i - 3).$$

(b) Suggest a test, based on the z -transformation, for the hypothesis that k bivariate normal distributions have the same correlation coefficient.

$$\text{Partial ans. } \chi^2 = \sum_i (n_i - 3)(z_i - z_0)^2$$

$$= \sum_i (n_i - 3)z_i^2 - \left\{ \sum_i (n_i - 3)z_i \right\}^2 / \sum_i (n_i - 3) \text{ with } (k-1) \text{ d.f.}$$

17.4 A six-faced die was thrown 300 times, and the number of points obtained at each throw was recorded. In this way the following frequency distribution was formed. Use these data to test whether the die was unbiased.

Number of points per throw	1	2	3	4	5	6
Frequency	31	52	46	40	54	77

$$\text{Partial ans. } \chi^2 = 24.52.$$

17.5 The following table gives the number of weed seeds in 196 1-lb. packets of a variety of pulses and also the frequencies of the different classes as obtained by fitting a Poisson distribution.

Number of weed seeds	Observed frequency	Expected frequency
0	7	10.78
1	33	31.28
2	54	45.35
3	37	43.84
4	34	31.78
5	16	18.43
6	8	8.91
7	5	3.69
8	1	1.34
9	1	0.43
10 or more	0	0.17
Total	196	196.00

Test for the goodness of fit.

$$\text{Partial ans. } \chi^2 = 5.039.$$

17.6 Suppose two drugs are administered to each of n patients suffering from headaches. The reactions are summarised in the following table :

Drug 1

		Good	Not good	Total
Drug 2	Good	a	b	$a+b$
	Not good	c	d	$c+d$
Total		$a+c$	$b+d$	n

Starting from the marginals, $a+b$ and $a+c$, show that a comparison of the two drugs can be made by means of the statistic $(b-c)^2/(b+c)$, which is approximately a χ^2 with 1 d.f.

17.7 Suggest a suitable large-sample test for comparing the probabilities, say p_1 and p_2 , corresponding to two categories of a population with $k (> 2)$ categories. (The sample frequencies f_i , $i=1, 2, \dots, k$, may be assumed to be distributed in the multinomial form.)

[Hint : $E(f_i) = np_i$, $\text{var}(f_i) = np_i(1-p_i)$ and $\text{cov}(f_i, f_j) = -np_i p_j$, for $i \neq j$.]

17.8 Writing $p_i = a_i/T_a$, $\bar{p} = T_a/n = \sum_i T_i p_i / \sum_i T_i$ and $\bar{q} = 1 - \bar{p}$, show that the formula for χ^2 in the $2 \times k$ case (vide Section 17.11) can be reduced to the form

$$\chi^2 = \frac{1}{\bar{p}\bar{q}} \sum_i T_i (p_i - \bar{p})^2 = \frac{1}{\bar{p}\bar{q}} [\sum_i T_i p_i^2 - n \bar{p}^2].$$

Deduce also a third alternative form :

$$\chi^2 = T_a T_b \sum_i \left[\left(\frac{a_i}{T_a} - \frac{b_i}{T_b} \right)^2 \right] / T_i.$$

17.9 Suppose a linear combination of independent χ^2 statistics, $\chi_1^2, \chi_2^2, \dots, \chi_k^2$ (with d.f.'s v_1, v_2, \dots, v_k , respectively)—say, $\sum_i a_i \chi_i^2$, where $a_i > 0$ for each i —is to be approximated with a new statistic of the form $a \chi^2$ with d.f. v (say). What should be the values of a and v in order that the mean and variance of the approximating statistic may be the same as those of the original statistic?

$$\text{Ans. } a = \sum_i a_i^2 v_i / \sum_i a_i v_i, v = (\sum_i a_i v_i)^2 / \sum_i a_i^2 v_i.$$

1710 A manufacturer of watches claims that not more than 2 per cent of his products are defective. A retail dealer buys a batch of 720 watches from the manufacturer and finds on inspection that 26 of the watches are defective. Would you consider the manufacturer's claim justified by the data? *Partial ans* $\tau = 3.088$

1711 With the above data, obtain 95% confidence limits for the true percentage of defective watches *Ans* 2.25% and 4.97%

1712 A random sample of 826 students taken from Calcutta colleges in 1950 contained 143 women, while a random sample of 1,214 students taken in 1960 included 385 women. Examine whether there has been a significant progress among women in respect of collegiate education *Partial ans* $\tau = -7.291$

1713 A sample of 80 pigs was given a certain diet, and the gain in weight over a period of 20 days was recorded for each pig. The sample mean and the sample standard deviation came out to be 32.12 lb and 11.59 lb. Test whether the corresponding population values may be supposed to be 30 lb and 10 lb, respectively.

Partial ans $\tau(\text{for mean}) = 1.636$, $\tau(\text{for s.d.}) = 2.011$.

1714 The correlation coefficient between brother's height and sister's height for 53 brother-sister pairs was found to be 0.585. Is this coefficient significantly smaller than 0.8? *Partial ans* $\tau = -2.938$

1715 The correlation coefficient between sitting height and stature was found to be 0.7854 for a group of 70 adult Europeans. For a group of 39 adult Indians, on the other hand, the coefficient was 0.5209. Do the two coefficients differ significantly?

Partial ans $\tau = 2.332$

1716 1,072 schoolboys were classified according to intelligence, and at the same time their economic conditions were recorded. The results are shown in the following table. Judge whether there is any association between intelligence and economic conditions.

Economic conditions	Intelligence			Dull
	Excellent	Good	Mediocre	
Good	48	199	181	82
Not good	81	185	190	105

Partial ans $\chi^2 = 9.735$

17.17 During a smallpox epidemic, the following data were collected on the basis of a survey of 222 persons vaccinated against the disease. Do you think that the standard of vaccination affects the power to resist the disease?

	<i>Attacked with smallpox</i>	<i>Not attacked</i>	Total
<i>Well vaccinated</i>	33	120	153
<i>Badly vaccinated</i>	18	51	69
Total	51	171	222

$$\text{Partial ans. } \chi^2 = 0.549.$$

17.18 Supposing that the data of *Exercise 10.12* correspond to a random sample from an infinite population, test for the mutual independence of *A*, *B* and *C*. Also, test whether *B* may be considered independent of *A* and *C*, taken jointly.

$$\text{Partial ans. } \chi^2\text{'s have d.f.'s 4 and 3.}$$

17.19 There are two sections in a class, having 120 and 100 pupils, respectively. The following table gives their results in the half-yearly and annual examinations :

Section I			Section II		
Half-yearly Exam.			Half-yearly Exam.		
Annual Exam.			Annual Exam.		
	Passed	Failed		Passed	Failed
	Passed	48	Failed	21	8
	8	52		6	65

(a) For each section, test if the annual exam. results have any association with the results of the half-yearly exam.

$$\text{Partial ans. } \chi_1^2 = 53.571, \chi_2^2 = 42.739.$$

(b) Test whether the two sections may be regarded as random samples from the same population. *Partial ans.* $\chi^2 = 11.371$ (3 d.f.)

SUGGESTED READING

- [1] Cochran, W. G. "The χ^2 test of goodness of fit", *Ann. Math. Stat.*, 23, pp. 315-345, 1952.
- [2] Cochran, W. G. "Some methods of strengthening the common χ^2 tests", *Biometrika*, 41, pp. 417-451, 1954.

- [3] Fisher, R A *Statistical Methods for Research Workers* (Ch 4)
Oliver and Boyd, 1954
- [4] Goulden, C H *Methods of Statistical Analysis* (Chs 15, 16)
John Wiley, 1952, and Asia Publishing House, 1959
- [5] Irwin, J O "A note on the subdivision of χ^2 into components", *Biometrika*, 36, pp 130-134, 1949
- [6] Kimball, A W "Short-cut formulas for the exact partition of χ^2 in contingency tables", *Biometrics*, 10, pp 452-458, 1954
- [7] Maxwell, A E *Analysing Qualitative Data* (Chs 1-1),
Methuen, 1961
- [8] Mood, A M and Graybill, F A *Introduction to the Theory of Statistics* (Chs 10-12) McGraw-Hill, 1963, and Kōgakusha
- [9] Rao, C R *Advanced Statistical Methods in Biometric Research* (Chs 5, 6) John Wiley, 1952
- [10] Yule, G U and Kendall, M G *An Introduction to the Theory of Statistics* (Chs 17-20) Charles Griffin, 1950

18

NON-PARAMETRIC METHODS

18.1 Introduction

Most of the standard statistical techniques are optimum under certain standard assumptions, e.g. independence, homoscedasticity and normality.

Statistical methods have been called *robust* by G.E.P. Box if the inferences are not invalidated by the violation of the underlying assumptions. It is customary to justify the use of a normal theory criterion, in a situation where it cannot be guaranteed, by arguing that it is robust under non-normality. A fair number of enquiries have been made into the behaviour of standard tests when something other than the standard assumptions hold. When a remedy is required for non-normality, a procedure available is to transform the original variable. But this transformation must be based on some assumed form of the population.

Non-parametric methods are concerned with the treatment of standard statistical problems when the assumption of normality is replaced by general assumptions concerning the distribution function. Frequently, the variables are just assumed to come from a continuous distribution. Non-parametric methods, including K. Pearson's χ^2 test for goodness of fit, rank correlation and methods based on order statistics, provide a means of avoiding the normality assumption. But non-parametric methods do nothing to avoid the assumptions of independence or homoscedasticity.

Another term which has been freely interchanged with the term *non-parametric* is *distribution-free*. But it is better to maintain a distinction between the two. A statistical problem is parametric or non-parametric depending on whether we allow the parent distribution to depend on a number of parameters or leave it to be quite general, say just continuous. In other words, these terms depend on the formulation of the problem. On the other hand, if the method used to solve the problem depends neither on the form of the parent

distribution nor on its parameters, then the procedure is said to be distribution-free. If the method does not depend on the parameters of the parent distribution but depends on the form, we may call it a *parameter-free* procedure. Thus both parametric and non-parametric problems may or may not be distribution free. Distribution-free procedures were devised primarily for non-parametric problems. Hence we find these two terms being used interchangeably.

Non-parametric methods have certain advantages in that they require few assumptions, are simple to compute and can be used even in situations where actual measurements are unavailable and the data are obtained only as ranks.

18.2 Non-parametric estimation of location and dispersion

In non-parametric theory, the most frequently used measure of location is the population median θ , given by

$$F(\theta) = 0.5,$$

where the distribution function $F(x)$ is supposed to be continuous. The measure of dispersion used in non-parametric theory is the interquartile range,

$$\zeta_{3/4} - \zeta_{1/4},$$

where ζ_p is defined by $F(\zeta_p) = p$. (Thus the median is $\zeta_{1/2}$)

Point estimation

A point estimate of the p quantile ζ_p is given by the corresponding sample quantile

Let $x_{(r)}$ be the r th smallest sample observation in a sample of size n also called the r th *order statistic*. Then $x_{(r)}$ is a point estimate of the $r/(n+1)$ -quantile. If a population quantile relates to a p lying between $r/(n+1)$ and $(r+1)/(n+1)$, then its point estimate is obtained by linear interpolation from $x_{(r)}$ and $x_{(r+1)}$. Thus the sample median is a point estimate of population location and the sample interquartile range is a point estimate of population dispersion.

Interval estimation

Interval estimates for a population quantile can be easily obtained with the help of the binomial distribution.

Note that the probability is p for an observation to fall to the left of ζ_p and $(1-p)$ for an observation to lie to the right of ζ_p .

Then

$$P[x_{(s)} \geq \zeta_p] = \sum_{i=0}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \quad \dots \quad (18.1)$$

and

$$P[x_{(r)} \leq \zeta_p] = \sum_{i=r}^n \binom{n}{i} p^i (1-p)^{n-i} \quad \dots \quad (18.2)$$

define one-sided confidence intervals for ζ_p in terms of order statistics. Thus $(-\infty, x_{(s)})$ and $(x_{(r)}, \infty)$ are one-sided confidence intervals for ζ_p with confidence coefficients given by (18.1) and (18.2), respectively. If $s > r$, we obtain, from (18.1) and (18.2),

$$P[x_{(r)} \leq \zeta_p \leq x_{(s)}] = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}, \quad \dots \quad (18.3)$$

and this provides a two-sided confidence interval for ζ_p obtained from the order statistics $x_{(r)}$, $x_{(s)}$, with confidence coefficient given by (18.3). For the median, we have to take $p = .50$.

Ex. 18.1 Let us obtain a confidence interval for the median for a sample of size 10 and based on $x_{(3)}$ and $x_{(8)}$.

We have, from (18.3),

$$\begin{aligned} P[x_{(3)} \leq \zeta_{1/2} \leq x_{(8)}] &= \sum_{i=3}^7 \binom{10}{i} \left(\frac{1}{2}\right)^{10} \\ &= 1 - \sum_{i=0}^2 \binom{10}{i} \left(\frac{1}{2}\right)^{10} - \sum_{i=8}^{10} \binom{10}{i} \left(\frac{1}{2}\right)^{10} \\ &= 1 - \frac{7}{2^6} \simeq 0.90. \end{aligned}$$

Hence $(x_{(3)}, x_{(8)})$ is approximately a 90% confidence interval for $\zeta_{1/2}$.

18.3 Tolerance interval

Let L_1 and L_2 be two functions of sample values with $L_1 < L_2$, such that the random interval (L_1, L_2) has a probability β of containing at least $100\gamma\%$ of the population ; that is to say,

$$P\left[\int_{L_1}^{L_2} f(x) dx \geq \gamma\right] = P[F(L_2) - F(L_1) \geq \gamma] = \beta. \quad \dots \quad (18.4)$$

This interval (L_1, L_2) is called a $100\gamma\%$ tolerance interval with probability β . The functions L_1 , L_2 are called the lower and upper tolerance limits. For general statistics L_1 and L_2 , the probability β in (18.4) depends on the form of the distribution function $F(x)$. Hence

generally tolerance intervals are not distribution free. However, if each of L_1 and L_2 is one of the order statistics, then the probability β does not depend on the form of $F(x)$, i.e., tolerance intervals based on order statistics are distribution-free tolerance intervals. The equation

$$P[F(x_{(n)}) - F(x_{(r)}) \geq \gamma] = \sum_{i=0}^{s-r-1} \binom{n}{i} \gamma^i (1-\gamma)^{n-i}$$

determines β as a function of r , s , n and γ only, with $L_1=x_{(r)}$ and $L_2=x_{(s)}$.

In mass production of articles, a certain amount of variation from the aimed-at value is inevitable. After production has started, statisticians often calculate tolerance intervals, on the basis of a sample drawn from the production line, which cover with a given probability (β) a certain fraction (γ) of the items being produced.

18.4 Non-parametric tests for location

We shall now consider some non parametric tests for the location parameter (median) of a population or for comparing the locations of two populations. In these tests, we do not make the usual normality assumptions about the parent population(s).

Sign test (test for the location parameter θ of a population)

We have n observations x_1, x_2, \dots, x_n , such that the x 's are mutually independent and each x comes from a population (not necessarily the same) which is continuous in the vicinity of θ , i.e., $P[x_i < \theta] = P[x_i > \theta] = 1/2$, $i=1, 2, \dots, n$.

To test $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$ (or $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$), where θ_0 is some specified number, we replace each sample observation exceeding θ_0 by a *plus* sign and each sample observation less than θ_0 by a *minus* sign. Sample values equal to θ_0 are ignored. The null hypothesis can be tested by testing the equivalent null hypothesis that the plus and minus signs come from a binomial population with parameter value 0.5. The critical value is based on the binomial distribution with $m = (\text{number of plus signs} + \text{number of minus signs})$ and $p=0.5$. Because of the discreteness of the binomial distribution, the critical values may not correspond exactly to the stated levels of significance. One-tailed cumulative binomial probabilities are given in Table VI in the Appendix. To get the two-tailed probability, the tabular value is to be doubled.

Ex. 18·2 Suppose that we want to test the hypothesis that the median length (θ) of ear-head of a variety of wheat is $\theta_0=9\cdot9$ cm. against the alternatives that $\theta \neq 9\cdot9$ cm., with $\alpha=0\cdot05$, on the basis of the following 20 ear-head measurements :

$$\begin{array}{ccccccccccccc} 9\cdot3, & 8\cdot8, & 10\cdot7, & 11\cdot5, & 8\cdot2, & 9\cdot7, & 10\cdot3, & 8\cdot6, & 11\cdot3, & 10\cdot7 \\ - & - & + & + & - & - & + & - & + & + \\ 11\cdot2, & 9\cdot0, & 9\cdot8, & 9\cdot3, & 9\cdot9, & 10\cdot3, & 10\cdot0, & 10\cdot1, & 9\cdot6, & 10\cdot4 \\ + & - & - & - & - & + & + & + & - & + \end{array}$$

We first determine the signs, which are posted below the measurements. We find 9 minus and 10 plus signs and one measurement equal to $\theta_0=9\cdot9$. So we are to test whether 9 minus and 10 plus signs support the hypothesis $H_0 : \theta_0=9\cdot9$ or, equivalently, to judge how likely are 9 successes (the smaller of the two sign frequencies) to occur in 19 trials from a binomial distribution with $p=0\cdot5$. The critical region for the two-sided test is given by

$$r > K_{\alpha/2} \text{ and } r < K'_{\alpha/2},$$

where r = number of minus signs and $K_{\alpha/2}$ is the smallest integer and $K'_{\alpha/2}$ is the largest integer such that

$$\sum_{x=K_{\alpha/2}}^n \binom{n}{x} \left(\frac{1}{2}\right)^n \leq \alpha/2$$

and

$$\sum_{x=0}^{K'_{\alpha/2}} \binom{n}{x} \left(\frac{1}{2}\right)^n \leq \alpha/2.$$

From Table VI in the Appendix, we find that $K_{0.025}=15$ and $K'_{0.025}=4$ for $n=19$ and $p=0\cdot5$. Since for this example $r=9$, the null hypothesis is to be accepted.

In the above example, the critical region for the one-sided test against the alternatives $H : \theta > 9\cdot9$ cm. will be given by $r > K_\alpha$, where K_α is the smallest integer such that

$$\sum_{x=K_\alpha}^{19} \binom{19}{x} \left(\frac{1}{2}\right)^{19} \leq 0\cdot05.$$

If $n > 25$, then the normal approximation to the binomial may be used to perform the test. In that case, the probability of r or fewer successes in n trials will be approximately given under the null hypothesis by $\Phi(\tau)$, where

$$\tau = \frac{r-n/2}{\sqrt{n/4}} = \frac{2r-n}{\sqrt{n}}.$$

Now, we consider some non-parametric tests for comparing the locations of two populations. First, we consider the sign test which may be regarded as an alternative to the paired *t* test.

Sign test (paired sample test)

The paired *t* test for the significance of the observed difference of means in paired samples assumes that all the paired differences are independently and normally distributed with a common variance. But sometimes the assumptions of normality and homoscedasticity are too strong. And in those cases we can apply the sign test by supposing that the median of the population differences is θ (the populations need not be the same for all pairs, but the population of differences is assumed to be continuous in the vicinity of its median θ).

To test the null hypothesis $H_0: \theta = \delta$ against appropriate one-sided or two sided alternatives, we replace the quantities $x_{1i} - x_{2i} - \delta$ by their *sigs* and then proceed exactly as in the one-sample sign test (x_1 and x_2 being the values corresponding to the *i*th pair).

Ex 18.3 Consider the problem in Ex 16.11

The gains in weight (*d*) for the ten boys are

$$6, 8, 1, 3, 3, 1, 3, -2, 4, -2$$

The null hypothesis is $H_0: \theta = 0$ and the alternatives are $H: \theta > 0$, where θ is the median of the population of differences. There are two minus signs in 10 non zero values. On the null hypothesis, the expected number of minus signs (or plus signs) among the differences in a sample of 10 pairs is 5. The sampling distribution of the number of minus signs is the binomial distribution with probability of a minus sign 0.5. From Table VI we find that the probability of 2 or fewer minus signs is 0.547. So the null hypothesis is accepted at the 5% level.

For large sample size, say for $n > 25$, the normal approximation to the binomial may be used.

Wilcoxon signed rank test (paired sample test)

Here also our problem is the same as in the paired sample sign test.

This test for paired observations is more powerful than the sign test for paired observations since the former takes account of the magnitude of the difference between the members of a pair.

The observed differences $d_i = x_{1i} - x_{2i} - \delta$ are ranked in increasing order of absolute magnitude and then the ranks are given the signs of the corresponding differences.

We assume that the differences d_1, d_2, \dots, d_n are mutually independent and that all d 's come from continuous populations (not necessarily the same) that are symmetric about zero.

If the null hypothesis is true, then we would expect the sum of the positive ranks to be approximately equal to the absolute value of the sum of the negative ranks.

Let T be the smaller sum of the two absolute rank-sums. Compare T with the critical rank-sum T_0 given in Table VII in the Appendix for specific probabilities and for specified n (the number of pairs which give non-zero differences). If $T \leq T_0$, then H_0 is to be rejected at the specified level; if $T > T_0$, then there is not sufficient evidence to doubt the null hypothesis at the specified level.

For $n > 25$, T is approximately normally distributed under H_0 , with

$$\begin{aligned} E(T) &= n(n+1)/4 \\ \text{and} \quad \text{var}(T) &= n(n+1)(2n+1)/24. \end{aligned} \quad \left. \right\} \quad \dots \quad (18.5)$$

Ex. 18.4 Consider the problem of Ex. 18.3. The differences and signed ranks are as follows :

d_i	6	8	1	3	3	1	3	-2	4	-2
Signed rank	9	10	1.5	6	6	1.5	6	-3.5	8	-3.5

The sums are 48 and -7; hence T becomes 7. In Table VII we have, for $n=10$ and $p=0.025$ (one-sided), $T_0=8$. Since $T < T_0$, we may conclude that the null hypothesis that there is no effect of diet is rejected in favour of the (one-sided) alternative hypotheses at the 2.5% level.

Mann-Whitney U-test : (test for location of two independent populations)

The sign test and the signed-rank test are applicable when the observations are paired. A useful test procedure, when the observations of the two samples are independent (not paired), for testing whether their location parameters are the same, i.e. for testing $H_0 : \theta_1 = \theta_2$, is given by the *U*-test of Mann-Whitney and Wilcoxon.

The appropriate null hypothesis in this case is that the two independent samples come from identical populations and the alternative is that one population is shifted to the right (or left) or that the populations differ only in location.

We take independent samples of sizes n_1 and n_2 from the two populations and rank the combined sample of size $n=n_1+n_2$. If the two populations are identical, then we would expect that the samples will intermingle in a regular way. On the other hand, if there be any sizeable difference between the location parameters, then most of the lower ranks will be occupied by the observations from one sample, while most of the higher ranks will be occupied by the observations from the other sample.

Let R_1 and R_2 represent the sums of the ranks of the observations from the two samples of sizes n_1 and n_2 , respectively. Then

$$R_1 + R_2 = n(n+1)/2,$$

where $n=n_1+n_2$.

The statistic used for making the test is

$$U = n_1n_2 + n_1(n_1+1)/2 - R_1 \quad (18.6a)$$

or, equivalently,

$$U = n_1n_2 + n_2(n_2+1)/2 - R_2 = n_1n_2 - U \quad (18.6b)$$

Under H_0 , both the samples come from the same population. So R_1 is the sum of n_1 positive integers selected at random from the first n positive integers. Then, under H_0 ,

$$\left. \begin{aligned} E(R_1) &= n_1(n+1)/2 \text{ and } \text{var}(R_1) = n_1n_2(n+1)/12, \\ \text{and hence} \end{aligned} \right\} \quad (18.7)$$

$$E(U) = n_1n_2/2 \text{ and } \text{var}(U) = n_1n_2(n+1)/12$$

For values of n_1 and n_2 which are moderately large (say 9 or more), under H_0 , U is approximately normal with mean and variance given by (18.7). For small values of n_1 and n_2 (none larger than 8), Mann and Whitney have given a table of exact probabilities. For n_2 (the size of the larger sample) between 9 and 20 and $n_1 \leq 20$, Auble has given a table of critical values of U . If the computed value of U is less than or equal to the tabulated value, then we reject the null hypothesis at the stated level of significance for a one-tailed test. For a two-tailed test, the level of significance is to be doubled.

Formulae (18.6a) and (18.6b) yield different values of U . It is the smaller of the two which is needed for performing the test with the help of Table VIII.

Ex. 18.5 The following are the numbers of defective items produced by workman *A* and workman *B*. Use the *U*-test with $\alpha=0.02$ to test the null hypothesis that the samples are drawn from identical populations against the alternative that the populations differ in location only.

Workman *A* : 26, 27, 31, 26, 19, 21, 20, 25, 30 ;

Workman *B* : 23, 28, 26, 24, 22, 19.

Here

$$n_1=6, n_2=9, n=n_1+n_2=15$$

and

$$R_1=42.5, U=32.5; R_2=77.5, U'=21.5.$$

From Table VIII we find that for (size of larger sample) $n_2=9$ and (size of smaller sample) $n_1=6$ for a two-tail test at level 0.02, the significant value is 7. Since 21.5 (the smaller of *U* and *U'* for this problem) is greater than 7, we have no reason to believe that the samples are drawn from identical populations.

Median test

Like the *U*-test, the median test also tests the null hypothesis that two independent samples are from identical populations against the alternative that they have different location parameters. This test is sensitive to differences in location.

Let there be two samples of sizes n_1 and n_2 from the two populations under comparison. We order the $n=n_1+n_2$ observations in the two samples combined and determine the combined sample median $\hat{\theta}$. Then we count the number of observations in each of the two samples that lie below $\hat{\theta}$ and the number of those which do not lie below $\hat{\theta}$. These can be put in a 2×2 contingency table :

	Observations		Total
	$<\hat{\theta}$	$\geq\hat{\theta}$	
Sample 1	m_1	n_1-m_1	n_1
Sample 2	m_2	n_2-m_2	n_2
Total	m_1+m_2	$n-n_1-n_2$	n

If n_1 and n_2 are small, one may obtain the exact probability of

obtaining this table, viz

$$P(m_1, m_2) = \frac{\binom{n_1}{m_1} \binom{n_2}{m_2}}{\binom{n}{m_1 + m_2}}, \quad (188)$$

whereas for moderately large n_1, n_2 (say each greater than 10), we may use the χ^2 statistic with 1 d.f., where

$$\chi^2 = \frac{n[m_1(n_2 - m_2) - m_2(n_1 - m_1)]^2}{n_1 n_2 (m_1 + m_2)(n - m_1 - m_2)} \quad (189)$$

[*vide* formula (17.38)]

Ex 18.6 Perform the median test for the data shown in Ex 18.5

In this case the sample median (combined) = 25, and we have the following table

	Observations		Total
	<25	>25	
Workman A	3	6	9
Workman B	4	2	6
Total	7	8	15

The exact probability of getting such a table, using (18.8), is

$$\frac{\binom{9}{3} \binom{6}{4}}{\binom{15}{7}} = \frac{140}{715}$$

The probability of getting less likely distributions (in one direction) than the observed one will be obtained from the following two tables

	<25	>25	Total
A	2	7	9
B	5	1	6
Total	7	8	15

and

	<25	>25	Total
A	1	8	9
B	6	0	6
Total	7	8	15

The exact probabilities of these two distributions are

$$\frac{\binom{9}{2} \binom{6}{5}}{\binom{15}{7}} = \frac{24}{715} \text{ and } \frac{\binom{9}{1} \binom{6}{6}}{\binom{15}{7}} = \frac{1}{715}.$$

So the exact probability of getting the observed distribution or less likely ones in either direction is, by symmetry,

$$2\left[\frac{140}{715} + \frac{24}{715} + \frac{1}{715}\right] = .462.$$

As $.462 > .05$, we accept the null hypothesis that the two populations are identical.

The chief objection to the exact method is the computational labour involved. But in cases where the probability of the observed table exceeds the level of significance (as in the present case, where it is $\frac{140}{715} \simeq .2$), we need not obtain the probabilities of more extreme cases, since the sum of probabilities of the observed table and the more extreme cases will also exceed the level of significance.

18.5 Two-sample non-parametric tests for dispersion

Let x_1, x_2, \dots, x_m and y_1, y_2, \dots, y_n be two samples of independent observations drawn from two populations with distribution functions $F(x)$ and $G(y)$, respectively. We assume that F and G are continuous and that they are the same in all respects, except perhaps for the scale parameter. So we test the null hypothesis $H_0: F=G$ against the alternative $H: F \neq G$. (The c.d.f.s F and G differ only in scale values.) We consider the test of Mood and Sukhatme.

Mood's rank test for dispersion

Let the m observations from $F(x)$ and the n observations from $G(y)$ be ranked from 1 to $(m+n)$ and let W be the sum of squares of the deviations of the y ranks from the average rank $\frac{m+n+1}{2}$:

$$W = \sum_{i=1}^n \left(r_i - \frac{m+n+1}{2} \right)^2,$$

where r_i is the rank of y_i in the combined sample of $(m+n)$ observations. If the x 's are more (less) dispersed relative to the y 's, W will be relatively small (large). We reject the hypothesis H_0 if W is too

large or too small As shown by Mood, under the null hypothesis,

$$\text{and } \begin{aligned} E(W) &= n(s^2 - 1)/12 \\ \text{var}(W) &= mn(s+1)(s^2 - 4)/180, \end{aligned} \quad \} \quad (18.10)$$

where, for short, we write s for $m+n$.

For large samples, the normal approximation may be used for testing H_0 , and then W is taken to be approximately normal with mean and variance given by (18.10)

Sukhatme's test for dispersion

The test statistic may be defined as

$$T = \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m \phi(x_i, y_j),$$

where $\phi(x, y) = 1$ if $\begin{cases} \text{either } 0 < x < y \\ \text{or } y < x < 0 \end{cases}$
 $= 0$ otherwise

We reject H_0 : $F=G$ if T is either too large or too small. The mean and variance of T , under H_0 , as given by Sukhatme, are

$$\text{and } \left. \begin{array}{l} E(T) = 1/4 \\ \text{var}(T) = (m+n+7)/48mn \end{array} \right\} \quad (18.11)$$

For large samples, the normal approximation may be used to test H_0 . T is taken to be approximately normal with mean and variance given by (18.11), for large n .

The above two tests assume knowledge about the relative location of the two populations. It is, however, possible to modify the tests for dispersion without making any assumption about the location parameters.

18.6 A general non-parametric test for two independent samples

In the above two sections, we have considered two-sample non-parametric tests for special alternatives, e.g. that the two populations differ only in respect of location or dispersion. In this section, we consider a general two sample test for testing the null hypothesis that two independent samples come from identical populations against the alternative that the two populations differ in any manner whatsoever—in location, in dispersion, in skewness, in kurtosis or in any other way.

If we want to test whether the two populations differ only in one particular respect, say either in location or in dispersion, then we should use a test for location or dispersion. The test being considered here will be less powerful in disclosing differences of a particular kind, for it is a test for any sort of difference and not for a particular type of difference.

Wald-Wolfowitz run test

When we wish to test whether two independent samples have been drawn from the same population against the alternative that the two populations differ (in any manner), this test is used. The test assumes that the underlying population distribution is continuous.

Let us draw a random sample (observations denoted by x 's) of size n_1 from the first population and a random sample (observations denoted by y 's) of size n_2 from the second population. We then arrange the $n=n_1+n_2$ observations from the two samples combined in order of magnitude ; thus we might have the arrangement :

$$y \underline{x} \underline{x} \underline{x} y \underline{y} \underline{x} \underline{y} \underline{x} \dots$$

A *run* is a sequence of values of the same kind bounded by values of the other kind. Thus, in the above sequence, we have a run of one y followed by a run of three x 's ; this in turn is followed by a run of two y 's, and so on. Let r be the total number of runs in the group of n observations. Then if the two samples are from the same population, the two samples are expected to be thoroughly mixed and hence r is expected to be large ; whereas r is expected to be relatively small if the populations are not the same. Hence we reject H_0 if r is too small.

To obtain the sampling distribution of r , we observe that there are $\binom{n_1+n_2}{n_1} = \binom{n_1+n_2}{n_2}$ different possible arrangements of the n_1 x 's and n_2 y 's in a line and all these arrangements are equally likely under the null hypothesis. Next we find the number of arrangements of n_1 x 's and n_2 y 's giving a total of r runs. Let $r=2d$ (even), then we must have d runs of x 's and d runs of y 's. To get d runs of x 's we are to divide the n_1 x 's in d groups and find all ordered d -part partitions of n_1 things. This is obtained (with the help of generating function) as follows : The required number of d -part partitions of n_1 things is

the coefficient of t^d in

$$(t+t^2+t^3+\dots)^d = t^d (1-t)^{-d} \\ = t^d \sum_{i=0}^{\infty} \binom{d-1+i}{d-1} t^i$$

and is $\binom{n_1-1}{d-1}$. In a similar way, the number of d part partitions of n_2 's is $\binom{n_2-1}{d-1}$. Hence the total number of ways of getting $r=2d$ runs is $2\binom{n_1-1}{d-1}\binom{n_2-1}{d-1}$, since d runs of x 's and d runs of y 's can be combined in two ways to give $r=2d$.

$$\text{So } P[r=2d] = 2 \frac{\binom{n_1-1}{d-1} \binom{n_2-1}{d-1}}{\binom{n_1+n_2}{n_1}} \quad (18.12)$$

By a similar argument, we have, for r odd,

$$P[r=2d+1] = \frac{\binom{n_1-1}{d} \binom{n_2-1}{d-1} + \binom{n_1-1}{d-1} \binom{n_2-1}{d}}{\binom{n_1+n_2}{n_1}} \quad (18.13)$$

To perform the test of the null hypothesis at the level α , we find r_0 (as close as possible) such that

$$P[r \leq r_0] = \alpha$$

and reject H_0 if the observed r does not exceed r_0 .

Tables of critical values of r , based on (18.12) and (18.13), are given by Swed and Eisenhart. Any value of r which is equal to or smaller than that shown in Table IX is significant at the .05 level.

For large values of n_1 and n_2 , the sampling distribution of r is approximately normal with mean and variance given, under H_0 , by

$$E(r) = \frac{2n_1 n_2}{n_1 + n_2} + 1 \quad \left. \begin{aligned} \text{and} \quad \text{var}(r) &= \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \end{aligned} \right\} \quad (18.14)$$

An approximate test can be performed by using the above approximation if (1) n_1 and n_2 are both larger than 10 or (2) either n_1 or n_2 is larger than 20.

Since the underlying distribution is assumed to be continuous, no ties should occur. But in practice, due to measurement approximations, ties may occur. If ties are within the same sample, then there is no problem as the number of runs is not affected. But if there be ties among observations from both the samples, we cannot get a unique value of r . In that case we break ties in all possible ways and find the corresponding r 's. If all these different r 's lead to the same conclusion at the selected level α , than there is no problem.

There is, however, some difficulty when the different r 's associated with different ways of breaking lead to different decisions. If the number of ties between observations from the two samples is large, then the run test is not to be recommended.

Ex. 18.7 Scores on a clerical aptitude test administered to a batch of 6 Secretariat and 7 Directorate clerks are given below. Test whether the two groups of clerks have the same score distribution in the population.

Scores for Secretariat Clerks	40,	35,	52.	60,	46,	55
Scores for Directorate Clerks	47,	56.	42,	57,	50,	57, 62

We arrange the scores from the two groups in order of magnitude, noting the groups to which each score belongs :

35, 40, 42, 46, 47, 50, 52, 55, 56, 57, 57, 60, 62
S S D S D D S S D D D S D

Thus we have $n_1=6$, $n_2=7$, $n=6+7=13$ and 4 runs of S 's and 4 runs of D 's, giving $r=8$. The critical value of r at the 5% level, from Table IX, is 3. Since the observed value (8) is greater than the critical value (3), we accept the null hypothesis at the 5% level.

18.7 One-sample run test for randomness

In order to arrive at some conclusion about a population on the basis of a sample, we need a random sample. To test the hypothesis that a sample is random, we use the *order* in which the observations were originally obtained. The technique to be used here is based on the theory of runs. For example, in sampling inspection we have runs of defective and non-defective items, in tossing a coin we have runs of heads and tails. The total number of runs in a sample of a

given size indicates whether a sample is random or not. Too few or too many runs may indicate a time trend or systematic short-term cyclical fluctuations in the observations.

We find the number of runs (r) in the group of n observations in the sample. The observations may be heads or tails in a coin-tossing experiment, the good or bad items in a sampling inspection of items from a lot or the measurements below the sample median (−) and the measurements above the sample median (+).

The sampling distribution of r arising from random sampling is known (Section 18.6). Using this, we decide whether a given sample has more or fewer runs than expected under random sampling.

Ex 18.8 The following data were obtained on the sexes of the 15 students standing in the queue in front of the admission counter of a university. Ascertain whether the arrangement of sexes was a random one or not.

Order of 15 males (m) and females (f) in the queue

$m f m m m f f m m m m f m f f$

We have (using the notation of Section 18.6) $n_1=9$, $n_2=6$, $n=9+6=15$ and 4 runs of m 's and 4 runs of f 's, giving $r=4+4=8$. The critical value of r at the 5% level, from Table IX, is 4. Since the observed value is greater than the critical value, we accept the null hypothesis that in the queue the order of the sexes was a random one.

18.8 A non-parametric measure and a test of association

Spearman's rank correlation coefficient (*vide* Section 13.2) is a measure of association based on ranks. It can be used to test whether the two variables X and Y are independent. We make no assumption about the distributions of X and Y . Like the product moment correlation coefficient, its value also ranges from -1 to 1 . A value of $+1$ indicates perfect agreement and a value of -1 indicates perfect disagreement between the two series of ranks.

The formula for calculating r_R when there is no tie is given in (13.1), while that for tied ranks is given in (13.3). Under the assumption that the individuals, who were ranked, were randomly

drawn from some population, we can test the null hypothesis that the two variables X and Y are not associated in the population. If the null hypothesis is true, then for a given rank order of the Y values (X values) all possible rank orders of the X values (Y values) are equally likely to occur. For n individuals there are $n!$ possible rankings of X values, and since they are equally likely, the probability of the occurrence of any particular ranking of the X values with a given ranking of Y values is $\frac{1}{n!}$.

The probability of the occurrence of any given value of r_R , under H_0 , is proportional to the number of permutations of X values giving rise to that value of r_R , since to each ranking of Y values there corresponds a value of r_R .

For $n=2$, only two values of r_R are possible, viz. +1 and -1, and each has the probability of occurrence $\frac{1}{2}$ under H_0 . For $n=3$, the possible values of r_R are -1, $-\frac{1}{2}$, $\frac{1}{2}$ and +1, with respective probabilities of occurrence $\frac{1}{6}$, $\frac{1}{3}$, $\frac{1}{3}$ and $\frac{1}{6}$ under H_0 . Table X of the Appendix gives critical values of r_R for n from 4 to 30, which were arrived at by a similar method. If an observed value of r_R equals or exceeds the tabulated value, then that value of r_R is significant (for a one-sided test) at the stated level.

The table may be used for two-sided tests also. In this case, the hypothesis of independence is rejected whenever $|r_R|$ is greater than the tabulated value. The associated level of significance is then double the value given in the table.

Kendall has shown that when n is 10 or larger, the statistic

$$t = r_R \sqrt{n-2} / \sqrt{1-r_R^2} \quad \dots \quad (18.15)$$

may be taken to be distributed as a Student's t with $n-2$ d.f.

Ex. 18·9 For the data of Ex. 13·1, let us find out whether the rankings by the two judges are independent or not.

We have already seen in Ex. 13·1 that $r_R=0.394$, with $n=10$. By referring to Table X, we find that the 5% value of r_R for $n=10$ is 0.564. So we do not reject the hypothesis of independence in ranking at the 5% level of significance.

Ex. 18·10 Consider the data of Ex. 13·2 and test the hypothesis of independence in ranking by the supervisors.

Here, as already obtained in Ex 13.2, $r_R=0.956$ with $n=12$. Since $n > 10$, we may use here the large sample test given by (18.15)

$$\begin{aligned} t &= 0.956 \sqrt{10} / \sqrt{1 - (0.956)^2} \\ &= 0.956 \sqrt{\frac{10}{1 - 0.913936}} \\ &= 0.956 / \sqrt{0.0086064} \\ &= 0.956 / 0.09277 = 10.30 \text{ with } 10 \text{ d.f} \end{aligned}$$

This is significant at the 5% level. The same result is obtained by performing the exact test and using Table X.

Questions and exercises

18.1 State clearly the difference between a parametric and a non parametric problem. What is meant by 'robustness' of a statistical procedure?

18.2 What are the differences between non parametric, parameter free and distribution free procedures?

18.3 Describe how one can obtain a point estimate of a population quantile and its confidence interval with confidence coefficient $1-\alpha$.

18.4 What is meant by a $100\gamma\%$ tolerance interval with probability β ? What is a distribution free tolerance interval and how can it be obtained?

18.5 Discuss the sign test for the location parameter of a population. Show how it can be adapted to the case of paired samples. Which one of the two tests, (i) sign test and (ii) Wilcoxon signed rank test, for paired samples is more powerful? Explain why it is so.

18.6 Explain the Mann Whitney U test for two independent samples. Give the large-sample approximation to the test.

18.7 Explain why no ties should theoretically occur in the discussion of a non parametric test and yet why ties are found in practice. How are these dealt with?

18.8 Give some two-sample non parametric tests for dispersion along with their large-sample approximations.

18.9 Discuss the Wald-Wolfowitz run test. Show how the theory of runs may be used to test for the randomness of a sample.

18.10 Explain the use of Spearman's rank correlation coefficient as a test of association. Give the large-sample approximation to the test.

18.11 To determine the mileage of a type of truck, 6 trucks were run and the mileage of each obtained with a gallon of gasoline was as follows :

$$21, 19, 22, 18, 20, 24.$$

Use the sign test to examine whether the average number of miles run with a gallon of gasoline by trucks of this type is 20, the alternative hypotheses being that it is greater than 20.

Ans. H_0 is accepted at 5% level.

18.12 *Quantile test.* Consider the data of Ex. 18.2 and test whether the 40th percentile of ear-head distribution is 9.1 cm. against the alternative that it is different from 9.1 cm.

[*Hint :* Let r be the number of sample observations less than the specified p -quantile value, while s observations are greater than it. Then r has the binomial distribution with parameters $n=r+s$ and p . The critical value is based on this binomial distribution.]

The sign test is a particular case of the quantile test with $p=0.50$.]

Ans. H_0 is accepted at 5% level.

18.13 A manufacturer of electric bulbs claims that he has developed a new production process which will increase the mean efficiency (in suitable units) from the present value of 9.03. The results obtained from an experiment with 15 bulbs from the new process are given below :

9.29	9.76	8.93
10.15	12.05	9.02
8.69	12.38	10.87
11.25	9.08	10.00
11.47	10.25	11.56

Do we have reasons to believe that the efficiency has been increased ?

Ans. H_0 is rejected at 5% level.

1814 Below are given the marks obtained by a group of 20 students in a subject in a college test and in the subsequent public examination. Test at the 1% level whether the group has improved its mean performance from the college test to the public examination by using (a) the sign test and (b) the Wilcoxon signed-rank test.

Serial No.	Marks obtained in		Serial No.	Marks obtained in	
	College test	Public examination		College test	Public examination
1	183	133	11	123	126
2	175	193	12	121	141
3	134	170	13	175	103
4	170	164	14	133	126
5	183	199	15	144	146
6	167	160	16	109	155
7	120	168	17	165	162
8	175	158	18	144	161
9	126	162	19	164	182
10	187	176	20	125	119

Partial ans (a) 9 minus signs, accept H_0 ,
 (b) $T=82$, $N=20$, critical value is 43

1815 A firm is advertising that it has been successful in designing a new home automatic clothes washer which is more effective in removing dirt than the most popular washer now in use. And in support of its claim, it is also displaying the following data of the dirt removed (in suitable units) by the most popular washer and the new washer for 14 equally-sized and equally-soiled loads of clothes which were washed with the same soap and for the same length of time, 7 loads being washed by each washer.

Dirt removed by	
Popular washer	New washer
13	10
10	11
9	12
12	13
11	9
10	14
8	12

Do you have reasons to believe that the firm's claim is genuine ? (Use both the median test and the Mann-Whitney test.)

Partial ans. (a) Exact probability for median test is 0.283 ;
 (b) $U=33.5$ for $N=14$.

18.16 In a class of 12 newly admitted students, the arrangement of students with father living and those with father dead according to the order of admission is

$$LLDLLLDDLDLL,$$

where L denotes 'father living' and D denotes 'father dead'.

Does the above arrangement throw any doubt on the randomness of arrangement of the two types of students ?

Ans. $r=7$, $n_1=8$, $n_2=4$; H_0 is accepted at the 5% level.

18.17 Using Spearman's rank correlation, test for the association between the judgements of the two judges in *Exercise 13.7*.

Ans. H_0 is rejected at the 5% level, but accepted at the 1% level.

SUGGESTED READING

- [1] Auble, D. "Extended tables for the Mann-Whitney statistic", *Bulletin of the Institute of Educational Research at Indiana University*, 1, No. 2, 1953.
- [2] Hollander, M. and Wolfe, D. A. *Nonparametric Statistical Methods* (Chs. 3, 4). John Wiley, 1973.

- [3] Mann, H. B. and Whitney, D. R. "On a test of whether one of two random variables is stochastically larger than the other", *Ann. Math. Stat.*, 18, pp. 52-54, 1947.
- [4] Mood, A. M. and Graybill, F. A. *Introduction to the Theory of Statistics* (Ch. 16) McGraw-Hill, 1950, and Kōgakusha.
- [5] Olds, E. G. "The 5% significance levels for sums of squares of rank differences and a correction", *Ann. Math. Stat.* 28, pp. 117-118, 1949.
- [6] Siegel, S. *Nonparametric Statistics for the Behavioral Sciences* (Chs. 4—6, 9). McGraw-Hill, 1956.
- [7] Sukhatme, B. V. "On some two sample non-parametric tests for variance", *Ann. Math. Stat.* 28, pp. 188-194, 1957.
- [8] Swed, F. S. and Eisenhart, C. "Tables for testing randomness of grouping in a sequence of alternatives", *Ann. Math. Stat.*, 14, pp. 66-87, 1943.
- [9] Wilcoxon, F. and Wilcox, R. A. *Some Rapid Approximate Statistical Procedures*. Lederle Laboratories, American Cyanamid Co., 1964.

APPENDICES

A

ELEMENTARY THEORY OF ERRORS

A1 Introduction

The theory of errors was developed by Laplace, Gauss and others in the beginning of the 19th century. The Gaussian or the so-called normal distribution plays a very important rôle in the theory. The method of maximum likelihood is used in developing the concept of the most probable value of any physical quantity, obtained from repeated measurements.

Let us consider repeated measurements of a physical quantity by means of an experimental process which is as uniform as possible. It is a matter of common experience that the measurements are not all identical, but fluctuate around the true (unknown) value of the quantity. The difference of each observed value from the unknown true value is called the *experimental error* or *error of observation*. These errors are caused by a large number of uncontrollable (known or unknown) factors. Since the errors of observations are uncontrollable or random in nature, they are also called *random* or *accidental errors*.

Mathematically, if x is the measurement of a physical quantity whose true value is μ , then

$$e = x - \mu \quad \dots \quad (A1)$$

is called the error of measurement.

If we take n measurements, denoted by x_1, x_2, \dots, x_n , the errors of measurement are

$$e_i = x_i - \mu, \quad i = 1, 2, \dots, n. \quad \dots \quad (A2)$$

A2 Normal law of errors

That the distribution of errors follows a normal law can be deduced from Gauss's postulate of arithmetic mean, which states : "When any number of equally good direct measurements of an unknown magnitude μ is given, then the best estimate or the most probable value of μ is their arithmetic mean." The postulate of

arithmetic mean can, in its turn, be deduced from the following four elementary axioms

Axiom 1 The best estimate of μ , say $\psi(x_1, x_2, \dots, x_n)$, is a simple function, possessing a single-valued continuous derivative everywhere

Axiom 2 $\psi(x_1, x_2, \dots, x_n)$ is a symmetric function of x_1, x_2, \dots, x_n

Axiom 3 $\psi(x_1, x_2, \dots, x_n)$ is independent of the origin of measurement, i.e.,

$$\psi(x_1 + h, x_2 + h, \dots, x_n + h) = \psi(x_1, x_2, \dots, x_n) + h, \text{ for any } h,$$

Axiom 4 $\psi(x_1, x_2, \dots, x_n)$ is independent of the units of measurement, i.e.,

$$\psi(kx_1, kx_2, \dots, kx_n) = k\psi(x_1, x_2, \dots, x_n)$$

From *Axioms 1* and *4*, we have

$$\begin{aligned} k\psi(x_1, x_2, \dots, x_n) &= \psi(kx_1, kx_2, \dots, kx_n) \\ &= \psi(0, 0, \dots, 0) + k \sum_i x_i \left[\frac{\partial \psi}{\partial x_i} \right]_{(0, 0, \dots, 0)} (kx_1 - 0, kx_2 - 0, \dots, kx_n - 0), \\ &\quad 0 < k < 1 \end{aligned}$$

Making $k \rightarrow 0$, we get $\psi(0, 0, \dots, 0) = 0$, and hence dividing by k and making $k \rightarrow 0$ again, we have

$$\psi(x_1, x_2, \dots, x_n) = \sum_i v_i \left[\frac{\partial \psi}{\partial x_i} \right]_{(0, 0, \dots, 0)}$$

By *Axiom 2*, the constants $\left[\frac{\partial \psi}{\partial x_i} \right]_{(0, 0, \dots, 0)}$ must all be the same, say all equal to c . Then

$$\psi(x_1, x_2, \dots, x_n) = c \sum_i x_i$$

By *Axiom 3*

$$c \sum_i (x_i + h) = c \sum_i x_i + h$$

or

$$c = \frac{1}{n}$$

Hence $\psi(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_i x_i$,

the arithmetic mean

Now, if $\phi(e)$ be the probability density of the error $e = x - \mu$, then the joint probability density of the n independent observations

x_1, x_2, \dots, x_n is

$$f(x_1, x_2, \dots, x_n) = \phi(e_1)\phi(e_2)\dots\phi(e_n).$$

According to Gauss's postulate, this f attains its maximum when

$$\mu = \frac{x_1 + x_2 + \dots + x_n}{n},$$

i.e. when $\sum_i e_i = 0$.

We then have, differentiating $\log f$ with respect to μ ,

$$\sum_i \frac{\partial}{\partial e_i} \log \phi(e_i) \cdot \frac{\partial e_i}{\partial \mu} = 0$$

or $\sum_i \frac{\phi'(e_i)}{\phi(e_i)} = 0 \quad [\text{since } \frac{\partial e_i}{\partial \mu} = -1 \text{ for each } i]$

or $\sum_i F(e_i) = 0,$

where $F(e_i) = \frac{\phi'(e_i)}{\phi(e_i)}.$

Since, according to the postulate, this leads to the solution $\mu = \bar{x}$ or $\sum_i e_i = 0$, we have

$$\sum_i \{F'(e_i) + \lambda\} de_i = 0,$$

where λ is a Lagrangian multiplier.

It follows that

$$F'(e) + \lambda = 0$$

or $F(e) + \lambda e + c = 0,$

where c is a constant. Since $\sum_i F(e_i) = 0$ with $\sum_i e_i = 0$, we shall have

$$c = 0.$$

Thus

$$\frac{\phi'(e)}{\phi(e)} = -\lambda e$$

or $\log \phi(e) = -\frac{\lambda}{2} e^2 + \log A,$

where $\log A$ is the constant of integration.

Thus we have

$$\phi(e) = A \exp\left[-\frac{\lambda}{2} e^2\right] \quad (\text{A3})$$

λ must be positive, for otherwise the integral $\int_{-\infty}^{\infty} \phi(e) de$ will not converge. By putting

$$\int_{-\infty}^{\infty} \phi(e) de = 1$$

or $A \int_{-\infty}^{\infty} \exp\left[-\frac{\lambda}{2} e^2\right] de = 1,$

we have $A = \sqrt{\frac{\lambda}{2\pi}} = \frac{h}{\sqrt{\pi}},$

where $h^2 = \lambda/2$

Thus we have, finally,

$$\phi(e) = \frac{h}{\sqrt{\pi}} \exp[-h^2 e^2] \quad (\text{A4})$$

Hence the distribution of the error (e) is normal with mean 0 and standard deviation $1/\sqrt{2}h$. This h is sometimes called the *index* (or *modulus*) of precision, since the larger the value of h , the smaller is the value of the variance and the more precise will be the set of measurements.

A3 Most probable value

It can now be seen that if x_1, x_2, \dots, x_n be a set of equally precise measurements on a physical quantity whose true value is μ and if the index of precision be h , then the maximum-likelihood estimate of μ is \bar{x} . Here the likelihood function L is given by

$$L(\mu) = \frac{h^n}{(\sqrt{\pi})^n} \exp\left[-h^2 \sum_{i=1}^n (x_i - \mu)^2\right]$$

Hence $\log L = -h^2 \sum_i (x_i - \mu)^2 + \text{a constant independent of } \mu$, so that the likelihood equation $\frac{\partial \log L}{\partial \mu} = 0$ gives

$$\mu = \bar{x} \quad (\text{A5})$$

Thus \bar{x} is the maximum likelihood estimate of μ . It has been called the *most probable value* of μ in the theory of errors.

If, however, the measurements x_1, x_2, \dots, x_n are not equally precise and if their indices of precision be h_1, h_2, \dots, h_n , respectively, then the likelihood function is

$$L(\mu) = \frac{h_1 h_2 \dots h_n}{(\sqrt{\pi})^n} \exp \left[-\sum_i h_i^2 (x_i - \mu)^2 \right].$$

Here the maximum-likelihood estimate of μ is

$$\hat{\mu} = \frac{\sum_i h_i^2 x_i}{\sum_i h_i^2}. \quad \dots \quad (\text{A6})$$

Hence the most probable value of μ is now a weighted average of the measurements, the weights being proportional to the squares of the indices of precision.

On the other hand, suppose we take a linear function of observations, say

$$z = a_1 x_1 + a_2 x_2 + \dots + a_n x_n.$$

Since x_1, x_2, \dots, x_n are independently normally distributed with common mean μ and variances $\frac{1}{2h_1^2}, \frac{1}{2h_2^2}, \dots, \frac{1}{2h_n^2}$, z also will be normally distributed with mean $\zeta = \mu \sum_i a_i$ and variance $\frac{a_1^2}{2h_1^2} + \frac{a_2^2}{2h_2^2} + \dots + \frac{a_n^2}{2h_n^2}$ (*vide Exercise 14.11*). If H is the index of precision of z , we have

$$\frac{1}{H^2} = \frac{a_1^2}{h_1^2} + \frac{a_2^2}{h_2^2} + \dots + \frac{a_n^2}{h_n^2}. \quad \dots \quad (\text{A7})$$

Thus z has the probability density

$$f(z) = \frac{H}{\sqrt{\pi}} \exp[-H^2(z - \zeta)^2]. \quad \dots \quad (\text{A8})$$

Note that $\hat{\mu}$ of (A6) is also a linear function, with

$$a_i = h_i^2 / \sum_i h_i^2, \quad \sum_i a_i = 1.$$

Hence the most probable value $\hat{\mu}$ is itself normally distributed with mean μ and index of precision H , given by

$$\frac{1}{H^2} = \frac{1}{\sum_i h_i^2}.$$

A4 Measures of error

To give a general idea of the magnitude of errors in any given case, it is customary to employ one of the following measures

Error function

The probability that an error lies between $-r$ and $+r$ is given by

$$\begin{aligned} P[-x \leq e \leq +x] &= \int_{-x}^x \frac{h}{\sqrt{\pi}} \exp[-h^2 e^2] de \\ &= \frac{2h}{\sqrt{\pi}} \int_0^x \exp[-h^2 e^2] de \\ &= \frac{2}{\sqrt{\pi}} \int_0^{hx} \exp[-y^2] dy \\ &= \psi(hx) \end{aligned}$$

where

$$\psi(r) = \frac{2}{\sqrt{\pi}} \int_0^r \exp[-y^2] dy \quad (\text{A9})$$

The function $\psi(r)$ is called the error function of r and is sometimes denoted by $\text{erf}(x)$

Root mean square error

The root mean square error is the standard deviation of the error distribution. It is denoted by σ . Clearly

$$\sigma = \frac{1}{h\sqrt{2}} \quad (\text{A10})$$

Probable error (PE)

This is defined as a number γ such that the probability that the error lies between $-\gamma$ and $+\gamma$ is equal to $\frac{1}{4}$. From the table of normal distribution (i.e. Table I of Appendix B) it is seen that the probable error (PE) is given by

$$PE = \gamma = 0.6745\sigma, \quad (\text{A11})$$

Average error (η)

The mean absolute deviation of the error distribution is defined to be the *average error (η)*. Thus

$$\begin{aligned}\eta &= \frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} |e| \exp[-h^2 e^2] de \\ &= \frac{2h}{\sqrt{\pi}} \int_0^{\infty} |e| \exp[-h^2 e^2] de \\ &= \frac{1}{h\sqrt{\pi}} \int_0^{\infty} \exp[-y] dy, \quad \text{putting } y = h^2 e^2 \\ &= \frac{1}{h\sqrt{\pi}} = \sqrt{\frac{2}{\pi}} \cdot \sigma = 0.7979\sigma.\end{aligned}\dots \quad (\text{A12})$$

SUGGESTED READING

- [1] Scarborough, J. B. *Numerical Mathematical Analysis* (Chs. 16, 17). Johns Hopkins, 1962, and Oxford Book Co., 1964.
- [2] Whittaker, E and Robinson, G. *The Calculus of Observations* (Chs. 8, 9). Blackie, 1944.

B

STATISTICAL TABLES-

N.B. —For an explanation of the terms and symbols used in the tables the reader is referred to the following sections of the text

- 1 Section 9 15 (for Table I)
- 2 Section 14 6 (for Tables II V)
- 3 Section 18 4 (for Tables VI VIII)
- 4 Section 18 6 (for Table IX)
- 5 Section 18 8 (for Table X)

TABLE I. ORDINATES AND AREAS OF THE DISTRIBUTION OF NORMAL DEVIATE*

τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$
.00	.3989423	.5000000	.51	.3502919	.6949743	1.01	.2395511	.8437524
.01	.3989223	.5039894	.52	.3484925	.6984682	1.02	.2371320	.8461358
.02	.3988625	.5079783	.53	.3466677	.7019440	1.03	.2347138	.8484950
.03	.3987628	.5119665	.54	.3448180	.7054015	1.04	.2322970	.8508300
.04	.3986233	.5159534	.55	.3429439	.7088403	1.05	.2298821	.8531409
.05	.3984439	.5199388	.56	.3410458	.7122603	1.06	.2274696	.8554277
.06	.3982248	.5239222	.57	.3391243	.7156612	1.07	.2250599	.8576903
.07	.3979661	.5279032	.58	.3371799	.7190427	1.08	.2226535	.8599289
.08	.3976677	.5318814	.59	.3352132	.7224047	1.09	.2202508	.8621434
.09	.3973298	.5358564	.60	.3332246	.7257469	1.10	.2178522	.8643339
.10	.3969525	.5398278	.61	.3312147	.7290691	1.11	.2154582	.8665005
.11	.3965360	.5437953	.62	.3291840	.7323711	1.12	.2130691	.8686431
.12	.3960802	.5477584	.63	.3271330	.7356527	1.13	.2106856	.8707619
.13	.3955854	.5517168	.64	.3250623	.7389137	1.14	.2083078	.8728568
.14	.3950517	.5556700	.65	.3239724	.7421539	1.15	.2059363	.8749281
.15	.3944793	.5596177	.66	.3208638	.7453731	1.16	.2035714	.8769756
.16	.3938684	.5635595	.67	.3187371	.7485711	1.17	.2012135	.8789995
.17	.3932190	.5674949	.68	.3165929	.7517478	1.18	.1988631	.8809999
.18	.3925315	.5714237	.69	.3144317	.7549029	1.19	.1965205	.8829768
.19	.3918060	.5753454	.70	.3122539	.7580363	1.20	.1941861	.8849303
.20	.3910427	.5792597	.71	.3100603	.7611479	1.21	.1918602	.8868606
.21	.3902419	.5831662	.72	.3078513	.7642375	1.22	.1895432	.8887676
.22	.3894038	.5870644	.73	.3056274	.7673049	1.23	.1872354	.8906514
.23	.3885286	.5909541	.74	.3033893	.7703500	1.24	.1849373	.8925123
.24	.3876166	.5948349	.75	.3011374	.7733726	1.25	.1826491	.8943502
.25	.3866681	.5987063	.76	.2988724	.7763727	1.26	.1803712	.8961653
.26	.3856834	.6025681	.77	.2965948	.7793501	1.27	.1781038	.8979577
.27	.3846627	.6064199	.78	.2943050	.7823046	1.28	.1758474	.8997274
.28	.3836063	.6102612	.79	.2920038	.7852361	1.29	.1736022	.9014747
.29	.3825146	.6140919	.80	.2896916	.7881446	1.30	.1713686	.9031995
.30	.3813878	.6179114	.81	.2873689	.7910299	1.31	.1691468	.9049021
.31	.3802264	.6217195	.82	.2850364	.7938919	1.32	.1669370	.9065825
.32	.3790305	.6255158	.83	.2826945	.7967306	1.33	.1647397	.9082409
.33	.3778007	.6293000	.84	.2803438	.7995458	1.34	.1625551	.9098773
.34	.3765372	.6330717	.85	.2779849	.8023375	1.35	.1603833	.9114920
.35	.3752403	.6368307	.86	.2756182	.8051055	1.36	.1582248	.9130850
.36	.3739106	.6405764	.87	.2732444	.8078498	1.37	.1560797	.9146565
.37	.3725483	.6443088	.88	.2708640	.8105703	1.38	.1539483	.9162067
.38	.3711539	.6480273	.89	.2684774	.8132671	1.39	.1518308	.9177356
.39	.3697277	.6517317	.90	.2660852	.8159399	1.40	.1497275	.9192433
.40	.3682701	.6554217	.91	.2636880	.8185887	1.41	.1476385	.9207302
.41	.3667817	.6590970	.92	.2612863	.8212136	1.42	.1455641	.9221962
.42	.3652627	.6627573	.93	.2588805	.8238145	1.43	.1435046	.9236415
.43	.3637136	.6664022	.94	.2564713	.8263912	1.44	.1414600	.9250663
.44	.3621349	.6700314	.95	.2540591	.8289439	1.45	.1394306	.9264707
.45	.3605270	.6736448	.96	.2516443	.8314724	1.46	.1374165	.9278550
.46	.3588903	.6772419	.97	.2492277	.8339768	1.47	.1354181	.9292191
.47	.3572253	.6808225	.98	.2468095	.8364569	1.48	.1334353	.9305634
.48	.3555325	.6843863	.99	.2443904	.8389129	1.49	.1314684	.9318879
.49	.3538124	.6879331	1.00	.2419707	.8413447	1.50	.1295176	.9331928

TABLE I (Contd.)

τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$
151	1275830	9344783	201	0529192	9777844	251	0170947	993963
152	1256646	9157445	202	0518636	9783083	252	0166701	994137
153	1237628	9369916	203	0508239	9788217	253	0162545	994296
154	1218775	9382198	204	0498001	9793248	254	0158476	994451
155	1200090	9394292	205	0487920	9798178	255	0154493	994611
156	1181573	9406201	206	0477996	9803007	256	0150596	994764
157	1163225	9417924	207	0468226	9807738	257	0146782	994911
158	1145048	9429466	208	0458611	9812372	258	0143051	99506
159	1127042	9440826	209	0449148	9816911	259	0139401	99520
160	1109208	9452007	210	0439836	9821356	260	0135830	99533
161	1091548	9463011	211	0430674	9825708	261	0132337	99547
162	1074061	9473839	212	0421661	9829970	262	0128921	99560
163	1056748	9484491	213	0412795	9834142	263	0125581	99573
164	1039611	9494974	214	0404076	9838226	264	0122315	99581
165	1022649	9505285	215	0395500	9842224	265	0119122	99597
166	1005864	9515428	216	0387069	9846137	266	0116001	99601
167	0989255	9525403	217	0378779	9849966	267	0112951	99621
168	0972823	9535213	218	0370629	9853713	268	0109969	9963
169	0956568	9544860	219	0362619	9857379	269	0107056	9964
170	0940491	9554345	220	0354746	9860966	270	0104209	9965
171	0924591	9563671	221	0347009	9864474	271	0101428	9966
172	0908870	9572818	222	0339408	9867906	272	0098712	9967
173	0893326	9581849	223	0331939	9871263	273	0096058	9968
174	0877961	9590705	224	0324603	9874545	274	0093466	9969
175	0862773	9599408	225	0317397	9877755	275	0090936	9970
176	0847764	9607961	226	0310319	9880894	276	0088465	9971
177	0832932	9616364	227	0303370	9883962	277	0086052	9971
178	0818278	9624620	228	0296546	9886962	278	0083697	9972
179	0803801	9632730	229	0289847	9889893	279	0081398	9973
180	0789502	9640697	230	0283270	9892759	280	0079155	9974
181	0775379	9648521	231	0276816	9895559	281	0076965	9975
182	0761433	9656205	232	0270481	9898296	282	0074829	9975
183	0747663	9663750	233	0264265	9900969	283	0072744	9976
184	0734068	9671159	234	0258166	9903581	284	0070711	9977
185	0720649	9678432	235	0252182	9906133	285	0068728	9978
186	0707404	9685572	236	0246313	9908625	286	0066793	9978
187	0694333	9692581	237	0240556	9911060	287	0064907	9979
188	0681436	9699460	238	0234910	9913437	288	0063067	9980
189	0668711	9706210	239	0229374	9915758	289	0061274	9980
190	0656158	9712834	240	0223945	9918025	290	0059525	998
191	0643777	9719334	241	0218624	9920237	291	0057821	998
192	0631566	9725711	242	0213407	9922397	292	0056160	998
193	0619524	9731966	243	0208294	9924506	293	0054541	998
194	0607652	9738102	244	0203284	9926564	294	0052963	998
195	0595947	9744119	245	0198374	9928572	295	0051426	998
196	0584409	9750021	246	0193563	9930531	296	0049929	998
197	0573038	9755808	247	0188850	9932443	297	0048470	998
198	0561831	9761482	248	0184233	9934309	298	0047050	998
199	0550789	9767045	249	0179711	9936128	299	0045666	998
200	0539910	9772499	250	0175283	9937903	300	0044318	998

TABLE I (Contd.)

τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$
3.01	.0043007	.9986938	3.21	.0023089	.9993363	3.41	.0011910	.9996752
3.02	.0041729	.9987361	3.22	.0022358	.9993590	3.42	.0011510	.9996869
3.03	.0040486	.9987772	3.23	.0021649	.9993810	3.43	.0011122	.9996982
3.04	.0039276	.9988171	3.24	.0020960	.9994024	3.44	.0010747	.9997091
3.05	.0038098	.9988558	3.25	.0020290	.9994230	3.45	.0010383	.9997197
3.06	.0036951	.9988933	3.26	.0019641	.9994429	3.46	.0010030	.9997299
3.07	.0035836	.9989297	3.27	.0019010	.9994623	3.47	.0009689	.9997398
3.08	.0034751	.9989650	3.28	.0018397	.9994810	3.48	.0009358	.9997493
3.09	.0033695	.9989992	3.29	.0017803	.9994991	3.49	.0009037	.9997585
3.10	.0032668	.9990324	3.30	.0017226	.9995166	3.50	.0008727	.9997674
3.11	.0031669	.9990646	3.31	.0016666	.9995335	3.51	.0008426	.9997759
3.12	.0030698	.9990957	3.32	.0016122	.9995499	3.52	.0008135	.9997842
3.13	.0029754	.9991260	3.33	.0015595	.9995658	3.53	.0007853	.9997922
3.14	.0028835	.9991553	3.34	.0015084	.9995811	3.54	.0007581	.9997999
3.15	.0027943	.9991836	3.35	.0014587	.9995959	3.55	.0007317	.9998074
3.16	.0027075	.9992112	3.36	.0014106	.9996103	3.56	.0007001	.9998146
3.17	.0026231	.9992378	3.37	.0013639	.9996242	3.57	.0006814	.9998215
3.18	.0025412	.9992636	3.38	.0013187	.9996376	3.58	.0006575	.9998282
3.19	.0024615	.9992886	3.39	.0012748	.9996505	3.59	.0006343	.9998347
3.20	.0023841	.9993129	3.40	.0012322	.9996631	3.60	.0006119	.9998409

*Abridged from Table 1 of *Biometrika Tables for Statisticians*, Vol. I, with the kind permission of the Biometrika Trustees.

TABLE II. DISTRIBUTION OF NORMAL DEVIATE
Values of τ_α

α	0.05	0.025	0.01	0.005
τ_α	1.645	1.960	2.326	2.576

TABLE III χ^2 DISTRIBUTION*

Values of χ^2_v

α	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.688	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.706	22.164	24.433	26.509	55.759	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

For larger values of v , the variable $\sqrt{2X^2} - \sqrt{2v-1}$ may be used as a normal deviate.

*Abridged from Table 8 of *Biometrika Tables for Statisticians*, Vol. I, with the kind permission of the Biometrika Trustees.

TABLE IV. *t* DISTRIBUTION**Values of t_{α, v}*

$\frac{\alpha}{v}$	0.05	0.025	0.01	0.005
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
∞	1.645	1.960	2.326	2.576

*Abridged from Table 12 of *Biometrika Tables for Statisticians*, Vol. I, with the kind permission of the Biometrika Trustees.

TABLE V *F* DISTRIBUTION^a

Values of $F_{0.05} \cdot v_1 \cdot v_2$

v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.8	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3	
2	18.51	19.00	19.16	19.25	19.30	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.47	19.48	19.49	19.50		
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.36	
6	5.59	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.67	
7	5.32	4.46	4.07	3.84	3.69	3.58	3.57	3.59	3.53	3.50	3.44	3.38	3.34	3.31	3.30	3.27	3.23	3.23	
8	5.12	4.26	3.86	3.63	3.48	3.37	3.32	3.30	3.29	3.27	3.23	3.20	3.18	3.15	3.12	3.08	3.04	3.03	
9	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.80	2.75	2.70	2.65	2.62	2.62	
10	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	
11	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.55	2.51	2.47	2.43	2.38	2.34	
12	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.30	2.25	2.20	
13	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	
14	4.54	3.68	3.32	3.03	2.85	2.74	2.61	2.59	2.54	2.48	2.40	2.35	2.28	2.23	2.19	2.15	2.11	2.07	
15	4.49	3.63	3.24	3.01	2.85	2.74	2.61	2.55	2.49	2.42	2.35	2.28	2.20	2.15	2.11	2.06	2.01	1.96	
16	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.42	2.35	2.28	2.20	2.15	2.11	2.06	2.01	1.96	
17	4.41	3.55	3.16	2.93	2.77	2.63	2.54	2.45	2.39	2.32	2.25	2.18	2.11	2.04	1.99	1.95	1.90	1.84	
18	4.38	3.52	3.13	2.90	2.74	2.62	2.51	2.42	2.32	2.22	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78	
19	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.42	2.32	2.22	2.15	2.07	2.01	1.99	1.95	1.90	1.84	1.78	
20	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.39	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78	
21	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.32	2.22	2.12	2.04	2.01	1.92	1.84	1.79	1.74	1.68	1.62	
22	4.23	3.37	2.98	2.74	2.59	2.47	2.36	2.29	2.22	2.12	2.04	2.01	1.92	1.84	1.79	1.74	1.68	1.62	
23	4.20	3.34	2.95	2.71	2.56	2.45	2.33	2.27	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.68	1.62	
24	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.25	2.18	2.12	2.04	1.96	1.89	1.81	1.75	1.70	1.65	1.59	
25	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.17	2.10	2.04	1.96	1.88	1.81	1.75	1.66	1.61	1.55	1.49	
26	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.09	2.02	1.94	1.88	1.81	1.75	1.67	1.61	1.55	1.49	1.39	
27	3.94	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.94	1.88	1.81	1.75	1.67	1.61	1.55	1.49	1.43	1.32	
28	3.84	3.00	2.60	2.37	2.21	2.01	1.94	1.88	1.81	1.75	1.67	1.61	1.55	1.49	1.43	1.39	1.32	1.30	

For other values of v_1 and v_2 , one may use linear interpolation taking $1/v_1$ and $1/v_2$ as the independent variables.

TABLE V (Contd.)

Values of $F_{0.01; v_1, v_2}$

$\frac{v_1}{v_2}$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	.40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	
2	9850	99.00	99.17	99.25	99.30	99.33	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.48	99.49	99.49	99.50	
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.50	26.41	26.32	26.22	26.13	
4	11.32	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	
5	9.75	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.75	7.62	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.86	
6	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.91	5.82	5.74	5.65	
8	11.22	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.56	3.51	3.43	3.35	3.27	3.18	
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.84	2.75	2.66	2.57	
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.01	
31	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.80	
40	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	
60	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	
120	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	

For other values of v_1 and v_2 , one may use linear interpolation, taking $1/v_1$ and $1/v_2$ as the independent variables.

* Abridged from Table 18 of Biometrika Tables for Statisticians, Vol. I, with the kind permission of the Biometrika Trustees.

TABLE VI CUMULATIVE BINOMIAL PROBABILITIES OF r OR FEWER SUCCESSES IN n INDEPENDENT TRIALS WITH $p=0.5$

$\backslash r$	0	1	2	3	4	5	6	7	8	9	10	11	12
4	062	312	688										
5	031	188	500										
6	016	109	344	656									
7	008	062	227	500									
8	004	035	145	363	637								
9	002	020	090	254	500								
10	001	011	055	172	377	623							
11		006	033	113	274	500							
12		003	019	073	194	387	613						
13		002	011	046	133	291	500						
14		001	006	029	090	212	395	605					
15			004	018	059	151	304	500					
16			002	011	038	105	227	402	598				
17			001	006	025	072	166	315	500				
18			001	004	015	048	119	240	407	593			
19			002	010	032	084	180	324	500				
20			001	006	021	058	132	252	412	588			
21			001	004	013	039	095	192	332	500			
22				002	008	026	067	143	262	416	584		
23				001	005	017	047	105	202	339	500		
24				001	003	011	032	076	154	271	419		
25					002	007	022	054	115	212	345	500	

When $r > n/2$, use the fact that cumulative probability for r equals $1 -$ cumulative probability for $(n-r-1)$

TABLE VII. CRITICAL VALUES OF T IN THE WILCOXON SIGNED-RANK TEST*

n	Significance level of test		
	One-tailed :	0.025	0.01
	Two-tailed :	0.05	0.02
6		1	
7		2	0
8		4	2
9		6	3
10		8	5
11		11	7
12		14	10
13		17	13
14		21	16
15		25	20
16		30	24
17		35	28
18		40	33
19		46	38
20		52	43
21		59	49
22		66	56
23		73	62
24		81	69
25		90	77
			68

*Adapted from Table 2 of Wilcoxon, F. and Wilcox, R. A. (1964) : *Some Rapid Approximate Statistical Procedures*, Lederle Laboratories, American Cyanamid Co., Pearl River, New York, with the kind permission of the authors and the publishers.

TABLE VIII. CRITICAL VALUES OF U FOR THE
MANN-WHITNEY TEST*

(a) Significance level 0.01 for a one-tailed test and 0.02 for a two-tailed test :

$n_1 \backslash n_2$	9	10	11	12	13	14	15	16	17	18	19	20
1												
2					0	0	0	0	0	0	1	1
3	1	1	1	2	2	2	3	3	4	4	4	5
4	3	3	4	5	5	6	7	7	8	9	9	10
5	5	6	7	8	9	10	11	12	13	14	15	16
6	7	8	9	11	12	13	15	16	18	19	20	22
7	9	11	12	14	16	17	19	21	23	24	26	28
8	11	13	15	17	20	22	24	26	28	30	32	34
9	14	16	18	21	23	26	28	31	33	36	38	40
10	16	19	22	24	27	30	33	36	38	41	44	47
11	18	22	25	28	31	34	37	41	44	47	50	53
12	21	24	28	31	35	38	42	46	49	53	56	60
13	23	27	31	35	39	43	47	51	55	59	63	67
14	26	30	34	38	43	47	51	56	60	65	69	73
15	28	33	37	42	47	51	56	61	66	70	75	80
16	31	36	41	46	51	56	61	66	71	76	82	87
17	33	38	44	49	55	60	66	71	77	82	88	93
18	36	41	47	53	59	65	70	76	82	88	94	100
19	38	44	50	56	63	69	75	82	88	94	101	107
20	40	47	53	60	67	73	80	87	93	100	107	114

TABLE VIII (Contd.)

(b) Significance level 0.05 for a one-tailed test and 0.10 for a two-tailed test :

$n_1 \backslash n_2$	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7		9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

*Adapted and abridged from Auble, D. (1953) : "Extended tables for the Mann-Whitney statistic", *Bulletin of the Institute of Educational Research at Indiana University*, I, No. 2, with the kind permission of the author and the publishers.

TABLE IX CRITICAL VALUES OF r IN THE RUN TEST*
Significance level 0.05

$n_1 \backslash n_2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2										2	2	2	2	2	2	2	2	2	2
3			2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3
4			2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12		2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10
13		2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10
14		2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11
15		2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	12
16		2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12
17		2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12
18		2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13
19		2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13
20		2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	14

*Adapted from Swed, F S and Eisenhart, C (1943). "Tables for testing randomness of grouping in a sequence of alternatives", *Annals of Mathematical Statistics*, 14, pp 83-86, with the kind permission of the author and the editor.

TABLE X. CRITICAL VALUES OF r_R , THE SPEARMAN RANK CORRELATION COEFFICIENT*

n	Significance level (one-tailed test)	
	0.05	0.01
4	1.000	
5	.900	1.000
6	.829	.943
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746
12	.506	.712
14	.456	.645
16	.425	.601
18	.399	.564
20	.377	.534
22	.359	.508
24	.343	.485
26	.329	.465
28	.317	.448
30	.306	.432

*Adapted from Olds, E. G. (1938) : "Distributions of sums of squares of rank differences for small numbers of individuals", *Annals of Mathematical Statistics*, 9, pp. 133-148, and from Olds, E. G. (1949) : "The 5% significance levels for sums of squares of rank differences and a correction", *Annals of Mathematical Statistics*, 20, pp. 117-118, with the kind permission of the author and the editor.

INDEX

- Alternative hypotheses, 413-414
Array distribution (*see* conditional distribution)
 Array mean, 318, 324-326
 Array variance, 319
Association, 277-278, 318
 — absolute, 278, 280
 — and causal relationship, 289-290
 — coefficient of, 279
 — complete, 278, 279
 — joint, 287
 — measures of, 279-281, 283-285
 — 287-289
 — multiple, 287
 — negative, 278
 — partial, 288
 — perfect, 278, 284
 — positive, 278
 — total, 288
Asymmetrical distribution, 225
Asymmetry (*see* skewness)
Attribute, definition, 162, 275
Attributes, independent, 277, 282, 286
Average, 179, 189
Average error, 537

 b_1, b_2 coefficients, 229
Band chart, 148-149
Bar diagram, 149-153
Bernoulli, J., 100, 235
Bernoulli's theorem, 130
Beta function, 30-31
Binomial distribution, 235-242, 259-260
 — cumulative probability of, 269
 — fitting of, 239-242
 — moments of, 236-238
 — recursion relation concerning moments of, 238-239
Binomial series, 8-9
Bivariate data, 297-299
Bivariate frequency distribution, 299
Bivariate interpolation, 70-72
Bivariate normal distribution, 318-320
Box, G. E. P., 507
Cauchy-Schwarz inequality, 12
Central limit theorem, 471
Central moments, definition, 215
Central tendency, definition, 178-179
Charlier checks, 218-219
Chebyshev's inequality, 128
Chebyshev's lemma, 127-128
Chi-square distribution, 388-391
Class boundaries, 170-171
Class interval, 170
Class limits 170
Cochran, W. G., 446, 489
Coefficient of colligation, 279
Coefficient of contingency, Karl Pearson's, 284
 — Tschuprow's, 284
Coefficient of variation, 210
Co-factor, 16
Collection of data, 141-142
Column diagram, 172-173
Component-parts chart (*see* band chart)
Compound probability, 107-111
Concentration, area of, 211
 — coefficient of, 211
 — curve of, 210-212
Conditional distribution, 276, 299, 318-319
Conditional probability, 107-109
Confidence coefficient, 410
Confidence limits (interval), definition, 410
Consistency, 403
Convergence in probability (stochastic), 127-130, 403
Convergence of iteration method, 86-87
 — of Newton-Raphson method, 87-88
Correlation, 300-301
Correlation coefficient (simple), definition, 301-303
 — limitations of, 320-322
 — properties of, 303-305, 320
Correlation index, 322-323
Correlation ratio, 325-326
Counting numbers, 5
Covariance, 123, 302
Cox, G. M., 446

- Multiple correlation, 340 343
 — in terms of total and partial correlations, 348 349
- Multiple regression, 333-340, 342 343
- Multivariate data, 333
- Multivariate normal distribution, 353-355
- Negative binomial distribution, 252-253
- Neyman, J., 415
- Nonparametric methods 507
- Normal deviate definition, 257, 386-388
- Normal distribution, 254 263, 268
 — fitting of, 260 262
 — importance of, 262-263
 — properties of, 255-258
- Normal law of errors, 531-534
- Numerical differentiation, 72-73
- Numerical integration, definition, 74
 — Gregory's formula, 94
 — Euler Maclaurin formula, 80-81
 — relative accuracy of quadrature formulae, 77-79
 — Simpson's one third rule, 75-76
 — Simpson's three eighths rule, 94
 — trapezoidal rule, 74-75
 — Weddle's rule, 76-77
- Numerical solution of equations, definition, 82 83
 — Horner's method, 88-91
 — iteration method, 86 87
 — method of false position, 83
 — Newton Raphson method, 84, 87
- Ogive (see cumulative frequency diagram)
- Operators, 45-49, 92
 — A , E , 45-49
 — ∇ , δ , μ , 92
- Order statistic, 508 510
- Orthogonal transformation, 18, 396-397
- Paired t test, 456-458
- Parameter, definition, 236, 379
- Parameter free procedure, 508
- Partial correlation, 343 346
 — of higher order in terms of coefficients of lower order, 351
 — of lower order in terms of coefficients of higher order, 352
- Partial regression coefficient, 337, 346-347
- Pearson, E S., 415
- Pearson Karl, 265, 284, 488
- Pearsonian chi-square (see frequency chi-square)
 — test for goodness of fit, 489-492
 — test for homogeneity, 493-494
 — test for independence, 495-497
 — simplified formulae, 497-498
- Pearsonian curve, points of inflection of, 268
- Pearsonian curve, type I, 266
 — type II, 268
 — type III, 267
 — type IV, 267
 — type V, 267
 — type VI, 267
 — type VII, 268
- Pearsonian differential equation, 265-266
- Pearsonian system of curves, 265-268
- Pictorial diagram, 153
- Pie diagram, 155-156
- Platykurtic curve, 229
- Point estimation, theory of, 401-406
- Point of inflection, 24
- Poisson distribution, 243 249, 259-260
 — cumulative probability of, 269
 — fitting of, 248 249
 — moments of, 246 247
 — recursion relation concerning moments of 247 248
- Population, definition, 232
- Power function, 416
- Presentation of data, 143-157
- Probability, an axiomatic approach, 115 117
 — classical definition, 99-100, 112-115
- Probability density function, 233
- Probability in continuum (see geometrical probability)
- Probability mass function, 233
- Probability, meaning of, 98
- Probability sampling, 377
- Probable error, 380, 536
- Product notation, 4-5

- Quadratic form, 17-18
 - non-negative, 18
 - positive definite, 18
 - positive semi-definite, 18
 - rank of, 17
- Quantile, 208, 508
- Quantile test, 525
- Quartile, 208
- Quartile deviation, 208-209
- Random order, 101
- Random sampling, 104, 110, 232, 377-379
- Random sampling, simple : with replacements, 126, 378
 - simple : without replacements, 126, 378-379
- Random variable, 117-127
- Rank, 359
- Rank correlation coefficient, 359-367
 - Kendall's, 364-367
 - Spearman's, 359-364, 522-523
- Range, 199
- Ratio chart, 146-148
- Rational numbers, 6
- Real number system, 5-7
- Rectangular distribution, 253-254
- Regression, 310-317, 318
- Regression coefficient, 312
- Regression lines, 310-312, 318-319
 - properties of, 312-315
- Relative accuracy of quadrature formulae, 77-79
 - one-third rule, 78-79
 - three-eighths rule, 79
 - trapezoidal rule, 78
 - Weddle's rule, 79
- Relative dispersion, 209-210
- Relative frequency, 115, 163, 166
- Remainder terms in interpolation formulae, 69-70
 - Bessel's, 70
 - Lagrange's, 70
 - Newton's, 70
 - Stirling's, 70
- Residual variance, 313, 342
- Robust procedure, 507
- Root-mean-square deviation, 202
- Root-mean-square error, 536
- Rounding off, 36
- Run test for randomness, 521-522
- Sample, 232
- Sampling distribution, definition, 379-380
 - of sample mean from normal population, 396-397
 - of sample variance from normal population, 396-398
 - with binomial variables, 385
 - with Poisson variables, 385-386
- Sampling fluctuations, 188, 379
- Scatter diagram, 300-302
- Scedasticity, 319
- Scrutiny of data, 142-143
- Semi-interquartile range (*see* quartile deviation)
- Semi-logarithmic chart (*see* ratio chart)
- Separation of symbols, 49
- Sequence, definition, 7
 - bounded, 7
 - convergent, 7
 - unbounded, 7
- Series, definition, 7
 - absolutely convergent, 8
 - conditionally convergent, 8
 - convergent, 7-8
 - divergent, 7-8
 - oscillatory, 8
- Set, definition, 5
 - empty (or null), 5
- Sheppard's corrections, 221-223
- Sign test, 510-512
- Significant figures, 37
- Signed rank test, 512-513
- $\sin^{-1} \sqrt{p}$ transformation for proportion, 484
- Skewness, 224-227
- Smoking and lung cancer, 290-293
- Spearman, C., 359, 522
- Standard deviation, 202-205
- Standard error, definition, 380
 - for central moments, 480
 - for coefficient of variation, 480
 - or correlation coefficient, 481
 - for g_1, g_2 coefficients, 480
 - for mean, 381-383

- Standard error (*contd.*)
 — for median, 481
 — for proportion, 384
 — for quantiles, 481
 — for variance, 480
- Standard error of estimate, 314, 342
- Statistic, definition 379
- Statistics (plural), definition, 139
 — (singular) definition, 139
- Statistical map, 156-157
- Statistical quality control, 206
- Step diagram, 176
- Stirling's approximation, 34, 94
- Subset, 5
 — proper, 5
- Sufficiency, 405-406
- Sukhatme's test for dispersion, 518
- Sum notation, 34
- Symmetrical distribution, 224-225
- t* distribution, definition, 392-393
- Tabular representation of data 144-145
- Test of significance, theory of, 411-417
- Tests of significance, relating to binomial distribution, 427-430
 — bivariate normal distribution, 455-459
 — correlation (multiple), 464
 — correlation (partial), 464
 — correlation (simple), 455-456
 — independence of two attributes, 432-435
 — more than two univariate normal distributions, 450-455
 — one univariate normal distribution 470-472
 — regression, 451
 — pearman's rank correlation, 522-523
 — two univariate normal distributions, 442-450
- Theoretical distributions, 232-235
- Tie, 361
- Tolerance interval, 509-510
- Total correlation, 344
- Total probability, 102-107
- Unbiased estimate of mean, 381-383
 — of proportion, 383-384
 — of regression coefficient, 460
 — of variance, 409, 435
- Unbiasedness, 402, 417
- Variables, 161-162
 — continuous, 164
 — discrete, 164
 — independent, 122, 318
 — random (or stochastic), 117
- Variance, definition, 119-120, 125-126, 215
- Variates (*ie* variables)
- Vectors, 15
 — column, 15
 — linearly dependent, 15
 — linearly independent, 15
 — row, 15
- Wald Wolfowitz run test, 519-521
- Wallis' formula, 94
- Weak law of large numbers, 127-130
- Weighted mean, 191-192
- Wilcoxon, F, 512
- \sqrt{x} transformation for Poisson variable, 485
- Yates' correction for continuity, 493-499
- Yates, F, 498
- z transformation for correlation coefficient, 484-487

ERRATA

P. 84, line 9	<i>for</i> $ \theta < 1$	<i>read</i> $0 < \theta < 1$
P. 106, line 15	<i>for</i> values of m	<i>read</i> values of $m \geq 2$
P. 266, eqn. (9.80)	<i>for</i> y_0	<i>read</i> C
P. 365, eqn. (13.6)		<i>read</i> $\tau = \frac{\sum(a_{ij} b_{ij})}{\sqrt{\sum a_{ij}^2} \sqrt{\sum b_{ij}^2}}$
P. 374, <i>Exercise 13.3</i>	<i>for</i> r_I	<i>read</i> r_R
P. 375, <i>Exercise 13.8</i>	<i>Ans.</i> is $r_R = 0.342$, $\tau = 0.299$	
P. 422, line 9	<i>for</i> Hence...	<i>read</i> Hence $C = t_{\alpha/2, n-1}$ and.....
P. 453, line 3	<i>for</i> $\sum_i \sum_j x_{ij}^2 / n$	<i>read</i> $\sum_i \sum_j x_{ij}^2$
P. 492, col. (4) of table	<i>for</i> $(3)^2/(2)$	<i>read</i> $(2)^2/(3)$
512, line 9 from bottom	<i>for</i> .0547	<i>read</i> .055