

**Fundamentals
of
STATISTICS
*VOLUME TWO***

**A. M. Goon
M. K. Gupta
B. Dasgupta**

Fundamentals of STATISTICS

VOLUME TWO

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

B. DASGUPTA, M.Sc.

*Department of Statistics
Presidency College, Calcutta*



CALCUTTA
THE WORLD PRESS PRIVATE LTD.

**Atindra Mohan Goon, 1931—
Milan Kumar Gupta, 1932—
Bhagabat Dasgupta, 1933—**

**PUBLISHED BY S. BHATTACHARJEE FOR THE WORLD PRESS PRIVATE LTD., 37A,
COLLEGE STREET, CALCUTTA-700073 AND PRINTED IN INDIA BY OFFSET, BY N. K.
MITTER AT THE INDIAN PRESS PVT. LTD., 99A LENIN SARANI, CALCUTTA-700019**

*IN THE DEPARTMENTS OF STATISTICS OF
PRESIDENCY COLLEGE AND CALCUTTA UNIVERSITY*

PREFACE

The 5th edition of Volume One of *Fundamentals of Statistics* was brought out almost a year ago. Volume Two (4th edition) of the book has also been out of print for some months now and this has necessitated the publication of this newer edition.

We should point out once again that Volume Two does no more than supplement the material covered in Volume One. For while Volume One is concerned with the general concepts and the general principles and techniques of statistics that are applicable to a wide variety of situations, Volume Two presents, by and large, some special techniques and also methods meant for some special fields of application.

Volume Two comprises Parts Three and Four of the book. Of these, Part Three deals with the technique of analysis of variance, the design and analysis of experiments and the design and analysis of sample surveys. Part Four presents methods that are required to handle some of the major problems arising in the fields of demography, psychology and education, economics and manufacturing industries.

A whole chapter (Chapter 19) has been devoted to analysis of variance in view of its importance as a general tool for data analysis. The next two chapters deal with the question of properly planning statistical enquiries so as to reach valid and reliable conclusions. Of these, Chapter 20 has to do with situations where the investigator can effect a good degree of control over the experimental conditions, while situations where his task is simply to collect data as they occur in nature are the subject-matter of Chapter 21.

In the field of demography, some of the topics of prime importance are mortality, fertility, morbidity, population growth and population projection, which have been treated in Chapter 22. Scaling procedures and test theory are the topics from the field of psychology and education that have received attention in this book (Chapter 23). Among problems in economics that have been

treated here are index numbers, time-series analysis and demand analysis (Chapters 24-26). In the case of large-scale manufacturing industries, quality control is a problem that has assumed particular importance in modern times. This problem has been considered here in its two aspects: process control and product control (Chapter 27).

We have included, as appendices to the main body of the volume, important statistical tables and a discussion of the statistical system in India and also the sources, scope and limitations of Indian official statistics.

Particular care has been taken in the formulation of the examples and exercises. Mostly based on Indian data, they are expected to provide the reader a better understanding of the subject-matter.

The need for a new edition has enabled us to subject the volume to a rather thorough revision. We have again borne in mind not only the syllabi of the courses in statistics of different Indian universities but also the requirements of the research worker, for whom the volume (nay, the whole book) is expected to prove a reliable guide. In preparing this edition, we have added some new material to make the volume all the more useful. The reader's attention may be drawn particularly to the discussions on analysis of covariance for a general complete block design (Subsection 20.13.3), series of experiments (Section 20.17), further mortality rates (Subsections 22.3.4-22.3.8), morbidity rates (Section 22.7), preliminary adjustment of time-series data (Section 25.2), changing seasonal patterns (Section 25.6), 3σ control limits and probability limits (Section 27.4). At the same time, some of the items included in earlier editions have been discussed in greater detail, while errors that came to, or were brought to, our notice have been removed and minor changes made here and there to improve the exposition.

We should put on record the help and encouragement we have received from our teachers, friends and colleagues in preparing the successive editions of this volume. Dr. S. Chakraborty, Mr. A. K. De, Mr. C. R. Malakar and Dr. S. P. Mukherji deserve thanks for making available to us some valuable data that have been used in Chapters 22 and 24. We are thankful also to our colleague Mr. B.

Das and to Rahul Mukherji, a student of ours, for their suggestions that have considerably enhanced the value of the volume.

Our thanks are also due to The World Press for the utmost care with which they undertake the publication of our books.

THE AUTHORS

CONTENTS

CHAPTER		PAGES
<i>Part Three : ANALYSIS OF VARIANCE AND DESIGNS</i>		<i>1—178</i>
19 ANALYSIS OF VARIANCE		3—48
<p>Introduction. Linear model. A theorem of importance in Model I analysis. Tests of general linear hypotheses. Analysis of one-way classified data. Analysis of two-way classified data with one observation per cell. Analysis of two-way classified data with m observations per cell. Application of the technique of analysis of variance in the study of relationship. Effects of violations of the assumptions made in the analysis of variance.</p>		
20 DESIGNS OF EXPERIMENTS		49—128
<p>Terminology in experimental designs. Principles of design. Choice of size and shape of plots and blocks. Completely randomised design. Randomised block design. Latin square design. Graeco-Latin square. Cross-over design. Factorial experiments : a 2^n-experiment, a 2^3-experiment. A 2^n-experiment in 2^k blocks per replicate. Factorial experiments in a single replicate. Split-plot design. Analysis of covariance : analysis of covariance for a one-way layout with one concomitant variable, analysis of covariance for an RBD with one concomitant variable, analysis of covariance for any complete block design. Testing the homogeneity of a group of regression coefficients. Some facts about analysis of covariance. Missing-plot technique. Series of experiments.</p>		
21 DESIGNS OF SAMPLE SURVEYS		129—178
<p>Introduction. Basic principles of sample surveys. Advantages of sample survey over complete census. Different steps in a large-scale sample survey. Biases in surveys. Technique of random sampling. Types of population and types of sampling. Simple random sampling. Stratified random sampling. Multistage sampling. Systematic sampling. Multiphase sampling. Double sampling. Purposive sampling. Sampling with probability proportional to size. Quota sampling. Some mathematical methods for errors in measurement. National Sample Surveys.</p>		
<i>Part Four : METHODS FOR SOME SPECIAL FIELDS OF APPLICATION</i>		<i>179—395</i>
22 VITAL STATISTICS METHODS		181—244
<p>Introduction. Rates of vital events. Measurement of mortality : crude death rate, specific death rate, standardised death rate.</p>		

CHAPTER		PAGES
	comparative mortality index, cause-of-death rate, maternal mortality rate, infant mortality rate, case fatality rate. Life table : description, construction of a life table, abridged life table, King's method, Greville's method and method of Reed and Merrell, uses of a life table. Measurement of fertility : crude birth rate, general fertility rate, age-specific fertility rate, total fertility rate. Measurement of population growth : crude rate of natural increase and vital index, gross reproduction rate, net reproduction rate. Measurement of morbidity : morbidity incidence rate, morbidity prevalence rate. Graduation formulae used in vital statistics : graduation of population data, logistic curve, fitting a logistic curve, graduation of mortality rates. Makeham's graduation formula, fitting Makeham's formula. Population projection.	
23 STATISTICAL METHODS FOR PSYCHOLOGY AND EDUCATION		245—279
	Introduction. Some scaling procedures : scaling individual test-items in terms of difficulty, scaling of test-scores in several tests, scaling of rating or ranking in terms of the normal curve, scaling of qualitative answers to a questionnaire, scaling of judgements of a number of products : product scale. Test theory : the linear model of test theory, definition of parallel tests, definition of true score, error variance (standard error of measurement), definition of reliability, effect of test length on the reliability of the test, practical methods of estimating test reliability, validity, effect of test length on test parameters. Intelligence tests and <i>IQ</i> .	
24 INDEX NUMBERS		280—304
	Introduction. Problems in the construction of index numbers : purpose of the index, choice of the base period, choice of the commodities, collection of data, method of combining data, choice of weights, interpretation of the index. Errors in index numbers. Tests for index numbers. Chain index. Relative merits and demerits of chain-base and fixed-base methods. Cost of living index number. Cost of living index number and Laspeyres' and Paasche's formulæ. Two important index number series. Uses of index numbers.	
25 ANALYSIS OF TIME SERIES		305—341
	Introduction. Preliminary adjustments of time-series data. Components of time series. Measurement of secular trend. Measurement of seasonal fluctuations. Changing seasonal patterns. Measurement of cyclical fluctuations. Effect of moving averages on cyclical and random components of a time series. Different schemes which account for oscillations in a stationary time series. Serial correlation and correlogram. Correlation between two time series : lag correlation.	

CHAPTER	PAGES
26 DEMAND ANALYSIS	342—356
<p>Introduction. Demand and supply curves. Price-elasticity of demand and supply. Determination of demand curves from market data. Engel's law and the Engel curve. Income-elasticity of demand. Different forms of the Engel curve. Variation in household size and composition.</p>	
27 STATISTICAL QUALITY CONTROL	357—395
<p>Introduction. Different types of quality-measures. Rational sub-groups and the technique of control charts. 3σ control limits and probability limits. Control charts for mean, s.d. and range : control charts for mean, control charts for s.d., control charts for range. Control charts for number defective and fraction defective : control charts for number defective, control charts for fraction defective, control charts for per cent defective. Control charts for number of defects. Two types of control charts. Natural tolerance limits and specification limits Advantages of process control. Sampling inspection by attributes : single sampling plans, double sampling plans, multiple sampling plans, sequential sampling inspection plans, comparison of the three types of plans. Sampling inspection by variables : underlying principles, variables inspection with known s.d., variables inspection with unknown s.d.</p>	
Appendices	397—431
A INDIAN OFFICIAL STATISTICS	399—415
<p>Introduction. Population statistics. Agricultural statistics. Price statistics. Industrial statistics. Trade statistics. Labour statistics. Transport and communications statistics. Miscellaneous statistics.</p>	
B STATISTICAL TABLES	416—426
I Ordinates and areas of the distribution of normal deviate.	
II Distribution of normal deviate : Values of τ_a .	
III X^2 distribution : Values of $X^2_{a,v}$.	
IV t distribution : Values of $t_{a,v}$.	
V F distribution : Values of $F_{a;v_1, v_2}$.	
VI Random sampling numbers.	
VII Factors useful in the construction of control charts.	
INDEX	427—431

“Statistics is essentially an applied science. Its only justification lies in the help it can give in solving a problem.”

P. C. Mahalanobis

PART THREE

ANALYSIS OF VARIANCE AND DESIGNS

19

ANALYSIS OF VARIANCE

19.1 Introduction

The total variation present in a set of observable quantities may, under certain circumstances, be partitioned into a number of components, associated with the nature of classification of the data. The systematic procedure for achieving this is called the *analysis of variance*. With the help of the technique of analysis of variance, it will be possible for us to perform certain tests of hypotheses and to provide estimates for components of variation.

Consider random samples of students of Class IX from each of 3 secondary schools (selected at random out of all secondary schools in Calcutta). A certain intelligence test is applied to the selected students and their performances, as determined by the test scores, are noted. The total variation is measured by the sum of squares of deviations of scores from the mean score. In this case, there are two sources of variation present into which the total variation may be partitioned. First, the scores within a school differ and it is true for all the schools. Secondly, there may be an effect due to schools, i.e. the mean scores for the three different schools may vary. Hence, in the present example, the total variation is partitioned into two components : within schools and between schools. This analysis of variance will serve two other purposes—we can test the hypothesis that the mean scores of all students of Class IX are equal for all Calcutta secondary schools ; we can also estimate the two variance-components present here (*vide* Sections 19.5 to 19.7).

19.2 Linear model

Let y_1, y_2, \dots, y_n be n observable quantities. In all cases, we shall assume the observed value to be composed of two parts :

$$y_i = \mu_i + e_i, \quad \dots \quad (19.1)$$

where μ_i is the *true value* and e_i the *error*. The true value μ_i is that part which is due to assignable causes, and the portion that remains is the error, which is due to various chance causes. The true value

μ_i is again assumed to be a linear function of k unknown quantities, $\tau_1, \tau_2, \dots, \tau_k$, called *effects*:

$$\alpha_i = a_{i1}\tau_1 + a_{i2}\tau_2 + \dots + a_{ik}\tau_k, \quad \dots \quad (19.2)$$

where a_{ij} 's are known, each being usually taken to be 0 or 1.

This set-up, which is fundamental to analysis of variance, is called the *linear model*.

It is possible that there may be association between errors of successive measurements, but we shall assume that the errors ϵ_j 's are always independent random variables. These are also assumed to have expectation zero and to be homoscedastic. We shall call a model in which all the effects τ_j 's are unknown constants, which we call parameters, a *fixed-effects model* or Model I or *linear hypothesis model*. It is often the case that one of the τ_j 's is a constant with $a_{ij}=1$ for that j and all i . Such a τ_j is called a *general mean* or an *additive constant*. A model in which all the τ_j 's are random variables, except possibly the additive constant, is called a *random-effects model* or Model II or *variance-components model*. Finally, a model in which at least one τ_j is a random variable and at least one τ_j is a constant (not an additive constant) is called a *mixed model*.

There is implicit in the model an assumption that the effects are linearly connected. Further, for tests of significance the errors are assumed to be normally distributed with zero mean and a constant variance σ^2_ϵ .

19.3 A theorem of importance in Model I analysis

We shall now state a theorem of sufficient importance in the discussion of analysis of variance. In various cases of the Model I analysis of variance, the distributions of the constituent sums of squares and their independence may be deduced from this theorem.

A theorem on least squares

Let the random variables y_1, y_2, \dots, y_n be independently normally distributed with common variance σ^2 and let

$$E(y_i) = a_{i1}\tau_1 + a_{i2}\tau_2 + \dots + a_{ik}\tau_k \quad \dots \quad (19.3) \\ (i=1, 2, \dots, n),$$

where a_{ij} 's are elements of a specified matrix A and τ_j 's are unknown parameters. Let rank $A=r$.

$$\text{If } S_1^2 = \min \sum_i (y_i - a_{i1}\tau_1 - \dots - a_{ik}\tau_k)^2$$

with respect to the τ_j 's, then S_1^2/σ^2 is a χ^2 with $(n-r)$ d.f.

Suppose the τ_j 's are subject to s independent conditions, viz.

$$R : \begin{cases} h_{11}\tau_1 + h_{12}\tau_2 + \dots + h_{1k}\tau_k = g_1, \\ h_{21}\tau_1 + h_{22}\tau_2 + \dots + h_{2k}\tau_k = g_2, \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ h_{s1}\tau_1 + h_{s2}\tau_2 + \dots + h_{sk}\tau_k = g_s. \end{cases}$$

$$\text{If } S_2^2 = \min \sum_i (y_i - a_{i1}\tau_1 - \dots - a_{ik}\tau_k)^2,$$

when minimised with respect to the τ_j 's subject to the conditions in R above, then S_2^2/σ^2 is a χ^2 with $(n-r+t)$ d.f., where t is the number of vectors $(h_{i1}, h_{i2}, \dots, h_{ik})$, $i=1, 2, \dots, s$, depending on the rows of the matrix A .

It is true that $(S_2^2 - S_1^2)/\sigma^2$ is a χ^2 with t d.f. It is also true that $(S_2^2 - S_1^2)/\sigma^2$ and S_1^2/σ^2 are independent χ^2 's with t and $(n-r)$ d.f., respectively.

19.4 Tests of general linear hypotheses

A linear hypothesis H_0 , corresponding to the linear model (19.3), specifies the values of one or more linear functions of the parameters, say

$$H_0 : \begin{cases} l_{11}\tau_1 + l_{12}\tau_2 + \dots + l_{1k}\tau_k = b_1, \\ l_{21}\tau_1 + l_{22}\tau_2 + \dots + l_{2k}\tau_k = b_2, \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ l_{m1}\tau_1 + l_{m2}\tau_2 + \dots + l_{mk}\tau_k = b_m. \end{cases}$$

The above m linear functions can be assumed to be independent. Suppose the parameters in the model (19.3) are known to satisfy the s linear restrictions R given in Section 19.3. These conditions also can be taken as independent.

It is necessary that the vectors $(l_{i1}, l_{i2}, \dots, l_{ik})$, $i=1, 2, \dots, m$, in H_0 be linearly dependent on the vectors $(a_{i1}, a_{i2}, \dots, a_{ik})$, $i=1, 2, \dots, n$, and $(h_{i1}, h_{i2}, \dots, h_{ik})$, $i=1, 2, \dots, s$, in order that H_0 may be tested.

Let, as before, rank $A=r$ and t be the number of independent vectors in $(h_{i1}, h_{i2}, \dots, h_{ik})$, $i=1, 2, \dots, s$, that are linearly dependent.

dent on the rows of \mathbf{A} . And let $t+m$ be the number of independent vectors in $(l_{i1}, l_{i2}, \dots, l_{ik})$, $i=1, 2, \dots, m$, and $(h_{i1}, h_{i2}, \dots, h_{ik})$, $i=1, 2, \dots, s$, that are linearly dependent on $(a_{i1}, a_{i2}, \dots, a_{ik})$, $i=1, 2, \dots, n$. Then, by the theorem of Section 19.3, it follows that $\sigma^2\chi^2_R$, which is the minimum value of

$$\sum_i (y_i - a_{i1}\tau_1 - \dots - a_{ik}\tau_k)^2$$

when τ_j 's are subject to the conditions R , is distributed as a $\sigma^2\chi^2$ with $(n-r+t)$ d.f. Similarly, $\sigma^2\chi^2_{R+H_0}$, which is the minimum value of

$$\sum_i (y_i - a_{i1}\tau_1 - \dots - a_{ik}\tau_k)^2$$

when τ_j 's are subject to the conditions in R and H_0 , is distributed as a $\sigma^2\chi^2$ with $(n-r+t+m)$ d.f.

Hence $\chi^2 = \chi^2_{R+H_0} - \chi^2_R$ is distributed as a χ^2 with m d.f. under H_0 , i.e. only if H_0 be true. Then a test for H_0 is provided by

(1) $\chi^2_{R+H_0} - \chi^2_R$, which is a χ^2 with m d.f., if σ^2 is known, or by

(2) $F = \frac{[\chi^2_{R+H_0} - \chi^2_R]/m}{\chi^2_R/(n-r+t)}$, which is distributed as an F with m and $(n-r+t)$ d.f., if σ^2 is not known.

19.5 Analysis of one-way classified data

Let there be n observations, classified into k classes, A_1, A_2, \dots, A_k , the number of observations in the i th class being n_i . Let y_{ij} be the j th observation in the i th class, $i=1, 2, \dots, k$; $j=1, 2, \dots, n_i$. The scheme of classification is given below :

		Classes			
A_1		A_2	A_k
y_{11}		y_{21}	y_{k1}
y_{12}		y_{22}	y_{k2}
\vdots		\vdots			\vdots
y_{1n_1}		y_{2n_2}	.		y_{kn_k}

We may consider that these k classes are the *only* classes in which we are interested. In that case, we have a fixed-effects model and our method of analysis will be as follows.

Here our linear model is

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad \dots \quad (19.4)$$

where μ_i is the fixed effect due to the i th class or the mean of the i th class in the population and ϵ_{ij} 's are the errors which are supposed to be independently and normally distributed with zero mean and variance σ^2 . Denoting $\sum n_i \mu_i / n$ by μ , called the general effect, and $\mu_i - \mu$ by β_i , called the additional effect due to the i th class over the general effect, we can write (19.4) as

$$y_{ij} = \mu + \beta_i + \epsilon_{ij}, \quad \dots \quad (19.4a)$$

where, obviously,

$$\sum_i n_i \beta_i = 0.$$

The least-square estimates of μ and β_i 's are obtained by minimising

$$\sum_i \sum_j (y_{ij} - \mu - \beta_i)^2,$$

the normal equations being

$$\sum_i \sum_j y_{ij} = n\mu + \sum_i n_i \beta_i$$

and $\sum_j y_{ij} = n_i \mu + n_i \beta_i \quad (i=1, 2, \dots, k).$

Solving these equations, we have, since $\sum n_i \beta_i = 0$,

$$\hat{\mu} = y_{00},$$

the grand mean of the observations, and

$$\hat{\beta}_i = y_{i0} - y_{00} \quad (i=1, 2, \dots, k),$$

y_{i0} being the mean of the i th class.

In the model (19.4a), each observation is represented as the sum of three components. Similarly, the analysis of variance partitions the raw sum of squares of the observations, $\sum_i \sum_j y_{ij}^2$, into three components —sum of squares due to the general effect, sum of squares due to the class-effects and sum of squares due to error. We may write

$$y_{ij} = \hat{\mu} + \hat{\beta}_i + \hat{\epsilon}_{ij}$$

or

$$y_{ij} = y_{00} + (y_{i0} - y_{00}) + (y_{ij} - y_{i0}).$$

Squaring both the sides and summing over all the observations,

we get

$$\sum_{i,j} y_{ij}^2 = \sum_i y_{i0}^2 + \sum_j (y_{i0} - y_{00})^2 + \sum_{i,j} (y_{ij} - y_{i0})^2$$

(since the three sums of product terms on the right-hand side vanish),

or $\sum_{i,j} y_{ij}^2 - ny_{00}^2 = \sum_i n_i (y_{i0} - y_{00})^2 + \sum_{i,j} (y_{ij} - y_{i0})^2,$

or $\sum_{i,j} (y_{ij} - y_{00})^2 = \sum_i n_i (y_{i0} - y_{00})^2 + \sum_{i,j} (y_{ij} - y_{i0})^2,$

or total sum of squares = sum of squares due to class-effects
+ sum of squares due to error,

or, in short,

$$\text{total } SS = SSA + SSE. \quad \dots \quad (19.5)$$

Now, the total sum of squares is computed from n quantities like $(y_{ij} - y_{00})$, of which only $n-1$ are independent, since

$$\sum_{i,j} (y_{ij} - y_{00}) = 0.$$

Hence it is said to carry $n-1$ degrees of freedom (d.f.).

Similarly, the sum of squares due to class-effects is obtained by squaring k quantities like $(y_{i0} - y_{00})$, satisfying one condition :

$$\sum_i n_i (y_{i0} - y_{00}) = 0.$$

Thus it carries $k-1$ d.f.

Lastly, the sum of squares due to error is calculated by squaring n quantities like $(y_{ij} - y_{i0})$, which satisfy k conditions :

$$\sum_j (y_{ij} - y_{i0}) = 0 \quad (i=1, 2, \dots, k).$$

Hence this sum of squares is based on $n-k$ d.f.

So the degrees of freedom are also additive :

$$n-1 = (k-1) + (n-k).$$

Dividing an SS by its d.f., we get the corresponding variance or mean square (MS).

Thus $SSA/(k-1) = \text{mean square due to class-effects} = MSA$

and $SSE/(n-k) = \text{mean square due to error} = MSE.$

Now, SSA and SSE add up to total SS and the corresponding d.f.'s $(k-1)$ and $(n-k)$ add up to total d.f. $(n-1)$, but the MSA and MSE will not add up to total MS . So, though the procedure is called an analysis of variance, it is actually an analysis of SS .

Now, by partitioning the total SS and total $d.f.$ into two parts, we shall be able to test the hypothesis that the k class means are equal, i.e. the hypothesis $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ or its equivalent hypothesis in terms of β_i 's, viz. $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$.

To obtain the appropriate test we calculate the expectations of the MSA and MSE .

From the model (19.4a), we have

$$y_{i0} = \mu + \beta_i + e_{i0}$$

and

$$y_{00} = \mu + e_{00}.$$

$$\text{Then } SSE = \sum_{i,j} (y_{ij} - y_{i0})^2 = \sum_{i,j} (e_{ij} - e_{i0})^2 = \sum_{i,j} e_{ij}^2 - \sum_i n_i e_{i0}^2.$$

Taking expectations, we have

$$\begin{aligned} E(SSE) &= \sum_{i,j} E(e_{ij}^2) - \sum_i n_i E(e_{i0}^2) \\ &= n\sigma_e^2 - \sum_i n_i (\sigma_e^2/n_i) = n\sigma_e^2 - k\sigma_e^2 \\ &= (n-k)\sigma_e^2. \end{aligned}$$

$$\text{Thus } E(MSE) = E[SSE/(n-k)] = \sigma_e^2. \quad \dots \quad (19.6)$$

$$\text{Again, } SSA = \sum_i n_i (y_{i0} - y_{00})^2 = \sum_i n_i (\beta_i + e_{i0} - e_{00})^2.$$

$$\begin{aligned} \text{Hence } E(SSA) &= \sum_i n_i \beta_i^2 + E[\sum_i n_i (e_{i0} - e_{00})^2], \text{ since } \beta_i E(e_{i0} - e_{00}) = 0 \\ &= \sum_i n_i \beta_i^2 + E[\sum_i n_i e_{i0}^2 - n e_{00}^2] \\ &= \sum_i n_i \beta_i^2 + [\sum_i n_i \cdot \sigma_e^2/n_i - n \cdot \sigma_e^2/n] \\ &= \sum_i n_i \beta_i^2 + (k-1)\sigma_e^2. \end{aligned}$$

$$\begin{aligned} \text{Thus } E(MSA) &= E[SSA/(k-1)] = \sigma_e^2 + (\sum_i n_i \beta_i^2)/(k-1) \\ &= \sigma_e^2 + \phi(\beta_1, \beta_2, \dots, \beta_k), \quad \dots \quad (19.7) \end{aligned}$$

where $\phi(\beta_1, \beta_2, \dots, \beta_k)$ is a variance-like function of the β_i 's. This function takes the value zero when the null hypothesis $H_0 : \beta_1 = \dots = \beta_k = 0$ is true ; otherwise, it is a positive quantity. Thus MSA gives us an unbiased estimate of σ_e^2 when H_0 is true ; otherwise, its expectation is greater than σ_e^2 . On the other hand, MSE is always an unbiased estimate of σ_e^2 . If the null hypothesis H_0 be true, $E(MSA) = E(MSE)$. The ratio $F = MSA/MSE$ will thus give us a test for the null hypothesis. So to know whether an observed F is significantly greater than 1 or not, we are to derive the distribution

of $F = MSA/MSE$ under H_0 . This follows from the fact that SSA/σ^2 and SSE/σ^2 are independently distributed as χ^2 's with $(k-1)$ and $(n-k)$ d.f., respectively, when H_0 is true. (This result is obtainable by an application of the theorem of Section 19.3.) Hence $F = MSA/MSE$ follows the F distribution with $(k-1)$, $(n-k)$ d.f.

Thus the hypothesis H_0 is rejected at a specified level α if for the given values

$$F = \frac{MSA}{MSE} > F_{\alpha ; (k-1), (n-k)},$$

where $F_{\alpha ; (k-1), (n-k)}$ is the upper α -point of the F distribution with $(k-1)$, $(n-k)$ d.f. Otherwise H_0 is accepted.

In the analysis of variance, the values of the SS , d.f., MS and F are usually exhibited in a table—called the analysis of variance table.

TABLE 19.1
ANALYSIS OF VARIANCE FOR ONE-WAY CLASSIFIED DATA

Source of variation	d.f.	SS	MS	F	F at level 1% 5%
Between classes	$k-1$	$\sum_i n_i (y_{i0} - \bar{y}_{00})^2$ $= SSA$	$\sum_i n_i (y_{i0} - \bar{y}_{00})^2 / (k-1)$ $= MSA$	$\frac{MSA}{MSE}$	$F_{01 ; (k-1), (n-k)}$
Error	$n-k$	$\sum_{ij} (y_{ij} - \bar{y}_{i0})^2$ $= SSE$	$\sum_{ij} (y_{ij} - \bar{y}_{i0})^2 / (n-k)$ $= MSE$		$F_{05 ; (k-1), (n-k)}$
Total	$n-1$	$\sum_{ij} (y_{ij} - \bar{y}_{00})^2$			

If the previous null hypothesis H_0 is rejected, then we may test for the equality of two class means, say the hypothesis $H_{01} : \mu_i = \mu_j$, with the help of

$$t = \frac{\bar{y}_{i0} - \bar{y}_{j0}}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \text{ with } (n-k) \text{ d.f.}$$

If $n_1 = n_2 = \dots = n_k = n_0$, then this reduces to the simple form :

$$t = \frac{\bar{y}_{i0} - \bar{y}_{j0}}{\sqrt{2MSE/n_0}} \text{ with } k(n_0 - 1) \text{ d.f.},$$

since $n = n_0 k$ now.

If $|t| = \frac{|y_{i0} - y_{i'0}|}{\sqrt{2MSE/n_0}} > t_{\alpha/2, k(n_0-1)}$, then H_{01} is rejected at the level α . That is to say, H_{01} is rejected at the level α if

$$|y_{i0} - y_{i'0}| > t_{\alpha/2, k(n_0-1)} \times \sqrt{2MSE/n_0}.$$

Thus, to compare the class means two at a time, we are to calculate $t_{\alpha/2, k(n_0-1)} \times \sqrt{2MSE/n_0}$, called the *critical difference* or the *least significant difference (lsd)*, and if the difference between the observed class means, i.e. $|y_{i0} - y_{i'0}|$, is greater than the *lsd*, then $H_{01} : \mu_i = \mu_{i'}$ is rejected at the $100\alpha\%$ level ; otherwise, it is accepted. $t_{\alpha/2, k(n_0-1)}$ is the upper $\alpha/2$ -point of the *t* distribution with $k(n_0-1)$ d.f.

The above model was called the *fixed-effects model* because the k classes in the experiment were the only classes in which we were interested. But the situation may be like this : There are a large number of classes and we want to know from an experiment whether all these class-effects are equal or not. Now, due to considerations of cost, time or space, it is not possible to include in our experiment all the available classes. We can include only a sample of these classes, and we want to infer about all the classes, whether included in the experiment or not, from the results of the classes included in the experiment. Then the β_i 's of the previous fixed model will not be fixed parameters but will be random variables, as the model is now a random one. In the random model we shall consider *balanced cases*, because tests for the random model are known only for balanced cases. A one-way classification is called balanced if the numbers of observations under different categories are the same. Higher-order classifications are balanced if the numbers of observations in cells are equal. The analysis of variance table will remain the same as in the corresponding balanced case of the fixed model. But in this random model, if we find that effects due to different classes are not the same, we cannot apply the *t* test to find out which classes differ, as we have not included all the classes in the experiment.

In this case, the model is

$$y_{ij} = \mu + b_i + e_{ij}, \quad i=1, 2, \dots, k; \quad j=1, 2, \dots, r, \dots \quad (19.8)$$

with the $(k+r)$ random variables b_i 's and e_{ij} 's being completely

independent, b_i 's being normal with mean zero and variance σ_b^2 . e_{ij} 's are, as before, normal with mean zero and variance σ_e^2 .

The variance of an observation, y_{ij} , is

$$\sigma_y^2 = \sigma_b^2 + \sigma_e^2.$$

So σ_b^2 and σ_e^2 are called the *components of the variance* of y and the model a variance-components model.

In the random-effects model the observations (y_{ij}) have the same expectation μ and the same variance $\sigma_y^2 = \sigma_b^2 + \sigma_e^2$. But the observations are not statistically independent. This dependence can be expressed in terms of the intra-class correlation coefficient, which is the ordinary correlation coefficient between any two observations, y_{ij} and $y_{i'j}$ ($j \neq j'$), of the same class :

$$\rho = E(y_{ij} - \mu)(y_{i'j} - \mu) / \sigma_y^2 = E[(b_i + e_{ij})(b_{i'} + e_{i'j})] / \sigma_y^2 = \sigma_b^2 / (\sigma_b^2 + \sigma_e^2).$$

The appropriate hypothesis for the equality of all the class means in the case of the random model will be $H_0 : \sigma_b^2 = 0$.

From (19.8), we have

$$y_{i0} = \mu + b_i + e_{i0}, \text{ where } e_{i0} = \sum_j e_{ij} / r,$$

$$\text{and } y_{00} = \mu + b_0 + e_{00}, \text{ where } b_0 = \sum_i b_i / k, e_{00} = \sum_i e_{i0} / k.$$

$$\text{Then } SSE = \sum_{i,j} (y_{ij} - y_{i0})^2 = \sum_{i,j} (e_{ij} - e_{i0})^2.$$

So, as in equation (19.6), here also we have

$$E(MSE) = \sigma_e^2. \quad \dots \quad (19.9)$$

$$\text{Next, } SSA = r \sum_i (y_{i0} - y_{00})^2 = r \sum_i (b_i - b_0 + e_{i0} - e_{00})^2.$$

Taking expectations and noting that b_i 's are independent of e_{ij} 's, we have

$$\begin{aligned} E(SSA) &= E[r \sum_i (b_i - b_0)^2] + E[r \sum_i (e_{i0} - e_{00})^2] \\ &= E[r \sum_i b_i^2 - nb_0^2] + E[r \sum_i e_{i0}^2 - ne_{00}^2], \text{ where } rk = n \\ &= n\sigma_b^2 - n\sigma_b^2/k + (k-1)\sigma_e^2 \\ &= (n-r)\sigma_b^2 + (k-1)\sigma_e^2 = r(k-1)\sigma_b^2 + (k-1)\sigma_e^2. \end{aligned}$$

Hence

$$E(MSA) = E[SSA/(k-1)] = \sigma_b^2 + r\sigma_e^2. \quad \dots \quad (19.10)$$

Thus we see that if H_0 be true, then $E(MSE) = E(MSA)$, and hence a test for H_0 can be obtained by using the statistic $F = MSA/MSE$.

It can be shown that when H_0 is true, SSA/σ_e^2 and SSE/σ_e^2 are independently distributed as χ^2 's with $(k-1)$ and $(n-k)$ d.f., respectively. Thus $F=MSA/MSE$ follows the F distribution with $(k-1)$, $(n-k)$ d.f., when H_0 is true.

Estimates of the components of variance are obtained in the following manner :

$$\delta_e^2 = MSE, \delta_b^2 + r\delta_e^2 = MSA \text{ and so } \delta_b^2 = (MSA - MSE)/r.$$

(If δ_b^2 comes out negative according to this formula, then it is taken to be zero.)

We present below the $E(MS)$'s for the two models in the case of balanced one-way classified data in the same table for the purpose of comparison :

TABLE 19.2
E(MS)'S UNDER BALANCED MODEL I AND MODEL II FOR
ONE-WAY CLASSIFIED DATA

Source of variation	d.f.	SS	MS	E(MS)	
				Model I	Model II
Between classes	$k-1$	SSA	MSA	$\sigma_e^2 + r \sum_i \beta_i^2 / (k-1)$	$\sigma_e^2 + r\sigma_b^2$
Error	$n-k$	SSE	MSE	σ_e^2	σ_e^2
Total	$n-1$			—	

Computational procedure for the analysis of one-way classified data under fixed model :

(1) Calculate total for each class : $T_{10}, T_{20}, \dots, T_{k0}$,

where

$$T_{i0} = \sum_{j=1}^{n_i} y_{ij}.$$

(2) Calculate the grand total : $T_{00} = \sum_i \sum_j y_{ij} = \sum_i T_{i0}$.

(3) Calculate the raw total SS : $\sum_i \sum_j y_{ij}^2$.

(4) Calculate $\sum_i T_{i0}^2 / n_i$.

(5) Calculate correction factor : T_{00}^2 / n .

(6) Total SS = $\sum_i \sum_j y_{ij}^2 - T_{00}^2 / n$ = value obtained in step (3)—that in step (5).

(7) $SSA = \sum_i T_{10}^2/n_i - T_{00}^2/n$ = value obtained in step (4)—that in step (5).

(8) $SSE = \text{total } SS - SSA$ = value obtained in step (6)—that in step (7).

It may be noted that sometimes calculations may be simplified by a change of base and scale of the observations. This will not affect the tests, though the estimates will change. The above procedure can be easily adapted to the balanced case of the random or fixed model.

The reader is referred to Ex. 16·9 for an illustration of the analysis of one-way classified data.

19.6 Analysis of two-way classified data with one observation per cell.

We can plan an experiment in such a way as to study the effects of two factors in the same experiment. For each factor there will be a number of classes or levels. In the fixed-effects model, there will be only fixed levels of the two factors. We shall first consider the case of one observation per cell (or combination).

Let the factors be A and B and the levels A_1, A_2, \dots, A_p and B_1, B_2, \dots, B_q . Let y_{ij} be the observation under the i th level of A and the j th level of B . The observations can be represented as follows :

TABLE 19.3
TABLE OF OBSERVATIONS

	B_1	B_2	\dots	B_j	\dots	\dots	B_q	Total	Mean
A_1	y_{11}	y_{12}		y_{1j}			y_{1q}	T_{10}	y_{10}
A_2	y_{21}	y_{22}		y_{2j}			y_{2q}	T_{20}	y_{20}
\vdots	\vdots	\vdots		\vdots			\vdots	\vdots	\vdots
A_i	y_{i1}	y_{i2}		y_{ij}			y_{iq}	T_{i0}	y_{i0}
\vdots	\vdots	\vdots		\vdots			\vdots	\vdots	\vdots
A_p	y_{p1}	y_{p2}		y_{pj}			y_{pq}	T_{p0}	y_{p0}
Total	T_{01}	T_{02}		T_{0j}			T_{0q}	T_{00}	—
Mean	y_{01}	y_{02}		y_{0j}			y_{0q}	—	y_{00}

Here the mathematical model may be written as

$$y_{ij} = \mu_{ij} + \epsilon_{ij}, \dots \quad (19.11)$$

where ϵ_{ij} 's are independently normally distributed with mean zero and variance σ^2 . Corresponding to the above table of observations (Table 19.3), we can form a table of expected values of observations (Table 19.4).

TABLE 19.4
TABLE OF EXPECTATIONS

	B_1	$B_2, \dots, B_j, \dots, B_q$	Mean	Difference
A_1	μ_{11}	$\mu_{12}, \dots, \mu_{1j}, \dots, \mu_{1q}$	μ_{10}	$\mu_{10} - \mu = \alpha_1$
A_2	μ_{21}	$\mu_{22}, \dots, \mu_{2j}, \dots, \mu_{2q}$	μ_{20}	$\mu_{20} - \mu = \alpha_2$
\vdots	\vdots	\vdots	\vdots	\vdots
A_i	μ_{i1}	$\mu_{i2}, \dots, \mu_{ij}, \dots, \mu_{iq}$	μ_{i0}	$\mu_{i0} - \mu = \alpha_i$
\vdots	\vdots	\vdots	\vdots	\vdots
A_p	μ_{p1}	$\mu_{p2}, \dots, \mu_{pj}, \dots, \mu_{pq}$	μ_{p0}	$\mu_{p0} - \mu = \alpha_q$
Mean	μ_{01}	$\mu_{02}, \dots, \mu_{0j}, \dots, \mu_{0q}$	μ	
Difference	$\mu_{01} - \mu$ $= \beta_1$	$\mu_{02} - \mu$ $= \beta_2$	$\mu_{0j} - \mu$ $= \beta_j$	$\mu_{0q} - \mu$ $= \beta_q$

Now, we can think of μ_{ij} as composed of the following parts :

$$\begin{aligned} \mu_{ij} &= \mu + (\mu_{i0} - \mu) + (\mu_{0j} - \mu) + (\mu_{ij} - \mu_{i0} - \mu_{0j} + \mu) \\ &= \mu + \alpha_i + \beta_j + \gamma_{ij}, \text{ say,} \end{aligned} \dots \quad (19.11a)$$

where μ is a constant general effect, present in all the observations ;

$$\alpha_i = \mu_{i0} - \mu$$

is an effect due to the i th level of the factor A , which is common to all the observations belonging to this level of A ;

$$\beta_j = \mu_{0j} - \mu$$

is an effect due to the j th level of the factor B , which is common to all the observations belonging to this level of B ; and

$$\gamma_{ij} = \mu_{ij} - \mu_{i0} - \mu_{0j} + \mu$$

is called the *interaction* between the i th level of A and the j th level

of B . It is an effect peculiar to the combination (A_i, B_j) . It is not present in the i th level of A or in the j th level of B if not taken together. If the joint effect of A_i and B_j is different from the sum of the effects due to A_i and B_j taken individually, we say that there is interaction and it is measured by γ_{ij} .

From the table of expected values (Table 19.4) for the fixed-effects model, it is clear that

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i \gamma_{ij} = 0, \quad \sum_j \gamma_{ij} = 0.$$

for all j for all i

The observation y_{ij} in the (i, j) th cell may thus be expressed as—

$$\begin{aligned} y_{ij} = & \text{a constant general effect } (\mu) \\ & + \text{an effect due to the } i\text{th level of } A \ (\alpha_i) \\ & + \text{an effect due to the } j\text{th level of } B \ (\beta_j) \\ & + \text{interaction between } A_i \text{ and } B_j \ (\gamma_{ij}) \\ & + \text{error } (e_{ij}). \end{aligned}$$

In the case of two-way classified data with one observation per cell, the interaction (γ_{ij}) cannot be estimated, and we shall assume for the fixed-effects model that there is no interaction, i.e. $\gamma_{ij}=0$ for all i and j . So the model reduces to

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad (19.11b)$$

with $\sum_i \alpha_i = \sum_j \beta_j = 0$ and e_{ij} 's being independently normally distributed with mean zero and variance σ_e^2 .

The least-square estimates, obtained by minimising

$$\sum_i \sum_j (y_{ij} - \mu - \alpha_i - \beta_j)^2,$$

are

$$\hat{\mu} = y_{00},$$

$$\hat{\alpha}_i = y_{i0} - y_{00}$$

and

$$\hat{\beta}_j = y_{0j} - y_{00}.$$

In the model, each observation is the sum of four components, and the analysis of variance partitions the raw SS , $\sum_i \sum_j y_{ij}^2$, also into four components— SS due to general effect, SS due to factor A , SS due to factor B and SS due to error—as follows :

$$y_{ij} = y_{00} + (y_{i0} - y_{00}) + (y_{0j} - y_{00}) + (y_{ij} - y_{i0} - y_{0j} + y_{00}).$$

Squaring both sides and summing over i and j , we get

$$\begin{aligned}\sum_i \sum_j y_{ij}^2 &= pq(y_{00})^2 + q \sum_i (y_{i0} - y_{00})^2 + p \sum_j (y_{0j} - y_{00})^2 \\ &\quad + \sum_i \sum_j (y_{ij} - y_{i0} - y_{0j} + y_{00})^2\end{aligned}$$

or $\sum_i \sum_j (y_{ij} - y_{00})^2 = q \sum_i (y_{i0} - y_{00})^2 + p \sum_j (y_{0j} - y_{00})^2 + \sum_i \sum_j (y_{ij} - y_{i0} - y_{0j} + y_{00})^2.$

In words,

total $SS = SS$ due to factor $A + SS$ due to factor $B + SS$ due to error,
or, in short,

$$\text{total } SS = SSA + SSB + SSE. \quad \dots \quad (19.12)$$

The corresponding partitioning of the total $d.f.$ is as follows :

$$pq-1 = (p-1) + (q-1) + (p-1)(q-1).$$

Dividing an SS by its $d.f.$, we get the corresponding MS .

By partitioning the total SS and the total $d.f.$ into three components each, we shall be able to test the following two hypotheses—

$$H_{01} : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$$

and $H_{02} : \beta_1 = \beta_2 = \dots = \beta_q = 0$

for the equality of the effects of the different levels of A and of the different levels of B , respectively. To derive appropriate test for the hypotheses, we find the expectations of the mean squares. It can be shown that

$$E(MSA) = \sigma_e^2 + q \sum_i \alpha_i^2 / (p-1), \quad \dots \quad (19.13)$$

$$E(MSB) = \sigma_e^2 + p \sum_j \beta_j^2 / (q-1) \quad \dots \quad (19.14)$$

and $E(MSE) = \sigma_e^2. \quad \dots \quad (19.15)$

If $H_{01} : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ is true, $E(MSA) = E(MSE)$, and hence $F = MSA/MSE$ will give the test of H_{01} . So a test for the hypothesis of equality of the effects of the different levels of A is provided by this F , which follows the F distribution with $(p-1)$, $(p-1)(q-1)$ $d.f.$. This result is obtained by an application of the theorem of Section 19.3. Thus the null hypothesis will be rejected at the $100\alpha\%$ level if (and only if)

$$F = \frac{MSA}{MSE} > F_{\alpha ; (p-1), (p-1)(q-1)}.$$

Similarly, $H_{02} : \beta_1 = \beta_2 = \dots = \beta_q = 0$, for the equality of the effects of the different levels of B , is rejected at the $100\alpha\%$ level if

$$F = \frac{MSB}{MSE} > F_{\alpha ; (q-1), (p-1)(q-1)}$$

These calculations are shown in the following analysis of variance table (Table 19.5).

After performing the F test, we can test for the equality of the means of any pair of A classes or any pair of B classes with the help of the t test.

TABLE 19.5
ANALYSIS OF VARIANCE FOR TWO-WAY CLASSIFIED DATA
WITH ONE OBSERVATION PER CELL

Source of variation	d.f.	SS	MS	F	F at level 1% 5%
Between the levels of A	$p-1$	$q \sum_i (y_{i0} - \bar{y}_{00})^2$ $= SSA$	$SSA/(p-1)$ $= MSA$	MSA MSE	$F_{0.1:(p-1),(q-1)(p-1)}$
Between the levels of B	$q-1$	$p \sum_j (y_{0j} - \bar{y}_{00})^2$ $- SSB$	$SSB/(q-1)$ $= MSB$	MSB MSE	$F_{0.1:(q-1),(p-1)(q-1)}$
Error	$(p-1)(q-1)$	$\sum_{i,j} (y_{ij} - \bar{y}_{i0} - \bar{y}_{0j} + \bar{y}_{00})^2 = SSE$	$SSE/(p-1)(q-1)$ $= MSE$		
Total	$pq-1$	$\sum_{i,j} (y_{ij} - \bar{y}_{00})^2$			

In the above case, we assumed that we had only p levels of A and q levels of B . But it may be that the total number of levels of A is greater than p and that of B is greater than q . Then in the model α_i 's, β_j 's and γ_{ij} 's are not fixed parameters but are supposed to be random variables. Thus, in the random model we assume that

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}, \dots \quad (19.16)$$

and that the a_i 's, b_j 's, c_{ij} 's and ϵ_{ij} 's are independently normal with zero means and respective variances σ_a^2 , σ_b^2 , σ_{ab}^2 and σ_e^2 . The two hypotheses that can be tested in this case of random model are $H_{01} : \sigma_a^2 = 0$ and $H_{02} : \sigma_{ab}^2 = 0$. In this case, we need not assume that interaction effects are all zero as we did in the fixed model, though we cannot test or estimate it. The expectations of the *MSs* in the case of random model are :

$$E(MSA) = \sigma_a^2 + \sigma_{ab}^2 + q\sigma_e^2, \quad \dots \quad (19.17)$$

$$E(MSB) = \sigma_b^2 + \sigma_{ab}^2 + p\sigma_e^2 \quad \dots \quad (19.18)$$

and $E(MSE) = \sigma_e^2 + \sigma_{ab}^2. \quad \dots \quad (19.19)$

Thus we see that the test for the effects of *A* classes or that for the effects of *B* classes will be the same here as in the fixed model. And the corresponding *F* statistics have each the *F* distribution under the appropriate null hypothesis.

By equating the observed *MSs* to their expectations, we get as estimates of the components of variance the following :

$$\left. \begin{aligned} \hat{\sigma}_e^2 + \hat{\sigma}_{ab}^2 &= MSE, \\ \hat{\sigma}_a^2 &= (MSB - MSE)/p, \\ \hat{\sigma}_b^2 &= (MSA - MSE)/q. \end{aligned} \right\} \quad \dots \quad (19.20)$$

(If any of these formulae leads to a negative value for the corresponding estimate, then the estimate is taken to be zero.)

In the case of mixed model, we assume that the levels of one of the factors, say *A*, have all been included in the experiment and those of the other factor, in this case *B*, are not all included in the experiment. Then the α_i 's are fixed parameters and the β_j 's are supposed to be random variables. The interaction effects also become random due to the sampling of the β_j 's.

We assume that

$$y_{ij} = m_{ij} + \epsilon_{ij},$$

where the errors ϵ_{ij} 's are independently distributed with mean zero and variance σ_e^2 and also ϵ_{ij} 's are statistically independent of the m_{ij} 's.

Subdividing m_{ij} , we get the following model equation :

$$y_{ij} = \mu + \alpha_i + b_j + c_{ij} + e_{ij}, \quad \dots \quad (19.21)$$

with $\sum_i \alpha_i = 0$; $\sum_j c_{ij} = 0$ for all j .

The random variables b_j 's and c_{ij} 's have zero means, but they are not independent.

We define

$$\sigma_a^2 = \frac{1}{p-1} \sum_i \alpha_i^2, \quad \sigma_b^2 = \text{var}(b_j), \quad \sigma_c^2 = \frac{1}{p-1} \sum_j \text{var}(c_{ij}).$$

The expectations of the MSs are as follows :

$$\left. \begin{aligned} E(MSA) &= \sigma_a^2 + \sigma_c^2 + q\sigma_b^2, \\ E(MSB) &= \sigma_a^2 + p\sigma_b^2, \\ E(MSE) &= \sigma_c^2 + \sigma_b^2. \end{aligned} \right\} \quad \dots \quad (19.22)$$

These expectations lead to the following unbiased estimates :

$$\hat{\sigma}_a^2 + \hat{\sigma}_c^2 = MSE; \quad \hat{\sigma}_b^2 = (MSB - MSE)/p \text{ if } \sigma_c^2 = 0;$$

while μ and α_i 's are estimated as in the fixed-effects model.

We assume that the b_j 's, c_{ij} 's and e_{ij} 's are jointly normal for performing the tests of hypotheses.

We further make the following simplifying assumption about the symmetry of the covariance matrix of m_{ij} :

The variances of the elements m_{ij} are all equal and, similarly, the $p(p-1)/2$ covariances of m_{ij} , $m_{ij'}$ are all equal. σ_a^2 and σ_c^2 depend on these variances and covariances.

This assumption of symmetry may not be always desirable, but it is made in order that MSA/MSE may have an F distribution.

The test for the equality of the effects due to the p levels of A , i.e. for $H_{01} : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$, is provided by $F = MSA/MSE$, which has the F -distribution with $(p-1)$, $(p-1)(q-1)$ d.f., when H_{01} is true. In this case, we need not assume absence of interaction.

The test for the equality of the effects due to the levels of B , i.e. for $H_{02} : \sigma_b^2 = 0$, is provided by $F = MSB/MSE$, which also has the F distribution with $(q-1)$, $(p-1)(q-1)$ d.f. when H_{02} is true, if we assume that $\sigma_c^2 = 0$, i.e. that there is no interaction present.

The $E(MS)$ s, under the different models for two-way classified data with one observation per cell, are given below :



TABLE 19.6
E(MS)s UNDER DIFFERENT MODELS FOR TWO-WAY
CLASSIFIED DATA WITH ONE OBSERVATION PER CELL

Source of variation	d.f.	SS	MS	E(MS)		
				Model I	Model II	Mixed model
Between the levels of A	$p-1$	SSA	MSA	$\sigma_e^2 + \frac{q \sum a_i^2}{p-1}$	$\sigma_e^2 + \sigma_{ab}^2 + q\sigma_a^2$	$\sigma_e^2 + \sigma_{ab}^2 + q\sigma_a^2$
Between the levels of B	$q-1$	SSB	MSB	$\sigma_e^2 + \frac{p \sum \beta_j^2}{q-1}$	$\sigma_e^2 + \sigma_{ab}^2 + p\sigma_b^2$	$\sigma_e^2 + p\sigma_b^2$
Error	$(p-1)(q-1)$	SSE	MSE	σ_e^2	$\sigma_e^2 + \sigma_{ab}^2$	$\sigma_e^2 + \sigma_{ab}^2$
Total	$pq-1$			—	—	—

(Under the mixed model, the factor A is fixed in the above table.)

Computational procedure for the analysis of two-way classified data with one observation per cell :

- (1) Calculate total for each of the p A-classes : $T_{10}, T_{20}, \dots, T_{p0}$.
- (2) Calculate total for each of the q B-classes : $T_{01}, T_{02}, \dots, T_{0q}$.
- (3) Calculate grand total : $T_{00} = \sum_i T_{i0} = \sum_j T_{0j} = \sum_i \sum_j y_{ij}$.
- (4) Calculate raw total SS : $\sum_i \sum_j y_{ij}^2$.
- (5) Calculate correction factor : T_{00}^2/p .
- (6) Calculate $\frac{1}{q} \sum_i T_{i0}^2$.
- (7) Calculate $\frac{1}{p} \sum_j T_{0j}^2$.
- (8) Total SS = $\sum_i \sum_j y_{ij}^2 - T_{00}^2/pq = (4) - (5)$.
- (9) $SSA = \frac{1}{q} \sum_i T_{i0}^2 - T_{00}^2/pq = (6) - (5)$.
- (10) $SSB = \frac{1}{p} \sum_j T_{0j}^2 - T_{00}^2/pq = (7) - (5)$.
- (11) SSE = total SS - SSA - SSB = (8) - (9) - (10).

Ex. 19.1 An experiment was conducted to determine the effects of different dates of planting and different methods of planting on the yield of sugar-cane (plot size : 120' x 10'). The data below show the yields of sugar-cane in md. for four different dates and three methods of planting :

Method of planting	October	November	February	March
I	7.10	3.69	4.70	1.90
II	10.29	4.79	4.58	2.64
III	8.30	3.58	4.90	1.80

Carry out an analysis of variance for the above data.

Let y_{ij} be the yield of sugar-cane for the i th method ($i=1, 2, 3$) and the j th date ($j=1, 2, 3, 4$).

The grand total and totals for the 3 methods and the 4 dates are shown below :

Method of planting	1	Date of planting 2	3	4	Total
1	7.10	3.69	4.70	1.90	17.39
2	10.29	4.79	4.58	2.65	22.31
3	8.30	3.58	4.90	1.80	18.58
Total	25.69	12.06	14.18	6.35	58.28

In this case,

$$\sum \sum y_{ij}^2 = 355.5096,$$

$$\text{correction factor} = \frac{T_{00}^2}{pq} = \frac{(58.28)^2}{12} = 283.0465,$$

$$\frac{\sum T_{i0}^2}{q} = \frac{(17.39)^2 + (22.31)^2 + (18.58)^2}{4} = 286.3412,$$

and

$$\frac{\sum T_{0j}^2}{p} = \frac{(25.69)^2 + (12.06)^2 + (14.18)^2 + (6.35)^2}{3} = 348.9382.$$

Therefore,

$$\text{total } SS = \sum_i \sum_j y_{ij}^2 - \frac{T_{00}^2}{pq} = 355.5096 - 283.0465 \\ = 72.4631,$$

$$SS \text{ due to methods} = \frac{\sum_i T_{0j}^2}{q} - \frac{T_{00}^2}{pq} = 286.3412 - 283.0465 \\ = 3.2947,$$

$$SS \text{ due to dates} = \frac{\sum_j T_{0j}^2}{p} - \frac{T_{00}^2}{pq} = 348.9382 - 283.0465 \\ = 65.8917,$$

and $SSE = \text{total } SS - SS \text{ due to methods} - SS \text{ due to dates}$
 $= 72.4631 - 3.2947 - 65.8917 = 3.2767.$

TABLE 19.7
 ANALYSIS OF VARIANCE
 FOR THE DATA OF EX. 19.1

Source of variation	d.f.	SS	MS	F	F at level 1% 5%
Due to methods	2	3.2947	1.6473	3.02	10.92 5.14
Due to dates	3	65.8917	21.9639	40.22	9.78 4.76
Error	6	3.2767	0.5461		
Total	11	72.4631			

The observed F for methods of planting, being smaller than $F_{0.05; 2,6}$, is insignificant at both the levels. But the F for dates of planting is greater than $F_{0.01; 3,6}$ and hence is significant at both the levels. This indicates that the different methods of planting affect the mean yield of sugar-cane in the same manner. But the mean yield differs with different dates of planting.

If the four dates of planting included in the above experiment be the only dates in which we are interested, then the next question that arises is : which one of the four dates will give us the maximum mean yield ? To answer this question we compute the critical

difference at, say, the 5% level.

$$t_{0.025, 6} \times \sqrt{2MSE/3} \\ = 2.447 \times \sqrt{0.3641} = 2.447 \times 0.6034 = 1.48.$$

The mean yields of sugar-cane for the four dates of planting arranged in order of magnitude, are :

October :	8.56,
February :	4.73,
November :	4.02,
March :	2.12.

Thus we find that October gives the maximum mean yield and March the minimum. February and November show no significant difference, but their mean yield is significantly less than that for October and significantly more than that for March

19.7 Analysis of two-way classified data with m observations per cell

In the preceding section, it was seen that we cannot obtain an estimate of, or make a test for, the interaction effect in the case of two-way classification with just one observation per cell. This is possible, however, if some or all of the cells contain more than one observation. For simplicity, we shall assume that there is an equal number (m) of observations in each cell. The m observations in the (i, j) th cell will be denoted by $y_{ij1}, y_{ij2}, \dots, y_{ijk}, \dots, y_{ijm}$. Thus a typical observation is y_{ijk} —the k th observation for the i th level of A and the j th level of B ($i=1, 2, \dots, p$; $j=1, 2, \dots, q$; $k=1, 2, \dots, m$). The mathematical model is given by

$$y_{ijk} = \mu_{ij} + e_{ijk}, \quad \dots \quad (19.23)$$

where μ_{ij} is the true value for the (i, j) th cell and e_{ijk} is the error. e_{ijk} 's are assumed to be independently normally distributed with mean zero and variance σ^2_e . The decomposition of μ_{ij} into different parts is the same as in (19.11a). In the fixed-effects model, we again take

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad \dots \quad (19.24)$$

where $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$.
for all j for all i

The least-square estimates, obtained by minimising

$$\sum_i \sum_j \sum_k (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2,$$

are

$$\hat{\mu} = y_{000}, \hat{\alpha}_i = y_{i00} - y_{000}, \hat{\beta}_j = y_{0j0} - y_{000}$$

and

$$\hat{\gamma}_{ij} = y_{ij0} - y_{i00} - y_{0j0} + y_{000}.$$

The analysis of variance is based on the relation

$$\begin{aligned} \sum_i \sum_j \sum_k (y_{ijk} - y_{000})^2 &= mq \sum_i (y_{i00} - y_{000})^2 + mp \sum_j (y_{0j0} - y_{000})^2 \\ &\quad + m \sum_i \sum_j (y_{ij0} - y_{i00} - y_{0j0} + y_{000})^2 \\ &\quad + \sum_i \sum_j \sum_k (y_{ijk} - y_{ij0})^2 \end{aligned}$$

or, in words,

total $SS = SS$ due to factor $A + SS$ due to factor $B + SS$ due to interaction of A and $B + SS$ due to error

or, in short,

$$\text{total } SS = SSA + SSB + SS(AB) + SSE.$$

The corresponding partitioning of the total $d.f.$ is as follows :

$$mpq - 1 = (p-1) + (q-1) + (p-1)(q-1) + pq(m-1).$$

By partitioning the total SS and the total $d.f.$ into these components, we shall be able to test the following three hypotheses—

$$H_{01} : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0,$$

$$H_{02} : \beta_1 = \beta_2 = \dots = \beta_q = 0$$

and $H_{03} : \gamma_{ij} = 0$, for all i and j .

H_{03} is the new hypothesis that we are able to test by taking $m (> 1)$ observations per cell, and it expresses the independence of the two factors A and B . The appropriate tests are suggested by the following $E(MS)$ s. It can be shown that

$$E(MSA) = \sigma_e^2 + \frac{mq}{p-1} \sum_i \alpha_i^2, \quad \dots \quad (19.25)$$

$$E(MSB) = \sigma_e^2 + \frac{mp}{q-1} \sum_j \beta_j^2, \quad \dots \quad (19.26)$$

$$E(MS(AB)) = \sigma_0^2 + \frac{m}{(p-1)(q-1)} \sum_{i,j} \gamma_{ij}^2 \dots \quad (19.27)$$

and $E(MSE) = \sigma_0^2.$... (19.28)

If H_{03} is true, $E(MS(AB)) = E(MSE)$, and hence

$$F = \frac{MS(AB)}{MSE}$$

will give the test for H_{03} . This F follows the F distribution with $(p-1)(q-1)$, $pq(m-1)$ d.f. when H_{03} is true. Thus H_{03} is rejected at the $100\alpha\%$ level if (and only if)

$$F = \frac{MS(AB)}{MSE} > F_{\alpha; (p-1)(q-1), pq(m-1)}.$$

If H_{04} is rejected, the tests for H_{01} and H_{02} are not worth making, for if a particular level of A is found to be the best and if interaction is present, then there is no knowing that this will be the best for each level of B . And when H_{01} is true and interaction is present, then in the presence of a particular level of B the effects of the levels of A will differ. Similarly for the factor B . So, in the case of presence of interaction, it is reasonable to test whether the levels of A differ significantly in presence of a particular level of B . This is done by making an analysis of variance for the one-way classified data obtained by taking a particular level of B , but all levels of A .

Similarly, we test for the levels of B in the presence of a particular level of A .

If H_{03} is accepted, the tests for H_{01} and H_{02} can be performed as follows :

H_{01} is rejected at the $100\alpha\%$ level if

$$F = \frac{MSA}{MSE} > F_{\alpha; (p-1), pq(m-1)}.$$

Similarly, H_{02} is rejected at the $100\alpha\%$ level if

$$F = \frac{MSB}{MSE} > F_{\alpha; (q-1), pq(m-1)}.$$

These calculations are shown in the following analysis of variance table :

TABLE 19.8
ANALYSIS OF VARIANCE FOR TWO-WAY CLASSIFIED DATA
WITH m OBSERVATIONS PER CELL

Source of variation	d.f.	SS	MS	F	F at level 1% 5%
Between the levels of A	$p-1$	$mq \sum_i (y_{i00} - \bar{y}_{000})^2$ $= SSA$	$SSA/(p-1)$ $= MSA$	$\frac{MSA}{MSE}$	
Between the levels of B	$q-1$	$mp \sum_j (y_{0j0} - \bar{y}_{000})^2$ $= SSB$	$SSB/(q-1)$ $= MSB$	$\frac{MSB}{MSE}$	
Interaction AB	$(p-1)(q-1)$	$m \sum_{i,j} (y_{ij0} - \bar{y}_{000} - \bar{y}_{0j0} + \bar{y}_{i00})^2 = SS(AB)$	$\frac{SS(AB)}{(p-1)(q-1)}$ $= MS(AB)$	$\frac{MS(AB)}{MSE}$	
Error	$pq(m-1)$	$\sum_{i,j,k} (y_{ijk} - \bar{y}_{000})^2$ $= SSE$	$SSE/pq(m-1)$ $= MSE$		
Total	$mpq-1$	$\sum_{i,j,k} (y_{ijk} - \bar{y}_{000})^2$		—	

If the interaction effects are not significant, we can find the best A -level and the best B -level with the help of t -tests. On the other hand, if they are found to be significant, there will not be a single A -level or a single B -level that will be the best in all situations. In this case, one will have to compare for each level of B the different levels of A and for each level of A the different levels of B .

In the random model, we assume

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk}, \quad \dots \quad (19.29)$$

where a_i 's, b_j 's, c_{ij} 's and e_{ijk} 's are independently normal, with zero means and respective variances σ_a^2 , σ_b^2 , σ_{ab}^2 and σ_e^2 . Now our hypotheses are $H_{01}: \sigma_a^2 = 0$, $H_{02}: \sigma_b^2 = 0$ and $H_{03}: \sigma_{ab}^2 = 0$. The partitioning of total SS and total d.f. is the same as in the fixed model. The $E(MS)$'s are now

$$E(MSA) = \sigma_a^2 + m\sigma_{ab}^2 + mq\sigma_e^2, \quad \dots \quad (19.30)$$

$$E(MSB) = \sigma_b^2 + m\sigma_{ab}^2 + mp\sigma_e^2, \quad \dots \quad (19.31)$$

$$E(MS(AB)) = \sigma_{ab}^2 + m\sigma_{ab}^2, \quad \dots \quad (19.32)$$

$$E(MSE) = \sigma_e^2. \quad \dots \quad (19.33)$$

In this case, H_{01} will be rejected at the $100\alpha\%$ level if (and only if)

$$F = \frac{MSA}{MS(AB)} > F_{\alpha; (p-1), (p-1)(q-1)},$$

H_{02} will be rejected at the $100\alpha\%$ level if

$$F = \frac{MSB}{MS(AB)} > F_{\alpha; (q-1), (p-1)(q-1)},$$

and H_{03} will be rejected at the $100\alpha\%$ level if

$$F = \frac{MS(AB)}{MSE} > F_{\alpha; (p-1)(q-1), pq(m-1)}.$$

In each of the above cases, the F statistic follows the F distribution with appropriate *d.f.* under the null hypothesis.

Here also the test for H_{01} and H_{02} will be performed only if H_{03} is accepted.

It is thus seen that while in the fixed-effects model the same error variance is used for all the tests, the random-effects model gives rise to two error variances, of which one, MSE , is used for H_{03} while the other, $MS(AB)$, is used for both H_{01} and H_{02} .

By equating the observed MS s to their expectations, we obtain as point estimates of the components of variance the following quantities :

$$\hat{\sigma}_e^2 = MSE, \quad \dots \quad (19.34)$$

$$\hat{\sigma}_{AB}^2 = \frac{MS(AB) - MSE}{m}, \quad \dots \quad (19.35)$$

$$\hat{\sigma}_A^2 = \frac{MSA - MS(AB)}{mq} \quad \dots \quad (19.36)$$

and $\hat{\sigma}_B^2 = \frac{MSB - MS(AB)}{mp}. \quad \dots \quad (19.37)$

Here again, if any of the estimates for $\hat{\sigma}_e^2$, $\hat{\sigma}_A^2$ or $\hat{\sigma}_B^2$ turns out negative according to these formulæ, then it should be taken equal to zero.

Lastly, let us consider the case of the mixed model. Of the two factors, let us assume that A refers to the fixed effects and B to the random effects. Then β_j 's and γ_{ij} 's will be random variables, while α_i 's will be fixed parameters. We shall assume that

$$y_{ijk} = m_{ij} + \epsilon_{ijk},$$

where the errors ϵ_{ijk} 's are independently distributed with zero means and variance σ_e^2 and the m_{ij} 's are statistically independent of the ϵ_{ijk} 's.

Subdividing the cell means m_{ij} 's, we get the following model equation :

$$y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + e_{ijk}, \quad \dots \quad (19.38)$$

with $\sum_i \alpha_i = 0, \quad \sum_i c_{ij} = 0 \text{ for all } j.$

The random variables b_j 's and c_{ij} 's have zero means, but they are not independent.

We define

$$\sigma_a^2 = \frac{1}{p-1} \sum_i \alpha_i^2, \quad \sigma_b^2 = \text{var}(b_j)$$

and $\sigma_{ab}^2 = \frac{1}{p-1} \sum_i \text{var}(c_{ij}).$

Then under the mixed model, we have

$$\left. \begin{aligned} E(MSA) &= \sigma_a^2 + m\sigma_b^2 + mq\sigma_e^2, \\ E(MSB) &= \sigma_a^2 + mp\sigma_b^2, \\ E(MS(AB)) &= \sigma_a^2 + m\sigma_{ab}^2, \\ E(MSE) &= \sigma_e^2. \end{aligned} \right\} \quad \dots \quad (19.39)$$

These expectations lead to the following unbiased estimates ($m > 1$) :

$$\hat{\sigma}_b^2 = \frac{MSB - MSE}{mp},$$

$$\hat{\sigma}_{ab}^2 = \frac{MS(AB) - MSE}{m}$$

and $\hat{\sigma}_e^2 = MSE,$

while μ and α_i 's are estimated as in the fixed-effects model.

We assume that the b_j 's, c_{ij} 's and e_{ijk} 's are jointly normal ; this assumption is needed for performing the following tests of hypotheses :

The hypothesis $H_{03} : \sigma_{ab}^2 = 0$, relating to the absence of interaction, may be tested with the statistic $MS(AB)/MSE$, and it will be rejected if

$$F = \frac{MS(AB)}{MSE} > F_{a-1, (p-1)(q-1), pq(m-1)}.$$

If H_{03} is not rejected, we test for H_{02} and H_{01} .

The hypothesis $H_{02} : \sigma_b^2 = 0$, regarding the equality of the effects due to the levels of the random factor B , is rejected if

$$F = \frac{MSB}{MSE} > F_{a-1, (q-1), pq(m-1)}.$$

Under the hypothesis H_{01} : all $\alpha_i=0$, MSA and $MS(AB)$ are statistically independent and have the same expectation. But $\frac{MSA}{MS(AB)}$ will not have, in general, the F distribution. An approximate F -test, with $(p-1)$ and $(p-1)(q-1)$ d.f., may be performed for H_{01} with the ratio $MSA/MS(AB)$. Scheffé suggests an exact test in this case based on Hotelling's T^2 -statistic. However, the approximate F -test for H_{01} will be exact if we make one further simplifying assumption about the variances and covariances of m_{ij} 's, viz. that all variances of m_{ij} are equal and all covariances of m_{ij}, m_{rj} are also equal.

Thus in the mixed model also, we have two error variances, one of which, $MS(AB)$, is used for testing the hypothesis about the fixed-effects factor A , while the other, MSE , is used for testing the hypotheses about the random-effects factor B and the interaction AB .

We present below the $E(MS)$'s under different models for two-way classified data with $m(>1)$ observations per cell.

TABLE 19.9
E(MS)'S UNDER DIFFERENT MODELS FOR TWO-WAY
CLASSIFIED DATA WITH m OBSERVATIONS PER CELL

Source of variation	d.f.	MS	E(MS)		
			Model I	Model II	Mixed model
Between the levels of A	$p-1$	MSA	$\sigma_e^2 + mq \frac{\sum \alpha_i^2}{p-1}$	$\sigma_e^2 + m\sigma_{ab}^2 + mq\sigma_e^2$	$\sigma_e^2 + mq\sigma_e^2$
Between the levels of B	$q-1$	MSB	$\sigma_e^2 + mp \frac{\sum \beta_j^2}{q-1}$	$\sigma_e^2 + m\sigma_{ab}^2 + mp\sigma_e^2$	$\sigma_e^2 + mp\sigma_e^2$
Interaction AB	$(p-1)(q-1)$	$MS(AB)$	$\sigma_e^2 + \frac{m \sum \gamma_{ij}^2}{(p-1)(q-1)}$	$\sigma_e^2 + m\sigma_{ab}^2$	$\sigma_e^2 + m\sigma_{ab}^2$
Error	$pq(m-1)$	MSE	σ_e^2	σ_e^2	σ_e^2
Total	$pq-1$	—	—	—	—

(Under the mixed model, factor A refers to the fixed effects and σ^2 's are as defined above for the mixed model.)

It is seen from the above table that if the interaction effects be absent, then $E(MS(AB))=E(MSE)$ under all the three models. Hence there are some who advocate that when the hypothesis of no interaction effects is accepted, the interaction and error lines be pooled together to form a new error. And this new (pooled) error is used to test for the main effects. But according to others, this is a questionable practice. According to this school, the pooling of the interaction SS with the error SS will be justified only if the interaction is *known* to be absent, and in that case the interaction component is not to be included in the model. According to them, if the interaction is wrongly assumed to be zero, then it will tend to swell the expectation of the pooled mean square.

Computational procedure for the analysis of two-way classified data with m observations per cell :

- (1) Calculate total for each of the pq cells of the table :

$$T_{ijk} = \sum_k y_{ijk}, \quad i = 1, 2, \dots, p; j = 1, 2, \dots, q.$$

- (2) Calculate total for each of the p A-classes :

$$T_{i00} = \sum_j T_{ijk}, \quad i = 1, 2, \dots, p.$$

- (3) Calculate total for each of the q B-classes :

$$T_{0j0} = \sum_i T_{ijk}, \quad j = 1, 2, \dots, q.$$

- (4) Calculate the grand total :

$$T_{000} = \sum_i T_{i00} = \sum_j T_{0j0} = \sum_i \sum_j T_{ijk}.$$

- (5) Calculate row total SS : $\sum_i \sum_j \sum_k y_{ijk}^2$.

- (6) Calculate $\frac{T_{000}^2}{mpq}$.

- (7) Calculate $\frac{\sum_i T_{i00}^2}{mq}$.

- (8) Calculate $\frac{\sum_j T_{0j0}^2}{mp}$.

- (9) Calculate $\frac{\sum_i \sum_j T_{ijk}^2}{m}$.

$$(10) \quad \text{Total } SS = \sum_i \sum_j \sum_k y_{ijk}^2 - \frac{T_{000}^2}{mpq} = (5) - (6).$$

$$(11) \quad SSA = \frac{\sum T_{i00}^2}{mq} - \frac{T_{000}^2}{mpq} = (7) - (6).$$

$$(12) \quad SSB = \frac{\sum T_{0j0}^2}{mp} - \frac{T_{000}^2}{mpq} = (8) - (6).$$

$$(13) \quad SS(AB) = \frac{\sum \sum T_{ij0}^2}{m} - \frac{\sum T_{i00}^2}{mq} - \frac{\sum T_{0j0}^2}{mp} + \frac{T_{000}^2}{mpq}$$

$$= \left(\frac{\sum \sum T_{ij0}^2}{m} - \frac{T_{000}^2}{mpq} \right) - SSA - SSB$$

$$= (9) - (6) - (11) - (12).$$

$$(14) \quad SSE = \text{total } SS - SSA - SSB - SS(AB)$$

$$= (10) - (11) - (12) - (13).$$

Ex. 19.2 An experiment was conducted to determine the effects of five different varieties of cowpeas (V_1, V_2, \dots, V_5) and three different spacings, viz. 4", 8" and 12" (S_1, S_2 and S_3) apart in a row, with rows 3' apart, and also to see whether the varieties behave differently at different spacings. The data below give the yield of each of 4 plots taken for each variety-spacing combination :

Variety	Spacing											
	S_1				S_2				S_3			
V_1	56	45	43	46	60	50	45	48	66	57	50	50
V_2	61	58	55	56	60	59	54	54	59	55	51	52
V_3	63	53	49	48	65	56	50	50	66	58	52	55
V_4	65	61	60	63	60	58	56	60	59	53	48	55
V_5	60	61	50	53	62	68	67	60	73	77	77	65

Carry out an analysis of variance for the above data.

Let y_{ijk} denote the yield of the k th plot for the i th variety at the j th spacing ($i=1, 2, 3, 4, 5$; $j=1, 2, 3$; $k=1, 2, 3, 4$).

The sub-totals for the five varieties, the three spacings and the fifteen variety-spacing combinations, and the grand total are shown below :

Varieties	S_1	Spacings S_2	S_3	Total
V_1	190	203	223	616
V_2	230	227	217	674
V_3	213	221	231	665
V_4	249	234	209	692
V_5	224	257	292	773
Total	1,106	1,142	1,172	3,420

In this case,

$$\sum \sum \sum y_{ijk}^2 = 198,184,$$

$$\text{correction factor} = \frac{T_{000}^2}{mpq} = \frac{(3,420)^2}{60} = 194,940,$$

$$\frac{\sum T_{i00}^2}{mq} = \frac{(616)^2 + (674)^2 + \dots + (773)^2}{12} = 196,029.1667,$$

$$\frac{\sum T_{0j0}^2}{mp} = \frac{(1,106)^2 + (1,142)^2 + (1,172)^2}{20} = 195,049.2,$$

$$\text{and } \frac{\sum \sum T_{ij0}^2}{m} = \frac{(190)^2 + (203)^2 + \dots + (292)^2}{4} = 197,013.5.$$

$$\text{Therefore, } \text{total } SS = \sum \sum \sum y_{ijk}^2 - \frac{T_{000}^2}{mpq}$$

$$= 198,184 - 194,940 = 3,244,$$

$$\begin{aligned} SS \text{ due to varieties} &= \frac{\sum T_{i00}^2}{mq} - \frac{T_{000}^2}{mpq} \\ &= 196,029.1667 - 194,940 = 1,089.1667, \end{aligned}$$

$$\begin{aligned} SS \text{ due to spacings} &= \frac{\sum T_{0j0}^2}{mp} - \frac{T_{000}^2}{mpq} \\ &= 195,049.2 - 194,940 = 109.2, \end{aligned}$$

SS due to variety \times spacing interaction

$$\begin{aligned} &= \left(\frac{\sum \sum T_{ijn}^2}{m} - \frac{T_{000}^2}{mpq} \right) - SS \text{ due to varieties} - SS \text{ due to spacings} \\ &= (197,013.5 - 194,940) - 1,089.1667 - 109.2 = 875.1833, \\ \text{and } SSE &= \text{total } SS - SS \text{ due to varieties} - SS \text{ due to spacings} \\ &\quad - SS \text{ due to variety } \times \text{ spacing interaction} \\ &= 3,244 - 1,089.1667 - 109.2 - 875.1833 = 1,170.4500. \end{aligned}$$

Assuming that the fixed-effects model holds in the present case we shall test the mean squares for varieties, spacings and interaction, against the error mean square.

TABLE 19.10
ANALYSIS OF VARIANCE OF DATA ON YIELD OF COW-PEAS

Source of variation	d.f.	SS	MS	F	F at level 1% 5%
Due to varieties	4	1,089.1667	272.292	10.469	3.79 2.59
Due to spacings	2	109.2	54.6	2.099	5.13 3.21
Due to interaction	8	875.1833	109.398	4.206	2.95 2.16
Error	45	1,170.4500	26.010		
Total	59	3,244		—	

Thus the observed *F* for variety-spacing interaction is significant at both the 1% and the 5% levels. Hence we do not test for the effects of varieties and spacings. (This would also be the case with random and mixed models.) If they are to be tested, then each should be tested at each level of the other.

Under the random model, the estimates for the variance-components are :

$$\delta^2_{\text{varieties}} = \frac{272.292 - 109.398}{12} = \frac{162.894}{12} = 13.574;$$

$$\delta^2_{\text{spacing}} = \frac{54.6 - 109.398}{20}, \text{ i.e. } 0;$$

$$\delta^2_{\text{interaction}} = \frac{109.398 - 26.010}{4} = \frac{83.388}{4} = 20.847.$$

19.8 Application of the technique of analysis of variance in the study of relationship

In the analysis of variance discussed so far, the a_{ij} 's in (19.2) are the values of "indicator variables" and are usually 1 or 0, according as the effect τ_j occurs in the i th observation or not. If the a_{ij} 's are values taken not by indicator variables but by independent variables (in which case y_i 's are called dependent variables), then we have a problem in *regression analysis*. If there be a_{ij} 's of both types, i.e. both indicator variables and independent variables, then we have a problem in *analysis of covariance*. The technique of the analysis of variance is also applicable to these problems. In this section we shall consider some regression problems, and the analysis of covariance problems will be treated in the next chapter.

We now consider the systematic procedure for testing the relationship between two variables. Suppose, corresponding to each level of the independent variable x , which is assumed to be non-stochastic, we have some observations on the dependent variable y as follows :

x_1	x_2	...	x_p
y_{11}	y_{21}	...	y_{p1}
y_{12}	y_{22}	...	y_{p2}
\vdots	\vdots	\vdots	\vdots
y_{1n_1}	y_{2n_2}	...	y_{pn_p}

Any column is an y -array for fixed x . The first question to be asked about the data is : do the available observations provide any evidence that the two variables x and y are related in their movements ? To answer this question we assume that

$$y_{ij} = \mu_i + \epsilon_i,$$

where μ_i 's are the column effects and ϵ_i 's are independently normally distributed, each with mean zero and variance σ^2_ϵ . If the values of y do not depend on the values of x , then we expect $\mu_1 = \mu_2 = \dots = \mu_p$, which is the null hypothesis for testing the absence of relationship.

This case is the same as that of one-way classified data (fixed-effects model), which has been discussed in Section 19.5. So we can write down the analysis of variance table as follows :

TABLE 19.11
ANALYSIS OF VARIANCE FOR TESTING THE RELATIONSHIP
BETWEEN TWO VARIABLES

Source of variation	d.f.	SS	MS	F
Between arrays	$p-1$	$\sum_i n_i (y_{i0} - \bar{y}_{00})^2 = SSB$	$SSB/(p-1) = MSB$	$\frac{MSB}{MSW}$
Within arrays	$n-p$	$\sum_i \sum_j (y_{ij} - \bar{y}_{i0})^2 = SSW$	$SSW/(n-p) = MSW$	
Total	$n-1$	$\sum_i \sum_j (y_{ij} - \bar{y}_{00})^2$		—

The null hypothesis of absence of any relationship between x and y will be rejected at the level α if the value of F obtained in the above table exceeds $F_{\alpha; (p-1), (n-p)}$. In the above test, we have made no assumption about the form of relationship—whether linear or non-linear. We have tested for the presence of *any* relationship and the rejection of the null hypothesis would suggest that there is some relationship.

After the relationship is established, the next step will be to find the appropriate regression function. And at first we try to find out whether the simplest function, linear regression, fits the observed data. So the null hypothesis is now $H_0 : \mu_i = \alpha + \beta x_i$, with the same observational equations as in the previous case :

$$y_{ij} = \mu_i + e_{ij},$$

α and β being parameters.

We shall make use of the theorem of Section 19.3 for testing H_0 .

$$\begin{aligned} S_1^2 &= \min \sum_i \sum_j (y_{ij} - \mu_i)^2, \text{ when minimised with respect to } \mu_i \text{'s} \\ &= \sum_i \sum_j (y_{ij} - \bar{y}_{i0})^2, \text{ and it has } (n-p) \text{ d.f.} \end{aligned}$$

$$\begin{aligned} S_2^2 &= \min \sum_i \sum_j (y_{ij} - \mu_i)^2, \text{ when minimised w.r.t. } \mu_i \text{'s subject to} \\ &\quad \text{the conditions } \mu_i = \alpha + \beta x_i \\ &= \min \sum_i \sum_j (y_{ij} - \alpha - \beta x_i)^2, \text{ when minimised w.r.t. } \alpha \text{ and } \beta \\ &= \sum_i \sum_j [y_{ij} - \bar{y}_{00} - b(x_i - \bar{x})]^2, \text{ and it has } (n-2) \text{ d.f.} \end{aligned}$$

The least-square estimates are :

$$\hat{a} = y_{00} - b\bar{x}, \quad \hat{b} = b = \frac{\sum_i \sum_j (y_{ij} - y_{00})(x_i - \bar{x})}{\sum_i n_i (x_i - \bar{x})^2}, \quad \dots \quad (19.40)$$

where

$$\bar{x} = \frac{\sum_i n_i x_i}{n}.$$

Now,

$$S_2^2 = \sum_i \sum_j (y_{ij} - y_{00})^2 - b^2 \sum_i n_i (x_i - \bar{x})^2$$

is, under H_0 , a $\sigma^2 \chi^2$ with $(n-2)$ d.f., and

$$S_1^2 = \sum_i \sum_j (y_{ij} - y_{i0})^2$$

is a $\sigma^2 \chi^2$ with $(p-2)$ d.f.

$$\begin{aligned} \text{Then } S_2^2 - S_1^2 &= \sum_i \sum_j (y_{ij} - y_{00})^2 - \sum_i \sum_j (y_{ij} - y_{i0})^2 - b^2 \sum_i n_i (x_i - \bar{x})^2 \\ &= \sum_i n_i (y_{i0} - y_{00})^2 - b^2 \sum_i n_i (x_i - \bar{x})^2 \end{aligned}$$

is a $\sigma^2 \chi^2$ with $(p-2)$ d.f. under H_0 , and

$$\begin{aligned} F &= \frac{S_2^2 - S_1^2}{p-2} / \frac{S_1^2}{n-p} \\ &= \frac{\sum_i n_i (y_{i0} - y_{00})^2 - b^2 \sum_i n_i (x_i - \bar{x})^2}{\sum_i \sum_j (y_{ij} - y_{i0})^2} \cdot \frac{n-p}{p-2} \end{aligned} .$$

is an F with $(p-2), (n-p)$ d.f. under H_0 and is used to test H_0 .

The hypothesis of linearity of regression is rejected at the level α if for our data the F as obtained above exceeds $F_{\alpha; (p-2), (n-p)}$.

We can see the entire picture in this case if we partition the total SS as follows :

$$\begin{aligned} \sum_i \sum_j (y_{ij} - y_{00})^2 &= [\sum_i \sum_j (y_{ij} - y_{i0})^2] + [\sum_i n_i (y_{i0} - y_{00})^2 - b^2 \sum_i n_i (x_i - \bar{x})^2] \\ &\quad + [b^2 \sum_i n_i (x_i - \bar{x})^2], \end{aligned}$$

or total $SS = SSW + SS(DLR) + SS(LR)$, $\dots \quad (19.41)$

where $SS(DLR) = SS$ due to deviation from linear regression

and $SS(LR) = SS$ due to linear regression.

$$\begin{aligned} \text{Then } F &= \frac{S_2^2 - S_1^2}{S_1^2} \cdot \frac{n-p}{p-2} = \frac{SS(DLR)}{SS} \cdot \frac{n-p}{p-2} \\ &= \frac{MS(DLR)}{MSW} \end{aligned}$$

is the statistic used to test for the linearity of regression. If the MS

due to deviation from linear regression is significantly different from (i.e. greater than) the MS due to error, then we have reason to doubt the linearity of regression.

If we denote $\alpha + \beta x_i$ by Y_i , where α and β are the least-square estimates of α and β , then it is easy to verify that

$$SS(LR) = b^2 \sum_i n_i (x_i - \bar{x})^2$$

$$= \sum_i n_i (Y_i - y_{00})^2$$

and

$$SS(DLR) = \sum_i n_i (y_{10} - y_{00})^2 - b^2 \sum_i n_i (x_i - \bar{x})^2$$

$$= \sum_i n_i (y_{10} - Y_i)^2.$$

The analysis of variance table for testing the linearity of regression is given below :

TABLE 19.12

ANALYSIS OF VARIANCE FOR TESTING LINEARITY OF REGRESSION

Source of variation	d.f.	SS	MS	F
Due to linear regression	1	$\sum_i n_i (Y_i - y_{00})^2 = SS(LR)$	$MS(LR)$	
Due to deviation from linear regression	$p-2$	$\sum_i n_i (y_{10} - Y_i)^2 = SS(DLR)$	$MS(DLR)$	$F = \frac{MS(DLR)}{MSW}$
Between arrays	$p-1$	$\sum_i n_i (y_{10} - y_{00})^2 = SSB$	MSB	
Within arrays	$n-p$	$\sum_i \sum_j (y_{ij} - y_{10})^2 = SSW$	MSW	
Total	$n-1$	$\sum_i \sum_j (y_{ij} - y_{00})^2$	—	—

If the F at the previous stage is significant, then the hypothesis $H_0 : \mu_i = \alpha + \beta x_i$ fails to account for the relationship between x and y . We may then try various hypotheses regarding the form of the relationship. In this way, we may examine successively whether a polynomial in x of degree k , $P_k(x)$, will be able to explain the relationship, for $k=2, 3$, etc., but $k < (p-1)$, where p is the number of y -arrays.

*This is a sub-total line.

We give below the test procedure for testing the null hypothesis

$$H_0 : \mu_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k,$$

that the relationship between x and y can be explained by a polynomial of degree k .

Let $\hat{\alpha}$ and $\hat{\beta}_j$, for $j=1, 2, \dots, k$, be the least-square estimates of α and β_j 's. Also, let

$$\hat{P}_k(x_i) = \hat{\alpha} + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \dots + \hat{\beta}_k x_i^k.$$

Then the total SS can be partitioned as follows :

$$\begin{aligned} \sum_{i,j} (y_{ij} - y_{00})^2 &= \sum_i \sum_j [y_{ij} - y_{i0} + y_{i0} - \hat{P}_k(x_i) + \hat{P}_k(x_i) - y_{00}]^2 \\ &= \sum_i \sum_j (y_{ij} - y_{i0})^2 + \sum_i n_i [y_{i0} - \hat{P}_k(x_i)]^2 + \sum_i n_i [\hat{P}_k(x_i) - y_{00}]^2 \\ &= SSW + SS(DR) + SSR, \text{ say,} \quad \dots \quad (19.42) \end{aligned}$$

where SSR is the SS due to polynomial regression of degree k and $SS(DR)$ i.e. the SS due to deviation from regression. We have an analysis of variance similar to that of Table 19.12.

TABLE 19.13
ANALYSIS OF VARIANCE FOR TESTING

$$H_0 : \mu_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k$$

Source of variation	d.f.	SS	MS	F
Due to regression $P_k(x)$	k	$\sum_i n_i [\hat{P}_k(x_i) - y_{00}]^2 = SSR$	MSR	
Due to deviation from regression	$p-k-1$	$\sum_i n_i [y_{i0} - \hat{P}_k(x_i)]^2 = SS(DR)$	$MS(DR)$	$F = \frac{MS(DR)}{MSW}$
Between arrays	$p-1$	$\sum_i n_i (y_{i0} - y_{00})^2 = SSB$	MSB	
Within arrays	$n-p$	$\sum_i \sum_j (y_{ij} - y_{i0})^2 = SSW$	MSW	
Total	$n-1$	$\sum_i \sum_j (y_{ij} - y_{00})^2$		—

The test for $H_0 : \mu_i = \alpha + \beta_1 x_i + \dots + \beta_k x_i^k$ is given by

$$F = \frac{MS(DR)}{MSW}, \text{ which is an } F \text{ with } (p-k-1), (n-p) \text{ d.f.}$$

H_0 is rejected at the level α if $F > F_{\alpha ; (p-k-1), (n-p)}$.

Tests for multiple linear regression model

Let us suppose that we have a set of k independent variables, x_1, x_2, \dots, x_k , and the dependent variable y . As the model, we take

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + e_i, \quad \dots \quad (19.43)$$

$$(i=1, 2, \dots, n),$$

where e_i 's are independently normal, each with mean 0 and variance σ_e^2 . We are interested in the null hypothesis H_0 : all $\beta_j = 0$, which means that there is no dependence of y on the k fixed variates x_1, x_2, \dots, x_k . For simplifying the determination of the least-square estimates of the constants α and β_j 's, we can write the above model equation in the form

$$y_i = \alpha' + \beta_1 x'_{1i} + \beta_2 x'_{2i} + \dots + \beta_k x'_{ki} + e_i, \quad \dots \quad (19.43a)$$

where $x'_{ji} = x_{ji} - \bar{x}_j$.

Now, by an application of the theorem of Section 19.3, we have

$$S_1^2 = \min \sum_i (y_i - \alpha' - \beta_1 x'_{1i} - \dots - \beta_k x'_{ki})^2 \text{ w.r.t. } \alpha' \text{ and } \beta_j's$$

$$= \sum_i (y_i - \bar{y})^2 - \sum_{j=1}^k b_j P_j, \quad \text{which has } (n-k-1) \text{ d.f.,}$$

where $P_j = \sum_{i=1}^n (x_{ji} - \bar{x}_j) y_i$ and b_j is the least-square estimate of β_j ,

and $S_2^2 = \min \sum_i (y_i - \alpha' - \beta_1 x'_{1i} - \dots - \beta_k x'_{ki})^2$, when minimised
w.r.t. α' and β_j 's, subject to the condition H_0
 $= \sum_i (y_i - \bar{y})^2, \quad \text{which has } (n-1) \text{ d.f.}$

Thus

$$S_2^2 - S_1^2 = \sum_{j=1}^k b_j P_j \quad (\text{having } k \text{ d.f.})$$

$$= SS \text{ due to multiple linear regression}$$

$$= SSR, \text{ say,}$$

and $S_1^2 = SS \text{ due to error}$
 $= SSE, \text{ say.}$

Hence if for our data

$$F = \frac{SSR/k}{SSE/(n-k-1)} > F_{\alpha; k, (n-k-1)},$$

we reject H_0 at the level α .

TABLE 19.14
**A ANALYSIS OF VARIANCE FOR TESTING FOR THE MULTIPLE
 LINEAR REGRESSION**

Source of variation	d.f.	SS	MS	F
Due to multiple linear regression	k	$\sum_j b_j P_j$	MSR	$F = \frac{MSR}{MSE}$
Error	$n-k-1$	$\sum_i (y_i - \bar{y})^2 - \sum_j b_j P_j$	MSE	
Total	$n-1$	$\sum_i (y_i - \bar{y})^2$		—

A slightly different problem in connection with multiple linear regression is the following :

Given a set of $(p+q)$ independent variables, one may want to know whether a particular group of q independent variables has any effect on the prediction of the dependent variable, after already having fitted the other p independent variables. And, without any loss of generality, the former group of q variables may be taken to be the last q of the $(p+q)$ variables. Then the hypothesis is

$$H_0 : \beta_{p+1} = \beta_{p+2} = \dots = \beta_{p+q} = 0,$$

the linear model being

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \beta_{p+1} x_{p+1,i} + \dots + \beta_{p+q} x_{p+q,i} + e_i, \quad \dots \quad (19.44)$$

where e_i 's are independently normally distributed, each with mean 0 and variance σ_e^2 .

Under H_0 , the above model reduces to

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i. \quad \dots \quad (19.45)$$

Also,

$$S_1^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{j=1}^{p+q} b_j P_j \quad \text{has } (n-p-q-1) \text{ d.f.},$$

where $P_j = \sum_{i=1}^n (x_{ji} - \bar{x}_j) y_i$,

and $S_2^2 = \sum (y_i - \bar{y})^2 - \sum_{j=1}^p b_j^* P_j \quad \text{has } (n-p-1) \text{ d.f.}$

Then

$$S_e^2 - S_1^2 = \sum_{j=1}^{p+q} b_j P_j - \sum_{j=1}^p b_j^* P_j \text{ has } q \text{ d.f.}$$

Here b_j 's are the least-square estimates of β_j 's for the model (19.44), and b_j^* 's are the least-square estimates of β_j 's for the restricted model (19.45).

Hence H_0 is rejected at the level α if for the data

$$F = \frac{\frac{\sum_{j=1}^{p+q} b_j P_j - \sum_{j=1}^p b_j^* P_j}{n-p-q-1}}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{j=1}^{p+q} b_j P_j}{q}} > F_{\alpha; q, (n-p-q-1)}$$

The analysis of variance table is as follows :

TABLE 19.15

ANALYSIS OF VARIANCE FOR TESTING THE EFFECT OF
INTRODUCTION OF NEW VARIABLES IN THE REGRESSION EQUATION

Source of variation	d.f.	SS	MS	F
Due to multiple linear regression of y on x_1, x_2, \dots, x_p	p	$\sum_{j=1}^p b_j^* P_j$	MSR_p	
Due to multiple linear regression of y on x_{p+1}, \dots, x_{p+q} , after fitting x_1, x_2, \dots, x_p	q	$\sum_{j=1}^{p+q} b_j P_j - \sum_{j=1}^p b_j^* P_j$	$MSR_{q/p}$	$F = \frac{MSR_{q/p}}{MSE}$
Due to multiple linear regression of y on x_1, x_2, \dots, x_{p+q}	$p+q$	$\sum_{j=1}^{p+q} b_j P_j$	MSR_{p+q}	
Error	$n-p-q-1$	$\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{j=1}^{p+q} b_j P_j$	MSE	
Total	$n-1$	$\sum_{i=1}^n (y_i - \bar{y})^2$		—

The hypothesis that will be considered next under the multiple linear regression model is the following :

$$H_0 : \beta_j = c,$$

for any particular j , c being any given value, including 0.

For this, we refer to the normal equations for estimating the β_j 's, which can be written in the matrix notation as

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{P}, \quad \dots \quad (19.46)$$

where $\mathbf{A} = \left(\sum_{l=1}^n x'_{il} x_{jl} \right)_{k \times k}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$

and $\mathbf{P} = \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_k \end{pmatrix}$ with $P_j = \sum_i x'_{ij} y_i$.

Then

$$\hat{\boldsymbol{\beta}} = \mathbf{A}^{-1} \mathbf{P}, \quad \dots \quad (19.47)$$

which shows that $\hat{\beta}_j$ is a linear function of y_1, y_2, \dots, y_n and hence is normally distributed, with

$$\left. \begin{array}{l} E(\hat{\beta}_j) = \beta_j, \\ \text{var}(\hat{\beta}_j) = c_{jj} \sigma^2 \\ \text{cov}(\hat{\beta}_j, \hat{\beta}_{j'}) = c_{jj'} \sigma^2, \end{array} \right\} \quad \dots \quad (19.48)$$

and

where $(c_{jj'})_{k \times k} = \mathbf{A}^{-1}$.

Thus the statistic for testing $H_0 : \beta_j = c$ is

$$t = \frac{\hat{\beta}_j - c}{\sqrt{c_{jj} MSE}},$$

which is distributed as a t with $(n-k-1)$ d.f., MSE being the error MS of the analysis of variance table.

Next, consider the hypothesis

$$H_0 : \beta_j = \beta_{j'}$$

The test of this hypothesis is also given by the t -statistic with $(n-k-1)$ d.f., where

$$t = \frac{\hat{\beta}_j - \hat{\beta}_{j'}}{\sqrt{(c_{jj} - 2c_{jj'} + c_{j'j'})MSE}}.$$

19.9 Effects of violations of the assumptions made in the analysis of variance

In this section, we shall make certain comments on the effects of violations of the underlying assumptions of (i) normality of the errors and also of the random effects, (ii) independence and (iii) homoscedasticity of the errors.

In Model I, the normality assumption is needed only for hypothesis-testing and interval estimation. Thus all (point) estimates and their estimated variances remain valid even under non-normality. Heteroscedasticity and correlation of errors do not bias the estimators. For other models too, the estimators of variance-components remain unbiased even with non-normal random effects.

Robustness against non-normality of the tests on means and the lack of it in the case of tests on variances lead us to expect that tests and confidence intervals in the case of Model I will be robust to non-normality, while those in other models, which are mainly concerned with variances, will not be robust.

Investigations have shown that the effects of heteroscedasticity, which are large in the case of Model I, can be reduced by using experiments with equal cell-frequencies.

The effects of the stochastic dependence among the errors may completely vitiate the tests. As a remedy, the use of randomisation should be taken into consideration while allocating the treatments to different experimental units.

Transformations of the observations are often used to reduce non-normality or heteroscedasticity.

The study of the robustness of the analysis of variance methods has led to the search for distribution-free methods for analysis of variance. Such distribution-free methods exist and are completely robust for any continuous distribution and compare favourably with the classical normal-theory procedures.

Questions and exercises

19.1 What is a 'linear model'? Clearly bring out the differences among 'fixed', 'mixed' and 'random' models.

19.2 What is meant by the term 'linear hypothesis'? How is such a hypothesis tested?

19.3 Show that for a two-way classified data with one observation per cell and satisfying model (19.11b) the following is true :

$$\sum_{i,j} (y_{ij} - \mu - \alpha_i - \beta_j)^2 = pq(y_{00} - \mu)^2 + q \sum_i (y_{i0} - y_{00} - \alpha_i)^2 + p \sum_j (y_{0j} - y_{00} - \beta_j)^2 + \sum_{i,j} (y_{ij} - y_{i0} - y_{0j} + y_{00})^2.$$

Use the above relation to obtain the least-square estimates of the parameters in (19.11b). Use this to obtain also SSE , SSA and SSB .

19.4 State how the formulation of the model and that of the null hypothesis depend on whether an effect is a fixed or a random effect.

19.5 Discuss the problem of selecting valid error in relation to a two-way layout with $m (> 1)$ observations per cell, under the various models.

19.6 In what respects do analysis of variance, regression analysis and analysis of covariance differ ?

19.7 Use the technique of analysis of variance for testing (i) the linearity of regression and (ii) that a group of q independent variables, out of a totality of $(p+q)$ independent variables, have no effect on the prediction of y .

19.8 How would you interpret an observed F which is less than one ? Why are negative estimates of variance-components replaced by zeroes ?

19.9 Show that for the random model (19.8), the following is a consistent estimate of the intra-class correlation coefficient $\rho = \sigma_b^2 / \sigma_s^2$:

$$\frac{MSA - MSE}{MSA + (\bar{\lambda} - 1)MSE}$$

19.10 What are the assumptions that are made in the analysis of variance ? State how violations of these assumptions affect the analysis.

19.11 Below are given the yields in gm. per plot (plot size = $\frac{1}{100}$ acre) for three varieties of seed cotton :

Variety 1	Variety 2	Variety 3
77	109	46
10	106	70
63	137	71
84	79	65
95	134	61
81	78	40
88	126	47
101	98	73

(a) Write out the analysis of variance table.

- (b) Test if the varieties differ significantly among themselves.
 (c) If the result of (b) is affirmative, determine which varieties differ in the case of fixed model.
 (d) If the result of (b) is affirmative, obtain an estimate of the variability of the varietal effects in the case of random model.

Partial ans. $F=17.11.$

19.12 Information relating to weight at birth (in lb.) of boys at a number of primary schools is given below. Analyse the data.

School	A	B	C	D	E	F
Number of boys	112	69	128	97	62	78
Mean weight per boy	6.132	6.261	6.345	6.112	6.320	5.927
Standard deviation (the divisor used is sample size and not d.f.)	0.763	0.812	0.752	0.733	0.835	0.743

Partial ans. $F=3.06.$

19.13 The determination of visual acuity at three different distances (say A, B and C) was the subject of a recent experiment. Four different subjects chosen at random from a large group were used for this purpose. The data recorded were as follows :

Subject	A	Distance	
		B	C
1	12	16	30
2	5	10	18
3	7	28	33
4	10	26	51

- (a) Analyse the above two-way classified data.
 (b) Test for the effects of subjects and distances at the 5% level of significance.
 (c) Estimate the variability due to subjects.
 (d) Determine which distances differ, if any. Is it possible to do the same with the subjects ?

Partial ans. F (for subjects) = 3.55 ;
 F (for distances) = 12.93.

19.14 The following data show the birth-weights of babies born, classified according to the age of mother and the order of gravida, there being three observations per cell.

**BIRTH-WEIGHTS (in lb.) OF BABIES BORN IN A
NURSING HOME AT HOWRAH**

Age of mother Order of gravida \	15—20	20—25	25—30	30—35	35 and over
1	5·1, 5·0, 4·8	5·0, 5·1, 5·3	5·1, 5·1, 4·9	4·9, 4·9, 5·0	5·0, 5·0, 5·0
2	5·2, 5·2, 5·4	5·3, 5·3, 5·5	5·3, 5·2, 5·2	5·2, 5·0, 5·5	5·1, 5·3, 5·0
3	5·8, 5·7, 5·9	6·0, 5·9, 6·2	5·8, 5·9, 5·9	5·8, 5·5, 5·5	5·9, 5·4, 5·5
4	6·0, 6·0, 5·9	6·2, 6·5, 6·0	6·0, 6·1, 6·0	6·0, 5·8, 5·5	5·8, 5·6, 5·5
5 and over	6·0, 6·0, 6·0	6·0, 6·1, 6·3	5·9, 6·0, 5·8	5·9, 6·0, 5·5	5·5, 6·0, 6·2

Test whether the age of mother and order of gravida significantly affect the birth-weight.

Partial ans. F (for age of mother) = 96·413 ;
 F (for order of gravida) = 9·335.

SUGGESTED READING

- [1] Anderson, R. L. and Bancroft, T. A. *Statistical Theory in Research* (Chs. 13-15, 21-23). McGraw-Hill, 1952.
- [2] Bowker, A. H. and Lieberman, G. J. *Engineering Statistics* (Ch. 10). Asia Publishing House, 1962.
- [3] Goon, A. M., Gupta, M. K. and Dasgupta, B. *An Outline of Statistical Theory*, Volume 2 (Ch. 7). World Press, 1973.
- [4] Goulden, C. H. *Methods of Statistical Analysis* (Ch. 5). Asia Publishing House, 1959.
- [5] Guenther, W. C. *The Analysis of Variance*. Prentice-Hall, 1964.
- [6] Hald, A. *Statistical Theory with Engineering Applications* (Ch. 16). John Wiley, 1952.
- [7] Kendall, M. G. and Stuart, A. *The Advanced Theory of Statistics*, Volume 3 (Chs. 35-37). Charles Griffin, 1966.
- [8] Rao, C. R. *Advanced Statistical Methods in Biometric Research* (Ch. 2, 3). John Wiley, 1952.

- [9] Rao, C. R. *Linear Statistical Inference and Its Applications* (Ch. 4). John Wiley, 1965, and Wiley Eastern.
- [10] Scheffé, H. *The Analysis of Variance* (Chs. 3, 4, 7, 8). John Wiley, 1961.
- [11] Steel, R. G. D. and Torrie, J. H. *Principles and Procedures of Statistics* (Chs. 7-9, 14, 15). McGraw-Hill, 1960.

20

DESIGNS OF EXPERIMENTS

The theoretical aspects of the analysis of variance technique were discussed in Chapter 19. A number of commonly used experimental designs will be considered in this chapter. We first consider the terminology used in experimentation and the basic principles of experimental designs.

20.1 Terminology in experimental designs

Before discussing the principles of designs, it is proper to explain the terminology used in this context. The terms commonly used are *experiment*, *treatment*, *experimental unit*, *experimental error* and *precision*.

Experiment is a means of getting an answer to the question that the experimenter has in mind. This may be to decide which of several pain-relieving drugs that are available in the market is the most effective or whether they are equally effective. An experiment may be planned to compare the Chinese method of cultivation with the standard method used in India. In planning an experiment, we clearly state our objectives and formulate the hypotheses we want to test.

Treatment—The different procedures under comparison in an experiment are the different treatments. E.g., in an agricultural experiment, the different varieties of a crop or the different manures will be the treatments. In a dietary or medical experiment, the different diets or medicines, etc., are the treatments.

Experimental unit—In carrying out an experiment, we should be clear as to what constitutes the experimental unit. An experimental unit is the material to which is applied the treatment and on which the variable under study is measured. In an agricultural field experiment, the plot of land, and not the individual plant, will be the experimental unit ; in a feeding experiment of cows, the whole cow is the experimental unit ; in human experiments in which the treatment affects the individual, the individual will be the experimental unit.

Experimental error—A fundamental phenomenon in replicated experiments is the variation in the measurements made on different

experimental units even when they get the same treatment. A part of this variation is systematic and can be explained, whereas the remainder is to be taken to be of the random type. The unexplained random part of the variation is termed the experimental error. This is a technical term and does not mean a mistake, but includes all types of extraneous variation due to (i) inherent variability in the experimental units, (ii) errors associated with the measurements made and (iii) lack of representativeness of the sample to the population under study.

The experimental error provides a basis for the confidence to be placed in the inference about the population. So it is important to estimate and control the experimental error. An estimate of the experimental error can only be obtained by *replication*, and it is controlled by the principle of *local control*, to be explained shortly.

The *precision* of an experiment is measured by the reciprocal of the variance of a mean :

$$1/\sigma_x^2 = n/\sigma^2.$$

As n , the replication number, increases, precision also increases. Another means of increasing precision is to control σ^2 ; the smaller the value of σ^2 , the greater the precision.

20.2 Principles of design

Designing an experiment means deciding how the observations or measurements should be taken to answer a particular question in a valid, efficient and economical way. The design and the final analysis go together; they are inseparable in the sense that if an experiment is properly designed, then there will exist an appropriate way of analysing the data. From an ill-designed experiment no conclusion can be drawn.

Though most of the recent advances in the efficient design and analysis of experiments arose in an effort to meet the needs of agricultural research, they are also generally applicable to other branches of research. Modern experiments are designed so that we can get the data for verifying the hypotheses in as economical a way as possible. The application of the technique of analysis of variance is appropriate only when the data conform to the basic set-up of the analysis of variance. The analysis of the data will be meaningless if the assumptions in the analysis of variance are not fulfilled. So the layout and the method of analysis are co-ordinated in the design of experiments.

Even now it is not uncommon to encounter a research worker who collects his data in any way he can and then comes to a statistician for help in establishing his conjectures. The desirable course for him would be to consult a statistician before planning the experiment, and thus deciding the manner in which the data should be collected for the specific purpose and the form the analysis would take.

As an extreme example, consider the following experiment for comparing the effectiveness of two different tranquillizers that are available in the market. Tranquillizer X is applied to a group of female patients of hospital A and tranquillizer Y is applied to a group of male patients of hospital B . It is found from the data collected that the average effect for tranquillizer X is superior to that of tranquillizer Y . The hospital authorities may say that this difference reflects the sex-difference, while the druggists may say that this difference is due to differences in the tranquillizers. A statistician will, however, politely say that the effects of the tranquillizers and sex-differences are completely entangled or mixed up and one cannot be separated from the other. If the experimenter insists on a decision, the statistician will have to say that no conclusion can be drawn from this experiment.

The application of designs has reduced, if not completely eliminated, the cases where an experiment is conducted and data collected without first conceiving a method of statistical analysis.

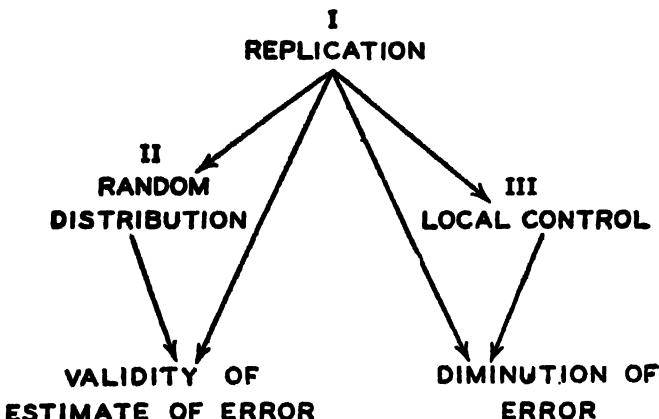


Fig. 20.1 Fisher's diagram.

The three basic principles of experimental design, namely, the indispensability of *replication* and that of *randomisation* and the desirability of *local control*, were developed by R. A. Fisher. Fisher illustrated the function of the principles, from which modern experimental designs have been evolved, in the above diagram (Fig. 20.1).

Randomisation

The principle of randomisation, as advocated by Fisher, is essential for a valid estimate of the experimental error and also to minimise bias in the results. We mentioned that one of the assumptions in the model of the analysis of variance is the independence of errors. If we consider agricultural experiments, it is a fact that soil fertility is not distributed at random and nearby plots happen to be correlated. Randomisation is a simple device to achieve this independence of errors.

In the words of Cochran and Cox, "Randomisation is analogous to insurance in that it is a precaution against disturbances that may or may not occur, and that may or may not be serious if they do occur."

However, randomisation by itself is not sufficient for the validity of the experiment. Consider an experiment for comparing two diets for children and suppose there are only two children available for the experiment. If the two children are different in initial conditions, say in the type of family, initial weight, etc., then even if the two diets be equally effective, the one applied to the superior child will give a better result in spite of random allocation of the diets to the children. So randomisation forms only a basis of a valid experiment. In order to ensure validity, it is necessary to have more than one child of each type and then to make the allocation of diets at random. Thus randomisation plus replication will be necessary for the validity of the experiment.

It must be explicitly understood that separate randomisation for every replication and experiment is necessary.

Replication

The second essential feature of an experiment is replication. A treatment is repeated a number of times in order to obtain a more reliable estimate than is possible from a single observation. In the previous example, if we have more than two children, we can plan

the experiment so that no particular diet is favoured or disfavoured in the experiment, i.e. each diet is applied approximately equally often to all types of experimental units.

Since the error of the experiment arises from the differences between experimental units of the same treatment, that are not due to differences between the replicates, there is no other way but replication to get an estimate of the error of the experiment. It is apparent from Fisher's diagram that the function of replication is two-fold : (a) along with randomisation, it provides an estimate of the error to which comparisons are subjected, and (b) along with local control, it reduces the experimental error.

The most effective way to increase the precision of an experiment is to increase the number of replications. In field experiments, precision can be increased by an increase of plot size. However, it has been found that, for the same amount of land, increased replication of small plots is more effective than using larger plots less frequently. Of course, replication beyond a limit may be impractical. Since

$$\sigma_s = \frac{\sigma}{\sqrt{n}},$$

decrease in σ_s is proportional to the square-root of the number of replications—this is true if the variations due to replicates have been removed from error. The number of replications in a particular case depends on the variability of the material, cost of taking observations, etc. A rule-of-thumb is to get about 10 d.f. for the experimental error ; and generally one should not use less than 4 replications.

Replication broadens the scope of the experiment by including different types of experimental units. Replication in space and time is also necessary in order to sample different soil and climatic conditions.

Local control

The third principle, a desirable one, is called local or error control. As already mentioned, replication is used with local control to reduce the experimental error. In a replicated experiment, the randomisation may be restricted in such a manner that a portion of the total variation may be eliminated from the error, the variation that is irrelevant in making comparisons.

In the simplest case, the experimental units are divided into homogeneous groups or blocks. The variation among these blocks is eliminated from the error and thereby efficiency is increased. We shall see afterwards that the random allocation of treatments to the experimental units may be restricted in different ways in order to control experimental error.

Another means of controlling error is the use of confounded designs when the number of treatment combinations is very large, as in some factorial experiments. The use of one or more auxiliary variables for an analysis of covariance will also reduce experimental error. These we shall discuss in later sections.

The choice of the size and shape of experimental units and of blocks has also some effect on the error of the experiment.

Besides the above three principles, there are some other general principles in designing an experiment. Familiarity with the treatments and experimental material is an asset. Selection of the experimental site should be carefully done. Within-block variability should be reduced.

20.3 Choice of size and shape of plots and blocks

In field experiments, the size and shape of plots as well as those of blocks influence the experimental error. The total available experimental area remaining fixed, an increase in the size of plots will automatically decrease the number of plots and indirectly increase the block size while reducing the number of blocks. In order to reduce the flow of experimental material from one plot to another, it is customary to leave out strips of land between consecutive plots and also between blocks—these non-experimental areas are known as *guard areas*. So, as the number of plots increases, the number of guard areas, and hence the amount of non-experimental area, also increases. This fact should be kept in view while deciding on the size of plots.

An important investigation on the effect of size and shape of plot and block was conducted by H. F. Smith. He conducted uniformity trial experiments with the same crop, and then harvesting the crop in small units, he found that the variance per unit area for plots of area x units was approximately given by $V_s = V_1/x^b$, where b is a

soil characteristic. $b=1$ means that the units making the plot of size x units are not correlated and then $V_s = V_1/x$, so that an increase in plot size increases the precision of the experiment. $b=0$ means that the units of the plot are perfectly correlated and then $V_s = V_1$, so that there is no gain in precision by increasing plot size. Usually, $0 < b < 1$ and an increase in plot size increases the precision of the experiment, provided we use the same number of plots.

Long and narrow plots have been found to be relatively more precise.

The size and shape of a block will ordinarily be determined by the size and shape of plots and the number of plots in a block. It is desirable from the point of view of error control to have small variation among the plots within a block and large variation among the blocks. When definite fertility contours are present, the maximum precision will be obtained by arranging the plots in a block with their long sides parallel to the direction of the fertility gradient and by taking blocks one after another in the direction of the gradient.

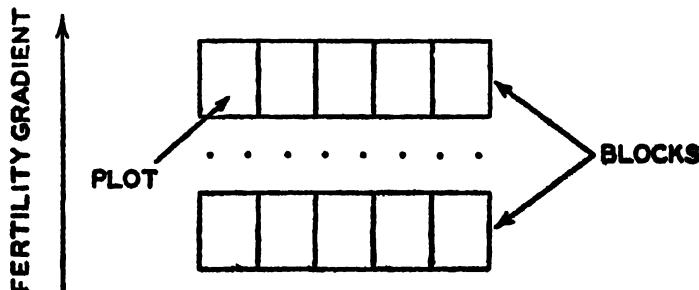


Fig. 20.2 Orientation of plot and block.

In the absence of any knowledge of fertility contours, it is better to use square plots and, generally speaking, it is best to have small blocks ; otherwise, the plots within a block will not be homogeneous.

In the following sections, we shall use the principles stated in the previous section in designing an experiment and then shall use the technique of the analysis of variance for analysing the data. We shall consider only the Model I analysis.

20.4 Completely randomised design (CRD)

The simplest design using the two essential principles of replication and randomisation is the *CRD*. Suppose that we have t treatments (or t levels of a factor) under comparison and the i th treatment is to be replicated r_i times, for $i=1, 2, \dots, t$. Then the total number of experimental units necessary for this experiment is $n = \sum_{i=1}^t r_i$. In the *CRD*, we allocate the t treatments completely at random to the n units subject to the condition that the i th treatment appears in r_i units, for $i=1, 2, \dots, t$. A particular case of this is equal replication for different treatments, where $r_1=r_2=\dots=r_t=r$, so that $n=rt$.

Layout

The term *layout* refers to the placement of treatments to the experimental units according to the conditions of the design.

Randomisation may be carried out by using a random number table. Let us obtain the layout for a *CRD* with three treatments, the number of replications used being 5, 4 and 3, respectively. We number the experimental units, in any convenient way, from 1 to 12 (the total number of experimental units). We then get a random permutation of the experimental units. To the first 5 of the units in the random permutation we apply treatment 1, to the next 4 units treatment 2 is applied, and treatment 3 is applied to the remaining 3 experimental units. An alternative method of getting the layout of a *CRD*, when the total number of experimental units is small is the method suggested by Steel and Torrie [13]. In the present example, this will mean that we draw twelve 3-digited numbers from a random sampling number table and then rank them. We break ties by using additional digits. These ranks give a random permutation of the plots 1 to 12. We allot, as before, treatment 1 to the first five plots, treatment 2 to the next four plots and treatment 3 to the remaining three plots in this random order of the plots.

Analysis

We use the following model :

observation from the j th replicate of the i th treatment
 = general effect + {effect due to the i th treatment} + {random error component}

or, symbolically,

$$y_{ij} = \mu + \tau_i + e_{ij}, \quad \dots \quad (20.1)$$

where μ and τ_i 's are a set of constants with $\sum_i \tau_i = 0$, and ϵ_{ij} 's are independently normally distributed with mean zero and variance σ^2 .

We are interested in testing $H_0 : \tau_1 = \tau_2 = \dots = \tau_t$ against the alternatives that τ 's are not all equal. The analysis in the present case is the same as that of one-way classified data considered in Section 19.5. The analysis of variance table is given below :

TABLE 20.1
ANALYSIS OF VARIANCE TABLE FOR A CRD

Source of variation	d.f.	SS	MS	F
Treatments	$t-1$	$\sum_i r_i (y_{i0} - \bar{y}_{00})^2 = SST$	MST	$F = \frac{MST}{MSE}$
Error	$n-t$	$\sum_i \sum_j (y_{ij} - \bar{y}_{00})^2 = SSE$	MSE	
Total	$n-1$	$\sum_i \sum_j (y_{ij} - \bar{y}_{00})^2$		—

We reject H_0 at the level α if $\frac{MST}{MSE} > F_{\alpha ; (t-1), (n-t)}$; otherwise H_0 is accepted. When H_0 is rejected, we may be interested in finding out which of the treatment effects differ significantly. This can be done by using *t*-tests and comparing all possible pairs τ_i, τ_j . This procedure can be simplified by computing the critical difference when the number of replications is the same for each treatment.

Advantages and disadvantages

The CRD is useful in small preliminary experiments and also in certain types of animal or laboratory experiments where the experimental units are homogeneous. There is complete flexibility in the number of treatments and the number of their replications, which may vary from treatment to treatment. This feature also simplifies the analysis when data on some experimental units or on an entire treatment are missing. The CRD provides maximum d.f. for the estimation of experimental error. (The precision of small experiments increases with error d.f.)

The main objection against the CRD is that the principle of local control has not been used in this design. Owing to this, the experimental error is inflated by the presence of the entire variation among

experimental units except the part which is attributable to treatments. We can, as we shall see in the next section, group the experimental units in a manner that will take out a part of the variance among these groups from the experimental error and thereby will reduce the experimental error. The *CRD* is seldom used in field experiments because the plots are not homogeneous. The *CRD* may be used in a chemical or a baking experiment where the experimental units are the parts of the thoroughly mixed chemical or powder.

20.5 Randomised block design (RBD)

The *CRD* will seldom be used if the experimental units are not alike. For in that case the variation among the units will vitiate the test of significance of the treatment effects. The simplest design which enables us to take care of the variability among the units is the *RBD*. This is also the simplest design using all the three principles enunciated by Fisher.

Suppose we want to compare the effects of t treatments, each treatment being replicated an equal number of times, say r times. Then we need $n=rt$ experimental units, and these units are not perhaps homogeneous. The *RBD* consists of two steps. The first step is to divide the units into r more or less homogeneous groups. In each group or block we take as many units as there are treatments. Thus the number of blocks is the same as the common replication number (r). The same technique should be applied to the units of a block. Variation in technique, if any, should be made between the blocks. In agricultural field experiments sometimes a fertility gradient is present. In such a situation, it is advisable to place the blocks across the gradient in order to get homogeneous material for a block and to obtain major differences between blocks. Familiarity with the nature of the experimental units is necessary for an effective blocking of the material.

The second step is to assign the treatments at random to the units of a block. This randomisation has to be done afresh for each block. This is the difference of *RBD* from *CRD*. In an *RBD* randomisation is restricted within a homogeneous block.

With this design each treatment will have the same number of replications. If we want additional replications for some treatments, each of these may be applied to more than one unit in a block.

Layout

Let us obtain the layout of an *RBD* with 5 treatments, each replicated 3 times. So we need 15 units, which are to be grouped into 3 blocks of 5 plots each. We conveniently number the treatments and also the units in a block. Then, following any method of drawing a random sample (as used for the layout of *CRD*), we get a random permutation of the digits from 1 to 5, say 4, 3, 1, 5, 2, for the units of block I. Then we apply treatment number 1 to unit 4, treatment number 2 to unit 3 and so on, finally treatment number 5 to unit 2, of block I. We find another random permutation for block II and so on for the other block.

Analysis

The analysis of this design is the same as that of two-way classified data with one observation per cell. We use the following model : observation for the i th treatment from the j th block

$$= \text{general effect} + \left\{ \begin{array}{l} \text{jth block effect} \\ \text{iith treatment effect} \end{array} \right\} + \left\{ \begin{array}{l} \text{random error component} \end{array} \right\}$$

or, symbolically,

$$y_{ij} = \mu + \beta_j + \tau_i + e_{ij}, \quad \dots \quad (20.2)$$

where μ , β_j 's and τ_i 's are constants with $\sum_j \beta_j = \sum_i \tau_i = 0$, and e_{ij} 's are independently normal with mean zero and variance σ^2_e . The hypothesis of interest is

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_t$$

the alternatives being that τ 's are not all equal. The analysis is the same as that of two-way classified data with one observation per cell considered in Section 19.6. The analysis of variance table will be as follows :

TABLE 20.2
ANALYSIS OF VARIANCE TABLE FOR AN RBD

Source of variation	d.f.	SS	MS	F
Blocks	$t-1$	$t \sum_j (y_{0j} - \bar{y}_{00})^2 = SSB$	MSB	
Treatments	$t-1$	$t \sum_i (y_{i0} - \bar{y}_{00})^2 = SST$	MST	$F = MST/MSE$
Error	$(t-1)(t-1)$	$\sum_i \sum_j (y_{ij} - \bar{y}_{i0} - \bar{y}_{0j} + \bar{y}_{00})^2 = SSE$	MSE	
Total	$rt-1$	$\sum_i \sum_j (y_{ij} - \bar{y}_{00})^2$		-

H_0 is rejected at the level α if

$$\frac{MST}{MSE} > F_{\alpha ; (t-1), (r-1)(t-1)}.$$

Otherwise, H_0 is accepted. Obtaining the critical difference at the level α when the number of replications is the same for each treatment, we can test for the significance of the difference between any two treatment means when H_0 is rejected.

Extra replications for a treatment in an *RBD* will mean that their number is some multiple of r and that the treatment occurs equally often in the different blocks. The standard error of the difference of two such treatment means will be $\sigma \sqrt{\frac{1}{r_1} + \frac{1}{r_2}}$ as in a *CRD*, and not, $\sigma \sqrt{\frac{2}{r}}$ as in the case of an equal-replication *RBD*.

A hypothesis can be framed for block effects and can be tested. But, generally, it is of no interest. If the block effects are significant, then the experimenter may be supposed to have removed the variation among units. Very large block differences may also be due to heteroscedasticity of error and may often be taken care of by a transformation of the variable. Non-significant block effects may mean that either the experimenter was not successful in eliminating variation among units and thereby reducing experimental error or that the units were homogeneous.

Advantages and disadvantages

The *RBD* has many advantages over other designs. It is quite flexible. It is applicable to a moderate number of treatments. If extra replication is necessary for some treatments, these may be applied to more than one unit (but to the same number of units) per block. Since variability among replicates can be eliminated from experimental error, it is not necessary to use continuous blocks. It also enables us to use different techniques to different blocks, though the technique should be the same within a block. The analysis is straightforward and remains so if due to accident data on an entire block or treatment be missing. If data from individual units be missing, then we can use Yates' *missing-plot technique* (*vide* Section 20.14) to estimate the values and perform the test. By grouping the units, we obtain greater precision than is obtainable with the *CRD*.

This is the most popular design with experimenters in view of its simplicity, flexibility and validity. No other design has been used so frequently as the *RBD*. If satisfactory results can be obtained with this design, then we shall not use other complicated designs.

The chief disadvantage is that if the blocks are not internally homogeneous, then a large error term will result. As usually occurs in field experiments, with the increase in the number of treatments, the block size increases and so one has a lesser control over error, for the block will include material of a more heterogeneous nature. In such cases, special types of incomplete block designs are used to reduce the block size.

20.6 Latin square design (LSD)

The principle of 'local control' was used in the *RBD* by grouping the units in one way, i.e. according to blocks. This grouping can be carried one step forward and we can group the units in two ways, each way corresponding to a source of variation among the units, and get the *LSD*. This design is used with advantage in agricultural field experiments where the fertility contours are not always known. Then the *LSD* eliminates the initial variability among the units in two orthogonal directions. The *LSD* has also been used successfully in industry and in the laboratory.

In this design, the number of treatments equals the common replication number of a treatment. So letting m stand for the number of treatments as well as the number of replications for each treatment, the total number of experimental units needed for this design is $m \times m$. These m^2 units are arranged in m rows (one source of variation) and m columns (second source of variation). Then the m treatments are allotted to these m^2 units at random, subject to the condition that each treatment occurs once and only once in each row and in each column.

This arrangement of units and allocation of treatments to units makes the m rows similar to m complete blocks of an *RBD* (the same is true also of the m columns).

The *LSD* is actually an incomplete three-way layout, where all the three factors, rows, columns and treatments, are at the same number of levels (m). For a complete three-way layout with each factor at m levels, we need m^3 experimental units. But in the *LSD*

we take observations on only m^2 of these m^3 units according to the plan stated above.

As an example, let us consider a 4×4 Latin square for comparing four varieties of a crop. We take a rectangular field divided into $4 \times 4 = 16$ plots, arranged in four rows and four columns. We represent the varieties by A , B , C and D . Then the following is a particular 4×4 Latin square :

				Columns
Rows	<i>D</i>	<i>C</i>	<i>B</i>	<i>A</i>
	<i>C</i>	<i>B</i>	<i>A</i>	<i>D</i>
	<i>B</i>	<i>A</i>	<i>D</i>	<i>C</i>
	<i>A</i>	<i>D</i>	<i>C</i>	<i>B</i>

Layout

In connection with the random choice of a Latin square, we first define the following :

The totality of *LSDs* obtained from a single *LSD* by permuting the rows, columns and letters (treatments) is called a *transformation set*. An $m \times m$ Latin square with the m letters A , B , C , in the natural order occurring in the first row and in the first column is called a *standard square* (square in the canonical form). Thus the standard square corresponding to the square cited above is

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>B</i>	<i>C</i>	<i>D</i>	<i>A</i>
<i>C</i>	<i>D</i>	<i>A</i>	<i>B</i>
<i>D</i>	<i>A</i>	<i>B</i>	<i>C</i>

From a standard $m \times m$ Latin square, we may obtain $m!(m-1)!$ different *LSDs* by permuting all the m columns and the $(m-1)$ rows except the first row. Hence there are in all $m!(m-1)!$ different *LSDs* with the same standard square. Thus the total number of different *LSDs* in a transformation set is $m!(m-1)!$ times the number of standard *LSDs* in the set.

As in all other designs, the necessity of randomisation applies to the *LSD* also. In order to give all $m \times m$ *LSDs* equal probability of being selected, we select with equal probability one standard square from all the standard $m \times m$ *LSDs* and then randomise the columns and rows, excluding the first row. More detailed instructions and tables of standard *LSDs* are given in the introduction to Tables XV and XVI of Fisher and Yates' *Statistical Tables for Biological, Agricultural, and Medical Research*.

Two $m \times m$ Latin squares are said to be *orthogonal* if, when these are superimposed, every one of the m^2 pairs of numbers occurs once and once only. A set of $m \times m$ Latin squares is called orthogonal if every pair of them is orthogonal.

Analysis

We shall denote by y_{ijk} the observation on the treatment combination where the factor A is at the i th level (i th row), B is at the j th level (j th column) and C is at the k th level (k th treatment). The triplets (i, j, k) take only m^3 of the possible m^3 values that are dictated by the particular LSD used. If we denote this set of m^3 possible values by D , then (i, j, k) takes values from D , or, symbolically $(i, j, k) \in D$. Then our linear model is

$$\text{with } \left. \begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \tau_k + \epsilon_{ijk}, \quad (i, j, k) \in D \\ \sum_i \alpha_i &= \sum_j \beta_j = \sum_k \tau_k = 0, \end{aligned} \right\} \dots \quad (20.3)$$

and the m^3 random variables ϵ_{ijk} are assumed to be independently normal with mean zero and variance σ_e^2 . α , β and τ stand for the effects due to the factors A , B and C .

The hypothesis of interest here is about zero effects of the treatments (levels of factor C), $H_0 : \text{all } \tau_k = 0$.

The least-square estimates of the effects, obtained by minimising $\sum_{(i, j, k) \in D} (y_{ijk} - \mu - \alpha_i - \beta_j - \tau_k)^2$ subject to the conditions in the model, are $\hat{\mu} = y_{000}$, $\hat{\alpha}_i = y_{i00} - y_{000}$, $\hat{\beta}_j = y_{0j0} - y_{000}$ and $\hat{\tau}_k = y_{00k} - y_{000}$.

TABLE 20.3
ANALYSIS OF VARIANCE TABLE FOR AN $m \times m$ LSD

Source of variation	d.f.	SS	MS	F
Rows	$m-1$	$m \sum_i (y_{i00} - y_{000})^2 = SSR$	MSR	
Columns	$m-1$	$m \sum_j (y_{0j0} - y_{000})^2 = SSC$	MSC	
Treatments	$m-1$	$m \sum_k (y_{00k} - y_{000})^2 = SST$	MST	$F = \frac{MST}{MSE}$
Error	$(m-1)(m-2)$	$\sum (y_{ijk} - y_{i00} - y_{0j0} - y_{00k} + 2y_{000})^2 = SSE$	MSE	
Total	$m^3 - 1$	$\sum (y_{ijk} - y_{000})^2$		—

Σ in SSE and total SS is over $(i, j, k) \in D$.

H_0 is rejected at the level α if

$$\frac{MST}{MSE} > F_{\alpha ; (m-1), (m-1)(m-2)};$$

otherwise, H_0 is accepted.

The estimate of the standard error of each treatment mean is $\sqrt{\frac{MSE}{m}}$, while that for the difference of two treatment means is $\sqrt{\frac{2MSE}{m}}$. The critical difference at level α for testing the differences between treatment means taken two at a time is $t_{\alpha, (m-1)(m-2)} \sqrt{\frac{2MSE}{m}}$.

Ex. 20.1 The following is a 5×5 Latin square for data taken from a manurial experiment with sugarcane. The five treatments were as follows :

A : no manure,

B : an inorganic manure,

C, D and *E* : three levels of farm-yard manure.

TABLE 20.4

PLAN AND YIELD OF SUGARCANE (IN SUITABLE UNITS) PER PLOT

Row	Column				
	I	II	III	IV	V
I	<i>A</i> 52.5	<i>E</i> 46.3	<i>D</i> 44.1	<i>C</i> 48.1	<i>B</i> 40.9
II	<i>D</i> 44.2	<i>B</i> 42.9	<i>A</i> 51.3	<i>E</i> 49.3	<i>C</i> 32.6
III	<i>B</i> 49.1	<i>A</i> 47.3	<i>C</i> 38.1	<i>D</i> 41.0	<i>E</i> 47.2
IV	<i>C</i> 43.2	<i>D</i> 42.5	<i>E</i> 67.2	<i>B</i> 55.1	<i>A</i> 45.3
V	<i>E</i> 47.0	<i>C</i> 43.2	<i>B</i> 46.7	<i>A</i> 46.0	<i>D</i> 43.2

Analyse the above data to find out if there are any treatment effects.

The five row totals are : 231.9, 220.3, 222.7, 253.3 and 226.1 ; the five column totals are : 236.0, 222.2, 247.4, 239.5 and 209.2 ; the five treatment totals are : 242.4, 234.7, 205.2, 215.0 and 257.0.

The grand total is 1,154.3.

The correction factor = $\frac{(1,154.3)^2}{25} = 53,296.3333$.

$$\begin{aligned}\text{Total } SS &= (52.5)^2 + (46.3)^2 + \dots + (46.0)^2 + (43.2)^2 \\ &\quad - 53,296.3333 \\ &= 54,273.51 - 53,296.3333 = 977.1767.\end{aligned}$$

$$\begin{aligned}\text{Row } SS &= \frac{(231.9)^2 + (220.3)^2 + (222.7)^2 + (253.3)^2 + (226.1)^2}{5} \\ &\quad - 53,296.3333 \\ &= \frac{267187.09}{5} - 53,296.3333 = 53,437.4180 - 53,296.3333 \\ &= 141.0847.\end{aligned}$$

$$\begin{aligned}\text{Column } SS &= \frac{(236.0)^2 + (222.2)^2 + (247.4)^2 + (239.5)^2 + (209.2)^2}{5} \\ &\quad - 53,296.3333 \\ &= \frac{267400.49}{5} - 53,296.3333 = 53,480.0980 - 53,296.3333 \\ &= 183.7647.\end{aligned}$$

$$\begin{aligned}\text{Treatment } SS &= \frac{(242.4)^2 + (234.7)^2 + (205.2)^2 + (215.0)^2 + (257.0)^2}{5} \\ &\quad - 53,296.3333 \\ &= \frac{268222.89}{5} - 53,296.3333 = 53,644.5780 - 53,296.3333 \\ &= 348.2447.\end{aligned}$$

$$\begin{aligned}\text{Error } SS &= \text{Total } SS - \text{Row } SS - \text{Column } SS - \text{Treatment } SS \\ &= 304.0826.\end{aligned}$$

TABLE 20.5
ANALYSIS OF VARIANCE TABLE FOR THE LSD

Source of variation	d.f.	SS	MS	F
Rows	4	141.0847	35.2712	
Columns	4	183.7647	45.9412	
Treatments	4	348.2447	87.0612	3.436
Error	12	304.0826	25.3402	
Total	24	977.1767	—	

As $F_{0.01; 4,18} = 5.41$ and $F_{0.05; 4,18} = 3.26$, the hypothesis of no treatment effect is accepted at the 1% level but is rejected at the 5% level.

Advantages and disadvantages

The effect of grouping the units in two ways—according to rows and according to columns—is to eliminate from the error two major sources of variation that are not relevant to the comparisons (among the different treatments) we are interested in. Thus the *LSD* is an improvement over the *RBD* in controlling error by planned grouping, just as the *RBD* is an improvement over the *CRD*.

As has already been observed, the *LSD* is an incomplete 3-way layout. The advantage over the corresponding complete 3-way layout is that only $1/m$ of the m^3 observations are needed.

In field experiments the plots are laid out in a square. But there may be cases when the *LSD* may be used even with the plots in a continuous line, e.g. when the fertility gradient is also along the line.

A serious limitation of the *LSD* is that the number of replicates must be the same as the number of treatments. As a result, squares larger than 12×12 are seldom used, for then the size of the square becomes too large and thus the square does not remain homogeneous. On the other hand, small squares provide only a few *d.f.* for the error, and so we must use a number of such squares (i.e. replicate the *LSD*). The most commonly used sizes are 5×5 to 8×8 .

Another disadvantage is that the analysis depends heavily on the assumption that there are no interactions present.

Also, the analysis is not so simple when there are missing observations.

20.7 Graeco-Latin square

This is another name for a pair of orthogonal Latin squares superimposed on one another, the treatments being represented by Greek letters in one square and Latin letters in the other. In this arrangement, every Greek letter (Latin letter) occurs once in each row and once in each column and once with each Latin letter (Greek letter).

An example of a 3×3 Graeco-Latin square is the following :

A_γ	B_β	C_α
B_α	C_γ	A_β
C_β	A_α	B_γ

To obtain a random square, arrange the rows and columns at random. Then assign the Latin letters and the Greek letters at random.

An $m \times m$ Graeco-Latin square is actually an incomplete 4-way layout with all the four factors at the same level (m), and observations are taken on only m^2 of the possible m^4 treatment combinations.

The analysis of variance table will have five components : rows, columns, Latin letters, Greek letters and error. The *d.f.s* of the first four components will be $(m-1)$ each, while that of error will be $(m-1)(m-3)$. The *SSs* are obtained and the analysis is performed in the usual way.

This design has not been used often. It has the same disadvantage as the *LSD* in case interactions are present. However, Graeco-Latin squares find an application in the construction of certain other designs.

20.8 Cross-over design

A design that resembles the *LSD* but is suitable in dairy husbandry and biological assay when the number of treatments is small is the *cross-over design* (also known as the *change-over design*). The simplest case is of two treatments, *A* and *B*. The number of replicates must be a multiple of two. The experimental units are grouped into pairs. Sometimes one member of each pair is superior to the other and this superiority is about the same for all pairs. Let us call the units in a pair 'good' and 'poor'. Then the treatment *A* is applied to the 'good' members of half of the pairs selected at random from all pairs and *A* is applied to the 'poor' members of the remaining half of the pairs. Thus each treatment is exposed to the same type of units equally frequently.

As an example of the cross-over design, we may cite the following :

Row	Pair					
	1	2	3	4	5	6
Good	<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>B</i>	<i>A</i>
Poor	<i>B</i>	<i>A</i>	<i>B</i>	<i>A</i>	<i>A</i>	<i>B</i>

Randomisation has led to the allotment of treatment *A* to the good units of pairs 1, 3 and 6. The analysis of variance table is as follows :

TABLE 20.6
ANALYSIS OF VARIANCE FOR CROSS-OVER DESIGN

Source of variation	d.f.	SS	MS	F
Pairs (columns)	5	$\sum_i C_i^2/2 - G^2/12 = SSC$	MSC	
Good vs. poor (rows)	1	$\sum_j R_j^2/6 - G^2/12 = SSR$	MSR	
Treatments	1	$\sum_k T_k^2/6 - G^2/12 = SST$	MST	$F = MST/MSE$
Error	4	by subtraction = SSE	MSE	
Total	11	$\sum y^2 - G^2/12$		

C_i , R_j , T_k and G are the i th column total, j th row total, k th treatment total and grand total, respectively. SSE is obtained by subtraction. The hypothesis of equality of treatment effects is rejected at the level α if

$$\frac{MST}{MSE} > F_{\alpha, 1, 4};$$

otherwise, it is accepted.

This design may be used with any number of treatments, subject to the condition that the number of replicates must be a multiple of the number of treatments.

The cross-over design may be used with advantage for large animals (man, cow, etc.), where each animal gives a replicate (column) and the two treatments are applied after some time-lag so that there are no carry-over effects of the application of the first treatment. To half of the animals selected at random, treatment A is applied first and B is applied after noting the result of A and after some time-lag. To the remaining half, B is applied first and then treatment A .

20.9. Factorial experiments

Experiments where the effects of more than one factor, say variety, manure, etc., are considered together are called factorial experiments, while experiments with one factor, say only variety or manure, may be called simple experiments. Consider a simple case of factorial experiment. The yield of a crop depends on the particular variety of crop being used and also on the particular manure

applied. We may have two simple experiments, one for the varieties and one for the manures. The first experiment will give information on whether the different varieties of the crop are equally effective or there are some varieties which will give higher yields than the rest. Similar type of information may be obtained from the second simple experiment about the manures. Though the experiment with varieties will be performed in the presence of a particular manure (not all the manures) and the experiment with manure will be performed with a particular variety (not all the varieties), they will not give us any information about the dependence or independence of the effects of the varieties on those of the manures. The only way to know about the behaviour of the different varieties in the presence of different manures (or *vice versa*) is to have all possible combinations of the varieties and manures in the same experiment, i.e. to conduct a factorial experiment with the two factors, variety and manure.

If there are p different varieties, then we shall say that there are p levels of the factor 'variety'. Similarly, the second factor 'manure' may have q levels, i.e. there may be q different manures or different doses of the same manure. Then this factorial experiment will be called a $p \times q$ -experiment. As a different example, the two factors may be two different manures, say nitrogen and phosphate, and at p and q different doses, respectively. Then this will also give a $p \times q$ -experiment. We shall consider only the simplest cases, viz. cases of n factors each at 2 levels, or what are known as 2^n -experiments, where n is any positive integer greater than or equal to 2.

20.9.1 A 2^n -experiment

Let us consider two factors, A and B , each at 2 levels. Following Yates, we denote by a or b one of two levels at which the corresponding factor (denoted by capital letter) occurs, and for definiteness we shall call this the second level. The first level of A or B will be signified by the absence of the corresponding letter in the treatment combination. Now with 2 factors, each at 2 levels, there will be $2 \times 2 = 4$ treatment combinations. They are enumerated below :

- (1) : A and B both at first levels,
- a : A at second level and B at first level,
- b : A at first level and B at second level,
- ab : A and B both at second levels.

These four treatment combinations may be compared using a *CRD* or an *RBD* or an *LSD*. For a 2^2 -experiment in r randomised blocks, the analysis will be the same as stated in Section 20.5, with the number of treatment combinations $t=4$. And the analysis of a 2^3 -experiment in a Latin square design will be the same as in Section 20.6, with $m=4$. In a factorial experiment one is more interested in the separate tests about main effects and interactions, which are performed by splitting the treatment *SS* carrying 3 *d.f.* into 3 orthogonal components, each carrying a single *d.f.* and each associated either with a main effect or an interaction.

Main effect and interaction effect

The symbols $[a]$ and (a) will be used to denote the *total* and *mean* (respectively) of all the observations receiving the treatment combination a . The letters A , B and AB , when they refer to numbers, will be used to stand for main effects due to factors A and B and the interaction of the two factors.

Consider the effect of A . We may say that the effect of changing factor A from its first level to a in the presence of the first level of factor B is given by $(a)-(1)$, and the effect of changing factor A from its first level to a in the presence of the second level of factor B is given by $(ab)-(b)$. These two effects are known as the *simple effects* of the factor A . If the factors A and B are independent in their effects, then we expect the above two simple effects to be equal, and an average of these two simple effects is defined as the *main effect* due to A . Thus the main effect of the factor A is

$$A = \frac{1}{2}\{(ab)-(b)+(a)-(1)\}.$$

This is simplified by writing it in the following form :

$$A = \frac{1}{2}(a-1)(b+1), \quad \dots \quad (20.4)$$

where the right-hand side is to be expanded algebraically and then the treatment combinations are to be replaced by corresponding treatment means. From the first form of the main effect, we find that A is a linear function of the four treatment means with the sum of the coefficients of the linear function equal to zero ($\frac{1}{2}-\frac{1}{2}+\frac{1}{2}-\frac{1}{2}=0$). Such a linear function of the treatment means with sum of coefficients equal to zero is called a *contrast* (or a comparison)

of the treatment means. Thus the main effect of A (also the main effects of B and interaction AB) is a contrast of the treatment means. Here and in what follows we consider only the case of treatments having equal replication numbers.

If the two factors are not independent, then the above two simple effects of A will not be the same. And one half of the difference of the first simple effect from the second is taken to be a measure of this dependence or *interaction*. Thus the *two-factor interaction* (or the *first-order interaction*) between the factors A and B is

$$AB = \frac{1}{2}\{(ab) - (b) - (a) + (1)\},$$

the simplified version of this being,

$$AB = \frac{1}{2}(a-1)(b-1), \quad \dots \quad (20.5)$$

where the expression on the right-hand side is to be expanded algebraically and then the treatment combinations are to be replaced by the corresponding treatment means.

It is easy to verify that AB is a contrast of the treatment means. The coefficients of the contrasts A and AB satisfy another relation, viz. the sum of products of the corresponding coefficients of the contrasts A and AB is equal to zero ; i.e. $\frac{1}{2} \cdot \frac{1}{2} + (-\frac{1}{2})(-\frac{1}{2}) + (\frac{1}{2})(-\frac{1}{2}) + (-\frac{1}{2})(\frac{1}{2}) = 0$. Such a pair of contrasts are said to be *orthogonal contrasts*.

Next, we define the two simple effects of the factor B and then give the definition of the main effect of B and the interaction BA .

The effect of changing factor B from its first level to b in presence of the first level of factor A is given by $(b) - (1)$, and the effect of changing factor B from its first level to b in presence of the second level of factor A is given by $(ab) - (a)$. Then the main effect of the factor B is

$$B = \frac{1}{2}\{(ab) - (a) + (b) - (1)\}$$

$$\text{or } B = \frac{1}{2}(a+1)(b-1), \quad \dots \quad (20.6)$$

and the interaction of the factor B with the factor A is

$$BA = \frac{1}{2}\{(ab) - (a) - (b) + (1)\}$$

$$\text{or } BA = \frac{1}{2}(a-1)(b-1), \quad \dots \quad (20.7)$$

where in the second forms of B and BA , the right-hand side is to be expanded algebraically and then the treatment combinations are to

be replaced by the corresponding treatment means. Now interaction BA is the same as interaction AB , so that the interaction does not depend on the order of the factors. And it is also easy to verify that the main effect of factor B is a contrast of treatment means and is orthogonal to each of A and AB .

The above three orthogonal contrasts defining the main effects and interaction can be easily obtained from the following table, which gives the signs with which to combine the treatment means and also the divisor. The first line gives the general mean

$$M = \frac{1}{4}\{(ab) + (a) + (b) + (1)\}. \quad \dots \quad (20.8)$$

TABLE 20.7

TABLE OF SIGNS AND DIVISORS GIVING M , A , B AND AB
IN TERMS OF TREATMENT MEANS

Effect	(1)	Treatment mean (a)	(b)	(ab)	Divisor
M	+	+	+	+	4
A	-	+	-	+	2
B	-	-	+	+	2
AB	+	-	-	+	2

The rule to write down the signs of the main effect of a factor is the following : Give to each of the treatment means a *plus* sign where the corresponding factor is at the second level and a *minus* sign where it is at the first level. Or, for the system of notation that we have adopted, give a plus sign to the treatment combinations containing the corresponding small letter and a minus sign where the corresponding small letter is absent. The signs of a two-factor interaction are obtained by combining the corresponding signs of the two main effects. (Two opposite signs will give a minus sign, two identical signs will give a plus sign in the interaction.)

SS due to factorial effects and tests of factorial effects

The factorial effects, main and interaction, are orthogonal contrasts. We can obtain the *SS* due to these factorial effects by multiplying the squares of the factorial effects by a suitable

quantity. These SSs, each having a single d.f., will add up to the SS due to treatments carrying 3 d.f.

It is convenient to obtain the factorial effects and their SSs from the treatment totals rather than from the treatment means. We define the *factorial effect totals* as follows :

$$\left. \begin{aligned} [A] &= [ab] - [b] + [a] - [1], \\ [B] &= [ab] + [b] - [a] - [1], \\ [AB] &= [ab] - [b] - [a] + [1]. \end{aligned} \right\} \dots \quad (20.9)$$

Then the SS due to any main effect or the interaction effect is obtained by multiplying the square of the effect total with the reciprocal of $4r$, where r is the common replication number. Thus

$$\left. \begin{aligned} \text{SS due to main effect of } A &= [A]^2/4r, \text{ with 1 d.f. ;} \\ \text{SS due to main effect of } B &= [B]^2/4r, \text{ with 1 d.f. ;} \\ \text{SS due to interaction } AB &= [AB]^2/4r, \text{ with 1 d.f.} \end{aligned} \right\} \dots \quad (20.10)$$

The general rule for obtaining the SS (carrying 1 d.f.) due to a contrast among t treatment totals (T_i 's) is as follows :

Let $z = \sum_{i=1}^t l_i T_i$, with $\sum_i r_i l_i = 0$ and r_i being the replication number for the i th treatment. Then the SS due to the contrast z is given by

$$SS_z = z^2 / (\sum_i r_i l_i^2). \quad \dots \quad (20.11)$$

It is then a simple matter to express the factorial effect totals or the SSs in terms of the factorial effects, main or interaction, remembering that a factorial effect total is $2r$ times the corresponding factorial effect. Thus the factorial effects are as follows :

$$\left. \begin{aligned} \text{main effect of } A &= [A]/2r, \\ \text{main effect of } B &= [B]/2r, \\ \text{interaction } AB &= [AB]/2r, \end{aligned} \right\} \dots \quad (20.12)$$

and the SS due to a factorial effect is $r \times (\text{factorial effect})^2$.

The test for the significance of any factorial effect, main effect or interaction, may now be obtained by computing

$$F = \frac{MS \text{ due to factorial effect}}{MSE},$$

where MSE is the error MS of the analysis of variance table of

the corresponding design. This F follows the F -distribution with $1, 3(r-1)$ d.f. Hence the hypothesis of absence of a factorial effect is rejected at the level α if for our data

$$F > F_{\alpha ; 1, 3(r-1)};$$

otherwise, the hypothesis is accepted. $3(r-1)$ is the error d.f. for a 2^2 -experiment conducted in an RBD with r blocks.

TABLE 20.8
ANALYSIS OF VARIANCE TABLE FOR A 2^2 -EXPERIMENT
IN r RANDOMISED BLOCKS

Source of variation	d.f.	SS	MS	F
Blocks	$r-1$	SS (Blocks)	MS (Blocks)	
Main effect A	1	$[A]^2/4r$	MS_A	MS_A/MSE
, , B	1	$[B]^2/4r$	MS_B	MS_B/MSE
Interaction AB	1	$[AB]^2/4r$	$MS(AB)$	$MS(AB)/MSE$
Error	$3(r-1)$	by subtraction	MSE	
Total	$4r-1$	$\sum_{i,j} (y_{ij} - \bar{y}_{\text{tot}})^2$		—

The above tests of significance may be simplified by computing the estimate of the standard error of a factorial effect total or factorial effect.

Standard error of a factorial effect total = $\sqrt{4r\sigma_e^2}$ } ... (20.13)
and standard error of a factorial effect mean = $\sqrt{\sigma_e^2/r}$, }

since each factorial effect total is nothing but a linear function of $4r$ independent observations with coefficients ± 1 and common variance σ_e^2 . Thus,

the estimate of standard error of a factorial effect total = $\sqrt{4r.MSE}$
and the estimate of standard error of a factorial effect = $\sqrt{MSE/r}$,
where MSE is the error MS of the analysis of variance table.

Then a factorial effect total must numerically exceed $t_{\alpha/2, 3(r-1)}, \sqrt{4rMSE}$ for significance at the level α , whereas a factorial

effect must exceed numerically $t_{\alpha/2, 3(n-1)} \sqrt{MSE/r}$ for significance at the level α .

Yates' method of computing factorial effect totals

Yates gives a systematic method of obtaining the various effect totals for any 2^n -experiment without writing down the algebraic expressions. We shall describe it for the 2^3 -experiment, but it can be easily extended to the case of any 2^n -experiment.

The steps are as follows :

(i) First, write down the 4 treatment combinations systematically in the first column, starting with the treatment combination (1) and then introducing the letters a, b in turn. After introducing a letter, write down its combination with all the previous treatment combinations and then introduce a new letter. Repeat it until all the letters (n letters in the case of a 2^n -experiment) have been exhausted.

(ii) Next, write down the treatment total from all the replicates in the second column against the appropriate treatment combination.

(iii) The first two columns we get from the observed data. For obtaining column 3, we break the even number of values in the second column into consecutive pairs (1, 2 ; 3, 4 ; etc). Then in the first half of the third column we write down the sums of the values in these pairs in order and in the second half of the third column we write down in order the differences of the values in the pairs in the second column (the first member subtracted from the second member of a pair).

(iv) We next break the values in the third column into consecutive pairs and put the sums and differences of the members of these pairs in order in the fourth column.

For a 2^3 -experiment the fourth column values give the factorial effect totals corresponding to the treatment combinations occurring in the corresponding position of the first column.

For a 2^n -experiment we are to repeat n times the operations of columns 3 and 4 and then the values in the $(n+2)$ nd column will be the factorial effect totals, the first entry in the last column being always the grand total.

TABLE 20.9
YATES' METHOD FOR A 2²-EXPERIMENT

Treatment combination (1)	Total (2)	(3)	(4)
(1)	[1]	[1]+[a]	[1]+[a]+[b]+[ab]=grand total
a	[a]	[b]+[ab]	[a]-[1]+[ab]-[b]=[A]
b	[b]	[a]-[1]	[b]+[ab]-[1]-[a]=[B]
ab	[ab]	[ab]-[b]	[ab]-[b]-[a]+[1]=[AB]

Ex. 20.2 A 2²-experiment in six randomised blocks was conducted in order to obtain an idea of the interaction : spacing \times number of seedlings per hole, along with the effects of different types of spacing and different numbers of seedlings per hole, while adopting the Japanese method of cultivation.

The levels of the two factors are :

$S:$ $\left\{ \begin{array}{l} 8'' \text{ spacings in between,} \\ 10'' \text{ spacings in between,} \end{array} \right.$

and $N:$ $\left\{ \begin{array}{l} 3 \text{ seedlings per hole,} \\ 4 \text{ seedlings per hole.} \end{array} \right.$

The field plan and yield of dry *Aman* paddy (in kg.) are given below :

Block 1				Block 2				Block 3			
(1)	s	ns	n	ns	(1)	s	n	(1)	n	s	ns
117	106	109	114	114	120	117	114	111	117	114	106
Block 4				Block 5				Block 6			
ns	n	s	(1)	ns	s	(1)	n	n	(1)	ns	s
98	121	112	108	75	97	73	38	58	81	105	117

Analyse the data to find out if there are any significant treatment effects—main or interaction.

We apply Yates' method to find the total effects.

TABLE 20.10
YATES' METHOD FOR THE ABOVE 2^a-EXPERIMENT

Treatment combination (1)	Total yield from all blocks (2)	(3)	(4)	Main and interaction effects
(1)	610	1172	2442—grand total	
n	562	1270	-104=[N]	-8·667=N
s	663	-48	98=[S]	8·167=S
ns	607	-56	-8=[NS]	-0·667=NS

We next perform the randomised block analysis.

The six block totals are : 446, 465, 448, 439, 283 and 361.

The treatment totals are : [1]=610, [n]=562, [s]=663 and [ns]=607.

Raw total $SS=259,024$;

$$\text{Correction factor} = \frac{(2442)^2}{24} = \frac{5963364}{24} = 248,473\cdot 5 ;$$

$$\text{Total } SS=259,024 - 248,473\cdot 5 = 10,550\cdot 5 ;$$

$$\begin{aligned}\text{Block } SS &= \frac{(446)^2 + \dots + (361)^2}{4} - 248,473\cdot 5 \\ &= \frac{1018976}{4} - 248,473\cdot 5 = 254,744 - 248,473\cdot 5 \\ &= 6,270\cdot 5 ;\end{aligned}$$

$$\begin{aligned}\text{Treatment } SS &= \frac{(610)^2 + \dots + (607)^2}{6} - 248,473\cdot 5 \\ &= \frac{1495962}{6} - 248,473\cdot 5 = 249,327 - 248,473\cdot 5 \\ &= 853\cdot 5 ;\end{aligned}$$

$$\text{Error } SS=10,550\cdot 5 - 6,270\cdot 5 - 853\cdot 5 = 3,426\cdot 5 .$$

$$\text{Also, } SS \text{ due to } N = \frac{(-104)^2}{24} = 450\cdot 667 ;$$

$$SS \text{ due to } S = \frac{(98)^2}{24} = 400\cdot 167 ;$$

$$SS \text{ due to } NS = \frac{(-8)^2}{24} = 2\cdot 667 .$$

TABLE 20.11
ANALYSIS OF VARIANCE TABLE FOR THE 2^a-EXPERIMENT

Source of variation	d.f.	SS	MS	F	
Blocks	5	6,270.5	1,254.1		
N	1	450.667	450.667	1.973	
S	1	400.167	400.167	1.752	$F_{.05; 1, 15} = 4.54$
NS	1	2.667	2.667	<1	
Error	15	3,426.5	228.433		
Total	23	10,550.5		—	

There are no significant main or interaction effects present in the above experiment, as in each of the cases the computed value of F is less than the corresponding theoretical value at the 5% level.

20.9.2 A 2³-experiment

We now consider the case of three factors A , B and C , each at 2 levels, where a , b and c will denote the second levels of the factors, respectively. The $2 \times 2 \times 2 = 8$ treatment combinations written in the systematic order are :. (1), a , b , ab , c , ac , bc , abc .

The 8 treatment combinations may be compared in any of the designs—CRD, RBD or LSD. The analysis will be the same as in the corresponding design, the number of treatments being $t=8$ in CRD and RBD and $m=8$ in LSD. The treatment SS has 7 d.f. We next divide it into 7 orthogonal contrasts of the 8 treatment means (or totals) with the help of the main effects and interactions. In a three-factor experiment there are three main effects— A , B , C ; three first-order interactions— AB , AC , BC ; and one second-order (or three-factor) interaction— ABC .

Main effects and interactions

The factor A has the following 4 simple effects :

The effect of changing factor A from its first to its second level in the presence of the first levels of factors B and C is given by (a)—(1); the effect of changing factor A from its first to its second

level in the presence of the second level of B and the first level of C is given by $(ab) - (b)$; the effect of changing factor A from its first to its second level in the presence of the first level of B and the second level of C is $(ac) - (c)$; the effect of changing factor A from its first to its second level in the presence of the second levels of factors B and C is $(abc) - (bc)$.

Similarly for the factors B and C .

As in a 2^3 -experiment, here also the main effect of A is defined to be the average of the above four simple effects :

$$A = \frac{1}{4}\{(abc) - (bc) + (ac) - (c) + (ab) - (b) + (a) - (1)\}$$

or $A = \frac{1}{4}(a-1)(b+1)(c+1), \dots \quad (20.14)$

where the right-hand side is to be expanded algebraically and treatment combinations are to be replaced by treatment means.

The interaction of A with B is next obtained separately at the two levels of C ;

$$AB \text{ (when } C \text{ is at the first level)} = \frac{1}{4}\{(ab) - (b) - (a) + (1)\}$$

and $AB \text{ (when } C \text{ is at the second level)} = \frac{1}{4}\{(abc) - (bc) - (ac) + (c)\}.$

From the average of these two we get the AB interaction, and half the difference of the first from the second gives the interaction of AB with C or the ABC interaction.

Thus

$$AB = \frac{1}{4}\{(abc) - (bc) - (ac) + (c) + (ab) - (b) - (a) + (1)\}$$

and $ABC = \frac{1}{4}\{(abc) - (bc) - (ac) + (c) - (ab) + (b) + (a) - (1)\}$

or, equivalently,

$$AB = \frac{1}{4}(a-1)(b-1)(c+1) \dots \quad (20.15)$$

and $ABC = \frac{1}{4}(a-1)(b-1)(c-1), \dots \quad (20.16)$

where the right-hand sides are to be expanded algebraically and treatment combinations are to be replaced by treatment means. From the four simple effects of A , we may also obtain AC and ACB interactions by first obtaining AC (when B is at first level) and AC (when B is at its second level). Here also ABC is the same three-factor interaction for all permutations of the letters. The main effects of B and C and the interaction BC may be derived starting

from the simple effects of B and simple effects of C . These 7 effects due to the main effects and the interactions are mutually orthogonal contrasts of the treatment means. We can verify this from the following table of signs :

TABLE 20.12

TABLE OF SIGNS AND DIVISORS GIVING M , A , B , C , AB , AC , BC AND ABC IN TERMS OF TREATMENT MEANS

Effect	(1)	(a)	(b)	Treatment mean				Divisor
	(ab)	(c)	(ac)	(bc)	(abc)			
M	+	+	+	+	+	+	+	8
A	-	+	-	+	-	+	-	4
B	-	-	+	+	-	-	+	4
C	-	-	-	-	+	+	+	4
AB	+	-	-	+	+	-	-	4
AC	+	-	+	-	-	+	-	4
BC	+	+	-	-	-	-	+	4
ABC	-	+	+	-	+	-	-	4

The rules of obtaining the signs of effects and two-factor interactions are the same as those stated for Table 20.7 for a 2^3 -experiment. The signs of ABC may be obtained by combining the signs of AB and C (or of AC and B or of BC and A).

SS due to factorial effects and tests of significance of factorial effects

We define factorial effect totals as in the 2^3 -experiment by combining the 8 treatment totals with the signs given in the above table. Thus

$$[A] = [abc] - [bc] + [ac] - [c] + [ab] - [b] + [a] - [1],$$

and similarly the other effect totals are obtained.

The SS due to a factorial effect is obtained by multiplying the square of the corresponding effect total by the reciprocal of $8r$, where r is the common replication number. Thus,

$$SS \text{ due to main effect } A = [A]^2 / 8r, \text{ with } 1 \text{ d.f.}$$

The test for the significance of any factorial effect, main effect or interaction may now be obtained by computing

$$F = \frac{MS \text{ due to factorial effect}}{MSE},$$

where MSE is the error MS of the analysis of variance table of the corresponding design. This F follows the F distribution with $1, 7(r-1)$ d.f. Hence the hypothesis of the absence of the factorial effect is rejected at the level α if for our data

$$F > F_{\alpha ; 1, 7(r-1)};$$

otherwise, the hypothesis is accepted. $7(r-1)$ is the error d.f. for a 2^3 -experiment conducted in r randomised blocks.

TABLE 20.13
ANALYSIS OF VARIANCE TABLE FOR A 2^3 -EXPERIMENT
IN r RANDOMISED BLOCKS

Source of variation	d.f.	SS	MS	F
Blocks	$r-1$	SS (Blocks)	MS (Blocks)	
Main effect A	1	$[A]^2/8r$	MSA	MSA/MSE
" " B	1	$[B]^2/8r$	MSB	MSB/MSE
" " C	1	$[C]^2/8r$	MSC	MSC/MSE
Two-factor interaction				
AB	1	$[AB]^2/8r$	$MS(AB)$	$MS(AB)/MSE$
" " AC	1	$[AC]^2/8r$	$MS(AC)$	$MS(AC)/MSE$
" " BC	1	$[BC]^2/8r$	$MS(BC)$	$MS(BC)/MSE$
Three-factor interaction				
ABC	1	$[ABC]^2/8r$	$MS(ABC)$	$MS(ABC)/MSE$
Error	$7(r-1)$	SSE	MSE	
Total	$8r-1$	$\sum_{i,j} (y_{ij} - \bar{y}_{\text{tot}})^2$		—

The above seven F tests may be replaced by computing the estimate of the standard error of a factorial effect total in the 2^3 -experiment, which is $\sqrt{8r}MSE$, and then the factorial effect total must numerically exceed $t_{\alpha/2, 7(r-1)}\sqrt{8r}MSE$ for its significance at the level α .

Yates' method of computing factorial effect totals for a 2³-experiment

We follow the instructions given in the case of a 2³-experiment and obtain one more column, as in Table 20.14.

TABLE 20.14
YATES' METHOD FOR A 2³-EXPERIMENT

Treatment combination (1)	Total (2)	(3)	(4)	(5)
(1)	[1]	[1]+[a]	[b]+[ab]+[1]+[a]	[bc]+[abc]+[c]+[ac]+[b] +[ab]+[1]+[a]=grand total
a	[a]	[b]+[ab]	[bc]+[abc]+[c]+[ac]	[abc]-[bc]+[ac]-[c]+[ab] -[b]+[a]-[1]=[A]
b	[b]	[c]+[ac]	[ab]-[b]+[a]-[1]	[bc]+[abc]-[c]-[ac]+[b] +[ab]-[1]-[a]=[B]
ab	[ab]	[bc]+[abc]	[abc]-[bc]+[ac]-[c]	[abc]-[bc]-[ac]+[c]+[ab] -[b]-[a]+[1]=[AB]
c	[c]	[a]-[1]	[b]+[ab]-[1]-[a]	[bc]+[abc]+[c]+[ac]-[b] -[ab]-[1]-[a]=[C]
ac	[ac]	[ab]-[b]	[bc]+[abc]-[c]-[ac]	[abc]-[bc]+[ac]-[c]-[ab] +[b]-[a]+[1]=[AC]
bc	[bc]	[ac]-[c]	[ab]-[b]-[a]+[1]	[bc]+[abc]-[c]=[ac]-[b] -[ab]+[1]+[a]=[BC]
abc	[abc]	[abc]-[bc]	[abc]-[bc]-[ac]+[c]	[abc]-[bc]-[ac]+[c]-[ab] +[b]+[a]-[1]=[ABC]

Orthogonality of a design and confounding

We have already defined orthogonal contrasts. Now we consider their practical utility. Suppose we have a random sample of n independent observations x_1, x_2, \dots, x_n from a normal population with variance σ^2 . If we consider two contrasts that are orthogonal,

$$\left. \begin{aligned} A &= \sum_i \lambda_i x_i \\ B &= \sum_i \mu_i x_i, \\ \text{with } \sum_i \lambda_i = 0, \sum_i \mu_i = 0 \text{ and } \sum_i \lambda_i \mu_i = 0, \end{aligned} \right\} \quad . \quad (20.17)$$

then we have

$$\text{cov}(A, B) = \sigma^2 \sum_i \lambda_i \mu_i = 0. \quad . \quad (20.17a)$$

This means that if we use A and B to estimate two different effects, then the errors in the two estimates will not be related as A and B will be distributed independently. These estimates also are then said to be orthogonal. Yates defines *orthogonality of a design* as the property which ensures that the different effects will be capable of separate estimation and testing without any entanglement. If our data arise from an orthogonal design, then we are not involved in any difficulties in making independent estimation and tests of effects.

The *CRD*, *RBD* and *LSD* give us orthogonal designs. But the difficulty in conducting a factorial experiment in an *RBD* or *LSD* is that, as the number of factors and/or that of levels of the factors increase, the number of treatment combinations to be compared increases too. This in turn necessitates the use of large-sized blocks or squares to accommodate all the treatment combinations. E.g., in a 2^6 -experiment there should be 32 plots in a block. But it has been found that the experimental error increases with an increase in the size of a block or square, for then it becomes less effective in controlling the heterogeneity of the units. A remedy has been found out : this is to divide a replicate (a complete block) into a number of equal blocks (incomplete blocks) and then to allocate the treatment combinations to these blocks so that only the unimportant treatment comparisons get mixed up or entangled with the block comparisons. These treatment comparisons are then said to be *confounded* or mixed up with block effects ; these effects cannot be separately tested or estimated. But the remaining treatment effects, which are not confounded with the block effects, are still capable of separate estimation and testing. Since in a confounded design we lose information on some of the treatment comparisons, these should be the least important comparisons and usually they are the highest-order interactions. It is easy to interpret simple interactions—first-order or second-order. But as the order increases, the interpretation becomes difficult, and high-order interactions are also of little or no importance to the experimenter.

Confounding in experimental designs is then a term to denote an arrangement of the treatment combinations in the blocks in which less important treatment effects are purposively confounded with the blocks. This non-orthogonality is not a defect of the design ; it is

deliberately introduced in order to get better estimates and tests on the important treatment comparisons.

We shall consider in detail the simplest case of confounding in a 2^n -experiment, where each replicate will be divided into two equal-sized blocks and the highest-order interaction will be confounded. In a 2^n -experiment it is possible to reduce the block size by using 2^k blocks (k being a positive integer) in a replicate. If $k > 1$, then more than one treatment comparison will be confounded. Actually, a 2^n -experiment in 2^k blocks (or blocks of 2^{n-k} plots each) confounds $(2^k - 1)$ treatment comparisons.

Confounding in a 2^3 -experiment

There are 2^3 or 8 treatment combinations under comparison in such an experiment, and suppose we decide to use blocks of 4 plots each. Then we need two blocks to give a complete replicate.. We are to divide the 8 treatment combinations into two groups of four treatments each and allot the two groups to the two blocks at random. Referring to Table 20.12, we find that the interaction ABC depends on

$$(abc) + (a) + (b) + (c) - (1) - (ab) - (ac) - (bc).$$

Let us apply the four treatments with plus signs in ABC in one block and the remaining four with minus signs in ABC in the other block. Thus abc, a, b, c go to block 1, whereas $(1), ab, ac, bc$ go to block 2, say. Then the contrast measuring the interaction ABC also contains block effects—effect of block 1 minus effect of block 2. So we say that ABC is mixed up or confounded with block effects and as such we lose information on ABC . On the other hand, the other six contrasts of the treatments, viz. A, B, C, AB, AC and BC , will have each two treatments from block 1 (block 2) with plus signs and two treatments with minus signs. And so they will contain no block effects and, being orthogonal to ABC , will also be orthogonal to blocks. Thus, in the above allocation of 8 treatments to the two blocks, no difficulties arise in the estimation or testing of the main and first-order interaction effects.

The above procedure is quite general, and in a 2^n -experiment we can confound a single $d.f.$ due to any effect by selecting the appropriate interaction and applying the treatment combination with plus signs in that interaction effect in one block and the

treatment combinations with minus signs in the other block. This will ensure the confounding of that interaction with blocks and the orthogonality of the remaining effects to blocks.

Confounding may be of two types—complete and partial. In *complete confounding*, we confound the same interaction in all the replications and so lose information on that from all the replications, whereas the unconfounded effects are orthogonal to the blocks of the replicates and can be obtained and tested as in a complete block design. But for the effect which is completely confounded, we do not have a separate component in the analysis of variance table ; it appears along with the block component.

Thus the allocation of the treatments to the two blocks of each replicate (before randomisation) of a 2^3 -experiment in r replicates with ABC completely confounded will be as follows :

Replicate	
Block 1	Block 2
a	ab
b	ac
c	bc
abc	(1)

The first two columns of the analysis of variance table will be as follows :

Source	d.f.
Blocks	$2r - 1$
A	1
B	1
C	1
AB	1
AC	1
BC	1
Error	$6(r - 1)$
Total	$8r - 1$

All SSs are computed in the usual way and tests for A , B , C , AB , AC and BC are obtained with the help of MSE . Note that there is no separate entry for ABC , which has been completely confounded with the blocks. This component is contained in the $(2r-1)$ d.f. due to blocks. We may use Yates' method for obtaining the total effects corresponding to the main effects and first-order interaction effects. Then, of course, we do not use the value of ABC given by that method.

Ex. 20.3 For a factorial experiment with three factors, N , P and K each at two levels, the design and yield per plot are given below. Analyse the experiment.

	Replicate 1				Replicate 2				
Block 1	(1)	pk	nk	np	p	npk	n	k	Block 3
	25	24	32	30		32	42	46	
Block 2	n	k	nPK	p	nk	(1)	np	pk	Block 4
	30	32	36	27		34	44	30	
	Replicate 3				Replicate 4				
Block 5	nPK	k	n	p	np	(1)	pk	nk	Block 7
	30	32	28	26		32	34	39	
Block 6	(1)	pk	nk	np	nPK	n	p	k	Block 8
	24	20	28	36		45	41	29	

This is a 2^3 -experiment conducted in four replicates and each replicate has been divided into blocks of four plots each. Thus this is an example of a confounded 2^3 -experiment. By referring to Table 20.12, we find that interaction NPK has been completely confounded with blocks.

We apply Yates' method for obtaining the six unconfounded treatment effects and then to find their SSs. The value for NPK will not be used, as it is completely confounded, and hence will occur along with the block component.

TABLE 20.15
YATES' METHOD FOR A 2³-EXPERIMENT

Treatment combination	Total from all replicates	(1)	(2)	(3)	Mean effects	$SS = []^2 / 32$
(1)	127	272	514	1059		
<i>n</i>	145	242	545	63=[<i>N</i>]	3.9375= <i>N</i>	124.0312
<i>p</i>	114	273	32	-31=[<i>P</i>]	-1.9375= <i>P</i>	30.0312
<i>np</i>	128	272	31	33=[<i>NP</i>]	2.0625= <i>NP</i>	34.0312
<i>k</i>	138	18	-30	31=[<i>K</i>]	1.9375= <i>K</i>	30.0312
<i>nk</i>	135	14	-1	-1=[<i>NK</i>]	-0.0625= <i>NK</i>	0.0312
<i>pk</i>	119	-3	-4	29=[<i>PK</i>]	1.8125= <i>PK</i>	26.2812
<i>npk</i>	153	34	37	41		

The eight block totals are : 111, 125, 159, 144, 116, 108, 146 and 150.

Grand total=1,059.

Raw $SS=36,381$.

$$\text{Corrected total } SS = 36,381 - \frac{(1059)^2}{32} = 36,381 - 35,046.2833 \\ = 1,334.7167;$$

$$\text{Block } SS = \frac{(111)^2 + (125)^2 + \dots + (146)^2 + (150)^2}{4} - 35,046.2833 \\ = \frac{142899}{4} - 35,046.2833 = 35,724.75 - 35,046.2833 \\ = 678.4667.$$

From Table 20.15,

$$\text{Treatment } SS = \frac{[N]^2 + [P]^2 + \dots + [NK]^2 + [PK]^2}{r \cdot 2^3} \\ = \frac{(63)^2 + (-31)^2 + \dots + (-1)^2 + (29)^2}{32} \\ = \frac{7822}{32} = 244.4375;$$

$$\text{Error } SS = \text{total } SS - \text{block } SS - \text{treatment } SS \\ = 1,334.7167 - 678.4667 - 244.4375 \\ = 411.8125.$$

TABLE 20.16
ANALYSIS OF VARIANCE TABLE FOR A 2ⁿ-EXPERIMENT
WITH *NPK* COMPLETELY CONFOUNDED

Source of variation	d.f.	SS	MS	F	
Blocks	7	678.4667	96.9238		
<i>N</i>	1	124.0312	124.0312	5.421	
<i>P</i>	1	30.0312	30.0312	1.313	$F_{.01, 1, 18} = 8.29$
<i>K</i>	1	30.0312	30.0312	1.313	$F_{.05, 1, 18} = 4.41$
<i>NP</i>	1	34.0312	34.0312	1.487	
<i>NK</i>	1	0.0312	0.0312	<1	
<i>PK</i>	1	26.2812	26.2812	1.149	
Error	18	411.8125	22.8784		
Total	31	1,334.7167		—	

From the above analysis of variance table, we find that only the main effect of *N* is significant at the 5% level. Other treatment effects are not significant at the 5% level.

We next compare the unconfounded and completely confounded 2ⁿ-experiments by defining the *information* of an effect contained in the experiment as the reciprocal of the variance of its estimator.

In the case of an unconfounded design, the replicate is itself a block and in this case we shall denote the error variance by σ^2 . In a completely confounded design, a block is a half-replicate, two blocks make up a replicate. In this case we denote the error variance by $\sigma_{1/2}^2$. Thus σ^2 and $\sigma_{1/2}^2$ are the error variances for a complete block design (unconfounded) and an incomplete block design (a block containing only half the treatment combinations, as is the case in a complete confounding), respectively. And it is expected that $\sigma_{1/2}^2 < \sigma^2$, since the smaller blocks will have greater control over error than the complete blocks which are large.

The variance of the estimator of an effect, main or interaction, in a 2ⁿ-experiment in *r* replicates without confounding is $\sigma^2/r \cdot 2^{n-2}$. Whereas the variance of the estimator of each unconfounded effect in

a 2^n -experiment in r replicates, completely confounding the highest-order interaction, is $\sigma_{1/2}^2/r \cdot 2^{n-2}$. Thus the information about each effect in an unconfounded design is $r \cdot 2^{n-2}/\sigma^2$, whereas the information about each unconfounded effect in a completely confounded design is $r \cdot 2^{n-2}/\sigma_{1/2}^2$. Since, as has already been observed, $\sigma_{1/2}^2$ will be smaller than σ^2 , the completely confounded design contains more information about the unconfounded effects than the unconfounded design does. But the former design contains zero information about the effect that has been completely confounded, whereas we get information amounting to $r \cdot 2^{n-2}/\sigma^2$ about this from the unconfounded design.

Sometimes it may be that we are not sure whether the highest-order interaction is really absent or unimportant. In such cases we shall be unwilling to sacrifice the entire information on this. We shall, instead, distribute the loss among more than one interaction and shall get some information on each of them. This is achieved by a partially confounded design, which we shall discuss now.

Partial confounding in a 2^3 -experiment

We illustrate this technique with a 2^3 -experiment, though we shall rarely have any occasion to use a confounded design for such a small experiment. Here we have four interactions, viz. AB , AC , BC and ABC . We take four replications and two blocks of size four in each replicate. We allot the 8 treatments to the blocks of a replicate so that AB is confounded in replicate 1, AC in replicate 2, BC in replicate 3 and ABC in replicate 4. The layout, before randomisation, will look like the following :

Block 1	Block 2	Block 3	Block 4
(1) ab c abc	a b ac bc	(1) b ac abc	a ab c bc
Replicate 1 AB confounded		Replicate 2 AC confounded	
Block 5	Block 6	Block 7	Block 8
(1) a bc abc	b ab c ac	(1) b ac bc	a b c abc
Replicate 3 BC confounded		Replicate 4 ABC confounded	

The above design is an example of a *partially confounded 2³-experiment* with all the interactions partially confounded.

In the above design the main effects A , B , C are not confounded in any replicate, so they are estimated from all 4 replicates. The experiment contains $8/\sigma_{1/2}^2$ information about each of the main effects. But each interaction is confounded in one replicate and left unconfounded in three others. Thus we can estimate this interaction from those replicates where it is not confounded ; e.g., AB will be estimated from replicates 2, 3 and 4. So only three replicates contain information about the confounded interactions, and the amount of information for them is $6/\sigma_{1/2}^2$. Thus the relative information of each partially confounded interaction with respect to the unconfounded main effects is $6/8$ or $3/4$, which is the same as the proportion of replicates giving information about the confounded interaction.

The table below summarises the amount of information contained in various types of 2³-experiment in four replications

TABLE 20.17
AMOUNT OF INFORMATION IN DIFFERENT 2³-EXPERIMENTS

Effect	Amount of information		
	Unconfounded design	ABC completely confounded	AB , AC , BC and ABC partially confounded
A			$8/\sigma_{1/2}^2$
B			$8/\sigma_{1/2}^2$
C			$8/\sigma_{1/2}^2$
AB	$8/\sigma^2$ each	$8/\sigma_{1/2}^2$ each	$6/\sigma_{1/2}^2$
AC			$6/\sigma_{1/2}^2$
BC			$6/\sigma_{1/2}^2$
ABC		Zero	$6/\sigma_{1/2}^2$

Since usually $\sigma_{1/2} < \sigma$, the confounded experiments will contain more information on unconfounded effects than the unconfounded experiments will. In a partially confounded design, we get some information on confounded effects, though the information is less than that for an unconfounded effect.

The first two columns of the analysis of variance table in the case of a partially confounded 2^8 -experiment, partially confounding all the interactions using four replicates, will be as follows :

Source	d.f.
Blocks	7
<i>A</i>	1
<i>B</i>	1
<i>C</i>	1
<i>AB</i>	1
<i>AC</i>	1
<i>BC</i>	1
<i>ABC</i>	1
Error	17
Total	31

The Block SS is computed from the 8 block totals and grand total. SSs due to the main effects *A*, *B*, *C*, which are not confounded with blocks, are computed using data from all four replicates, whereas the SS due to any confounded interaction is obtained from those replicates where that particular interaction is not confounded.

We may obtain a table of the following type to get the different total effects—main and interaction.

TABLE 20.18
TABLE FOR OBTAINING EFFECTS IN A PARTIALLY
CONFOUNDED DESIGN

(1) Treatment combination	(2) Total from all replicates	(3) Total from replicates where <i>AB</i> is not confounded	(4) Total from replicates where <i>AC</i> is not confounded	(5) Total from replicates where <i>BC</i> is not confounded	(6) Total from replicates where <i>ABC</i> is not confounded
(1)					
<i>a</i>					
<i>b</i>					
<i>ab</i>					
<i>c</i>					
<i>ac</i>					
<i>bc</i>					
<i>abc</i>					

$[A]$, $[B]$ and $[C]$ are obtained from column 2 of the above table,
 $[AB]$ is obtained from column 3,
 $[AC]$ is obtained from column 4,
 $[BC]$ is obtained from column 5,
and $[ABC]$ is obtained from column 6.

If there are in all $4r$ replicates and the interactions are partially confounded only in r of the replicates each, the SSs will be as follows :

$$SS \text{ due to } A = [A]^2 / 32r,$$

$$SS \text{ due to } B = [B]^2 / 32r,$$

$$SS \text{ due to } C = [C]^2 / 32r,$$

$$SS \text{ due to } AB = [AB]^2 / 24r,$$

$$SS \text{ due to } AC = [AC]^2 / 24r,$$

$$SS \text{ due to } BC = [BC]^2 / 24r,$$

and $SS \text{ due to } ABC = [ABC]^2 / 24r.$

Each of the above SSs carries 1 d.f

Ex. 20.4 The plan and yield per plot (in a suitable unit) of a 2^3 field experiment on wheat are given below, the treatments being all combinations of two levels of dung (0, d), two levels of potash (0, k) and two levels of superphosphate (0, p). Analyse the data.

	Replicate 1				Replicate 2				Block 3
Block 1	(1) 42	pkd 48	kd 36	p 45	pd 50	(1) 58	k 50	pkd 41	
Block 2	d 53	k 55	pd 55	kp 41	pk 44	p 58	kd 41	d 43	Block 4
Replicate 3									
Block 5	pkd 52	pk 54	d 42	(1) 56	(1) 47	pk 34	kd 42	pd 50	Block 7
Block 6	k 43	pd 57	p 52	kd 39	p 52	pkd 52	k 39	d 44	
Replicate 4									

Since each replicate has been divided into 2 blocks, one effect has been confounded in each replicate. Replicate 1 confounds KD , replicate 2 confounds PD , replicate 3 confounds PK , and PKD has been confounded in replicate 4.

The 8 block totals are : 171, 204, 199, 186, 204, 191, 173 and 187.
Grand total=1,515.

$$\begin{aligned}\text{Block } SS &= \frac{(171)^2 + (204)^2 + \dots + (173)^2 + (187)^2 - (1515)^2}{4} - \frac{32}{32} \\ &= \frac{288049}{4} - \frac{2295225}{32} \\ &= 72,012.25 - 71,725.78125 = 286.46875 ;\end{aligned}$$

Raw SS=73,141 ;

Total **SS**=73,141 - 71,725.78125 = 1,415.21875.

Next, to obtain the treatment **SS**, we form the following table :

TABLE 20.19

TABLE FOR OBTAINING THE MAIN EFFECTS AND INTERACTIONS

(1) Treatment combination	(2) Total from all replicates	(3) Total from replicates 1, 2 and 3	(4) Total from replicates 1, 2 and 4	(5) Total from replicates 1, 3 and 4	(6) Total from replicates 2, 3 and 4
(1)	203	156	147	145	161
p	207	155	155	149	162
k	187	148	144	137	132
pk	173	139	119	129	132
d	182	138	140	139	129
	212	162	155	162	157
kd	158	116	119	117	122
pkd	193	141	141	152	145

The total effects due to P, K and D are obtained from column (2) of Table 20.19 :

$$\begin{aligned}[P] &= -[1] + [p] - [k] + [pk] - [d] + [pd] - [kd] + [pkd] \\ &= -203 + 207 - 187 + 173 - 182 + 212 - 158 + 193 \\ &= -730 + 785 = 55,\end{aligned}$$

$$\begin{aligned}[K] &= -[1] - [p] + [k] + [pk] - [d] - [pd] + [kd] + [pkd] \\ &= -203 - 207 + 187 + 173 - 182 - 212 + 158 + 193 \\ &= -804 + 711 = -93,\end{aligned}$$

$$\begin{aligned}[D] &= -[1] - [p] - [k] - [pk] + [d] + [pd] + [kd] + [pkd] \\ &= -203 - 207 - 187 - 173 + 182 + 212 + 158 + 193. \\ &= -770 + 745 = -25.\end{aligned}$$

The total effect due to PK is obtained from column (4) of Table 20.19 as $[PK]=[1]-[p]-[k]+[pk]+[d]-[pd]-[kd]+[pkd]$

$$\begin{aligned} &= 147 - 155 - 144 + 119 + 140 - 155 - 119 + 141 \\ \therefore &= 547 - 573 = -26. \end{aligned}$$

The total effect due to PD is obtained from column (5) of Table 20.19 as $[PD]=[1]-[p]+[k]-[pk]-[d]+[pd]-[kd]+[pkd]$

$$\begin{aligned} &= 145 - 149 + 137 - 129 - 139 + 162 - 117 + 152 \\ &= 596 - 534 = 62. \end{aligned}$$

The total effect due to KD is obtained from column (6) of Table 20.19 as $[KD]=[1]+[p]-[k]-[pk]-[d]-[pd]+[kd]+[pkd]$

$$\begin{aligned} &= 161 + 162 - 132 - 132 - 129 - 157 + 122 + 145 \\ &= 590 - 550 = 40. \end{aligned}$$

The total effect due to PKD is obtained from column (3) of Table 20.19 as $[PKD]=-[1]+[p]+[k]-[pk]+[d]-[pd]-[kd]+[pkd]$

$$\begin{aligned} &= -156 + 155 + 148 - 139 + 138 - 162 - 116 + 141 \\ &= -573 + 582 = 9. \end{aligned}$$

Next, we compute the Treatment SS :

$$SS \text{ due to } P = \frac{[P]^2}{32} = \frac{(55)^2}{32} = \frac{3025}{32} = 94.53125,$$

$$SS \text{ due to } K = \frac{[K]^2}{32} = \frac{(93)^2}{32} = \frac{8649}{32} = 270.28125,$$

$$SS \text{ due to } D = \frac{[D]^2}{32} = \frac{(25)^2}{32} = \frac{625}{32} = 19.53125,$$

$$SS \text{ due to } PK = \frac{[PK]^2}{24} = \frac{(-26)^2}{24} = \frac{676}{24} = 28.16667,$$

$$SS \text{ due to } PD = \frac{[PD]^2}{24} = \frac{(62)^2}{24} = \frac{3844}{24} = 160.16667,$$

$$SS \text{ due to } KD = \frac{[KD]^2}{24} = \frac{(40)^2}{24} = \frac{1600}{24} = 66.66667,$$

$$SS \text{ due to } PKD = \frac{[PKD]^2}{24} = \frac{(9)^2}{24} = \frac{81}{24} = 3.37500,$$

Treatment SS = sum of SSs due to P, K, D, PK, KD, PD and PKD
 $= 642.71876.$

SS due to error = total SS - block SS - treatment SS
 $= 1,415.21875 - 286.46875 - 642.71876$
 $= 1,415.21875 - 929.18751 = 486.03124.$

TABLE 20.20
**ANALYSIS OF VARIANCE TABLE FOR THE PARTIALLY
 CONFOUNDED 2^3 -EXPERIMENT**

Source of variation	d.f.	SS	MS	F	
Blocks	7	286.46875	40.92411		
P	1	94.53125	94.53125	3.306	
K	1	270.28125	270.28125	9.454	
D	1	19.53125	19.53125	< 1	$F_{.01; 1,17} = 8.40$
PK	1	28.16667	28.16667	< 1	$F_{.05; 1,17} = 4.45$
KD	1	66.66667	66.66667	2.332	
PD	1	160.16667	160.16667	5.607	
PKD	1	3.37500	3.37500	< 1	
Error	17	486.03124	28.59007		
Total	31	1,415.21875		—	

From the above table it is seen that, among the interactions, only interaction PD is significant at the 5% level. The main effect K is also significant at the 1% level.

20.10 A 2^n -experiment in 2^k blocks per replicate

We have considered the case of confounding a 2^n -experiment in 2 blocks (of equal sizes) per replicate. This necessitated the confounding of a factorial effect carrying 1 d.f. in a replicate and this effect is usually the highest-order interaction. If we confound the same effect in all replicates, then we have complete confounding of that effect ; otherwise, we have partial confounding.

Now a 2^n -experiment may also be conducted in 2^k blocks ($k=2, 3, \dots$ and blocks of equal sizes) per replicate. Then each block will receive 2^{n-k} treatment combinations. In each replicate there will be 2^k block totals, giving rise to $(2^k - 1)$ orthogonal block contrasts. These $(2^k - 1)$ orthogonal block contrasts in a replicate will be identical with $(2^k - 1)$ orthogonal treatment contrasts. That is why we say that a 2^n -experiment in 2^k blocks in a replicate

confounds $(2^k - 1)$ factorial effects with blocks. The particular set of $(2^k - 1)$ d.f. that is confounded in a replicate depends on the layout of that replicate. Depending on whether we have the same layout in each replicate or different layouts for different replicates, we have complete or partial confounding, respectively. Of the $(2^k - 1)$ factorial effects that are confounded in a replicate, we may select k factorial effects as we please subject to the restriction that none of these should be a *generalised interaction* of the others included in this set of k effects. The generalised interaction of two effects is the effect that is obtained by combining the letters of the two effects and neglecting a letter if it occurs twice. Thus the generalised interaction of *ABCD* and *BDEF* is *ACEF* and is obtained as follows :

$$ABCDBDEF = AB^2CD^2EF = ACEF.$$

It can be shown that if in a replicate two interactions (say, *ABCD* and *BDEF*) are confounded, then their generalised interaction (in this case, *ACEF*) is also automatically confounded. So in deciding which set of $(2^k - 1)$ factorial effects should be confounded in a replicate, we select k factorial effects (without including any generalised interaction in these k effects) and then the remaining $(2^k - 1 - k)$ factorial effects, which are the generalised interactions of the k effects selected, will be automatically confounded in that replicate. Afterwards we check that no main effects or lower-order interactions are included in these $(2^k - 1)$ confounded effects (if that is possible).

To get the layout of a 2^n -experiment in 2^k blocks in a replicate, we first decide on the factorial effects we want to confound in this replicate. Then we form the *intrablock subgroup* (or *principal block*) of the replicate. It is that block which contains the treatment combination (1) and other $(2^{n-k} - 1)$ treatment combinations, each having an even number of letters (including no letters) in common with each of the factorial effects confounded in that replicate. After obtaining the intrablock subgroup, the other $(2^k - 1)$ blocks of the replicate are obtained one by one by first including a treatment combination which has not appeared in the previous blocks constructed and then combining its letters with the letters of the treatment combinations of the intrablock subgroup and following the rule of rejecting a letter if it occurs twice.

Let us obtain the layout (before randomisation) of a 2^4 -experiment

in 2^3 blocks in a replicate. The $2^2=4$ blocks in the replicate will confound 3 factorial effects. Of these, we can select 2 effects, and the third one which will be their generalised interaction will be automatically confounded. Suppose, e.g., we select ABC and BCD for confounding, then $ABCBCD=AB^2C^2D=AD$ will also be confounded.

Next, we obtain the intrablock subgroup (by taking (1) and the treatment combinations having an even number of letters in common with each of ABC , BCD , AD) and the remaining three blocks.

Intrablock	Block 2	Block 3	Block 4
(1) bc abd acd	a abc bd cd	b c ad abcd	d bcd ab ac

Block 2 is obtained starting with, say, a which is not in the intra-block subgroup and then the other treatment combinations are abc , $aabd=bd$, $aacd=cd$. Then to form block 3 we take, say, b which is not present in either of the first two blocks and then get $bbc=c$, $babd=ad$, $bacd=abcd$. The remaining 4 treatment combinations form block 4. Afterwards, the treatment combinations in a block are randomised in the plots of the block.

The first two columns of the analysis of variance table of the above 2^4 -experiment in r replicates, completely confounding ABC , BCD , AD in each replicate, will be as follows :

Source	d. f.
Blocks	$4r-1$
Treatments	12
Error	$12r-12$
Total	$16r-1$

The block SS is computed from the $4r$ block totals and the grand total.

The treatment SS contains 12 d.f., excluding the 3 d.f. due to the three confounded effects, ABC , BCD and AD . This SS carrying 12 d.f. can be partitioned into 12 orthogonal components in the usual way.

20.11 Factorial experiments in a single replicate

We have seen that more than one replicate is necessary to get an estimate of the experimental error. We have also observed that as the order of an interaction increases, it becomes difficult to interpret the interaction, and also an experimenter is usually interested in the main effects and some lower-order interactions only.

In the case of a 2^n -experiment with a large number of factors, say $n=5$ or 6, and with a single replicate, we can pool some of the high-order interactions, say the 4-, 5- and 6-factor interactions, and use the pooled value to estimate the error of the experiment on the assumption that these high-order interactions are absent (or negligible). This error then can be used to perform tests about the main effects and the lower-order interactions.

In a 2^6 -experiment with a single replicate, we shall have the following components :

	<i>d.f.</i>
Main effects	6
2-factor interactions	15
3-factor interactions	20
Error (pooled 4-, 5- and 6-factor interactions)	22
Total	63

In the confounded case, some of the interaction effects will form the block component. Thus for a 2^6 -experiment in 4 blocks of 16 plots each (and in a single replicate) confounding $ABCD$, $CDEF$ and $ABEF$, the appropriate table will be the following :

	<i>d.f.</i>
Blocks	3
Main effects	6
2-factor interactions	15
3-factor interactions	20
Error*	19
Total	63

20.12 Split-plot design

In field experiments, sometimes a factor has to be applied to a large experimental unit. This is true when the different methods of ploughing or irrigation are to be compared. And in such cases it is

*This error is obtained from pooled 4-, 5- and 6-factor interactions, excluding $ABCD$, $CDEF$ and $ABEF$.

possible to introduce a second factor, which does not require large plots, with a small number of levels into the same experiment, at a little extra cost. This is done by splitting the plots (called whole plots) of the first factor into as many sub-plots as there are levels of the second factor.

A split-plot design with an *RBD* for the first set of treatments (called "the whole-plot treatments") is obtained by allotting the whole-plot treatments at random to the whole plots of a block and then randomising the second set of treatments (called "the sub-plot treatments") to the sub-plots within each whole plot.

The difference between the split-plot arrangement and the ordinary two-factor experiment in an *RBD* is that, while in the former case the randomisation is done separately for the whole plot treatments (to the whole plots of a block) and the sub-plot treatments (to the sub-plots of a whole plot), in the latter case all the combinations of the two factors are allotted at random to the plots of a block.

This enables us to test for the main effects of the sub-plot treatments and the interaction of the whole-plot treatments and the sub-plot treatments more efficiently than the main effects of the whole-plot treatments in a split-plot design. On the other hand, the main effects and the interaction are all tested equally efficiently in the two-factor experiment in an *RBD*.

There is another interpretation of the split-plot design which brings out its similarity with a confounded design. If the sub-plots are considered as plots and the whole plots as blocks, we find that the differences among the whole plots are the same as the differences among the levels of the whole-plot treatments. And so this design may be said to have confounded the main effects of the whole-plot treatments. In this respect, this design violates our recommendation in previous sections that the confounding in factorial experiments should preferably be restricted to higher-order interactions.

Layout

The p levels of the factor A are randomised according to the plan used—an *RBD* or an *LSD* for the factor A . The q levels of the factor B are then randomised inside each whole plot of factor A by dividing each whole plot into q sub-plots. This randomisation is carried out separately for each whole plot of a block (or a square).

Analysis

Suppose we have a factor A at p levels, which are arranged in an RBD using r blocks, and a second factor B at q levels, which are applied to the plots of a block after subdividing each plot into q sub-plots. So there are p whole plots in a block and q sub-plots in a whole plot. The model used is

$$y_{ijk} = \mu + b_i + \tau_j + e_{ij} + \gamma_k + \delta_{jk} + e'_{ijk} \quad \dots \quad (20.18)$$

$$(i=1, 2, \dots, r; j=1, 2, \dots, p \text{ and } k=1, 2, \dots, q),$$

where τ_j , γ_k and δ_{jk} are the fixed effects due to the j th level of A , k th level of B and the interaction between the j th level of A and the k th level of B , respectively, with

$$\sum_j \tau_j = \sum_k \gamma_k = \sum_{\substack{j \\ \text{all } k}} \delta_{jk} = \sum_{\substack{k \\ \text{all } j}} \delta_{jk} = 0.$$

The random components b_i , e_{ij} and e'_{ijk} are independently normal with zero means and respective variances σ_b^2 , σ_e^2 and $\sigma_{e'}^2$. Then the analysis can be done in two stages. At the first stage, we use the analysis of an RBD with p treatments in r blocks, but remembering that each plot value now is based on the total of q sub-plot values. Then the whole plot analysis is as follows :

Source	d.f.	SS
Blocks	$r-1$	$pq \sum_i (y_{i00} - \bar{y}_{000})^2 = SS(\text{Blocks})$
Whole-plot treatments (A)	$p-1$	$rq \sum_j (y_{0j0} - \bar{y}_{000})^2 = SSA$
Whole-plot error (E_I)	$(r-1)(p-1)$	$q \sum_{i,j} (y_{ij0} - \bar{y}_{i00} - \bar{y}_{0j0} + \bar{y}_{000})^2 = SSE_I$
Total between whole plots	$(rp-1)$	$q \sum_{i,j} (y_{ij0} - \bar{y}_{000})^2$

It can be shown that

$$E(MSA) = \sigma_b^2 + q\sigma_e^2 + \phi_1(\tau_1, \tau_2, \dots, \tau_p)$$

$$\text{and} \quad E(MSE_I) = \sigma_e^2 + q\sigma_{e'}^2,$$

where ϕ_1 is zero if $H_{01} : \tau_j = 0$, for all j , is true, otherwise $\phi_1 > 0$. Thus a test for H_{01} is provided by $F = MSA/MSE_I$, which follows an F distribution with $(p-1)$, $(r-1)(p-1)$ d.f.

The next stage of the analysis is the sub-plot analysis within the whole plots :

Source	d.f.	SS
Sub-plot treatments (B)	$q-1$	$rp \sum_k (y_{0k} - y_{00})^2 = SSB$
Interaction (AB)	$(p-1)(q-1)$	$r \sum_j \sum_k (y_{jk} - y_{0j} - y_{0k} + y_{00})^2 = SS(AB)$
Sub-plot error (E_{II})	$p(q-1)(r-1)$	$\sum_i \sum_j \sum_k (y_{ijk} - y_{0jk} - y_{ij0} + y_{00})^2 = SSE_{II}$
Total between sub-plots within whole plots	$rp(q-1)$	$\sum_i \sum_j \sum_k (y_{ijk} - y_{ij0})^2$

Putting both the parts together, we have the following analysis of variance for the split-plot design :

TABLE 20.21
ANALYSIS OF VARIANCE OF A SPLIT-PLOT DESIGN
WITH WHOLE- PLOT TREATMENTS IN r RANDOMISED BLOCKS

Source of variation	d.f.	SS	MS	$E(MS)$	F
Blocks	$r-1$	$SS(\text{Blocks})$	$MS(\text{Blocks})$		
Treatments (A)	$p-1$	SSA	MSA	$\sigma_e^2 + q\sigma_s^2 + \phi_1(\tau, s)$	$F = \frac{MSA}{MSE_I}$
Error (I)	$(r-1)(p-1)$	SSE_I	MSE_I	$\sigma_e^2 + q\sigma_s^2$	
Treatments (B)	$q-1$	SSB	MSB	$\sigma_e^2 + \phi_2(y_k's)$	$F = \frac{MSB}{MSE_{II}}$
Interaction (AB)	$(p-1)(q-1)$	$SS(AB)$	$MS(AB)$	$\sigma_e^2 + \phi_3(x_{jk}'s)$	$F = \frac{MS(AB)}{MSE_{II}}$
Error (II)	$p(q-1)(r-1)$	SSE_{II}	MSE_{II}	σ_e^2	
Total	$rpq-1$	Total SS		—	

If the whole-plot treatments (A) are applied to a $p \times p$ Latin square, then the whole-plot analysis will be that of an LSD. The sub-plot analysis will remain as above with $r=p$.

It can be shown that

$$E(MSB) = \sigma_e^2 + \phi_2(y_1, y_2, \dots, y_q),$$

$$E[MS(AB)] = \sigma_e^2 + \phi_3(x_{jk}'s)$$

and

$$E(MSE_{II}) = \sigma_e^2,$$

where $\phi_2=0$ if $H_{02} : \gamma_k=0$ (for all k) is true, otherwise $\phi_2 > 0$,

and $\phi_3=0$ if $H_{03} : \delta_{jk}=0$ (for all j, k) is true, otherwise $\phi_3 > 0$.

Thus a test for H_{02} is provided by $F = MSB/MSE_H$, which has an F -distribution with $(q-1)$, $p(q-1)(r-1)$ d.f. And a test for H_{03} is given by $F = MS(AB)/MSE_H$, which has also an F -distribution with $(p-1)(q-1)$, $p(q-1)(r-1)$ d.f.

Computational procedure for the analysis of a split-plot design :

- (1) Calculate the rp whole-plot totals :

$$T_{ij0}, \quad i=1, 2, \dots, r \text{ and } j=1, 2, \dots, p.$$

- (2) Calculate the p whole-plot treatment (A) totals :

$$T_{0j0}, \quad j=1, 2, \dots, p.$$

- (3) Calculate the r block totals : $T_{i00}, i=1, 2, \dots, r$.

(4) Calculate the pq totals for the pq A and B treatment combinations : $T_{0jk}, j=1, 2, \dots, p$ and $k=1, 2, \dots, q$.

- (5) Calculate the q sub-plot treatment (B) totals :

$$T_{00k}, \quad k=1, 2, \dots, q.$$

- (6) Calculate the grand total :

$$T_{000} = \sum_{i,j} T_{ij0} = \sum_j T_{0j0} = \sum_i T_{i00} = \sum_{j,k} T_{0jk} = \sum_k T_{00k}.$$

- (7) Calculate T_{000}^2/rpq .

- (8) Calculate raw total $SS = \sum_{i,j,k} y_{ijk}^2$.

- (9) SS (Blocks) = $\sum_i T_{i00}^2/pq - T_{000}^2/rpq$: obtained from (3) and

(7).

- (10) $SSA = \sum_j T_{0j0}^2/rq - T_{000}^2/rpq$: obtained from (2) and (7).

- (11) $SSE_I = \sum_{i,j} T_{ij0}^2/q - T_{000}^2/rpq - SS$ (Blocks) - SSA : obtained

from (1), (7), (9) and (10).

- (12) $SSB = \sum_k T_{00k}^2/rp - T_{000}^2/rpq$: obtained from (5) and (7).

- (13) $SS(AB) = \sum_{j,k} T_{0jk}^2/r - T_{000}^2/rpq - SSA - SSB$: obtained from

(4), (7), (10) and (12).

- (14) Total $SS = \sum_{i,j,k} y_{ijk}^2 - T_{000}^2/rpq$: obtained from (8) and (7).

(15) SSE_H = total SS - SS (Blocks) - SSA - SSE_I - SSB - $SS(AB)$: obtained by subtraction.

The estimates of standard errors for different types of comparison are :

Difference between two whole-plot treatment means : $\sqrt{2MSE_H/rq}$;

Difference between two sub-plot treatment means : $\sqrt{2MSE_H/rp}$;

Difference between two sub-plot treatment means at the same level of the whole-plot treatment : $\sqrt{2MSE_{II}/r}$;

Difference between two whole-plot treatment means at the same or different levels of the sub-plot treatment : $\sqrt{2[(q-1)MSE_{II}+MSE_I]/rq}$.

The ratio of the treatment difference to its standard error for the last type of comparison mentioned above does not follow a *t*-distribution. For an approximate test, see [2].

Advantages and disadvantages

The split-plot design has two errors, of which E_{II} is smaller than E_I . Hence usually, the *B* and *AB* effects will be estimated and tested more precisely than the *A* effects. The main advantage of the design is that often it is possible to introduce the second factor *B*, requiring small experimental material, along with *A* in a split-plot arrangement at little extra cost. If we have a choice for the allocation of factor *A* and factor *B* to the whole plots and split-plots, we shall apply the factor which is more important to the split-plots.

The disadvantages of this design are that the presence of two errors makes the analysis difficult and sometimes the error E_I may be too large.

Although the experimental error for sub-plot treatments and interaction is smaller than that for whole-plot treatment, it can be shown that the average experimental error over all treatment comparisons is the same for a split-plot design and the corresponding factorial experiment in an *RBD*.

Theoretically, the splitting of plots can be continued further. The split-plots may be split into split-split-plots and a third factor (*C*) may be allotted at random inside each split-plot, and so on. Efficiency increases with the decrease of plot-size. However, splitting beyond a stage is practically impossible, and the analysis also becomes complicated as the splitting continues. So repeated sub-division of plots is not carried out too far in practice. There is a variant of the split-plot design in which both factor *A* and factor *B* will be applied to large strips by dividing the experimental field into as many rows as the levels of one factor and as many columns as the levels of the other factor. Then the two factors will be applied at random to the rows and columns. This is helpful when both the factors require large plots.

Ex. 20.5 A variety-manurial experiment was conducted by allotting the three varieties V_1 , V_2 and V_3 at random to the plots of four randomised blocks and then, splitting each plot into four sub-plots, the four manures M_1 , M_2 , M_3 and M_4 were applied at random within each plot. The plan and yield are shown on the next page. Analyse the data to find out if there are any effects due to manure or variety or interaction between variety and manure.

We draw up the block-variety table for obtaining the whole-plot analysis :

Variety	Block				Total
	I	II	III	IV	
V_1	609	450	488	545	2092
V_2	920	870	833	1118	3741
V_3	1067	1072	1093	905	4137
Total	2596	2392	2414	2568	9970

$$\text{Block SS} = \frac{(2596)^2 + (2392)^2 + (2414)^2 + (2568)^2 - (9970)^2}{12} - \frac{99400900}{48}$$

$$= \frac{24882900}{12} - \frac{99400900}{48}$$

$$= 2,073,575 - 2,070,852.08333$$

$$= 2,722.91667.$$

$$\text{Variety SS} = \frac{(2092)^2 + (3741)^2 + (4137)^2 - (9970)^2}{16} - \frac{99400900}{48}$$

$$= \frac{35486314}{16} - 2,070,852.08333$$

$$= 2,217,894.62500 - 2,070,852.08333$$

$$= 147,042.54167.$$

$$\text{Error (I) SS} = \frac{(609)^2 + (920)^2 + \dots + (1118)^2 + (905)^2 - (9970)^2}{4} - \frac{99400900}{48}$$

\leftarrow variety SS - block SS

$$= \frac{8957010}{4} - 2,070,852.08333 - 147,042.54167 - 2,722.91667$$

$$= 2,239,252.5 - 2,220,617.54167$$

$$= 18,634.95840.$$

Field Plan and Yield

	V_1	V_2	V_3
Block I	M_1 94	M_4 440	M_3 250
	M_3 220	M_1 297	M_1 147
	M_3 185	M_2 218	M_3 248
	M_4 110	M_1 112	M_4 275

	V_1	V_2	V_3
Block II	M_1 135	M_4 160	M_4 370
	M_4 290	M_4 95	M_1 140
	M_3 180	M_2 124	M_2 340
	M_3 265	M_1 71	M_2 222

	V_1	V_2	V_3
Block III	M_1 78	M_3 196	M_3 235
	M_3 135	M_4 262	M_3 260
	M_4 130	M_1 155	M_1 115
	M_3 145	M_2 220	M_4 483

	V_1	V_2	V_3
Block IV	M_1 81	M_4 246	M_4 296
	M_3 175	M_3 191	M_3 250
	M_4 175	M_1 145	M_1 12
	M_3 114	M_4 323	M_4 450

Next, to obtain the manure SS and interaction SS, we draw up the variety-manure table :

Variety Manure \	V_1	V_2	V_3	Total
M_1	324	559	412	1395
M_2	665	900	1000	2565
M_3	593	1005	1009	2607
M_4	510	1277	1616	3403
Total	2092	3741	4137	9970

$$\begin{aligned} \text{Manure } SS &= \frac{(1395)^2 + (2565)^2 + (2607)^2 + (3403)^2 - (9970)^2}{12} \\ &= \frac{26902108}{12} - 2,070,852.08333 \\ &= 2,241,842.33333 - 2,070,852.08333 \\ &= 170,990.25. \end{aligned}$$

$$\begin{aligned} \text{Variety } \times \text{Manure } SS &= \frac{(324)^2 + (665)^2 + \dots + (1009)^2 + (1616)^2 - (9970)^2}{4} - \text{manure } SS - \text{variety } SS \\ &= \frac{9813866}{4} - 2,388,884.875 \\ &= 2,453,466.5 - 2,388,884.875 \\ &= 64,581.625. \end{aligned}$$

$$\text{Raw total } SS = 2,500,068.$$

$$\begin{aligned} \text{Total } SS &= 2,500,068 - 2,070,852.08333 \\ &= 429,215.91667. \end{aligned}$$

$$\text{Error (II) } SS = 25,243.62493, \text{ by subtraction.}$$

TABLE 20.22
ANALYSIS OF VARIANCE OF THE SPLIT-PLOT DESIGN

Source of variation	d.f.	SS	MS	F
Blocks	3	2,722.91667	907.63889	
Varieties	2	147,042.54167	73,521.27088	
Error (<i>I</i>)	6	18,634.95840	3,105.82640	
Manures	3	170,990.25000	56,996.75000	
Variety \times Manure	6	64,581.62500	10,763.60416	11.512
Error (<i>II</i>)	27	25,243.62493	934.94907	
Total	47	429,215.91667		

Since $F_{.01; 6,26} = 3.59$ and $F_{.01; 6,28} = 3.53$, we find that the *F* for interaction (which has 6,27 d.f.) is highly significant. So the hypothesis of no interaction effects is rejected at the 1% level. As such, we do not perform the test for main effects of *A* and *B*, and hence the corresponding *F*'s are not shown in the above table.

20.13 Analysis of covariance

This is an extension of the analysis of variance to cover the case where observations are taken on more than one variable from each experimental unit. Interest, however, centres on one of these (*y*, called the dependent variable) and the question is whether the variation of the dependent variable over the classes is due to class effects or due to its dependence on the other variables (*x*'s, called the independent or *concomitant variables*), which also vary from class to class. The analysis of covariance controls the experimental error by taking into consideration the dependence of *y* on *x*.

As simple examples where techniques of the analysis of covariance may be used, we may consider the following :

- (i) The yield of a crop may depend on the number of plants per plot, and we may consider number of plants as the concomitant variable and perform an analysis of covariance.
- (ii) In a study of the effect of drugs or diets on the growth

of animals, the growth may depend on the initial condition (say initial weight) of the animals and an analysis of covariance may be performed.

Analysis

Suppose that observations are taken according to some plan, say a one- or two-way layout, a Latin square or some other design, and that with each observation on y , the dependent variable, we also take observations on each of a number of concomitant variables, x_1, x_2, \dots, x_l . In the analysis of variance model, each y was expressed as the sum of two components—the true value $E(y)$ plus the error. In the analysis of covariance, the $E(y)$ is the sum of two components—one that would be present in an analysis of variance and the second is the linear combination of the values of the concomitant variables with regression coefficients (β 's).

Thus the model in the present case is

$$y_i = (a_{i1}\tau_1 + a_{i2}\tau_2 + \dots + a_{ik}\tau_k) + (\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_l x_{il}) + e_i^*, \quad \dots \quad (20.19)$$

where a_{ij} 's are known, x_{ij} is the value of the j th concomitant variable observed with y_i , β 's are the regression coefficients of y on the concomitant variables and τ 's are the effects (main, interaction, block or other effects) in the corresponding analysis of variance model. e_i^* is the random error component in the analysis of covariance model. For tests of significance, e_i^* 's are assumed to be independently normally distributed with zero means and common variance $\sigma_{e_i}^2$. We shall consider only the case of fixed effects.

The use of a '*' with the error e and variance σ_e^2 , of the corresponding analysis of variance model, is meant to stress the fact that these quantities in the two models need not be the same. The introduction of $\sum_k \beta_k x_{ik}$ in the analysis of covariance model may change their values. While discussing 'local control', we said that one way to control error is by the technique of analysis of covariance. This is because $\sigma_{e_i}^2$ of the analysis of covariance model will be smaller than σ_e^2 of the corresponding analysis of variance model, provided the β 's are not all zeros.

The least-square estimates of τ 's and β 's can be obtained in the usual manner, and the test for a linear hypothesis about any set of

effects can be derived following the procedure outlined under 'Tests of general linear hypotheses' in Section 19.4.

We next consider in detail the analyses for some simple models.

20.13.1 Analysis of covariance for a one-way layout with one concomitant variable

The model here is

$$y_{ij} = \mu_i + \beta(x_{ij} - x_{00}) + e_{ij}^* \quad \dots \quad (20.20)$$

$(i=1, 2, \dots, t; j=1, 2, \dots, r_i),$

where the e_{ij}^* 's are independently normal with zero means and variance σ_e^2 and $x_{00} = \sum_{i,j} x_{ij}/n$, where $n = \sum_i r_i$.

The least-square normal equations for μ_i and β are

$$\sum_i [y_{ij} - \mu_i - \beta(x_{ij} - x_{00})] = 0$$

and

$$\sum_i \sum_j [y_{ij} - \mu_i - \beta(x_{ij} - x_{00})](x_{ij} - x_{00}) = 0.$$

The least-square estimates are

$$\hat{\mu}_i^* = y_{i0} - \hat{\beta}(x_{i0} - x_{00})$$

(this $\hat{\mu}_i^*$ is $\hat{\mu}_i$ of the analysis of variance model minus the adjustment factor $\hat{\beta}(x_{i0} - x_{00})$ due to the introduction of x in the model),

$$\hat{\beta} = \frac{\sum_{i,j} (x_{ij} - x_{i0})(y_{ij} - y_{i0})}{\sum_{i,j} (x_{ij} - x_{i0})^2} = \frac{E_{xy}}{E_{xx}}, \text{ say,} \quad \dots \quad (20.21)$$

where

$$E_{xy} = \sum_{i,j} (x_{ij} - x_{i0})(y_{ij} - y_{i0})$$

and

$$E_{xx} = \sum_{i,j} (x_{ij} - x_{i0})^2.$$

Also, we define $E_{yy} = \sum_{i,j} (y_{ij} - y_{i0})^2$,

$$T_{xx} = \sum_i r_i (x_{i0} - x_{00})^2,$$

$$T_{xy} = \sum_i r_i (x_{i0} - x_{00})(y_{i0} - y_{00})$$

and

$$T_{yy} = \sum_i r_i (y_{i0} - y_{00})^2.$$

It is easy to verify that

$$\sum_{i,j} (y_{ij} - y_{00})^2 = T_{yy} + E_{yy},$$

$$\sum_{i,j} (x_{ij} - x_{00})^2 = T_{xx} + E_{xx}$$

and

$$\sum_{i,j} (x_{ij} - x_{00})(y_{ij} - y_{00}) = T_{xy} + E_{xy}.$$

These give the partitioning of the total SS due to y , the total SS due to x and the total sum of products (SP) of x and y , respectively.

The unrestricted residual SS obtained for the above model is

$$\begin{aligned} SSE^* &= \sum_i \sum_j [y_{ij} - \mu_i^* - \beta(x_{ij} - x_{00})]^2 \\ &= \sum_i \sum_j [y_{ij} - y_{i0} - \beta(x_{ij} - x_{i0})]^2 \\ &= \sum_i \sum_j (y_{ij} - y_{i0})^2 - 2\beta \sum_i \sum_j (x_{ij} - x_{i0})(y_{ij} - y_{i0}) + \beta^2 \sum_i \sum_j (x_{ij} - x_{i0})^2 \\ &= E_{ss} - 2\beta E_{sx} + \beta^2 E_{xx} = E_{ss} - \beta E_{sx}, \\ &= E_{ss} - E_{sx}^2/E_{xx}, \text{ and this has } (n-t-1) \text{ d.f.} \end{aligned}$$

The null hypothesis in the present case is $H_0: \mu_i$'s are all equal, which means that the effects due to the different classes, after considering the dependence of y on x , are the same.

The restricted residual SS (i.e. the residual SS under H_0) is

$$\begin{aligned} (SSE^*)' &= \text{the minimum value of } \sum_i \sum_j [y_{ij} - \mu - \beta(x_{ij} - x_{00})]^2 \\ &\quad \text{when minimised w.r.t. } \mu \text{ and } \beta \\ &= \sum_i \sum_j [y_{ij} - y_{00} - \beta^*(x_{ij} - x_{00})]^2, \text{ and this has } (n-2) \text{ d.f.,} \end{aligned}$$

$$\text{where } \hat{\mu} = y_{00} \text{ and } \hat{\beta}^* = \frac{\sum_i \sum_j (x_{ij} - x_{00})(y_{ij} - y_{00})}{\sum_i \sum_j (x_{ij} - x_{00})^2} = \frac{E'_{sx}}{E'_{xx}}, \text{ say,}$$

are the least-square estimates of μ (common value of μ_i 's) and β under H_0 .

Then it is easy to check that

$$\begin{aligned} (SSE^*)' &= \sum_i \sum_j (y_{ij} - y_{00})^2 - \hat{\beta}^* \sum_i \sum_j (x_{ij} - x_{00})(y_{ij} - y_{00}) \\ &= E'_{ss} - \hat{\beta}^* E'_{sx} = E'_{ss} - E'_{sx}^2/E'_{xx}. \end{aligned}$$

Thus, from the results of Section 19·4, it follows that the appropriate test statistic for testing H_0 is

$$F = \frac{(SSE^*)' - (SSE^*)}{(SSE^*)} \cdot \frac{n-t-1}{t-1};$$

and H_0 is rejected at the level α if

$$F > F_{\alpha; (t-1), (n-t-1)};$$

otherwise, it is accepted.

To put the above material in the form of an analysis of covariance table, we draw up a table similar to an analysis of variance table, but having some additional entries.

TABLE 20.23
ANALYSIS OF COVARIANCE FOR ONE-WAY CLASSIFIED
DATA WITH ONE CONCOMITANT VARIABLE

Source of variation	d. f.	SS_{xx}	SP_{xy}	SS_{yy}	Estimate of β	Adjusted	
						SS_{yy}	d. f.
Classes	$t-1$	T_{xx}	T_{xy}	T_{yy}			
Error	$n-t$	E_{xx}	E_{xy}	E_{yy}	E_{xy}/E_{xx}	SSE^*	$n-t-1$
Total	$n-1$	E'_{xx}	E'_{xy}	E'_{yy}	E'_{xy}/E'_{xx}	$(SSE^*)'$	$n-2$
Difference : Total - Error			—			$(SSE^*)' - SSE^*$	$t-1$

20.13.2 Analysis of covariance for an RBD with one concomitant variable

Now we consider the analysis of covariance for an *RBD*, i.e. for a two-way layout with one observation on y per cell and with one concomitant variable.

We take the model in the form

$$y_{ij} = \mu + \alpha_i + \theta_j + \beta(x_{ij} - x_{00}) + \epsilon_{ij}^* \quad \dots \quad (20.23)$$

$$(i=1, 2, \dots, r; j=1, 2, \dots, t),$$

where α_i , θ_j are the fixed block and treatment effects, β the regression coefficient and x_{ij} the value of the concomitant variable, and ϵ_{ij}^* 's are independently normal, each with mean zero and variance $\sigma_{\epsilon_{ij}}^2$. Also,

$$\sum_i \alpha_i = \sum_j \theta_j = 0.$$

The least-square estimates are

$$\hat{\mu} = y_{00},$$

$$\hat{\alpha}_i^* = (y_{i0} - y_{00}) - \beta(x_{i0} - x_{00}),$$

$$\hat{\theta}_j^* = (y_{0j} - y_{00}) - \beta(x_{0j} - x_{00})$$

$$\text{and } \hat{\beta} = \frac{\sum_{i,j} (x_{ij} - x_{i0} - x_{0j} + x_{00})(y_{ij} - y_{i0} - y_{0j} + y_{00})}{\sum_{i,j} (x_{ij} - x_{i0} - x_{0j} + x_{00})^2}.$$

$$= E_{xy}/E_{xx}, \text{ say.}$$

$\hat{\alpha}_i^*$, $\hat{\theta}_j^*$ are the corresponding estimates in the analysis of variance model minus adjustment factors due to the introduction of x .

The partitioning of the total sum of products of x, y is

$$\sum_{i,j} (x_{ij} - x_{00})(y_{ij} - y_{00}) = t \sum_i (x_{i0} - x_{00})(y_{i0} - y_{00}) +$$

$$r \sum_j (x_{0j} - x_{00})(y_{0j} - y_{00}) + \sum_{i,j} (x_{ij} - x_{i0} - x_{0j} + x_{00})(y_{ij} - y_{i0} - y_{0j} + y_{00})$$

or, symbolically,

$$\text{total } (SP_{xz}) = B_{xz} + T_{xz} + E_{xz}.$$

Similarly,

$$\sum_{i,j} (x_{ij} - x_{00})^2 = t \sum_i (x_{i0} - x_{00})^2 + r \sum_j (x_{0j} - x_{00})^2 + \sum_{i,j} (x_{ij} - x_{i0} - x_{0j} + x_{00})^2$$

$$\text{or } \text{total } (SS_{zz}) = B_{zz} + T_{zz} + E_{zz};$$

$$\text{and } \text{total } (SS_{yy}) = B_{yy} + T_{yy} + E_{yy}.$$

The unrestricted residual SS obtained for the above model is

$$\begin{aligned} SSE^* &= \sum_{i,j} [y_{ij} - \hat{\mu} - \hat{\alpha}_i^* - \hat{\theta}_j^* - \hat{\beta}(x_{ij} - x_{00})]^2 \\ &= \sum_{i,j} (y_{ij} - y_{00})^2 - t \sum_i (y_{i0} - y_{00})^2 - r \sum_j (y_{0j} - y_{00})^2 - \hat{\beta} E_{xz}, \\ &= \text{total } (SS_{yy}) - B_{yy} - T_{yy} - \hat{\beta} E_{xz}, \\ &= (SSE \text{ for RBD}) - \hat{\beta} E_{xz}, \text{ with } (r-1)(t-1)-1 \text{ d.f.} \end{aligned}$$

$\hat{\beta} E_{xz}$, is the reduction in error SS due to the regression of y on x .

Thus

$$SSE^* = E_{yy} - \hat{\beta} E_{xz}, \text{ with } \{(r-1)(t-1)-1\} \text{ d.f.}$$

The null hypothesis to be tested is H_0 : all θ_j 's are equal, which means that the effects due to the treatments after considering the regression of y on x are the same.

The restricted residual SS (i.e. the residual SS under H_0) is

$$(SSE^*)' = \text{minimum value of } \sum_{i,j} [y_{ij} - \mu - \alpha_i - \beta(x_{ij} - x_{00})]^2$$

when minimised w.r.t. μ, α_i 's and β

$$= \sum_{i,j} (y_{ij} - y_{00})^2 - t \sum_i (y_{i0} - y_{00})^2 - \beta^* E'_{xz}, \text{ with } r(t-1)-1 \text{ d.f.},$$

β^* being the least-square estimate of β under H_0 and being given by

$$\beta^* = E'_{xz} / E'_{zz},$$

where

$$E'_{zz} = E_{zz} + T_{zz},$$

$$E'_{yy} = E_{yy} + T_{yy},$$

and

$$E'_{xz} = E_{xz} + T_{xz}.$$

Thus, from the general theory of Section 19.4, it follows that the

appropriate test statistic for testing H_0 is

$$F = \frac{(SSE^*)' - SSE^*}{SSE^*} \cdot \frac{(t-1)(t-1)-1}{(t-1)};$$

and H_0 is rejected if the above $F > F_{\alpha; (t-1), (t-1)(t-1)-1}$; otherwise, H_0 is accepted at the level α .

The corresponding analysis of covariance table is shown below :

TABLE 20.24
ANALYSIS OF COVARIANCE FOR AN RBD
WITH ONE CONCOMITANT VARIABLE

Source of variation	d. f.	SS_{xx}	SP_{xy}	SS_{yy}	Estimate of β	Adjusted	
						SS_{yy}	d. f.
Blocks	$t-1$	B_{xx}	B_{xy}	B_{yy}			
Treatments	$t-1$	T_{xx}	T_{xy}	T_{yy}			
Error	$(t-1)(t-1)$	I_{xx}	I_{xy}	I_{yy}	F_{xy}/E_{xx}	SSE^*	$(t-1)(t-1)$ -1
Treatments + Error	$t(t-1)$	E'_{xx}	E'_{xy}	E'_{yy}	E'_{xy}/E'_{xx}	$(SSE^*)'$	$t(t-1)-1$
Difference : (Treatments + Error) - Error				—		$(SSE^*)' - (SSE^*)$	$t-1$

20.13.3 Analysis of covariance for any complete block design

The computations for any complete block design are the same as those for the RBD. The steps to be followed are :

- (1) Set up the appropriate analysis of covariance table with columns for SS_{xx} , SP_{xy} , and SS_{yy} .
- (2) Compute $SSE^* = E_{yy} - E_{yy}^2/E_{xx}$, with $v_1 = (\text{d.f. for } E_{yy}, -1)$ as d.f.
- (3) Obtain MSE^* .
- (4) Compute ($\text{Treatments} + \text{Error}$) line in the table and obtain E'_{xx} , E'_{xy} , E'_{yy} .
- (5) Obtain $(SSE^*)' = E'_{yy} - E'_{yy}^2/E'_{xx}$, with $v_2 = (\text{d.f. for } E'_{yy}, -1)$ as d.f.
- (6) Obtain $(SSE^*)' - (SSE^*)$ with $d.f. v_2 - v_1 = (t-1)$,
- (7) Obtain $F = \frac{(SSE^*)' - SSE^*}{SSE^*} \cdot \frac{v_1}{(t-1)}$ with d.f. $(t-1)$, v_1 for testing H_0 : all θ_j 's are equal.

Ex. 20·6 An experiment on sugar-cane conducted in four randomised blocks, using plots of size 37' \times 12' each, gave the following values of number of plants per plot (x) and weight of cane in kg. (y). The data on number of plants provide a basis for error control through the analysis of covariance. The three treatments used were :

- Manures : (1) Nitrogen—350 lb./acre as ammonium sulphate— N .
(2) Phosphorous—450 lb./acre as superphosphate— P .
(3) Potash—150 lb./acre as sulphate of potash— K .

PLANT NUMBER (x) AND WEIGHT OF CANE IN KG. (y)
FOR THREE TREATMENTS : N , P AND K

Block	Treatment						Total	
	N		P		K		x	y
	x	y	x	y	x	y		
1	41	122	41	81	42	80	124	283
2	40	120	50	80	38	82	128	282
3	38	138	46	79	54	65	138	282
4	41	121	42	73	40	58	123	254
Total	160	501	179	315	174	285	513	1101

The relevant computations are shown below :

$$T_{xx} = \frac{(160)^2 + (179)^2 + (174)^2}{4} - \frac{(513)^2}{12} = \frac{87917}{4} - \frac{263169}{12}$$

$$= 21,979.25 - 21,930.75 = 48.50.$$

$$B_{xx} = \frac{(124)^2 + (128)^2 + (138)^2 + (123)^2}{3} - 21,930.75$$

$$= 65933/3 - 21,930.75 = 21,977.6667 - 21,930.75 = 46.9167$$

$$\text{Total } (SS_{xx}) = (41)^2 + (40)^2 + \dots + (54)^2 + (40)^2 - 21,930.75$$

$$= 22,191 - 21,930.75 = 260.25.$$

$$T_{yy} = \frac{(501)^2 + (315)^2 + (285)^2}{4} - \frac{(1101)^2}{12} = \frac{431451}{4} - \frac{1212201}{12}$$

$$= 107,862.75 - 101,016.75 = 6,846.00.$$

$$B_{yy} = \frac{(283)^2 + (282)^2 + (282)^2 + (254)^2}{3} - 101,016.75$$

$$= 303653/3 - 101,016.75$$

$$= 101,217.6667 - 101,016.75 = 200.9167.$$

$$\begin{aligned}\text{Total } (SS_{xx}) &= (122)^2 + (120)^2 + \dots + (65)^2 + (58)^2 - 101,016.75 \\ &= 108,509 - 101,016.75 = 7,492.25.\end{aligned}$$

$$\begin{aligned}\text{Total } (SP_{xy}) &= (41 \times 122) + \dots + (40 \times 58) - 513 \times 1101/12 \\ &= 46,418 - 564813/12 = 46,418 - 47,067.75 = -649.75. \\ T_{xy} &= \frac{(160 \times 501) + (179 \times 315) + (174 \times 285)}{4} - 47,067.75 \\ &= 186135/4 - 47,067.75 = 46,533.75 - 47,067.75 = -534.00.\end{aligned}$$

$$\begin{aligned}B_{xy} &= \frac{(124 \times 283) + (128 \times 282) + (138 \times 282) + (123 \times 254)}{3} \\ &- 47,067.75 = \frac{141346}{3} - 47,067.75 = 47,115.3333 - 47,067.75 \\ &= 47.5833.\end{aligned}$$

The SSs and SPs are entered in the following table :

TABLE 20.25
ANALYSIS OF COVARIANCE FOR THE DATA OF EX. 20.6

Source of variation	d. f.	SS_{xx}	SP_{xy}	SS_{yy}	b	Adjusted SS_{yy} d.f.
Blocks	3	46.9167	47.5833	200.9167		
Treatments	2	48.5000	-534.0000	6846.0000		
Error	6	164.8333	-163.3333	445.3333	-0.9909	283.4863 5
Total	11	260.2500	-649.7500	7492.2500		
Treatments + Error	8	213.3333	-697.3333	7291.3333	-3.2668	5011.8902 7
Difference : (Treatments + Error) - Error	.			-		4728.4039 2

Since

$$F = \frac{4728.4039/2}{283.4863/5} = \frac{2364.2019}{56.697} = 41.6987$$

is greater than $F_{0.01; 2,5} = 13.27$, it would seem that there are real treatment differences after adjustment has been made for the differences in the number of plants per plot.

20.13.4 Testing the homogeneity of a group of regression coefficients

Suppose we have p groups of observations on (x, y) . The observations in the i th group may be labelled (x_{ij}, y_{ij}) , for $j=1, 2, \dots, n_i$ and $i=1, 2, \dots, p$. We can then have p regression equations (considering the regression of y on x) as follows :

$$E(y_{ij}) = \alpha_i + \beta_i(x_{ij} - x_{i0}). \quad \dots \quad (20.24)$$

Then, under the assumption that y_{ij} 's are independently normal with $\text{var}(y_{ij}) = \sigma^2$ for all groups, we may be interested in the null hypothesis H_0 : all β_i 's are equal or, in other words, in the hypothesis that the p regression lines are parallel to one another. We shall use the general procedure of Section 19.4 in deriving the test-statistic.

The least-square estimates of α_i 's and β_i 's are

$$\hat{\alpha}_i = y_{i0}, \quad \hat{\beta}_i = \frac{\sum_j (x_{ij} - x_{i0})(y_{ij} - y_{i0})}{\sum_j (x_{ij} - x_{i0})^2} = \frac{B_i}{A_i} = b_i, \text{ say.}$$

Then the unrestricted residual SS is

$$\begin{aligned} S_1^2 &= \sum_i \sum_j [y_{ij} - y_{i0} - b_i(x_{ij} - x_{i0})]^2 \\ &= \sum_i \sum_j (y_{ij} - y_{i0})^2 - \sum_i b_i \sum_j (x_{ij} - x_{i0})(y_{ij} - y_{i0}) \\ &= \sum_i C_i - \sum_i b_i B_i, \text{ say} \\ &= \sum_i (C_i - b_i B_i) \\ &= \sum_i (\text{unrestricted residual SS for } i\text{th group}), \text{ with } \sum_{i=1}^p (n_i - 2) \text{ d.f.} \end{aligned}$$

Next, we obtain the restricted (under H_0) residual SS, which is

$$\begin{aligned} S_2^2 &= \text{minimum value of } \sum_i \sum_j [y_{ij} - \alpha_i - \beta(x_{ij} - x_{i0})]^2 \text{ when minimised} \\ &\quad \text{w.r.t. } \alpha_i \text{'s and } \beta, \text{ where } \beta \text{ is the common value of } \beta_i \text{'s} \\ &\quad \text{under } H_0 \\ &= \sum_i \sum_j [y_{ij} - y_{i0} - b(x_{ij} - x_{i0})]^2 \\ &= \sum_i \sum_j (y_{ij} - y_{i0})^2 - b \sum_i \sum_j (x_{ij} - x_{i0})(y_{ij} - y_{i0}) \\ &= \sum_i C_i - b \sum_i B_i = C_i - b B_i, \text{ say, with } \sum_i (n_i - 1) - 1 \text{ d.f.,} \end{aligned}$$

where the least-square estimates under H_0 of α_i 's and β are

$$\hat{\alpha}_i = y_{i0}, \quad \hat{\beta} = \frac{\sum_i \sum_j (x_{ij} - x_{i0})(y_{ij} - y_{i0})}{\sum_i \sum_j (x_{ij} - x_{i0})^2} = \frac{\sum_i B_i}{\sum_i A_i} = \frac{B_i}{A_i} = b, \text{ say.}$$

Thus the test of H_0 is obtained by using the statistic

$$F = \frac{S_{\bar{s}}^2 - S_1^2}{S_1^2} \cdot \frac{\sum (n_i - 2)}{p-1}$$

$$= \frac{(C_t - bB_t) - (C_t - \sum_i b_i B_i)}{C_t - \sum_i b_i B_i} \cdot \frac{n-2p}{p-1}, \text{ with } (p-1), (n-2p) \text{ d.f.}$$

where $n = \sum_{i=1}^p n_i$.

The above test may be systematically performed with the help of the following two tables :

TABLE 20.26
TABLE OF PRELIMINARY COMPUTATIONS

Group	d.f.	SS_{xx}	SP_{xy}	SS_{yy}	b	Adjusted SS_{yy}	$d.f.$
1	$n_1 - 1$	A_1	B_1	C_1	$b_1 = B_1/A_1$	$C_1 - b_1 B_1$	$n_1 - 2$
2	$n_2 - 1$	A_2	B_2	C_2	$b_2 = B_2/A_2$	$C_2 - b_2 B_2$	$n_2 - 2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
p	$n_p - 1$	A_p	B_p	C_p	$b_p = B_p/A_p$	$C_p - b_p B_p$	$n_p - 2$
Total	$n-p$	$A_t = \sum_i A_i$	$B_t = \sum_i B_i$	$C_t = \sum_i C_i$	$b = B_t/A_t$	$C_t - b B_t$	$n-p-1$

TABLE 20.27
TEST OF SIGNIFICANCE

Source of variation	d.f.	SS	MS	F
Difference : (Total - Within groups)	$p-1$	By subtraction $= S_{\bar{s}}^2 - S_1^2$ $= (S_{\bar{s}}^2 - S_1^2)/(p-1) = MSR$	$(S_{\bar{s}}^2 - S_1^2)/(p-1) = MSR$	$F = \frac{MSR}{MSE}$
Within groups	$n-2p$	$\sum_i (C_i - b_i B_i) = S_{\bar{s}}^2$	$S_{\bar{s}}^2/(n-2p) = MSE$	
Total	$n-p-1$	$C_t - b B_t = S_{\bar{s}}^2$		—

The hypothesis of homogeneity of β_i 's (i.e. H_0) is rejected at the level α if for the data

$$F = \frac{MSR}{MSE} > F_{\alpha; (p-1), (n-2p)}$$

otherwise, H_0 is accepted.

20.13.5 Some facts about analysis of covariance

It is said that the analysis of covariance for increasing the precision of treatment comparisons is valid only if the treatments do not affect the values of the concomitant variables. The adjusted class mean in the case of model (20.20) is estimated by

$$\hat{\mu}_i^* = y_{i0} - \beta(x_{i0} - x_{00}).$$

The effect of the adjustment $\beta(x_{i0} - x_{00})$ is to change y_{i0} to the value that would be expected if there were the same x mean for all classes. So, if x 's are affected by classes, then a part of the class effect will be removed by this adjustment. An F -test of the x -values ($F = MS(T_{xx})/MS(E_{xx})$) gives information on this. If this F is not significant, then the adjusted class differences may be attributed to the different classes. But when the F for x -values is significant, the experimenter should be cautious. For differences in the adjusted class effects may really be due to the dependence of y on x . If, however, the adjusted class effects do not differ significantly, then this may be due to the adjustment which might have cancelled class effects.

We have introduced the component $\beta(x_{ij} - x_{00})$ in the model on the assumption that the x 's do affect the y -values. If one wants to verify this before proceeding with the final analysis, one may do so with the help of the test statistic

$$t = \beta \sqrt{E_{xx}/MSE^*}, \text{ with } d.f. \text{ equal to the } d.f. \text{ of } SSE^*.$$

This is a test for $H_0 : \beta = 0$. So if H_0 is rejected, then we proceed with the analysis of covariance. Otherwise, we do not. For in the latter case an analysis of variance will be appropriate.

If the hypothesis $H_0 : \text{all } \mu_i$'s are equal, is rejected, one can compare all possible pairs to find out which of these differ. For this we need the estimate of the standard error of

$$(\hat{\mu}_i^* - \hat{\mu}_{i'}^*),$$

where $(\hat{\mu}_i^* - \hat{\mu}_{i'}^*) = (y_{i0} - y_{i'0}) - \beta(x_{i0} - x_{i'0})$.

$$\text{Now, } \text{var}(\hat{\mu}_i^* - \hat{\mu}_{i'}^*) = \sigma^2 \left[\frac{1}{r_i} + \frac{1}{r_{i'}} + \frac{(x_{i0} - x_{i'0})^2}{E_{xx}} \right]. \quad \dots \quad (20.25)$$

Even for simple layouts this exact value is different for different pairs of $\hat{\mu}_i^*$, $\hat{\mu}_{i'}^*$ due to the factor $(x_{i0} - x_{i'0})^2$. Finney shows that the average value of this variance for an RBD with model (20.23) is

$$\text{var}(\hat{\theta}_j^* - \hat{\theta}_{j'}^*) = \frac{2\sigma^2}{r} \left[1 + \frac{T_{xx}}{(t-1)E_{xx}} \right]. \quad \dots \quad (20.26)$$

20.14 Missing-plot technique

After conducting an experiment according to some plan, we may find the yields from some of the plots missing. There may be various causes behind missing values, viz. accident, attack of pests, negligence on the part of the observer, or the value may be suspicious so that it is wise to treat it as absent.

The correct procedure, then, is to write down the observational equations for the available observations and to perform a least squares analysis. But this gives rise to normal equations which are difficult to solve owing to the absence of certain observations. Yates considered a method of estimating the missing values, inserting the estimates and analysing the data. The technique of using the estimates of missing values gives results identical with those obtained by the correct procedure.

The general procedure when k values are missing is as follows : Let x_1, x_2, \dots, x_k denote the k missing values. Write down the SSE using the x_i 's and the available data. Then SSE will be a quadratic expression in x_1, x_2, \dots, x_k , say $E(x_1, x_2, \dots, x_k)$. The estimates of the x_i 's are obtained by minimising $E(x_1, x_2, \dots, x_k)$ with respect to the x_i 's. Let $x_1^*, x_2^*, \dots, x_k^*$ be the values that minimise $E(x_1, x_2, \dots, x_k)$. Then $E(x_1^*, x_2^*, \dots, x_k^*)$ is the correct value of the error SS and it has $(\nu_t - k)$ d.f., where ν_t is the error d.f. of the corresponding complete design with no missing values.

The next step is to obtain the minimum value of (\therefore treatment $SS +$ error SS) as a function of x_1, x_2, \dots, x_k , i.e. to minimise $T(x_1, x_2, \dots, x_k) + E(x_1, x_2, \dots, x_k)$. Let $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ be the values that minimise this. Then $T(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k) + E(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$ has $(\nu_t + \nu_e - k)$ d.f., where ν_t is the treatment d.f. for the complete design. The correct value of the treatment SS is obtained as follows :

$$T(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k) + E(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k) - E(x_1^*, x_2^*, \dots, x_k^*),$$

and it has ν_t d.f.

Then to compare the treatment effects, we compute an F with the corrected treatment MS and corrected error MS . We have to assume that all the ν_t treatment contrasts are estimable even when the k values are missing.

An alternative procedure is to perform an analysis of covariance with as many concomitant variables as the number of missing values.

Thus if k values are missing, we introduce k concomitant variables X_1, X_2, \dots, X_k , where X_i takes the value 0 for the i th missing plot and the value 1 for all other plots. For each missing plot, the y -value is taken as 0. Then the analysis of covariance with these values of y and X_1, X_2, \dots, X_k will give the correct value of treatment SS as $(SSE^*)' - SSE^*$ and the correct value of error SS as (SSE^*) .

20.15 Series of experiments

In many experimental situations, it becomes necessary to repeat an experiment over time (for a number of seasons or years) and/or over space (a number of places). This repetition (or replication) of the experiment broadens the scope of the experiment in the sense that our recommendations will be applicable for a number of seasons and/or a number of places. A single experiment performed in one place for only one season will provide recommendations for that place and for that season. It may not be applicable to other places or for other seasons. In the case of agricultural experiments, for example, there may be present treatment \times place interaction and/or treatment \times season interaction. So the results of a series of experiments, performed over different places for different seasons with the same set of treatments, will have wider applicability.

We shall consider the simplest case of repetition of experiments of identical structure over a number of places (it is the same for a number of seasons). Let us consider randomised block experiments with t treatments in r blocks and conducted in p places. The analysis of the experiment in a place is based on the linear model (20.2). Before attempting a combined analysis for the p experiments, it is necessary to perform the analysis for the p places separately (according to Table 20.2) and interpret the results separately. It may be of interest to find out whether differences among the treatments are the same in the different places so that a 'best' treatment may be recommended for all places, or whether different treatments are to be recommended for different places.

We next consider the analysis of the combined experiments considering the places as a random sample from the population of places. The model for this combined analysis may be written as

$$y_{ijt} = \mu + p_t + b_{jt} + \tau_i + c_{it} + e_{ijt} \quad \dots \quad (20.27)$$

where $i=1, 2, \dots, t$; $j=1, 2, \dots, r$; $k=1, 2, \dots, p$.

Here ρ_k is the random place effect ; b_{jk} is the effect of the j th block at place k ; the i th treatment effect at place k has been broken into two components : (i) τ_i , constant for all places, and (ii) c_{ik} , the treatment \times place interaction effect. The assumptions made are that c_{ik} are independently normally distributed with zero mean and variance σ_c^2 for all i and k and that they are also independent of e_{ijk} 's. e_{ijk} 's are also assumed to be independently normally distributed with zero mean and constant variance σ^2 .

The analysis of variance under model (20.27) is given below :

TABLE 20.28
ANALYSIS OF VARIANCE OF SERIES OF EXPERIMENTS
UNDER MODEL (20.27)

Source	<i>d.f.</i>	<i>SS</i>	<i>MS</i>	<i>E(MS)</i>
Places	$p-1$	<i>SSP</i>		
Blocks within places	$p(r-1)$	<i>SSB</i>		
Treatments	$t-1$	<i>SST</i>	<i>MST</i>	$\sigma^2 + r\sigma_c^2 + \rho p\sigma_T^2$
Treatment \times Place	$(t-1)(p-1)$	<i>SS(TP)</i>	<i>MS(TP)</i>	$\sigma^2 + r\sigma_c^2$
Pooled error	$p(r-1)(t-1)$	<i>SSE</i>	<i>MSE</i>	σ^2
Total	$prt-1$			—

σ_T^2 is the variance due to treatments. We test for the absence of treatment \times place interaction using the F statistic $MS(TP)/MSE$, while treatment effects are tested by computing the F statistic $MST/MS(TP)$. If interactions are present, we may use the test for treatment effects to know whether, in addition, there are consistent differences among the treatment effects.

The above F -test for treatments will be vitiated if σ_T^2 is not constant and this will happen if the effectiveness of treatments is not the same from place to place. Similarly, the F -test for interactions will be vitiated if σ^2 is not constant for the different places.

The F -test for treatments will be based on the statistic MST/MSE if the different places are considered fixed.

The SSs and $d.f.$'s for the blocks within places and the pooled error are obtained by pooling the SSs and $d.f.$'s of the blocks and the errors of the p analysis of variance tables for the p places (i.e. from p tables like Table 20.2) used in the preliminary analysis. The other three components of Table 20.28 are obtained in the usual manner by first forming a $p \times t$ table for places and treatments.

The analysis for a series of similar $t \times t$ Latin square designs conducted in p places will be carried out as above with blocks within places with $p(r-1)$ $d.f.$ replaced by two components—rows within places and columns within places each with $p(t-1)$ $d.f.$, and the $d.f.$'s for pooled error and total being replaced by $p(t-1)(t-2)$ and (pt^2-1) , respectively.

Questions and exercises

20.1 Define the following terms which occur in the design of experiments : *treatment, experimental unit, experimental error.*

20.2 What are the three basic principles of design ? Explain them.

20.3 State how the following techniques help to control the error of an experiment :

- (a) the grouping of experimental units into homogeneous blocks,
- (b) confounding and
- (c) the use of concomitant variables.

20.4 How do the size and shape of plots and blocks affect the results of a field experiment ?

20.5 Clearly state the restrictions that are being imposed on the number of treatments and the number of replications of a treatment as we pass from a *CRD* to an *RBD* and then to an *LSD*. Also state how at the sacrifice of flexibility greater control over error is achieved in the above designs.

20.6 Give the layout and analysis of each of the following designs :

- (a) randomised block design, (b) Latin-square design.

20.7 Give an example where you think that a cross-over design will be useful. Write down the analysis of variance table of this design.

20.8 What is a factorial experiment? In what respects is it different from a number of single-factor experiments (this number being equal to the number of factors in the factorial experiment)?

20.9 Define the terms *main effects* and *interaction effects* in relation to a 2^3 -experiment?

20.10 What is a treatment contrast? When are two such contrasts said to be orthogonal? Show that in an *RBD* every block contrast is orthogonal to every treatment contrast. Show that in a 2^3 -experiment the main effects and interaction effects are mutually orthogonal. How would you obtain the SS due to a main effect or an interaction effect in a 2^3 -experiment?

20.11 Give in detail the analysis of a 2^3 -experiment conducted in randomised blocks.

20.12 Give the expression for the total effect, the main effect, SS due to an effect and the standard error of an effect for a 2^n -experiment.

20.13 What is meant by confounding in a factorial experiment? Why is confounding used even at the cost of loss of information on the confounded effects? Explain the terms *complete confounding* and *partial confounding*.

20.14 Give in detail the analysis of a partially confounded 2^3 -experiment. Give the expressions for the standard errors of the unconfounded and the confounded effects.

20.15 Obtain an appropriate system of confounding in a 2^5 -experiment in 2^2 blocks and obtain the intra-block sub-group.

20.16 Indicate the analysis of a 2^n -experiment in a single replicate. (For detailed discussion, consider a 2^4 -experiment.)

20.17 What is a split-plot design? Why is it said that this design confounds main effects? Give the analysis of this design.

20.18 Illustrate the use of the technique of analysis of covariance in reducing error as it is applied to the *RBD*.

20.19 How would you test for the homogeneity of a number of regression coefficients by an analysis of covariance?

20.20 Obtain the layouts of the following designs:

(a) A *CRD* with three treatments, *A*, *B* and *C*, the replication numbers being 6, 5 and 10, respectively.

(b) An *RBD* with five treatments in four blocks.

(c) A 6×6 *LSD*.

(d) A 2^3 -experiment in which the highest-order interaction is completely confounded.

(e) A split-plot design, with five levels of the whole-plot treatment and three levels of the sub-plot treatment, in two replicates.

20.21 Show that for Table 20.24,

$$(SSE^*)' - (SSE^*) = \left(T_{ss} - \frac{T_{ss}}{T_{ss}} \right)^2 + \left(\frac{T_{ss}}{T_{ss}} - \frac{E_{ss}}{E_{ss}} \right)^2 \cdot \frac{T_{ss} E_{ss}}{E_{ss}}.$$

Interpret the significance of the two components on the right-hand side of the above relation.

20.22 Consider an RBD with one missing value. Perform the appropriate analysis using Yates' technique for missing values.

[Hint : Let x represent the missing value and let B' , T' be the totals of the block and the treatment for which this value is missing. Then

$$x^* = \frac{rB' + tT' - G'}{(r-1)(t-1)} \text{ and } \bar{x} = \frac{B'}{t-1},$$

where G' is the grand total for the available $(rt-1)$ plot yields.]

20.23 The following data were obtained from an experiment using the treatments : 0.32% of Blitox, 0.16% of Dithane z-78, 0.09% of Brestan-60 and control. After sowing rhizomes of the mat-grass *Cyperus tagetum Roxb.* in four plots in each of three villages, the above four treatments were applied at random to the plots in a village after 30 days of sowing. The yields in gm. of 1 sq. feet cuttings per plot after 120 days are given below. Analyse the data to find out if there are any significant treatment effects.

YIELD (IN GM.) OF 1 SQ. FT. CUTTINGS AFTER 120 DAYS

Treatment	I	Village	III
		II	
Blitox	678.2	510.2	531.2
Dithane z-78	703.2	689.5	611.2
Brestan-60	736.8	574.2	573.7
Control	556.4	510.2	500.0

Partial ans. $F=6.914$.

20.24 A 4×4 Latin-square experiment was conducted to compare the effects of four spacings, A , B , C and D , on the yield of millet. The plan and yields are given below :

Rows	Columns			
	1	2	3	4
1	A 231	B 280	C 285	D 284
2	B 284	A 246	D 283	C 271
3	C 275	D 282	A 258	B 258
4	D 259	C 271	B 289	A 275

Test whether the different spacings are equally effective ; and in case they are not so, compare the spacings pairwise.

Partial ans $F=2.285$.

20.25 The following table gives the plan and the yields of a manuriel experiment involving three factors N , P , K , each at two levels :

	Block 1				Block 2			
	o	fk	nk	np	k	p	n	npk
Replicate 1	145	191	300	240	189	272	160	317
Replicate 2	226	159	240	182	266	300	233	278
Replicate 3	186	173	170	213	209	93	224	245
Replicate 4	182	175	156	183	293	226	248	269

Analyse the data and write a report.

Partial ans. $[N]=385$, $[P]=293$, $[K]=23$, $[NK]=342$,
 $[PK]=-91$, $[NP]=-183$, $[NPK]=-119$
and $MSE=2,194.8483$.

20.26 The following data relate to the yields of an experiment in 2 replications of 5 varieties of corn, each in three generations. For each replicate a randomised block of 5 plots was used, with all the three generations of each variety being accommodated in three subplots of a single plot.

Block I

Variety Number				
3	2	1	4	5
^a 50	^a 48	^a 40	^c 45	^b 50
.....
^c 48	^b 46	^c 48	^a 46	^a 48
.....
^b 45	^c 42	^b 46	^b 48	^c 45

Block II

Variety Number				
4	3	1	5	2
^c 48	^a 45	^b 43	^b 46	^c 41
.....
^a 50	^b 46	^c 51	^a 49	^a 50
.....
^b 40	^c 41	^a 45	^c 41	^b 46

Analyse the data completely to test for the differential effect of generations and their interaction with varieties.

Partial ans. $F(\text{interaction})=2.067$, $F(\text{generations})=1.3049$, $F(\text{varieties}) < 1$.

20.27 In the year 1964-65, in each of the three villages of a district of West Bengal, four treatments were applied at random to four plots 90 days after sowing the plots with rhizomes of the mat-grass *Cyperus tagetum Roxb.* The yield (y) in gm. of 1 sq. ft. cuttings after 120 days and infection value (x) after 90 days were recorded for each plot. Analyse the data given below, to find out whether the treatments had any effect, after eliminating the dependence of y on x .

**YIELD (y) AND INFECTION VALUE (x) FOR THE
4 TREATMENTS IN 3 VILLAGES**

Treatment	Village		
	I	II	III
Blitox	$x=7.8$ $y=642$	$x=11.3$ $y=695$	$x=11.9$ $y=730$
Dithane z-78	$x=3.1$ $y=722$	$x=5.6$ $y=757$	$x=5.2$ $y=738$
Brestan-60	$x=6.1$ $y=762$	$x=4.7$ $y=767$	$x=8.3$ $y=759$
Control	$x=3.3$ $y=625$	$x=11.2$ $y=643$	$x=12.4$ $y=655$

Partial ans $F=12.06$, with d.f. 3, 5.

SUGGESTED READING

- [1] Anderson, R. L. and Bancroft, T. A. *Statistical Theory in Research* (Chs. 18, 20, 21). McGraw-Hill, 1952.
- [2] Chakravarti, I. M., Laha, R. G. and Roy, J. *Handbook of Methods of Applied Statistics*, Vol. II (Chs. 5, 7, 9 and Section 6.8). John Wiley, 1967.
- [3] Cochran, W. G. and Cox, G. M. *Experimental Designs* (Chs. 1—7, 14). Asia Publishing House, 1959.
- [4] Federer, W. T. *Experimental Design* (Chs. 1—7, 9 10, 14, 16). Macmillan, 1963, and Oxford & I.B.H , 1967.
- [5] Fisher, R.A. *The Design of Experiments*. Oliver and Boyd, 1947.
- [6] Goulden, C. H. *Methods of Statistical Analysis* (Chs. 5, 9, 11, 12). Asia Publishing House, 1959.
- [7] Kempthorne, O. *The Design and Analysis of Experiments* (Chs. 1—3, 5—11, 13—15, 28). John Wiley, 1965, and Wiley Eastern.
- [8] Mann, H. B. *Analysis and Design of Experiments*. Dover, 1949.
- [9] Nandi, H. K. "Analysis of covariance", *Calcutta Stat. Assocn. Bulletin*, 4, No. 14, pp 79-82, Dec. 1951.
- [10] ——"Analysis of serial experiments", *Calcutta Stat. Assocn. Bulletin*, 5, No. 17, pp. 43-46, Oct. 1953.
- [11] Quenouille, M. H. *The Design and Analysis of Experiment* (Chs. 1—3). Charles Griffin, 1953.

- [12] Scheffé, H. *The Analysis of Variance* (Ch. 6). John Wiley, 1961.
- [13] Steel, R. G. D. and Torrie, J. H. *Principles and Procedures of Statistics* (Chs. 6—8, 11, 12, 15). McGraw-Hill, 1960.
- [14] Yates, F. *The Design and Analysis of Factorial Experiments* (Chs. 1—4, 16). Imperial Bureau of Science, Tech. Com. No. 35, 1937.

21

DESIGNS OF SAMPLE SURVEYS

21.1 Introduction

The use of sampling in making inferences about an aggregate (or *population*) is possibly as old as civilisation itself. When one has to make an inference about a large lot and it is not practicable to examine each and every individual from the lot, one invariably takes recourse to sampling ; i.e., one examines only a few members of the lot and, on the basis of this sample information, one makes decisions about the whole lot. Thus, if a person wants to purchase a basket of oranges, he will examine one or two oranges from the basket and on that basis make his decision about the whole basket. This is *inductive inference*, i.e. inference about the whole from a knowledge of a part.

Thus we see that most of our enquiries in practice are *sampling enquiries*. Sampling may become inevitable because we may have limited resources in terms of money and/or man-hours, or it may be preferred because of practical convenience. In many cases we shall find that a complete count or complete *census* is practically inconvenient or impossible.

Sampling enquiries may be broadly classified into two groups :

(a) The enquiries which can be answered by conducting a sampling experiment, suitably designed or controlled by the experimenter. Thus, if we want to know which of 5 given varieties of rice is expected to give the maximum yield in the long run or whether a new drug is more effective than an old drug in curing a disease, we have to conduct an experiment with a sample of experimental plots or with a sample of patients, suitably controlled, and we can base our conclusions upon the experimental data. This group of experiments has been discussed in detail in Chapter 20, under the heading *Designs of experiments*.

(b) The enquiries which can be answered by conducting a *sample survey*. Here the individuals to be sampled occur in nature and cannot be subjected to any experimental control. Individuals

are sampled as they appear in nature and the required information is obtained from them. Thus, if we want to know the extent of unemployment or the expenditure pattern amongst the middle-class families in Calcutta, we have to conduct a sample survey. Here also we shall encounter the problem of planning or designing the sample survey suitably, as regards the method of sampling, size of sample, etc., so that at a given level of cost (in terms of money or man-hours) maximum accuracy may be attained in making decisions. This is the group of sampling enquiries with which we are concerned in this chapter.

21.2 Basic principles of sample surveys

The two basic principles for the design of a sample survey are (1) *validity* and (2) *optimisation*. The principle of optimisation takes into account the factors of (a) *efficiency* and (b) *cost*.

By validity of a sample design we mean that the sample should be so selected that the results could be interpreted objectively in terms of probability. In other words, valid tests or estimates about the population characteristics must be available. The principle will be satisfied by selecting a so-called *probability sample*, which ensures that there is some definite, preassigned probability for each individual of the aggregate to be included in the sample.

Efficiency is measured by the inverse of the sampling variance of the estimator. Cost is measured by expenditure incurred in terms of money or man-hours. The principle of optimisation ensures that a given level of efficiency will be reached with minimum cost or that the maximum possible efficiency will be attained with a given level of cost. More generally, we can introduce the idea of a loss function, associated with the difference $T - \theta$ (where θ is the population characteristic to be estimated and T is the estimate based on the sample information) and the cost of sampling, and define an optimum design to be one for which the expected loss is a minimum.

Thus designing a sample survey in a particular situation involves the following problems : (1) determination of the type of sampling and (2) determination in an optimum manner of the details of the survey.

The first problem is solved from two considerations : (a) convenience—both in the identification of sample individuals and collection of data and in the analysis of sample data—and (b) efficiency, as measured by the inverse of the sampling variance. We would choose that particular type of sampling which would ensure maximum efficiency with a given level of expenditure.

The second problem is solved by expressing the cost and variance of the estimator as functions of F_1, F_2, \dots, F_p , the free or flexible variables determining the details of the sample design. The cost function $C(F_1, F_2, \dots, F_p)$ and the variance function $V(F_1, F_2, \dots, F_p)$ are determined theoretically or empirically. C or V may involve certain unknown characteristics which must be determined by taking a so-called small-scale *pilot survey* previous to the final survey. Finally, one determines the optimum values of F_1, F_2, \dots, F_p by Lagrange's method of undetermined multipliers.

21.3 Advantages of sample survey over complete census

Sampling is the selection of a part of an aggregate of material or population to represent the whole *population*. The part of the population selected by sampling is called a *sample*. It is from the sample that we make inferences about the population in which we are interested.

A survey carried out on a properly selected representative sample is called a *sample survey* or *sample census*, as opposed to a *complete enumeration* or *complete census*, in which the whole population is enumerated or surveyed.

In many cases, we undertake a sample survey in preference to a complete census because of the following considerations :

(1) There is a *reduction of cost* either in terms of money or in terms of man-hours. Although the cost per individual may be larger in a sample survey, the total cost is expected to be smaller. In many cases our resources may be limited or it may be necessary that the results of the survey should be available within a specified time limit. In such cases it is imperative to adopt a sample survey rather than a complete census.

(2) There is generally *greater scope* in a sample survey than in a census. Some enquiries may require highly trained personnel or specialised equipment for collection of data, thus making a census practically inconceivable. Thus in a sample survey we may have greater coverage both in respect of the information collected and in respect of the geographical, demographic or other boundaries taken into account.

(3) A sample survey generally gives data of a *better quality* than a complete census, because in a sample survey it may be possible to employ better-trained personnel, effect better supervision or use better equipment than is possible or feasible in a complete census. The errors in the estimates due to sampling are likely to be more than compensated by better control of non-sampling errors. (Sources of different types of non-sampling errors will be discussed in some detail in Section 21.5.)

(4) What is more important, there is no way of *gauging the magnitude of (non-sampling) errors* to which the estimates from a complete census may be subject. On the other hand, a properly designed sample will itself give an idea of the magnitude of the sampling errors involved in its estimates.

(5) It should also be remembered that in some cases a complete census is ruled out by the nature of the population. If there is a population which is *infinite* and/or *hypothetical*, like the population of all the throws that may be made with a coin, sampling is the only course available. Again, if the *enumeration* is by its nature *destructive*, e.g. when we want to know the average life in hours of a type of electric bulb, one must have recourse to a sample and a rather small sample at that.

However, when time and cost are not important factors for consideration or when detailed information is wanted for all the sub-classes into which the population may be divided or when the population size is not large, a complete enumeration may be more appropriate than any sampling procedure. But even in such situations, sampling methods may be used concurrently to get advance information well ahead of the processing of the complete enumeration data as well as to assess the quality of these data.

21.4 Different steps in a large-scale sample survey

Conducting a large-scale sample survey involves three main stages : (a) planning stage, (b) execution stage and (c) analysis and reporting stage.

The *planning stage* consists of the following steps :

(1) Defining the objectives—The objectives of the survey must be clearly stated. Some of the objectives may be immediate and some far-reaching. Along with the objectives, the planner should take into account the available resources in terms of money and man-power, the time-limit within which the survey results must be available and the accuracy desired in the set of estimates to be prepared.

(2) Defining the population—The *population* or the aggregate of individuals to which the survey results would apply must be clearly and unambiguously defined. The geographical, demographic and other boundaries of the population must be specified so that no ambiguity arises regarding the coverage of the survey.

(3) Determination of the data to be collected—This must be done in conformity with the objectives of the survey. Once the nature of the data to be collected in the survey is decided upon, one must prepare a *questionnaire* or *schedule of enquiry*. Generally, a draft questionnaire is first prepared and tried over a number of individuals to discover any ambiguity or defect in framing the questions. The questionnaire is revised and finalised in the light of the trial data. The questions should be brief, practical and as objective as possible, and they must not leave much scope for guessing on the part of the interviewer.

(4) Deciding on the method of collection of data—One must decide whether the *interview method* (i.e. house-to-house enquiry for the collection of data) or the *mail questionnaire method* (i.e. mailing of the questionnaires to individuals of the population for filling in and returning them) is to be adopted. Although the latter method is less costly, there is in it large scope for non-response and it is only practicable amongst educated people interested in the particular survey. In the cases where the information is to be collected by observation, one must decide upon the method of measurement—eye estimation

or exact measurement, the type of equipment or instrument to be used and similar other things.

(5) **Choice of sampling unit (s.u.)**—The *sampling unit* is the ultimate unit to be sampled for the purpose of the survey. Thus, in an agricultural yield or acreage survey, one must decide whether a village or a plot or a small cut in a plot is to be the ultimate s.u., or in a socio-economic survey, whether a family or a member of a family is to be the ultimate s.u. In some surveys the ultimate s.u. may be sampled by a number of stages. In these cases there would be a hierarchy of s.u.'s. Once the s.u. is decided upon, one must see whether a *sampling frame*, i.e. a complete list of the s.u.'s included in the population, is available. If it is available, it must be scrutinised to see whether it is adequate, complete, accurate, up-to-date and not subject to any duplication. On the basis of the scrutiny, the frame must be corrected if this is necessary. If a frame is not available, naturally one must prepare a frame before one can actually draw the sample.

(6) **Designing the survey**—This is the most important step in planning a sample survey. Designing a survey means

- (i) deciding whether unrestricted random sampling or a variant of that should be used in the survey under consideration;
- (ii) choosing the flexible variables in the sample, if any, in an optimum manner; and
- (iii) deciding upon the details of a pilot or exploratory survey, if one is necessary for the main design.

The design should take into account the available resources and the time-limit, if any, besides the accuracy desired. Any relevant practical considerations should also be taken into account.

(7) **Drawing the sample**—The technique of random sampling, involving the use of a random sampling number series or some other random method, will be discussed in detail in a later section.

(8) **Training of personnel**—The investigators and supervisors should be adequately trained for the job before the survey is actually undertaken.

The *execution stage* involves the identification of the sampled individuals in the field and the filling in of the questionnaires.

The *analysis and reporting stage* again consists of the following steps :

(1) Scrutiny of data—The filled-in questionnaires should be carefully scrutinised to find out whether the data furnished are plausible and whether data on different items are mutually consistent. If doubt or suspicion arises on any questionnaire, it should be sent back to the field staff for re-survey.

(2) Tabulation of data—Whether hand-tabulation or machine-tabulation is to be taken recourse to depends upon the quantity of data. For a large-scale survey involving several thousands of individuals, machine-tabulation is expected to be more economical and quicker. Use of *code numbers* for qualitative characters is essential for machine-tabulation. Data for a questionnaire are to be transferred to punched cards, and sorters and tabulators are to be used for obtaining a set of *primary tables*.

(3) Statistical analysis—The primary tables may be further utilised for deriving necessary estimates for population characteristics or for testing hypotheses, if any. Some tables may be derived from the primary tables to bring to light some special features of the data.

(4) Reporting—The report should incorporate a detailed statement regarding all the stages of the survey and should present all the statistical information collected in a neat tabular form. The data should be properly interpreted, conclusions derived and recommendations made. The technical aspects of the design of the survey, e.g. the types of estimators used and their margins of error, may be presented in a separate chapter.

(5) Storing of information for future surveys—At the completion of the survey, arrangements should be made for proper storing of the information for possible use in designing future surveys.

21.5 Biases in surveys

Almost all subjective sampling methods, where a good deal of choice is left to the sampler, give rise to biases of some form or other. The following are the main type of biases in surveys :

I. Procedural biases—These are common to both complete enumeration and sampling methods. The following are the different types of procedural biases :

(i) Response biases—These biases have their origin in the responses furnished by the respondents ; for example, wrong answers arising from pride, called *prestige bias*, by virtue of which one may over-state one's education or occupation or under-state one's age. Also, one may give wrong answers for the protection of self-interest, e.g. may make an under-statement of income or production and an over-statement of expenses. Response bias may also be due to preference for certain figures. For example, in the age returns individuals usually are found to have preferences for even numbers or for multiples of 5.

(ii) Observational biases—Where the variate value is obtained by observation, psychological factors may sometimes influence the returns given. For example, in eye-estimating the yield-rate or the crop-condition factor, the crop-reporter almost invariably under-estimates, whereas eye-estimation of acreage generally results in over-estimation.

(iii) Biases arising from non-response—Non-response may arise if the respondent is not found at home even after repeated calls, or if he either refuses or fails to furnish the information. Since non-response leads to a section of the population with certain peculiar characteristics being excluded, it generally results in biases of some form.

(iv) Interviewer biases—Answers given by suggestions from the interviewer may be influenced by the interviewer's beliefs and prejudices in interpreting some questions.

II. Sampling biases—These are the biases that have their origin in sampling and are absent in complete enumeration. The following types may be distinguished :

(i) Bias due to defective sampling technique—Purposive or judgment sampling, in which the sampler tries deliberately to choose a representative sample, has been found generally to involve some bias. If a proper random process is not strictly adhered to, the investigator may allow his desire to obtain a certain result to influence his selection. For example, for getting a sample of wheat plants growing in a field, it might be thought that a satisfactory method would be to throw a hoop in the air at random and to select all the plants over which it fell. But this might give biased results since the hoop might tend to be caught by the taller ears of wheat.

(ii) Bias due to substitution—Investigators often substitute one convenient member of the population when difficulties arise in enumerating another. Thus in a house-to-house enquiry, the next house may be chosen when there is no reply from the one that was originally to be included in the sample. This will necessarily lead to an over-representation of the types that are occupied all day.

(iii) Bias due to faulty demarcation of s.u.'s (*border effect*)—In area surveys, the location of areas by means of a pair of random co-ordinates, though theoretically ensures a random sample, will in practice do so only if the field work is done with complete objectivity. In a crop-cutting survey, e.g., there may be an inclination on the part of the investigator to include some good plants in the sample, thus resulting in over-estimation. The amount of bias, however, decreases if we take relatively large areas since the errors in the demarcation of boundaries become of decreasing importance as the size of the unit is increased.

(iv) Constant bias due to wrong choice of the statistic—For example, in estimating the population variance with a sample of independent observations, the sample variance is a biassed statistic,

whereas $s'^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ is unbiased.

21.6 Technique of random sampling

The technique of random sampling is of fundamental importance in the application of statistics, for the whole of sampling theory is based on the assumption of random sampling, the essence of which is that each of the individuals included in the population has an equal chance of being selected.

The first attempt towards drawing a random sample may be made by lottery. This is done by constructing a miniature population which can be easily handled and then drawing individuals from it, each time shuffling it thoroughly before the next drawing is made. In practice, a ticket may be prepared for each sampling unit bearing its identification mark, say by putting on it its serial number, and these tickets may be placed in similar containers, usually small metallic cylinders, and thrown into a rotating drum in which they are thoroughly mixed or randomised before each drawing. Similarly,

we can draw a random sample of houses by taking a pack of cards, as similar as possible, making one card correspond to one of the houses by writing on it the number of the house in the street, and then drawing a sample of cards, each time shuffling the cards before the next drawing is made.

But it should be realised that these methods lack the property of strict randomness. First, it is not practically possible to have cards or cylinders of exactly similar shape, size and weight. Secondly, the writing of numbers with ink may weight the cards differentially. Furthermore, the practical difficulties in preparing such a miniature population, when the population size is large, are immense, and lack of care may often lead to non-random samples.

These difficulties can be overcome if we have a series of random numbers (i.e. a series in which the digits 0, 1, 2,, 9 occur randomly). The problem of constructing the miniature population will then reduce to attaching to each unit of the population an ordinal number. We can then choose a number of digits from any part of the series which is already randomised and hence get a random sample. It is this possibility that has led to the construction of random sampling number series.

Definition of a random sampling number series

A random sampling number series is an arrangement, which may be looked upon either as linear or as rectangular, in which each place has been filled in with one of the digits 0, 1,, 9. The digit occupying any place is selected at random from these ten digits and independently of the digits occurring in other positions.

Advantages of random sampling numbers

If we use random sampling numbers for drawing random samples, we need not construct a miniature population. Also, the numbering of the sampling units can be done in any convenient manner.

Secondly, randomisation of the numbers being done once for all, the tedious process^s of randomisation of the miniature population (viz. through shuffling, rotating, etc.) each time before the next drawing is made is not necessary. Any part of the series can be used for a random sample of numbers and the problem is simply to interpret these numbers in terms of individuals of the population.

Lastly, a random sampling number series can be used for any enumerable population, so that a series of random numbers has a wide range of application.

Different sets of random sampling numbers and their construction

Mention may be made of four different sets of random sampling numbers :

1. Tippett's series (*Tracts for Computers*, No. 15. Cambridge University Press), comprising 41,600 numbers arranged in fours.

These numbers were obtained by taking down digits from census reports in a random manner.

2. Fisher and Yates' series (in *Statistical Tables for Biological, Agricultural and Medical Research*), comprising 15,000 digits arranged in twos.

Fisher and Yates obtained their random numbers from the 15th to the 19th digits of Thompson's 20-figure logarithmic tables. In choosing from these digits, an element of randomness was introduced by using playing cards for the selection of half pages of the tables and of a column (of 50 digits) between the 15th and the 19th and, finally, for allotting these digits to the 50 places in a block.

3. Kendall and Smith's series (*Tracts for Computers*, No. 24. Cambridge University Press), comprising 100,000 digits grouped in twos and fours and in 100 separate blocks of 1,000 digits. 5 out of these 100 blocks are indicated as unsuitable for sampling requiring fewer than 1,000 digits.

The authors used a specially constructed machine—a refined version of the common roulette wheel used in gambling.

4. *A Million Random Digits* by Rand Corporation (Free Press, Illinois). This series has been obtained through a mechanical device which is similar to that of Kendall and Smith, but in which further technical improvements have been incorporated.

Tests applied to random sampling number series

To examine whether any series is really random, the following tests may be applied. The tests may be applied to the whole series or any part of it, because a set of numbers may be perfectly random when considered as a part of the whole series, but may not be so when considered as a part of a certain block of the series.

(1) *Frequency test* : Here the observed frequencies of the 10 digits from 0 to 9 are obtained and tested against the expected frequencies on the basis of the hypothesis that the set of numbers is random. The appropriate statistic is a Pearsonian χ^2 with 9 d.f.

(2) *Serial test* : Here the series is considered to be composed of two-digit numbers. The frequencies of all the 100 possible numbers, viz. 00, 01,, 99, are obtained and the hypothesis of randomness is tested by using the appropriate Pearsonian χ^2 with 99 d.f.

(3) *Gap test* : In this test, we first pick out the successive zeros (or the successive occurrences of any other digit) and find the gaps between them. The frequencies of such gaps are obtained and the hypothesis of randomness is tested by using an appropriate Pearsonian χ^2 .

(4) *Poker test* : Consider here the series to be made up of four-digit (or five-digit) numbers. There are five possibilities, viz. all 4 digits same ; 3 digits same and 1 different ; 2 digits same and 2 different ; 2 groups, each of 2 identical digits ; and all 4 digits different. The frequencies of all these five types are obtained and the hypothesis of randomness of numbers is tested by an appropriate Pearsonian χ^2 with 4 d.f.

The tables in common use have been found satisfactory in the light of the above tests.

The use of random sampling numbers for drawing a random sample from a population may be illustrated with the following example :

Ex. 21.1 Draw a random sample of size 10 without replacements from a population of 121 boys numbered from 1 to 121.

We shall take three-digit numbers from the table of random numbers in the Appendix row-wise from the beginning of the 6th line of the first page. To ensure equal probability for each individual, we shall take numbers from 001 to 968 (the greatest three-digit multiple of 121) and shall ignore other three-digit numbers. We shall divide the number by 121 and take the remainder. The remainder, of course, varies from 000 to 120. The remainders 001 to 120 will correspond to the boys with the same numbers, whereas the remainder 000 will correspond to the 121st boy. Since the sampling

is without replacements, a boy once selected cannot be selected again. The selection is done in a tabular form as shown below :

TABLE 21.1
SHOWING THE SELECTION OF A RANDOM SAMPLE OF
SIZE 10 WITHOUT REPLACEMENTS FROM A
POPULATION OF 121 BOYS

Number taken from the table	Remainder when divided by 121	The number of the boy selected
991	Rejected	—
734	008	8
905	058	58
593	049	49
257	015	15
743	017	17
480	117	117
971	Rejected	—
258	016	16
019	019	19
436	073	73
376	013	13

(Alternatively, we could add 1 to the remainder and make it correspond to the serial number of the boy to be selected)

21.7 Types of population and types of sampling

In the first place, the population may be *finite* or *infinite*. By a finite population we shall mean a population which contains a finite number of members. Such, for instance, is the population of heights of 500 boys in a college, or the population of books in a college library. Similarly, by an infinite population we shall mean a population containing an infinite number of members. Such, for instance, is the population of pressures at various points of the atmosphere or of yields of a particular crop at various points in an agricultural field. In many cases the number of members in a population is so large as to be practically infinite, e.g. the human population of India or the population of visible stars.

Secondly, the population may be *existent* or *hypothetical*. The population of concrete existent objects will be called an existent population. But the population may also be hypothetically constructed; e.g., the outcome of the tossing of a coin an infinite number of times represents a hypothetical population of heads and tails. Here the population is to be conceived of as having no existence in reality, but only in imagination.

Sampling is first broadly classified as *subjective* and *objective*. Any type of sampling which depends upon the personal judgment or discretion of the sampler himself is called subjective. But the sampling method which is fixed by a sampling rule or is independent of the sampler's own judgment is objective sampling. Any haphazard or deliberate selection will result in subjective sampling. The main difficulty with subjective sampling is that the sampler is ignorant of the degree of representativeness of his sample or the accuracy of the final estimate. There may be a bias and an unknown bias at that.

Objective sampling is further sub-divided as *non-probabilistic*, *probabilistic* and *mixed*. In non-probabilistic objective sampling, there is a fixed sampling rule but there is no probability attached to the mode of selection; e.g., selecting every 10th individual from a list, starting with the first, or selecting every 10th line in a potato-field. If, however, the selection of the first individual is made in such a manner that each of the first 10 gets an equal chance of being selected, it becomes a case of mixed sampling—partly probabilistic, partly non-probabilistic. On the other hand, if for each individual there is a definite preassigned probability of being selected, the sampling is said to be probabilistic. Probabilistic sampling is also called *random sampling*. If, in particular, each individual of the population has an equal chance of being selected, then the sampling is called *unrestricted random sampling* or *simple random sampling*. Simple random sampling is said to be with or without replacements according as any individual once selected is returned to the population or not.

21.8 Simple random sampling

The simplest and the most commonly used type of probability sampling is simple random sampling. In this kind of sampling, each member of the population has the same probability of being

included in the sample. Simple random sampling may be with or without replacements. This type of sampling has been discussed in Sections 14.1—14.4.

If we are interested in estimating the population mean μ from a simple random sample of size n drawn from a population of size N , let $T = \sum_{i=1}^n \lambda_i x_i$ be the best linear unbiased estimator, where x_i is the value of the variate for the i th sample individual. We know (*vide* Section 14.3) that

$$E(x_i) = \mu,$$

$\text{var}(x_i) = \sigma^2$, where σ^2 is the population variance,

and $\text{cov}(x_i, x_j) = 0$ for sampling with replacements (SRSWR), and

$$= -\frac{\sigma^2}{N-1} \text{ for sampling without replacements (SRSWOR).}$$

Thus,

$$\begin{aligned} E(T) &= \sum_i \lambda_i E(x_i) \\ &= \mu \sum_i \lambda_i. \end{aligned}$$

For T to be unbiased, $\sum_i \lambda_i = 1$.

Again,

$$\begin{aligned} \text{var}(T) &= \sum_i \lambda_i^2 \text{ var}(x_i) + 2 \sum_{i < j} \lambda_i \lambda_j \text{ cov}(x_i, x_j) \\ &= \sigma^2 \sum_i \lambda_i^2 - \frac{2\sigma^2}{N-1} \sum_{i < j} \lambda_i \lambda_j \\ &= \sigma^2 \left(1 + \frac{1}{N-1}\right) \sum_i \lambda_i^2 - \frac{\sigma^2}{N-1} (\sum_i \lambda_i)^2 \text{ for SRSWR, and} \\ &= \sigma^2 \sum_i \lambda_i^2 \text{ for SRSWOR.} \end{aligned}$$

In either case, for $\text{var}(T)$ to be a minimum, we have to minimise $\sum_i \lambda_i^2$, subject to $\sum_i \lambda_i = 1$. This occurs when

$$\lambda_i = 1/n \text{ for each } i, \quad \dots \quad (21.1)$$

so that $T = \bar{x}$ is the best linear unbiased estimator. It also follows that

$$\text{var}(\bar{x}) = \frac{\sigma^2}{n} \quad \dots \quad (21.2)$$

for SRSWR, and

$$\text{var}(\bar{x}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \quad \dots \quad (21.3)$$

for SRSWOR. The factor $\frac{N-n}{N-1}$ is called the *finite population correction* (f.p.c.) and may be neglected if N is very large compared to n .

In case σ^2 is unknown, it can be estimated from the sample and its unbiased estimator is (*vide* Section 16.5)

$$s'^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \quad \dots \quad (21.3a)$$

for SRSWR, x_i being the value of the variate for the i th sample individual.

For SRSWOR, the unbiased estimator of σ^2 is

$$\frac{N-1}{N} \cdot s'^2. \quad \dots \quad (21.3b)$$

In SRSWOR from a finite population, many authors would define the population variance as*

$$S^2 = \frac{1}{N-1} \sum_{a=1}^N (X_a - \bar{X})^2. \quad \dots \quad (21.3c)$$

Since $(N-1)S^2 = N\sigma^2$, equation (21.3) would take the form

$$\text{var}(\bar{x}) = \frac{S^2}{n} \cdot \frac{N-n}{N}, \quad \dots \quad (21.3d)$$

the ratio $1 - \frac{n}{N}$ now serving as the f.p.c.

Further, in this case s'^2 would be an unbiased estimator of S^2 , so that an unbiased estimator of $\text{var}(\bar{x})$ would be

$$\widehat{\text{var}}(\bar{x}) = \frac{s'^2}{n} \cdot \frac{N-n}{N} = s'^2 \left(\frac{1}{n} - \frac{1}{N} \right). \quad \dots \quad (21.3e)$$

If we want to estimate the population proportion p of members of the population possessing a certain character A , its unbiased estimator from a sample of size n is the sample proportion f/n of sample individuals possessing that character and the standard error ($\sigma_{f/n}$) is (*vide* Section 14.4)

$$\sigma_{f/n} = \sqrt{\frac{pq}{n}} \quad \dots \quad (21.4)$$

* X_a is the value of x for the a th member of the population.

for SRSWR, and

$$\sigma_{fis} = \sqrt{\frac{pq}{n} \times \frac{N-n}{N-1}} \quad \dots \quad (21.5)$$

for SRSWOR, where $q = 1 - p$.

For estimating the standard error (s.e.), p may be replaced by its unbiased estimator in (21.4) or (21.5), as the case may be.

Ex. 21.2 The standard deviation of the marks obtained in mathematics by 121 boys is found to be 12.5. Find the standard error of the estimator of population mean for a random sample of size 10 (a) taken with replacements and (b) taken without replacements.

The estimator of the population mean is the sample mean \bar{x} . The s.e. of \bar{x} for SRSWR is

$$\begin{aligned} & \frac{\sigma}{\sqrt{n}} \\ &= \frac{12.5}{\sqrt{10}} = \frac{12.5}{3.1623} = 3.95, \end{aligned}$$

, and the s.e. of \bar{x} for SRSWOR is

$$\begin{aligned} & \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \\ &= \frac{12.5}{\sqrt{10}} \sqrt{\frac{121-10}{121-1}} \\ &= \frac{12.5}{\sqrt{10}} \sqrt{\frac{111}{120}} \\ &= 12.5 \sqrt{0.925} = 12.5 \times 0.3011 \\ &= 3.80. \end{aligned}$$

21.9 Stratified random sampling

In this method, before drawing the random sample, one divides the population, say Π , into several strata or sub-populations, say $\Pi_1, \Pi_2, \dots, \Pi_k$, which are relatively homogeneous within themselves and the means of which are as widely different as possible. The sample, say Σ , is composed of k partial samples, say $\Sigma_1, \Sigma_2, \dots, \Sigma_k$, drawn at random from the corresponding strata, generally without replacements.

Stratified random sampling has a number of merits relative to simple random sampling. (1) In many situations stratified sampling will be administratively more convenient. In taking a sample of villages from the whole of West Bengal, we may take the districts as strata. This will facilitate the organisation of field work, for the existing administrative set-up at the district level may be used for this purpose. (2) Again, stratified sampling will be more representative in the sense that here we can ensure that some individuals from each of the sub-populations (strata) will be included in the sample. (3) Stratified sampling, moreover, has the merit of supplying not only an estimate for the population as a whole, but also separate estimates (with estimates of their standard errors) for the individual strata. (4) Since a portion of the variability identifiable as between-strata variance is eliminated in stratified random sampling, it is more efficient than simple random sampling. If the between-strata variance is large, the within-strata variance, which provides the estimate for error, will be small as compared with the variance for the whole population. That is why we try to make each particular stratum as homogeneous as possible, while making the strata as different from each other as possible.

Let the population, consisting of N individuals with mean μ and standard deviation σ , be stratified into k strata, the number of individuals in the i th stratum being N_i , with mean μ_i and standard deviation σ_i , so that

$$\sum_i N_i = N$$

and

$$\mu = \frac{\sum_i N_i \mu_i}{N}.$$

We take a sample of size n , by selecting at random and without replacements n_i individuals from the i th stratum. Let us denote by x_{ij} the value of x for the j th selected individual from the i th stratum, for $i=1, 2, \dots, k$ and $j=1, 2, \dots, n_i$. Let $T = \sum_i \sum_j \lambda_{ij} x_{ij}$, λ_{ij} 's being constants, be the best linear unbiased estimator of μ . Thus

$$\begin{aligned} E(T) &= \sum_i \sum_j \lambda_{ij} E(x_{ij}) \\ &= \sum_i \sum_j \lambda_{ij} \mu_i = \sum_i \mu_i \sum_j \lambda_{ij}, \end{aligned}$$

on being equated to $\mu = \frac{\sum N_i \mu_i}{N}$, gives

$$\sum_i \lambda_{ij} = \frac{N_i}{N}.$$

Again,

$$\begin{aligned}\text{var}(T) &= \text{var}\left\{\sum_i \sum_j \lambda_{ij} x_{ij}\right\} \\ &= \sum_i \text{var}\left(\sum_j \lambda_{ij} x_{ij}\right) \quad (\text{since samples from different strata are independent}) \\ &= \sum_i \left\{ \sum_j \lambda_{ij}^2 \text{var}(x_{ij}) + 2 \sum_{j < j'} \lambda_{ij} \lambda_{ij'} \text{cov}(x_{ij}, x_{ij'}) \right\} \\ &= \sum_i \left\{ \sum_j \lambda_{ij}^2 \sigma_i^2 - 2 \sum_{j < j'} \lambda_{ij} \lambda_{ij'} \frac{\sigma_i^2}{N_i - 1} \right\} \\ &= \sum_i \left\{ \frac{\sigma_i^2 N_i}{N_i - 1} \sum_j \lambda_{ij}^2 - \frac{\sigma_i^2}{N_i - 1} \sum_j \lambda_{ij}^2 - \frac{2\sigma_i^2}{N_i - 1} \sum_{j < j'} \lambda_{ij} \lambda_{ij'} \right\} \\ &= \sum_i \left\{ \frac{\sigma_i^2 N_i}{N_i - 1} \sum_j \lambda_{ij}^2 - \frac{\sigma_i^2}{N_i - 1} (\sum_j \lambda_{ij})^2 \right\}.\end{aligned}$$

Now, $\sum_j \lambda_{ij} = N_i/N$ is fixed. Hence $\text{var}(T)$ is a minimum when $\sum_j \lambda_{ij}^2$ is a minimum for fixed $\sum_j \lambda_{ij}$. This happens when

$$\lambda_{ij} = \lambda_{i0} = \frac{N_i}{N n_i}.$$

Hence the best linear unbiased estimator of μ is

$$T = \sum_i \sum_j \frac{N_i}{N n_i} x_{ij} = \sum_i N_i \bar{x}_i / N, \quad \dots \quad (21.6)$$

and

$$\begin{aligned}\text{var}(T) &= \frac{1}{N^2} \sum_i N_i^2 \text{var}(\bar{x}_i) \\ &= \frac{1}{N^2} \sum_i N_i^2 \cdot \frac{\sigma_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i - 1}. \quad \dots \quad (21.7)\end{aligned}$$

Writing

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \mu_i)^2 = N_i \sigma_i^2 / (N_i - 1),$$

we have

$$\text{var}(T) = \frac{1}{N^2} \sum_i N_i \cdot \frac{S_i^2}{n_i} (N_i - n_i). \quad \dots \quad (21.8)$$

This is the so-called *variance function*.

To determine n_1, n_2, \dots, n_k in an optimum manner, i.e. in such a manner that for given cost the variance of the estimator is minimised, we assume a simple *cost function*, viz.

$$C = a + \sum_i c_i n_i, \quad \dots \quad (21.9)$$

where C is the total cost, a the overhead cost and c_i the cost per unit for the i th stratum. Thus n_1, n_2, \dots, n_k should be such that for $C=C_0$ (given), $\text{var}(T)$ is a minimum. To solve the problem, we take the function

$$\begin{aligned} F &= \text{var}(T) + \lambda C \\ &= \frac{1}{N^2} \sum_i N_i \cdot \frac{S_i^2}{n_i} (N_i - n_i) + \lambda (a + \sum_i c_i n_i), \end{aligned}$$

where λ is an undetermined multiplier. The equations

$$\frac{\partial F}{\partial n_i} = 0, \quad \text{for } i = 1, 2, \dots, k,$$

along with the equation

$$C = C_0,$$

will determine the n_1, n_2, \dots, n_k and λ that minimise $\text{var}(T)$ for given C .

Here

$$\frac{\partial F}{\partial n_i} = -\frac{1}{N^2} \cdot N_i^2 \cdot \frac{S_i^2}{n_i^2} + \lambda c_i = 0$$

gives

$$n_i \propto \frac{N_i S_i}{\sqrt{c_i}}$$

or

$$n_i = \lambda' \frac{N_i S_i}{\sqrt{c_i}}, \quad \dots \quad (21.10)$$

where λ' involves λ .

Finally,

$$a + \lambda' \sum_i c_i \frac{N_i S_i}{\sqrt{c_i}} = C_0$$

or

$$\lambda' = \frac{C_0 - a}{\sum_i N_i S_i \sqrt{c_i}}. \quad \dots \quad (21.11)$$

In particular, if $c_1 = c_2 = \dots = c_k$, fixing the cost is equivalent to fixing the total sample size, i.e. making $n_1 + n_2 + \dots + n_k = n$. Here

the optimum values of n_1, n_2, \dots, n_k are given by

$$n_i \propto N_i S_i$$

or $n_i = \lambda'' N_i S_i, \dots \quad (21.12)$

where $\lambda'' \sum_i N_i S_i = n$

or $\lambda'' = \frac{n}{\sum_i N_i S_i}. \dots \quad (21.13)$

This is *Neyman's formula for optimum allocation*.

If, further,

$$S_1 = S_2 = \dots = S_k,$$

or if the differences among the S_i 's are ignored, then we have

$$n_i \propto N_i. \dots \quad (21.14)$$

This is *Bowley's formula for proportional allocation*.

Formulas (21.10) and (21.12) for optimum allocation involve S_i 's and c_i 's, which are generally unknown. Thus, to use these formulæ one has to estimate S_i 's and c_i 's before the survey may be undertaken. This is done by means of a preliminary survey or *pilot survey*, which is a survey of a relatively smaller scale conducted for the purpose of estimating the constants involved in the main survey-design.

Two points should be remembered in this connection :

(1) A pilot survey which one has to undertake to estimate S_i 's and c_i 's involves some extra cost. With this added cost, one might increase the precision by taking a larger sample with proportional allocation, for which the variations among the S_i 's and the c_i 's would be ignored. So one must consider whether the extra cost for optimisation is worth while.

(2) Generally in sample surveys, more than one variable are involved. Now, an optimum sample with respect to one variable may not be so for another. If, however, there is a hierarchy of importance in the variables involved, one may take an optimum sample with respect to the most important variable.

Stratification is not justified if (1) stratification itself is too costly and/or (2) the between-strata variance is not large enough to effect a sufficient gain in the accuracy as compared with simple random sampling.

Ex. 21.3 2,010 cultivators' holdings in U.P. are stratified according to size. The number of holdings (N_i), mean area under wheat per holding (μ_i) and s.d. of area under wheat per holding (σ_i) are given below for each stratum :

Stratum No.	Holding size (acres)	Number of holdings (N_i)	Mean area under wheat per holding (μ_i)	s.d. of area under wheat per holding (σ_i)
1	0— 40	394	5·4	8·3
2	41— 80	461	16·3	13·3
3	81—120	391	24·3	15·1
4	121—160	334	34·5	19·8
5	161—200	169	42·1	24·5
6	201 and above	261	57·9	31·2

A sample of 100 holdings is taken to estimate the mean area under wheat per holding by (a) simple random sampling, (b) stratified random sampling with proportional allocation and (c) stratified random sampling with optimum allocation. Compare the standard errors of the estimators in the three cases.

The standard deviation σ of area under wheat per holding for the whole population is given by

$$\sigma^2 = \frac{\sum N_i \sigma_i^2}{N} + \frac{\sum N_i (\mu_i - \bar{\mu})^2}{N},$$

where $\bar{\mu} = \frac{\sum N_i \mu_i}{N} = \frac{52,893.0}{2,010} = 26.3$.

Thus

$$\sigma^2 = 340.44 + 271.16 = 611.60,$$

or $\sigma = 24.73$.

The standard error of the estimator for mean in simple random sampling, ignoring the f.p.c., is

$$\begin{aligned} \text{s.e. random} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{24.73}{\sqrt{100}} = 2.473. \end{aligned}$$

The standard error of the estimator for mean in stratified random sampling with proportional allocation is given by

$$(s.e._{prop})^2 = \frac{1}{N^2} \sum_i N_i^2 \cdot \frac{\sigma_i^2}{n_i},$$

where $n_i = \frac{n}{N} \cdot N_i$, the f.p.c. being ignored.

$$\begin{aligned} \text{Thus } (s.e._{prop})^2 &= \frac{1}{N^2} \sum_i N_i \sigma_i^2 \\ &= 340.44/100 = 3.4044, \end{aligned}$$

and $s.e_{prop} = 1.845$.

Lastly, the standard error of the estimator for mean in stratified random sampling with optimum allocation is given by

$$(s.e._{opt})^2 = \frac{1}{N^2} \sum_i N_i^2 \cdot \frac{\sigma_i^2}{n_i},$$

where $n_i = n \cdot \frac{N_i \sigma_i}{\sum_i N_i \sigma_i}$, the f.p.c. being ignored.

$$\text{Thus } (s.e._{opt})^2 = \frac{1}{N^2 n} (\sum_i N_i \sigma_i)^2,$$

$$\begin{aligned} \text{or } s.e._{opt} &= \frac{1}{N \sqrt{n}} \sum_i N_i \sigma_i \\ &= \frac{17.02}{\sqrt{100}} = 1.702. \end{aligned}$$

21.10 Multistage sampling

In multistage sampling, the material to be sampled is regarded as composed of a number of first-stage (or primary) sampling units, each of which is made up of a number of second-stage (or secondary) sampling units, each of which, in its turn, is made up of a number of third-stage (or tertiary) units, and so on, until we reach the ultimate sampling units in which we are interested. The sampling is also carried out in stages. At the first stage, the first-stage sampling units are sampled by some suitable random method. At the second stage, a sample of second-stage units is selected from each of the selected first-stage units, again by some suitable random method. Further stages may be added, if necessary, to get a sample

of the ultimate sampling units. For example, to get a sample of crop-fields growing paddy in West Bengal, one may first get a sample of districts, then a sample of villages from each selected district and finally a sample of crop-fields from each selected village.

Multistage sampling introduces a flexibility into the sampling procedure which is lacking in the simpler methods. It enables existing divisions and sub-divisions of the material to be taken as sampling units at different stages. The construction of a second-stage frame is necessary only for the selected first-stage units. This means a great saving in operational costs, particularly if the survey covers a large area including under-developed pockets. Thus, in selecting a number of households from the whole of Indian Union, it is an impossible task, from the stand-point of both administration and field-work, to take a simple random sample. It is much simpler and practicable to select a sample of villages and then a sample of households from each selected village. Multistage sampling, however, is in general less efficient than some suitable single-stage process.

The mode of analysis of data in multistage sampling may be illustrated with two-stage sampling. First, let us assume for simplicity that the numbers of second-stage units in all first-stage units are equal.

Suppose there are M first-stage units numbered 1, 2, ..., M , each consisting of N_0 second-stage units. Let m first-stage units be selected, and from each chosen first-stage unit let n_0 second-stage units be selected. Let the sampling be simple random at each stage, and suppose the finite population corrections are negligible. With no loss of generality, the selected first-stage units may be numbered from 1 to m and the selected second-stage units from the i th selected first-stage unit may be numbered from 1 to n_0 . Let x_{ij} denote the value of the variable under enquiry for the j th second-stage unit in the i th first-stage unit, for $i=1, 2, \dots, m$ and $j=1, 2, \dots, n_0$.

The appropriate model for analysis is the random model for one-way classified data (*vide* Section 19.5), the data being classified by the first-stage units, but the given m first-stage units are only a sample of all M first-stage units in which we are interested. Thus

$$x_{ij} = \mu + b_i + e_{ij}, \quad \dots \quad (21.15)$$

where μ is the grand mean, b_i the population mean of the i th first-

stage unit (μ_i) taken as a deviation from the grand mean and e_{ij} the residual, so that

$$E(b_i) = 0 \text{ and } E(e_{ij}) = 0.$$

It is also assumed that $\text{var}(b_i) = \sigma_b^2$ and $\text{var}(e_{ij}) = \sigma_e^2$.

From the analysis of variance table, we have (vide Section 19.5)

$$E(MSB) = E\left\{\frac{1}{m-1} \sum_i n_0 (x_{i0} - x_{00})^2\right\} = \sigma_b^2 + n_0 \sigma_e^2 \quad \dots \quad (21.16)$$

and $E(MSE) = E\left\{\frac{1}{m(n_0-1)} \sum_i \sum_j (x_{ij} - x_{i0})^2\right\} = \sigma_e^2. \quad \dots \quad (21.17)$

Thus MSE and $(MSB - MSE)/n_0$ are the unbiased estimators of σ_e^2 and σ_b^2 , respectively.

If we are interested in estimating the population grand mean μ , the best linear unbiased estimator is x_{00} . Now,

$$x_{00} = \mu + b_0 + e_{00},$$

where $b_0 = \frac{1}{m} \sum_i b_i$

and $e_{00} = \frac{1}{mn_0} \sum_i \sum_j e_{ij},$

so that $E(x_{00}) = \mu \quad \dots \quad (21.18)$

and $\text{var}(x_{00}) = \frac{\sigma_b^2}{m} + \frac{\sigma_e^2}{mn_0}$, ignoring the f.p.c. $\dots \quad (21.19)$

Thus, by substituting in (21.19) the unbiased estimators of σ_b^2 and σ_e^2 , we get an unbiased estimator of the variance of the estimator, x_{00} , from the sample, viz.

$$\frac{MSB - MSE}{mn_0} + \frac{MSE}{mn_0} = \frac{MSB}{mn_0}. \quad \dots \quad (21.20)$$

We can choose m and n_0 in an optimum manner, using the variance function

$$V = \frac{\sigma_b^2}{m} + \frac{\sigma_e^2}{mn_0}$$

and taking the cost function as

$$C = a_0 + c_1 m + c_2 mn_0, \quad \dots \quad (21.21)$$

where a_0 is the overhead cost.

c_1 is the cost per first-stage unit sampled,

and c_2 is the cost per second-stage unit sampled.

With given cost C_0 , we can determine m , n_0 and the undetermined multiplier λ from the following equations :

$$\left. \begin{aligned} \frac{\partial V}{\partial m} + \lambda \frac{\partial C}{\partial m} &= 0, \\ \frac{\partial V}{\partial n_0} + \lambda \frac{\partial C}{\partial n_0} &= 0 \end{aligned} \right\}$$

and

$$C = C_0,$$

or

$$\left. \begin{aligned} \frac{\sigma_s^2}{m^2} + \frac{\sigma_e^2}{m^2 n_0} &= \lambda(c_1 + c_2 n_0), \\ \frac{\sigma_e^2}{m n_0^2} &= \lambda c_2 m \end{aligned} \right\}$$

and

$$C = C_0.$$

These lead to

$$n_0 = \frac{\sigma_e}{\sigma_s} \sqrt{\frac{c_1}{c_2}} \quad \dots \quad (21.22)$$

and

$$m = \frac{C_0 - a_0}{c_1 + c_2 n_0}. \quad \dots \quad (21.23)$$

For estimating the quantities involved in the solution, viz σ_s and σ_e , and a_0 , c_1 and c_2 , one has to undertake a pilot survey.

Next, let us suppose that the numbers of second-stage units in different first-stage units are different, other conditions remaining the same. Let N_i be the number of second-stage units in the i th first-stage unit, for $i=1, 2, \dots, M$. Also, let in the sample n_i second-stage units be selected from the i th selected first-stage unit, $i=1, 2, \dots, m$. Suppose that sampling is simple random at both stages and that the f.p.c. may be ignored. In the linear model

$$x_{ij} = \mu + b_i + e_{ij},$$

μ now represents the mean of the M means, viz.

$$\mu = \frac{1}{M} \sum_{i=1}^M \mu_i,$$

$$\therefore b_i = \mu_i - \mu$$

and

$$e_{ij} = x_{ij} - \mu_i,$$

so that

$$E(b_i) = 0 \text{ and } E(e_{ij}) = 0.$$

As before, the sample mean

$$\begin{aligned} x_{00} &= \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \\ &= \frac{1}{n} \sum_i n_i x_{i0}, \quad \text{where } n = \sum_{i=1}^m n_i, \end{aligned} \quad \dots \quad (21.24)$$

is an unbiased estimator of μ .

But since μ is different from the grand mean, viz. $\frac{1}{N} \sum_{i=1}^M N_i \mu_i$, where $N = \sum_{i=1}^M N_i$, the sample mean would be a biased estimator of the grand mean. However, if N_i 's are almost uncorrelated with μ_i 's, the distorted weighting does not matter greatly and the bias may be supposed to be negligible in large samples. But if N_i 's are highly correlated with μ_i 's, the bias may be considerable even in large samples. Some alternative estimator which is unbiased may be considered ; e.g., $x'_{00} = \frac{M}{m} \cdot \frac{1}{N} \sum_{i=1}^m N_i x_{i0}$ will be an unbiased estimator of the population grand mean, but this will often have very poor precision if N_i 's vary considerably.

To obtain the variance of x_{00} and to estimate it from the sample, we proceed exactly as before. We have

$$\text{var}(x_{00}) = \frac{\sigma_x^2}{n^2} \sum_i n_i^2 + \frac{\sigma_\epsilon^2}{n}. \quad \dots \quad (21.25)$$

From the analysis of variance table of one-way classified data, we have (vide Section 19.5)

$$E(MSB) = E\left\{ \frac{1}{m-1} \sum_i n_i (x_{i0} - x_{00})^2 \right\} = \sigma_\epsilon^2 + \frac{n - \sum_i n_i^2/n}{m-1} \cdot \sigma_\epsilon^2 \quad \dots \quad (21.26)$$

$$\text{and } E(MSE) = E\left\{ \frac{1}{n-m} \sum_i \sum_j (x_{ij} - x_{i0})^2 \right\} = \sigma_\epsilon^2, \quad \dots \quad (21.27)$$

so that MSE and $\frac{(MSB - MSE)(m-1)}{n - \sum_i n_i^2/n}$ are unbiased estimators of σ_ϵ^2 and σ_x^2 , respectively.

Thus an unbiased estimator of the variance of the sample mean is given by

$$\frac{(MSB - MSE)(m-1) \sum_i n_i^2/n^2}{n - \sum_i n_i^2/n} + \frac{MSE}{n}. \quad \dots \quad (21.28)$$

The mode of analysis may be easily generalised if the number of stages is more than two.

Ex. 21.4 To determine the yield-rate of paddy in a district of West Bengal, 6 villages were selected at random, 3 plots were selected in each selected village and 2 circular cuts were taken at randomly-chosen points in each selected plot. The yields, in suitable units, were obtained as follows :

Plot No.	Village 1		Village 2		Village 3		Village 4		Village 5		Village 6	
	cut 1	cut 2										
1	16	8	6	5	18	8	13	13	17	15	12	15
2	26	31	5	16	3	10	7	6	11	21	11	17
3	11	13	16	5	16	20	7	2	17	25	8	10

Give an estimate of the s.e. of the estimator of mean yield.

Let x_{ijk} be the yield of the k th cut in the j th plot of the i th village ($i=1, 2, \dots, 6; j=1, 2, 3; k=1, 2$). Let us make a change of base and write

$$u_{ijk} = x_{ijk} - 15.$$

We have

$$T_{000} = \sum_i \sum_j \sum_k u_{ijk} = -80$$

and

$$\sum_i \sum_j \sum_k u_{ijk}^2 = 1,732.$$

In the following table we write down the values $T_{ij0} = \sum_k u_{ijk}$ for all (i, j) combinations :

SHOWING VILLAGE- PLOT SUB-TOTALS

Plot No.	Village					
	1	2	3	4	5	6
1	-6	-19	-4	-4	2	-3
2	27	-9	-17	-17	2	-2
3	-6	-9	6	-21	12	-12
Total	15	-97	-15	-42	16	-17

$$\text{Thus } SS \text{ between villages} = \frac{\sum T_{i00}^2}{6} - \frac{T_{000}^2}{36}$$

$$= \frac{4,128}{6} - 177.78$$

$$= 510.22,$$

$$SS \text{ between plots (within villages)} = \frac{\sum \sum T_{ij0}^2}{2} - \frac{\sum T_{i00}^2}{6}$$

$$= \frac{2,720}{2} - 688$$

$$= 672,$$

$$SS \text{ between cuts (within plots)} = \frac{\sum \sum \sum u_{ijk}^2}{2} - \frac{\sum \sum T_{ij0}^2}{2}$$

$$= 1,732 - 1,360$$

$$= 372.$$

As such, $MSA = MS \text{ between villages}$

$$= \frac{510.22}{5} = 102.04,$$

$$MSB = MS \text{ between plots (within villages)}$$

$$= \frac{672}{12} = 56,$$

and $MSE = MS \text{ between cuts (within plots)}$

$$= \frac{372}{18} = 20.67.$$

Hence the estimate of the variance of the estimator of mean yield is

$$\frac{MSA}{36} = \frac{102.04}{36} = 2.8344,$$

and the corresponding estimate of s.e. is $\sqrt{2.8344} = 1.68$.

21.11 Systematic sampling

A frequently used method of sampling when a list of the sampling units is available is systematic sampling. Suppose the N units of the population are numbered from 1 to N and a sample of size n is to be selected such that $\frac{n}{N} = \frac{1}{k}$, k being an integer. Systematic sampling

then consists in selecting at random a unit from the first k units and also selecting every subsequent k th unit. This is a case of mixed sampling, which is partly probabilistic and partly non-probabilistic. This is probabilistic since the first member of the sample is selected at random (with equal probabilities) from the first k units and non-probabilistic since the other members in the sample are fixed by the choice of the first member. We may as well have a completely random start, the first unit being chosen from the N units at random and then every k th unit being taken in a circular manner until the whole list is exhausted. This is called *circular systematic sampling*.

The apparent advantages of this method over simple random sampling are the following :

(1) It is much easier and quicker to draw a systematic sample and the work may be done by laymen. (2) Intuitively, systematic sampling seems likely to give more precise estimates than simple random sampling. In effect, the method stratifies the population into n strata of k units each and one unit is selected from each stratum.

The method, however, has many disadvantages. The estimator of the population mean is the sample mean. The variance of the estimator is the variance of the k possible estimates from each of k possible systematic samples with one of the first k units of the population as the first member. If $x_{10}, x_{20}, \dots, x_{k0}$ are the k possible estimates, which are equally likely, the variance of the estimator is given by

$$\sigma_{\bar{x}}^2 = \frac{1}{k} \sum_i (x_{i0} - \bar{x}_{00})^2. \quad \dots \quad (21.29)$$

The variance, however, cannot be unbiasedly estimated from a single sample. A way out is to make use of the method of *interpenetrating samples*, where two or more independent samples are taken from the population. In the present case, e.g., if p estimates of x_{00} , say \bar{x}_i ($i=1, 2, \dots, p$), are available, then our combined estimate will be $\bar{x} = \frac{1}{p} \sum_{i=1}^p \bar{x}_i$, whose variance will be estimated by

$$\frac{1}{p(p-1)} \sum_{i=1}^p (\bar{x}_i - \bar{x})^2.$$

The method may give highly biased estimates if there are some

periodic features in the list and the sampling interval k is equal to, or is a multiple of, the period. Suppose we have a list of individuals such that the variate value for every 10th individual is large (or small) compared to those for the others. It can be easily seen that the estimates would be highly biased if in drawing the samples systematically those individuals happened to be selected.

In the same way as in sampling from a list, systematic sampling may also be adopted for sampling material continuously distributed over time or in space, by taking sampling units at equal intervals over the material. For example, the products coming out continuously from a manufacturing process may be sampled systematically by selecting products manufactured at a fixed interval of time. Again, in sampling from plants growing in rows, one may divide the whole area into a number of equal rectangular blocks, choose a plant at random from the first block and draw plants exactly from similar spots from other blocks. In some cases of area sampling, specially in forest surveys, it may be convenient to divide the whole area into a number of parallel strips and select a number of strips either systematically or at random. This type of sampling is called *line sampling*.

21.12 Multiphase sampling

It is sometimes convenient and economical to collect certain items of information from a sample constituting only a part of the original sample. This is termed two-phase sampling. Further phases may be added, if necessary.

Multiphase sampling has several advantages. If the number of units required to give the desired accuracy in different items is widely different or if the cost of collection of data for different items is different, multiphase sampling may be suitably adopted. Also, the information gathered in earlier phases may be utilised as a basis for sampling, say for stratification, in subsequent phases, thus resulting in a large saving in cost. Multiphase sampling differs structurally from multistage sampling, in that in the former the sampling unit is the same in all the phases, whereas in the latter there is a hierarchy of sampling units in different stages. For example, in drawing a random sample of households for a

family-budget enquiry amongst the middle-class families in Calcutta, we may take a sample from all households to classify the households into middle-class and non-middle-class groups. In the second phase, we may draw for the family-budget enquiry a sample out of the sample of middle-class households obtained in the first phase.

21.13 Double sampling

If the variable y under enquiry is very costly to enumerate and there is another variable x , which is highly correlated with y and relatively much cheaper and easier to enumerate, then we may have recourse to double sampling.

A first sample, of a relatively small size n , is taken in which both x and y are enumerated to find a suitable relationship between x and y . In the second sample, of a relatively large size m , only x is enumerated to find the estimate of μ_x , say \bar{x}_m . Then the estimate of the population mean of y is obtained by using \bar{x}_m in the relationship obtained from the first sample. Sampling is supposed to be simple random without replacements.

We may, however, take more than one auxiliary variable, if necessary.

For example, in the estimation of yield of dry jute fibre (y), which is costly and time-consuming to enumerate, we may take as the auxiliary variable the weight of green jute plants. Also, in forest surveys, in estimating the timber volume (y) of trees one can take the eye-estimation value (x) as the auxiliary variable.

In particular, instead of taking a second sample, one may completely enumerate x to obtain the true value of μ_x .

The following different cases may be discussed :

Case I. The relationship between y and x is of the form

$$y_i = Rx_i + \epsilon_i, \quad \dots \quad (21.30)$$

where ϵ_i is the residual. If ϵ_i is normally distributed with mean zero and variance proportional to x_i , the maximum-likelihood estimator of R (i.e. μ_y/μ_x), from the first sample, is

$$\frac{\sum y_i}{\sum x_i} = \frac{\bar{y}_n}{\bar{x}_n}. \quad \dots \quad (21.31)$$

The estimator, however, is not unbiased, but is consistent.

The estimator of μ_{y_1} , obtained by using the estimator of μ_x from the second sample, is

$$\hat{\mu}_y = (\bar{y}_n/\bar{x}_n)\bar{x}_m. \quad \dots \quad (21.32)$$

If x were completely enumerated, the estimator would be

$$\hat{\mu}_y = (\bar{y}_n/\bar{x}_n)\mu_x. \quad \dots \quad (21.33)$$

These are the so-called *ratio estimators*. The estimators are consistent.

Let us find the variance of the estimator in large samples when μ_x is known. Here

$$\begin{aligned}\hat{\mu}_y - \mu_y &= (\bar{y}/\bar{x})\mu_x - \mu_y = \mu_x \{(\bar{y}/\bar{x} - \mu_y)/\mu_x\} \\ &= \frac{\mu_x}{\bar{x}} \{(\bar{y} - R\bar{x})\},\end{aligned}$$

where R is the population value of the ratio. In large samples, neglecting the variation of \bar{x} from μ_x , we have

$$\hat{\mu}_y - \mu_y \approx \bar{y} - R\bar{x} = \frac{1}{n} \sum_{i=1}^n (y_i - Rx_i).$$

Hence

$$\begin{aligned}\text{var}(\hat{\mu}_y) &\approx \text{var}(\bar{y} - R\bar{x}) \\ &= \frac{1}{n} \cdot \frac{N-n}{N-1} \cdot \text{var}(y_i - Rx_i) \\ &= \frac{1}{n} \cdot \frac{N-n}{N-1} \left\{ \frac{1}{N} \sum_{i=1}^N (y_i - Rx_i)^2 \right\} \\ &= \frac{1}{n} \cdot \frac{N-n}{N-1} \left[\frac{1}{N} \sum_i \{(y_i - \mu_y) - R(x_i - \mu_x)\}^2 \right] \\ &= \frac{1}{n} \cdot \frac{N-n}{N-1} \{s_y^2 - 2R s_{yx} + R^2 s_x^2\}. \quad \dots \quad (21.34)\end{aligned}$$

The estimator of this variance would be

$$\widehat{\text{var}}(\hat{\mu}_y) = \frac{1}{n} \cdot \frac{N-n}{N} \{s'_y^2 - 2\hat{R}s'_{yx} + \hat{R}^2 s'_x^2\}, \quad \dots \quad (21.35)$$

where $\hat{R} = \bar{y}/\bar{x}$ is the estimator of the ratio and

$$s'_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2,$$

$$s'_{yx} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$$

and

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

If μ_x is unknown and is estimated by \bar{x}_m , from a second sample of size m , then

$$\hat{\mu}_y = (\bar{y}_n / \bar{x}_n) \bar{x}_m,$$

so that

$$\begin{aligned}\hat{\mu}_y - \mu_y &= (\bar{y}_n / \bar{x}_n) \bar{x}_m - \mu_y \\ &= (\bar{y}_n / \bar{x}_n) \mu_x - \mu_y + (\bar{y}_n / \bar{x}_n) \bar{x}_m - (\bar{y}_n / \bar{x}_n) \mu_x \\ &= (\mu_x / \bar{x}_n) \{ \bar{y}_n - R\bar{x}_n \} + (\bar{y}_n / \bar{x}_n) \{ \bar{x}_m - \mu_x \}.\end{aligned}$$

In large samples, neglecting the variations of \bar{x}_n from μ_x and of \bar{x}_n / \bar{y}_n from R , we have

$$\hat{\mu}_y - \mu_y \approx (\bar{y}_n - R\bar{x}_n) + R(\bar{x}_m - \mu_x).$$

Since the two samples are drawn independently,

$$\begin{aligned}\text{var}(\hat{\mu}_y) &\approx \text{var}(\bar{y}_n - R\bar{x}_n) + R^2 \text{var}(\bar{x}_m - \mu_x) \\ &= \frac{1}{n} \cdot \frac{N-n}{N-1} \{ \sigma_y^2 - 2R\sigma_{yx} + R^2\sigma_x^2 \} + R^2 \cdot \frac{\sigma_x^2}{m} \cdot \frac{N-m}{N-1}.\end{aligned}\quad \dots \quad (21.36)$$

This variance also can be similarly estimated from the sample.

Case 2. The relationship between y and x is linear and of the form

$$y_i = A + Bx_i + \epsilon_i. \quad \dots \quad (21.37)$$

If ϵ_i is normally distributed with mean zero and constant variance σ_ϵ^2 , then the maximum-likelihood estimators of A and B from the first sample are, respectively,

$$a = \bar{y}_n - b\bar{x}_n \quad \dots \quad (21.38)$$

and $b = \frac{s_y}{s_x}. \quad \dots \quad (21.39)$

The estimator of μ_y , obtained by using the estimator of μ_x from the second sample, is

$$\hat{\mu}_y = \bar{y}_n + b(\bar{x}_m - \bar{x}_n). \quad \dots \quad (21.40)$$

If μ_x is known, the estimator reduces to

$$\hat{\mu}_y = \bar{y}_n + b(\mu_x - \bar{x}_n). \quad \dots \quad (21.41)$$

These are the so-called *regression estimators*. The estimators are consistent.

Again, the population values A and B in (21.37) are

$$A = \mu_y + B\mu_x,$$

and

$$B = \rho \frac{\sigma_y}{\sigma_x}.$$

Thus

$$\bar{y}_n = \mu_y + B(\bar{x}_n - \mu_x) + \epsilon_n. \quad \dots \quad (21.41a)$$

Substituting this value of \bar{y}_n in (21.41), we get

$$\hat{\mu}_y = \mu_y + (b - B)(\mu_x - \bar{x}_n) + \hat{\epsilon}_n.$$

In large samples, neglecting the variation of b from B , we get

$$\hat{\mu}_y - \mu_y \approx \hat{\epsilon}_n,$$

so that

$$\text{var}(\hat{\mu}_y) \approx \text{var}(\hat{\epsilon}_n)$$

$$= \frac{\sigma_y^2}{n} \cdot \frac{N-n}{N-1}.$$

$$\begin{aligned} \text{Also, } \sigma_y^2 &= \frac{1}{N} \sum_{i=1}^N (y_i - A - Bx_i)^2 \\ &= \frac{1}{N} \sum_i \left\{ (y_i - \mu_y) - \rho \frac{\sigma_y}{\sigma_x} (x_i - \mu_x) \right\}^2 \\ &= \frac{1}{N} \sum_i (r_i - \mu_y)^2 + \rho^2 \frac{\sigma_y^2}{\sigma_x^2} \cdot \frac{1}{N} \sum_i (x_i - \mu_x)^2 \\ &\quad - 2\rho \frac{\sigma_y}{\sigma_x} \cdot \frac{1}{N} \sum_i (r_i - \mu_y)(x_i - \mu_x) \\ &= \sigma_y^2 + \rho^2 \sigma_y^2 - 2\rho^2 \sigma_y^2 \\ &= \sigma_y^2 (1 - \rho^2). \end{aligned}$$

Hence the variance of the regression estimator in the large-sample case, when μ_x is known, is given by

$$\text{var}(\hat{\mu}_y) = \frac{\sigma_y^2}{n} (1 - \rho^2) \cdot \frac{N-n}{N-1}. \quad \dots \quad (21.42)$$

The estimator of the variance from the sample would be

$$\widehat{\text{var}}(\hat{\mu}_y) = \frac{s_y^2}{n} (1 - r^2) \frac{N-n}{N}. \quad \dots \quad (21.43)$$

If μ_x is estimated from a second sample, we have, from (21.40) and (21.41a),

$$\begin{aligned} \hat{\mu}_y &= \mu_y + B(\bar{x}_n - \mu_x) + b(\bar{x}_m - x_n) + \hat{\epsilon}_n \\ &= \mu_y + B(\bar{x}_n - \mu_x) + b(\mu_x - \bar{x}_n) + b(\bar{x}_m - \mu_x) + \hat{\epsilon}_n \\ &= \mu_y + (b - B)(\mu_x - \bar{x}_n) + b(\bar{x}_m - \mu_x) + \hat{\epsilon}_n. \end{aligned}$$

In large samples, neglecting the variation of b from B , we have

$$\hat{\mu}_y - \mu_y \approx B(\bar{x}_m - \mu_x) + \epsilon_m,$$

so that

$$\begin{aligned} \text{var}(\hat{\mu}_y) &\approx B^2 \cdot \frac{\sigma_x^2}{m} \cdot \frac{N-m}{N-1} + \frac{\sigma_x^2}{n} \cdot \frac{N-n}{N-1} \\ &= \rho^2 \cdot \frac{\sigma_x^2}{m} \cdot \frac{N-m}{N-1} + \sigma_y^2 \cdot \frac{(1-\rho^2)}{n} \cdot \frac{N-n}{N-1}. \end{aligned} \quad \dots \quad (21.44)$$

The estimator of this variance from the sample would be

$$\widehat{\text{var}}(\hat{\mu}_y) = r^2 \frac{s'_x^2}{m} \cdot \frac{N-m}{N} + s'_y^2 \frac{(1-r^2)}{n} \cdot \frac{N-n}{N}. \quad \dots \quad (21.45)$$

Let us now compare the efficiencies of the three methods of sampling, viz. simple random sampling, random sampling with ratio estimate and random sampling with regression estimate when μ_x is a known value. We have

$$V_1 = \frac{\sigma_y^2}{n} \cdot \frac{N-n}{N-1} \quad \text{for simple random sampling,}$$

$$V_2 = \frac{(\sigma_y^2 - 2R\sigma_{yx} + R^2\sigma_x^2)}{n} \cdot \frac{N-n}{N-1} \quad \text{for ratio estimator}$$

$$\text{and} \quad V_3 = \frac{\sigma_y^2(1-\rho^2)}{n} \cdot \frac{N-n}{N-1} \quad \text{for regression estimator.}$$

Clearly, $V_3 < V_1$, unless the correlation between y and the auxiliary variable x is zero. Thus random sampling with regression estimator is, for all practical purposes, more efficient.

Again,

$$\begin{aligned} V_1 - V_2 &= \frac{2R\sigma_{yx} - R^2\sigma_x^2}{n} \cdot \frac{N-n}{N-1} \\ &= \frac{2R\rho\sigma_y\sigma_x - R^2\sigma_x^2}{n} \cdot \frac{N-n}{N-1}, \end{aligned}$$

which is positive if

$$\begin{aligned} \rho &> \frac{1}{2} R \frac{\sigma_x}{\sigma_y} \\ &= \frac{1}{2} \cdot \frac{\sigma_x/\mu_x}{\sigma_y/\mu_y} \\ &= \frac{1}{2} \cdot \frac{c_x}{c_y}. \end{aligned}$$

Thus random sampling using ratio estimator may be more efficient or less efficient than simple random sampling, depending upon the correlation between y and x .

Thirdly,

$$\begin{aligned} V_2 - V_3 &= \frac{R^2 \sigma_x^2 - 2R\rho\sigma_x\sigma_y + \rho^2\sigma_y^2}{n} \cdot \frac{N-n}{N-1} \\ &= \frac{(R\sigma_x - \rho\sigma_y)^2}{n} \cdot \frac{N-n}{N-1} \end{aligned}$$

This is positive, unless

$$\rho = R \frac{\sigma_x}{\sigma_y} = \frac{c_x}{c_y}.$$

Hence, for all practical purposes, random sampling with regression estimator is more efficient than that with ratio estimator.

21.14 Purposive sampling

The term 'purposive sampling' has been used in several slightly different senses in connection with subjective methods of sampling.

In the most general sense, it means selecting individuals according to some purposive principle. For example, an observer who wishes to take a sample of oranges from a lot runs his eyes over the whole lot and then chooses average oranges—average in size, shape, weight or whatever other quality he may have in his mind. It has been claimed that the purposive method is more likely to give a typical or representative sample. But it may be pointed out that the method in most cases may involve some bias of unknown magnitude. Moreover, the method cannot provide an estimate of the error involved. Also, the method, although it may tell more about the mean of the population, would probably give a wrong idea about the variability since the observer has deliberately chosen values near the mean.

In a most restricted sense, the method refers to a particular sampling procedure adopted by Gini and Galvani with Italian census data. At one time, there was a good deal of controversy over the question whether this method provides a more representative sample than the random method. Suppose we want to estimate the population mean μ_y of y and suppose from census data the mean μ_x ,

of a control variable x correlated with y , is known. The method then consists in selecting a sample of size n by trial and error, for which the sample mean of x , say \bar{x}_n , is approximately equal to μ_x . That is, for the sample,

$$\bar{x}_n = \mu_x \pm e,$$

where e is a pre-assigned small quantity.

It is claimed that since y , the variable under enquiry, is correlated with x , the sample would provide an estimate of μ_y sufficiently near to its true value. The number of control variables may be more than one.

The method of purposive sampling as described above cannot provide an estimate of the standard error. Neyman, however, gave the method some benefit by making it probabilistic in allowing all the possible samples satisfying the requirement $\bar{x}_n = \mu_x \pm e$ to have equal probabilities of being selected. However, it was demonstrated by Neyman that even with this modification, the method is in general less efficient than stratified random sampling, stratification being with respect to the values of the control variable, if not less efficient than simple random sampling. It is only in a highly restricted class of situations, which rarely materialise in practice, that purposive sampling may give a more efficient estimator than stratified random sampling.

21.15 Sampling with probability proportional to size

In many surveys the sampling units vary in size. In surveys where household is the convenient unit, the household-size may vary from 1 to 25 or more. In a multistage sample survey, where the first-stage units are the villages, they may differ considerably in size, measured either by area or by population. So it is natural to suppose that a more representative sample will be obtained if a sample is taken with probability proportional to size (PPS) than a sample selected with equal probability. This technique has found its principal use in multistage sampling, but it is applicable in other situations too. In area sampling for yield determination, if we have areas demarcated on a map such as fields, fields may be selected with probability proportional to size by the simple procedure of locating random points on the map. In analogy with stratified

sampling, it may be said that under certain circumstances PPS sampling is expected to give greater precision of estimators than equal-probability sampling.

The practical procedure of PPS sampling consists in associating with each unit of the population a number of random numbers equal (or proportional) to its size. This is illustrated in the following example.

Ex. 21.5 There are 10 villages from which a sample of size 3 is to be taken with PPS, the measure of size being the village population. The population figures are shown in column 2 of the table below :

Village	Size	Cumulative total
1	165	165
2	690	855
3	1131	1986
4	907	2893
5	582	3475
6	2057	5532
7	973	6505
8	692	7197
9	1738	8935
10	988	9923

Cumulative total method

In this method we first take the cumulative totals of the population figures. Since the last cumulative total (i.e. the total population of all the villages taken together) is 9923, we choose 3 random numbers between 0001 and 9923. Supposing the first random number is 1705, it will mean that the first sample member is village 3, since 1705 lies between 855 and 1986.

Lahiri's method

D. B. Lahiri has provided a method of PPS selection that does not call for cumulation of the sizes. It requires at each drawing selecting at random one of the numbers, say u , from 1 to N and also, independently, one of the numbers, say v , from 1 to x_u (a number greater than or equal to the maximum of the sizes). In case $v \leq x_u$, the size of the u th unit, unit u will be included in the sample, while u will be rejected and the above process will be repeated in case $v > x_u$.

As a result, the probability that u will be selected in a given draw is $\frac{1}{N} \cdot \frac{x_u}{x_0}$. Since the probability that *some* unit will be selected at a given draw is

$$\sum_u \frac{1}{N} \cdot \frac{x_u}{x_0} = \frac{T_x}{Nx_0},$$

where $T_x = \sum_u x_u$, the probability for x_u to be chosen at an *effective* draw is

$$\frac{1}{N} \cdot \frac{x_u}{x_0} / \frac{T_x}{Nx_0} = x_u / T_x,$$

which is indeed proportional to x_u .

When the population size is large and a relatively large sample is to be taken, Lahiri's method will lead to a considerable saving in time.

If in a population there are N sampling units and if y_i and x_i be, respectively, the variate value and a measure of size of the i th sampled individual, then probability of selection for u th population unit is

$$p_u = \frac{x_u}{\sum_{u=1}^N x_u}.$$

Then, y_u denoting the value of y for the u th population unit,

$$E(y_i/p_i) = \sum_{u=1}^N \frac{y_u}{p_u} \cdot p_u = \sum_{u=1}^N y_u = Y \text{ (say),}$$

and $\text{var}(y_i/p_i) = E(y_i/p_i - Y)^2 = E(y_i^2/p_i^2) - Y^2$

$$= \sum_{u=1}^N (y_u^2/p_u^2) \times p_u - Y^2 = \sum_{u=1}^N y_u^2/p_u - Y^2.$$

Hence the best linear unbiased estimator of the *population total*, based on a sample of size n , is

$$T = \frac{1}{n} \sum_{i=1}^n y_i/p_i \quad \dots \quad (21.46)$$

and $\text{var}(T) = \frac{\text{var}(y_i/p_i)}{n} = \frac{1}{n} \left\{ \sum_{u=1}^N y_u^2/p_u - Y^2 \right\}. \quad \dots \quad (21.47)$

Since the sample values y_i/p_i are n independent unbiased estimators of Y with the same variance, an unbiased estimator of $\text{var}(T)$ is

$$\widehat{\text{var}}(T) = \frac{1}{n(n-1)} \left(\sum_{i=1}^n y_i^2/p_i^2 - nT^2 \right). \quad \dots \quad (21.48)$$

21.16 Quota sampling

If the sampling frames for the different strata into which the population may be divided are not available and are costly to construct, it may be possible to fix up a sample quota for each stratum and to continue sampling until the necessary quota for each stratum is filled up. The objective is to gain the benefits of stratification as far as possible without the high costs that may be incurred in any attempt to have recourse to probabilistic sampling. The method has been found useful in many socio-economic and opinion surveys. The method suffers from two major difficulties : (1) The method may involve biases due to non-response, because the non-responding individuals may come from a particular section of the population with some special characteristics ; (2) sampling theory cannot be applied to quota sampling, which contains no element of probability sampling.

21.17 Some mathematical methods for errors in measurement

Each sample observation is liable to errors of measurement. We do not expect that a sample observation will be equal to the correct value. What we do expect is that for a large sample these individual errors will cancel out and the mean value of the sample observations will approximate the mean of the true values.

Let u'_{ia} be the observed value of the i th sampling unit in the a th repetition, while u_i is its true value and ϵ_{ia} is the error ; i.e., let

$$u'_{ia} = u_i + \epsilon_{ia}. \quad \dots \quad (21.49)$$

If the sample is a random one, we expect

$$E_i(u'_{ia}) = u_i,$$

$$E_i(\epsilon_{ia}) = 0$$

and $E(u'_{ia}) = E(u_i) = \mu$, say, the population mean of true values,

where E_i denotes the conditional expectation for given i and E denotes the unconditional expectation.

In the above model, ϵ is called the *random sampling error* or, simply, the *sampling error*.

When the sampling is not perfectly random, another kind of error, which is called *bias*, may arise. This type of error is not stochastic with expectation zero, but it contributes a constant component of error to each sampling unit. In this case, we can write

$$u'_{ia} = u_i + \beta_i + \epsilon_{ia}, \quad \dots \quad (21.50)$$

where ϵ_{ia} is the stochastic component of error attributable to random sampling with $E(\epsilon_{ia})=0$ and β_i is the constant bias component with $E(\beta_i)=\beta$, say.

$$E(u'_{ia}) = u_i + \beta_i,$$

and

$$E(u'_{ia}) = \mu + \beta.$$

The component of error $E(u'_{ia}) - \mu = \beta$ is called the bias. The bias may be positive or negative. The essence of the bias is that it forms a constant component of error which does not decrease as the sample size increases, whereas the random sampling errors tend to cancel out as the sample size increases and the sample estimators generally converge in probability to the corresponding population values as the sample size increases.

The presence of constant bias (i.e. $\beta_i = \beta$ for all i) does not in general affect the variance of the estimators. Let us consider the variance of the sample mean \bar{u}'_n based on a sample of size n .

$$\begin{aligned} \text{var}(\bar{u}'_n) &= \text{var}(\bar{u}_n) + \text{var}(\bar{\beta}_n) + \text{var}(\bar{\epsilon}_n) \\ &= \frac{\sigma_u^2}{n} + \frac{\sigma_\epsilon^2}{n}, \end{aligned} \quad \dots \quad (21.51)$$

where σ_u^2 is the population variance of true value and σ_ϵ^2 is the population variance of random sampling error, provided ϵ_i is independent of u_i .

If, however, ϵ_i is correlated with u_i with correlation ρ , the formula (21.51) will reduce to

$$\text{var}(\bar{u}'_n) = \frac{1}{n} (\sigma_u^2 + \sigma_\epsilon^2 + 2\rho\sigma_u\sigma_\epsilon). \quad \dots \quad (21.52)$$

21.18 National Sample Surveys (NSS)

The National Sample Surveys are the biggest set of sample surveys in India being conducted by the Government of India. The NSS were initiated in 1950 to conduct sampling enquiries with a view to pro-

viding the Government and other organisations with socio-economic data which can be used for planning for national development and for research purposes. It is a continuing survey, being carried out in the form of rounds, the survey period in a round varying from 3 months to a complete year. The kind of data collected changes from one round to another and includes a variety of topics, like national income, consumer expenditure, small-scale industries, distribution of land-holdings, employment and unemployment, estimation of acreage and yield-rates of cereal crops, economic condition of agricultural labourers, etc. As such, it is a multipurpose survey, data on widely different topics being collected in the same survey. A multipurpose survey is more economical than a series of unipurpose surveys, provided that the enquiries to be included in a multipurpose survey are not so numerous and diversified as to overburden the investigators. In the NSS, the field work is done by specially trained investigators by personally interviewing sample households or persons or by direct observation or by harvesting crop in randomly located circular cuts in sample plots (in the case of crop surveys). The reference period may be a day, a week, a month or a year, depending upon the characteristic under consideration.

The Central Statistical Organisation (CSO) is responsible for deciding upon the coverage of the survey and the methodology to be used. The major portion of the field work is conducted by the Directorate of NSS, Government of India. The technical work relating to the NSS, the processing and analysis of data and the preparation of the final reports, was previously entrusted to the Indian Statistical Institute, but this too has now been taken over by the NSS Directorate.

The sample design in the NSS has also undergone changes from one round to another. The general sample design is a stratified two-stage one, where villages are the first-stage units and households and clusters of plots form the second-stage units, in socio-economic enquiries and crop surveys, respectively. In yield surveys, the crop-plots and circular cuts in them form the third-stage and fourth-stage units. Villages are generally selected by the circular systematic method, with equal probability after proper stratification and arrangement.

Two special features of the NSS may be mentioned :

(1) In the NSS, the practice has been to use a moving reference period, which is the day, the week, the month or the year preceding the date of investigation. This makes it possible to get estimates of averages over the whole period of the survey. For the characteristics which are subject to highly seasonal variation, these estimates of averages are more meaningful than those based on a fixed reference period. The method may also provide measures of seasonal variation for the characteristics under consideration.

(2) The NSS data are collected for two so-called independent *interpenetrating sub-samples*. The data are also collected by two teams of investigators. The method helps in analysing the total variation into components, such as sampling variation, variation due to investigators and interaction between investigators and samples.

Questions and exercises

21.1 Discuss the basic principles of sample surveys. What are the advantages of sample surveys over complete census ?

21.2 Discuss the different steps in a sample survey with special reference to any sample survey recently conducted in India.

21.3 Discuss the possible sources of bias in the following procedures :

(1) A basket of oranges is sampled by taking some oranges from the top.

(2) A mixture of sand and sugar is sampled by taking a quantity from the bottom.

(3) A sample of digits is taken by opening a page of five-figure logarithmic tables at random and taking down the last three digits of the logarithms of all numbers in the order in which they occur on that page.

(4) A sample of digits is taken by opening a page of a telephone directory at random and taking the digits in the telephone numbers in the order in which they occur on that page.

(5) Investigators collecting data on the size of families in a town conduct a house-to-house enquiry of the households selected

at random, during the working hours of the day, ignoring those houses from which there is no reply.

(6) A sample of opinions is obtained about a topical event by the mail questionnaire method.

21.4 What are random sampling numbers? Describe the different random sampling number series and describe their methods of construction. Describe the different tests for randomness generally applied to these series.

21.5 A population of N units is stratified into k strata, there being N_i units in the i th stratum. If n_i units are drawn at random without replacements from the i th stratum, the samples from the different strata being independent, obtain the best linear unbiased estimator for the population mean and the variance of the estimator.

Considering a linear cost function, $C = a_0 + \sum_{i=1}^k c_i n_i$, a_0 being the overhead cost and c_i the cost per unit for the i th stratum, obtain the optimum values of n_i 's such that for given cost the variance is minimised. Describe also the nature of the pilot survey to be undertaken in this case.

21.6 Distinguish between two-stage sampling and stratified random sampling.

For two-stage sampling, where the first-stage units are of equal size, obtain an estimator of the population mean. Also obtain the expression for the variance of the estimator. How will you estimate the variance from the sample?

What modifications are necessary if the first-stage units are of different sizes?

21.7 Describe the following methods of sampling with suitable examples :

- (1) systematic sampling,
- (2) multistage sampling,
- (3) multiphase sampling,
- (4) double sampling

and (5) purposive sampling.

21.8 Write a note on the nature, the coverage and the survey design of the National Sample Surveys of India.

21.9 The following are the marks obtained by a group of 43 students in a science test :

47	26	45	19	7	30	27	23	12
48	35	28	26	15	36	23	26	29
46	37	39	28	29	37	8	30	36
28	32	29	23	28	21	13	24	
37	38	22	27	32	24	20	13	

- (a) Draw a random sample of size 10 from this group
- (i) with replacements and (ii) without replacements.
- (b) In each case, give an estimate of the average number of marks per student in the whole group.
- (c) Also give in each case an estimate of the standard error.

21.10 In stratified sampling with replacements, suppose the strata sizes are equal and that there is equal allocation. Show that

$$\text{var}(\bar{y}_{st}) = \text{var}(\bar{y}) - \frac{1}{nk} \sum_{i=1}^k (\mu_i - \mu)^2,$$

where $\text{var}(\bar{y})$ is the variance of the unbiased estimator \bar{y} in the case of SRSWR. Hence comment.

21.11 Show how one can select one out of two units at random by tossing a biased coin twice. Extend this procedure for selecting one unit at random from (1) a population of 3 units, (2) a population of 4 units, (3) a population of 6 units. What is the least number of tosses that one will need to make in each of these cases?

21.12 Indicate how one can select at random a sample of 15 points (each co-ordinate being correct to the nearest millimetre) from the following regions :

- (a) a rectangular area whose sides are 98 cm. and 48 cm.;
- (b) an elliptical region defined by

$$\frac{x^2}{1024} + \frac{y^2}{625} \leq 1,$$

x and y being the distances of a point in cm., along the principal axes, from the centre. State the limitations of the methods, if any.

21.13 In selecting a sample of words from a dictionary, the following procedure is used : A page is selected at random from all the pages in the dictionary. Next, one of the two columns is chosen

at random. Third, a number R is drawn at random from 1 to M , a number greater than or equal to the maximum number of words in any column. If it is less than or equal to the number of words in the column, include the R th word of the column in the sample. Otherwise, repeat the operation, starting from the selection of a page till one word is chosen. Then repeat the entire procedure till n different words are obtained. Show that this procedure is equivalent to SRSWOR.

21.14 Give an outline of a sample survey you would conduct if you were to study the living conditions of college students in Calcutta.

21.15 Explain the method of sampling you would recommend for the following cases :

(a) To determine the average retail price of fish in the Calcutta markets.

(b) To determine the average yield of paddy in a district of West Bengal.

(c) To have a sample of middle-class families in Calcutta for a family-budget enquiry.

(d) To have a sample of newspaper readers in Calcutta for an opinion survey.

21.16 The N_i 's, σ_i 's (in kg.) and c_i 's (in Rs.) are given for 5 strata into which a population is divided in a certain crop survey. Obtain the optimum values of n_i 's and the corresponding variance of the estimator if the population mean is to be estimated and if the total approved cost of the survey is Rs. 5,000/- and the overhead cost is Rs. 550/-

i	N_i	σ_i (in kg.)	c_i (in Rs.)
1	3,780	28.5	3.50
2	5,260	18.6	2.75
3	8,200	27.6	2.25
4	4,160	27.2	3.00
5	2,980	16.8	2.50
		24,380	

Partial ans. $n_1=274$, $n_2=281$, $n_3=718$, $n_4=242$, $n_5=151$.

21.17 In a multistage survey, 11 first-stage units were selected, with 4 second-stage units in each first-stage unit and 8 third-stage units in each second-stage unit. The following mean squares were obtained—

MS between first-stage units : 335·6 ;

MS between second-stage units : 296·8 ;

MS between third-stage units : 134·2.

Evaluate the standard error of the sample mean.

Assuming a cost function of the form

$$C = 30·4 + 2·8m + 1·3mn + 0·6mn\bar{p},$$

where m , n and \bar{p} stand for the numbers of first-stage, second-stage and third-stage units, respectively, determine the optimum values of n and \bar{p} for a given cost. *Ans.* s.e. = 0·976, $n_{opt} = 6$, $\bar{p}_{opt} = 4$.

21.18 Show that (21.29) may be expressed as

$$\frac{\sigma^2}{n} \left[1 + (n-1)\rho_c \right],$$

where σ^2 is the population variance and ρ_c is the correlation coefficient between pairs of sample units (intraclass correlation coefficient).

Hence show that

$$-\frac{1}{n-1} \leq \rho_c \leq 1.$$

Show further that the relative efficiencies of systematic sampling compared to SRSWR and SRSWOR in estimating the population mean are, respectively,

$$\frac{1}{1 + (n-1)\rho_c} \text{ and } \frac{N-n}{N-1} \cdot \frac{1}{1 + (n-1)\rho_s}.$$

Hence indicate how the units in the population should be arranged in order that systematic sampling may be highly efficient.

21.19 Show that in (21.31), if the variance of ϵ_i is proportional to x_i^2 , then the estimate of the ratio R will be

$$\frac{1}{n} \sum_{i=1}^n (y_i/x_i).$$

21.20 The number of labourers x (in thousands) and the quantity of raw materials y (in lakhs of bales) are given below for 20 jute mills :

<u>Serial No. of Mill</u>	<u>x</u>	<u>y</u>
1	368	31
2	384	33
3	361	37
4	347	39
5	403	43
6	529	61
7	703	68
8	396	42
9	473	41
10	509	49
11	512	31
12	503	29
13	472	38
14	429	41
15	387	40
16	376	38
17	412	42
18	385	45
19	297	32
20	633	54

Draw a sample of 5 mills with PPS, taking x as the size. Estimate the total amount of raw material consumed by the 20 mills and also its standard error.

SUGGESTED READING

- [1] Cochran, W. G. *Sampling Techniques* (Chs. 1—3, 5—8, 10—13). John Wiley, 1963, and Wiley Eastern.
- [2] Deming, W. E. *Some Theory of Sampling* (Chs. 1, 2, 4—6). John Wiley, 1950.
- [3] Raj, D. *Sampling Theory*. McGraw-Hill, 1968, and Tata McGraw-Hill.
- [4] —— *The Design of Sample Surveys* (Chs. 1—10). McGraw-Hill, 1972.
- [5] Moser, C. A. *Survey Methods in Social Investigations*. William Heinemann, 1958.

- [6] Murthy, M. N. *Sampling Theory and Methods* (Chs. 1—3, 5, 7, 9—11, 13—15). Statistical Publishing Society, 1967.
- [7] Stuart, A. *Basic Ideas of Scientific Sampling*. Charles Griffin, 1962.
- [8] Yates, F. *Sampling Methods in Censuses and Surveys* (Chs. 1—3, 6—8). Charles Griffin, 1960.
- [9] Yule, G. U. and Kendall, M. G. *Introduction to the Theory of Statistics* (Chs. 16, 23). Charles Griffin, 1953.

PART FOUR

**METHODS FOR SOME
SPECIAL FIELDS OF APPLICATION**

22.1 Introduction

The term *vital statistics* signifies either the data or the methods applied in the analysis of the data which provide a description of the vital events occurring in given communities. By *vital events*, again, we mean such events of human life as birth, death, sickness, marriage, divorce, etc.

The raw data of vital statistics are generally obtained from the following sources :

(1) *Census*. Population censuses are now undertaken in almost all countries, generally at ten-year intervals. A census may be defined as an enumeration at a specified time of individuals inhabiting a specified area, during which particulars are collected regarding age, sex and some social, economic, ethnic or familial characteristics of the individuals.

(2) *Vital statistics registers*. In many countries, there is a system of registering the occurrence of every important vital event under legal requirement. For instance, when a child is born, the matter has to be reported to the proper authorities, together with such information as the age of mother, religion of parents, etc. Similarly, every death occurring in the community gets automatically recorded, because the disposal of the body requires a death certificate from the authorities.

Data on vital events may also be obtained from hospital records and from specially conducted population surveys.

In the following discussion, we shall be concerned with birth, death and sickness (or morbidity)—the three most important vital events. It will be assumed that we have from *census data* for the given community the total size of the population and also its distribution with respect to different characters (e.g. age and sex) corresponding to different *points of time*, while from *registers* we have data regarding the number of births and the number of deaths occurring during different *periods*. When it comes to the number of

cases of a disease or a group of diseases, as also the number of deaths therefrom, on the other hand, it will be assumed that the needed data have been obtained from *hospital records*.

In order to determine the population at a time (say t) subsequent to a census or between two censuses, one may use a number of procedures. A very common method is to make use of birth and death statistics as well as the statistics of immigration and emigration. The population P_t at time t is then obtained as

$$P_t = P_0 + (B - D) + (I - E),$$

where P_0 = population recorded at last census, B = total number of births in the intervening period, D = total number of deaths during the period, I = total immigration into the region during the period and E = total emigration from the region during the period.

When migration statistics are absent or unreliable, one would make use of a suitable mathematical formula for P_t , like the exponential law or the logistic law (*vide* Section 22.8), and then determine P_t by fitting this formula to the available census figures.

22.2 Rates of vital events

The raw data of vital statistics are given in the form of frequencies of vital events, perhaps classified according to certain characters such as age, sex, occupation, etc. These absolute numbers have numerous uses for administrative purposes. But to a statistician these raw materials alone will not be enough for an intelligent study of problems. The statement that in country A 20,000 people died in a certain year, while in country B 12,000 died in the same year conveys no particularly useful information. It is also necessary at least to know the population size of each country to have an idea as to their relative mortality situations. By relating the two—the number of deaths to the population size, we have a *rate* (in this case a *death rate*).

The general definition of a rate is as follows :

$$\text{Rate of a vital event} = \frac{\text{Number of cases of the vital event}}{\text{Total number of persons exposed to the risk of occurrence of the event}} \dots \quad (22.1)$$

It is obvious that a rate refers to (1) a particular type of vital event (e.g. birth or death), (2) a particular geographical region (e.g.

India or West Bengal) and (3) a particular period (e.g. the year 1970). The second and third points may not always be mentioned explicitly but may have to be understood from the context.

The number of persons exposed to the risk of a vital event is usually the population of the given area during the given period or some segment of that population. The population during any period, however, does not remain the same throughout. One will, therefore, use the population either at the beginning of the period or at the end. A more correct procedure would be to use the *mean population* during the period :

$$\frac{1}{t_2 - t_1} \cdot \int_{t_1}^{t_2} P_t dt, \quad \dots \quad (22.2)$$

where (t_1, t_2) denotes the given period, the population P_t being assumed for simplicity to be an integrable function of time t . The mid-period population

$$P_{(t_1+t_2)/2}$$

will give an approximation to this figure (and would be equal to this figure if P_t were a linear function of t).

A rate, according to the above definition, will be a proper fraction. For ease of understanding, the fraction is generally multiplied by a constant, which for most rates is 1,000. Vital statistics rates are thus generally expressed 'per thousand of population'.

A vital statistics rate is sometimes looked upon as an estimate of the probability that a person exposed to the risk of the vital event during the given period will actually experience the event. This interpretation cannot, however, be given to all such rates.

22.3 Measurement of mortality

22.3.1 Crude death rate

The simplest type of rate used in the measurement of mortality is the *crude death rate (CDR)*, which is defined as follows :

$$m = 1,000 \times \frac{D}{P}, \quad \dots \quad (22.3)$$

where m = crude death rate per 1,000 of population ;

D = number of deaths (from all causes) which occurred

in the population of the given region during the given period;

P =total population of the given region during the given period.

It is the most widely used of vital statistics rates. This rate has a simple interpretation, for it gives the number of deaths that occur, on the average, per 1,000 people in the community. Further, it is relatively easy to compute, requiring only the total population size and the total number of deaths. Besides, it is a probability rate in the true sense of the term. It represents the chance of dying for a person belonging to the given population, because the whole population may be supposed to be exposed to the risk of dying of something or other.

However, it has also some serious drawbacks. In using the *CDR*, we ignore the fact that the chance of dying is not the same for the young and the old or for males and females, and the fact that it may also vary with respect to race, occupation or locality of dwelling. Because of this, the *CDR* is unsuitable as an index of relative mortality in *different places* unless the populations of the places compared have substantially identical age- and sex-distributions, a condition which is seldom fulfilled. Thus a population composed of a high proportion of old people will naturally show a higher *CDR* than one with a high proportion of the young although, taken separately, they may have the same mortality in each of the two age-groups.

Under most circumstances, the *CDR* may well be used for comparing the mortality situations of the same place at *different times*, provided the periods compared are not too far apart, because in a stable, large community the age- and sex-composition of the population changes very slowly.

22.3.2 Specific death rate

A *specific death rate (SDR)* is a death rate computed for a specific segment of the community. Thus an *SDR* is given by

$$\frac{1,000 \times \frac{\text{Number of deaths which occurred in the specified section of the population during the given period in the given region}}{\text{Total number of persons in the specified section of the population in the given period in the given region}}}{\dots} \quad (22.4)$$

Usually, death rates are made specific only with respect to age and sex. If ${}_nD_x$ is the number of deaths between ages x and $x+n-1$ last birthday (or l.b.d.) among residents in a community during a period, and if ${}_nP_x$ is the number of persons in the same age-group in the community during the period,* then the *age-specific death rate* for the age-group is

$${}^n m_x = 1,000 \times \frac{{}_n D_x}{{}_nP_x}. \quad \dots \quad (22.5)$$

The formula for an *annual* age-specific death rate (for which $n=1$) is written simply as

$$m_x = 1,000 \times \frac{D_x}{P_x}, \quad \dots \quad (22.6)$$

where D_x = number of deaths among persons aged x l.b.d.;

P_x = number of persons aged x l.b.d.

Let ${}_nF_x$ and ${}_nD_x$ denote the number of males aged x to $x+n-1$ and the number of deaths occurring to such males. Then the *SDR* for males aged between x and $x+n-1$ will be

$${}^n m_x = 1,000 \times \frac{{}_n D_x}{{}_n P_x}. \quad \dots \quad (22.7)$$

This is a death rate specific for both *age* and *sex*. The age-specific death rates for females are defined in a similar manner.

The *SDRs* are the true and best measures of mortality, because they furnish a really meaningful idea of the probability that a person of a certain specified kind will die within the given period. For general purposes, death rate specific for age and sex is one of the most widely used types of death rate. It also supplies one of the essential components required for constructing *life tables* and *net reproduction rates* (*vide* Sections 22.4 and 22.6).

Specificity by age and sex eliminates differences in death rates arising from variation in population composition in respect of these characters. To this extent, such *SDRs* can be compared from one geographical area to another. However, this does not eliminate differences due to other characters which might also be important.

*In each such symbol, the *lower suffix* denotes the beginning of the particular age-interval and the *lower prefix* the width of the interval; the *upper prefix*, if any, denotes a particular sex and the *upper suffix*, if any, a particular community.

In order to get a clear insight into the forces of mortality, death rates ought to be made specific for some other factors, besides age and sex, e.g. race (white, non-white, etc.), occupation and locality of dwelling (urban and rural).

We give below the *CDR* and *SDRs* for rural India for the year 1957-58, which have been estimated from National Sample Survey data.

TABLE 22.1
CRUDE DEATH RATE AND DEATH RATES SPECIFIC FOR
AGE AND SEX FOR RURAL INDIA FOR 1957-58

Age-group	Death rate per thousand persons		
	Males	Females	All
0	198·0	182·5	190·3
1—4	42·6	45·4	44·0
5—14	5·5	5·5	5·5
15—24	3·5	5·4	4·5
25—34	4·2	6·4	5·3
35—44	5·8	5·4	5·6
45—54	12·8	8·0	10·5
55—64	32·2	21·0	26·6
65—	72·9	54·7	63·5
All ages	19·6	18·8	19·2 (=CDR)

Source : National Sample Survey (Report No. 76)—*Fertility and Mortality Rates in India*.

The table shows that mortality is high among infants, then it decreases with increasing age and attains a minimum somewhere in the age-group 15-24. But from then on, it rises steadily until it reaches a peak in the old ages. This is true for both males and females and also applies to the populations of almost all countries. Secondly, at most ages the mortality for males is seen to be greater than that for females. This, again, is almost a universal phenomenon.

22.3.3 Standardised death rate

To study the differences in the mortality experiences of two communities, or even in the mortality experiences of the same community over two periods lying wide apart, it is necessary to compare their *SDRs*, specificity being achieved with respect to such characters as age, sex, etc. This procedure, however, involves an unwieldy mass of data whose significance may not be readily grasped. Secondly, one series may have higher *SDRs* than the other for some of the segments, but lower *SDRs* for the other segments. In such a case, one will not be able to make a general statement of the form : "Mortality is higher (or lower) in *A* than in *B*."

What is wanted, then, is a single index of mortality—some sort of average of the death rates for the various segments of the population. The simplest composite figure of this type is, of course, the *CDR*. Assuming that specificity is achieved with respect to age alone, the *CDRs* for *A* and *B* may be written as

$$m^a = \frac{\sum m_x^a \cdot P_x^a}{\sum P_x^a} \text{ and } m^b = \frac{\sum m_x^b \cdot P_x^b}{\sum P_x^b}$$

However, m^a and m^b are not comparable, as has been pointed out in Section 22.3.1. For m^a and m^b may be unequal even when $m_x^a = m_x^b$ for each x , simply because the proportions

$$P_x^a / \sum P_x^a \text{ and } P_x^b / \sum P_x^b$$

may not be the same, i.e. because the age-distributions of the two populations may not be identical.

To eliminate this defect, it is necessary to use for both *A* and *B* the same set of weights in taking weighted averages of the series of *SDRs*. This is done by considering a third population, called a *standard population*. Supposing the number of persons of age x in the standard population is P'_x , the weighted average of the *SDRs* for *A*, called the *standardised death rate* or *adjusted death rate* (*STDR*) for *A*, will be

$$\sum m_x^a \cdot P'_x / \sum P'_x. \quad .. \quad (22.8)$$

This *age-adjusted* death rate is the *CDR* which would be observed in the standard population if it experienced the age-specific death rates of the community in question. A death rate may be adjusted for

characters other than age and may be similarly interpreted. For instance, in case the death rates for A are specific for both age and sex, the $STDR$ for A is

$$\therefore \left[\sum_s m_s^a \cdot {}^m P_s^i + \sum_s m_s^a \cdot {}^f P_s^i \right] / \left[\sum_s {}^m P_s^i + \sum_s {}^f P_s^i \right]. \quad \dots \quad (22.9)$$

An $STDR$ is easy to compute and to explain. Further, if the $SDRs$ of one community are proportional to those of the other, this will also be reflected in their $STDRs$. On the other hand, the choice of the standard population may influence the comparison of two $STDRs$. However, this difficulty will not be serious if the standard chosen is not far removed in its population composition from the communities being compared. The usual procedure is to take as standard the actual population (or the *life-table stationary population*) of a bigger community of which A and B are parts. For instance, in comparing Assam and West Bengal in respect of mortality, one may take the population of the whole of India or of Eastern India as standard.

Indirect standardisation

Besides the direct method of computing $STDR$ by using, say, formula (22.8), there is an indirect method. The use of formula (22.8) requires that the number of persons and the $SDRs$ for all segments of the given population (say A) be known. In some cases, however, we may have a population classified according to age, for instance, but the $SDRs$ for the individual age-groups may not be available. Only the total number of deaths, and hence the CDR , may be known.

In such a case, let the age-specific death rates for the standard community be given, which are denoted by m_s^i .

Let us look for an adjustment factor C such that

$$CDR \times C = STDR,$$

$$\text{i.e. } \frac{\sum_s m_s^a P_s^i}{\sum_s P_s^i} \times C = \frac{\sum_s m_s^a P_s^i}{\sum_s P_s^i}.$$

Obviously, this C is to be equal to

$$\frac{\sum_s m_s^a P_s^i / \sum_s P_s^i}{\sum_s m_s^a P_s^i / \sum_s P_s^i}.$$

But this factor cannot be evaluated exactly with the type of data we have in hand, since m_x^a for individual ages x are not available. The usual practice is, therefore, to replace the unknown m_x^a by the known figures m_x^s . C is thus approximated by

$$C' = \frac{\sum m_x^s P_x^s / \sum P_x^s}{\sum m_x^s P_x^a / \sum P_x^a} \quad \dots \quad (22.8a)$$

and, correspondingly, the $STDR$ is approximated by

$$CDR \times C'. \quad \dots \quad (22.8b)$$

The computation of the $STDR$ by adjusting the CDR in this manner is called *indirect standardisation* of specific death rates.

In general, the indirect method leads to almost the same value for the $STDR$ as the direct method would. And the two methods would be exactly equivalent if the specific death rates for the given community happened to be proportional to the specific death rates for the community taken as standard.

In the following table, we have the specific death rates for rural Madras and rural Madhya Pradesh for 1957-58, specificity being achieved with respect to both age and sex. These are taken from *Fertility and Mortality Rates in India* of the National Sample Survey (Report No. 76). For comparing these two sets of rates, we may take as standard the life-table stationary population for the whole of rural India, 1957-58, as given in Tab^l 22.6. The population figures are given in cols. (5)-(6) of the following table, the figures being reduced to a cohort of $I_0=1,000$ for both males and females.

The $STDR$ for rural Madras is then the weighted average of the figures in cols. (1) and (2), the life-table figures in cols. (5) and (6) being taken as the weights. This is

$$(736,415.0 + 667,982.2) / (45,232 + 46,571) = 1,404,397.2 / 91,803 \\ = 15.3 \text{ (per thousand).}$$

Similarly, the $STDR$ for rural Madhy.. Pradesh is

$$(1,183,953.8 + 1,054,009.7) / 91,803 = 2,237,963.5 / 91,803 \\ = 24.4 \text{ (per thousand).}$$

TABLE 22.2
SPECIFIC DEATH RATES FOR RURAL MADRAS AND RURAL
MADHYA PRADESH AND LIFE-TABLE STATIONARY
POPULATION FOR RURAL INDIA, 1957-58

(0) Age l.b.d.	Specific death rates for rural Madras		Specific death rates for rural Madhya Pradesh		Life-table stationary population for rural India	
	(1) Male	(2) Female	(3) Male	(4) Female	(5) Male	(6) Female
0	163.9	140.6	203.8	165.4	897	903
1—4	26.2	26.3	50.4	60.2	3,104	3,116
5—14	3.0	4.2	5.9	5.1	7,174	7,194
15—24	2.7	3.1	3.1	5.2	6,871	6,812
25—34	4.4	3.6	4.3	6.4	6,616	6,447
35—44	5.8	5.1	7.3	7.3	6,317	6,089
45—54	15.9	8.1	12.4	8.5	5,797	5,690
55—64	22.6	8.9	48.8	34.4	4,655	4,945
65—	53.9	49.0	107.2	63.4	3,801	5,375
Total	—	—	—	—	45,237	46,571

The two *STDRs* indicate that the age-sex specific death rates of rural Madras would result on the average in about 15 deaths per 1,000 if they *operated* on the life-table stationary population of the whole of rural India, while the age-sex specific death rates of rural Madhya Pradesh would result on the average in about 24 deaths per thousand. A precise idea is thus obtained from the *STDRs* regarding the comparative mortality situations of the two regions.

22.3.4 Comparative mortality index

The use of the *STDR* in making a comparison of mortality over time sometimes gives rise to difficulties. In this case the population of a part of the time period, generally at the start of the period, would be taken as standard. But the resulting *STDR* values may give an unrealistic picture, for the age-sex distribution of the current population may be widely different from that of the

standard. The comparative mortality index (*CMI*) has been introduced to meet this objection. Here use is made of a shifting set of weights in taking a weighted average of *SDRs*. Thus the *CMI* for a given period will be given by the formula

$$CMI = \sum_{x} w_x m_x / \sum_{x} w_x m_x^*$$

where

$$w_x = \frac{1}{2} \left[\frac{P_x^*}{\sum_x P_x^*} + \frac{P_x}{\sum_x P_x} \right], \quad \dots \quad (22.10)$$

P_x^* and P_x being the population figures at age x for the standard and the given period, respectively, and m_x^* and m_x the *SDRs* at age x for the periods

We may be required to compare the mortality of a community in successive years. This may be achieved by forming ratios of the corresponding *CMIs*.

22.3.5 Cause-of-death rate

This rate is used to measure the contribution to the total mortality of a community that is made by a specified cause of death, say a specified disease or accidents.

The (crude) cause-of-death rate for cause i , denoted by m^i , is, by definition,

$$m^i = 100,000 \times \frac{D^i}{P}, \quad \dots \quad (22.11)$$

where D^i =total number of deaths from cause i occurring in the given period in the given community

and P =total population of the given community in the given period.

This rate has the multiplier 100,000, instead of the usual 1,000, so that in any given case the computed rate does not appear as a small fraction.

It is an over-all index of the attrition of the population as a whole from the given cause. Further, it is the measure that serves as the basis for many public-health programmes and also as an index of their success and failure. Moreover, it is simple to calculate.

However, it suffers from the same defect as the crude death rate does, for it does not take into account the age-sex composition of the population. Second, cause of death being subject to the greatest

degree of reporting errors, the computed rate is also likely to be highly unreliable. What is more, unlike the *CDR*, it is not a probability rate, for the whole population may not always be regarded as the population exposed to the risk of death from a given cause. For instance, in the case of lung cancer, which is an old-age disease, only the population above, say, 45 years of age should be so regarded.

22.3.6 Maternal mortality rate

This rate is defined by the formula

$$1,000 \times \frac{D_p}{B}, \quad \dots \quad (22.12)$$

where D_p = total number of deaths from puerperal causes among the female population in the given period in the given community

and B = total number of live births occurring in the given period in the community.

This rate may be looked upon as an alternative to, or a refined version of, the corresponding cause-of-death rate.

First, here note is taken of the fact that only the part of the female population that goes through conception some time during the period, and not the whole population, is exposed to the risk of dying from puerperal causes (i.e. causes relating to child-birth). This population may be taken to be approximately the number of mothers giving birth to live-born children plus the number of those delivered of dead foetuses. Now, foetal deaths are almost universally poorly registered. Moreover, most countries do not maintain data on the number of mothers but rather on the number of live births. These are the reasons why maternal mortality rate has as its denominator the number of live births.

As against its merit as a measure of the effect of puerperal diseases on the mortality of women, this rate may often be erroneous. For one thing, puerperal causes of death are generally subject to a large margin of reporting errors. Secondly, live births are generally subject to a greater degree of under-registration than maternal deaths. As such, the maternal mortality rate will tend to be overstated to some extent. The effect of the overstatement, however, generally happens to be minor.

22.3.7 Infant mortality rate

The infant mortality rate (*IMR*), too, is an alternative to, and in a sense an improvement upon, the age-specific death rate for age 0 l.b.d.—in other words, upon the death rate for infants (i.e. children under 1 year of age). It is defined as

$$IMR = 1,000 \times \frac{D_0}{B}, \quad \dots \quad (22.13)$$

where D_0 =number of deaths among children of age 0 l.b.d.

and B =number of live births.

The age-specific death rate for age 0 l.b.d., which has the same numerator, has for its denominator the number of infants. However, it is well known that infants are grossly under-enumerated in a population census. As such, the age-specific death rate tends to be highly overstated. Moreover, estimates of population by age are seldom obtainable annually. This is why the *IMR* is generally used, in lieu of the *ASDR* m_0 , as the measure of infant mortality.

The *IMR* has a number of advantages. It does away with the need for the data of population censuses or estimates. For the same reason the *IMR* can be computed for any population and for any time period, provided only the number of infant deaths and the number of live births are available. The same cannot be said of the corresponding *ASDR*, for in the case of a small area an estimate of the population of age 0 l.b.d may not be found. The *IMR* has been called the most *sensitive* of all measures of mortality. For in most countries the great risk of death under 1 year of age is not equalled at any other part of the life span, except at very old ages. But unlike deaths at very old ages, infant deaths are highly responsive to improvements in environmental and medical conditions. No wonder, then, the *IMR* serves as an excellent index of the general healthiness of the community.

As to its drawbacks, it will be apparent that the *IMR* is not a probability rate in the true sense of the term. For the numerator and denominator of the *IMR* are not strictly related. The deaths under 1 year in a given calendar year include those of some children born in the previous year; moreover, some of the deaths among the current year's births during the first year of life may occur in the following calendar year. (Another way of putting this is to say that

a child born, say, on January 1 remains exposed to the risk of death under 1 year of age (in the current year) for a full one year, while one born on December 1 remains so exposed for 31 days only.) If fertility and mortality are stable, these two types of errors tend to cancel each other, but their effect may be considerable when fertility and mortality are changing fairly rapidly. The more serious drawback arises from the under-registration of live births. The definitions of live birth and still birth vary from country to country and also over time. There is also found a reluctance to register as live-born those infants who, though born alive, die immediately after birth. Live births are thus under-registered, while infant deaths are more completely registered. This leads to an *IMR* being larger than what it should be. This is why it has been said that it is possible to lower the *IMR* without saving a single life simply by improving the birth registration system.

The following table shows the *IMRs* for a number of countries of the world for the year 1969 and brings into clear relief the abject backwardness of India in the field of health and hygiene.

TABLE 22.3
INFANT MORTALITY RATES FOR SOME
COUNTRIES FOR THE YEAR 1969

Country	<i>IMR</i> per 1,000 live births
Australia	18·0
Japan	15·0
India	139·0
UAR	117·0
Ghana	156·0
United Kingdom	18·8
Sweden	12·9
Canada	22·0
USA	21·2
Guatemala	89·0
Chile	100·0

22.3.8 Case fatality rate

As the name indicates, this rate is intended to measure the fatality, or importance as a killer, of a given disease. The formula for the rate is

$$1,000 \times \frac{D^i}{C^i}, \quad \dots \quad (22.14)$$

where D^i = number of deaths among cases of the disease i
and C^i = total number of cases of the disease i .

Provided age, sex, occupation, etc., are taken into account in its computation, this may be regarded as the most refined specific death rate. For, in the strictest sense, those who have a specified disease are the ones truly exposed to the risk of dying of that disease. Further, it is a truly probability rate. The case fatality rate for T.B., e.g., represents the probability that a person suffering from T.B. in a given period will die of that disease in that period. Because of its bearing on prognosis, this rate is of the greatest interest to clinicians.

However, the computation of this rate is often beset with difficulties because of the non-availability of the relevant data. Generally, these rates are computed on the basis of the case records of big hospitals. But rates computed in this way are to be taken with a pinch of salt. For one thing, the cases of a disease that are treated in a hospital are the more serious ones, so that case fatality rates computed from hospital data tend to be unduly high. On the other hand, the type of treatment given in a hospital is often different from the average treatment given outside. This too may make the case fatality rate from hospital data different from the true rate for the community at large.

22.4 Life table

Suppose an investigator, who is studying the mortality prevailing in a community during a given period, asks : "If 100,000 babies born at the same time experienced throughout their lifetime the given mortality, how many would reach age 1, how many would reach age 10, 20, 30, etc.? Further, when the life of all these 100,000 would run its course, what would be the average longevity per person?" The answers to such questions are given in a *life*

table. A life table thus presents in a more vivid way than the simple death rates can the mortality experience of a community during a given period.

22.4.1 Description

A life table gives, for integral values of age in years (denoted by x), the values of the following functions :

(1) l_x , the number of persons who attain (or rather are expected to attain) exact age x out of an assumed number of births l_0 (called the *cohort* or *radix* of the life table).

(2) d_x , the number of persons, among the l_x persons reaching age x , who die before reaching age $x+1$. Thus

$$d_x = l_x - l_{x+1}.$$

(3) q_x , the probability that a person of exact age x will die before reaching age $x+1$. It follows that

$$q_x = d_x / l_x.$$

Some tables contain, besides q_x , another function $p_x = 1 - q_x$, which is the probability that a person of precise age x will survive till his next birthday.

(4) L_x , the number of years lived, in the aggregate, by the cohort of l_0 persons between ages x and $x+1$. Thus

$$L_x = \int_0^1 l_{x+t} dt,$$

an approximate value of which is given by

$$\frac{l_x + l_{x+1}}{2} = l_x - \frac{1}{2} d_x,$$

provided l_{x+t} is approximately a linear function of t between $t=0$ and $t=1$.

Since the width of the interval $(x, x+1)$ is unity, it is clear that L_x may also be interpreted as the average size of the cohort between ages x and $x+1$.

*This is equivalent to the condition that the d_x deaths occurring in the age-interval $(x, x+1)$ are approximately uniformly distributed over this interval.

The function L_x may be interpreted in yet a third way. Suppose in a community every year there are exactly l_0 births, these being distributed uniformly throughout the year, and that the death rate at each age remains the same—same as that given by the q_x column of the life table. Further, let there be no migration. Under these conditions, ultimately (after 100 years or thereabouts) the population will be of the same size from year to year and will have the same age-distribution, the number of persons between ages x and $x+1$ being always given by L_x . A population with constant size and constant age-composition or constant age- and sex-composition over time is called *stationary*. The L_x column is, therefore, said to give the age-distribution of the *life-table stationary population*.

[The idea of a *stable population* is closely related to that of a stationary population. A population is said to be stable if it has a fixed age- and sex-distribution and if the same mortality and fertility are experienced at each age, it being assumed that there is no migration. For a stable population the over-all birth and death rates must remain constant. Hence the rate of increase of the population must also be constant for such a population, so that the compound interest law of growth will be applicable.

If the over-all birth and death rates in a stable population happen to be equal, so that the size of the population also remains constant, then the stable population becomes a stationary population.]

(5) T_x , the number of years lived by the cohort after attaining age x or the total future lifetime of the l_x persons who reach age x . We have, then,

$$T_x = L_x + L_{x+1} + L_{x+2} + \dots$$

(6) e_x^0 , the average number of years lived after age x by each of the l_x persons who attain that age. It is called the (complete) expectation of life at age x and is obtained from the relation

$$e_x^0 = \frac{T_x}{l_x}.$$

e_0^0 , the expectation of life at age 0, is the average age at death, or the average longevity, of a person belonging to the given community.

[A closely related concept is that of the *curiate expectation of life*, denoted by e_x , which represents the average number of *complete* years of life lived after age x by any of the l_x persons who attain age x . We have

$$e_x = \sum_{t=1}^{\infty} l_{x+t}/l_x,$$

so that

$$e_x^0 \simeq e_x + \frac{1}{2}.]$$

22.4.2 Construction of a life table

The pivotal column of a life table is the q_x column, as will be apparent from the following discussion. Suppose we have the value of q_x for every x from 0 upwards. We can then start with a suitable cohort, say one of 100,000 (l_0) births. Multiplying l_0 by q_0 , we get $l_0 q_0 = d_0$. Then $l_1 = l_0 - d_0$. Again, $d_1 = l_1 q_1$, $l_2 = l_1 - d_1$, and so on. Having obtained the values in the l_x column, we can then fill in the other columns, viz. L_x , T_x (for which we start from the bottom of the table and get the values successively by using the relation $T_x = L_x + T_{x+1}$) and e_x^0 , by means of the relations stated above.

If the probability that a person belonging to the age-group x to $x+1$ will die while in that age-group is denoted by m'_x , then

$$m'_x = \frac{d_x}{L_x} \simeq \frac{d_x}{l_x - \frac{1}{2}d_x},$$

i.e. $m'_x \simeq \frac{2q_x}{2-q_x},$

or $q_x \simeq \frac{2m'_x}{2+m'_x}.$

The probabilities m'_x are estimated by the observed age-specific death rates (m_x) for the community, where we now take

$$m_x = D_x/P_x \text{ (without the multiplier 1,000).}$$

Hence the q_x values can be determined, if the m_x values are known, by using the approximate relation

$$q_x \simeq \frac{2m_x}{2+m_x}. \quad \dots \quad (22.15)$$

For the early years of life, the values of m_x are usually not so reliable owing to defects in census records. Besides, the assumption underlying (22.15) that deaths are distributed uniformly over the years of age is not valid for the early ages, especially for age 0 : mortality is generally very high in the first few weeks after birth and then it diminishes sharply. It is, therefore, necessary to have alternative formulæ for q_x for $x=0, 1, 2$, say. We shall consider an alternative formula for q_0 based on registration data alone. Here the assumption will be made that the effect of migration is negligible, which is probably legitimate at age 0. This formula is due to Kuczynski [6].

Note that in order to survive the first year of age, a child must survive till the end of the calendar year in which it is born and then live long enough in the next calendar year to attain exact age 1. Hence, denoting the probabilities of these two events by p' and p'' , respectively, we have

$$p_0 = p' \cdot p''. \quad \dots \quad (22.16)$$

The probabilities p' and p'' are estimated by

$$\text{and } \left. \begin{array}{l} (B_0 - D'') / B_0 \\ (B^{-1} - D' - D'') / (B^{-1} - D'), \end{array} \right\} \quad \dots \quad (22.17)$$

respectively, where

B^{-1} =number of children born in the preceding calendar year,

B_0 =number of children born in the current calendar year,

D' =number of children born and deceased in the preceding calendar year,

D'' =number of children born in the preceding calendar year and deceased in the current calendar year before reaching age 1 and

D''' =number of children born and deceased in the current calendar year.

More elaborate formulæ for q_x at young ages are given in the book by Anderson and Dow [1] (Ch. 20).

For the sake of illustration, we give below (on pp. 201-202) the

life tables for India, for the decade 1951-60, separately for males and females. (It should be noted that in these tables L_0 is not even approximately equal to $(l_0 + l_1)/2$. It is computed by a more complicated formula, because the assumption of uniform distribution of deaths, underlying the approximation $L_x \approx (l_x + l_{x+1})/2$, is not at all legitimate for $x=0$.)

22.4.3 Abridged life table

The type of life table considered above, where the age-interval is a year throughout the table and the various functions are evaluated for every year of age, is customarily called *complete life table*. As opposed to this type of table, there are *abridged life tables*. The abridgement may be of two kinds. In the first form of abridgement, the functions are evaluated for single years of age, as in a complete table, but these are now given, for the greater part of the table, at intervals of 5 years or 10 years. In the second form, the function values are stated, for the major part of the table, for 5-year or 10-year age-groups, and hence this type is obtained through a condensation of a complete table rather than through the omission of some of its rows.

We shall discuss some methods of constructing abridged tables. The method of G. King is meant for the first type of abridgement, while the method of T.N.E. Greville and the one due to L. J. Reed and M. Merrell are intended for the second type.

22.4.4 King's method

Suppose the life table functions q_x , l_x and e_x^0 are to be given at 5-year intervals in the abridged table. Then the first step would be to compute probabilities of death q_x at the pivotal ages by the usual procedure. Next, one has to form

$$p_x = 1 - q_x$$

for the pivotal ages.

To evaluate the next life table function, l_x , at the pivotal ages, we note that

$$l_{x+5} = l_x \times {}_5 p_x \quad \text{or} \quad \log l_{x+5} = \log l_x + \log {}_5 p_x,$$

so that it is necessary to estimate ${}_5 p_x$ from the available p_x values.

TABLE 22.4
ALL-INDIA LIFE TABLE—MALES
(1951-60)

<i>x</i>	<i>l_x</i>	<i>d_x</i>	<i>q_x</i>	<i>L_x</i>	<i>T_x</i>	<i>e_x</i> ⁰
0	100000	15322	.15322	88509	4188830	41·89
1	84678	2552	.03014	82404	4100321	48·42
2	82126	1950	.02374	80401	4017917	48·92
3	80176	1473	.01837	78886	3937513	49·11
4	78703	1098	.01393	77751	3838627	49·03
5	77605	807	.01046	77202	3780876	48·72
6	76798	588	.00765	76504	3703674	48·23
7	76210	428	.00562	75996	3627170	47·59
8	75782	321	.00423	75622	3551174	46·86
9	75461	255	.00338	75334	3475552	46·06
10	75206	226	.00300	75093	3400218	45·21
11	74980	226	.00301	74867	3325125	44·35
12	74754	247	.00330	74631	3250258	43·48
13	74507	291	.00391	74362	3175627	42·62
14	74216	358	.00483	74037	3101265	41·79
15	73858	367	.00497	73675	3027228	40·99
16	73491	371	.00505	73306	2953553	40·19
17	73129	374	.00512	72933	2880247	39·59
18	72746	378	.00520	72557	2807314	38·59
19	72368	381	.00527	72178	2734757	37·79
20	71987	384	.00533	71795	2662579	36·99
21	71603	391	.00546	71408	2590784	36·18
22	71212	402	.00564	71011	2519376	35·38
23	70810	413	.00583	70604	2448365	34·58
24	70397	424	.00603	70185	2377761	33·78
25	69973	437	.00625	69755	2307576	32·98
26	69536	451	.00649	69311	2237821	32·18
27	69085	467	.00676	68852	2168510	31·39
28	68618	484	.00706	68376	2099658	30·60
29	68134	505	.00741	67882	2031282	29·81
30	67629	534	.00790	67362	1963400	29·03
31	67095	582	.00867	66804	1896038	28·26
32	66513	631	.00949	66198	1829234	27·50
33	65882	685	.01040	65540	1763036	26·76
34	65197	740	.01135	64827	1697496	26·04
35	64457	798	.01238	64058	1632669	25·33
36	63659	859	.01349	63230	1568611	24·64
37	62800	921	.01466	62340	1505381	23·97
38	61879	981	.01585	61389	1443041	23·32
39	60898	1030	.01691	60383	1381652	22·69
40	59868	1074	.01794	59331	1321269	22·07
41	58794	1115	.01897	58237	1261938	21·46
42	57679	1154	.02001	57102	1203701	20·87
43	56525	1190	.02106	55930	1146599	20·24
44	55335	1223	.02214	54723	1090669	19·71
45	54110	1257	.02323	53482	1035946	19·15
46	52853	1287	.02435	52210	982464	18·59
47	51566	1317	.02554	50908	930254	18·04
48	50249	1347	.02681	49576	879346	17·50
49	48902	1377	.02816	48214	829770	16·97

TABLE 22.4 (Contd.)

x	l_x	d_x	q_x	L_x	T_x	e_x^0
50	47525	1407	.02961	46822	781556	16.45
51	46118	1437	.03117	45400	734734	15.93
52	44681	1467	.03283	43948	689334	15.43
53	43214	1494	.03458	42467	645386	14.93
54	41720	1519	.03642	40961	602919	14.45
55	40201	1542	.03836	39430	561958	13.98
56	38659	1562	.04040	37878	522528	13.52
57	37097	1578	.04255	36308	484650	13.06
58	35519	1591	.04480	34724	448342	12.62
59	33928	1600	.04716	33128	413618	12.19
60	32328	1605	.04964	31526	380490	11.77
61	30723	1605	.05224	29921	348964	11.36
62	29118	1600	.05496	28318	319043	10.96
63	27518	1591	.05780	26723	290725	10.56
64	25927	1576	.06077	25139	264002	10.18
65	24351	1556	.06390	23573	238863	9.81
66	22795	1532	.06721	22029	215290	9.44
67	21263	1503	.07069	20512	193261	9.09
68	19760	1469	.07433	19026	172749	8.74
69	18291	1430	.07816	17576	153723	8.40
70	16861	1386	.08218	16168	136147	8.07
71	15475	1337	.08639	14807	119975	7.75
72	14138	1284	.09081	13496	105172	7.44
73	12854	1227	.09545	12241	91676	7.13
74	11627	1166	.10300	11044	79435	6.83
75	10461	1102	.10539	9910	68391	6.54
76	9359	1036	.11072	8841	58481	6.25
77	8323	968	.11631	7839	49640	5.96
78	7355	898	.12215	6906	41801	5.68
79	6457	828	.12826	6043	34895	5.40
80	5629	758	.13466	5250	28852	5.13
81	4871	689	.14135	4527	23602	4.85
82	4182	622	.14884	3871	19075	4.56
83	3560	561	.15764	3280	15204	4.27
84	2999	505	.16826	2747	11924	3.98
85	2494	452	.18121	2268	9177	3.68
86	2042	402	.19700	1841	6909	3.38
87	1640	354	.21614	1463	5068	3.09
88	1286	308	.23914	1132	3605	2.80
89	978	261	.26651	848	2473	2.53
90	717	214	.29876	610	1625	2.27
91	503	169	.33640	419	1015	2.02
92	334	127	.37994	271	596	1.78
93	207	89	.42989	163	325	1.57
94	118	57	.48676	90	162	1.37
95	61	34	.55106	44	72	1.18
96	27	47	.62330	19	28	1.04
97	10	7	.70399	7	9	0.90
98	3	2	.79364	2	2	0.67
99	1	1	.89276	1	—	—

Source : *Life Tables, 1951-60, Census of India, 1961 Census. Registrar-General, India.*

TABLE 22.5
ALL-INDIA LIFE TABLE—FEMALES
(1951-60)

<i>x</i>	<i>l_x</i>	<i>d_x</i>	<i>q_x</i>	<i>L_x</i>	<i>T_x</i>	<i>e_x⁰</i>
0	100000	13826	·13826	89631	4055487	40·55
1	86174	3119	·03620	86390	3965856	46·02
2	83055	2378	·02863	80950	3882466	46·75
3	80677	1797	·02227	79100	3801516	47·12
4	78880	1343	·01702	77708	3722416	47·19
5	77537	991	·01278	77042	3644708	47·01
6	76546	723	·00945	76185	3567666	46·61
7	75823	527	·00695	75560	3491481	46·05
8	75296	391	·00519	75101	3415921	45·37
9	74905	305	·00407	74753	3340820	44·60
10	74600	261	·00350	74470	3266067	43·78
11	74339	251	·00338	74214	3191597	42·93
12	74088	267	·00361	73955	3117383	42·08
13	73821	310	·00420	73666	3043428	41·23
14	73511	380	·00517	73321	2969762	40·40
15	73131	388	·00530	72937	2896441	39·61
16	72743	391	·00538	72548	2823504	38·81
17	72352	394	·00544	72155	2750956	38·02
18	71958	395	·00549	71761	2678801	37·23
19	71563	396	·00554	71365	2607040	36·43
20	71167	399	·00560	70968	2535675	35·63
21	70768	401	·00566	70568	2464707	34·83
22	70367	403	·00573	70166	2394139	34·02
23	69964	406	·00580	69761	2323973	33·22
24	69558	410	·00590	69353	2254212	32·41
25	69148	434	·00628	68931	2184859	31·60
26	68714	497	·00724	68466	2115928	30·79
27	68217	579	·00849	67928	2047462	30·01
28	67638	661	·00977	67308	1979534	29·27
29	66977	742	·01108	66606	1912226	28·55
30	66235	825	·01245	65823	1845620	27·86
31	65410	906	·01385	64957	1779797	27·21
32	64504	986	·01528	64011	1714840	26·59
33	63518	1062	·01672	62987	1650829	25·99
34	62456	1136	·01819	61888	1587842	25·42
35	61320	1190	·01940	60725	1525954	24·89
36	60130	1219	·02027	59521	1465229	24·37
37	58911	1235	·02097	58294	1405708	23·86
38	57676	1245	·02159	57054	1347414	23·36
39	56431	1253	·02221	55805	1290360	22·87
40	55178	1258	·02279	54549	1234555	22·37
41	53920	1255	·02328	53293	1180006	21·88
42	52665	1250	·02374	52040	1126713	21·39
43	51415	1244	·02420	50793	1074673	20·90
44	50171	1237	·02466	49553	1023880	20·41
45	48934	1234	·02522	48917	974327	19·91
46	47700	1239	·02598	47081	926010	19·41
47	46461	1248	·02686	45837	878929	18·92
48	45219	1257	·02780	44585	833092	18·43
49	43956	1266	·02880	43923	788507	17·94

TABLE 22.5 (Contd.)

x	l_x	d_x	q_x	L_x	T_x	e_x^0
50	42690	1274	.02984	42053	745184	17·46
51	41416	1283	.03099	40775	703131	16·98
52	40133	1292	.03220	39487	662356	16·50
53	38841	1302	.03352	38190	622869	16·04
54	37539	1312	.03496	36883	584679	15·58
55	36227	1322	.03648	35566	547796	15·12
56	34905	1331	.03812	34240	512290	14·67
57	33574	1341	.03995	32904	477990	14·24
58	32233	1348	.04183	31559	445086	13·81
59	30885	1352	.04376	30209	413527	13·39
60	29533	1351	.04574	28858	383318	12·98
61	28182	1347	.04778	27509	354460	12·58
62	26835	1339	.04989	26166	326951	12·18
63	25496	1328	.05203	24892	300785	11·80
64	24168	1314	.05437	23511	275953	11·42
65	22854	1297	.05676	22206	252442	11·05
66	21557	1277	.05925	20919	230236	10·68
67	20280	1254	.06184	19653	209317	10·32
68	19026	1228	.06455	18412	189664	9·97
69	17798	1199	.06736	17199	171252	9·62
70	16599	1167	.07030	16016	154053	9·28
71	15492	1132	.07336	14866	138037	8·94
72	14300	1095	.07654	13753	123171	8·61
73	13205	1055	.07986	12678	109418	8·29
74	12150	1012	.08331	11644	96740	7·96
75	11198	968	.08691	10654	85096	7·64
76	10170	922	.09066	9709	74442	7·32
77	9248	874	.09455	8811	64733	7·00
78	8374	826	.09861	7961	55922	6·68
79	7548	776	.10283	7160	47961	6·35
80	6772	726	.10722	6409	40801	6·02
81	6046	676	.11178	5708	34392	5·69
82	5370	629	.11712	5056	28684,	5·34
83	4741	587	.12384	4448	23628	4·98
84	4154	551	.13254	3879	19180	4·62
85	3603	518	.14382	3344	15301	4·25
86	3085	488	.15828	2841	11957	3·88
87	2597	458	.17652	2368	9116	3·51
88	2139	426	.19914	1926	6748	3·15
89	1713	388	.22674	1519	4822	2·81
90	1325	344	.25992	1153	3303	2·49
91	981	294	.29928	834	2150	2·19
92	687	237	.34542	569	1316	1·92
93	450	180	.39894	360	747	1·66
94	270	124	.46044	208	387	1·43
95	146	77	.53052	108	179	1·23
96	69	42	.60978	48	71	1·03
97	27	19	.69882	18	23	0·85
98	8	6	.79824	5	5	0·63
99	2	2	.90864	—	—	—

Source : *Life Tables, 1951-60, Census of India, 1961 Census.* Registrar-General, India.

For the first pivotal age, $\log p_{x+1}$ is evaluated from Newton's forward formula as follows : Ignoring differences higher than the third, we have

$$\log p_{x+1} = \log p_x + \cdot2\Delta \log p_x - \cdot08\Delta^2 \log p_x + \cdot048\Delta^3 \log p_x,$$

$$\log p_{x+2} = \log p_x + \cdot4\Delta \log p_x - \cdot12\Delta^2 \log p_x + \cdot064\Delta^3 \log p_x,$$

$$\log p_{x+3} = \log p_x + \cdot6\Delta \log p_x - \cdot12\Delta^2 \log p_x + \cdot056\Delta^3 \log p_x,$$

$$\log p_{x+4} = \log p_x + \cdot8\Delta \log p_x - \cdot08\Delta^2 \log p_x + \cdot032\Delta^3 \log p_x.$$

Hence we get

$$\begin{aligned}\log \bar{p}_x &= \sum_{i=0}^4 \log p_{x+i} \\ &= 5 \log p_x + 2\Delta \log p_x - \cdot4\Delta^2 \log p_x + \cdot2\Delta^3 \log p_x \\ &= 2 \cdot 4 \log p_x + 3 \cdot 4 \log p_{x+5} - \log p_{x+10} + \cdot2 \log p_{x+15}, \quad \dots \quad (22.18)\end{aligned}$$

noting that

$$\Delta' \log p_x = (E^5 - 1)' \log p_x.$$

For the remaining pivotal ages, one uses Newton's forward formula based on p_{x-5} , and the differences corresponding to p_{x-5} , as follows :

$$\log p_x = \log p_{x-5} + 4 \log p_{x-5},$$

$$\log p_{x+1} = \log p_{x-5} + 1 \cdot 2\Delta \log p_{x-5} + \cdot12\Delta^2 \log p_{x-5} - \cdot032\Delta^3 \log p_{x-5},$$

$$\log p_{x+2} = \log p_{x-5} + 1 \cdot 4\Delta \log p_{x-5} + \cdot28\Delta^2 \log p_{x-5} - \cdot056\Delta^3 \log p_{x-5},$$

$$\log p_{x+3} = \log p_{x-5} + 1 \cdot 6\Delta \log p_{x-5} + \cdot48\Delta^2 \log p_{x-5} - \cdot064\Delta^3 \log p_{x-5},$$

$$\log p_{x+4} = \log p_{x-5} + 1 \cdot 8\Delta \log p_{x-5} + \cdot72\Delta^2 \log p_{x-5} - \cdot048\Delta^3 \log p_{x-5}.$$

Hence

$$\begin{aligned}\log \bar{p}_x &= 5 \log p_{x-5} + 7\Delta \log p_{x-5} + 1 \cdot 6\Delta^2 \log p_{x-5} - \cdot2\Delta^3 \log p_{x-5} \\ &= -2 \log p_{x-5} + 3 \cdot 2 \log p_x + 2 \cdot 2 \log p_{x+5} - \cdot2 \log p_{x+10}, \quad \dots \quad (22.19)\end{aligned}$$

Having obtained these, one forms the sum

$$N'_{x \bar{B}_1} = \sum_{i=1}^5 l_{x+i}$$

for each pivotal age x . These sums are similar to those involved in (22.18) and (22.19). The formula corresponding to (22.18), for the first pivotal age, is

$$\begin{aligned}N'_{x \bar{B}_1} &= 5l_x + 3\Delta l_x - \cdot4\Delta^2 l_x + \cdot2\Delta^3 l_x \\ &= 1 \cdot 4l_x + 4 \cdot 4l_{x+5} - l_{x+10} + \cdot2l_{x+15}, \quad \dots \quad (22.20)\end{aligned}$$

and the formula corresponding to (22.19), for the other pivotal ages, is

$$\begin{aligned} N'_{x \bar{5}} &= 5l_{x-5} + 84l_{x-5} + 2\cdot64^2l_{x-5} - 2\cdot24^3l_{x-5} \\ &= -2l_{x-5} + 2\cdot2l_x + 3\cdot2l_{x+5} - 2l_{x+10}. \quad \dots \quad (22.21) \end{aligned}$$

In case the formula gives a negative value (this will happen for very high values of x), $N'_{x \bar{5}}$ will be taken to be zero.

By taking cumulative totals of $N'_{x \bar{5}}$, starting from the end of the table, the values of

$$N'_x = \sum_{i=1}^{\infty} l_{x+i} = N'_{x \bar{5}} + N'_{x+5} \quad \dots \quad (22.22)$$

are obtained.

Lastly, one evaluates e_x^0 for the pivotal ages by using the fact that

$$\begin{aligned} e_x^0 &= \frac{\int_0^{\infty} l_{x+t} dt}{l_x} \approx \frac{\frac{1}{2} l_x + N'_x}{l_x} \\ &= \cdot5 + N'_x/l_x. \quad \dots \quad (22.23) \end{aligned}$$

22.4.5 Greville's method and method of Reed and Marrell

It is first necessary to describe the different symbols that are used in an abridged table of the second type. For an age-interval extending from exact age x to exact age $x+n$, such a table would give the values of the following functions :

(1) l_x , the number of persons, out of a cohort of l_0 persons, living at the beginning of the interval.

(2) $\pi q_x = 1 - l_{x+n}/l_x$, the probability that a member of the cohort living at age x will die before reaching age $x+n$.

(3) πd_x , the number of persons dying in the age-interval, which equals $l_x \times \pi q_x$.

(4) $\pi L_x = \int_0^n l_{x+t} dt$, which may be interpreted as the total number of years lived by the cohort while in the given age-group or as the number of members of the life-table stationary population belonging to the age-group.

(5) $T_x = \int_0^{\infty} l_{x+t} dt$, which is the total number of years lived by

the cohort while at age x and thereafter, or the number of members of the life-table stationary population of age x or above. This is obtained by taking cumulative totals of the ${}_n L_x$ values, starting from the bottom of the table and using the relation $T_x = {}_n L_x + T_{x+n}$.

(6) e_x^o , the expectation of life at age x , which equals T_x/l_x .

The basic feature of the construction of a life table of this type is the estimation of ${}_n q_x$ from the observed age-specific death rates ${}_n m_x$.

The simplest relation between ${}_n q_x$ and ${}_n m_x$, obtained by assuming that l_x is a linear function in the given age-interval, is

$${}_n q_x = \frac{2n \cdot {}_n m_x}{2 + n \cdot {}_n m_x}. \quad \dots \quad (22.24)$$

This is similar to (22.15).

Greville uses more precise equations of the same general form. In a life table, we have

$$\begin{aligned} {}_n m_x &= \frac{{}_n d_x}{{}_n L_x} \\ &= (l_x - l_{x+n})/(T_x - T_{x+n}) \\ &= -\frac{d}{dx} \log_e (T_x - T_{x+n}) \\ &= -\frac{d}{dx} \log_e ({}_n l_x) \end{aligned}$$

or
$${}_n L_x = C \exp \left[- \int {}_n m_x dx \right]. \quad \dots \quad (22.25)$$

Now, from the Euler-Maclaurin formula, we have

$$\begin{aligned} T_x &= \sum_{i=0}^{\infty} {}_n L_{x+i} \\ &= \frac{1}{n} \left[\int_x^{\infty} {}_n L_t dt + \frac{n}{2} \cdot {}_n L_x - \frac{n^2}{12} \cdot \frac{d}{dt} ({}_n L_t) \Big|_{x+} + \dots \dots \right] \\ &= C \left[\frac{1}{n} \int_x^{\infty} \exp \left[- \int {}_n m_t dt \right] dt + \frac{1}{2} \exp \left[- \int {}_n m_x dx \right] \right. \\ &\quad \left. + \frac{n}{12} \cdot {}_n m_x \exp \left[- \int {}_n m_x dx \right] + \dots \dots \right]. \end{aligned}$$

Differentiating T_x and using (22.25), we have, approximately,

$$\begin{aligned} l_x &= C \left[\frac{1}{n} \cdot \exp \left[- \int {}_n m_x dx \right] + \frac{1}{2} \cdot {}_n m_x \cdot \exp \left[- \int {}_n m_x dx \right] \right. \\ &\quad \left. + \frac{n}{12} \left({}_n m_x^2 - \frac{d}{dx} {}_n m_x \right) \exp \left[- \int {}_n m_x dx \right] \right] \\ &= {}_n L_x \left[\frac{1}{n} + \frac{1}{2} \cdot {}_n m_x + \frac{n}{12} \left({}_n m_x^2 - \frac{d}{dx} {}_n m_x \right) \right], \end{aligned}$$

so that

$$\begin{aligned} {}_n q_x &= \frac{{}_n m_x \cdot {}_n L_x}{l_x} \\ &= \frac{2n \cdot {}_n m_x}{2 + n \cdot {}_n m_x + \frac{n^2}{6} \left({}_n m_x^2 - \frac{d}{dx} {}_n m_x \right)}. \end{aligned} \quad \dots \quad (22.26)$$

If it is assumed that ${}_n m_x$ is an exponential function :

$${}_n m_x = BC^x,$$

then $\frac{d}{dx} {}_n m_x = k \cdot {}_n m_x$, where $k = \log_e C$.

Hence (22.26) may be written as

$${}_n q_x = \frac{2n \cdot {}_n m_x}{2 + {}_n m_x \left[n + \frac{n^2}{6} ({}_n m_x - k) \right]}. \quad \dots \quad (20.27)$$

A slight variation in the value of k is found to have little effect on the value of ${}_n q_x$, except at the older ages and the very young ages, where one in any case uses a different set of formulæ. Hence k may be assumed to be constant throughout the table. (It has been found that in most cases k lies between .080 and .104). One may estimate C from an average of the values

$$({}_n m_{x+n} / {}_n m_x)^{1/n}$$

and hence obtain an estimate of k .

In constructing an abridged life table by Greville's method, the probabilities of death for the first few ages are found by any of the procedures involving birth and death statistics, as in a complete table. These probabilities will give a value of l_x with which to start the abridged calculations. Then one would complete the l_x and ${}_n d_x$ columns by means of the formulæ

$${}_n d_x = l_x \times {}_n q_x, \quad l_{x+n} = l_x - {}_n d_x.$$

As to the ${}_nL_x$ column, two distinct methods may be followed. In the first method, it is assumed that the death rate ${}_n m_x$ has the same value in the observed population as in the life-table population, and use is made of the relation

$${}_nL_x = {}_n d_x / {}_n m_x. \quad \dots \quad (22.28a)$$

The other method uses the relation

$${}_nL_x = \int_0^n l_{x+t} dt,$$

which is approximated by numerical integration, e.g. by a formula like

$${}_nL_x = \frac{n}{2} (l_x + l_{x+n}) + \frac{n}{24} ({}_n d_{x+n} - {}_n d_{x-n}). \quad \dots \quad (22.28b)$$

This method, although less direct, in practice gives more accurate results. For the terminal age-group, the value is

$${}_\infty L_x = \frac{l_x}{{}_\infty m_x}. \quad \dots \quad (22.28c)$$

The values of T_x and ϵ_x^0 are then computed by the formulæ given in (5) and (6) above.

Instead of starting with an explicit assumption about the l_x function as Greville did, Reed and Merrell empirically obtained a relationship between ${}_n m_x$ and ${}_n q_x$. They studied Glover's 1910 life tables and found that a satisfactory equation is

$${}_n q_x = 1 - \exp[-n \cdot {}_n m_x - an^3 \cdot {}_n m_x^2], \quad \dots \quad (22.29)$$

where a may be taken to be -0.008 . Reed and Merrell found that for groupings as broad as 10 years, the formula works satisfactorily for all ages from 5 years to the end of life. Even for the age-group 2-4 (1.b.d), it is possible to employ equation (22.29).

For the ages 0 and 1, the under-enumeration of population and the consequent over-estimation of death rates have to be taken into account. By examining a series of U.S. life tables, it was found that the correction needed is dependent on the value of ${}_n m_x$: the greater under-enumeration of population is present in the larger values of ${}_n m_x$ rather than in the smaller ones. Equations were, therefore, derived of the form

$${}_n q_x = 1 - \exp[{}_n m_x (a + b \cdot {}_n m_x)],$$

where a and b were determined by the method of least squares, based on residuals of the form "log, p /observed q ". The equations obtained for the U.S. tables were

$$q_0 = 1 - \exp[-m_0(0.9539 - 0.5509l_0)] \quad \dots \quad (22.30)$$

and $q_1 = 1 - \exp[-m_1(0.9510 - 1.921l_1)]. \quad \dots \quad (22.31)$

In Reed and Merrell's procedure, the l_s column was obtained in the usual way. For ${}_nL_x$, with x in the first 10 years of life, the following formulæ were obtained :

$$\left. \begin{array}{l} L_0 = 0.276l_0 + 0.724l_1, \\ L_1 = 0.410l_1 + 0.590l_2, \\ {}_4L_1 = 0.034l_0 + 1.184l_1 + 2.782l_5, \\ {}_3L_2 = -0.021l_0 + 1.384l_2 + 1.637l_5, \\ {}_5L_5 = -0.003l_0 + 2.242l_5 + 2.761l_{10}. \end{array} \right\} \quad \dots \quad (22.32)$$

These were derived by fitting equations of the form

$${}_nL_x = al_0 + bl_x + cl_{x+n},$$

with $a+b+c=n$, to the values from a series of U.S. tables.

For ages beyond 10, ${}_nL_x$ was determined in terms of the area under a parabola. T_x may then be obtained by taking cumulative totals of ${}_nL_x$ or, more directly, from formulæ in terms of l_s . From age 5 to the end of life, for 5-year age intervals

$$T_x = -0.20833l_{x-5} + 2.5l_x + 0.20833l_{x+5} + 5 \sum_{a=1}^{\infty} l_{x+5a} \quad \dots \quad (22.33a)$$

and for 10-year age intervals

$$T_x = 4.16667l_x + 0.83333l_{x+10} + 10 \sum_{a=1}^{\infty} l_{x+10a}. \quad \dots \quad (22.33b)$$

Formula (22.33a) results from the assumption that ${}_nL_x$ is equal to the area between x and $x+n$ under a parabola through the four points $(x-n, l_{x-n})$, (x, l_x) , $(x+n, l_{x+n})$ and $(x+2n, l_{x+2n})$. On the other hand, (22.33b) is based on the assumption that ${}_nL_x$ is the area between x and $x+n$ under a parabola through the three points (x, l_x) , $(x+n, l_{x+n})$ and $(x+2n, l_{x+2n})$.

The following abridged life tables, which relate to the population of rural India, 1957-58, have been constructed by using the method of Reed and Merrell.

TABLE 22.6
ABRIDGED LIFE TABLES FOR RURAL INDIA, 1957-58*
MALES

Years of age x to $x+n$	$n\bar{m}_x$	nq_x	\bar{l}_x	$n\bar{d}_x$	T_x	e_x^0
0	·1802	·142731	100000	14273	4523066	45·23
1—5	·0417	·138520	85727	11875	4433400	51·72
5—15	·0055	·053734	73852	3968	4123043	55·83
15—25	·0095	·034480	69884	2410	3405662	48·73
25—35	·0042	·041259	67474	2784	2718560	40·29
35—45	·0058	·056598	64690	3661	2057009	31·80
45—55	·0128	·121293	61029	7402	1425297	23·35
55—65	·0317	·281283	53627	15084	845615	15·77
65—75	·0727	·536649	38543	20684	380098	9·86
75—85	·1700	·855026	17859	15270	102600	5·75
85—95	·3973	·994660	2589	2575	10939	4·23
95—	·9289	1·000000	14	14	15	1·07

FEMALES

Years of age x to $x+n$	$n\bar{m}_x$	nq_x	\bar{l}_x	$n\bar{d}_x$	T_x	e_x^0
0	·1672	·134191	100000	13419	4657175	46·57
1—5	·0444	·145946	86581	12636	4566891	52·75
5—15	·0055	·053734	73945	3973	4255264	57·55
15—25	·0054	·052779	69972	3693	3535912	50·53
25—35	·0056	·054689	66279	3625	2854714	43·07
35—45	·0061	·059454	62654	3725	2209966	35·27
45—55	·0087	·083870	58929	4942	1601037	27·17
55—65	·0208	·190594	53987	10290	1034000	19·12
65—75	·0497	·403544	43697	17634	537460	12·30
75—85	·1189	·728038	26063	18975	187543	7·20
85—95	·2843	·969487	7088	6872	31873	4·50
95—	·6796	1·000000	216	216	317	1·47

*Based on "Abridged Life Tables for Rural India, 1957-1958" by A. K. De and R. K. Som, *The Milbank Memorial Fund Quarterly*, 42, pp. 96-108.

22.4.6 Uses of a life table

Although the primary purpose of a life table is to present a clear picture of the mortality prevailing in a given population group, it may be put to other important uses.

It may be used in the measurement of population growth—in the computation of net reproduction rate, in particular—and in population projection, i.e. in estimating what the size and age-composition of the population will be at some future date.

Different columns of the life tables of two or more population groups may also be compared to determine relative mortality. The l_s columns, the q_s columns, the L_s columns or the e_0^s columns of the life tables may be thus compared. (In case the l_s or L_s columns are used, the size of l_0 must be the same for the tables to have comparability.) The most familiar of such comparisons (although a rough one) is in regard to e_0^s , the average longevity per member of a population.

A life table is useful from the points of view of business and Government as well. It is employed by life insurance companies in determining rates of premium for policies of persons of different ages, while the Government or a firm may use it for the determination of rates of retirement benefits for its employees.

22.5 Measurement of fertility

22.5.1 Crude birth rate

The simplest way of measuring fertility is to relate the number of births to the total population. Since it is only a *live birth* that signifies an addition to the existing population, live births alone are considered in measuring fertility, thus excluding *still births*. The formula for the above-mentioned measure, called a *crude birth rate (CBR)* is, therefore,

$$i' = 1,000 \times \frac{B}{P}, \quad \dots \quad (22.34)$$

where i' = crude birth rate per 1,000 of population ;

B = number of live births which occurred in the given region during the given period ;

P = total population of the given region during the given period.

The *CBR* per year is estimated at 27.6 for India for the year 1969.

This simple rate is, however, not an adequate measure of fertility, as it is calculated without paying any regard to the age- and sex-composition of the community.

For one thing, it cannot be called a probability rate, since the whole population cannot be supposed to be at the risk of experiencing the particular type of vital event we are considering here. Only females and only those between certain ages are really liable to this risk. Among such females, again, the risk varies from one age-group to another—a woman of 25 is certainly under a greater risk than a woman of 40.

22.5.2 General fertility rate

By relating the number of live births to the number of females in the *child-bearing ages*, the *general fertility rate (GFR)* is obtained. The formula for the *GFR* is thus

$$i = 1,000 \times \frac{B}{\sum_{\omega_1}^{\omega_2} fP_x}, \quad \dots \quad (22.35)$$

where i =general fertility rate per 1,000 females in child-bearing ages ;

B =number of live births in the given region during the given period ;

fP_x =number of females of age x l.b.d. in the given region during the given period ; and

ω_1, ω_2 =lower and upper limits of the female reproductive period.

The computation of the *GFR* requires that a decision be taken beforehand as to which years of life of a woman should be included in the child-bearing (or reproductive) period. Although the practice varies in this respect, the generally adopted method is to take $\omega_1=15$ and $\omega_2=49$. Births to mothers under 15 and above 49 are so rare that they are not recorded separately but are included in the age-groups 15 and 49, respectively.

The *GFR* shows how much the women in child-bearing ages have added to the existing population through births. It takes into account the sex-composition of the population, and also the age-composition to a certain extent. Yet it is calculated without proper regard to the age-composition of the female population in child-bearing ages. As such, two populations may show quite different *GFRs*, although they may have the same fertility in each one-year age-group.

22.5.3 Age-specific fertility rate

To form a better idea as to the fertility situation obtaining in a community, it is necessary to compute a fertility rate for each age-group of mothers separately. Fertility rates specific for age are obtained according to the same principle as is followed in computing specific death rates. Thus the specific fertility rate for the age-group x to $x+n-1$ is

$$i_x = 1,000 \times \frac{B_x}{fP_x}, \quad \dots \quad (22.36)$$

where B_x = number of live births to women of age x to $x+n-1$ in the given region during the given period and

fP_x = number of women of age x to $x+n-1$ in the region during the given period.

In the case of an *annual* age-specific fertility rate, $n=1$ and here one writes simply

$$i_x = 1,000 \times \frac{B_x}{fP_x}. \quad \dots \quad (22.37)$$

Fertility data for all countries show that usually specific fertility starts from a low point, rises to a peak somewhere in the age-group 20-29 l.b.d. and thereafter steadily declines. The fertility curve is, therefore, a highly positively skew curve. This point will be apparent from the following table of estimated fertility rates for rural India.

TABLE 22.7
FERTILITY RATES SPECIFIC FOR AGE OF MOTHER,
RURAL INDIA, 1957-58

Age-group	Age-specific fertility per 1,000 females
15-19	143.9
20-24	263.6
25-29	244.3
30-34	188.3
35-39	127.9
40-44	49.6
45-49	17.6

Source : The National Sample Survey (Report No. 76)—*Fertility and Mortality Rates in India*.

22.5.4 Total fertility rate

Age-specific fertility rates give a true picture of the fertility situation prevailing in a community. However, their use in comparing the fertility situations of two regions (or of the same region for two different periods) is not easy. Very likely, the rates will be higher for some age-groups, but lower for the remaining age-groups, in one region than in the other. One may not, in such a case, readily say that fertility as such is higher (or lower) in one region than in the other.

To be practically useful, age-specific fertility rates have, therefore, to be combined into a single quantity. For this purpose a standardised fertility rate may be employed, which is to be computed by the same method as is used in the computation of a standardised death rate. A much simpler method is to add up the annual age-specific rates and take the sum, called the *total fertility rate (TFR)*, as an index of the overall fertility of the community. Thus

$$TFR = \sum_{w_1}^{w_2} i_x. \quad \dots \quad (22.38)$$

The *TFR* is a hypothetical figure : it shows how many children *would be* born to 1,000 women if none of them died before reaching the end of the reproductive period and if all were subject to the observed specific fertility rates throughout this period.

When only quinquennial, instead of annual, fertility rates are available, an approximate value of the *TFR* is given by

$$5 \times \sum_5 i_x,$$

the sum being taken over all five-year age-groups in the reproductive period. From Table 22.7, we have, approximately,

$$\sum_5 i_x = 1,035.2.$$

Hence the *TFR* for rural India for the year 1957-58 would be about

$$5 \times 1035.2 = 5,176$$

per thousand females.

22.6 Measurement of population growth

When measures of mortality and fertility are obtained, a question that naturally arises is whether the tendency of the given population,

as indicated by these measures, is to increase, to decrease or to remain stable. Our next concern is, therefore, to devise measures of population growth on the assumption that current mortality and fertility will also continue to prevail in future.

22.6.1 Crude rate of natural increase and vital index

The simplest measure of population growth is the *crude rate of natural increase*, which is obtained by subtracting the *CDR* from the *CBR*. The *CBR* gives the proportion by which the population increases through births, while the *CDR* represents the proportion by which it decreases through deaths. The difference of the two, therefore, shows the net gain (or loss) in the population size through births and deaths taken together.

The following table shows the estimated *CBR*, the *CDR* and the crude rate of natural increase per annum for India for different parts of this century. The figures in the last two rows are based on the registration data for a few States where the registration system is relatively good. The others are estimated from census data.

TABLE 22.8
ANNUAL DEATH RATE, BIRTH RATE AND RATE OF
GROWTH OF THE INDIAN POPULATION

Years	<i>CBR</i>	<i>CDR</i>	<i>Crude rate of natural increase</i>
1901—10	49·2	42·6	6·6
1911—20	48·1	47·2	0·9
1921—30	46·4	36·3	10·1
1931—40	45·2	31·2	14·0
1941—50	39·9	27·4	12·5
1951—60	41·7	22·8	18·9
1965	29·6	9·9	19·7
1969	27·6	10·3	17·3

Source : (a) *Vital Statistics of India for 1961*. Registrar-General, India, 1964.
 (b) *Vital Statistics of India for 1969*. Registrar-General, India, 1973.

An alternative measure of the same type is the ratio of the total number of births to the total number of deaths (sometimes multiplied by 100), which is called the *vital index*. This is, of course, identically equal to the ratio of the *CBR* to the *CDR*.

Simple as they are, both these measures are considered unsuitable as indices of population growth, being subject to all the defects of the *CDR* and the *CBR*.

22.6.2 Gross reproduction rate

To get a proper measure of population growth, it is first of all necessary to take into account the age-sex composition of the population.

Our concern being to measure population growth, it is also appropriate that we should consider female births alone, since it is mainly through females that a population increases. Our age-specific fertility rates will then be given by

$$f_i_x = \frac{f B_x}{f P_x}, \quad \dots \quad (22.39)$$

where $f B_x$ is the number of female births to women of age x during the given period in the given community. Summing these rates for all ages in the reproductive period, a measure of population growth, called the *gross reproduction rate (GRR)*, is obtained. Thus

$$GRR = \sum_{x=1}^{x=5} f_i_x. \quad \dots \quad (22.40)$$

Like the *TFR*, the *GRR* is a hypothetical figure. It indicates the number of daughters who *would be* born, on the average, to each of a group of females beginning life together, supposing none of them died before reaching the end of the child-bearing period, if they experienced throughout this period the current level of fertility as represented by f_i_x .

If the given fertility rates are for quinquennial age-groups, viz.

$$f_i_x = \frac{f B_x}{f P_x},$$

then the *GRR* will be approximately given by

$$5 \times \sum_{x=1}^5 f_i_x,$$

the sum being taken over all quinquennial age-groups in the reproductive period.

In some cases births may be classified according to age of mother and according to sex. But the two-way classification of births with

respect to age of mother as well as sex may not be available. Here formula (22.40) cannot be applied, but an approximate value of the *GRR* can still be obtained if it can be assumed that the sex-ratio at birth, i.e. the ratio of the number of male births to the number of female births, remained sensibly constant over all ages of mother. Here we shall have, approximately,

$$\frac{fB_x}{B_x} = \text{a constant, say, } k.$$

Then

$$k = \frac{\frac{w_2}{w_1} fB_x}{\frac{w_2}{w_1} B_x} = \frac{fB}{B},$$

so that

$$fB_x = B_x \times \frac{fB}{B} \text{ and } f_i_x = i_x \times \frac{fB}{B}.$$

An estimate of the *GRR* will, therefore, be given by

$$\frac{fB}{B} \times \frac{\frac{w_2}{w_1} f_i_x}{\frac{w_2}{w_1}}. \quad \dots \quad (22.41)$$

$\frac{w_2}{w_1} f_i_x$, it should be noted, is just the *TFR* except for the usual multiplier 1,000.

For India, the sex-ratio at birth may be taken to be 105 males to 100 females. Hence for the year 1957-58, for which the *TFR* is approximately 5,176 per thousand females, the *GRR* is estimated at

$$5.176 \times \frac{100}{205}$$

or 2.4.

22.6.3 Net reproduction rate

The principal drawback of the *GRR* is that it does not take cognisance of the fact that some of the females who are assumed to begin life together may die before reaching age 15, some may die between ages 15 and 16, and so on. In other words, the *GRR* takes into account current fertility only but ignores current mortality.

To take into consideration the factor of mortality in measuring population growth, we may, to begin with, construct a life table for females on the basis of the observed age-specific death rates for females, f_m_x . The values in the L_x column of the table (denoted by $f L_x$ in this case) give the mean size of the cohort of $f l_0$ females in the age interval x to $x+1$ for varying x . Hence

$$f i_x \cdot f L_x$$

gives the number of female children that would be born to the cohort at age x l.b.d. The sum of these values,

$$\sum_{w_1}^{w_2} f i_x \times f L_x,$$

is the total number of female children that are expected to be born to the $f l_0$ females during their life-time. Our new measure of population growth is

$$\frac{1}{f l_0} \sum_{w_1}^{w_2} f i_x \times f L_x \quad \dots \quad (22.42)$$

and is called the *net reproduction rate* (*NRR*). The *NRR* is also a hypothetical figure : it shows how many females *would be* born, on the average, per member of a group of females beginning life together, if they were subject to the observed rates of mortality and fertility throughout their life-time.

Usually, the *NRR* is computed by the formula

$$\frac{1}{f l_0} \sum_{w_1}^{w_2} f i_x \times f l_x = \sum_{w_1}^{w_2} f i_x \times f p_0.$$

But this should be regarded only as an approximation to the value given by (22.42). The quantities $f l_x / f l_0 = f p_0$ are called the *survivorship values* for females.

With quinquennial fertility rates $f i_x$, an estimate of the *NRR* is obtained as

$$\frac{1}{f l_0} \sum f i_x \times f L_x,$$

where $f L_x = f L_x + f L_{x+1} + \dots + f L_{x+4}$.

Obviously, the *NRR* cannot be greater than the *GRR*. The latter

may be regarded as a limit above which the *NRR* cannot be raised, with fertility as it is, simply by reducing mortality.

The *NRR* is an excellent gauge for measuring the balance of births and deaths. It indicates how many future mothers would be born to present mothers according to the current levels of fertility and mortality. If the *NRR*=1, then it may be said that current fertility and mortality are such that a group of newly-born females will easily replace itself in the next generation. In such a case the population may be said to have a tendency to remain constant in size. It may be said to show a tendency to increase or decrease according as the $NRR >$ or < 1 , for in that case a group of females is expected to be replaced by a larger or a smaller number of females in the next generation, in the light of the given rates of fertility and mortality. It is in this sense that the *NRR* may be looked upon as a good index of population growth.

Useful as they are, the *NRR* as also the *GRR* should be used with caution. Both are based on the values f_{i_x} obtained from a short period of observation (such as a year). But these values, of necessity, relate to different generations of mothers. Thus these rates, in effect, use different generation values of f_{i_x} to forecast the number of births that may occur to a single generation.

The *NRR*, not to speak of the *GRR*, should not be used for forecasting future population changes. For one thing, it does not take the factor of migration into account. A more important point to note is that rates of fertility and mortality are quite unlikely to be the same in future as at present. Thirdly, the *NRR*, as well as the *GRR*, ignores the actual age-sex distribution of the population. Thus despite the fact that the actual age-sex distribution determines the reproductive capacity of a population, the *NRR* and *GRR* give theoretical numbers of births based on a hypothetical life table population, whose age-sex composition may be completely different from that of the actual population.

The *GRR* and the *NRR* for rural India are being computed below on the basis of the observed age-specific fertility rates for the year 1957-58 (*vide* Table 22.7) and the life table for females for the decade 1951-60 (*vide* Table 22.5), which is so adjusted that the size of the cohort at age 0 becomes 1,000.

TABLE 22.9
DETERMINATION OF GROSS AND NET REPRODUCTION
RATES FOR RURAL INDIA

(1) Age in years	(2) Age-specific fertility rate	(3) Female life-table stationary population	(4) col. (2) × col. (3)
15—19	0·1439	3608	519·2
20—24	0·2636	3508	924·7
25—29	0·2443	3392	828·7
30—34	0·1883	3197	602·0
35—39	0·1279	2914	372·7
40—44	0·0496	2602	129·1
45—49	0·0176	2291	40·3
Total	1·0352	—	3,416·7

The sex-ratio at birth for the country may be supposed to be 105 males to 100 females. Hence from the above table, we get

$$GRR = 5 \times 1·0352 \times \frac{100}{205} = 2·52$$

and $NRR = \frac{3,416\cdot7}{1,000} \times \frac{100}{205} = 1·67.$

22.7 Measurement of morbidity

In most countries there is no system of maintaining regular records of morbidity (i.e. sickness). Whatever data are available come from records of big hospitals. For some purposes, the number of cases of sickness or the number of persons involved will be of primary interest. But there are also many tasks that call for the use of rates for measuring morbidity, e.g. a comparison among communities or a study of time-trends.

When we try to construct such rates, a number of problems crop up. First, there is the problem of definition of sickness. While there is a clear-cut distinction between the living and the dead, no such

line of demarcation can be said to exist between sickness and health, except in the case of acute illness. This is why we have to go by definitions or standards of good health and also by standards of diagnosis. Since such definitions or standards vary from community to community, the rates of morbidity of different communities may not be comparable. Secondly, we have to take note of the fact that illness is a state that continues for a period of time. As such, any case of illness observed during a given interval may be classified into one of 4 categories : (i) illness that began before the period but terminated during the period ; (ii) illness that began before the period and terminated after the period ; (iii) illness that began as well as terminated during the period and (iv) illness that began during the period and terminated after the period. We may, then, have one type of morbidity rate considering new cases of the disease and another type considering all current cases. Thirdly, during a given period an individual may have more than one case of sickness (morbid condition) either concurrently or separated by time intervals that are greater than those indicating relapses. Different measures of morbidity may, then, be obtained by taking the total number of illnesses in the community and by taking the number of persons involved. Generally, the first type is considered* of primary interest.

22.7.1 Morbidity incidence rate

The term 'incidence' relates to the emergence of new cases of illness, and this rate is defined in terms of new cases of illness observed during a period, i.e. cases falling under categories (iii) and (iv) above.

The morbidity incidence rate (*MIR*) is given by

$$MIR = 1,000 \times \frac{I}{P}, \quad \dots \quad (22.43)$$

where I =total number of new cases of illness in a given period in a given community

and P =total population of the community during the period.

An *MIR* may either be a crude rate (when it relates to the whole population) or an age-specific rate (when it relates to a specific age-

group). Again, an *MIR* may relate to a specific type of illness (or injury) rather than all kinds of illness.

Apart from the difficulty in computing an *MIR* for lack of reliable data, it should be remembered that an *MIR* cannot be given a probability interpretation because of the way it has been defined.

22.7.2 Morbidity prevalence rate

The term 'prevalence' relates to cases of illness prevalent or existing during the given period, and the morbidity prevalence rate (*MPR*) is, therefore, based on a pooling of the categories (i)—(iv) considered earlier. The rate is thus defined by

$$MPR = 1,000 \times \frac{C}{P}, \quad \dots \quad (22.44)$$

where C =number of cases of illness observed to exist in the given community during the given period

and P =total population.

Here, too, we may have a crude *MPR* or an age-specific *MPR*. Again, an *MPR* may relate to a specific kind of illness rather than all kinds of illness.

Usually, an *MPR* relates to a short interval of time, such as a day or a week, whereas an *MIR* generally relates to a longer period. In cases of acute illness of short duration, like influenza and typhoid fever, the *MPR* would approximate the *MIR*, provided the period of observation is long enough.

22.8 Graduation formulæ used in vital statistics

22.8.1 Graduation of population data

We have seen that in the measurement of mortality or fertility for a given period, we need an estimate of the population preferably for the mid-point of this period. To get such an estimate, in cases where the registration figures are unreliable, and also to forecast future population changes, we need suitable graduation formulæ for population data. A very satisfactory formula is represented by the logistic curve. We shall see how this is developed as a suitable model for population growth and shall also consider some methods of fitting this curve.

22.8.2 Logistic curve

Suppose a population has the size P at time t and the size $P + \Delta P$ at time $t + \Delta t$. The rate of increase of the population at time t is $\frac{dP}{dt} = \lim_{\Delta t \rightarrow 0} \frac{\Delta P}{\Delta t}$. We may consider the *relative growth rate* of P ,

which is

$$\frac{1}{P} \cdot \frac{dP}{dt}, \quad \dots \quad (22.45)$$

and examine its behaviour as a function of time.

If it is assumed that

$$\frac{1}{P} \cdot \frac{dP}{dt} = r, \text{ a constant,} \quad \dots \quad (22.46)$$

then by solving this differential equation, which is equivalent to

$$\frac{d \log P}{dt} = r,$$

we get the following functional form of P :

$$\log P = \int r dt = a + rt \quad \dots \quad (22.47)$$

or $P = A e^{rt}$,

where A is some positive constant.

Thus, with a constant relative growth rate (supposed to be positive), the population follows the compound interest law. When $t \rightarrow -\infty$, $P \rightarrow 0$, whereas in case $t \rightarrow \infty$, $P \rightarrow \infty$ also. This second result appears unrealistic, because for a region with limited means of sustenance, it is unthinkable that the population can increase without limits.

When the relative growth rate is supposed to be changing with time, it is proper to relate these changes to the changes in P . A very plausible assumption for a population that is growing in an area of fixed limits would be that the relative growth rate gradually decreases as t and P increase. One of the simplest forms of decreasing functions of P is $r(1-kP)$, where r and k are positive constants. In this case, the differential equation for P takes the form

$$\frac{1}{P} \cdot \frac{dP}{dt} = r(1-kP) \quad \dots \quad (22.48)$$

or $\left(\frac{1}{P} + \frac{k}{1-kP}\right) \cdot \frac{dP}{dt} = r.$

This gives, on integration,

$$\log P - \log(1 - kP) = rt + C$$

or $\frac{P}{1 - kP} = Ae^{rt}$

or $P = \frac{1}{k + \frac{1}{A}e^{-rt}}$, ... (22.49)

where also A is a positive constant.

When $t \rightarrow -\infty$, $P \rightarrow 0$. On the other hand, when $t \rightarrow \infty$, $P \rightarrow \frac{1}{k}$.

If we denote this upper limit to the population size by L , then the equation may be written as

$$P = \frac{L}{1 + \frac{L}{A}e^{-rt}}. \quad \dots \quad (22.49a)$$

Let β be the value of t for which P is $L/2$; then we have

$$\frac{L}{2} = \frac{L}{1 + \frac{L}{A}e^{-r\beta}}$$

or $A = Le^{-r\beta}$.

Making this substitution in (22.49a), we have

$$P = \frac{L}{1 + e^{r(\beta-t)}}. \quad \dots \quad (22.50)$$

This is the form in which the equation to the logistic curve is generally expressed.

In order to study the properties of this curve, we see that the differential equation (22.48) has the form

$$\frac{dP}{dt} = rP \left(1 - \frac{P}{L}\right).$$

Since r , P and $1 - P/L$ are all positive quantities, $\frac{dP}{dt}$ is also positive, so that P is, according to the logistic law, continuously increasing with t . We have also

$$\frac{d^2P}{dt^2} = r \cdot \frac{dP}{dt} \left(1 - \frac{P}{L}\right) + rP \left(-\frac{1}{L} \cdot \frac{dP}{dt}\right) = r \left(1 - \frac{2P}{L}\right) \frac{dP}{dt}.$$

Hence $\frac{d^2P}{dt^2}$ is positive, zero or negative according as P is less than, equal to or greater than $L/2$. The critical value $L/2$ occurs, as we have already seen, when $t=\beta$. Thus the curve has a point of inflexion at $t=\beta$ and is concave upwards for $t<\beta$ and convex upwards for $t>\beta$. Again, note that $\frac{dP}{dt}=0$ for $P=0$ and $P=L$, which values correspond to $t\rightarrow-\infty$ and $t\rightarrow\infty$, respectively. Hence the logistic curve has two asymptotes, viz. $P=0$ and $P=L$. The curve is shaped like an elongated S (see Fig. 22.1).

22.8.3 Fitting a logistic curve

To fit a logistic curve to a set of data, we have to estimate the constants L , r and β from the observed figures. It will be assumed that population figures are given for N equidistant points of time, say for $t=0, 1, 2, \dots, N-1$. The population at time t will be denoted by P_t . Here we shall discuss two methods of fitting the curve, one due to R. Pearl and L. J. Reed and the other to E. C. Rhodes.

Method of Pearl and Reed

Since there are three unknown constants, these can be determined in such a way as to make the logistic curve pass through any three selected points (t, P_t) . These points should be so selected that the whole range of observations is more or less evenly covered. It will be supposed that these three points are equidistant on the time scale, so that these may be denoted by (i, P_i) , $(i+n, P_{i+n})$ and $(i+2n, P_{i+2n})$ or, through a change of origin of t , by $(0, P_0)$, (n, P_n) and $(2n, P_{2n})$. Since the curve is to pass through these points, we have

$$\left. \begin{aligned} \frac{1}{P_0} &= \frac{1+e^{rB}}{L}, \\ \frac{1}{P_n} &= \frac{1+e^{r(B-n)}}{L} \\ \text{and } \frac{1}{P_{2n}} &= \frac{1+e^{r(B-2n)}}{L}. \end{aligned} \right\} \dots \quad (22.51)$$

Writing

$$d_1 = \frac{1}{P_0} - \frac{1}{P_n}$$

and

$$d_2 = \frac{1}{P_n} - \frac{1}{P_{2n}},$$

we get from (22.51)

$$d_1 = \frac{1}{L} \cdot e^{rB} (1 - e^{-rn})$$

and

$$d_2 = \frac{1}{L} \cdot e^{r(B-n)} (1 - e^{-rn}),$$

whence

$$e^{rn} = d_1/d_2,$$

or

$$r = \frac{1}{n} (\log d_1 - \log d_2). \quad \dots \quad (22.52)$$

Further,

$$1 - d_2/d_1 = L d_1 / e^{rB},$$

or

$$\frac{d_1^2}{d_1 - d_2} = \frac{e^{rB}}{L} = \frac{1}{P_0} - \frac{1}{L},$$

or

$$\frac{1}{L} = \frac{1}{P_0} - \frac{d_1^2}{d_1 - d_2}. \quad \dots \quad (22.53)$$

We estimate r and L from equations (22.52) and (22.53), respectively. Using these estimates and the relation

$$e^{rB} = \frac{L}{P_0} - 1$$

$$\text{or } \beta = \frac{1}{r} \log \left(\frac{L}{P_0} - 1 \right), \quad \dots \quad (22.54)$$

we finally determine β .

The values of L , β and r obtained in this way will, of course, be only rough estimates. Pearl and Reed suggest a method, based on the least-square principle, by which these can be improved upon.

Denoting the estimates found by the above 'method of three selected points' by L_0 , r_0 and β_0 , we may write

$$L = L_0 + \delta_L,$$

$$r = r_0 + \delta_r,$$

and

$$\beta = \beta_0 + \delta_\beta,$$

where δ_L , δ_r , and δ_β are the errors in the estimates. The population size P , regarded as a function of L , r and β , say

$$f(L, r, \beta) = \frac{L}{1 + e^{r(\beta - i)}},$$

may then be written as

$$\begin{aligned} P &\simeq f(L_0, r_0, \beta_0) + \delta_L \left(\frac{\partial f}{\partial L} \right)_0 + \delta_r \left(\frac{\partial f}{\partial r} \right)_0 + \delta_\beta \left(\frac{\partial f}{\partial \beta} \right)_0, \\ &= f_0 + \delta_L x + \delta_r y + \delta_\beta z, \text{ say.} \end{aligned}$$

The errors δ_L , δ_r , and δ_β may be estimated by the method of least squares, which yields the normal equations :

$$\sum_i x_i (P_i - f_{0i}) = \delta_L \sum_i x_i^2 + \delta_r \sum_i x_i y_i + \delta_\beta \sum_i x_i z_i,$$

$$\sum_i y_i (P_i - f_{0i}) = \delta_L \sum_i x_i y_i + \delta_r \sum_i y_i^2 + \delta_\beta \sum_i y_i z_i,$$

$$\sum_i z_i (P_i - f_{0i}) = \delta_L \sum_i x_i z_i + \delta_r \sum_i y_i z_i + \delta_\beta \sum_i z_i^2,$$

where

$$x_i = \left(\frac{\partial P_i}{\partial L} \right)_0 = \frac{1}{1 + e^{r_0(\beta_0 - i)}},$$

$$y_i = \left(\frac{\partial P_i}{\partial r} \right)_0 = -\frac{L_0}{[1 + e^{r_0(\beta_0 - i)}]^2} \cdot (\beta_0 - i) e^{r_0(\beta_0 - i)},$$

$$z_i = \left(\frac{\partial P_i}{\partial \beta} \right)_0 = -\frac{L_0 r_0 e^{r_0(\beta_0 - i)}}{[1 + e^{r_0(\beta_0 - i)}]^3}$$

and the sums are taken over $i = 0, 1, 2, \dots, N-1$.

The process may be repeated to get still better estimates of L , r and β .

Method of Rhodes

If the observed population figures were given exactly by the logistic equation, then we would have, for $t=i-1$ and $t=i$,

$$\frac{1}{P_{i-1}} = \frac{1}{L} + \frac{e^{r(\beta-i+1)}}{L}$$

and

$$\frac{1}{P_i} = \frac{1}{L} + \frac{e^{r(\beta-i)}}{L},$$

so that

$$\frac{1}{P_i} = \frac{1-e^{-r}}{L} + e^{-r} \cdot \frac{1}{P_{i-1}}. \quad \dots \quad (22.55)$$

This relationship may be put in the form

$$y_i = A + Bx_i,$$

where

$$y_i = \frac{1}{P_i}, \quad x_i = \frac{1}{P_{i-1}}$$

and

$$A = \frac{1-e^{-r}}{L}, \quad B = e^{-r}. \quad \dots \quad (22.56)$$

Thus the two variables x and y should be exactly linearly related if the population precisely follows the logistic law. The problem is to estimate the constants A and B , assuming that the deviations of the points (x_i, y_i) from an exact linear relationship arise from errors in both x_i and y_i . The proper estimates of B and A are taken to be

$$b = \sqrt{\sum_{i=1}^{N-1} (y_i - \bar{y})^2 / \sum_{i=1}^{N-1} (x_i - \bar{x})^2} \quad \dots \quad (22.57)$$

and

$$a = \bar{y} - b\bar{x}, \quad \dots \quad (22.58)$$

$$\text{where } \bar{x} = \sum_{i=1}^{N-1} x_i / (N-1), \quad \bar{y} = \sum_{i=1}^{N-1} y_i / (N-1) = \bar{x} + \frac{1}{N-1} \left[\frac{1}{P_{N-1}} - \frac{1}{P_0} \right].$$

The constants L and r of the logistic equation are estimated from the estimates of A and B . Finally, β is estimated by noting that for the logistic curve

$$\beta = \frac{1}{r} \log_e \left(\frac{L}{P} - 1 \right) + t.$$

Taking $t=0, 1, 2, \dots, N-1$, and adding the corresponding equations, we get

$$\beta = \frac{1}{Nr} \sum_{i=0}^{N-1} \log_e \left(\frac{L}{P_i} - 1 \right) + \frac{N-1}{2}. \quad \dots \quad (22.59)$$

TABLE 22.10
CENSUS POPULATION OF U.S.A. AND POPULATION
ACCORDING TO FITTED LOGISTIC CURVE

Year	Census population (in millions)	Estimated population (in millions)
1800	5.908	5.924
1810	7.240	7.205
1820	9.638	9.719
1830	12.866	13.053
1840	17.069	17.431
1850	23.192	23.103
1860	31.443	30.328
1870	38.558	39.392
1880	50.156	50.255
1890	62.948	63.80
1900	75.995	77.571
1910	91.972	93.256
1920	105.711	109.457
1930	122.775	125.407
1940	131.669	140.383
1950	150.697	153.831
1960	179.323	165.432
1970	—	175.100
1980	—	182.926
1990	—	189.113
2000	—	193.915

We shall use Rhodes's method to fit a logistic curve to the U.S. population data obtained at the decennial censuses of 1800—1960. The observed population figures are shown in col. (2) of Table 22.10.

With $t = (\text{year} - 1800)/10$,

we have for these data

$$\sum_{i=1}^{10} y_i = \sum_{i=1}^{10} (1/P_i) = 0.5763117,$$

$$\sum_{i=1}^{10} x_i = \sum_{i=0}^{10} (1/P_i) = 0.7591301,$$

$$\sum_{i=1}^{10} y_i^2 = 0.0440893, \quad \sum_{i=1}^{10} x_i^2 = 0.0795508,$$

so that

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = \sum_{i=1}^{10} y_i^2 - \left(\sum_{i=1}^{10} y_i \right)^2 / 10 = 0.0233309,$$

and, similarly,

$$\sum_{i=1}^{16} (x_i - \bar{x})^2 = 0.0435334.$$

Hence, from (22.57) and (22.58),

$$b = \sqrt{\frac{0.0233309}{0.0435334}} = 0.73.0731$$

and $a = \frac{0.5763117 - b \times 0.7591301}{16} = 0.0012858125$,

giving $r = 0.3118748$ and $L = 208.3717$.

Also, using formula (22.59), we have

$$\begin{aligned}\beta &= \left[\sum_{i=0}^{16} \log \left(\frac{L}{P_i} - 1 \right) + 8 \times 17r \log e \right] / 17r \log e \\ &= (8.4625772 + 18.4205880) / 2.3025735 = 11.67527.\end{aligned}$$

The fitted logistic curve has, therefore, the equation

$$P_i = \frac{208.3717}{1 + e^{0.3118748(11.67527 - i)}}.$$

The population figures given by this equation are also shown in Table 22.10. The fitted curve is shown in Fig. 22.1.

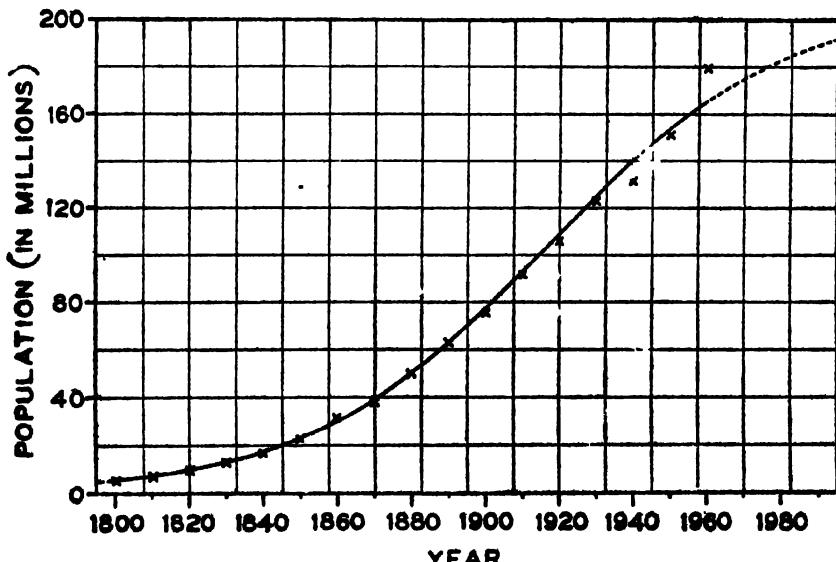


Fig. 22.1 A logistic curve fitted to the census population data of U.S.A.

22.8.4 Graduation of mortality rates

The age-specific death rates m_x for any community, as computed from census data and registration data, are found to be subject to various irregularities. For any mathematical work involving these rates, especially for the construction of life tables which take for their starting-point these rates, it is necessary to smooth out these irregularities. It thus becomes necessary to obtain some explicit expression for m_x as a function of x .

Actually, we shall consider here, instead of m_x , a related function called the *force of mortality* at age x .

Let l_x be the number of persons of exact age x and let $-\Delta l_x$ be the number of persons among them who die between age x and age $x+\Delta x$. The instantaneous death rate at age x , or the force of mortality at age x , is

$$\begin{aligned}\mu_x &= \lim_{\Delta x \rightarrow 0} \frac{1}{l_x} \cdot \frac{-\Delta l_x}{\Delta x} \\ &= -\frac{1}{l_x} \cdot \frac{dl_x}{dx}. \quad \dots \quad (22.60)\end{aligned}$$

On the other hand, denoting by d_x the number of deaths between age x and age $x+1$ and by L_x the number of persons in this age-group, we have

$$m_x = \frac{d_x}{L_x}.$$

Now,

$$\begin{aligned}\frac{dL_x}{dx} &= \frac{d}{dx} \int_0^1 l_{x+t} dt \\ &= \int_0^1 \left(\frac{d}{dx} l_{x+t} \right) dt, \quad \text{supposing that the function } l_{x+t} \text{ is sufficiently well-behaved} \\ &= l_{x+1} - l_x = -d_x.\end{aligned}$$

Hence

$$m_x = -\frac{1}{L_x} \cdot \frac{dL_x}{dx} \quad \dots \quad (22.61)$$

and, since this is approximately equal to $-\frac{1}{l_{x+1/2}} \cdot \frac{dl_{x+1/2}}{dx}$, we have the following approximate equality :

$$m_x \approx \mu_{x+1/2}. \quad \dots \quad (22.61a)$$

22.8.5 Makeham's graduation formula

Various attempts have been made to develop a suitable formula for μ_x . A very successful attempt has been that of the English actuary Makeham.

Makeham assumes that death occurs from one of two general causes. The first factor is *accidents*, whose effect may be supposed to be constant throughout the life span ; for although younger people are more active than older people and have greater recuperative power, they take greater risks. The second factor is the *decrease in the capacity to resist disease*. As regards this factor, one may assume that the force of mortality would vary inversely as a function $g(x)$, which represents the force of resistance to disease, if the factor of accidents were absent. One may, therefore, write

$$\mu_x = A + \frac{B}{g(x)}, \quad \dots \quad (22.62)$$

where $A > 0$, $B > 0$ and $g(x)$ is a decreasing function of x .

Makeham further assumes that in a short interval a person loses a constant proportion of such force of resistance to disease as he or she still has. He thus takes

$$\frac{1}{g(x)} \cdot \frac{dg(x)}{dx} = -r \quad (r > 0).$$

This leads to

$$g(x) = Ce^{-rx}$$

and $\mu_x = A + \frac{B}{Ce^{-rx}} = A + B'c^x, \quad \dots \quad (22.63)$

where A , B' and c are constants.

Because of (22.63), one gets a corresponding formula for l_x . One has

$$\log_e l_x = - \int \mu_x dx = -F - Ax - \frac{B'c^x}{\log_e c}$$

or $l_x = e^{-F - Ax - B'c^x} = k^x c^{c^x}, \text{ say.} \quad \dots \quad (22.64)$

This formula may be used to graduate the l_x figures in a life table.

Prior to Makeham's work, Gompertz had developed formulæ for μ_x and l_x , taking the force of resistance to disease into account

in the same way as Makeham did, but overlooking the factor of accidents. This had the effect of making $B=0$ and $s=1$ in the above formulae.

Makeham's modification has been found to be highly satisfactory for all ages from about 20 upwards.

22.8.6 Fitting Makeham's formula

We shall indicate the procedure to be followed in fitting Makeham's formula to a set of data. It will be assumed that the data relate to the l_x function rather than the μ_x (or m_x) function.

In Makeham's formula for l_x , there are four unknown constants, which can be determined from four independent equations. The estimates will be so determined that the resulting curve passes through four chosen points. For solving the equations, it will be convenient to make these points correspond to four equispaced values of x (say $x=0, n, 2n, 3n$, with a proper change of origin). In terms of the logarithms of l_x , we then have the following equations :

$$\left. \begin{array}{l} \log l_0 = \log k + \log g, \\ \log l_n = \log k + n \log s + c^n \log g, \\ \log l_{2n} = \log k + 2n \log s + c^{2n} \log g, \\ \log l_{3n} = \log k + 3n \log s + c^{3n} \log g. \end{array} \right\} \dots \quad (22.65)$$

For solving the equations, we first form the differences :

$$\left. \begin{array}{l} \Delta \log l_0 = n \log s + (c^n - 1) \log g, \\ \Delta \log l_n = n \log s + c^n (c^n - 1) \log g, \\ \Delta \log l_{2n} = n \log s + c^{2n} (c^n - 1) \log g, \end{array} \right\} \dots \quad (22.66)$$

and $\Delta^2 \log l_0 = (c^n - 1)^2 \log g,$ } ... (22.67)
 $\Delta^2 \log l_n = c^n (c^n - 1)^2 \log g.$ }

From this pair of equations, we also get

$$\Delta^2 \log l_n / \Delta^2 \log l_0 = c^n. \quad \dots \quad (22.68)$$

(The ratios $\Delta^2 \log l_{2n} / \Delta^2 \log l_n$, $\Delta^2 \log l_{3n} / \Delta^2 \log l_{2n}$, etc., are all equal to c^n . This fact provides a method of checking, by taking more than 4 equispaced values of x , whether Makeham's formula would be suitable for graduating any given set of values of l_x .)

An estimate of c is obtained from (22.68). Substituting this estimate in one of the equations of (22.67), we get an estimate of g . Next, substituting these estimates of c and g in some equation of (22.66), an estimate of s is obtained. Lastly, the substitution of these three estimates in some equation of (22.65) yields an estimate of k .

One may expect to get somewhat better estimates by using as much of the data as possible, and not just four observed values of l_s . Here one would use, instead of the logarithms of l_0 , l_n , l_{2n} and l_{3n} , the sums

$$S_0 = \sum_{s=0}^{n-1} \log l_s,$$

$$S_1 = \sum_{s=n}^{2n-1} \log l_s,$$

$$S_2 = \sum_{s=2n}^{3n-1} \log l_s$$

and $S_3 = \sum_{s=3n}^{4n-1} \log l_s.$

According to Makeham's formula,

$$\left. \begin{aligned} S_0 &= n \log k + \frac{n(n-1)}{2} \log s + \frac{c^n - 1}{c-1} \log g, \\ S_1 &= n \log k + \left[n^2 + \frac{n(n-1)}{2} \right] \log s + \frac{c^n(c^n - 1)}{c-1} \log g, \\ S_2 &= n \log k + \left[2n^2 + \frac{n(n-1)}{2} \right] \log s + \frac{c^{2n}(c^n - 1)}{c-1} \log g, \\ S_3 &= n \log k + \left[3n^2 + \frac{n(n-1)}{2} \right] \log s + \frac{c^{3n}(c^n - 1)}{c-1} \log g. \end{aligned} \right\} \dots \quad (22.69)$$

Also,

$$\left. \begin{aligned} \Delta S_0 &= n^2 \log s + \frac{(c^n - 1)^2}{c-1} \log g, \\ \Delta S_1 &= n^2 \log s + \frac{c^n(c^n - 1)^2}{c-1} \log g, \\ \Delta S_2 &= n^2 \log s + \frac{c^{2n}(c^n - 1)^2}{c-1} \log g. \end{aligned} \right\} \dots \quad (22.70)$$

Again,

$$\left. \begin{aligned} \Delta^2 S_0 &= \frac{(c^n - 1)^2}{c-1} \log g, \\ \Delta^2 S_1 &= \frac{c^n(c^n - 1)^2}{c-1} \log g, \end{aligned} \right\} \dots \quad (22.71)$$

and $\Delta^2 S_1 / \Delta^2 S_0 = c^n$

The estimates of c , g , s and k are obtained successively from (22.72), (22.71), (22.70) and (22.69) in the same way as the estimates were obtained from (22.65)–(22.68).

22.9 Population projection

The problem here is to predict, on the basis of the size and composition of the current population, what the size and composition of the population will be at some future date. Generally such projections are made with respect to age and sex, so that elaborate predictions have to be made of the population size for each separate age-group and separately for males and females. While it is possible to apply for this purpose some graduation formula like the logistic to each segment of the population, usually a different method is used in this case, which is based on registration data.

The starting-point for the projection may be either the latest census figures or the most current estimates. The age-distribution of the population is generally considered for 5-year age-groups, starting with the group 0–4 l.b.d. This type of grouping facilitates the computations since the projections are made in most cases for every fifth calendar year.

The method we shall discuss here is called the *component method*, where projections are first made separately for the three components that contribute to population changes, viz survivorship, migration and births. At the next step, these three projections are combined to guess what the net size and composition of the future population are likely to be.

Survivorship : It will be assumed that a life table has already been constructed on the basis of the observed (or assumed) mortality rates for the whole 5-year calendar period. From the life table we

then form the ratios

$${}_5L_{x+5}/{}_5L_x,$$

by means of which the population at the beginning of the calendar period is carried forward, with allowance for mortality, to the end of the period, when it will be 5 years older. Thus if the population at ages x to $x+4$ (l.b.d.) at the beginning of the period be ${}_5P_x$ and the population at ages x to $x+4$ at the end of the period be ${}_5P_x^{+5}$, then

$${}_5P_x^{+5} = {}_5P_x \times \frac{{}_5L_{x+5}}{{}_5L_x}. \quad \dots \quad (22.73)$$

To find the projected population in the age-group 0—4 l.b.d., one starts with an estimate of the number of births for the whole 5-year calendar period, i.e. with an estimate of $\sum_{i=1}^5 B^{+i}$. Then one estimates the population in the age-group 0—4 l.b.d. by means of the formula

$${}_5P_0^{+5} = (\sum_{i=1}^5 B^{+i}) \times \frac{{}_5L_0}{{}_5l_0}. \quad \dots \quad (22.74)$$

In case the number of births for each year of the calendar period has been estimated, one uses the more precise formula

$$\begin{aligned} {}_5P_0^{+5} &= \sum_{i=1}^5 B^{+i} \times \frac{L_{5-i}}{l_0} \\ &= B^{+1} \times \frac{L_0}{l_0} + B^{+2} \times \frac{L_1}{l_0} + \dots + B^{+5} \times \frac{L_0}{l_0}. \quad \dots \quad (22.74a) \end{aligned}$$

Migration : In this case an estimate is made of projected annual migration by regarding recent migration trends as typical for the community. Further, in line with recent experience, the distribution of net migration by age and sex may be kept unchanged.

To simplify the computations, the net migration during the 5-year calendar period is assumed to be concentrated on the last day of the period. In this way, births and deaths among migrants during the 5-year period are not taken into account. However, this is likely to introduce no serious error, for the number of births or of deaths will usually be small.

Births : It will be assumed that the projected data available consist of a distribution of females in 5-year age-groups within the reproductive period ($\{P_{15}, P_{20}, \dots, P_{45}\}$) and a corresponding set of fertility rates ($i_{15}, i_{20}, \dots, i_{45}$) for the same age-groups. Then the projected total number of births for each successive fifth calendar year is

$$B = \sum_i i_s \times P_s, \quad \dots \quad (22.75)$$

the sum being taken over all 5-year age-groups in the reproductive period. In case the projected fertility rates relate to births of both males and females combined, the number of male births and the number of female births may be estimated with the help of the sex-ratio at birth. As to the numbers of annual projected births for years intermediate between the successive fifth calendar years, these may be estimated by linear interpolation.

For a more detailed treatment of the subject of population projection, the reader is referred to the books by Cox [3] and Spiegelman [10].

Questions and exercises

22.1 Explain why the mortality situations of two places cannot usually be compared on the basis of crude death rates. Describe the construction of standardised death rates for this purpose. What is a CMI and how is it used ?

22.2 Describe the structure of a complete life table. Explain how the different columns of a life table may be computed on the basis of observed age-specific mortality rates.

22.3 How does an abridged life table differ from a complete life table ? Describe some methods of constructing an abridged table.

22.4 Derive, by starting from a suitable functional form for l_s , the formulae

$$(1) \quad L_s = (l_s + l_{s+1})/2$$

and (2) $L_s = (l_s - l_{s+1}) / (\log_l l_s - \log_l l_{s+1}) = d_s / \text{colog}_l p_s$.

Why is the first formula considered unsuitable for the early years of life, say for $s=0, 1$?

22.5 Show that

$$l_{x+1} + \int_0^1 t \left(-\frac{dl_{x+1}}{dt} \right) dt = \int_0^1 l_{x+1} dt.$$

Hence establish the formula for L_x as the total number of years lived by the cohort between age x and age $x+1$.

22.6 Show that the CDR for a life table stationary population, except for the multiplier 1,000, equals $1/e_0^0$.

22.7 Define reproduction rates. Explain how far they may be looked upon as indices of population growth.

22.8 What is meant by saying that the NRR for a country is 1.129? Show that for any community the NRR is necessarily less than the GRR.

22.9 The length of a female generation has been defined as the average age of mother at the birth of a female child Show that this may be taken as

$$\cdot \quad \sum_x (x + \frac{1}{2}) f_i_x f L_x / l_0 R_0$$

where R_0 is the net reproduction rate and the other symbols have their usual significance.

22.10 Examine the following statements :

(a) Birth rate in a year may be computed by relating the number of births occurring in the year to the number of marriages registered during the year.

(b) The relative effectiveness of public health measures of two countries may be gauged by comparing the expectations of life at birth ; the hazards of any two occupations in the same country may be compared through a comparison of the percentages of deaths in the two cases.

(c) The enumerated population and age-specific death rates for children have been recorded as follows :

Age l.b.d.	0	1	2	3	4	5
Population (millions)	4.584	8.864	6.423	6.768	5.970	8.014
Mortality rate (per thousand)	8.91	29.67	6.85	5.14	3.86	3.04

22.11 Starting from a suitable assumption regarding the relative growth rate of population, derive the logistic equation. Describe some method of fitting this curve.

It has been found that the logistic curve gives a bad fit to the population of U.S.S.R. How would you account for this?

22.12 What is meant by the force of mortality at age x ? Derive Makeham's formula, starting from suitable assumptions. Describe a method of fitting this formula.

22.13 With the help of the following data relating to New Zealand, 1958, determine the crude death rate and the age-specific death rates, separately for males and females.

Age	Population (000)		Number of deaths	
	Male	Female	Male	Female
0	29.8	28.5	807	609
1—4	109.3	104.9	192	138
5—9	126.1	120.7	88	65
10—19	198.2	189.7	182	82
20—29	150.8	142.7	247	117
30—39	156.9	151.0	284	203
40—49	139.5	138.3	565	425
50—59	110.0	106.7	1,230	746
60—69	70.1	80.9	2,083	1,464
70—79	45.4	54.5	3,308	2,650
80—	13.7	18.1	2,195	2,621
Total	1,149.8	1,136.8	11,181	9,120

Partial ans. $CDR = 8.881$ (per thousand).

22.14 A part of a life table is given here with most of the entries missing. On the basis of the available figures, supply the missing ones.

Age x	l_x	d_x	$1,000q_x$	L_x	T_x	e_x^0
10	93,102		0·62			
11			0·66			
12			0·72			
13			0·80			
14			0·90			
15			1·00			
16			1·12			
17			1·23			
18			1·33			
19			1·40		4,842,446	

Hence determine the probability (a) that a child of age 10 will live at least 5 years more, (b) that two children aged 10 and 11 will each live at least 5 years more, and (c) that of two children aged 10 and 11, at least one will die within 9 years.

22.15 In the 2nd and 3rd columns of the following table are given the age-specific death rates for Poland and Sweden for the year 1957. The figures in the 4th column give the age-distribution of a standard population adopted by the International Statistical Institute (ISI).

Age	Death rate (per thousand)		Number in ISI standard million
	Poland	Sweden	
0— 4	18·870	4·348	119,900
5—14	0·759	0·465	206,900
15—24	1·385	0·767	183,200
25—34	2·048	1·075	147,900
35—44	3·326	1·882	120,500
45—54	7·006	4·669	93,900
55—64	18·111	12·477	70,800
65—74	45·795	34·060	40,500
75 and above	124·258	116·433	16,400

Compute the standardised death rates for Poland and Sweden, taking the ISI population as standard. *Ans.* 9.210 ; 5.754.

22.16 The number of births occurring in New Zealand in 1958 is shown here classified according to age of mother, together with the female population in each age-group of the child-bearing period :

Age	Female population (000)	Number of births to mothers in the age-group
15—19	84.79	2,343
20—24	70.01	14,541
25—29	72.66	16,736
30—34	75.92	10,218
35—39	75.10	5,134
40—44	71.62	1,422
45—49	66.66	93
Total	516.76	50,487

The total population of New Zealand in 1958 was 2,285.8 thousand.

With the above information, determine (a) the crude birth rate, (b) the general fertility rate, (c) the age-specific fertility rates and (d) the total fertility rate for New Zealand, 1958. Also compute (e) the gross reproduction rate, assuming that the sex-ratio at birth was 104.5 male births to 100 female births in 1958.

Partial ans. (a) 22.09 (per thousand) ;
(b) 97.70 ; (d) 3,449.33 ; (e) 1.69.

22.17 The quinquennial fertility rates (computed on the basis of female births alone), for England and Wales, 1954, are shown in the following table, together with the survival factor for each 5-year age-group (which is the probability for a newborn female to survive till the mid-point of the age-group and is approximately equal to $\{L_x / 5^f l_0\}$) :

Age	Fertility rate (female births)	Survival factor
15—19	0·0108	0·969
20—24	0·0662	0·967
25—29	0·0675	0·963
30—34	0·0413	0·958
35—39	0·0216	0·952
40—44	0·0063	0·942
45—49	0·0004	0·928

Compute the *GRR* and *NRR* for England and Wales for 1954 on the basis of the above data.

Ans. 1·07 ; 1·03.

22.18 The population of India, as recorded in each of the last eight decennial censuses, is shown below :

Census Year	Population (millions)
1901	238·3
1911	252·0
1921	251·?
1931	278·9
1941	318·5
1951	361·0
1961	439·1
1971	547·0

Fit a logistic curve to the data. In case you find the fit to be unsatisfactory, suggest reasons for the same.

SUGGESTED READING

- [1] Anderson, J. L. and Dow, J. B. *Construction of Mortality and Other Tables* (Chs. 9, 18, 20). Cambridge Univ. Press, 1952.
- [2] Benjamin, B. *Health and Vital Statistics* (Chs. 2—6, 8). G. Allen & Unwin, 1968.
- [3] Cox, P. R. *Demography* (Chs. 6—8, 10—12, 14, 15). Cambridge Univ. Press, 1970.

- [4] Dublin, L. I., Lotka, A. J. and Spiegelman, M. *Length of Life* (Chs. 1, 12, 15). Ronald Press, 1949.
- [5] Jaffe, A. J. *Handbook of Statistical Methods for Demographers*. Bureau of the Census, U. S. Department of Commerce, 1951.
- [6] Kuczynski, R. R. *The Measurement of Population Growth* (Chs. 4—6). Sidgwick & Jackson, 1935.
- [7] Nair, K. R. "The Fitting of Growth Curves", *Statistics and Mathematics in Biology* (ed. Kempthorne *et al.*). Iowa State College Press, 1954.
- [8] Pearl, R. *Introduction to Medical Biometry and Statistics* (Chs. 7—9, 18). Saunders, 1940.
- [9] Rhodes, E. C. "Population Mathematics—III", *Journal of Royal Stat. Soc.* 103, pp. 362-87, 1940.
- [10] Spiegelman, M. *Introduction to Demography* (Chs. 2—5, 9, 12). Society of Actuaries, 1955.
- [11] Spurgeon, E. F. *Life Contingencies*. Cambridge Univ. Press, 1932.
- [12] Thompson, W. S. and Lewis, D. T. *Population Problems*. McGraw-Hill, 1965, and Tata McGraw-Hill.

23

STATISTICAL METHODS FOR PSYCHOLOGY AND EDUCATION

23.1 Introduction

Psychometry is the branch of psychology which deals with the measurement of psychological traits or mental abilities like intelligence, aptitude, interest, opinion, attitude or, simply, scholastic achievement. Educational statistics may be considered to be a part of psychometry where our main purpose is to rank a group of individuals according to their scholastic achievement. Although this task of ranking does not seem to present immediate problems, a close examination will reveal a number of pitfalls and weaknesses of the prevalent system. Statistics, however, has provided us with some techniques to remedy some of the defects of the old system.

Unlike physical or biological characteristics, psychological characteristics are rather abstract and hence can be measured only with some degree of unreliability. For the purpose of measurement, one has to develop a certain scale, which bears a strong analogy with a foot-rule used for measuring or comparing lengths. As on a foot-rule, equal distances on a psychological scale stand for empirically equal differences in the psychological trait being measured. But the zero-point of the psychological scale, unlike that of the foot-rule, is arbitrary. However, distances from the arbitrary zero are additive. In other words, a psychological scale is an *interval scale* and not a *ratio scale*, since there is no absolute zero-point on it.

23.2 Some scaling procedures

Most of the scaling procedures used for psychological or educational data are based on the assumption that the trait under consideration is normally distributed. The zero-point and the units of the scale are chosen arbitrarily, but the scale-units should be equal and remain stable throughout the scale. We shall discuss in this section some of the common scaling procedures used in psychology and education.

23.2.1 Scaling individual test-items in terms of difficulty

Here we have a number of items in a test administered to a large group of individuals. The proportion of individuals successful in each item is known. We assume in the construction of the difficulty scale that the ability (x) which the group of items is measuring is normally distributed with some mean μ and some s.d. σ . We can arbitrarily take the origin at μ and write $\mu=0$.

Let p_i be the proportion of individuals passing the i th item. We determine the point on the x -axis for which the area to the right of the ordinate is p_i . Let the point be $d_i \sigma$. Thus $d_i \sigma$ is the amount of ability required for passing the item and hence may be taken as a measure of difficulty (d_i) for the i th item. Thus an equal difference in d will mean an equal difference in ability required for passing the items.

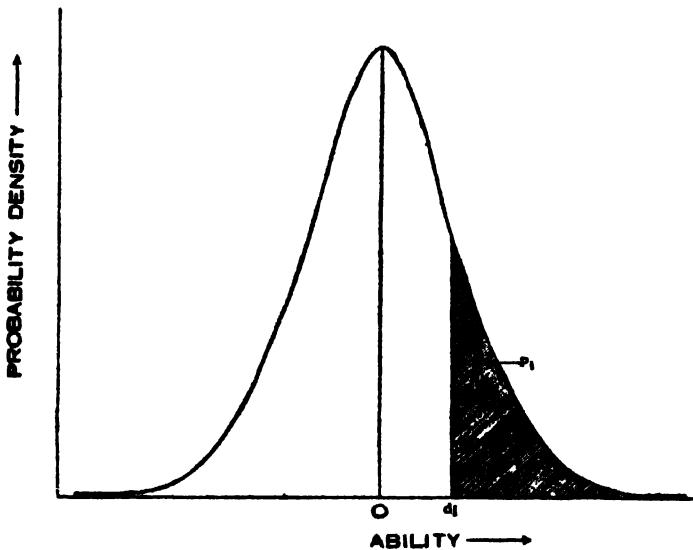


Fig. 23.1 Determining the difficulty-value of an item from the proportion of individuals passing the item.

Ex. 23.1 Suppose there are four items, A , B , C and D , passed, respectively, by 90%, 80%, 70% and 60% of the individuals. Compare the difference in difficulty between A and B with the difference in difficulty between C and D .

To find the difficulty value d_A of the item A we find the point, on the normal distribution with mean 0 and s.d. σ , the area to the right of which is 0.90. From the table of the areas under the normal probability curve (Table I, Appendix B), we have

$$d_A = -1.28\sigma.$$

Similarly, $d_B = -0.84\sigma$,

$$d_C = -0.52\sigma$$

and $d_D = -0.25\sigma$.

Hence $d_B - d_A = 0.44\sigma$, whereas $d_D - d_C = 0.27\sigma$.

Thus $d_B - d_A = 0.44\sigma$
 $d_D - d_C = 0.27\sigma$.
 $\therefore \frac{d_B - d_A}{d_D - d_C} = \frac{0.44\sigma}{0.27\sigma} = 1.63$.

The difficulty of B relative to A is 1.63 times greater than the difficulty of D relative to C .

23.2.2 Scaling of test-scores in several tests

The main defect of the prevalent system of ranking in scholastic tests consists in the adding of the raw scores of an individual on several tests to get his composite or total score and ranking all individuals on the basis of the total score. This is not a valid procedure since the same raw score x on different tests may involve different degrees of ability and hence may not be equivalent in different tests. Hence the raw scores have to be scaled under some assumption regarding the distribution of the trait which the test is measuring.

Percentile scaling

Here we assume that the distribution of the trait under consideration is rectangular, under which we shall have percentile differences equal throughout the scale. To determine the scale value corresponding to a score x on a test, we have to find the percentile position of an individual with score x , i.e. the percentage of individuals in the group having a score equal to or less than x , which can be easily obtained from the score-distribution assuming that 'score' is a continuous variable. Regardless of the form of the original raw scores distribution, the distribution of percentile scores will be rectangular. However, the distribution of raw scores is rarely rectangular, so that the basic assumption underlying the percentile scaling may not always be realistic. Thus, while using this scaling method one should beware of its limitations.

Z -scaling or σ -scaling

Here we assume that whatever differences that may exist in the forms of the raw score distributions may be attributed to chance or to the limitations of the test. In fact, the distributions of the traits under consideration are assumed to differ only in mean and s.d. Hence the scores on different tests should be expressed in terms of the scores in a hypothetical distribution of the same form as the trait-distribution with some arbitrarily chosen mean and s.d. The transformed scores are called *linear derived scores*. In particular, if the mean is arbitrarily taken to be zero and the s.d. to be unity, the scores are called *standard scores* or σ -scores or z -scores. To avoid negative standard scores, in linear derived scores the mean is generally taken to be 50 and the s.d. to be 10. If a particular test has raw score mean and s.d. equal to μ and σ , respectively, then the linear derived score corresponding to a score x on that test is given by

$$\frac{x-\mu}{\sigma} = \frac{w-50}{10}$$

or $w = 50 + 10 \cdot \frac{(x-\mu)}{\sigma} = 50 + 10z, \quad \dots \quad (23.1)$

where w is the linear derived score with mean 50 and s.d. 10 and z is the standard score.

This linear transformation changes only the mean and the s.d., while retaining the form of the original distribution.

 T -scaling

In this case we assume that the trait-distribution is normal. The raw score distribution may deviate from normality, but the deviations from normality are attributed to chance or to limitations of the tests. The mean and s.d. of the normal distribution of the trait may be arbitrarily taken to be 50 and 10, respectively. To get the scaled score corresponding to a raw score x , first we find, as in percentile scaling, the percentile position (P) of an individual with score x and then find the point (T) on a normal distribution with mean 50 and s.d. 10, below which the area is $P/100$. This is given by

$$\Phi\left(\frac{T-50}{10}\right) = \frac{P}{100}, \quad \dots \quad (23.2)$$

where $\Phi(\tau)$ is the area under the curve of the normal deviation from $-\infty$ to τ .

The scaled score obtained by this process is called *T-score* in memory of the psychologists Terman and Thorndyke. The scale is due to McCall.

Normalised scores are also expressed as *stanine* (standard nine) scores. The stanine scale takes nine values from 1 to 9, with mean 5 and s.d. 2. When a distribution is transformed to a stanine scale, the frequencies are distributed as follows :

TABLE 2.1
STANINE DISTRIBUTION

Stanine score	1	2	3	4	5	6	7	8	9
Percentage on each score (rounded)	4	7	12	17	20	17	12	7	4

A transformation is nonlinear if it changes the form of the distribution. Normalised scores and percentile scores are merely special cases of *nonlinear transformation* of the raw scores. For nonlinear transformation any form of distribution may be chosen.

Method of equivalent scores

Here we do not make any assumption about the distribution of the trait under consideration. The appropriate trait distribution is obtained by graduating the raw score distribution by an appropriate Pearsonian curve.

Let x and y be the scores on two tests, having probability-density functions $f(x)$ and $h(y)$, respectively, obtained by some process of graduation. Now, two scores on the two tests, x_i and y_i , are to be considered *equivalent*, in the sense that they bring into play equal amounts of the trait, if and only if

$$\int_{-\infty}^{x_i} f(x) dx = \int_{-\infty}^{y_i} h(y) dy. \quad \dots \quad (23.3)$$

For practical convenience, an equivalence curve may be obtained by computing a number of pairs of equivalent scores, (x_i, y_i) , and fitting to the corresponding set of points an appropriate curve, say $y = g(x)$.

Equivalent scores can also be obtained from the score distributions for x and y without going into the process of graduation. First, two ogives are drawn on the same graph paper. Two scores x_i and y_i with the same relative cumulative frequency are then regarded as equivalent (see Fig. 23.2).

For the purpose of comparison or combination, the raw scores on different tests may be converted into equivalent scores on a standard test. In this method the form of the distribution of equivalent (transformed) scores is the same as that of the standard test. If, however, the standard test score has a normal distribution, the method reduces to normalised scaling.

Ex. 23.2 The raw score distributions for Vernacular and English for a group of 500 students are given below. One of two students got 80 in Vernacular and 40 in English, while the other got 60 in both. Compare their performances by (i) percentile scaling, (ii) linear derived scores, (iii) T -scaling and (iv) equivalent scores (ogive method).

First, we have to remember that a score of 80 is to be considered as an interval from 79.5 to 80.5, and similarly for the other scores.

To obtain the percentile positions, we obtain the cumulative frequencies (less-than type) for both Vernacular and English. They are shown in Table 23.3.

Hence the percentile positions corresponding to 80.5 and 60.5 in Vernacular are given by

$$P_{80.5}(\text{Vern.}) = \frac{497 + 0.6}{500} \times 100 = 99.52$$

and

$$P_{60.5}(\text{Vern.}) = \frac{436 + 7.2}{500} \times 100 = 88.64.$$

Similarly, for English,

$$P_{80.5}(\text{Eng.}) = \frac{270 + 15.6}{500} \times 100 = 57.12$$

and

$$P_{60.5}(\text{Eng.}) = \frac{476 + 3.6}{500} \times 100 = 95.92.$$

TABLE 23.2
DISTRIBUTIONS OF SCORES IN VERNACULAR AND
ENGLISH OF A GROUP OF 500 STUDENTS

Score	Vernacular	Frequency	English
0— 4		3	
5— 9		6	
10—14		12	
15—19	6	23	
20—24	7	35	
25—29	18	45	
30—34	34	74	
35—39	56	72	
40—44	84	78	
45—49	74	53	
50—54	104	46	
55—59	53	29	
60—64	36	18	
65—69	16	5	
70—74	9	1	
75—79	0		
80—84	3		

TABLE 23.3
CUMULATIVE FREQUENCY DISTRIBUTIONS OF SCORES IN
VERNACULAR AND ENGLISH

Score	Vernacular	Cumulative frequency	English
0— 4	—	3	
5— 9	—	9	
10—14	—	21	
15—19	6	44	
20—24	13	79	
25—29	31	124	
30—34	65	198	
35—39	121	270	
40—44	205	348	
45—49	279	401	
50—54	383	447	
55—59	436	476	
60—64	472	494	
65—69	488	499	
70—74	497	500	
75—79			
80—84			

Hence the total scaled score for Student 1, getting 80 in Vernacular and 40 in English, is, by percentile scaling,

$$99.52 + 57.12 = 156.64,$$

and that of Student 2, getting 60 in both Vernacular and English, is,

$$88.64 + 95.92 = 184.56.$$

Thus we see that the relative performances of the two students are quite different although their total raw scores are equal.

For linear derived scores with mean 50 and s.d. 10, we require the means and s.d.s of scores in the two subjects. Denoting by x the score in Vernacular and by y the score in English, we have

$$\bar{x} = 47.07, \quad s_x = 11.32,$$

$$\bar{y} = 37.87 \quad \text{and} \quad s_y = 13.10.$$

Hence the w scores are given by

$$w_{80}(\text{Vern.}) = 50 + \frac{80 - 47.07}{11.32} \times 10 = 79.07,$$

$$w_{60}(\text{Vern.}) = 50 + \frac{60 - 47.07}{11.32} \times 10 = 61.40,$$

$$w_{40}(\text{Eng.}) = 50 + \frac{40 - 37.87}{13.10} \times 10 = 51.63$$

and

$$w_{60}(\text{Eng.}) = 50 + \frac{60 - 37.87}{13.10} \times 10 = 66.89.$$

As such, the total w -score of Student 1 is

$$79.07 + 51.63 = 130.70,$$

and that of Student 2 is

$$61.40 + 66.89 = 128.29.$$

Linear derived scores, however, show that Student 1 is slightly superior to Student 2.

Now, for T -scaling, percentile positions have to be converted into T -scores. We have

$$T_{80}(\text{Vern.}) = 50 + \tau_{.9952} \times 10 = 75.90,$$

$$T_{60}(\text{Vern.}) = 50 + \tau_{.8864} \times 10 = 62.08,$$

$$T_{40}(\text{Eng.}) = 50 + \tau_{.5718} \times 10 = 51.79$$

and $T_{60}(\text{Eng.}) = 50 + \tau_{.6682} \times 10 = 67.41.$

Hence the total *T*-score of Student 1 is

$$75.90 + 51.79 = 127.69,$$

and the total *T*-score of Student 2 is

$$62.08 + 67.41 = 129.49.$$

Thus *T*-scaling shows that Student 2 is slightly superior to Student 1.

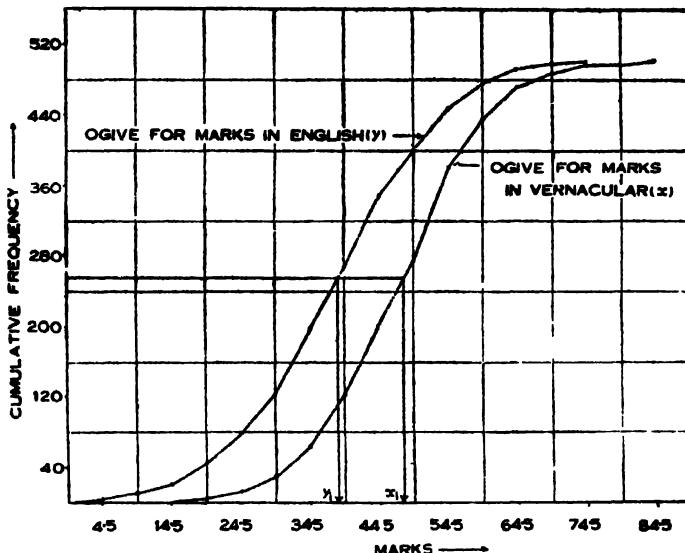


Fig. 23.2 Determination of equivalent scores in English and Vernacular from the ogives.

In the equivalent scores method, let us take Vernacular as the standard. From Fig. 23.2, we find that a score of 40 in English is equivalent to a score of 49.8 in Vernacular and a score of 60 in English is equivalent to a score of 66.9 in Vernacular.

Hence the total score of Student 1 in terms of Vernacular score is

$$80 + 49.8 = 129.8$$

and that of Student 2 is

$$60 + 66.9 = 126.9.$$

This method again shows that Student 1 is slightly superior to Student 2.

23.2.3 Scaling of rating or ranking in terms of the normal curve

In many psychological problems, individuals are rated or ranked by judges for their possession of some characteristics not readily measurable in terms of performance. Honesty, responsibility, tactfulness, etc., are examples of such traits. Suppose that there are two judges rating a group of individuals and that the frequency distributions of ratings for the two judges are known. The problem is to assign 'weights' or numerical scores to the ratings, so that the ratings of the two judges may be compared or combined.

Let us assume that the distribution of the trait (say x) is normal with mean 0 and s.d. 1. Now suppose that the individuals with trait values from x_1 to x_2 are given a particular rating. The scale value for the rating is taken to be the mean trait value of all these individuals and so is given by the formula :

$$\text{Scale value} = \frac{\int_{x_1}^{x_2} x \cdot \frac{1}{\sqrt{2\pi}} \exp[-x^2/2] dx}{\int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}} \exp[-x^2/2] dx}$$

$$= \frac{\left[-\frac{1}{\sqrt{2\pi}} \exp[-x^2/2] \right]_{x_1}^{x_2}}{\Phi(x_2) - \Phi(x_1)} = \frac{\phi(x_1) - \phi(x_2)}{\Phi(x_2) - \Phi(x_1)}, \quad \dots \quad (23.4)$$

$$\text{where } \phi(x) = \frac{1}{\sqrt{2\pi}} \exp[-x^2/2] \text{ and } \Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp[-u^2/2] du.$$

From the observed distribution of the ratings, it is easy to find $\Phi(x_1)$ and $\Phi(x_2)$, and hence $\phi(x_1)$ and $\phi(x_2)$.

The method is due to Likert and the scale is known as *Likert's scale*. This is also called the *category-scale method*.

If, on the other hand, the n individuals in the group are ranked by different judges, the scale values corresponding to the ranks can be obtained under the same assumptions as before, i.e. under the assumption of normality of the trait concerned.

Suppose there is no tie. Then the *percentile rank* (*PR*) of an individual with rank *R*, i.e. the percentage of individuals who are ranked below him, is given by

$$PR = 100 - \frac{100(R - \frac{1}{2})}{n} = P, \text{ say}, \quad \dots \quad (23.5)$$

since the rank *R* of the individual really represents the interval from $R - \frac{1}{2}$ to $R + \frac{1}{2}$. The scale value corresponding to this *PR* can now be obtained as the value of a normal deviate below which the area is $P/100$. In the case of tied ranks, the *PR* values can be obtained from the frequency distribution of ranks

Ex. 23.3 A group of 100 workers was rated by a supervisor on a five-point scale—*A*, *B*, *C*, *D* and *E*—with respect to efficiency, *A* being the highest rating and *E* the lowest. Obtain the scale value for each rating from the following frequency distribution of the ratings :

Rating	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
Frequency	5	24	45	23	3

Under the usual assumption of normality for the trait under consideration, we obtain, for the ratings, the scale values as follows :

Rating	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
Area covered by the rating $\Phi(x_2) - \Phi(x_1)$	0.05	0.24	0.45	0.3	0.03
Area below the rating $\Phi(x_1)$	0.95	0.71	0.26	0.03	0
Lower limit of the trait x_1	1.645	-553	-643	-1.881	$-\infty$
Upper limit of the trait x_2	∞	1.645	-553	-043	-1.381
Ordinate at the lower limit $\phi(x_1)$	-1031	-3424	-3244	0680	0
Ordinate at the upper limit $\phi(x_2)$	0	-1031	-3424	-3244	0680
Scale value $\frac{\phi(x_1) - \phi(x_2)}{\Phi(x_2) - \Phi(x_1)}$	2.062	0.997	-0.040	-1.115	-2.267

23.2.4 Scaling of qualitative answers to a questionnaire

The answers to the items in an attitude or personality test or a test of a similar type will be qualitative, e.g. 'Yes' and 'No', or 'Strongly approve', 'Approve', 'Undecided', 'Disapprove' and 'Strongly disapprove'. It is necessary to allot numerical scores to the answers so as to obtain the total scores of an individual measuring his attitude or personality. The method of scaling is exactly similar to Likert's rating scale described in Section 23.2.3. The questionnaire is first administered to a group of individuals and the frequency distribution of the answers is obtained. From the observed distribution, Likert's scale values are then obtained for different answers to the questionnaire.

23.2.5 Scaling of judgments of a number of products : product scale

It often happens that the ability or the trait in which we are interested cannot be expressed as a test score. This necessitates the construction of product scales. In such scales, excellence of performance is determined by comparing an individual's product with various standard products, the values of which are already determined by a number of competent and expert judges. Hand-writings, compositions, drawings, etc., are well-known examples.

We shall discuss the method of paired comparisons due to Thurstone. Suppose there are k standard products judged by a group of N judges. All possible pairs of products, $k(k-1)/2$ in all, are presented to a judge and he is to select one member of each pair in preference to the other. The data can be presented in the form of a proportion matrix :

		Product				
		1	2	k
Product	1	p_{11}	p_{21}	p_{k1}
	2	p_{12}	p_{22}	p_{k2}
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	k	p_{1k}	p_{2k}	p_{kk}

Here p_{ij} is the proportion of judges preferring the i th product to the j th one and $p_{ii}=1-p_{ii}$. By convention, $p_{ii}=1/2$.

Now, suppose that the distribution of difference in judgments (T) of the i th and j th products is normal with mean $S_i - S_j$ (the difference of their scale values) and s.d. σ_{i-j} . Thus

$$p_{ij} = \frac{1}{\sigma_{i-j}, \sqrt{2\pi}} \int_0^{\infty} \exp\left[-\frac{(T - (S_i - S_j))^2}{2\sigma_{i-j}^2}\right] dT$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \int_{-(S_i - S_j)/\sigma_{i-j},}^{\infty} \exp[-\tau^2/2] d\tau,$$

so that $S_i - S_j = -x_{ij} \sigma_{i-j}$, (23.6)

where x_{ij} is the value of the normal deviate the area to the right of which is p_{ij} . Equation (23.6) is known as Thurstone's *law of comparative judgment*. Assuming that the distribution of judgment for each product has the same s.d. σ and that judgments for any two products are uncorrelated, $\sigma_{i-j} = \sigma\sqrt{2}$, a constant.

Taking $\sigma_{i-j} = \sigma\sqrt{2}$ as the unit of the scale, we have

$$S_i - S_j = -r_{ij}. \quad \dots \quad (23.6a)$$

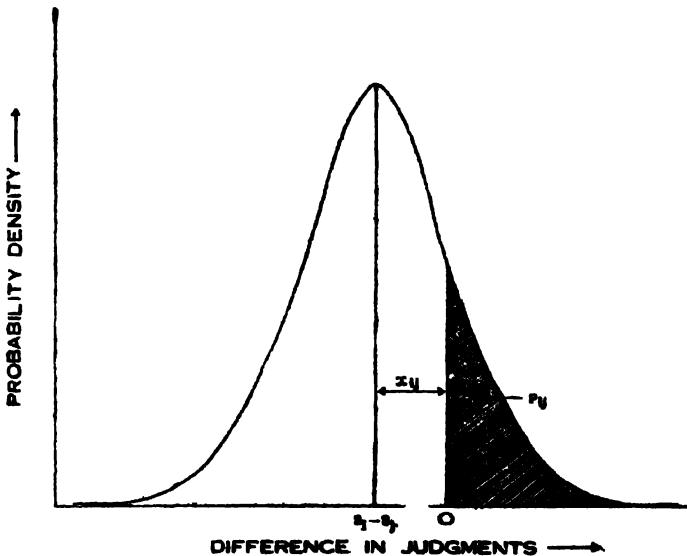


Fig. 23.3 Determining the difference of scale-values of judgments ($S_i - S_j$) from the proportion p_{ij} .

Thus we get the $(S_i - S_j)$ matrix :

		Product				
		1	2	k
		-	-	-	-	-
	1	$S_1 - S_1$	$S_2 - S_1$	$S_k - S_1$
Product	2	$S_1 - S_2$	$S_2 - S_2$	$S_k - S_2$
	⋮	⋮	⋮	⋮	⋮	⋮
	k	$S_1 - S_k$	$S_2 - S_k$	$S_k - S_k$

The column means give S_1, S_2, \dots, S_k as deviations from $\bar{S} = \frac{1}{k} \sum_{i=1}^k S_i$. If we take the origin at \bar{S} , then the column means provide us with the scale-values for the k products. Alternatively, we could take the origin at the minimum scale value and adjust the scale values accordingly.

Ex. 23·4 200 individuals were asked about their preferences for 4 different types of music. The proportion matrix is given below. Find the scale values.

		Music type			
		1	2	3	4
	1	·500	·770	·878	·892
Music type	2	·230	·500	·743	·845
	3	·122	·257	500	·797
	4	·108	·155	·203	·500

Under the usual assumption of normality of the distribution of difference in judgments with means $S_i - S_j$ and s.d. σ_{i-j} , and with the constant σ_{i-j} taken as the unit of the scale, we get the matrix of scale separations $S_i - S_j$ as follows :

SCALE SEPARATION MATRIX

		Music type			
		1	2	3	4
	1	0	·739	1·165	1·237
Music type	2	−·739	0	·653	1·015
	3	−1·165	−·653	0	·891
	4	−1·237	−1·015	−·831	0
Column mean		−·785	−·232	·247	·771

With the origin at S , the mean scale value, the column means give us the corresponding scale values for the four music types. With origin at S_1 , on the other hand, we get the following scale values :

Music type	1	2	3	4
Scale value	0	-553	1.032	1.556

23.3 Test theory

The measurements on the psychological characteristics considered in previous sections were collected by various types of methods such as tests, questionnaires or ratings. Whatever may be the method of obtaining measurements, we made the assumption, though not explicitly that the measurements were meaningful and reproducible. To be more exact, we assumed that the measuring instrument used would give us a stable and consistent measure of the trait if we remeasured the trait under identical conditions. Technically, this aspect of the accuracy is known as the *reliability* of the measuring instrument. The second requirement is that the measuring instrument measures the trait which it is intended to measure. And, technically, this is known as the *validity* of the measuring instrument.

With physical measurements these present no problems at all. For we know that if we use a non-flexible accurate measuring tape in the correct way, we shall get the exact length of an object, and this can be reproduced if remeasured under similar conditions. So physical measurements are, usually, always reliable and valid. But we are not so sure about psychological measurements. We have to verify in each case that we are getting reliable and valid measurements, and then only can we use them with confidence.

Before we actually discuss reliability and validity, we shall consider some simple results in test theory under a very simple model.

23.3.1 Linear model of test theory

We are interested in getting the true measure of an individual's performance on a test. By applying a measuring instrument what we get is the individual's raw score (obtained score) on the test.

We can consider various types of relationship between the true score of the i th individual (t_i) and his raw score (x_i). But the relationship that is usually adopted is the simplest one—a linear relationship. We assume that

$$x_i = t_i + e_i, \quad \text{for } i = 1, 2, \dots, n, \quad \dots \quad (23.7)$$

where $e_i = x_i - t_i$ is the error of measurement for the i th individual. The obtained score (x) does not equal the unknown true score (t). The difference ($x - t$), which may be due to various factors, is the error score (e).

In test theory we always consider only random errors (e). Constant or systematic errors are assumed to be absent in test theory. Since we consider only random errors, it is reasonable to make the following assumptions for the e_i 's :

$$\left. \begin{aligned} \mu_e &= 0, \\ \rho_{e,e} &= 0, \\ \rho_{e,g,h} &= 0. \end{aligned} \right\} \quad \dots \quad (23.8)$$

In words, the mean of error scores is zero, the correlation between true scores and error scores is zero, and the correlation between error scores from different testing occasions (or for two parallel tests, g and h , to be defined shortly) is zero. We note that under this model the estimates of μ_e , $\rho_{e,e}$, and $\rho_{e,g,h}$ will approach zero if the number of individuals (n) approaches infinity. In practice, however, the estimates are assumed to satisfy these relations for the given sample.

Since only random errors are considered, for a large number of cases (n large), the positive and negative errors of all magnitudes (small and large) will cancel each other with the result that the mean will be zero. Similarly, since only random errors are considered, there is no reason to expect any correlation between true scores and error scores for a large number of individuals. Large or small true scores will be expected to occur equally often with large or small error scores. This is reasonable for both positive and negative scores. Thus we assume $\rho_{e,e} = 0$. A similar argument will show that $\rho_{e,g,h} = 0$ is also a reasonable assumption.

23.3.2 Definition of parallel tests

Two tests are said to be *parallel* when it makes no difference which one is used. If g and h are two tests and if for the i th individual $t_{ig} \neq t_{ih}$, then we cannot say that it makes no difference whether we use test g or h . So, in order that g and h may be parallel tests, it is reasonable to assume that

$$t_{ig} = t_{ih}, \quad \text{for } i=1, 2, \dots, n; \quad \dots \quad (23.9)$$

i.e., the true score of any individual should be the same on the two tests.

Next, consistent with the definition of error scores (23.8), we assume about the error scores on two parallel tests that

$$\sigma_{e_g} = \sigma_{e_h}; \quad \dots \quad (23.10)$$

i.e., the standard deviations of errors on the two tests should be the same. Thus (23.9) and (23.10) define parallel tests in terms of unknown quantities. These can be expressed in terms of the distributions of the raw scores, using the relations (23.7), (23.8) and (23.9) as follows :

From (23.7), since $\mu_e = 0$, we have $\mu_t = \mu_x$ for any test. From (23.9), we have $\mu_{t_g} = \mu_{t_h}$, $\sigma_{t_g} = \sigma_{t_h}$ and $\rho_{t_g t_h} = 1$.

Also, from (23.7) and (23.8), we have $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$ for any test.

Then we have

$$\mu_{x_g} = \mu_{x_h} \text{ and } \sigma_{x_g} = \sigma_{x_h}, \quad \dots \quad (23.11)$$

for two parallel tests g and h .

Thus the means of raw scores on two parallel tests are equal ; and so are the standard deviations.

If we have more than two parallel tests (at least three—say g , h , and k), we have another condition to check, besides (23.11), before we can conclude that the tests g , h and k are parallel. And this condition is

$$\rho_{x_g x_h} = \rho_{x_g x_k} = \rho_{x_h x_k}, \quad \dots \quad (23.12)$$

the condition of equality of all inter-correlations between raw scores of the parallel tests.

Now we establish (23.12) by first obtaining an expression for $\rho_{x_g x_h}$ in terms of σ_t^2 and σ_e^2 .

$$\begin{aligned}
 \rho_{x_g x_h} &= \frac{\text{cov}(x_g, x_h)}{\sigma_{x_g} \cdot \sigma_{x_h}} \\
 &= \frac{\text{cov}(t_g, t_h) + \text{cov}(t_g, e_h) + \text{cov}(t_h, e_g) + \text{cov}(e_g, e_h)}{\sigma_{x_g} \cdot \sigma_{x_h}} \\
 &= \frac{\text{cov}(t_g, t_h)}{\sigma_{x_g}^2} \quad (\text{since } g, h \text{ are parallel tests, the remaining covariance terms are all zero and } \sigma_{x_g} = \sigma_{x_h}) \\
 &= \frac{\rho_{t_g t_h} \sigma_{t_g} \sigma_{t_h}}{\sigma_{x_g}^2} \\
 &= \sigma_{t_g}^2 / \sigma_{x_g}^2 \quad (\text{since } \rho_{t_g t_h} = 1 \text{ and } \sigma_{t_g} = \sigma_{t_h}, g \text{ and } h \text{ being parallel}).
 \end{aligned}$$

Thus, for two parallel tests g and h ,

$$\left. \begin{aligned}
 \rho_{x_g x_h} &= \sigma_{t_g}^2 / \sigma_{x_g}^2 \\
 &= \sigma_{t_h}^2 / \sigma_{x_h}^2 \quad (\text{since } \sigma_{t_g} = \sigma_{t_h}, \sigma_{x_g} = \sigma_{x_h}).
 \end{aligned} \right\} \quad \dots \quad (23.13)$$

Equation (23.13) easily establishes equation (23.12) for a number of parallel tests.

Thus, for three or more parallel tests the means of raw scores are equal ; so are the variances and the intercorrelations. In addition to satisfying these criteria, parallel tests should also be similar with respect to the content and nature of items, etc , which may be verified by expert judgment only.

23.3.3 Definition of true score

Equations (23.8) define error score. Then the true score (t) can be regarded as the difference ($x - e$) between the raw score and the error score. Thus, $t_i = x_i - e_i$.

Alternatively, we may define the true score of an individual as the limit of the average of the raw scores of the individual on a number of parallel tests when the number of parallel tests k approaches infinity, i.e.

$$t_i = \lim_{k \rightarrow \infty} \left[\sum_{s=1}^k x_{is} / k \right]. \quad \dots \quad (23.14)$$

With this definition of t , the error score is defined as the difference $x - t$; i.e., $e = x - t$.

23.3.4 Error variance (standard error of measurement)

From equations (23.7) and (23.8), we have

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2,$$

and from equation (23.13), we have

$$\sigma_t^2 = \rho_{x_g x_h} \sigma_x^2, \text{ where } g \text{ and } h \text{ are parallel tests.}$$

Thus, combining the above two relations, we get

$$\sigma_x^2 = \sigma_x^2 \rho_{x_g x_h} + \sigma_e^2$$

or $\sigma_e^2 = \sigma_x^2 (1 - \rho_{x_g x_h})$

or $\sigma_e = \sigma_x \sqrt{1 - \rho_{x_g x_h}}. \quad (23.15)$

Equation (23.15) gives the standard deviation of the error scores, which is technically known as the *standard error of measurement*.

23.3.5 Definition of reliability

We define reliability as the reproducibility of the measurements when remeasured under identical conditions. Spearman first introduced the term 'reliability'. The reliability of a test (a measuring instrument) is given by the correlation between the raw scores of the given test and a parallel test. Thus, if g be the given test and h any other test parallel to g , then the reliability of g is measured by $\rho_{x_g x_h}$ and will be denoted as ρ_{gg} .

From equation (23.13), we know that

$$\begin{aligned} \rho_{gg} &= \sigma_{x_g}^2 / \sigma_x^2 \\ &= 1 - \sigma_{e_g}^2 / \sigma_x^2 \end{aligned} \quad \left. \right\} \quad \dots \quad (23.16)$$

by virtue of the relation $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$.

Reliability can thus be defined as the ratio of the true score variance to the raw score variance or as the proportion of the raw score variance that is the true score variance. Reliability ranges from zero to one. $\rho_{gg} = 1$ when $\sigma_e = 0$. But $\sigma_e = 0$ if and only if all $e_i = 0$, since $\mu_e = 0$. Thus, a test is perfectly reliable ($\rho_{gg} = 1$) if $x_i = t_i$ for all i , and then the raw scores are the true scores. $\rho_{gg} = 0$ if $\sigma_e = 0$ (or, equivalently, if $\sigma_e = \sigma_x$), i.e. when $x_i = t_i + e_i$ for all i , and then the test is unreliable (here t denotes true score for all i). For any

test g , therefore,

$$0 \leq \rho_{gg} \leq 1.$$

It may be noted, however, that when the reliability is measured from a sample of individuals, one may obtain a negative coefficient.

23.3.6 Effect of test length on the reliability of the test

By the length of a test we mean the number of items in the test. Let us augment the length of the test by adding to it $(k-1)$ parallel tests of the same length. So the composite test is now made of k parallel tests of the same length and the length of the composite test is k times the length of the original test. The effects of this increase in length on the true score variance and raw score variance are the following :

Denoting the k parallel tests by g_1, g_2, \dots, g_k and the composite test by G , we have

$$\begin{aligned}\sigma_{GG}^2 &= \sigma_{(g_1+g_2+\dots+g_k)}^2 = \sum_i \rho_{g_i g_i} \sigma_{g_i}^2 \sigma_{g_i}^2 \\ &\quad (\text{summation over all } i, j=1, 2, \dots, k) \\ &= k^2 \sigma_{g_1}^2 \quad (\text{since the component tests are parallel,} \\ &\quad \rho_{g_i g_j} = 1 \text{ and } \sigma_{g_i} = \sigma_{g_j} \text{ for all } i, j). \quad \dots \quad (23.17)\end{aligned}$$

And

$$\begin{aligned}\sigma_{GG}^2 &= \sigma_{(g_1+g_2+\dots+g_k)}^2 = \sum_{i=1}^k \sigma_{g_i}^2 + \sum_{i \neq j} \rho_{g_i g_j} \sigma_{g_i} \sigma_{g_j}, \\ &= k\sigma_{g_1}^2 + k(k-1)\rho_{gg} \sigma_{g_1}^2, \quad \dots \quad (23.18)\end{aligned}$$

since $\rho_{g_i g_j} = \rho_{gg}$ (i.e. reliability) and $\sigma_{g_i} = \sigma_{g_j}$ for parallel tests g_i, g_j .

Using the equation (23.16), we may write down the reliability of a test whose length is increased k times (by adding $k-1$ parallel tests) as

$$\rho_{GG} = \sigma_{GG}^2 / \sigma_G^2,$$

which can be expressed in terms of ρ_{gg} , by using equations (23.17) and (23.18), as

$$\begin{aligned}\rho_{GG} &= \frac{k^2 \sigma_{g_1}^2}{k\sigma_{g_1}^2 [1 + (k-1)\rho_{gg}]} \\ &= \frac{k\rho_{gg}}{1 + (k-1)\rho_{gg}}, \quad \dots \quad (23.19)\end{aligned}$$

where ρ_{ss} is the reliability of the original test and ρ_{GG} is the reliability of the lengthened test G , whose length is equal to k times the length of g_1 .

Formula (23.19) is known as the general *Spearman-Brown formula*. In the usual case where $k=2$, the Spearman-Brown formula for doubled test length is

$$\rho_{GG} = \frac{2\rho_{ss}}{1+\rho_{ss}}. \quad \dots \quad (23.20)$$

The derivation of formulæ (23.19) and (23.20) involves the assumption that the additional test parts used in lengthening the original test are parallel to those in the original test.

The formula for determining k is obtained by solving equation (23.19) for k :

$$k = \frac{\rho_{GG}(1-\rho_{ss})}{\rho_{ss}(1-\rho_{GG})}, \quad \dots \quad (23.21)$$

where ρ_{ss} is the reliability of the original test and ρ_{GG} is the desired reliability of the lengthened test after the original test is lengthened k times.

Ex. 23.5 What would be the reliability coefficient when the original test of reliability 0.50 would be doubled in length?

We have in this case $\rho_{ss}=0.50$ and $k=2$. Then by equation (23.20) we get, as the reliability of the lengthened test,

$$\rho_{GG} = \frac{2 \times 0.50}{1 + 0.50} = 0.67.$$

Ex. 23.6 By what amount should the length of a test of reliability 0.66 be increased so as to get a reliability of 0.95 for the lengthened test?

Here $\rho_{ss}=0.67$ and $\rho_{GG}=0.95$. Then by equation (23.21), we have

$$k = \frac{0.95(1-0.67)}{0.67(1-0.95)} = \frac{0.95 \times 0.33}{0.67 \times 0.05} = \frac{0.3135}{0.0335} = 9 \text{ (approximately)}.$$

23.3.7 Practical methods of estimating test reliability

Reliability, as defined above and denoted by ρ_{ss} , is based on population data (an infinite number of individuals being tested). In practice, we have only a sample of finite size n and the correspond-

ing sample correlation estimates the reliability. There are available mainly four methods for estimating test reliability. These are :

(1) the parallel-test method, (2) the test-retest method, (3) the split-half method and (4) the Kuder-Richardson method.

Parallel-test method

Reliability was defined as the correlation between raw scores on two parallel tests. In this method, two tests are constructed satisfying as far as possible the conditions for parallelism. Then the two tests are administered to the same group with a suitable time lag and the reliability (ρ_{tt}) is estimated by the correlation (r_{tt}) between the raw scores of the parallel tests obtained from the sample.

For many situations, this is the best method of estimating test reliability. However, the ability measured should not change in the time interval between the administrations of the tests. For many scholastic achievement and mental ability tests, this condition is fulfilled. But there are cases where the ability tested will change, e.g. in performance tests like type-writing tests, athletic skills tests, etc., if the individuals continue practising during the interval between the two administrations.

The parallel-test reliability may also be obtained by administering both the tests at the same session. In this case also, the scores on the second test may be influenced either by familiarity with the material in the first test or by fatigue.

Generally speaking, parallel-test reliability will give a satisfactory result. But the difficulty is to construct two parallel tests. So when only one test is available, we are to use one of the other methods.

Test-retest method

This method consists in administering the same test twice after a suitable time interval to eliminate familiarity with the material, test fatigue, etc., and then finding the correlation between the test scores and retest scores. If, however, the individuals duplicate their first performance, then the reliability will be over-estimated by this method.

If the test is repeated immediately, the memory effect, practice and confidence will increase the scores on retesting. If sufficient time elapses before the second administration, then these effects will

be absent and the test-retest correlation will give an estimate of the stability of the test scores.

As in the parallel-test method, here also, the experimenter will have to adjust the time interval and control the activity of the individuals within the time interval so as to minimise the effects due to memory, fatigue, practice, etc.

The difficulty with both these methods is that sometimes it is difficult to get the individuals again after an interval of time. In such a case, we cannot apply either the same test twice or two parallel tests. For such cases, we have the following methods.

Split-half method

Here one test is applied once and then the score is divided into two equivalent halves, and the correlation between the scores on the half-tests estimates the reliability of each half-test. Then by Spearman-Brown formula (23.20) we may estimate the reliability of the original (full) test.

The test may be split into two parts in a number of ways. The commonest way is to split the test on the basis of odd-numbered and even-numbered items.

In many performance tests or personality tests, it is difficult to construct parallel tests or to retest with the same test. So the split-half method is regarded as the best method in such cases. The objection that is often raised is that there is no unique way of splitting the test and so no unique split-half correlation. In most *power tests* (where one does not emphasise the speed or quickness with which the work can be performed), the items are arranged in order of difficulty, and the odd-even split provides a unique estimate of reliability.

Rulon presented the following formula for estimating reliability from two subtest scores (of the same test) :

$$r_{ss} = 1 - \frac{s_d^2}{s_x^2}, \quad \dots \quad 23.22)$$

where s_x^2 is the variance of raw scores and s_d^2 is the variance of the difference of raw scores on the two halves of the test.

Similar results may be obtained by using the formula due to

Guttman, which is simpler to apply :

$$r_{ss} = 2 \left[1 - \frac{s_1^2 + s_2^2}{s_n^2} \right], \quad \dots \quad (23.23)$$

where s_1^2 and s_2^2 are the variances of raw scores on the two halves.

Equations (23.20), (23.22) and (23.23) will give the same reliability coefficient when $s_1^2 = s_2^2$, i.e. when the two halves have equal raw score variances. If $s_1^2 \neq s_2^2$, then the split-half reliability given by equation (23.20) will be the highest.

Kuder-Richardson method

We shall obtain the Kuder-Richardson formulæ for estimating test reliability by making the same assumptions as were made originally by Kuder and Richardson. Let us consider a test of length k which is made up of k parallel items. Then the raw score variance is given by

$$\sigma_x^2 = \sigma_{(x_1+x_2+\dots+x_k)}^2 = \sum_{g=1}^k \sigma_{x_g}^2 + \sum_{g \neq h} \rho_{x_g x_h} \sigma_{x_g} \sigma_{x_h}.$$

Since the items are all parallel, $\rho_{x_g x_h}$ will be equal to ρ_{gg} (reliability of item g) for all g and h , and σ_{x_g} will be the same for all g . Thus,

$$\sigma_x^2 = k\sigma_{x_g}^2 + k(k-1)\rho_{gg}\sigma_{x_g}^2,$$

so that the item reliability (ρ_{gg}) can be expressed as follows :

$$\rho_{gg} = \frac{\sigma_x^2 - \sum_{g=1}^k \sigma_{x_g}^2}{(k-1) \sum_{g=1}^k \sigma_{x_g}^2}, \text{ since } \sum_{g=1}^k \sigma_{x_g}^2 = k\sigma_{x_g}^2.$$

Next, to obtain the reliability of the test of k parallel items from ρ_{gg} , we apply the general Spearman-Brown formula (23.19) :

$$\begin{aligned} \rho_{GG} &= \frac{k\rho_{gg}}{1 + (k-1)\rho_{gg}} \\ &= k \frac{\sigma_x^2 - \sum_{g=1}^k \sigma_{x_g}^2}{(k-1) \sum_{g=1}^k \sigma_{x_g}^2} \cdot \frac{1}{1 + (k-1) \left[\left(\sigma_x^2 - \sum_{g=1}^k \sigma_{x_g}^2 \right) / (k-1) \sum_{g=1}^k \sigma_{x_g}^2 \right]} \\ &= \left[\frac{k}{k-1} \right] \cdot \left[\frac{\sigma_x^2 - \sum_{g=1}^k \sigma_{x_g}^2}{\sigma_x^2} \right]. \end{aligned} \quad \dots \quad (23.24)$$

This is the Kuder-Richardson "formula 20" for obtaining the reliability of a test of k parallel items in terms of k , s_x^2 and $s_{x_g}^2$. In practice, this is estimated by

$$r_{GG} = \left[\frac{k}{k-1} \right] \cdot \left[\frac{s_x^2 - \sum_{g=1}^k s_{x_g}^2}{s_x^2} \right], \quad \dots \quad (23.24a)$$

where s_x^2 is the sample variance of raw total scores and $s_{x_g}^2$ is the same for item g .

If the scoring of items be 1 for a correct response and 0 for a wrong response, then $s_{x_g}^2 = p_g(1-p_g)$, where p_g is the sample proportion of correct responses for item g . Then formula (23.24a) simplifies to

$$r_{GG} = \left[\frac{k}{k-1} \right] \cdot \left[\frac{s_x^2 - \sum_{g=1}^k p_g(1-p_g)}{s_x^2} \right]. \quad \dots \quad (23.25)$$

If in formula (23.24) we assume that the k parallel items are of equal difficulty, the scoring being 1 for a correct and 0 for a wrong response, with π as the common difficulty value for all items, then

$$s_{x_g}^2 = \pi(1-\pi) = \pi - \pi^2.$$

Now, the mean of obtained scores on the test is

$$\mu_x = k\pi.$$

Thus,

$$s_{x_g}^2 = \frac{\mu_x}{k} - \frac{\mu_x^2}{k^2}.$$

Then, from formula (23.24), we have

$$\begin{aligned} r_{GG} &= \left[\frac{k}{k-1} \right] \left[1 - \frac{k\sigma_{x_g}^2}{\sigma_x^2} \right] \\ &= \left[\frac{k}{k-1} \right] \left[1 - \frac{\mu_x - \mu_x^2/k}{\sigma_x^2} \right]. \end{aligned} \quad \dots \quad (23.26)$$

This is the Kuder-Richardson "formula 21" for obtaining the reliability of a test of k parallel items of equal difficulty in terms of k , σ_x^2 and μ_x . In practice, this is estimated by

$$r_{GG} = \left[\frac{k}{k-1} \right] \left[1 - \frac{\bar{x} - \bar{x}^2}{s_x^2} \right], \quad \dots \quad (23.26a)$$

where \bar{x} and s_x^2 are the sample mean and variance of raw total scores.

We have derived the Kuder-Richardson formulae under original assumptions. It is also possible to derive them under less restrictive conditions, as shown by Gulliksen [6].

The determination of reliability by the Kuder-Richardson formulae is also known as the method of *rational equivalence*.

23.3.8 Validity

In the previous section we considered one essential property of a measuring instrument—the reliability. Now we shall consider the second essential property—the validity. A psychological test (a measuring instrument) should not only be reliable, but it should also be valid. By this we mean that the test should measure what it is supposed or intended to measure. If we want to measure a trait *A* for a group of individuals with the test, we must be sure, before we can use the test confidently for that purpose, that it actually measures trait *A* and also measures it reliably. The term 'validity' is a relative term—a test is valid for a particular trait for a particular group or for a particular situation. We may use the same test for measuring different traits and then we must obtain its validity separately for each case.

As with the reliability of physical measurements, in the case of the validity of such measurements also, we face no great problem. But the situation is different with psychological measurements.

To estimate the validity of a test we must know which particular trait we want to measure. We make use of some known measure of the trait called the *criterion variable*. The validity of the test is then estimated by computing a coefficient (the *coefficient of validity*) which determines the relationship between the scores obtained on the test and the values of the criterion variable. The difficult part here is the proper choice of the criterion variable and getting measures on this variable which are to be compared with the scores on the test. Often it is difficult to get reliable measures on the true criterion. What we get are only approximate measures on the criterion variable. Depending upon the situation, the criterion scores may be of any of the following kinds : ratings by judges (experts who know the group) on the trait measured, scores on another valid test of the trait (we may validate a newly constructed test for trait *A* by selecting as the crite-

criterion variable the score on a well-established test for trait A), measures of later success (for a test for recruiting persons in a vocation), etc.

We discuss below the different concepts of validity :

Predictive validity

This type of validity arises when we use a test for selecting applicants for a particular course or job and the criterion variable is the degree of success at a later period, i.e. after the recruits have completed the course or have been on the job for a sufficient period. The criterion variable is the performance at that later period—grades or ratings on completion of the course or after a certain period of employment. A test has a high predictive validity if it can forecast efficiently later performance on a particular measurable aspect of life. And this is of importance in the selection or recruitment of individuals for different courses of study or training programmes or jobs.

Concurrent validity

Concurrent validity is obtained for tests for which the criterion variable is also available at the same time as the test results and we are not to wait as in the case of predictive validity. Tests are constructed for measuring a variable for which the result also may be obtained without waiting, because it is easier and sometimes saves time and expenditure, while giving the same result as the criterion variable.

Concurrent validity is used for diagnostic tests (e.g. in clinical diagnosis). Both types of validity (predictive and concurrent) are obtained by computing the correlation between the test scores and criterion scores, and the validity is the correlation coefficient.

Content validity

Sometimes tests are constructed to study the knowledge of the individuals on certain specific areas of study, say verbal ability, geometrical drawing ability, etc. There are a large number of items which measure these areas and, in a test, we have only a sample of these items. In content validity of a test, we try to ascertain how far the test covers the field of study under investigation or, in other words, how good the items of the test are as a sample from the totality of all items for that test.

It is, however, not possible to express content validity as a validity coefficient, as is possible with the previous two validities.

Construct validity

This is comparatively a new concept in validity theory. This concept is found useful when either there is no external criterion or it is difficult to obtain measurements on the criterion variables. This validity cannot be expressed in a single measure as the correlation between test scores and criterion scores. Validity in this case is demonstrated by showing that the predictions expected on the basis of theory may be confirmed by the test. Some of the common ways of establishing construct validity are the following :

- (1) Correlating different items or parts of the test. These correlations should be high if the test is measuring a unitary variable.
- (2) Correlating different tests which measure the same variable.

23.3.9 Effect of test length on test parameters

We have seen in Section 23.3.6 the effect of test length on true score variance (equation 23.17), on observed score variance (equation 23.18) and on reliability of a test (equation 23.19).

Using notations already introduced, it is easy to see the effect of test length on true score mean and observed score mean ;

$$\mu_{t_G} = k\mu_{t_g} \quad \dots \quad (23.27)$$

and $\mu_{x_G} = k\mu_{x_g}$ $\dots \quad (23.28)$

To find the effect of test length on the validity of a test, we first consider the case where the original test is lengthened by adding to it $(k-1)$ parallel tests of the same length and the original criterion variable is lengthened by adding to it $(l-1)$ parallel criterion variables of the same length, such that each pair of component test and criterion variable gives the same validity coefficient.

Let us denote the total test score by x_G :

$$x_G = x_{g_1} + x_{g_2} + \dots + x_{g_k}$$

and the total criterion score by y_H :

$$y_H = y_{h_1} + y_{h_2} + \dots + y_{h_l}$$

Now we obtain the correlation coefficient of augmented test scores with the augmented criterion variable scores.

$$\begin{aligned}
 \rho_{x_G y_H} &= \frac{\text{cov}(x_G, y_H)}{\sigma_{x_G} \cdot \sigma_{y_H}} \\
 &= \frac{\text{cov}(x_{g_1} + x_{g_2} + \dots + x_{g_k}, y_{h_1} + y_{h_2} + \dots + y_{h_l})}{\sqrt{\text{var}(x_{g_1} + x_{g_2} + \dots + x_{g_k}) \cdot \text{var}(y_{h_1} + y_{h_2} + \dots + y_{h_l})}} \\
 &= \frac{\sum_{g=1}^k \sum_{h=1}^l \rho_{x_g y_h} \sigma_{x_g} \sigma_{y_h}}{\{k\sigma_{x_g}^2 + k(k-1)\rho_{gg}\sigma_{x_g}^2\}^{1/2} \{l\sigma_{y_h}^2 + l(l-1)\rho_{hh}\sigma_{y_h}^2\}^{1/2}} \\
 &\quad \frac{k l \rho_{x_g y_h} \sigma_{x_g} \sigma_{y_h}}{\{k+k(k-1)\rho_{gg}\}^{1/2} \{l+l(l-1)\rho_{hh}\}^{1/2} \sigma_{x_g} \sigma_{y_h}} \\
 &= \frac{k l \rho_{x_g y_h}}{\{k+k(k-1)\rho_{gg}\}^{1/2} \{l+l(l-1)\rho_{hh}\}^{1/2}}, \quad \dots \quad (23.29)
 \end{aligned}$$

where $\rho_{x_g y_h}$ is the validity of the original test with the original criterion variable,

$\rho_{x_G y_H}$ is the validity of the lengthened test (lengthened k times) with the lengthened criterion variable (lengthened l times),
 ρ_{gg} is the reliability of the original test and
 ρ_{hh} is the reliability of the original criterion variable.

If the criterion variable is not lengthened, then the effect of increasing only the test length on the validity is obtained from (23.29) with $l=1$:

$$\rho_{x_G y_H} = \frac{k \rho_{x_g y_h}}{\{k+k(k-1)\rho_{gg}\}^{1/2}}. \quad \dots \quad (23.30)$$

23.4 Intelligence tests and IQ

Interest in the nature and measurement of intelligence is gradually increasing. Tests of intelligence and other mental qualities are being used in different spheres of life.

By intelligence is meant the capacity for relational and constructive thinking for the attainment of some goal. In the discussion of intelligence, Spearman's two-factor theory holds an important place. According to this theory, there is a common element, a *general factor*, in all our cognitive abilities—abilities that are concerned

with the intellectual aspects of mind. Spearman named this as the *g-factor* and this *g-factor* can be identified with intelligence. Besides the *g-factor*, which is present in all abilities, there is, according to Spearman, a *specific factor* for each ability. Spearman's theory was not, however, universally accepted. Thomson proposed a group-factor theory. According to Thomson, there are *group factors* each of which is present in a number of different abilities. Thus, while they are more restricted than Spearman's *g-factor*, they are less restricted than his specific factors. Some of the group factors are the following : (i) verbal ability ; (ii) numerical ability ; (iii) musical ability ; (iv) mechanical ability.

All attempts to describe intelligence by a recourse to physiology have failed. Though differences of opinion exist on the nature of intelligence, there is more or less general agreement as to the procedure for measuring intelligence. In an intelligence test, the following types of problems find a place : (i) synonyms and antonyms ; (ii) classification ; (iii) analogies ; (iv) number series ; etc.

Intelligence tests may be designed for application to individuals or for application to groups of individuals. One of the well-known individual tests is Binet's test. The revised version of this test is now being widely used for measuring intelligence of young children and for detecting mental deficiency. Group tests were first widely used by the U.S. Army authorities for recruitment, placement or promotion of personnel. The Alpha test was meant for the majority and the Beta test for illiterates or non-English-speaking persons.

Intelligence tests, like other tests, may again be verbal or non-verbal. The former demand the intelligent manipulation of ideas expressed in words, while the latter call for the intelligent manipulation of objects.

After constructing an intelligence test, we must check its reliability and validity by one of the methods discussed previously. When we are satisfied that the intelligence test is reliable and valid, we must compute some standard or *norm* which will aid us in assessing any given individual's score. We may compute either the mean and standard deviation or the percentile norms, standard scores or *T* scores for this purpose. It was in this connection that Binet introduced the

concept of *mental age*. An individual's mental age (*MA*) is the age at which an average person can pass the tests that the individual passes. Later, *mental ratio* (*MR*) was defined as

$$\text{mental ratio} = \frac{\text{mental age}}{\text{chronological age}}. \quad \dots \quad (23.31)$$

Thus, if a boy of 10 years possesses an *MA* of 9 years, then his *MR* is 0.9. He is thus a retarded child, his *MR* being less than 1. A child will be regarded as advanced if his *MR* exceeds 1, and he is of average intelligence if his *MR* equals 1.

The *intelligence quotient*, or *IQ* for short, has now replaced the *MR*. The *IQ* is defined as

$$\left. \begin{aligned} IQ &= 100 \times \frac{MA}{CA} \\ &= 100 \times MR. \end{aligned} \right\} \quad \dots \quad (23.32)$$

We now make some observations concerning the interpretation of *IQ* in its classical form. The *IQ* will be 100 (lower than 100/greater than 100) for all children who have the same (a lower/a higher) level of intellectual development as (than) the average child of the same age. It is necessary that the standard deviations of the *IQ* distributions of all age groups be approximately the same for the same *IQ* to have same relative position on the distributions for different ages. This is essential for an interpretation of an individual *IQ*. But as this is not fulfilled in many cases the present trend in standard tests is that the test is standardised and normalised into a set of normalised scores (called *IQ*-equivalents) for each age with mean 100 and standard deviation 15. Thus it is immaterial whether we use a *T*-scale or an *IQ*-equivalent scale for the norm.

The use of intelligence tests has shown that intelligence may be supposed to be normally distributed and that it depends on heredity. It has also been found that intelligence grows with age, which continues up to age 16 or 17, and then it remains steady. There is no evidence that intelligence and sex are related. It has also been found that different occupations require intelligence to varying degrees.

Intelligence tests have found many uses. They are used for

vocational guidance and selection, in the grading of pupils and in diagnosing mental deficiency.

Thus an intelligence test, properly constructed and standardised, is of immense use for various purposes.

Questions and exercises

23.1 What is the problem of measurement in education and psychology ? Explain clearly the terms *scaling*, *reliability* and *validity*, as used in problems of measurement in education and psychology.

23.2 Explain how you will combine the ranks of a number of subjects given by several judges.

23.3 Explain the different methods of combining and comparing scores in several tests, stating clearly the assumptions made in each method.

23.4 Describe how qualitative answers to a questionnaire may be scaled.

23.5 Explain the use of parallel tests in psychological studies.

23.6 Give an outline of the different methods of estimating reliability of a psychological test and give a comparative study of the coefficients of reliability obtained by these methods.

23.7 Obtain the general Spearman-Brown formula and explain how it is used for estimating reliability by the split-half method. What is the effect of increasing the length of a perfectly reliable test on its reliability ?

23.8 Derive, under suitable assumptions, the Kuder-Richardson formulae for estimating test reliability.

23.9 Define the term *validity* and discuss the different concepts of validity.

23.10 Discuss the effect of test length on different test parameters.

23.11 What are intelligence tests and how are they used in measuring intelligence ?

Define the terms 'mental age' and 'IQ' in this connection.

23.12 Four items are to be constructed so that they are equi-spaced on the difficulty scale. If the easiest item is passed by 80%

of the group and the most difficult item by 20%, find approximately the percentages of the individuals in the group passing the other two items.

Ans. 39% and 61%.

23.13 The frequency distributions of scores for two tests are given below :

Score	Frequency	
	Test A	Test B
0	5	1
1	7	2
2	10	4
3	18	8
4	20	10
5	12	16
6	10	22
7	8	25
8	5	9
9	3	2
10	2	1

Compare a score of 4 in test A with a score of 4 in test B, by
(i) percentile scaling, (ii) z-scaling, (iii) T-scaling and (iv) equivalent scores.

Partial ans. P_4 (Test A)=60; P_4 (Test B)=25;
Mean (Test A)=4.24; Mean (Test B)=5.61;
s.d. (Test A)=2.54; s.d. (Test B)=1.89.

23.14 Letter-grades *A*, *B*, *C*, *D* and *E* (*A* being the highest) are assigned by three supervisors to 50 workers in a factory. The frequency distributions of grades are given below :

Grade	Frequency		
	Supervisor 1	Supervisor 2	Supervisor 3
<i>A</i>	3	10	15
<i>B</i>	15	12	12
<i>C</i>	25	13	10
<i>D</i>	5	8	8
<i>E</i>	2	7	5

Find the numerical score corresponding to each grade for each supervisor.

Compare the performances of three workers whose grades are as follows :

Worker	Supervisor 1	Grades obtained from		
		Supervisor 2	Supervisor 3	
1	A	C	B	
2	B	C	A	
3	C	A	B	

Partial ans. Workers in descending order of performance : 1, 2, 3.

23.15 What would be the reliability coefficient if the original test of reliability 0.75 be increased three times in length ? By what amount should the length of the original test be increased so as to get a reliability of 0.95 ?

Ans. 0.90 ; 6 times.

23.16 Below are given the scores on odd-numbered items and even-numbered items in a clerical aptitude test of 100 items :

Serial No. of subject	Marks obtained in	
	odd-numbered items	even-numbered items
1	30	37
2	29	32
3	22	24
4	28	30
5	30	33
6	27	30
7	20	31
8	29	29
9	21	22
10	31	31
11	20	27
12	20	28
13	29	33
14	22	27
15	24	28
16	18	21
17	27	30
18	19	30
19	28	32
20	20	27

Obtain the test-reliability.

Ans. 0.83.

23.17 The following values were obtained by administering a 100-item geometrical drawing ability test to a group of 250 students of a local technical school :

mean score = 49.95,

s.d. = 12.53.

Obtain an estimate of test-reliability by the Kuder-Richardson method.

Ans. 0.85.

SUGGESTED READING

- [1] Bose, P. K. and Choudhury, S. B. "Scaling Procedures in Scholastic and Vocational Tests", *Sankhyā*, 15, pp. 197-206, 1955.
- [2] Freeman, F. S. *Theory and Practice of Psychological Testing* (Chs. 1, 3-5). Holt, Rinehart and Winston, 1963, and Oxford & IBH, 1965.
- [3] Garrett, H. E. *Statistics in Psychology and Education* (Chs. 4, 12, 13). Longmans, Green, 1966, and Vakils, Feffer and Simons, 1965.
- [4] Guilford, J. P. *Fundamental Statistics in Psychology and Education* (Chs. 6, 17-19). McGraw-Hill, 1956.
- [5] Guilford, J. P. *Psychometric Methods* (Chs. 7, 8, 11, 13, 14). McGraw-Hill, 1954.
- [6] Gulliksen, H. *Theory of Mental Tests* (Chs. 2, 7, 8, 15, 16, 19). John Wiley, 1950.
- [7] Knight, R. *Intelligence and Intelligence Tests* (Chs. 2, 3, 5-8). Methuen, 1959.
- [8] Magnusson, D. *Test Theory* (Chs. 1, 5, 6, 9, 10, 16). Addison-Wesley, 1967.

24

INDEX NUMBERS

24.1 Introduction

An index number may be defined as a measure of the average change in a group of related variables over two different situations. The group of variables may be the prices of a specified set of commodities, the volumes of production in different sectors of an industry, the marks obtained by a student in different subjects, and so on. The two different 'situations' may be either two different times or two different places.

The purpose of an index number and the problems faced in its construction may be well illustrated if we take the most commonly used index number, viz. the index number of prices. Changes in the prices of commodities have in present times attracted the attention of a great many people engaged in various capacities—employers, employees, trade union leaders, the government, and so on. The dearness allowances, and even the pays in certain cases, of employees of a number of commercial organisations are changed with a change in the prices of one or more of the commodities marketed. This necessitates the construction of a readily intelligible index that will reflect the change in the prices of commodities or in the cost of living. This purpose is served by the *consumer price index number* or, which is the same thing, the *cost of living index number*. Another important use to which a price index number is put is in the measurement of change in the general price level of a country. This is achieved by using the *wholesale price index number*.

Let p_0 and p_1 denote the prices of a commodity in suitable units in the two situations denoted by '0' and '1'. Any change in the price of the commodity from '0' to '1' may be expressed either in actual or in relative terms. Actual change is given by $p_1 - p_0$; the relative change is given by p_1/p_0 , which is called a *price relative*. Now, for each of the commodities marketed we have one of these two ways of reporting the price change. The problem is to combine these various individual changes in prices and get a measure of the overall

change in the prices of the set of commodities. The difficulty in dealing with actual changes is that for each commodity the change depends on the units in which the price is reported. Relative changes are better in this respect, being pure numbers and independent of the choice of units. A price index number is a sort of average of these individual price relatives, and it measures the price changes of all the commodities collectively.

24.2 Problems in the construction of index numbers

Let us discuss the various problems that arise in the construction of a price index number for any country. The problems may be enumerated as follows :

- (a) Purpose of the index.
- (b) Choice of the base period.
- (c) Choice of commodities.
- (d) Collection of data.
- (e) Method of combining data.
- (f) Choice of weights.
- (g) Interpretation of the index.

24.2.1 Purpose of the index

The purpose for which the index number is being constructed should be clearly and unambiguously stated, since most of the later problems will depend upon the purpose. For instance, if we want to construct an index number for measuring the change in the general price level, we have to take the wholesale prices of finished products, intermediate products, agricultural products, mineral products, etc. Similarly, the retail prices of consumer goods and the costs of services like electricity charges form the basis for the construction of a cost of living index number.

24.2.2 Choice of the base period

Suppose we want to compare the price levels of two time periods, say the price level of 1970 with that of 1949. We call the year 1970 the *current period* and the year 1949 the *base period*. The base period thus constitutes the basis of comparison. The price level of the base period is arbitrarily taken as 100 and the price level of the current period is expressed relative to that.

The base period should be a *normal* period in the recent past. It should be a normal period, i.e., the prices of that period should not be subject to boom or depression or effects of catastrophes like wars, floods, famines, etc. It is also desirable to select a base period which is not too far in the past, for then we may not get comparable figures. Market conditions, i.e. tastes and habits of people, may undergo some change, resulting in the replacement of old goods by new ones. Thus we find that when a base, on being used for a number of years, becomes a period in the remote past, it is to be shifted to a period in the recent past for all subsequent comparisons.

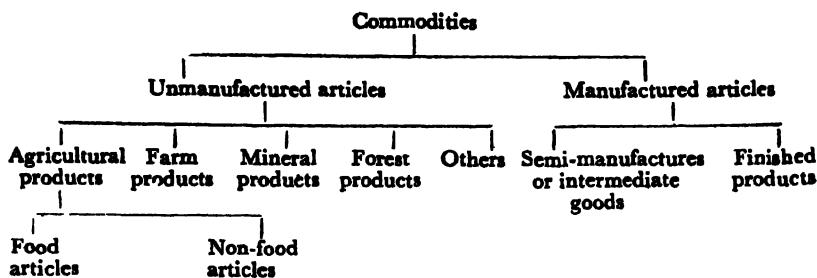
The base period should not be too short or too long. It should not be too short, e.g. a single day, because the prices for too short a period are highly unstable and unreliable. Again, it should not be too long, e.g. six years, for then the average price for that period may smooth out some important fluctuations.

The base period should be a period for which reliable figures are available and preferably a period of some economic importance for the country concerned. For example, the year 1951 may have some economic importance for India, being the inaugural year of India's five-year plans.

24.2.3 Choice of commodities

It is practically impossible to include the prices of all commodities of an economy in constructing a price index number. The reason is that it involves too much time, money and labour. Most of the practical users of index numbers require that current index numbers should be available quickly. So we are to take a suitable sample of commodities so that the index number may reflect the change in the price level, keeping in mind the purpose it is to serve. The selection of commodities should be by judgment sampling and not by random sampling, since to make the index representative of the price fluctuations we have to select the important and relevant commodities. Different groups of commodities display different patterns of price movements. So the commodities are to be classified into groups showing similar patterns of price fluctuations, and a number of commodities representative of each of these groups are to be selected. For constructing a wholesale price index number to measure the

change in the general price level, the commodities may be classified according to the following scheme :



The quality of the selected commodities should not vary much from period to period, and no commodity should disappear from the market. Reliable figures should also be available for the selected commodities.

The exact number of commodities included in the index should depend on the purpose of the index. Thus, if we construct an index number based on a few commodities most of which are food-items (which are known to be highly sensitive to price changes), the index may be useful for certain purposes, but cannot be used for measuring the change in the general price level. No rigid rule, however, can be laid down for the number of commodities to be included. But it may be stated that the number should not be too large or too small.

24.2.4 Collection of data

Makers of price index numbers take great pains to collect the necessary data in each period for all the commodities included in the index number. The price of a commodity at a particular period of time will vary from one market to another and also for different grades. So we are to collect prices of a commodity from a number of representative markets for a few important grades of the commodity. Each of these prices is referred to as a price quotation. In the case of wholesale price index numbers we are to collect wholesale prices of commodities, and for cost of living index numbers retail prices are required. As in all other cases of collection of statistical data, here too utmost care should be taken to get accurate data.

24.2.5 Method of combining data

The price fluctuations of different commodities are reflected in the price relatives. We want to represent the changes by means of a single number. So we are to consider some means of combining these individual price fluctuations. Although different commodities may have peculiar characteristics in their price fluctuations, it has been empirically found that, taken as a whole, the distribution of price relatives is bell-shaped with a marked central tendency, provided the base period is in the recent past. So we are justified in taking an appropriate measure of central tendency in combining the different price relatives.

Amongst the various measures of central tendency, the arithmetic mean and the geometric mean of price relatives are generally used.

Let us denote by p_{0i} the price of the i th commodity in the base period and by p_{1i} the price of this commodity in the current period.

If we use the arithmetic mean of price relatives for constructing the index number, then

$$I_{01} = \frac{\sum p_{1i}/p_{0i}}{k}, \quad \dots \quad (24.1)$$

where I_{01} is the index for the current period and \sum denotes summation over the k commodities. This is a *simple or unweighted index* using the arithmetic mean of price relatives.

Similarly, the formula for the index number using the simple geometric mean of price relatives will be :

$$I_{01} = (\prod p_{1i}/p_{0i})^{1/k}. \quad \dots \quad (24.2)$$

In the same way, the simple harmonic mean, median or mode of price relatives may be used.

So far we have considered some kind of average price relative to get the index number. We can also get a simple index number by comparing the simple aggregate of actual prices for the current period with that for the base period. Symbolically,

$$I_{01} = \frac{\sum p_{1i}}{\sum p_{0i}}. \quad \dots \quad (24.3)$$

This is called a *simple aggregative index*.

We are to multiply each formula by 100 to express the index in

the percentage form. However, this factor is generally omitted from the formula and introduced at the last stage.

Formula (24.3) has a serious drawback. It depends too much on the units in which the prices are quoted.

24.2.6 Choice of weights

The commodities included in the index number are not of equal importance. For instance, in constructing a wholesale price index number for India, 'rice' should have greater importance than 'tobacco'. So we must consider the problem of weighting the different commodities included in the index number according to their importance. If we ignore weights, we shall not get an unweighted or a simple index, but an inappropriately weighted index. E.g., the simple arithmetic mean of price relatives may be written in the following form :

$$I_{01} = \frac{\sum_i p_{1i}/p_{0i}}{k} = \frac{\sum_i p_{1i} \times \frac{1}{p_{0i}}}{\sum_i p_{0i} \times \frac{1}{p_{0i}}}$$

which is a *weighted aggregative index of prices*, each weight being the reciprocal of the base year price or the number of units of the commodity that can be purchased by one unit of money in the base period. It is also easily seen that in the simple average of price relatives, each relative influences the index number according to its percentage of increase or decrease over the base period. The influence which a commodity exerts on the simple aggregative index depends on the price per unit in which it is quoted.

Thus we must adopt a system of weighting for the price relatives or prices that will truly reflect the importance of each commodity. Since our index should not depend on the units in which the prices or quantities are reported, we shall weight the price relatives by values and the prices by quantities. The quantity used for determining weight may be the quantity of the commodity produced, marketed or sold, imported or exported. The prices and quantities required for the weights may relate either to the base period or to the current period.

If w_i be the weight attached to the price relative for the i th commodity, then the weighted arithmetic mean of the price relative

is given by

$$I_{01} = \frac{\sum_i \frac{p_{1i}}{p_{0i}} w_i}{\sum_i w_i}, \quad \dots \quad (24.4)$$

the weighted geometric mean by

$$I_{01} = \left\{ \prod_i \left(\frac{p_{1i}}{p_{0i}} \right)^{w_i} \right\}^{1/\sum_i w_i}, \quad \dots \quad (24.5)$$

and the weighted harmonic mean by

$$I_{01} = \frac{\sum_i w_i}{\frac{\sum_i p_{0i} w_i}{\sum_i p_{1i}}}. \quad \dots \quad (24.6)$$

Similarly, if w_i is the weight attached to the price of the i th commodity, then the weighted aggregative index is given by

$$I_{01} = \frac{\sum_i p_{1i} w_i}{\sum_i p_{0i} w_i}. \quad \dots \quad (24.7)$$

Now let us consider some particular weighted index numbers of prices.

If in (24.7) w_i be taken as q_{0i} , the base-period quantities, then we get

$$I_{01} = \frac{\sum_i p_{1i} q_{0i}}{\sum_i p_{0i} q_{0i}}, \quad \dots \quad (24.8)$$

which is known as *Laspeyres' formula*. This formula is also the same as (24.4) with w_i equal to $p_{0i} q_{0i}$, the base-period values.

Again, taking q_{1i} , the current-period quantities, as w_i in (24.7), we get

$$I_{01} = \frac{\sum_i p_{1i} q_{1i}}{\sum_i p_{0i} q_{1i}}, \quad \dots \quad (24.9)$$

which is known as *Paasche's formula*. This formula is also the same as (24.6) with w_i equal to $p_{1i} q_{1i}$, the current-period values.

Taking w_i as $(q_{1i} + q_{0i})/2$, the average of current-period and base-period quantities, in (24.7), we get

$$I_{01} = \frac{\sum_i p_{1i} (q_{1i} + q_{0i})}{\sum_i p_{0i} (q_{1i} + q_{0i})}, \quad \dots \quad (24.10)$$

which is known as the *Edgeworth-Marshall formula*. Irving Fisher

tested a large number of formulæ and selected the following formula which he obtained by crossing Laspeyres' and Paasche's formulæ geometrically :

$$I_{01} = \sqrt{\frac{\sum_i p_{1i} q_{0i}}{\sum_i p_{0i} q_{0i}} \times \frac{\sum_i p_{1i} q_{1i}}{\sum_i p_{0i} q_{1i}}} \quad \dots \quad (24.11)$$

This is known as *Fisher's ideal index number*, because it satisfies certain tests of consistency which Irving Fisher considered appropriate [vide Section 24.4].

In the majority of countries, the index numbers are computed using Laspeyres' formula or its equivalent, the weighted arithmetic mean of price relatives, the weights being the base-year values. This formula is simple to calculate and the necessary data may be easily obtained. The other most commonly used formula is the constant-weight aggregative or the constant-weight arithmetic mean of price relatives. The geometric mean of price relatives is not generally used in view of the difficulty involved in its calculation. Formulae involving current-period quantities are also not frequently used, since it is difficult to obtain those figures quickly.

24.2.7 Interpretation of the index

The interpretation will depend on the purpose of the index number. The wholesale price index number measures the change in the general price level from the base period to the current period, while the cost of living index number compares the amounts of money required to purchase the same basket of goods and services for the two periods.

Generally, the index numbers are expressed in percentage form and I_{00} , the index number for the base period, is taken as 100. Thus, the statement, "The wholesale price index number for India during June, 1971 with the year ended March, 1962 as the base is 184.8", means that, as compared with the price level during the year ended March, 1962, the price level during Jun. 1971 increased 1.848 times.

24.3 Errors in index numbers

The index numbers thus constructed will be subject to different types of errors. The errors are generally classified as : (a) formula error, (b) sampling error and (c) homogeneity error.

The formula error arises out of the choice of a particular formula in the construction of an index number. There cannot be any universally accepted formula which can measure the price changes with exactitude, and hence each formula is subject to some error inherent in the formula.

The sampling error arises from the selection of certain commodities out of the complete list of binary commodities, i.e. the commodities which are marketed in approximately the same quality in the current and base periods. Naturally, the sampling error decreases with an increase in the number of commodities included in the construction of the index number.

The third type of error is homogeneity error. This error arises from the fact that index numbers are calculated from data on binary commodities, whereas they should be based on all the commodities marketed in the base period and current period, including both binary and unique commodities. Since with the passage of time many old commodities disappear from the market and new commodities appear, the homogeneity error increases as the gap between the base period and the current period increases.

24.4 Tests for index numbers

Irving Fisher considered two tests of consistency which a price index number should satisfy, viz. the *time reversal test* and the *factor reversal test*.

Time reversal test

According to this, any formula to be accurate should be time consistent ; i.e., we should get the same picture of the change in the price level if the base period and the current period be interchanged. Consider a particular commodity, say rice. If the price of rice is doubled from 1938 to 1959, then the price relative for the period '59 with '38 as base is 2.00, while that of '38 with '59 as base will be 0.50. Thus one is the reciprocal of the other and the product is 1. This is obviously true for each individual price relative and, according to the time reversal test, it should be true for the index number. In symbols, this test says

$$I_{01} \cdot I_{10} = 1. \quad \dots \quad (24.12)$$

This test is satisfied by (24.2) and (24.3), by median and mode of

price relatives and by (24.10) and (24.11). Formulae (24.5) and (24.7) will satisfy this test if w_i 's are constants not depending on the base period or the current period

Factor reversal test

The value of a commodity is the product of the price per unit and the number of units of the commodity produced. The value of all commodities will be the sum of these products for various commodities. Thus the ratio of values for the two periods gives the value index (I_v)

$$I_v = \frac{\sum p_{1i} q_{1i}}{\sum p_{0i} q_{0i}} \quad \dots \quad (24.13)$$

According to this test, if the price and quantity factors in the price index formula (I_p) be interchanged so that a quantity index formula (I_q) is obtained then the product of these two indices should give the value index. Symbolically, one should have

$$I_p \cdot I_q = I_v \quad \dots \quad (24.14)$$

Fisher's 'ideal' formula is the only price index which satisfies (24.14). For this formula,

$$I_p = \sqrt{\frac{\sum_i p_{1i} q_{0i}}{\sum_i p_{0i} q_{0i}} \cdot \frac{\sum_i p_{1i} q_{1i}}{\sum_i p_{0i} q_{1i}}}.$$

$$I_q = \sqrt{\frac{\sum_i q_{1i} p_{0i}}{\sum_i q_{0i} p_{0i}} \cdot \frac{\sum_i q_{1i} p_{1i}}{\sum_i q_{0i} p_{1i}}},$$

while $I_v = \frac{\sum p_{1i} q_{1i}}{\sum p_{0i} q_{0i}}$

Obviously, $I_p \cdot I_q = I_v$ for this formula

24.5 Chain index

The index numbers we have considered so far are of the fixed-base type, i.e., the base period with which we compare the other time periods remains fixed with the progress of time. We have also noted that with the passage of time new commodities enter the market and old ones disappear; besides, the quality of the commodities may undergo a change. Also the relative importance of various commo-

dities, being dependent on the tastes and habits of the consumers, changes. If an index number is needed for comparing successive time periods—say 0, 1, 2, ..., n , it is not necessary to use a fixed base 0. We use the previous period as base for comparing any time period and construct what are called *link-indices*. There is no change in the method of calculation ; only the base period changes for each comparison and in each case it is the previous period. The symbol used for such an index for comparing the prices of period k with those of $(k-1)$ is $I_{k-1, k}$. Thus we construct n link indices— I_{01} ; I_{12} ; I_{23} ; ...; $I_{n-1, n}$. By multiplying successive links, i.e. by chaining, we obtain the *chain indices* as shown below :

$$\begin{aligned} I_{01}, \\ I'_{02} &= I_{01} \cdot I_{12}, \\ I'_{03} &= I_{01} \cdot I_{12} \cdot I_{23}, \\ &\dots \quad \dots \quad \dots \\ I'_{0n} &= I_{01} \cdot I_{12} \cdots \cdots I_{n-1, n}. \end{aligned} \quad \left. \right\} \quad \dots \quad (24.15)$$

These chain indices will not in general be equal to the corresponding fixed-base indices unless the formula used meets the so-called *circular test*. Stated symbolically, the test is

$$I_{01} \cdot I_{12} \cdots \cdots I_{n-1, n} \cdot I_{n0} = 1. \quad \dots \quad (24.16)$$

The time reversal test $I_{01} \cdot I_{10} = 1$ is a particular case of (24.16). Thus, if a formula satisfies the circular test, then

$$I_{01} \cdot I_{12} \cdots \cdots I_{n-1, n} = 1/I_{n0} = I_{0n}.$$

It can be easily verified that formulæ (24.2) and (24.3) satisfy the test. Formulæ (24.5) and (24.7) will also satisfy this test provided w_i 's are a set of constant weights. Formulæ (24.10) and (24.11) do not satisfy the circular test, although they satisfy the time reversal test.

The base period can be shifted to any convenient subsequent period if the formula satisfies the circular test, since I_{kn} can be calculated from the following relation, which follows from the circular test—

$$I_{kn} = \frac{I_{0n}}{I_{0k}}.$$

The practical advantage of a chain index is that the sample of commodities and/or the set of weights may be kept quite up-to-date in any index number. However, any change in the set of commodities or in the set of weights will upset the circular test.

24.6 Relative merits and demerits of chain-base and fixed-base methods

We have seen that the fixed-base index numbers become more and more inaccurate as the distance between the base period and the current period increases. As the chain-base index numbers are based on a number of link-indices, each of which is expected to be quite accurate, it is claimed that the chain-base index numbers are more accurate than the fixed-base ones, so far as long-term comparison is concerned. Also, a chain index fully utilises the information regarding prices and quantities of all the intervening periods between the base period and the current period, whereas a fixed-base index requires data concerning the base period and the current period only.

Some authorities, on the other hand, hold that since a chain index is obtained by multiplying a number of link-indices, it may involve a cumulative error, although none has put forward any convincing proof for the existence of such error.

Fixed-base index numbers are generally easier to calculate and are more easily understood by users of index numbers than chain-base index numbers.

24.7 Cost of living index number

A *cost of living index number* measures the relative change in the amount of money required to produce equivalent satisfaction under two different situations. The cost of living index number always relates to a designated group of people, e.g. the *normal class* of people in Calcutta. In practice, this index is constructed by comparing the consumer (retail) prices, for the two situations, of a fixed set of goods and services representing the consumption level (or the level of living) of the given group of people.

The cost of living index should cover the *food, clothing, fuel and lighting, house-rent* and *miscellaneous* groups. Each group should include a representative sample of the items of consumption. A separate index number is to be published for each of the major groups and a general index for all the groups combined. In calculating this index, weights are to be used proportional to the relative importance in consumption of the items in a group and also of the different groups. For each item there will be a number of price quotations covering different brands and markets. The price relative of an item is the

simple average of the price relatives for the different quotations of the item. A group index is an weighted average of the price relatives of the different items of the group, the weights being proportional to their consumption expenditure. The general index is in its turn the weighted average of the group indices, the weights being proportional to the consumption expenditure on the different groups.

The question of determining the list of items to be priced and their weights is very important. The items should represent the consumption level of the given group of people. This is found by means of a *family budget enquiry*. On the basis of this enquiry, a list of items representing the level of living can be determined. An obvious criterion for the selection of items is their importance. For a satisfactory picture of the price movements, all types of items having characteristic price movements should be included. Weights which are proportional to consumption expenditure are also determined from the family budget enquiry.

With a change in the consumption pattern, there arises a need for a new study of consumer purchases. Even in the normal course of events, economic changes sometimes outmode the old consumption pattern. As a result of wars and economic upheavals, very great changes occur in consumption pattern. Such changes in the pattern necessitate the undertaking of a fresh family budget enquiry on whose basis the items and weights have to be modified.

Ex. 24.1 With the following data relating to India, compute index numbers of wholesale crop prices for the year 1969-70, taking 1968-69 as base and using the Laspeyres', Paasche's, Edgeworth-Marshall and 'ideal' formulae.

WHOLESALE CROP-PRICES (UNIT : RS. PER QUINTAL)
IN 1968-69 AND 1969-70

Year	Rice	Wheat	Jowar	Barley	Maize	Gram
1968-69	119.00	82.56	56.00	55.62	60.58	83.42
1969-70	111.67	95.42	56.00	61.40	55.84	101.33

CROP-PRODUCTION (UNIT : THOUSAND METRIC TONS)
IN 1968-69 AND 1969-70

Year	Rice	Wheat	Jowar	Barley	Maize	Gram
1968-69	39,761	18,651	9,804	2,424	5,701	4,309
1969-70	40,490	20,093	9,721	2,716	5,674	5,546

Let p_{0i} , q_{0i} and p_{1i} , q_{1i} denote the prices and quantities for 1968-69 and 1969-70, respectively. Then

$$\sum_i p_{0i} q_{0i} = 7,660,056,$$

$$\sum_i p_{1i} q_{0i} = 7,672,622,$$

$$\sum_i p_{0i} q_{1i} = 7,971,866$$

and $\sum_i p_{1i} q_{1i} = 8,022,043.$

The wholesale price index according to the four formulae will then be as follows :

Laspeyres' formula

$$I_{01} = \frac{7,672,622}{7,660,056} \times 100 = 100.16.$$

Pasche's formula

$$I_{01} = \frac{8,022,043}{7,971,866} \times 100 = 100.63.$$

Edgeworth's formula

$$I_{01} = \frac{15,694,665}{15,631,922} \times 100 = 100.40.$$

Fisher's 'ideal' formula

$$I_{01} = \sqrt{100.16 \times 100.63} = 100.39.$$

Ex. 24.2 The group indices for wholesale prices in India, with the year ended August, 1939=100, and the corresponding weights for the week ended September 13, 1958, are shown below. Calculate the general index, using (a) weighted arithmetic mean and (b) weighted geometric mean.

Group	Weight	Index
Food articles	31	473.6
Industrial raw materials	18	510.2
Semi-manufactures	17	405.3
Manufactures	30	390.2
Miscellaneous	4	624.4

Using weighted arithmetic mean, the general index is

$$I_{01} = \frac{\sum \text{Weight} \times \text{Index}}{\sum \text{Weight}} = \frac{44,958.9}{100} = 449.6 \text{ (approx.)}.$$

Using weighted geometric mean,

$$\log I_{01} = \frac{\sum \text{Weight} \times \log (\text{Index})}{\sum \text{Weight}} = \frac{264.92976}{100} = 2.6492976,$$

so that $I_{01} = 446.0$ (approx.).

TABLE 24.1

RETAIL PRICES DURING 1939 AND DURING JULY, 1956, AND
WEIGHTS FOR DIFFERENT FOOD ARTICLES

Article (1)	Unit (2)	Price in Rs.		Weight (w_i) (5)	Price relative p_{1i}/p_{0i} (6)
		1939 (p_{0i}) (3)	July, 1956 (p_{1i}) (4)		
Rice	seer	.07	.45	22	6.4286
Fatni	Do.	.07	.50	6	7.1429
Wheat	Do.	.09	.51	3	5.6667
Jowar	Do.	.07	.45	1	6.4286
Bajra	Do.	.08	.46	4	5.7500
Turdal	Do.	.12	.54	4	4.5000
Gram	Do.	.11	.49	1	4.4545
Raw sugar	½ seer	.06	.27	1	4.5000
Sugar	Do.	.24	.88	5	3.6667
Tea	lb.	.61	2.52	2	4.1311
Fish, dry	dozen	.06	.21	3	3.5000
Fish, fresh	each	1.05	3.05	1	2.9048
Fish, prawn	dozen	.40	1.75	2	4.3750
Fish, breams	Do.	.12	.80	2	6.6667
Mutton	½ seer	.28	1.07	5	3.8214
Milk	Do.	.30	1.00	7	3.3333
Vanaspati	2 lb.	.75	2.50	2	3.3333
Salt	seer	.08	.15	1	1.8750
Chillies, dry	½ seer	.20	1.03	3	5.1500
Tamarind	Do.	.07	.36	2	5.1429
Turmeric	Do.	.12	.80	2	6.6667
Potatoes	Do.	.07	.24	1	3.4286
Onions	Do.	.05	.07	1	1.4000
Brinjals	Do.	.12	.22	5	1.8333
Pumpkins	Do.	.07	.22	5	3.1429
Oil, copoanut	Do.	.16	.85	2	5.3125
Oil, sweet	Do.	.13	.75	2	5.7692
Tea, readymade	cwt	.04	.06	5	1.5000
Total	—	—	—	100	—

Ex. 24.3 Suppose it is required to determine the cost of living index number for the working class people of Bombay city for July, 1956, with the year 1939 as base. For this purpose, it is first of all necessary to obtain individually the indices for the groups : food, fuel & light, clothing, house-rent and miscellaneous.

The basic data for the food group are given in the first five columns of Table 24.1. The last column of this table shows the price relatives. Taking the weighted arithmetic mean of the price relatives, the weights being taken from col. (5) of the said table, we obtain the food index :

$$I_{\text{food}} = 100 \times \frac{\sum_i \frac{p_{1i} w_i}{p_{0i}}}{\sum_i w_i} = 469.2.$$

Likewise, the indices of the other groups are found to be

$$I_{\text{fuel and light}} = 301.2,$$

$$I_{\text{clothing}} = 407.7,$$

$$I_{\text{house-rent}} = 106.3$$

and $I_{\text{miscellaneous}} = 346.7.$

For the general cost of living index number, the following weights are used :

food— 53

fuel & light— 8

clothing— 9

house-rent— 14

miscellaneous— 16.

Applying these weights to the group indices, we have finally the general cost of living index number, viz.

$$I = (53 \times 469.2 + 8 \times 301.2 + 9 \times 407.7 + 14 \times 106.3 + 16 \times 346.7) / 100 \\ = 379.8.$$

24.8 Cost of living index number and Laspeyres' and Paasche's formulæ

A cost of living index number may be defined as an index of change in the money required to get equal satisfaction in two different situations. Let $q_1^1, q_2^1, \dots, q_n^1$ be the series of quantities of a collection of n consumer goods and services which yield equivalent

satisfaction in the current year as compared with the base year series $q_1^0, q_2^0, \dots, q_n^0$. The cost of living index number I , for the current year relative to the base year, is given by

$$I = \frac{\sum p^1 q^1}{\sum p^0 q^0}$$

If p^1 's and p^0 's denote the consumer prices in the current year and in the base year, respectively.

This I is called the true cost of living index number. This I , however, cannot be possibly determined, since it is not possible in practice to determine the quantities q^1 's which would yield the same satisfaction as the q^0 's. The different formulæ given in subsections 24.2.5 and 24.2.6 would only approximate the true index I .

Stated in another way, the problem of measuring the true change in cost of living consists in identifying equal real incomes in two different situations and in determining the ratio of money values of these two real incomes. Strictly speaking, a separate index of this kind should be calculated for each distinguishable real income level. Let I_0 and I_1 be the true index numbers calculated on the basis of the real income levels prevailing in the base period and the current period, respectively. I_1 differs from I_0 due to the change of consumption pattern as a result of a change in real income level in the two situations—relatively increasing or decreasing the consumption of items which have advanced most in price and relatively decreasing or increasing the consumption of items which have advanced least. If the change in real income resulted in no change in consumption pattern, i.e. if income elasticity of demand (*vide* Chapter 26) were unity for all goods and services, I_0 and I_1 would be identically equal.

The difference between the results obtained by Laspeyres' formula (L) and Paasche's formula (P) may occur due to two reasons—the first being the same as the reason for which I_0 and I_1 differ and the second being a possible change in consumption pattern attributable to a change in relative prices. It is for this reason that $L > I_0$, the numerator of L being too large, because L assumes that consumers do not alter their consumption in response to relative price changes, buying more of the cheaper articles. For the same

reason, the denominator of P is too large, so that $P < I_1$. Thus it is said that Laspeyres' formula has an upward bias and Paasche's formula has a downward bias.

But this statement should be taken with some caution. If the price elasticity of demand (*vide* Chapter 26) were zero for all the commodities, then $L=I_0$ and $P=I_1$, so that

$$L-P=I_0-I_1=k, \text{ say.}$$

Let e represent the difference between L and P due to the second factor, so that

$$\begin{aligned} L-P &= \{(L-I_0)+(I_1-P)\} + (I_0-I_1) \\ &= e+k=d, \text{ say.} \end{aligned}$$

e is necessarily positive provided the tastes and preferences of consumers remain unchanged. k may be either positive or negative, so that d is either positive or negative. Thus it is quite possible that Paasche's formula would give a higher result than Laspeyres formula.

24.9 Two important index number series

- Index number of wholesale prices in India (revised series)*

Base : April, 1961—March, 1962=100.

Computation : weighted arithmetic mean of price relatives with weights proportional to the total values of quantities marketed during the base period. The number of items in each group and their respective percentage weights are :

Group	Number of items	Percentage weight
Food Articles	38	41·3
Liquor and Tobacco	3	2·5
Fuel, Power, Light & Lubricants	10	6·1
Industrial Raw Materials	25	12·1
Chemicals	11	7·9
Machinery and transport equipments	7	0·7
Manufactures	45	29·4
Total	1.	100·0

The above groups are further sub-divided into a number of sub-groups for which indices are computed.

The index is calculated weekly from once-a-week prices (on or about Friday), for 774 quotations on 139 items. For each variety, the price as well as the price relative is also published. This is issued by the office of the Economic Adviser to the Government of India and published in the weekly publication *Index Number of Wholesale Prices in India*. The table below gives the wholesale price index numbers for India for a number of years (averages of weekly index numbers) :

Year	Index
1965	129·2
1966	144·5
1967	166·2
1968	165·9
1969	168·8
1970	179·2

Cost of living index numbers, covering 25 towns in West Bengal, for five expenditure groups

The Bureau of Applied Economics and Statistics (formerly called the State Statistical Bureau) of West Bengal is currently publishing cost of living index numbers for 25 towns including Calcutta. The series have the year 1960 as base and have weights based on family budget enquiries conducted in 1960-61. The indices are constructed separately in respect of each of 5 monthly expenditure groups, namely, (i) up to Rs. 100, (ii) Rs. 101 to Rs. 200, (iii) Rs. 201 to Rs. 350, (iv) Rs. 351 to Rs. 700 and (v) above Rs. 700. In all, 468 price quotations are collected in respect of 87 items or sub-groups for which weights are determined, and they are divided into 5 major groups as follows :

Food—	28	items or subgroups
Clothing—	6	,,
Fuel & light—	9	,,
House-rent—	3	,,
Miscellaneous —	41	,,

Originally, the cost of living index numbers published by the Bureau, with August, 1939 as base, used the results of the surveys of the Indian Statistical Institute and the Bureau for determining the weights. In 1950-51, the Bureau conducted a regular full-scale family budget survey for the first time in 23 towns including Calcutta and prepared revised weights for the above-mentioned five monthly expenditure levels.

A fresh family-budget enquiry was undertaken in 1955-56 for Calcutta and 23 other towns of West Bengal for the above five expenditure levels. It was found that there had been a fall in the expenditure on food items and a rise in that of miscellaneous items as compared to 1950-51. There was thus a distinct shift in the pattern of consumption. There was a family budget survey in 1960-61 again for all the 23 towns in the 1950-51 series and for Purulia and Kalimpong. The present series is computed for these 25 centres.

The following table gives the weights for 1960-61 survey for Calcutta general holdings :

TABLE 24.2
**WEIGHTS FOR CALCUTTA GENERAL HOLDINGS, OBTAINED
 FROM 1960-61 FAMILY-BUDGET SURVEY**

Group	Monthly expenditure level (in Rs.)				
	1 - 100	101 - 200	201 - 350	351 - 700	701 and above
Food	62.22	59.75	54.31	47.48	42.71
Clothing	5.81	6.80	7.36	7.45	7.11
Fuel and light	5.82	5.39	4.91	4.82	4.06
Housing	11.96	10.64	10.50	11.83	12.83
Miscellaneous	14.19	17.42	22.92	28.42	33.29

Source : *A Brief Note on the Methodology of Construction of Consumer Price Index Numbers*. Bureau of Applied Economics & Statistics, West Bengal, 1972.

The index numbers are computed as weighted averages of price relatives. Monthly index numbers are published in the *Monthly*

Statistical Digest of West Bengal and Statistical Abstract (Annual), West Bengal. For Calcutta, however, a weekly index is published in the *Calcutta Gazette*.

Below are shown the index numbers for three consecutive years (averages of monthly indices) for all the 5 expenditure levels of Calcutta :

TABLE 24.4

COST OF LIVING INDEX FOR CALCUTTA : BASE (1960=100)
MONTHLY AVERAGES

Year	Monthly expenditure level (in Rs.)				
	1—100	101—200	201—350	351—700	701 and above
1972	188·4	185·9	180·7	176·5	175·3
1973	206·5	204·0	198·0	193·8	193·4
1974	263·1	259·0	248·7	240·5	240·5

24.10 Uses of index numbers

In addition to serving the basic purpose for which they are constructed, index numbers are also of use for the following purposes :

(a) *Purchasing power*

The purchasing power of money (say rupee) is the quantity of goods that a given quantity of money will buy. The reciprocal of a price index number is used to show the purchasing power of money. A price index is the amount of money required to purchase a fixed basket of goods, and the reciprocal of the price index—the purchasing power—represents the quantity of goods that can be purchased with a fixed amount of money. The purchasing power will be relative to the base period of the price index.

In 1963, the cost of living index number for the expenditure group (Rs. 351—Rs. 700) was 119·8 with November, 1950 as base. The purchasing power of the November, 1950 rupee for the said expenditure group was, therefore, $100·0/119·8$ or 0·835 in 1963. This means that in 1963, the November, 1950 rupee would purchase 0·835 times the amounts it could purchase in November, 1950.

(b) *Deflation*

Another use of index numbers is in adjusting a value series by dividing the series by a price index or by multiplying the series by the index of purchasing power. By this the unit of money is expressed in terms of the purchasing power in the base-year. This process, which is known as *deflation*, is not limited to value series only. Wages are deflated by cost of living index, departmental store sales by retail price index, population data by an index of population, and so on.

(c) *Indicator of general business conditions*

Index numbers are also used in studying the general business conditions. A company may plan its activities by studying the wholesale price index number. The index of industrial production may be studied to follow changes in the volume of production, etc.

Questions and exercises

24.1 Describe the different problems faced in constructing index numbers.

24.2 Discuss the different steps for constructing a wholesale price index number for India.

24.3 Discuss how you will proceed for constructing a cost of living index number for a given expenditure group in Calcutta.

24.4 What is a chain index? Discuss its advantages and disadvantages over a fixed-base index number.

24.5 What purpose is served by an index number? Show that the factor reversal test and time reversal test are not satisfied by Laspeyres' and Paasche's index numbers. Further show that both these tests are satisfied by Fisher's ideal index number.

24.6 Examine the important formulæ for the calculation of price index numbers in the light of the various tests devised for this purpose.

24.7 State the different uses of index numbers.

24.8 The table below gives the wholesale prices and quantities produced of a number of commodities in India. Calculate Laspeyres', Paasche's, Edgeworth-Marshall and Fisher's 'ideal' index numbers for the years 1952 to 1954 with 1951 as base

Commodity	1951		1952		1953		1954	
	p	q	p	q	p	q	p	q
Rice	16.87	20,964	17.50	22,537	17.50	27,769	16.73	24,209
Jowar	10.09	5,981	11.93	7,243	12.08	7,954	11.22	9,092
Bazra	10.07	2,309	13.33	3,142	13.33	4,475	11.45	3,555
Maize	21.75	2,043	15.67	2,825	14.70	2,991	10.77	2,944
Ragi	9.45	1,291	9.30	1,316	15.96	1,846	10.21	1,778
Wheat	18.60	6,085	23.67	7,382	21.93	7,890	16.42	8,539
Barley	20.66	2,330	17.29	2,882	18.80	2,905	10.17	2,786
Gram	24.09	3,334	19.01	4,142	19.60	4,756	12.30	5,125

p : price in Rs. per maund ;

q : quantity produced in thousand tons.

Partial ans. Indices for 1952 are 103.42; 103.50; 103.47; 103.46.

24.9 The following data relate to the wholesale prices of cereals at selected centres in India during two different weeks and the corresponding weights :

Item	Weight	Price (Rs. per maund)*	
		week ending 17-11-56	week ending 21-12-57
Rice	224	20.50	17.50
Wheat	106	18.50	17.40
Jowar	19	16.25	10.50
Bazra	10	15.50	12.44
Barley	10	13.00	11.25
Maize	9	13.00	11.06
Ragi	4	10.12	12.75

How did the wholesale prices of cereals in India during the week ending 21-12-57 compare with those in the week ending 17-11-56 ?

* *Partial ans.* The index number is 87.06.

24.10 The following data show the cost of living indices for the groups : Food, Clothing, Fuel and Light, House-rent, and Miscellaneous, with their respective weights, for middle class people of Calcutta in 1957. Obtain the general cost of living index number.



Mr. X was getting Rs. 250/- in 1939 and Rs. 429/- in 1957. State how much he ought to have received as extra allowance in 1957 to maintain his pre-war standard of living.

Base : 1939 = 100

<i>Group</i>	<i>Group index</i>	<i>Group weight</i>
Food	411.8	61.22
Clothing	544.8	4.51
Fuel and Light	388.0	6.58
House-rent	116.9	8.97
Miscellaneous	284.5	18.72

Ans. 365.95 ; Rs. 485/87 P.

24.11 The following data relate to the group indices and the corresponding weights (shown in brackets) for the menial class cost of living index numbers in Calcutta :

Year	Food (71.28)	Clothing (2.89)	Fuel & Light (9.27)	House-rent (6.69)	Miscellaneous (9.87)
1948	370.1	423.3	469.1	110.0	279.2
1949	387.2	440.4	469.8	115.8	287.1
1950	394.0	432.9	352.0	116.9	285.1
1951	396.7	551.4	366.0	116.9	291.8
1952	380.2	504.2	336.8	116.9	283.6

Calculate the general cost of living index for each of the above years.

The total wages and the number of workers employed in jute mills around Calcutta are given below .

Year	Total wages, (Rs. Lakhs)	Number of workers (000)
1948	2,076	319
1949	2,453	306
1950	2,246	291
1951	2,231	272
1952	2,552	275

Calculate the average nominal wages and real wages for the jute textile workers, using the general cost of living indices for the menial class people of Calcutta. *Partial ans.* Cost of living indices :

354.44 ; 368.36 ; 361.94 ; 369.25 ; 352.61.

SUGGESTED READING

- [1] Croxton, F. E. and Cowden, D. J. *Applied General Statistics* (Chs. 20-21). Prentice-Hall, 1967, and Prentice-Hall of India, 1969.
- [2] Dubois, E. N. *Essential Methods in Business Statistics* (Ch. 14). McGraw-Hill, 1964.
- [3] Greenwald, W. I. *Statistics for Economics* (Ch. 6). C. E. Merrit Books, 1963.
- [4] Mills, F. C. *Statistical Methods* (Ch. 13). Henry Holt, 1955.
- [5] Mudgett, B. D. *Index Numbers* (Chs. 1—7). John Wiley, 1951.

25

ANALYSIS OF TIME SERIES

25.1 Introduction

In this chapter we shall deal with statistical data which relate to successive intervals or points of time. These are referred to as *time series*. Examples of time series are yearly, quarterly or monthly production or consumption figures for a particular commodity, price of a commodity at different points of time, etc. Although the term 'time series' usually refers to economic data, and we too shall be concerned here with economic data, it equally applies to data arising in natural and other social sciences. Here the time sequence is of prime importance, and it requires special techniques for the analysis of the series. We analyse the past in order to understand the future better.

Symbolically, y_t denotes the value of the variable at time t ($t=1, 2, \dots, n$). In case the figures relate to n successive periods (and not points of time), t is to be taken as the mid-point of the t th period.

25.2 Preliminary adjustments of time series data

Before we subject the time series data to statistical analysis, we have to see that they represent a series of comparable figures over time. A series of figures may not be comparable or homogeneous for various reasons. It may be that the figures relate to geographical areas, which, however, changed from time to time. The series may relate to populations, which we know are always changing over time. The definitions of different terms and concepts also may change from time to time making the data non-comparable.

Industrial or mineral production data over different months are not homogeneous, since the number of days in different calendar months, as well as the number of working days, is not the same.

Figures given in monetary terms are not comparable over time, since with changes in the price-level the value of money, as measured by its purchasing power, changes. Thus the figures of wages or incomes or the money values of sales of goods have to be brought to a comparable basis, eliminating the effect of price-changes.

Thus the raw data have to be subjected to preliminary adjustments. The figures which are related to geographical areas or populations should be brought to per unit or per capita basis, dividing the figures by the geographical areas or the populations to which they relate. If the figures involve definitions of terms and concepts, adjustment factors have to be found out for any changes of definition over time.

Monthly production figures subject to calendar variation of number of working days should be made comparable by dividing each figure by the number of working days to convert the figures into per day basis. The figures given in monetary terms have to be expressed in terms of value of money in a certain base period. This will necessitate dividing or deflating the current figure by the index number of prices of the current period with the chosen base period. If the index number be I_{0t} in per cent form, 100 rupees in base period has the same purchasing power as I_{0t} rupees in the current period. Thus a figure x_t in money terms in the current period expressed in terms of base period purchasing power would be

$$x'_t = x_t \times \frac{100}{I_{0t}}.$$

25.3. Components of time series

A graphical representation of a time series reveals the changes over time. A series which exhibits no change during the period under consideration will give a horizontal line. However, usually we shall come across time series showing continual changes over time, giving us an overall impression of haphazard movement. A critical study of the series will, however, reveal that the changes are not totally haphazard and a part of it, at least, can be accounted for. The part which can be accounted for is the systematic one and the remaining part is the unsystematic or irregular. The systematic part may be attributed to several broad factors, viz. (1) secular trend, (2) seasonal variation and (3) cyclical variation. In a given time series, some or all of the above components may be present. Separation of the different components of a time series is of importance, because it may be that we are interested in a particular component or that we want to study the series after eliminating the effect of a

particular component. It may be noted that it is the systematic parts of the time series which may be used in forecasting.

In the classical or traditional approach, it is assumed that there is a multiplicative relationship among the four components ; i.e., any particular value (y_t) is considered to be the product of the factors attributable to secular trend (T_t), seasonal (S_t), cyclical (C_t) and irregular (I_t) components. Thus

$$y_t = T_t \times S_t \times C_t \times I_t. \quad \dots \quad (25.1)$$

Another approach is to assume y_t to be the sum of the four components :

$$y_t = T_t + S_t + C_t + I_t. \quad \dots \quad (25.2)$$

This model, however, is not generally used since it is considered inappropriate for most economic data. However, if y_t represents the logarithm of the original variable, then one may well use this simpler, additive model instead of the multiplicative model (25.1).

By secular trend (or, simply, *trend*) of time series we mean the smooth, regular, long-term movement of a series if observed long enough. Some series may exhibit an upward or a downward trend

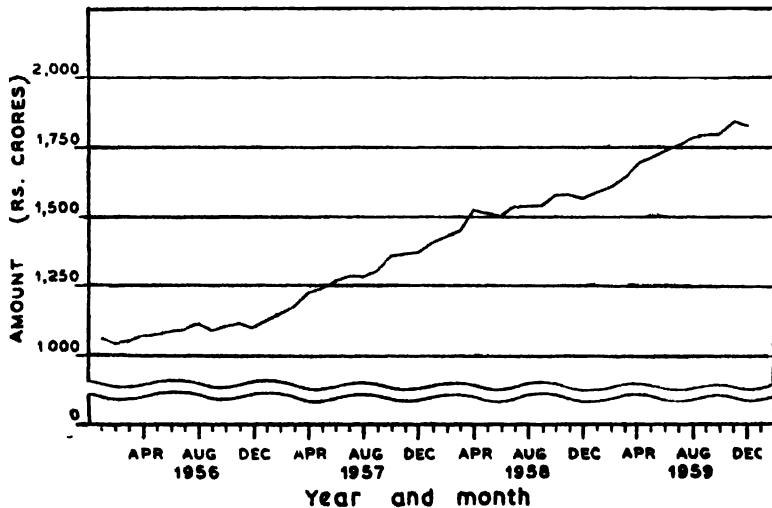


Fig. 25.1 Deposit liabilities of scheduled banks in India.

or may remain more or less at a constant level. Again, some series after a period of growth (decline) may reverse their course and enter

a period of decline (growth). But sudden or frequent changes are incompatible with the idea of trend. Fig. 25.1 illustrates a series exhibiting an upward trend, other components being almost absent.

By seasonal fluctuations we mean a periodic movement in a time series where the period is not longer than one year. A periodic movement in a time series is one which recurs or repeats at regular intervals of time (or periods). Examples of seasonal fluctuations may be found in the passenger traffic during the 24 hours of a day, sales of departmental stores during the 12 months of a year, issue of library books during the seven days of a week, and so on. The factors which mainly cause this type of variation in economic time series are the climatic changes of the different seasons and the customs and habits which the people follow at different times. E.g., the occurrence of a festival in a particular month will increase the sale of certain consumer goods in that month. The study and measure-

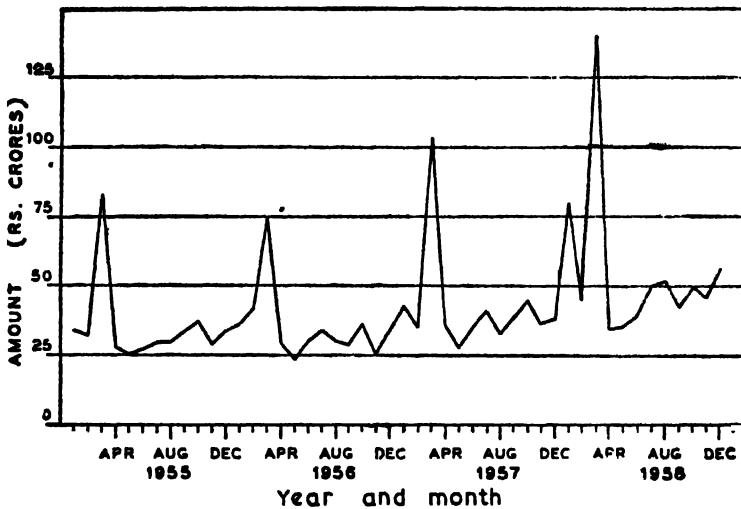


Fig. 25.2 Revenue expenditure and defence drawings, Govt of India

ment of this component is of prime importance in certain cases. E.g., the efficient running of any departmental store would necessitate a careful study of seasonal variation in the demand of the goods. Fig. 25.2 illustrates a series exhibiting marked seasonal variation, the other components being negligible.

By *cyclical fluctuations* we mean the oscillatory movement in a time series, the period of oscillation being more than a year (Fig. 25.3). One complete period is called a cycle. The cyclical fluctuations are not necessarily periodic, since the length of the cycle as also the intensity of fluctuations may change from one cycle to another. Every business man is familiar with the alternating periods of 'prosperity' (or 'boom') and 'depression' in business which follow one another in an irregular manner.

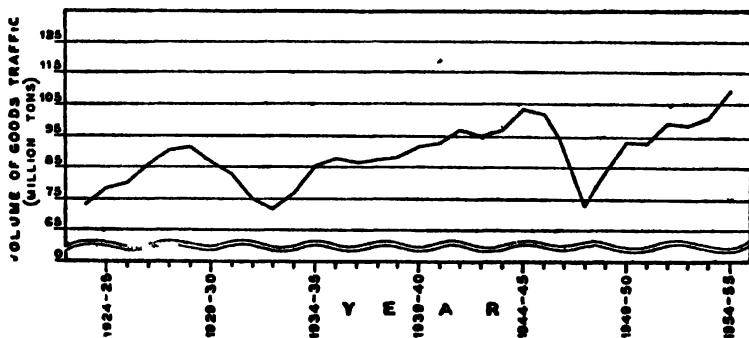


Fig. 25.3 Volume of goods traffic carried by Indian Railways.

Irregular fluctuations are those which are either wholly unaccountable

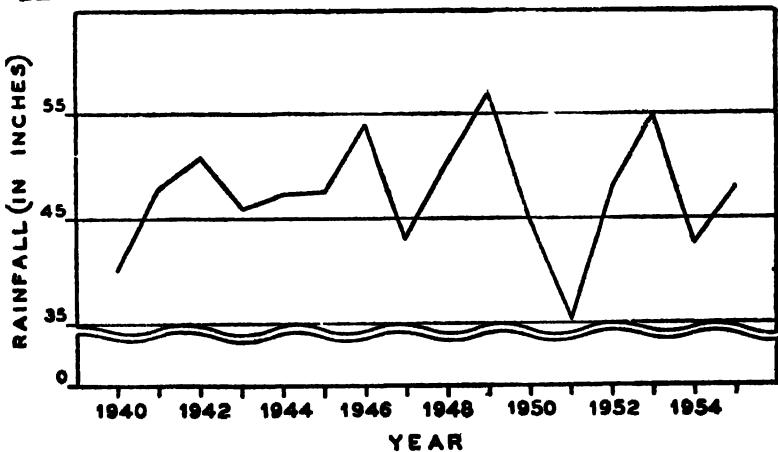


Fig. 25.4 Annual rainfall in Bihar.

or are caused by such unforeseen events as wars, floods, strikes, etc.

This category of movements includes all types of variation that are not accounted for by secular trend, seasonal or cyclical fluctuations (Fig. 25.4).

We now proceed to separate out the various components in a time series. We shall present the classical method, which assumes the multiplicative model (25.1). T_t is expressed in the same units in which y_t is reported. The other components are relatives, which are generally stated as percentages.

25.4 Measurement of secular trend

In order to measure trend, we are to eliminate from the time series the other three components, viz. seasonal fluctuations, cyclical fluctuations and irregular fluctuations. If the period of seasonal fluctuations be a year, then the yearly totals or yearly averages will be free from the seasonal effect. Thus, in determining trend from monthly data, it is customary to start with the yearly totals or averages, which are free from the seasonal effect. The monthly trend values can be obtained from the annual trend values by interpolation. To eliminate the other two components, viz. the cyclical and the irregular, we may consider the following methods :

Method of free-hand curve-fitting

In this method we first draw the line-diagram for the yearly data. Then we draw a free-hand smooth curve which seems to fit the data best. The method, however, is quite subjective, and its use therefore calls for sound judgment. The method is quite flexible, can be used for all types of trend, linear or non-linear, and requires a minimum of labour.

Method of moving averages

The moving average of period k of a time series gives us a new series of arithmetic means, each of k successive observations of the time series. We start with the first k observations. At the next stage, we leave the first and include the $(k+1)$ st observation. This process is repeated until we arrive at the last k observations. Each of these means is centred against the time which is the mid-point of the time interval included in the calculation of the moving average. Thus when k , the period of the moving average, is odd, the moving average values correspond to tabulated time values for which the time

series is given. When the period is even, the moving average falls midway between two tabulated values. In this case, we calculate a subsequent 2-item moving average to make the resulting moving average values correspond to the tabulated time periods.

A moving average with a properly selected period will smooth out cyclical fluctuations from the series and give an estimate of the trend. The central problem in this method is thus the selection of an appropriate period which will eliminate all fluctuations that draw the series away from the trend.

Cyclical fluctuations with a uniform period and a uniform amplitude (height) can be completely eliminated by taking a period of the moving average which is equal to (or a multiple of) the period of the cycles, provided the trend is linear. However, cycles in economic time series are not strictly periodic. The period and the amplitude generally vary from cycle to cycle. In such cases, the best results may be obtained by using a moving average whose period is equal to the average period of the cycles. This, however, will not completely eliminate the cycles.

- There will be further complications if the trend is non-linear. If the trend is concave upwards, a moving average will always overestimate the trend values. If the trend is convex upwards, a moving average will underestimate the trend values.

Like the graphical method, the method of moving averages is flexible ; the moving averages can adapt themselves to changing circumstances, i.e. any change in the trend will be faithfully reflected by them. But unlike the graphical method, this method has the merit of objectivity since the period of the moving averages can be more or less objectively determined. It should be noted, however, that since this method assumes no law of change, it cannot be used for forecasting purposes. Besides, in this process a number of trend values at each end of the series remain unestimated.

Ex. 25.1 In Table 25.1 data relating to the yield of wheat in India during the years 1947-48 to 1967-68 are given. The data show an increasing trend with a marked cyclical effect superimposed on it. In order to eliminate the cyclical fluctuations, and thereby determine the underlying trend, we may use the method of moving averages

TABLE 25.1
DETERMINATION OF TREND BY THE METHOD OF
MOVING AVERAGES FOR YIELD OF WHEAT
IN INDIA, 1947-48 to 1967-68

Year	Yield (000 tonnes)	3-year moving total	Trend value (3-year moving average)
1947-48	5,570	—	—
1948-49	5,650	17,510	5,836.7
1949-50	6,290	18,402	6,134.0
1950-51	6,462	18,837	6,279.0
1951-52	6,085	19,929	6,643.0
1952-53	7,382	21,357	7,119.0
1953-54	7,890	24,172	8,057.3
1954-55	8,900	25,550	8,516.7
1955-56	8,760	26,728	8,909.3
1956-57	9,068	25,826	8,608.7
1957-58	7,998	27,024	9,008.0
1958-59	9,958	28,280	9,426.7
1959-60	10,324	31,279	10,426.3
1960-61	10,997	33,393	11,131.0
1961-62	12,072	33,845	11,281.7
1962-63	10,776	32,701	10,900.3
1963-64	9,853	32,919	10,973.0
1964-65	12,290	32,567	10,855.7
1965-66	10,424	34,107	11,369.0
1966-67	11,393	38,384	12,794.7
1967-68	16,567 *	—	—

In the present case, the peak years are 1950-51, 1954-55, 1956-57, 1961-62 and 1964-65, so that the periods of the cycles are 4, 2, 5 and 3 years, respectively. Since the average period lies between 3 and 4 years, we may take for simplicity 3-point moving averages, which

will give the required trend values, for the years 1948-49 to 1966-67. The calculations are also shown in the table. The original data and the trend values are plotted in Fig. 25.5.

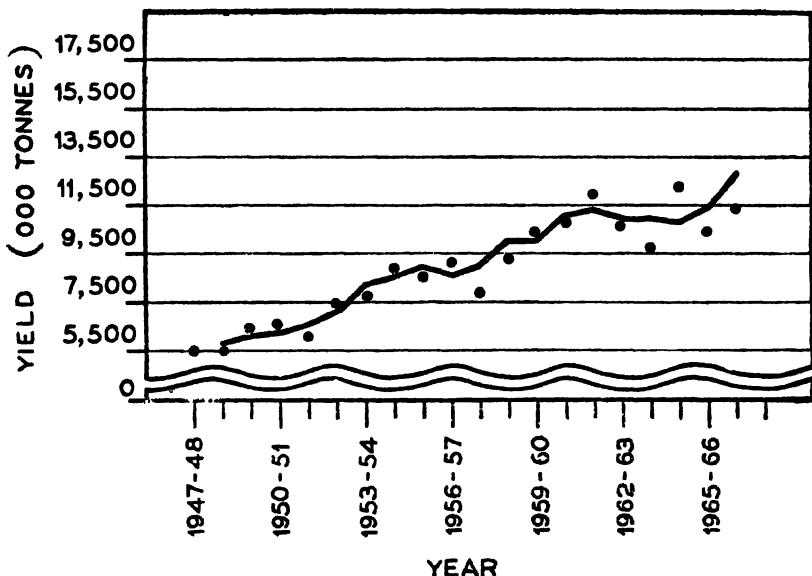


Fig. 25.5 Trend fitted by the method of moving averages to the data on yield of wheat in India.

Method of mathematical curves

This is perhaps the best and most objective method of determining trend. In this case, an appropriate type of trend equation is at first selected, and then the constants involved in the equation are estimated on the basis of the data in hand. Usually, a polynomial of a suitable degree is chosen either for the original variable or for a transformed variable and its constants determined by the method of least squares. The choice of the appropriate polynomial is facilitated by a graphical representation of the data, for which, apart from the usual arithmetic scales, semi-logarithmic or doubly-logarithmic scales may be used.

Supposing a polynomial of degree k in t is chosen to represent the trend T , viz.

$$T_t = a_0 + a_1 t + a_2 t^2 + \dots + a_k t^k. \quad (25.3)$$

the normal equations for determining the unknown constants $a_0, a_1, a_2, \dots, a_k$ will be

$$\left. \begin{aligned} \sum y &= a_0 + a_1 \sum t + a_2 \sum t^2 + \dots + a_k \sum t^k, \\ \sum ty &= a_0 \sum t + a_1 \sum t^2 + a_2 \sum t^3 + \dots + a_k \sum t^{k+1}, \\ \sum t^2y &= a_0 \sum t^2 + a_1 \sum t^3 + a_2 \sum t^4 + \dots + a_k \sum t^{k+2}, \\ &\dots \quad \dots \quad \dots \quad \dots \quad \dots \\ \sum t^ky &= a_0 \sum t^k + a_1 \sum t^{k+1} + a_2 \sum t^{k+2} + \dots + a_k \sum t^{2k}. \end{aligned} \right\} \dots \quad (25.4)$$

Using the estimates obtained from equations (25.4), we can get the trend value for any given time t by substituting that value of t in (25.3). Obviously, for linear trend,

$$T_t = a_0 + a_1 t,$$

and there will be two normal equations, viz

$$\sum y = a_0 + a_1 \sum t$$

$$\text{and } \sum ty = a_0 \sum t + a_1 \sum t^2.$$

For quadratic trend,

$$T_t = a_0 + a_1 t + a_2 t^2,$$

and the normal equations are

$$\sum y = a_0 + a_1 \sum t + a_2 \sum t^2,$$

$$\sum ty = a_0 \sum t + a_1 \sum t^2 + a_2 \sum t^3$$

$$\text{and } \sum t^2y = a_0 \sum t^2 + a_1 \sum t^3 + a_2 \sum t^4.$$

Usually, the successive points of time will be equidistant, the common difference being h , say. By taking as origin the mid-point of the period covered by the data, one can then make each sum of odd powers of t equal to zero. Further simplifications can be made if one takes h or $h/2$ as the new unit for t , according as the number of points is odd or even. The method is illustrated below.

Ex. 25.2 The first two columns of Table 25.2 give the data on the production of coal in India for a number of years. A graphical representation of the data indicates that a quadratic trend will be appropriate. The necessary calculations to fit a quadratic trend are shown in the other columns of the table.

TABLE 25.2
FITTING A QUADRATIC TREND TO THE DATA ON
PRODUCTION OF COAL IN INDIA

Year	Production (000 metric tons) y	t =Year-1962	ty	t^2y	t^3	t^4
1959	47,800	-3	-143,400	430,200	9	81
1960	52,593	-2	-105,186	210,372	4	16
1961	56,065	-1	-56,065	56,065	1	1
1962	61,370	0	0	0	0	0
1963	65,956	1	65,956	65,956	1	1
1964	62,440	2	124,880	249,760	4	16
1965	67,162	3	201,486	604,458	9	81
Total	413,386	0	87,671	1,616,811	28	196

Here

$$\sum t = 0, \sum t^3 = 0.$$

Hence the normal equations are :

$$413,386 = 7a_0 + 28a_1,$$

$$87,671 = 28a_1$$

and $1,616,811 = 28a_0 + 196a_1.$

From the second equation,

$$a_1 = 3,131 \cdot 11.$$

Solving the other two equations for a_0 and a_1 , we have

$$a_0 = 60,804 \cdot 34$$

and $a_2 = -437 \cdot 30.$

Therefore, the trend equation is given by

$$T_t = 60,840 \cdot 34 + 3,131 \cdot 11t - 437 \cdot 30t^2.$$

Table 25.3 and Fig. 25.6 show the fitted trend together with the observed series.

TABLE 25.3

**QUADRATIC TREND FITTED TO THE DATA ON PRODUCTION
OF COAL IN INDIA**

Year	t — Year — 1962	a_1t	a_2t^2	Trend $T_t = a_0 + a_1t + a_2t^2$	Production y
1959	-3	-9,393.33	-3,935.70	47,475.31	47,800
1960	-2	-6,262.22	-1,749.20	52,792.92	52,593
1961	-1	-3,131.11	-437.30	57,235.93	56,065
1962	0	0	0	60,804.34	61,370
1963	1	3,131.11	-437.30	63,498.15	65,956
1964	2	6,262.22	-1,749.20	65,317.36	62,440
1965	3	9,393.33	-3,935.70	66,261.97	67,162

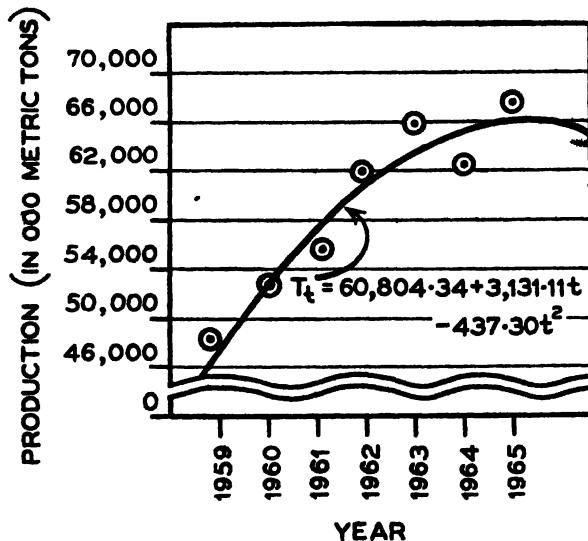


Fig. 25.6 Quadratic trend fitted to the data on production of coal in India.

In the above example we have an odd number of years. In the next example we shall consider data for an even number of years.

Ex. 25.3 Let us take the data of Table 25.4, which relate to the production of pure sulphuric acid in India for the years 1962—1967. In this case a linear trend seems to be appropriate. The necessary computations are done in the table below.

TABLE 25.4
FITTING A LINEAR TREND TO THE DATA ON PRODUCTION
OF PURE SULPHURIC ACID IN INDIA

Year	Production (tonnes)	$t = 2(\text{year} - 1964.5)$	ty	t^2	T_t
1962	469,464	-5	-2,347,320	25	503,388.97
1963	568,152	-3	-1,704,456	9	561,825.85
1964	679,740	-1	-679,740	1	620,262.73
1965	682,343	1	685,343	1	678,699.61
1966	689,738	3	2,069,214	9	737,136.49
1967	804,450	5	4,022,250	25	795,573.37
Total	3,896,887	0	2,045,291	70	-

Since $\sum t=0$, the normal equations are

$$3,896,887 = 6a_0$$

and $2,045,291 = 70a_1,$

so that $a_0 = 649,481.17$

and $a_1 = 29,218.44.$

The trend equation is, therefore,

$$T_t = 649,481.17 + 29,218.44t.$$

The trend values for the different years are shown in the last column of Table 25.4 and in Fig. 25.7.

Sometimes a time series plotted on semi-logarithmic graph paper may give approximately a straight line. Here the trend equation may be taken to be of the *exponential form*:

$$\left. \begin{aligned} T_t &= ab^t \\ \log T_t &= \log a + t \log b. \end{aligned} \right\} \quad \dots \quad (25.5)$$

or

Similarly, if the representation of the data on doubly-logarithmic paper gives approximately a straight line, we may use the following function to give the trend :

$$\begin{aligned} T_t &= at^b \\ \log T_t &= \log a + b \log t. \end{aligned} \quad (25.6)$$

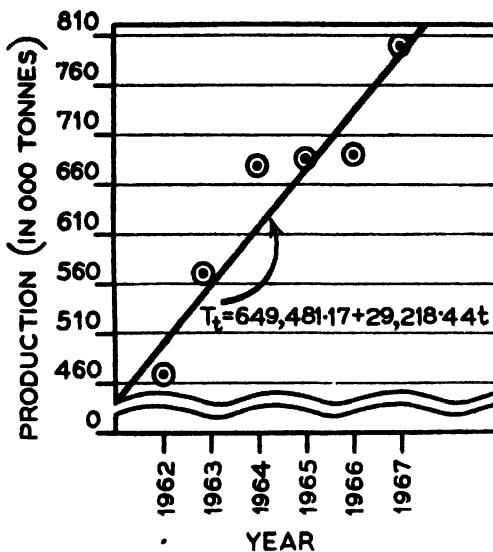


Fig. 25.7 Linear trend fitted to the data on production of pure sulphuric acid in India.

The constants a and b , in each case, may be determined by the least-square method, taking the second form of the corresponding equation.

Group-average method

Usually, the types of equation we have considered will explain trend in a majority of the cases. Occasionally, however, it may be necessary to consider more complicated trend equations. One such is the *modified exponential equation*

$$T_t = k + ab^t. \quad \dots \quad (25.7)$$

The curve approaches k as an upper limit if a is negative and approaches k as a lower limit if a is positive. To determine the

constants of the curve, the whole range of t covered by the data is divided into three equal parts, each including, say, m points of time. Equating the totals

$$S_1 = \sum_1^m y_i, \quad S_2 = \sum_{m+1}^{2m} y_i, \quad S_3 = \sum_{2m+1}^{3m} y_i,$$

to the totals of the corresponding trend values given by (25.7), three equations are obtained, viz.

$$S_1 = \sum_{i=1}^m (k + ab^i) = mk + ab \cdot \frac{1 - b^m}{1 - b},$$

$$S_2 = mk + ab^{m+1} \cdot \frac{1 - b^m}{1 - b}$$

and $S_3 = mk + ab^{2m+1} \cdot \frac{1 - b^m}{1 - b}.$

The three equations are now solved for the three unknowns k , a and b . The values will be found to be

$$b = \left(\frac{S_2 - S_3}{S_1 - S_2} \right)^{1/m},$$

$$a = \frac{(S_1 - S_2)(1 - b)}{b(1 - b^m)^2}$$

and $k = \frac{1}{m} \cdot \frac{S_1 S_3 - S_2^2}{S_1 - 2S_2 + S_3}.$

Two other curves, which can be reduced to the modified exponential form, are the *Gompertz curve* and the *logistic curve*.

Gompertz curve :

$$\begin{aligned} T_t &= ka^{b^t}, \\ \text{or} \quad \log T_t &= \log k + (\log a) b^t. \end{aligned} \quad \left. \right\} \quad . \quad (25.8)$$

$\log T_t$ being of the modified exponential form.

Logistic curve :

$$\begin{aligned} T_t &= \frac{k}{1 + e^{a+b t}}, \\ \text{or} \quad \frac{1}{T_t} &= \frac{1}{k} + \left(\frac{e^a}{k} \right) \left(e^b \right)^t, \end{aligned} \quad \left. \right\} \quad . \quad (25.9)$$

$\frac{1}{T_t}$ being of the modified exponential form.

Semi-average method

The method of semi-averages is nothing but the group-average method for estimation of parameters of mathematical curves. In the case of linear trend, the method reduces to dividing the series of values into two equal halves and plotting the average in each half against the middle of the period covered. Then the required linear trend is the straight line through the two points.

The method of mathematical curves is objective and, since it assumes a law of change, it can be used for forecasting purposes. The method, however, is rigid. If there are sharp changes in the trend, then to use this method the whole series is to be divided into a number of parts, and an appropriate trend equation has to be determined for each part separately. The method is most laborious unless one uses the simple linear or quadratic equations.

25.5 Measurement of seasonal fluctuations

The measurement of seasonal and/or cyclical variation may, in some cases, be as important as the measurement of trend. An understanding of seasonal fluctuations is necessary to plan business efficiently. The head of a departmental store, e.g., must know how the demand for different articles varies from month to month, so that he may provide for stocks in advance and thus keep pace with the demand.

We shall now describe different methods of isolating seasonal variation. For simplicity, we shall consider seasonal variation in monthly or quarterly data only, but the procedure for weekly, daily or hourly data will be quite similar.

Method of monthly (or quarterly) averages

This is a simple method of isolating seasonal variation. It is based on the assumption that the series contains neither a trend nor cyclical fluctuations but only seasonal and irregular fluctuations. Here the irregular variation may be eliminated by averaging the monthly (or quarterly) values over years. To express the averages as indices, they are shown as percentages of the grand mean, so that the total of the seasonal indices is 1,200 (for monthly data) or 400 (for quarterly data). For an additive model, the grand mean is subtracted from the monthly (or quarterly) averages to obtain the seasonal values, which in this case will add up to zero.

TABLE 25.5

IMPORTS OF RAW JUTE INTO CALCUTTA & MILL STATIONS (EXCLUDING IMPORTS BY ROAD) : PERCENTAGES OF 12-MONTH MOVING AVERAGES

(1) Year and month	(2) Import of raw jute (000 tons)	(3) 12-month moving total	(4) 2-point moving total of col. (3)	(5) Centred 12-month moving average	(6) Ratio to moving average = 100 × col. (2) col. (5)
1955					
Jan	103·4	—	—	—	—
Feb	105·5	—	—	—	—
Mar	89·5	—	—	—	—
Apr	69·2	—	—	—	—
May	55·6	—	—	—	—
Jun	48·7	—	—	—	—
Jul	42·4	1,038·2	2,103·2	87·63	48·38
Aug	47·0	1,065·0	2,146·2	89·43	53·15
Sep	87·3	1,081·2	2,188·0	91·17	95·76
Oct	10·9	1,106·8	2,229·7	92·90	113·99
Nov	143·9	1,122·9	2,253·4	93·89	153·26
Dec	138·9	1,130·5	2,265·9	94·41	147·12
1956					
Jan	10·2	1,135·4	2,284·1	95·17	136·81
Feb	121·8	1,148·7	2,324·1	96·84	125·78
Mar	115·1	1,175·4	2,342·9	97·62	117·90
Apr	85·3	1,167·5	2,339·1	97·46	87·52
May	63·2	1,171·6	2,337·9	97·41	64·88
Jun	53·6	1,166·3	2,329·4	97·06	55·22
Jul	55·7	1,163·1	2,327·8	96·99	57·43
Aug	74·5	1,164·7	2,308·6	96·19	77·45
Sep	79·4	1,143·9	2,258·5	94·10	84·38
Oct	110·0	1,114·6	2,211·2	92·13	119·39
Nov	148·6	1,108·8	2,205·4	91·89	150·83
Dec	135·7	1,118·9	2,227·7	92·82	146·20
1957					
Jan	131·8	1,123·1	2,242·0	93·42	141·09
Feb	101·0	1,101·7	2,224·8	92·70	108·95
Mar	85·8	1,097·8	2,199·5	91·65	93·62
Apr	67·3	1,097·7	2,195·5	91·48	73·57
May	75·4	1,100·5	2,198·2	91·59	82·32
Jun	63·7	1,102·8	2,203·3	91·80	69·39
Jul	59·9	1,103·2	2,206·0	91·92	65·17
Aug	53·1	1,098·7	2,201·9	91·75	57·88
Sep	75·5	1,098·9	2,197·6	91·57	82·45
Oct	109·9	1,106·0	2,204·9	91·87	119·62
Nov	141·4	1,108·4	2,214·4	92·27	153·25
Dec	138·0	1,099·0	2,107·4	87·81	157·16
1958					
Jan	132·2	1,096·1	2,195·1	91·46	144·54
Feb	96·5	1,106·6	2,202·7	91·78	105·14
Mar	86·0	1,104·1	2,210·7	92·11	93·36
Apr	74·4	1,116·7	2,220·8	92·53	80·40
May	77·8	1,128·0	2,244·7	93·53	83·18
Jun	54·3	1,148·0	2,276·0	94·83	57·26
Jul	57·0	—	—	—	—
Aug	63·6	—	—	—	—
Sep	73·0	—	—	—	—
Oct	122·5	—	—	—	—
Nov	152·7	—	—	—	—
Dec	158·0	—	—	—	—

Ratio-to-moving average method

As explained earlier, periodic fluctuations in a series are eliminated by taking a moving average of period equal to the period of the fluctuations. So from monthly data seasonal fluctuations can be removed by taking a 12-month moving average, which must again be centred by taking a further 2-point moving average. These moving averages will also eliminate some irregular variation and also a small part of the cyclical variation. The moving average values may, therefore, be supposed to give us estimates of the combined effects of trend and cyclical variation.

The ratios of the original values to the moving averages are, therefore, expected to represent the seasonal variation with a part of the irregular fluctuations ($S \times I'$). These ratios, one for each month except for 6 months in the beginning and 6 months at the end, are expressed as percentages. The different values for each month are then averaged so that the irregular fluctuations may be removed. If the variation in the set of values of a month is only due to irregular fluctuations, the values will vary only by small amounts, and the

TABLE 25.6
SHOWING PERCENTAGES OF MOVING AVERAGES AND
SEASONAL INDICES (*vide* TABLE 25.5)

Month Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1955	—	—	—	—	—	—	48·38	53·45	95·76	113·99	153·26	147·12
1956	136·81	125·78	117·90	87·52	64·88	55·22	57·43	77·45	84·38	119·39	150·83	146·20
1957	141·09	108·95	93·62	73·57	82·32	69·39	65·17	57·88	82·45	119·62	153·25	157·16
1958	144·54	105·14	93·36	80·40	83·18	57·26	—	—	—	—	—	—
Average (A.M.)	140·81	113·29	101·63	80·50	76·79	60·62	56·99	62·93	87·53	117·67	152·45	150·16
Adjusted seasonal index	140·6	113·2	101·5	80·4	76·7	60·6	56·9	62·9	87·4	117·5	152·3	150·0

$$\text{Adjustment factor} = \frac{1,200}{1,201·37} = 0·99886.$$

arithmetic mean may be used. If, however, there are some extreme values which are due to incomplete elimination of cyclical effect, one should use the median or modified mean, the modified mean being the arithmetic mean computed after ignoring the extreme values. These averages for the 12 months cannot be used as seasonal indices owing to the incomplete elimination of non-seasonal effects. This fact will be reflected in the total not being equal to 1,200. An adjustment is, therefore, made by multiplying each monthly average by the correction factor : $1,200 / (\text{total of unadjusted monthly averages})$. The scheme of calculations is given in Tables 25.5 and 25.6.

For the additive model, the moving averages are subtracted from the original values, and the deviations for a month are averaged over the years. The monthly (or quarterly) average deviations are finally adjusted so that the total of the seasonal values becomes zero.

Ratio-to-trend method

In this method, we first find an appropriate equation to determine trend values for various months. At the next step, we divide the original data month by month by the corresponding trend values and express them as percentages. The different values for a month are then averaged, as in the previous method. And finally these averages are adjusted to a total of 1,200. It may be noted that in this method we are trying to eliminate the irregular and cyclical variations by averaging. So this method is recommended for use either when cyclical variation is known to be absent or when it is not so pronounced even if present.

For the additive model, the trend values are subtracted from the original values and the other steps are the same as in the moving average method.

Considerable simplification in the calculations may be made by first fitting a trend equation to the yearly totals (or averages) and then obtaining the monthly trend values by a suitable modification of the equation. This is indicated in the following example

Ex. 25.4 The data relate to the revenue expenditure. Government of India, during the years 1953-.. to 1958-59 for the four quarters (Table 25.8).

First, we fit to the yearly totals a quadratic trend, which seems appropriate in this case.

TABLE 25.7
ANNUAL DATA RELATING TO REVENUE EXPENDITURE

Year	Revenue expenditure (lakhs of rupees)
1953-1954	22,543
1954-1955	23,813
1955-1956	26,157
1956-1957	29,251
1957-1958	39,905
1958-1959	51,990

TABLE 25.8
CALCULATION OF TREND-RATIOS

(1) Year & quarter	(2) Revenue expenditure (lakhs of rupees)	(3) Trend value	(4) Trend-ratio = $\frac{(2)}{(3)} \times 100$
1953-54	Apr—Jun	3,575	6,075·35
	Jul—Sep	4,342	73·67
	Oct—Dec	4,435	76·98
	Jan—Mar	10,191	179·49
1954-55	Apr—Jun	3,867	5,642·59
	Jul—Sep	4,404	77·86
	Oct—Dec	5,726	100·13
	Jan—Mar	9,816	168·38
1955-56	Apr—Jun	4,669	5,989·35
	Jul—Sep	5,327	85·95
	Oct—Dec	5,811	90·02
	Jan—Mar	10,350	153·08
1956-57	Apr—Jun	4,693	7,115·63
	Jul—Sep	5,640	75·01
	Oct—Dec	5,957	74·73
	Jan—Mar	12,961	152·99
1957-58	Apr—Jun	5,518	9,021·43
	Jul—Sep	6,887	71·59
	Oct—Dec	7,782	75·80
	Jan—Mar	19,718	179·87
1958-59	Apr—Jun	6,523	11,706·75
	Jul—Sep	9,808	78·46
	Oct—Dec	10,149	76·07
	Jan—Mar	25,510	179·24

Since we have an even number of years, we take a two-quarter period as unit (*vide* Table 25.4) and get the following equation :

$$T_t = 27,728·83 + 2,837·21t + 389·80t^2.$$

Our purpose is to obtain the quarterly trend values. The trend equation for the quarterly averages can be obtained by simply dividing the constants by 4, which thus reduces to

$$T_t = 6,932 \cdot 21 + 709 \cdot 30t + 97 \cdot 45t^2. \quad \dots \quad (25.10)$$

But in the above equations the unit of t is two quarters. Thus the trend equation for quarterly values may be obtained by writing $t/2$ for t in equation (25.10). The trend equation for quarterly values is thus

$$T_t = 6,932 \cdot 21 + 354 \cdot 65t + 24 \cdot 36t^2. \quad \dots \quad (25.11)$$

Again, the origin of the above equations is at the middle of the period covered, i.e. the end of the last quarter of 1955-56. But our trend values should correspond to the mid-points of the quarters. Thus for proper centring of the trend values, the origin must be shifted half a quarter to the right or to the left. If we want to shift the origin half a quarter to the right, i.e. to the middle of the first quarter of 1956-57, we have to write $t + \frac{1}{2}$ for t in equation (25.11). We then get the following equation :

$$\begin{aligned} T_t &= 6,932 \cdot 21 + 354 \cdot 65(t + \frac{1}{2}) + 24 \cdot 36(t + \frac{1}{2})^2 \\ &= 7,115 \cdot 63 + 379 \cdot 01t + 24 \cdot 36t^2. \quad \dots \quad (25.12) \end{aligned}$$

Putting $t=0$ in equation (25.12), we get the trend value for the first quarter of 1956-57. Putting $t=1, 2, 3, \dots$ and $t=-1, -2, -3, \dots$, we may get the trend values for the other quarters as well.

TABLE 23.9
CALCULATION OF SEASONAL INDICES FROM TREND-RATIOS
(*vide* TABLE 25.8)

Year	Quarter			
	Apr—Jun	Jul—Sep	Oct—Dec	Jan—Mar
1953-54	58.84	73.67	76.98	179.49
1954-55	68.53	77.86	100.13	168.38
1955-56	77.96	85.95	90.02	153.08
1956-57	65.95	75.01	74.73	152.99
1957-58	61.17	71.59	75.80	179.87
1958-59	55.72	78.46	76.07	179.24
Average (A.M.)	64.70	77.09	82.29	168.84
Adjusted seasonal index	65.8	78.5	83.8	171.9

$$\text{Adjustment factor} = \frac{400}{392.92} = 1.0180.$$

Method of link relatives

In this method, each monthly value is expressed as a percentage of the previous monthly value. This percentage, called a link relative, estimates approximately the ratio of successive seasonal indices $(100 \cdot \frac{S_t}{S_{t-1}})$. The link relatives for each month are then averaged, as in the previous methods. Taking the seasonal index for a month, say January, to be 100, the others can be obtained from the average link relatives by using the following chain relations :

$$S_{\text{Feb}} = S_{\text{Jan}} \times \frac{S_{\text{Feb}}}{S_{\text{Jan}}},$$

$$S_{\text{Mar}} = S_{\text{Feb}} \times \frac{S_{\text{Mar}}}{S_{\text{Feb}}},$$

$$\vdots$$

$$S_{\text{Dec}} = S_{\text{Nov}} \times \frac{S_{\text{Dec}}}{S_{\text{Nov}}}.$$

S_{Jan} obtained as

$$S_{\text{Jan}} = S_{\text{Dec}} \times \frac{S_{\text{Jan}}}{S_{\text{Dec}}}$$

may not be equal to 100, as assumed, since the other components, mainly the trend, may not be completely eliminated by the process of averaging the link relatives. A correction is, therefore, made by assuming a linear trend and by subtracting b , $2b$, \dots , $11b$ from the February, March, \dots , December indices, respectively, where

$$b = \frac{1}{12} \left(S_{\text{Dec}} \times \frac{S_{\text{Jan}}}{S_{\text{Dec}}} - 100 \right).$$

Finally, the indices are adjusted to a total of 1,200, as in the previous methods.

The method of link relatives was at one time extensively used, but now it is considered unsatisfactory because of its inability to eliminate the other effects efficiently.

The calculation of seasonal indices by the method of link relatives is illustrated in Table 25.10 with the data of Table 25.5.

It must be noted that the above methods are applicable to fixed seasonal patterns only. In case the seasonal pattern changes from year to year, the above methods must be suitably modified.

TABLE 25.10
ARRAYS OF LINK RELATIVES AND CALCULATION OF SEASONAL INDICES FOR THE DATA OF TABLE 25.5

Month Year	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
1955	—	102.128	84.754	77.318	80.347	87.590	87.064	112.736	182.636	121.306	135.883	96.525
1956	93.786	93.518	94.499	74.109	74.091	84.810	103.918	133.752	106.577	138.539	126.000	97.908
1957	97.126	76.631	84.950	78.438	112.036	84.483	94.035	88.648	142.185	145.563	128.662	97.595
1958	95.797	72.995	89.119	86.512	104.570	69.794	104.972	111.571	114.780	167.808	124.643	103.471
Average (A.M.)	95.553	86.326	88.330	79.094	92.761	81.669	97.497	111.677	136.544	143.304	128.800	98.875
Chain relatives	100	86.326	76.252	60.311	55.945	45.690	44.546	49.748	67.928	97.344	125.379	123.968
Trend correction	1.00	84.788	73.176	55.697	49.793	38.001	35.319	38.983	55.625	83.503	110.000	107.051
Adjusted seasonal index	144.2	122.3	105.6	80.3	71.8	54.8	51.0	56.2	80.2	120.5	158.7	154.4

$$\text{Trend correction : } b = \frac{123.964 \times 0.95553 - 100}{12} = 1.5379.$$

$$\text{Adjustment factor} = \frac{1,200}{831.936} = 1.44242.$$

25.6 Changing seasonal patterns

In our discussion in the previous section, we have assumed that the seasonal pattern is fixed, i.e. the seasonal indices for all the months (or some other sections of the year, as the case may be) remain unchanged over all the years under consideration. We have calculated only one set of seasonal indices, applicable for all the years.

Sometimes, however, the above assumption may not be correct. It may be legitimate to assume that the seasonal pattern itself is undergoing change from year to year. The changes may be due to climatic variations, changing tastes and preferences of people or economic factors like progressive measures undertaken by the government. The nature of changes in the seasonal pattern may be different in different situations ; viz. the changes may be slow and gradual showing some trend, or may be sudden or abrupt from one year to the next. Again, the changes may be only in the amplitude or intensity of variation or may be due to occurrence of a festival on different dates of the year (like Easter or Durgapuja) affecting the seasonal indices for two successive months, keeping other indices in tact. In all these cases, we have to calculate sets of seasonal indices appropriate for different years. Special methods have to be adopted in each case.

We shall discuss here the simplest case where the seasonal indices are undergoing change slowly and gradually showing some trend. In this case we adopt the method of moving average or the trend-ratio method. We calculate the ratio to moving averages or ratio to trend expressed in per cent form for all the years and months as in the case of a fixed seasonal pattern. Now we draw graphs, one for each month, plotting the ratios to moving averages (or ratios to trend) for the month against different years and pass through the set of points a free-hand curve (linear or non-linear) which seems to be appropriate. We then read off from the free-hand curves the unadjusted seasonal indices for different months for each year separately. Finally, the unadjusted seasonal indices are adjusted to a total of 1200 for each year separately. Thus we get sets of seasonal indices separately for each year. These seasonal indices are known as *moving seasonal indices*.

25.7 Measurement of cyclical fluctuations

We shall now consider briefly how the cyclical component of a time series is measured. The method we shall discuss is called the *residual method*. It consists in removing from the given time series the other three components, viz. trend, seasonal variation and irregular variation, in any order. According to the multiplicative model, we have

$$y_t = T_t \times S_t \times C_t \times I_t.$$

To get $C_t \times I_t$, it is necessary to remove T_t and S_t by division. This may be done in any of the following three ways, which will lead to the same result :

(i) y_t is first divided by the corresponding trend value T_t , and then by the corresponding seasonal index S_t , which is, of course, to be taken in the fractional and not in the percentage form. (E.g., an index of 89 is to be taken as 0.89 for this calculation.)

(ii) y_t is first divided by S_t and then by T_t .

(iii) The *normal value* $T_t \times S_t$ is first obtained, and y_t is then divided by the normal value.

At the final stage, it is necessary to remove I_t from $C_t \times I_t$ by some process of smoothing—generally, this is done by using moving averages of a suitable period.

A more sophisticated method of determining the cyclical component is the method of *periodogram analysis*. A brief account of the method is given below.

Periodogram analysis

Consider a time series from which trend and seasonal effects have been eliminated. Let u_t ($t=1, 2, \dots, n$) represent the residual series. We want to know whether u_t contains a harmonic term with period μ . Consider the quantities

$$A = \frac{2}{n} \sum_{t=1}^n u_t \cos \frac{2\pi t}{\mu} \quad \dots \quad (25.13)$$

and $B = \frac{2}{n} \sum_{t=1}^n u_t \sin \frac{2\pi t}{\mu}, \quad \dots \quad (25.14)$

where n is the number of terms in the time series. Let us write

$$R_\mu^2 = A^2 + B^2, \quad \dots \quad (25.15)$$

which is known as the *intensity* corresponding to the trial period μ .

Let us consider a simple model, according to which u_t is composed of two components, one periodic with period λ and amplitude a and the other an irregular component, say b_t . Thus

$$u_t = a \sin \frac{2\pi t}{\lambda} + b_t. \quad \dots \quad (25.16)$$

The second component is assumed to be uncorrelated with the first or similar periodic terms.

$$\text{Now, } A = \frac{2a}{n} \sum_i \sin^2 \frac{2\pi i}{\lambda} \cdot \cos \frac{2\pi i}{\mu} + \frac{2}{n} \sum_i b_i \cos \frac{2\pi i}{\mu} = \frac{2a}{n} \sum_i \sin \alpha i \cos \beta i$$

(putting $\alpha = 2\pi/\lambda$, $\beta = 2\pi/\mu$ and neglecting the second term)

$$= \frac{a}{n} \sum_i (\sin(\alpha - \beta)i + \sin(\alpha + \beta)i) \\ = \frac{a}{n} \left\{ \frac{\sin n \frac{(\alpha - \beta)}{2} \sin(n+1) \frac{(\alpha - \beta)}{2}}{\sin \frac{(\alpha - \beta)}{2}} + \frac{\sin n \frac{(\alpha + \beta)}{2} \sin(n+1) \frac{(\alpha + \beta)}{2}}{\sin \frac{(\alpha + \beta)}{2}} \right\},$$

remembering that

$$\sum_{i=0}^{n-1} \sin(\alpha + \beta i) = \frac{\sin \frac{n\beta}{2}}{\sin \frac{\beta}{2}} \sin \left(\alpha + \frac{n-1}{2} \beta \right).$$

For large n , the second term is always small; the first term will also be small unless β tends to α , i.e. unless μ , the trial period, approaches the true period λ . If β tends to α , then

$$A = a \sin(n+1) \frac{(\alpha - \beta)}{2} \cdot \frac{\sin n \frac{(\alpha - \beta)}{2}}{n \frac{(\alpha - \beta)}{2}} \Bigg/ \frac{\sin \frac{(\alpha - \beta)}{2}}{\frac{(\alpha - \beta)}{2}}$$

tends to $a \sin(n+1) \frac{(\alpha - \beta)}{2}$, $\dots \quad (25.17)$

since $\frac{\sin \theta}{\theta} \rightarrow 1$ as $\theta \rightarrow 0$.

Similarly,

$$B \rightarrow a \cos(n+1) \frac{(\alpha - \beta)}{2} \text{ as } \beta \rightarrow \alpha \quad \dots \quad (25.18)$$

and is small otherwise, so that

$$R_\mu^2 \rightarrow a^2 \text{ when } \beta \rightarrow \alpha, \quad \dots \quad (25.19)$$

i.e. when $\mu \rightarrow \lambda$, and is small otherwise.

We now take a number of trial periods μ round about the true period λ , which may be guessed by plotting the data on a graph paper, and calculate R_μ^2 in each case. Finally, we draw a graph plotting R_μ^2 against μ . The diagram, called a *periodogram*, is a simple device for finding the true cyclical period λ in a time series by equating it to that value of μ for which R_μ^2 attains a maximum.

Similarly, if the cyclical component is composed of several periodic terms, say with periods $\lambda_1, \lambda_2, \dots, \lambda_k$, R_μ^2 will remain small unless the trial period μ coincides with one of the true periods, in which case it attains a local maximum with value equal to the square of the amplitude of the periodic term concerned. This is shown in the figure below.

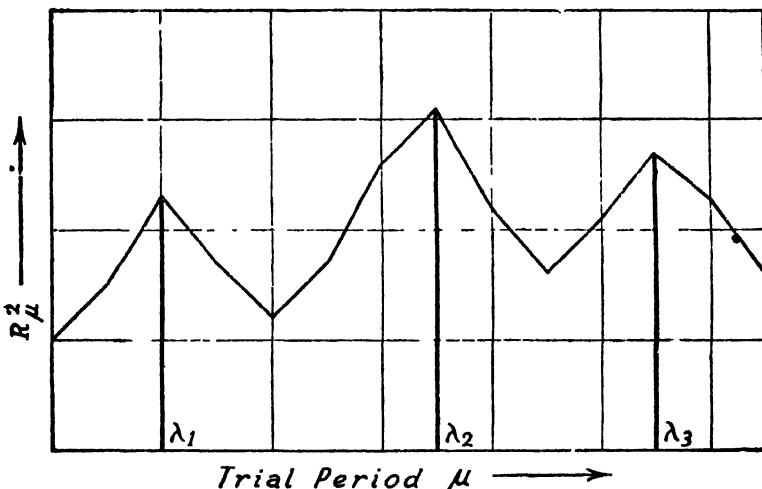


Fig 25.8 A typical periodogram.

25.8 Effect of moving averages on cyclical and random components of a time series

Suppose we have a time series y_t which is the sum of three components, a trend T_t , an oscillatory component C_t and a random component I_t , so that

$$y_t = T_t + C_t + I_t.$$

Here it is assumed that I_t s are such that $E(I_t) = 0$ and $\text{cov}(I_t, I_{t'}) = 0$.

If we determine the trend by a moving average, denoted by the operator M , then

$$M(y_t) = M(T_t) + M(C_t) + M(I_t).$$

Let us suppose that our method of trend determination is perfect so that

$$M(T_t) = T_t.$$

Thus

$$y_t - M(y_t) = C_t - M(C_t) + I_t - M(I_t).$$

We shall see that $M(C_t)$ and $M(I_t)$ are not necessarily zero, so that the moving average may distort the genuine oscillatory part of the residual series and introduce spurious oscillatory movements.

Consider the simple case when C_t is a sine term with periodicity λ and we take a simple moving average of period k . Thus

$$C_t = a \sin \frac{2\pi t}{\lambda},$$

so that

$$\begin{aligned} M(C_{(k+1)t}) &= a \cdot \frac{1}{k} \sum_{i=1}^k \sin \frac{2\pi i}{\lambda} \\ &= a \cdot \frac{1}{k} \cdot \frac{\sin \frac{k\pi}{\lambda}}{\frac{\sin \frac{\pi}{\lambda}}{\lambda}} \sin 2\pi \frac{(k+1)}{2\lambda}. \quad \dots \quad (25.20) \end{aligned}$$

Thus a simple k -period moving average will result in a sine series of the same period, but with amplitude reduced by the factor

$$\frac{1}{k} \cdot \frac{\sin \frac{k\pi}{\lambda}}{\frac{\sin \frac{\pi}{\lambda}}{\lambda}}.$$

The following special cases may be considered :

(1) If k is equal to or is a multiple of λ , then $\sin \frac{k\pi}{\lambda}$ is equal to zero, so that the cyclical component is completely eliminated by the moving average.

(2) We have

$$\frac{1}{k} \cdot \frac{\sin \frac{k\pi}{\lambda}}{\frac{\sin \frac{\pi}{\lambda}}{\lambda}} \rightarrow 0 \text{ as } k \rightarrow \infty;$$

so that if k is large compared to λ , then also the cyclical component is greatly eliminated.

(3) It is seen that

$$\frac{1}{k} \cdot \frac{\sin \frac{k\pi}{\lambda}}{\sin \frac{\pi}{\lambda}} = \frac{\sin \frac{k\pi}{\lambda} / \frac{k\pi}{\lambda}}{\sin \frac{\pi}{\lambda} / \frac{\pi}{\lambda}} \rightarrow 1 \text{ as } \frac{k\pi}{\lambda} \rightarrow 0.$$

Hence if k is small compared to λ , the moving average fails to eliminate the cyclical component.

As such, in the residual series we shall find that larger oscillations have almost disappeared, whereas only shorter oscillations will be found to reappear. Thus the process of moving average, in general, distorts the genuine oscillatory component of the time series, emphasising the shorter oscillations at the expense of the longer ones.

For the random element I_t , we have

$$M(I_t) = \frac{1}{k} \sum_{j=-[k/2]}^{[k/2]} I_{t+j}, \quad \dots \quad (25.21)$$

where $[k/2]$ is the greatest integer contained in $k/2$. Naturally, consecutive values of $M(I_t)$ will not be uncorrelated, since $M(I_a)$ and $M(I_b)$ have $k - (a - b)$ values of I_t in common and $M(I_a)$ and $M(I_b)$ will be correlated if $k > (a - b)$. Hence $M(I_t)$ will be a much smoother series than the original series. Thus the effect of taking a moving average of the random component would be to generate a spurious oscillatory series, provided the correlation between the successive members of the generated series is positive. This effect is generally known as the "Slutsky-Yule effect".

25.9 Different schemes which account for oscillations in a stationary time series

Time series may be broadly classified into two categories, viz. evolutive and stationary. In the former, different sections of the time series are dissimilar in one or more respects. A stationary time series may be divided into a number of sections which are unchanging in respect of their general structure. The oscillations in such a series may seem random or show tendencies of regularity, but in any case the series is on the whole the same in different sections.

Three different schemes or models may be considered which may account for oscillatory movements in a stationary time series :

(a) Effect of moving averages on the random component—We have seen that a moving average of a purely random series generates an oscillatory series with varying periods and amplitudes. It is quite possible that some of the observed oscillations in a time series are generated this way.

(b) Sum of a number of cyclical components—This is the classical approach. Here we attempt by periodogram analysis and harmonic analysis to represent an oscillatory series as the sum of a number of harmonic terms with varying periods and intensities. Thus, if u_t be the oscillatory series, and $\lambda_1, \lambda_2, \dots$ the different periods, then we have

$$\begin{aligned} u_t = & a_0 + a_1 \cos \frac{2\pi t}{\lambda_1} + a_2 \cos \frac{2\pi t}{\lambda_2} + \dots \\ & + b_1 \sin \frac{2\pi t}{\lambda_1} + b_2 \sin \frac{2\pi t}{\lambda_2} + \dots \quad \dots \quad (25.22) \end{aligned}$$

(c) Autoregression equations—If a series is such that the value corresponding to the time point $t+1$ depends on the previous $k+1$ values according to the relation

$$u_{t+1} = f(u_t, u_{t-1}, \dots, u_{t-k}) + I_{t+1}, \quad \dots \quad (25.23)$$

where f is a mathematical function and I a random variable, then the series is called autoregressive. In this case, under certain conditions the generated series is of the oscillatory type. The linear autoregression equations of the first and second orders are special cases of (25.23) and are of the form

$$(1) \quad u_{t+1} = \mu u_t + I_{t+1} \quad \dots \quad (25.24)$$

$$\text{and} \quad (2) \quad u_{t+1} = a u_t + b u_{t-1} + I_{t+1}, \quad \dots \quad (25.25)$$

respectively.

25.10 Serial correlation and correlogram

An observed series showing typical oscillatory movements may be due to any of the above schemes. We require some objective criterion for deciding which of them is applicable in particular cases. This criterion is provided by the so-called *correlogram*.

First, we define what are known as *serial correlations* or *autocorrelations* of different orders. A serial correlation (r_k) of order k is the correlation between u_t and u_{t+k} . From the original u_t series ($t=1, 2, \dots, n$) $n-k$ pairs of values are obtained with a lag of period k .

Thus

$$\begin{aligned}
 r_k &= \frac{\text{cov}(u_t, u_{t+k})}{\{\text{var}(u_t) \cdot \text{var}(u_{t+k})\}^{1/2}} \\
 &= \frac{\frac{1}{n-k} \sum_{t=1}^{n-k} u_t \cdot u_{t+k} - \frac{1}{(n-k)^2} \sum_{t=1}^{n-k} u_t \sum_{t=1}^{n-k} u_{t+k}}{\left\{ \frac{1}{n-k} \sum_{t=1}^{n-k} u_t^2 - \frac{1}{(n-k)^2} (\sum_{t=1}^{n-k} u_t)^2 \right\}^{1/2} \left\{ \frac{1}{n-k} \sum_{t=1}^{n-k} u_{t+k}^2 - \frac{1}{(n-k)^2} (\sum_{t=1}^{n-k} u_{t+k})^2 \right\}^{1/2}} \\
 &\dots \quad (25.26)
 \end{aligned}$$

Obviously, we have

$$r_0 = 1 \text{ and } r_{-k} = r_k.$$

The diagram obtained by plotting r_k against k on a graph paper and joining the points, each to the next, is called a correlogram.

It can be shown theoretically that the correlogram takes widely differing shapes under different schemes. In scheme (a), when oscillatory movement is generated by an m -point simple moving average of a random component I_t , where $E(I_t) = 0$, $\text{cov}(I_t, I_{t'}) = 0$ and $\text{var}(I_t) = \sigma^2$, it can be shown that

$$\rho_k = \begin{cases} 1 - \frac{k}{m} & \text{for } k \leq m \\ 0 & \text{for } k > m, \end{cases} \quad \dots \quad (25.27)$$

and

ρ_k being the theoretical value of the serial correlation of order k .

Thus the correlogram would be a straight line starting at $(0, 1)$ and ending at $(m, 0)$ and thereafter the correlogram would coincide

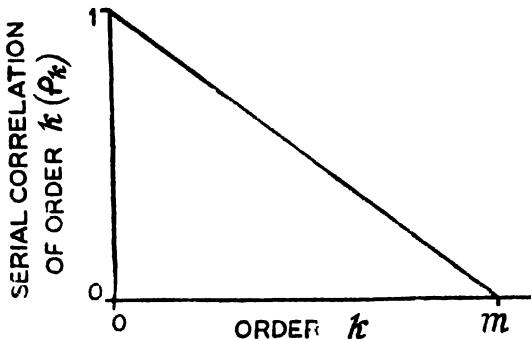


Fig. 25.9 Correlogram for oscillatory series generated by simple moving average of random component.

with the k -axis (Fig. 25.9). If however, the oscillations were generated

by an m -point weighted moving average, the correlogram would oscillate between the points $(0, 1)$ and $(m, 0)$ and thereafter would coincide with the k -axis.

In scheme (b), where the oscillatory movement is generated by the sum of a number of cyclical components represented by the sum of a number of harmonic terms with periods $\lambda_1, \lambda_2, \dots$, it can be shown that ρ_k would also be the sum of a number of harmonic terms, not necessarily with the same periods. In particular, if

$$u_t = a \sin \frac{2\pi t}{\lambda} + I_t, \quad \dots \quad (25.28)$$

ρ_k would be equal to $a \cos \frac{2\pi k}{\lambda}$ for $k > 0$, so that the correlogram would be a strictly periodic sinusoidal curve (Fig. 25.10). In any case, in scheme (b), the correlogram will take a sinusoidal form, which will not degenerate to the k -axis after some fixed point and will not be damped.

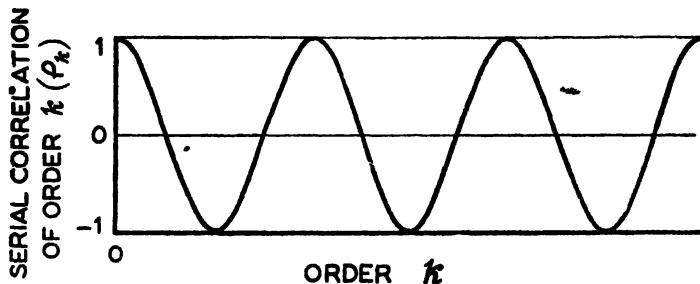


Fig. 25.10 Correlogram for oscillatory series generated by a cyclical term.

In scheme (c), where the oscillations are caused by autoregression, let us consider autoregressive equations of the first and second orders. For the equation of the first order, viz. $u_{t+1} = \mu u_t + I_{t+1}$, it can be shown that

$$\rho_k = \mu^k. \quad \dots \quad (25.29)$$

Hence the correlogram would take an exponential form. Since μ must be less than 1, so that the time series does not explode to

infinity, the curve would start at $(0, 1)$ and thereafter fall rapidly and tend to the k -axis asymptotically (Fig. 25.11).

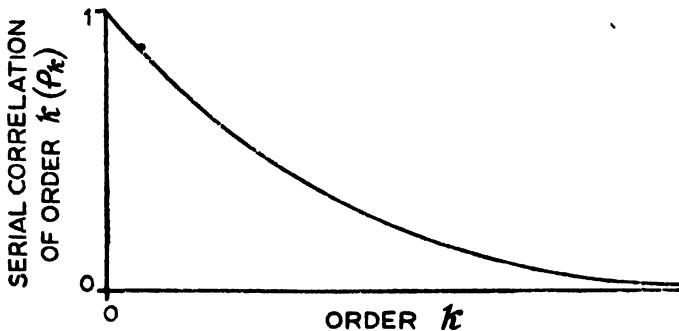


Fig 25.11 Correlogram for oscillatory series generated by an autoregressive scheme of first order.

For the equation of the second order, viz.

$$u_{t+1} = au_t + bu_{t-1} + I_{t+1},$$

the formula for ρ_k depends upon the nature of the roots of the quadratic

$$q^2 - aq - b = 0.$$

If the roots are real, say q_1 and q_2 ($q_1, q_2 < 1$ for practical purposes, otherwise the series would explode to infinity),

$$\rho_k = \frac{q_1^{-k}(1-q_2^2)q_1}{(q_1-q_2)(1+q_1q_2)} + \frac{q_2^{-k}(1-q_1^2)q_2}{(q_2-q_1)(1+q_1q_2)} \quad \dots \quad (25.30)$$

Here also, the correlogram starts at $(0, 1)$ and becomes asymptotic to the k -axis.

If the roots are imaginary, say $q_1 = p e^{i\theta}$ and $q_2 = p e^{-i\theta}$.

$$\rho_k = p^k \frac{\sin(k\theta + \psi)}{\sin \psi}, \quad \dots \quad (25.31)$$

where $\frac{1+p^2}{1-p^2} \tan \theta = \tan \psi$.

Hence the correlogram will be oscillatory in this case, but unlike in scheme (b), the oscillations will be damped (Fig. 25.12) owing to the presence of p^k ($p < 1$ for practical purposes).

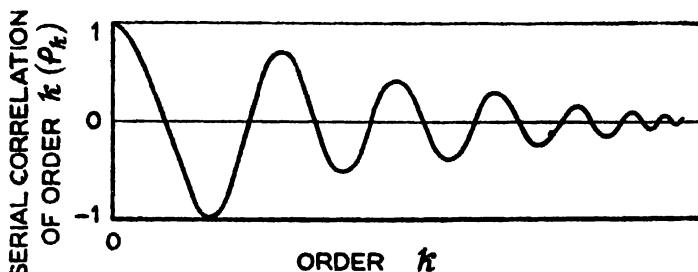


Fig. 25.12 Correlogram for oscillatory series generated by an autoregressive scheme of order two (Case 2).

Thus we see that the correlogram takes widely differing shapes under different schemes. Hence the correlogram provides a very useful criterion for discriminating between different schemes which can account for oscillatory movements in a time series.

25.11 Correlation between two time series : lag correlation

Correlation between two time series y_t and x_t may sometimes lead to misleading results, since both the series may have regular variations with respect to time and may show a high correlation although the two series do not have any causal relationship between each other. For example, the production of steel in India may show a high negative correlation with death-rates in India, since the two series are expected to have trends of opposite signs. This is called spurious (or nonsense) correlation. Similarly, owing to the effect of time on both x and y , a real correlation between x and y may be obliterated.

One can think of four possible situations :

- Actually there is no correlation, but owing to similar types of trends (i.e. both increasing or both decreasing) one may get a spuriously high positive correlation.
- Actually there is a negative correlation, but owing to similar types of trends correlation will be decreased or even one may get a small positive correlation.
- Actually there is no correlation, but owing to different types of trends (i.e. one increasing, but the other decreasing), one may get a spuriously high negative correlation.
- Actually there is a positive correlation, but owing to different

types of trends, the correlation will be decreased or even one may get a small positive correlation.

To remove this difficulty one may adopt any of the following procedures :

(1) One may calculate a partial correlation between x and y eliminating the effect of time on both.

(2) Before correlating x and y , the effect of time on both x and y may be eliminated either by taking trend-ratios or by taking link-relatives. That is, one may correlate $\frac{x_t}{T_x}$ and $\frac{y_{t-1}}{T_{y_{t-1}}}$, where

T_x and T_y denote the corresponding trends, or one may correlate $\frac{x_t}{x_{t-1}}$ and $\frac{y_{t-1}}{y_{t-2}}$.

Sometimes, the value of a variable x at time t may affect the value of another variable y at a later period, say at time $t+k$. For example, the production of raw-cotton in a certain year may affect the production of textiles in the next year. Here one has to calculate a lag-correlation of one year lag. In general, a lag-correlation with a lag of k periods or a lag-correlation of order k is the correlation between x_t and y_{t+k} .

Questions and exercises

25.1 Describe the different components of a time series. What purpose is served by analysing a time series ?

25.2 Discuss the different methods of determining trend in a time series. What are their relative merits and demerits ?

25.3 Discuss the different methods of obtaining measures of seasonal variation. Discuss their relative merits and demerits.

25.4 What is a periodogram ? Describe the method of periodogram analysis for determining the hidden periodicities in a time series.

25.5 Criticise the use of moving averages for determining trend. Establish the effect of eliminating trend by the method of moving averages on the other components of a time series.

25.6 Describe the different schemes for explaining the oscillations in a stationary time series. Explain the use of correlograms for discriminating between the above schemes.

25.7 Explain why the correlation between two time series sometimes leads to nonsensical results and state how you would tackle the problem.

25.8 Obtain the trend values for the following series by fitting a second-degree polynomial. Represent the trend values and the original data in a suitable diagram.

GOODS CARRIED BY INDIAN RAILWAYS DURING 1959-67

Year	Goods carried (000 metric tons)
1959-60	147,864
1960-61	157,640
1961-62	161,855
1962-63	180,090
1963-64	192,262
1964-65	195,062
1965-66	204,150
1966-67	202,697

25.9 The following table gives the yield-rate of rice in West Bengal for a number of years. Determine the trend values by means of moving averages of an appropriate period.

Year	Yield of rice (kg. per hectare)	Year	Yield of rice (kg. per hectare)
1951-52	920	1957-58	991
1952-53	971	1958-59	967
1953-54	1,243	1959-60	960
1954-55	959	1960-61	1,184
1955-56	1,025	1961-62	1,085
1956-57	1,082		

25.10 From the following table showing the monthly receipts of State Governments in India, obtain measures of seasonal variation.

TOTAL RECEIPTS OF STATE GOVERNMENTS IN INDIA (RS. CRORES)

Month Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1952	23	39	82	17	18	16	20	17	12	22	20	18
1953	25	26	105	20	22	20	26	18	23	29	15	16
1954	32	36	93	21	21	22	29	21	15	27	27	21
1955	32	42	99	24	24	23	29	24	21	32	28	21

25.11 The seasonal indices of the sales of garments of a particular type in a certain shop are given below :

Quarter	Seasonal index
Jan—Mar	97
Apr—Jun	85
Jul—Sep	83
Oct—Dec	135

If the total sales in the first quarter of a year be worth Rs. 15,000, determine how much worth of garments of this type should be kept in stock by the shop-owner to meet the demand for each of the other three quarters of the year.

Ans. Rs. 13,144 ; Rs. 12,835 ; Rs. 20,876.

SUGGESTED READING

- [1] Croxton, F. E. and Cowden, D. J. *Applied General Statistics* (Chs. 11—14, 16). Prentice-Hall, 1967, and Prentice-Hall of India, 1969.
- [2] Dubois, E. N. *Essential Methods in Business Statistics* (Ch. 13). McGraw-Hill, 1964.
- [3] Greenwald, W. I. *Statistics for Economics* (Chs. 7—10). C. E. Merrill Books, 1963.
- [4] Kendall, M. G. and Stuart, A. *The Advanced Theory of Statistics*, Vol. 3. (Chs. 45—47). Charles Griffin, 1966.
- [5] Lange, O. *Introduction to Econometrics* (Ch. 1). Pergamon Press, 1959.
- [6] Mills, F. C. *Statistical Methods* (Chs. 10—12). H. Holt, 1955.

26

DEMAND ANALYSIS

26.1 Introduction

By *demand* of a commodity we mean its absorption-capacity for the market or the quantity of the commodity that can be sold on the market. By *supply* of a commodity is meant its output or the quantity of the commodity which sellers supply to the market. One of the problems of demand analysis is to study the relationship between market price and demand on the basis of market data (also called time-series data). Another mode of study is to determine how demand varies with change in income on the basis of family-budget data (also called cross-section data).

26.2 Demand and supply curves

The traditional law of supply and demand states that demand, in general, varies inversely as price whereas supply, in general, varies in

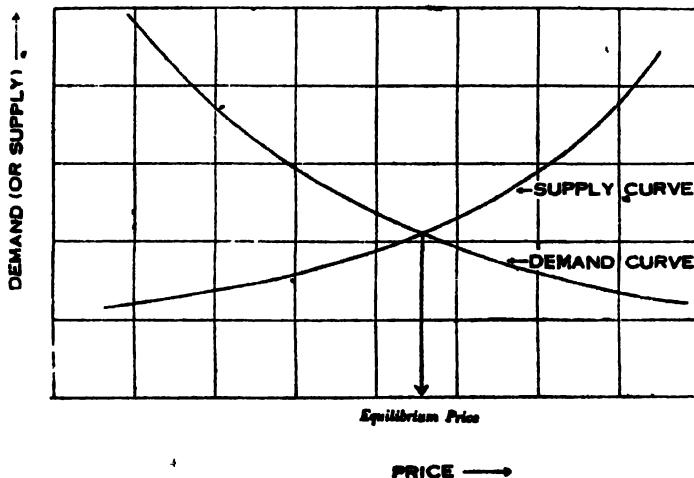


Fig. 26.1 Demand and supply curves.

the same direction as price. A mathematical formulation of the law of supply and demand was first given by A. A. Cournot. According

to Cournot, the demand D is a certain function of the price p , say $D=f(p)$, where $f(p)$ is a diminishing function, so that $f'(p) < 0$. On the other hand, the supply S is an increasing function of price, say $S=\phi(p)$, where $\phi'(p) > 0$. Of course, Marshall in his study took price as the dependent variable and D or s as the independent variable, and these mathematical laws of demand and supply are sometimes referred to as Cournot-Marshall laws. The curves for $D=f(p)$ and $S=\phi(p)$ may be called the demand and supply curves, respectively.

The law of supply and demand further states that market price forms on a level at which demand and supply are equal and, therefore, is determined by the point at which the demand and supply curves intersect. The price determined by the point of intersection is sometimes called the *equilibrium price* (Fig. 26.1).

26.3 Price-elasticity of demand and supply

Casual observations will show that the demand for some commodities is much more sensitive to price changes than is the demand for other goods. For some staple consumer's goods, demand will decrease only slightly when price increases, whereas for some luxury goods, only a slight increase in price may decrease its demand considerably. Thus commodities may be classified on the basis of their sensitivity to price changes. A measure of this sensitivity is provided by the price elasticity of demand. It is defined as the ratio of the relative change in demand to the relative change in price. If $D=f(p)$ be the demand function, the price elasticity of demand (η_p) is given by

$$\eta_p = - \frac{dD}{dp} = - \frac{p}{f(p)} \cdot \frac{df}{dp} = - \frac{d \log f}{d \log p}. \quad \dots \quad (26.1)$$

The negative sign is provided since the changes in demand and price are in opposite directions. If $\eta_p < 1$, the commodity is said to be elastic, while $\eta_p < 1$ means that the commodity is inelastic. When $\eta_p = 1$, the commodity is neither elastic nor inelastic ; it is then called an item of unitary elasticity. In general, luxury goods are elastic, while necessary goods are inelastic.

Price elasticity of supply can similarly be defined. The above definition relates to point elasticity of demand, and hence in

measuring it the definite point at which the elasticity is to be measured must be mentioned. If the elasticity is to be a constant, say, η_0 at all points of the demand curve, then

$$\frac{df}{dp} = -\frac{f}{p}\eta_0,$$

or
$$\frac{df}{f} = -\frac{dp}{p}\eta_0,$$

or
$$\log f = -\eta_0 \log p + \log c,$$

where $\log c$ is the constant of integration,

or
$$f = cp^{-\eta_0}. \quad \dots \quad (26.2)$$

Thus, the demand curve is then a simple hyperbolic curve.

26.4 Determination of demand curves from market data

It has already been stated that the market price at a particular point of time is the equilibrium price for which the demand is equal to the supply and is determined by the point of intersection of the demand curve and the supply curve. The market data, which are essentially in the nature of a time series, give the price of the commodity and the quantity sold at that price at different points of time. That the price of the commodity changes over time implies that either or both of the demand and the supply curves shift their positions. If both the curves remain fixed, the statistical data will not provide a sufficient number of points on the curves for their determination. If both the curves shift their positions, the market data would determine a curve which would give us a picture of the variations of the equilibrium price and the corresponding values of the demand (or supply) curve. If, however, the demand (supply) curve remains fixed and the supply (demand) curve shifts its position, the market data provide a number of points on the fixed demand (supply) curve and hence determine this curve. Thus, for the determination of the demand curve it has to be assumed that the demand curve remains relatively fixed and the supply curve shifts over the period under consideration. The assumption is more or less legitimate for staple consumer's goods, especially food articles. In many cases, where both demand and supply are variable, the market

data are not likely to trace out either the supply or the demand function closely. It may trace out a 'mongrel' function, which is a linear combination of both demand and supply functions. In any particular situation the econometrician has no way of distinguishing between a 'mongrel' result and the true demand curve.

Another difficulty that arises in the determination of the demand curve from time-series data is that other factors, besides the price of the commodity upon which the demand depends, also vary with time. The prices of related commodities, the national income, etc., are such factors. Thus, to determine the demand curve either such factors have to be taken explicitly or the effects of such factors upon the demand and the price have to be eliminated.

Further, in determining the demand curve any of the two variables, price and demand, may be taken as the independent variable, but it is to be noted that both the variables are subject to errors. Hence the ordinary least-square method for the determination of parameters involved in the demand function is not strictly valid.

The classical theory of consumer behaviour tells us a few things of interest regarding the form of the demand function to be used in practice. It starts with a utility function of unknown form, where utility (u) is expressed as a function of the quantities of goods (x_i) in the consumer's budget :

$$u = f(x_1, x_2, \dots, x_n). \quad \dots \quad (26.3)$$

To maximise u subject to the budget restriction

$$\sum_{i=1}^n p_i x_i = y, \quad \dots \quad (26.4)$$

where y is the income, we shall have $n-1$ equations of the form

$$\frac{\partial u}{\partial x_i} = p_i \\ \frac{\partial u}{\partial x_j} = p_j$$

or $\frac{\text{marginal utility of the } i\text{th good}}{\text{marginal utility of the } j\text{th good}} = \frac{p_i}{p_j}. \quad \dots \quad (26.5)$

These $n-1$ equations, together with the budget equation (26.4), can ordinarily be solved for x_i in terms of ratios of prices and of income as a ratio to price. These are homogeneous functions of degree zero

of the prices and income. Thus,

$$x_i = g_i \left(\frac{p_1}{p_i}, \frac{p_2}{p_i}, \dots, \frac{p_{i-1}}{p_i}, \frac{p_{i+1}}{p_i}, \dots, \frac{p_n}{p_i}, \frac{y}{p_i} \right) \dots \quad (26.6)$$

or $x_i = h_i \left(\frac{p_1}{p}, \frac{p_2}{p}, \dots, \frac{p_n}{p}, \frac{y}{p} \right), \dots \quad (26.6a)$

where p is a weighted arithmetic mean of all the prices or a measure of the general price-level.

Thus, y/p is the conventional measure of real income, obtained by deflating the normal income by the average of all prices of consumer goods.

Economic theory gives no suggestion as to whether the form of the functions in (26.6) and (26.6a) is linear or non-linear. Keeping the homogeneity restriction in mind, we can write the linear demand function as

$$x_{ii} = \alpha_{0i} + \alpha_{1i} \frac{p_{1i}}{p_{ii}} + \alpha_{2i} \frac{p_{2i}}{p_{ii}} + \dots + \alpha_{ni} \frac{p_{ni}}{p_{ii}} + \beta_{1i} \frac{y_i}{p_{ii}} \dots \quad (26.7)$$

or, alternatively, as

$$x_{ii} = \alpha_{0'i} + \alpha_{1'i} \frac{p_{1i}}{p_i} + \alpha_{2'i} \frac{p_{2i}}{p_i} + \dots + \alpha_{ni'} \frac{p_{ni}}{p_i} + \beta_{1'i} \frac{y_i}{p_i} \dots \quad (26.7a)$$

Linearity, it must be remembered, is a convenience and at times is accepted even against the reality. However, it is possible to introduce non-linearity in the variables in order to achieve a higher degree of realism. For example, introducing a second degree term of real income, we have

$$x_{ii} = \alpha_{0'i} + \alpha_{1'i} \frac{p_{1i}}{p_i} + \alpha_{2'i} \frac{p_{2i}}{p_i} + \dots + \alpha_{ni'} \frac{p_{ni}}{p_i} + \beta_{1'i} \frac{y_i}{p_i} + \beta_{2'i} \left(\frac{y_i}{p_i} \right)^2 \dots \quad (26.8)$$

Again, introducing demand functions of the constant elasticity type, we may write

$$x_{ii} = A_i \left(\frac{p_{1i}}{p_i} \right)^{\alpha_{1i}} \left(\frac{p_{2i}}{p_i} \right)^{\alpha_{2i}} \dots \left(\frac{p_{ni}}{p_i} \right)^{\alpha_{ni}} \left(\frac{y_i}{p_i} \right)^{\beta_i}$$

or $\log x_{ii} = \log A_i + \alpha_{1i} \log \left(\frac{p_{1i}}{p_i} \right) + \alpha_{2i} \log \left(\frac{p_{2i}}{p_i} \right) + \dots + \alpha_{ni} \log \left(\frac{p_{ni}}{p_i} \right) + \beta_i \log \left(\frac{y_i}{p_i} \right), \dots \quad (26.9)$

which is again a linear function in the logarithms of the variables.

The parameters of the function in each of the above cases can be determined by the method of least-squares. The least-square estimates will also be the maximum-likelihood estimates provided the errors are independently normally distributed.

Ex. 26.1 Index numbers of demand for agricultural products (y) and of prices of agricultural products (x) are given below for the years 1950-59. Obtain the price-elasticity of demand, assuming the following form of demand function :

$$Y = \alpha x^\beta.$$

Year	y	x
1950	102	89
1951	98	99
1952	100	100
1953	105	91
1954	117	93
1955	120	72
1956	120	75
1957	127	91
1958	118	91
1959	134	96

Here we are to fit a demand curve of the form

$$Y_i = \alpha x_i^\beta$$

or $\log Y_i = \log \alpha + \beta \log x_i.$

The constants α and β are to be estimated by the method of least-squares, i.e. by minimising

$$\sum_i (\log y_i - \log Y_i)^2.$$

The normal equations are :

$$\sum_i \log y_i = n \log \alpha + \beta \sum_i \log x_i.$$

and $\sum_i (\log y_i)(\log x_i) = \log \alpha \sum_i \log x_i + \beta \sum_i (\log x_i)^2,$

n being the number of years for which the data are tabulated here.

For these data, we have

$$\sum_t \log y_t = 20.5504,$$

$$\sum_t \log x_t = 19.5052,$$

$$\sum_t (\log y_t)(\log x_t) = 40.0771$$

and

$$\sum_t (\log x_t)^2 = 38.0657.$$

Substituting the above values in the normal equations and solving them, we get

$$\log \alpha = 2.07521 \text{ and } \beta = -0.33333,$$

so that

$$\alpha = 507.24.$$

Taking the constants up to three significant figures, the demand curve is

$$Y_t = 507x_t^{-0.333}.$$

The appropriate measure for the relative change in consumption to relative change in price is the price-elasticity of demand, viz.

$$\eta_p = -\frac{d \log Y}{d \log x}.$$

For the fitted curve,

$$\eta_p = -\beta = 0.333$$

Thus the demand of agricultural products, as shown by the data, is inelastic.

26.5 Engel's law and the Engel curve

We now proceed to discuss briefly the determination of demand curves on the basis of family-budget data. Here we base our estimates on a different type of variation—namely, over space, instead of over time, the variation arising out of inter-individual differences at a given point of time. A sample of such variations is called a cross-section sample. Here a cross-section sample will be a sample of family budgets showing expenditures on the main items of family consumption, together with information on family income, family consumption and other demographic, social and economic characteristics. The basic relationship to be derived here is between the expenditure on a particular item of consumption and the household income. The relationship is generally known as the *Engel curve*.

after Ernst Engel, who was the first to make a systematic study of family budgets. In the course of his studies, he observed that the proportion of expenditure on food decreases as the household income increases. This finding, repeatedly confirmed in later investigations, has become known as *Engel's law*.

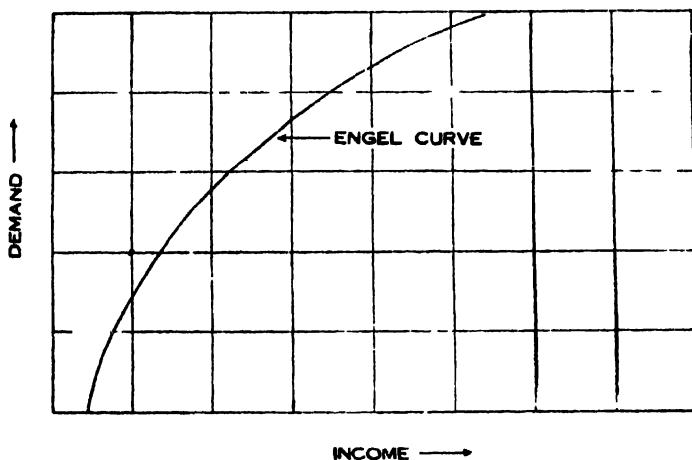


Fig. 26.2 An Engel curve.

- In the analysis of cross-section data, we assume that different people (families) are homogeneous, except for the measurable variables under study. The theory of demand would suggest that individual expenditure depends on the whole set of prices of goods in the budget and on income. Since in cross-section data the price variables, and indeed many other market variables, may be effectively held constant, we may consider income alone as the explanatory variable.

In practice, however, in determining the Engel curve, the total household expenditure is taken as the determining variable in many investigations, instead of the household income. It is contended that, compared to household expenditure, household income may be difficult to obtain and may be a poor indicator of the household standard of living. Besides, the total household expenditure may depend, in a complicated way, on the income expectations and the distribution of expenditure on various items may depend only on the total expenditure.

26.6 Income-elasticity of demand

We may define elasticity of demand with respect to income or its substitute, total expenditure. If e_i denotes the expenditure on the i th item and e_0 the total expenditure, then the elasticity of demand (η_s) with respect to total expenditure is defined as

$$\eta_s = \frac{\text{Relative change in the expenditure on the } i\text{th item}}{\text{Relative change in total expenditure}}$$

$$= \frac{\frac{\Delta e_i}{e_i}}{\frac{\Delta e_0}{e_0}} = \frac{e_0}{e_i} \cdot \frac{\Delta e_i}{\Delta e_0}. \quad \dots \quad (26.10)$$

If the Engel curve giving the relationship between e_i and e_0 is represented by the equation $e_i = f_i(e_0)$, where $\frac{df_i}{de_0} > 0$, then we have in the limit, as $\Delta e_0 \rightarrow 0$,

$$\eta_s = \frac{e_0}{f_i(e_0)} \cdot \frac{df_i}{de_0} = \frac{d \log f_i}{d \log e_0}, \quad \dots \quad (26.11)$$

since e_i varies in the same direction as e_0 .

As with price-elasticities of demand, items may also be classified as elastic or inelastic with respect to income-elasticities. Obviously, for staple articles of food η_s would be less than 1, whereas for luxury goods η_s would be greater than 1.

26.7 Different forms of the Engel curve

The two widely used forms are the straight line (on the arithmetic scale)

$$e_i = \alpha + \beta e_0, \quad \dots \quad (26.12)$$

used by Allen and Bowley, and the straight line on the doubly-logarithmic scale

$$\log e_i = \alpha + \beta \log e_0, \quad \dots \quad (26.13)$$

used by Stone and others.

While the former has the advantage of simplicity, the latter has the advantage that it provides a constant elasticity at all points on the curve.

The hyperbola

$$e_i = \alpha + \beta/e_0 \quad \dots \quad (26.14)$$

is also used. The curve has an initial income ($-\beta/\alpha$) below which

the item is not purchased and a saturation level α . The curve

$$\log e_i = \alpha + \beta/e_0 \quad \dots \quad (26.15)$$

is of a sigmoid shape, passing through the origin and having an asymptote.

Whichever form is used in a particular case, the determination of the parameters may be made simply by the method of least squares, the data being obtained from a family-budget survey covering a number of households with various expenditure levels, the tacit assumption being that the group of families is more or less homogeneous having identical structure of needs. In practice, the aggregate of households has to be determined for each stratum. Since the data are obtained in a short period of time, the other factors like price may be supposed to have remained constant over that period.

26.8 Variation in household size and composition

It is apparent that household standard of living depends upon the household size and its age- and sex-composition, besides the household income. These factors have to be considered explicitly, since it has been found erroneous to assume that the effects of these factors could be ignored by considering households of different sizes and composition.

First, let us consider the factor of household size. Here the working hypothesis is that the expenditure on an item per person depends upon the level of income (or total expenditure) per person. This will require that the function relating expenditure on the i th item (e_i) to the total expenditure (e_0) and family size (n), viz.

$$e_i = f_i(e_0, n) \quad \dots \quad (26.16)$$

is a homogeneous function of degree 1. It follows that

$$f_i(c e_0, cn) = c f_i(e_0, n), \quad \dots \quad (26.17)$$

where c is an arbitrary constant.

This will be apparent if (26.16) is written in the form

$$e_i/n = f_i(e_0/n). \quad \dots \quad (26.18)$$

In particular, for the hyperbolic form of the Engel curve,

$$e_i/n = \alpha + \beta n/e_0$$

or

$$e_i = n\alpha + \beta n^2/e_0, \dots \quad (26.19)$$

so that the initial income is $-n\beta/\alpha$ and the saturation level is $n\alpha$.

Besides the household size, the household composition with respect to age, sex, occupation, etc., may also affect the household consumption. Here it will be necessary to scale individuals of different age-groups, sexes or occupations with respect to the consumption of the item concerned. Thus, here n would mean the number of consumer units in a household instead of the number of members in the household. Naturally, a consumer unit in the consumption scale has to be properly defined.

Ex. 26.2 The table below gives the family-budget data for a few samples of four low-income classes of families of a country for a year :

	Yearly Income per Consumer Unit in Rs.			
	Below 600	600—	750—	1,050—
Number of households in the sample	136	179	111	22
Average number of consumer units per household	2.60	2.57	2.50	2.48
Average income per consumer unit	543.1	681.3	861.9	1,232.0
Average expenditure on food per consumer unit	291.8	331.6	374.4	407.1

Calculate the income-elasticity of demand for food, assuming that the demand function has constant elasticity.

The demand function here is

$$\log e_f = a + b \log e_0,$$

where e_f is the expenditure per consumer unit on food and e_0 is the income per consumer unit. The constants a and b are estimated by the method of least-squares, viz. by minimising

$$\sum_{i=1}^4 n_i (\log e_{fi} - a - b \log e_{0i})^2,$$

where n_i is the number of households at the i th income level. The

normal equations are

$$\sum_i n_i \log e_{f,i} = a \sum_i n_i + b \sum_i n_i \log e_{0i}$$

and

$$\sum_i n_i (\log e_{f,i}) (\log e_{0i}) = a \sum_i n_i \log e_{0i} + b \sum_i n_i (\log e_{0i})^2.$$

We make the following table :

TABLE 26.1
SHOWING THE CALCULATION OF INCOME-ELASTICITY

n_i	e_{0i}	$e_{f,i}$	$\log e_{0i}$	$\log e_{f,i}$	$\log e_{0i} \log e_{f,i}$	$(\log e_{0i})^2$
136	543·1	291·8	2.7349	2.4651	6.7418	7.4797
179	681·3	331·6	2.8333	2.5206	7.1416	8.0276
111	861·9	374·4	2.9355	2.5733	7.5539	8.6172
22	1232·0	407·1	3.0906	2.6097	8.0655	9.5518

Here we have

$$n = \sum_i n_i = 448,$$

$$\sum_i n_i \log e_{f,i} = 1129.4907,$$

$$\sum_i n_i \log e_{0i} = 1272.9408,$$

$$\sum_i n_i (\log e_{f,i}) (\log e_{0i}) = 3211.1551$$

and

$$\sum_i n_i (\log e_{0i})^2 = 3620.8284.$$

Substituting these values in the normal equations and solving the equations, we get

$$b = 0.45$$

and

$$a = 1.24.$$

The income-elasticity of demand (η_e) is given by

$$\eta_e = \frac{d \log e_f}{d \log e_0} = b = 0.45.$$

Obviously, the demand for food is inelastic.

Questions and exercises

26.1 Explain the meaning of elasticity of demand with respect to price. Given the demand for a commodity and the corresponding price, how will you calculate the elasticity of demand with respect to price ?

26.2 Describe a statistical law of demand and indicate the difficulties in its determination from time-series data.

26.3 Suppose you are asked to obtain the demand function for foodgrains in India. What variables will you include in explaining the demand ? How will you obtain the demand function on the basis of time-series data ?

26.4 What do you mean by income-elasticity of demand ? Given family-budget data, how would you estimate this elasticity ? What adjustments would you make for variation in the size of the family ?

26.5 Discuss the different forms of the Engel curve usually employed for fitting to family-budget data. In such fitting, how would you tackle the following complications ?

(a) Household expenditure on a particular item depends, besides depending on income, on the number of persons per family.

(b) Consumption of families of the same size differs because of varying age- and sex-composition.

26.6 Let d_1 and d_2 represent the demand of a commodity for two strata of a population. If η_1 and η_2 be the elasticities of demand with respect to national income for the two strata, show that the corresponding elasticity η for the two strata combined would be given by

$$\eta = \frac{\eta_1 d_1 + \eta_2 d_2}{d_1 + d_2}.$$

26.7 The following data represent per capita purchase of unhusked rice (q) in mds. and the retail price (p) in Rs. per md. for the years 1948-60. Obtain a linear demand function and calculate the price-elasticity of demand for each year at the average price prevailing in the year.

<i>Year</i>	<i>q</i>	<i>p</i>
1948	1.89	18.0
1949	1.88	20.1
1950	1.87	21.0
1951	1.60	24.2
1952	1.66	23.3
1953	1.72	23.8
1954	2.02	19.6
1955	1.82	16.8
1956	1.86	20.1
1957	1.93	22.4
1958	1.96	23.8
1959	1.99	22.9
1960	1.86	23.2

Partial ans. The demand function is $q = 2.199 - 0.0162p$.

26 8 The following table gives in rupees the monthly expenditure on clothing and the total monthly household expenditure of civilian staff employed in Defence Headquarters. Derive the Engel curve for clothing, assuming its form to be linear in doubly-logarithmic scale. Obtain the income-elasticity of demand for clothing.

Family group	Number of households	Average family size	Average monthly household expenditure (in Rs.)	on cloth. g	total
1	439	4.5	16.6	174.4	
2	361	5.0	17.1	198.3	
3	128	5.1	22.6	252.9	
4	784	5.4	24.7	297.4	
5	192	5.4	26.9	342.1	
6	49	5.2	27.6	387.5	
7	48	4.7	29	468.2	
8	40	6.6	47.7	570.1	
9	73	6.0	47.2	849.2	
10	45	6.4	71.1	1,253.3	

Partial ans. Income-elasticity of demand for clothing = 0.676

SUGGESTED READING

- [1] Klein, L. R. *An Introduction to Econometrics* (Chs. 1—2). Prentice-Hall, 1962, and Prentice-Hall of India, 1965.
- [2] Lange, O. *Introduction to Econometrics* (Ch. 2). Pergamon Press, 1959.
- [3] Prais, S. J. and Houthakker, H. S. *The Analysis of Family Budgets* (Chs. 7—11). Cambridge Univ. Press, 1955.
- [4] Schultz, H. *The Theory and Measurement of Demand* (Chs. 1—4). Univ. of Chicago Press, 1937.
- [5] Wold, H. and L. Jureén. *Demand Analysis*, John Wiley, 1951.

27

STATISTICAL QUALITY CONTROL

27.1 Introduction

By *statistical quality control (SQC)* we mean the various statistical methods used for the maintenance of quality in a continuous flow of manufactured products. In any manufacturing process, it is not possible to produce goods of exactly the same quality ; variation is inevitable. Certain small variation is natural to the process, being due to chance causes, and cannot be prevented ; this variation, therefore, is called *allowable*. Sometimes superimposed on this there will be variation which occurs when the process goes wrong, the causes of this variation being assignable ; such variation, therefore, is called *preventable*. The main purpose of *SQC* is to devise statistical methods for separating allowable variation from preventable variation, so that we may take appropriate steps as quickly as possible whenever assignable causes are operating in the process. In other words, an attempt is made to weed out systematic causes of variation as soon as they occur, so that the actual variation may be supposed to be due to the inevitable random causes alone.*

In the above type of problem, our aim is to control the manufacturing process so that the proportion of defective items is not excessively large. This is known as *process control*. In another type of problem, we like to ensure that *lots* of manufactured goods do not contain an excessively large proportion of defective items. This is known as *product or lot control*. The two are distinct problems, because, even when the process is in control, so that the proportion of defective products for the entire output over a long period will not be large, an individual lot of items may not be of satisfactory quality. Process control is achieved mainly through the technique of *control charts*, whereas product control is achieved through *sampling inspection*.

27.2 Different types of quality-measures

By quality we mean any characteristic of the finished products, of intermediate products or of raw materials which is of interest.

Many quality characteristics are measurable quantitatively and may be looked upon as variables, e.g. diameter of a bobbin, length of a screw, tensile strength of a yarn, chemical composition of a drug, life of an electric bulb, etc. All these are continuous variables, and generally the quality characteristics will be of this kind. Sometimes the characteristic may also be a discrete variable, e.g. the number of defects in a piece of cloth.

Often the quality characteristic cannot be measured and is expressed as an attribute. Here the items may be classified as good (or non-defective) and defective. Thus a bolt which does not fit the nut is defective. Also, an item which contains one or more defects is defective. Again, although the characteristic may be measurable, one may decide to treat it as an attribute for the sake of simplicity and economy. A manufacturer producing rods may classify a rod as defective if it is too long or too short and thus avoid recording its actual length.

27.3 Rational sub-groups and the technique of control charts

The central idea in Shewhart's control chart technique is the division of observations into what are called *rational sub-groups*. These are to be taken in such a way that variation within a sub-group may be attributed entirely to chance causes, while systematic variation, if it at all exists, can occur only from one sub-group to another. In statistical language, the products within a sub-group may be supposed to belong to a single homogeneous population; and the differences, if any, among the populations corresponding to different sub-groups will indicate the presence of systematic variation.

The most obvious basis for the selection of sub-groups is the order of production. Each sub-group will then consist of the product of a machine or a homogeneous group of machines for a short period of time, so that there cannot be any remarkable change in the cause system within that period. The use of such sub-groups would tend to reveal assignable causes of variation that come and go. However, there may be assignable causes that are not revealed merely by taking sub-groups in the order of production. E.g., two or more machines in a factory may have different patterns of variation. It may, therefore, be necessary to have different sub-groups for different

machines or for different spindles on the same machine, or for different operators or for different shifts.

The problem of process control then boils down to the use of methods that would enable us to judge whether the distributions of the given quality characteristic for the different sub-groups are identical or not. In case the distributions are identical, the process may be supposed to be in control. Otherwise, the process will be considered to be out of control and one will start looking for the source of trouble. This comparison has, of course, to be performed on the basis of suitable statistics for samples taken from the sub-groups.

Shewhart's control chart technique is a particular diagrammatic method of making this comparison and thus deciding whether the process is or is not affected by systematic variation. We first focus our attention on some parameter of the distribution, say θ . Let T be the corresponding statistic. If the process is in control, then θ must be the same from sub-group to sub-group and, consequently, the fluctuations in the values of T from sample to sample should be due to random variation alone. Supposing in such a case

$$E(T) = \mu_T$$

and

$$\text{var}(T) = \sigma_T^2,$$

one may take any value of T lying outside the limits $\mu_T - 3\sigma_T$ and $\mu_T + 3\sigma_T$ as an indication of the presence of systematic variation. The reason behind this argument lies in the fact that, in case T is normally distributed (and the process is stable),

$$P[|T - \mu_T| \leq 3\sigma_T] = 0.9973, \text{ approximately.}$$

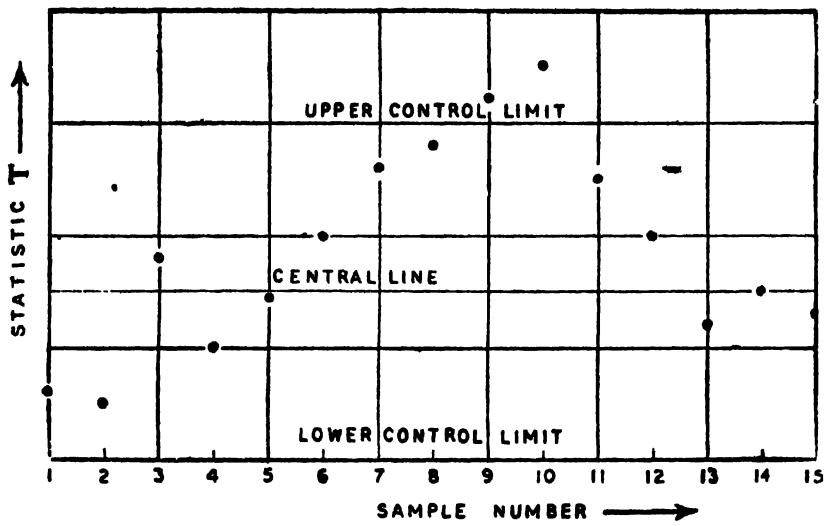
Even when T is non-normal, we have from Chebyshev's inequality

$$P[|T - \mu_T| \leq 3\sigma_T] > 8/9.$$

Thus, if the observed T_i lies between the limits $\mu_T - 3\sigma_T$ and $\mu_T + 3\sigma_T$, it is taken to be a fairly good indication of the non-existence of assignable causes of variation at the time when the i th sample was taken. If the observed T_i wanders outside the limits, one suspects the existence of assignable causes of variation and the process is supposed to be out of control. The obvious action is then to stop the process and to hunt for and remove the assignable causes. The

testing is, however, done by means of a horizontal chart where time is the abscissa and the values of the statistic T are plotted as ordinates. The *lower control limit* (LCL), $\mu_T - 3\sigma_T$, and the *upper control limit* (UCL), $\mu_T + 3\sigma_T$, are shown on the chart by means of horizontal lines. Generally, one also takes a line corresponding to the mean value μ_T , which is called the *central line*.

The Shewhart control chart technique consists in inspecting a fixed number of articles at regular intervals *during production*, measuring the associated statistic and then plotting them as ordinates on a horizontal chart, like Fig. 27.1, with a central line and a pair of control lines. If a plotted point falls within the control limits, then the process is assumed to be *in control* at that moment of production. If it falls outside the control limits, the process is said to be *out of control* at that moment and the presence of some assignable cause is indicated.



pical control chart.

From the above chart, e.g., it appears that the process has been out of control in the 9th and 10th samples.

Even though all the points are inside the control limits, indications of trouble or presence of assignable causes of variation in the

process are sometimes evidenced from unusual patterns or arrangements of points, e.g.

- (a) a series of points all falling close to one of the control limits,
- (b) a long series predominantly on one side of the central line or
- (c) a series of points exhibiting a trend.

There are two types of control chart : (1) Control chart with respect to a given standard—here our purpose is to discover whether the observed values of \bar{x} , s , p , etc., for samples of n items differ from the respective standard values \bar{x}' , σ' , p' , etc., by amounts greater than what should be attributed to chance. The standard values may be either established by authority as some desired or aimed-at values designated by specification or some economic standard levels provided by experience. These charts are used to maintain quality uniformly at the desired level. We may have a process in good control for a long time and we may know the type of population we are inspecting. We then use the standards to set control limits in-order to know about future production. (2) Control chart with no standard given—here we want to discover whether the observed values of \bar{x} , s , p , etc., for samples of size n vary amongst themselves by amounts greater than what should be attributed to chance. The common case that arises in quality control is one in which we do not have any prior knowledge about the process. We use the process to estimate the parameters involved in the lines of the control charts. In the language of Burr, “the control chart is the engineer’s stethoscope for the process” and we find from the process whether it is stable and what level it is maintaining. These charts are used to detect lack of constancy of the cause system.

So far as the size of the samples for different sub-groups is concerned, small samples at shorter intervals are always preferable to large samples at longer intervals. For (\bar{x}, R) or (\bar{x}, s) control charts, samples of size 4 to 8 are sufficiently good for the detection of lack of control. The successive samples are generally taken of equal size for variable control charts. For control charts for attributes, however, one has to take sufficiently large samples since the diagnostic power of charts for attributes is much less than that of charts for variables ; also it is easy and quick to examine products by “go” and “not go” gauges.

27.4 3-sigma control limits and probability limits

The limits on a control chart based on $\mu_T + 3\sigma_T$ and $\mu_T - 3\sigma_T$ are known as 3-sigma limits, as they are obtained after multiplying σ_T by 3. We have also noted in the last section that there is a probability associated with 3-sigma limits. But the real basis is not the value of the probability that a point charted will be inside the control limits (or fall outside the limits). It is said that experience indicates that the use of 3-sigma limits achieves control over the two types of error, viz. (i) looking for trouble when there is no trouble and (ii) failing to look for trouble when there is trouble. Also, it has other advantages —the limits are easy to obtain, tables are available and the two limits are symmetrically placed about the central line. In the United States, 3-sigma limits are mostly used.

The other point of view of setting limits on control charts advocates the use of what are known as probability limits. The upper and lower control limits should be so placed that, without any change of the population, the probability that a point will fall outside the limits is .002 (or some other suitable small value). If the statistic plotted is normal and the probability is equally distributed over the *UCL* and below the *LCL* (i.e. .001 in either direction), then the limits will be based on $\mu_T \pm 3.09\sigma_T$. The British use limits based on this principle. Difficulties arise in setting probability limits for *R*, *s*, *p* or *c* since their distributions are not even symmetrical. The probabilities that will be associated will also be approximate since they will be based on estimates of the standard values.

Considering all these, the 3-sigma limits are reasonably satisfactory, though they may not necessarily be the best always. In the construction of control charts, Shewhart chose 3-sigma limits. Charts using 3-sigma limits are called *Shewhart control charts*. We shall restrict our discussion to Shewhart control charts.

27.5 Control charts for mean, s.d. and range

Suppose we are dealing with a quality characteristic (*x*) like length, diameter or breaking strength—i.e. with a continuous variable. For manufactured articles subject solely to random variation, such a variable may be supposed to be normally distributed (being looked upon as the sum of a large number of independent components each

of which contributes a relatively negligible proportion to the total variability of x). This follows from the Central Limit Theorem. The different distributions of x for the different sub-groups are then all supposed to be of the normal type, the i th sub-group giving a distribution with mean μ_i and variance σ_i^2 , say. To examine whether the process is in control, we need see whether the μ 's and the σ 's are the same. The four types of situation that may be encountered here are :

- (a) the process is in control,
- (b) the mean is out of control but not the s.d.,
- (c) the s.d. is out of control but not the mean,
- (d) both the mean and the s.d. are out of control.

The appropriate statistics corresponding to μ and σ are \bar{x} and s . Hence the whole judgment regarding control or lack of it is based on control charts for \bar{x} and s . It is to be remembered, however, that the range R , in spite of its inferiority to s from the theoretical point of view, is simpler and easier to compute. Hence in quality control, the range is often preferred to the s.d. and one would frequently use charts for \bar{x} and R , instead of using charts for \bar{x} and s .

27.5.1 Control charts for mean

Case I : Standard given

For samples of size n per sub-group, we have for a stable system

$$E(\bar{x}) = \mu$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

assuming that the n observations in each sub-group are mutually independent.

Hence if the values for μ and σ are specified as \bar{x}' and σ' , the control chart for \bar{x} will be given by

$$LCL = \bar{x}' - 3 \frac{\sigma'}{\sqrt{n}} = \bar{x}' - A\sigma',$$

$$Central line = \bar{x}'$$

$$and \quad UCL = \bar{x}' + 3 \frac{\sigma'}{\sqrt{n}} = \bar{x}' + A\sigma',$$

$$where \quad A = 3/\sqrt{n}.$$

} .. (27.1)

Case 2 : Standards not given

Let there be m sub-groups and let the successive sample means be $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$; also, let the successive standard deviations be s_1, s_2, \dots, s_m and the successive ranges be R_1, R_2, \dots, R_m . Since μ and σ are unspecified, these are estimated from the samples themselves. Let

$$\bar{\bar{x}} = \sum_i \bar{x}_i / m,$$

$$\bar{s} = \sum_i s_i / m$$

and

$$\bar{R} = \sum_i R_i / m,$$

which are the pooled mean, the mean of sample standard deviations and the mean of sample ranges, respectively.

The relations

$$E(\bar{\bar{x}}) = \mu, \quad \dots \quad (27.2)$$

$$E(\bar{s}) = c_s \sigma \text{ (valid for a normal variable } x), \quad \dots \quad (27.3)$$

where

$$c_s = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sqrt{\frac{2}{n}}, \quad \dots \quad (27.4)$$

and $E(R) = d_s \sigma$. (valid for a normal variable x), $\dots \quad (27.5)$

where d_s is also a function of n but not as simple as c_s , provide us with an estimate for μ and two alternative estimates for σ , viz.

$$\hat{\mu} = \bar{\bar{x}}, \quad \dots \quad (27.6)$$

$$\hat{\sigma} = \bar{s} / c_s \quad \dots \quad (27.7)$$

and $\hat{\sigma} = \bar{R} / d_s \quad \dots \quad (27.8)$

In case one uses the estimates (27.6) and (27.7), the chart for \bar{x} will be based on

$$LCL = \bar{\bar{x}} - 3 \frac{\bar{s}}{c_s \sqrt{n}} = \bar{\bar{x}} - A_1 \bar{s}, \quad \left. \quad \right\} \quad \dots \quad (27.9)$$

$$Central \ line = \bar{\bar{x}}$$

and $UCL = \bar{\bar{x}} + 3 \frac{\bar{s}}{c_s \sqrt{n}} = \bar{\bar{x}} + A_1 \bar{s}, \quad \left. \quad \right\}$

where $A_1 = \frac{3}{c_s \sqrt{n}}$ and is tabulated, together with c_s , for different values of n in Table VII of Appendix B.

On the other hand, if one uses the estimates (27.6) and (27.8), the chart for \bar{x} will be given by

$$\left. \begin{aligned} LCL &= \bar{x} - 3 \frac{\bar{R}}{d_2 \sqrt{n}} = \bar{x} - A_2 \bar{R}, \\ Central line &= \bar{x} \\ UCL &= \bar{x} + 3 \frac{\bar{R}}{d_2 \sqrt{n}} = \bar{x} + A_2 \bar{R}, \end{aligned} \right\} \dots \quad (27.10)$$

and

where $A_2 = \frac{3}{d_2 \sqrt{n}}$ and is, again, given for different values of n in Table VII of Appendix B.

27.5.2 Control charts for s.d.

Case I : Standard given

For a normally distributed variable x , we have

$$E(s) = c_s \sigma$$

$$\text{and } s = \sigma \sqrt{\frac{n-1}{n} - c_s^2}. \quad \dots \quad (27.11)$$

If the standard value of σ is σ' , then the chart will be based on

$$\left. \begin{aligned} LCL &= c_s \sigma' - 3 \sigma' \sqrt{\frac{n-1}{n} - c_s^2} = B_1 \sigma', \\ Central line &= c_s \sigma' \\ UCL &= c_s \sigma' + 3 \sigma' \sqrt{\frac{n-1}{n} - c_s^2} = B_2 \sigma', \end{aligned} \right\} \dots \quad (27.12)$$

and

$$\begin{aligned} B_1 &= c_s - 3 \sqrt{\frac{n-1}{n} - c_s^2} \\ \text{and } B_2 &= c_s + 3 \sqrt{\frac{n-1}{n} - c_s^2} \end{aligned}$$

The values of B_1 , B_2 and c_s are to be obtained from Table VII in Appendix B for certain values of n .

Case 2 : Standard not given

In this case, one will use the estimate s/c_s for σ and get the control chart, on replacing $c_s\sigma'$ in (27.12) by s , from

$$LCL = s - 3 \frac{s}{c_s} \sqrt{\frac{n-1}{n} - c_s^2} = B_3 s, \quad \left. \begin{array}{l} \\ \end{array} \right\} \dots (27.12a)$$

$$\text{Central line} = s$$

$$\text{and } UCL = s + 3 \frac{s}{c_s} \sqrt{\frac{n-1}{n} - c_s^2} = B_4 s, \quad \left. \begin{array}{l} \\ \end{array} \right\}$$

where

$$B_3 = 1 - \frac{3}{c_s} \sqrt{\frac{n-1}{n} - c_s^2} \text{ and } B_4 = 1 + \frac{3}{c_s} \sqrt{\frac{n-1}{n} - c_s^2}.$$

The values of B_3 and B_4 may also be obtained from Table VII of Appendix B for different values of n .

In either case, if LCL , as given by the stated formula, comes out negative, then it is to be taken as zero. This is because in no case can s be a negative quantity.

27.5.3 Control charts for range*Case 1 : Standard given*

For a normally distributed variable x , we have

$$E(R) = d_2 \sigma$$

$$\text{and } \sigma_R = D\sigma, \quad \left. \begin{array}{l} \\ \end{array} \right\} \dots (27.13)$$

where D as well as d_2 is a function of n .

If the standard value of σ is given to be σ' , then the chart for R will be built on the basis of

$$\left. \begin{array}{l} LCL = d_3 \sigma' - 3D\sigma' = D_1 \sigma', \\ \text{Central line} = d_2 \sigma' \\ UCL = d_2 \sigma' + 3D\sigma' = D_2 \sigma', \end{array} \right\} \dots (27.13a)$$

$$\text{and } D_1 = d_3 - 3D \text{ and } D_2 = d_2 + 3D.$$

The values of D_1 and D_2 , as well as the values of d_3 , are obtainable for certain values of n from Table VII, Appendix B.

Case 2 : Standard not given

When no standard value of σ is specified, it is estimated by R/d_2 .

The chart will then be based on

$$\left. \begin{array}{l} LCL = \bar{R} - 3 \frac{D_3}{d_2} \bar{R} = D_3 \bar{R}, \\ \text{Central line} = \bar{R} \\ UCL = \bar{R} + 3 \frac{D_4}{d_2} \bar{R} = D_4 \bar{R}, \end{array} \right\} \dots \quad (27.13b)$$

and

where $D_3 = 1 - 3 \frac{D}{d_2}$, $D_4 = 1 + 3 \frac{D}{d_2}$. The values of these constants are, again, available from Table VII, Appendix B.

In either case, if LCL , according to the stated formula, comes out to be negative, then it is taken to be zero. For R by its very nature can never be a negative quantity.

27.6 Control charts for number defective and fraction defective

When the quality characteristic is an attribute, and each item is recorded as either defective or non-defective, to judge whether the process is in control, one has to ascertain whether the population fraction defective P is the same for all sub-groups. The judgment may be based either on the number of defectives, say d , in the sample or on the fraction defective $p = d/n$ in the sample, where n , as before, denotes the number of items inspected per sub-group.

27.6.1 Control charts for number defective

Case 1 : Standard given

Assuming that each random sample is taken with replacements or, even if taken without replacements, is taken from a practically infinite population, we may suppose that $d = np$ is distributed in the binomial form with

$$E(np) = nP$$

and $\sigma_{np} = \sqrt{nP(1-P)}$,

P being the same for all sub-groups if and only if the process is in control

Hence if p' be the specified standard value of P , the control chart will be constructed on the basis of

$$\left. \begin{array}{l} LCL = np' - 3\sqrt{np'(1-p')}, \\ \text{Central line} = np' \\ UCL = np' + 3\sqrt{np'(1-p')} \end{array} \right\} \dots \quad (27.14)$$

and

Case 2 : Standard not given

If no standard value is specified for P , it will have to be estimated from the samples themselves. The appropriate estimate is the mean fraction defective

$$\bar{p} = \sum_i p_i / m.$$

The lines on the control chart will then be

$$\left. \begin{aligned} LCL &= n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}, \\ \text{Central line} &= n\bar{p} \\ \text{and} \quad UCL &= n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}. \end{aligned} \right\} \dots \quad (27.14a)$$

Note that $n\bar{p}$ can never be negative. Hence if LCL , according to either of the above formulæ, comes out negative, then it is to be taken as zero.

27.6.2 Control charts for fraction defective**Case 1 : Standard given**

In case one constructs a control chart for p instead of $n\bar{p}$, one uses the relations

$$E(p) = P$$

$$\text{and} \quad \sigma_p = \sqrt{\bar{P}(1-P)/n}.$$

Supposing p' is the specified standard for P , the chart will consist of

$$\left. \begin{aligned} LCL &= p' - 3\sqrt{p'(1-p')/n} = p' - A\sqrt{p'(1-p')}, \\ \text{Central line} &= p' \\ \text{and} \quad UCL &= p' + 3\sqrt{p'(1-p')/n} = p' + A\sqrt{p'(1-p')}, \end{aligned} \right\} \dots \quad (27.15)$$

where $A = 3/\sqrt{n}$.

Case 2 : Standard not given

Here the common value P will be estimated by \bar{p} , and one will have

$$\left. \begin{aligned} LCL &= \bar{p} - A\sqrt{\bar{p}(1-\bar{p})}, \\ \text{Central line} &= \bar{p} \\ \text{and} \quad * \quad UCL &= \bar{p} + A\sqrt{\bar{p}(1-\bar{p})}. \end{aligned} \right\} \dots \quad (27.15a)$$

Here too one will have to remember that p can never be negative. Hence if LCL is found negative according to the above formulæ, then it is to be taken to be zero.

27.6.3 Control charts for percent defective

In this case we construct a control chart for $100p'$ instead of p . The formulæ for the three lines of percent defective chart can be written down from (27.15) and (27.15a) as follows :

Case 1 : Standard given

$$\left. \begin{array}{l} LCL = 100p' - 100A\sqrt{p'(1-p')}, \\ \text{Central line} = 100p' \\ \text{and} \quad UCL = 100p' + 100A\sqrt{p'(1-p')} \end{array} \right\} \dots (27.15b)$$

Case 2 : Standard not given

$$\left. \begin{array}{l} LCL = 100\bar{p} - 100A\sqrt{\bar{p}(1-\bar{p})}, \\ \text{Central line} = 100\bar{p} \\ \text{and} \quad UCL = 100\bar{p} + 100A\sqrt{\bar{p}(1-\bar{p})}. \end{array} \right\} \dots (27.15c)$$

A p -chart (or np -chart or $100p$ -chart) is advantageous because it may be used even for characters that are observed as variables. The cost of obtaining data on an attribute is usually less than that for obtaining data on variables. The cost of compiling a p -chart may also be less, since a p -chart may be used for any number of characteristics and may replace many pairs of (\bar{x} , s) or (\bar{x} , R) charts.

In case the sample size is constant, it is immaterial whether one uses the np -chart or the p -chart. If, however, the sample size varies, in the np -chart all three lines will vary with n and the resulting chart will be highly confusing, whereas in the p -chart the central line will be invariant. It is, therefore, simpler and preferable to use the p -chart (or $100p$ -chart) in case the sample size varies.

Instead of computing control limits for each sample size separately, two sets of limits may be computed based on the minimum and the maximum sample sizes. Action need not be taken for points lying within the inner set of limits, while action must be taken for points lying beyond the outer limits. For other points, action should be based on exact control limits.

The confusion in a p -chart (or np -chart or $100p$ -chart) with varying control limits can be avoided with some additional computation. For that, instead of plotting p in the control chart, one should plot

the standardised value, viz.

$$z = \frac{p - p'}{\sqrt{p'(1-p')/n}} \text{ or } \frac{p - \bar{p}}{\sqrt{\bar{p}(1-\bar{p})/n}}, \quad \dots \quad (27.16)$$

according as the standard value for p is specified or not, \bar{p} being the weighted mean of sample proportions with the sample sizes as weights. The central line as well as the control limits becomes invariant with n , since obviously here

$$\left. \begin{array}{l} LCL = -3, \\ Central \ line = 0 \\ UCL = 3. \end{array} \right\} \quad \dots \quad (27.17)$$

and

27.7 Control charts for number of defects

We are now concerned with cases where each item is observed for the number of defects it contains. The distinction between a defective and a defect is clear : a *defective* is an item that fails to fulfil one or more of the given specifications, a *defect* is any instance of the item's lack of conformity to specifications. Every defective item thus contains one or more defects. These defects may be the surface defects in a roll of paper or photographic film, the weak spots in a given length of a fibre or an insulated wire, the imperfections in, say, a 1-metre piece of cloth, the defective rivets in an aircraft, the loose screws or noisy hinges or exposed wires in a refrigerator, and so on.

In many manufactured articles, the opportunities for defects to occur are numerous, even though the probability for a defect to occur in any one spot is negligible. Hence the number of defects (c) may in most cases be supposed to be distributed in the Poisson form, say with parameter λ .

A control chart for c will then aim at detecting any differences that may exist among the Poisson distributions for the different sub-groups or, in other words, among the λ -values for the sub-groups.

Case 1 : Standard given

We know that for a Poisson variable c with parameter λ ,

$$E(c) = \lambda$$

and

$$\sigma_c = \sqrt{\lambda}.$$

Hence if a standard value for λ , say c' , is provided, then the control chart for c will be based on

$$\left. \begin{array}{l} LCL = c' - 3\sqrt{c'}, \\ \text{Central line} = c' \\ UCL = c' + 3\sqrt{c'} \end{array} \right\} \quad \dots \quad (27.18)$$

and

Case 2 : Standard not given

When no standard is specified, λ will have to be estimated from the observed c values. Supposing c_i is the c value for the sample taken from the i th sub-group ($i=1, 2, \dots, m$), the appropriate estimate of λ will be

$$\bar{c} = \sum_i c_i / m.$$

When this is substituted for c' in (27.18), we shall get the lines for the c -chart, viz.

$$\left. \begin{array}{l} LCL = \bar{c} - 3\sqrt{\bar{c}}, \\ \text{Central line} = \bar{c} \\ UCL = \bar{c} + 3\sqrt{\bar{c}}. \end{array} \right\} \quad \dots \quad (27.18a)$$

and

Note that c cannot be negative. Hence if LCL is negative according to the above formulæ, then it is to be taken equal to zero. The above formulæ relate to c -charts with samples of constant size from all sub-groups. In most cases, each sub-group sample will consist of a single article.

However, it is not necessary that different sub-groups should be of constant size. In the case of variable sub-group size, we obtain the number of defects per unit, i.e. $u = c/n$. The central line for a u -chart will be u' , which is the standard number of defects per unit. The limit lines will not be constant, but will vary with the sub-group size n . The lines for the u -chart, i.e. for a chart for the number of defects per unit with variable sample size, are

$$\left. \begin{array}{l} LCL = u' - 3\sqrt{u'/n}, \\ \text{Central line} = u' \\ UCL = u' + 3\sqrt{u'/n}. \end{array} \right\} \quad \dots \quad (27.19)$$

and

When u' is not specified, it is estimated by

$$\bar{u} = \frac{\sum u_i}{\sum n_i},$$

where u_i, n_i are, respectively, the number of defects and sample size for the i th sub-group. Substituting \bar{u} for u' in (27.19), we shall get the lines for the u -chart, viz.

$$\left. \begin{aligned} LCL &= \bar{u} - 3\sqrt{\bar{u}/n}, \\ \text{Central line} &= \bar{u} \\ UCL &= \bar{u} + 3\sqrt{\bar{u}/n}. \end{aligned} \right\} \dots \quad (27.20)$$

and

27.8 Two types of control chart

A control chart may be used either to determine whether past operations of a process have been in control or as a basis for action on future production.

The control limits for the first type of chart are computed solely on the basis of past data. Lack of control will generally be indicated by points lying outside these limits.

Control limits may also be applied as a basis for action on future production. But in this case a revision of the trial limits may be in order, for some of the points may lie outside the limits and indicate lack of control. All the points may not be assumed to come from a stable distribution. It is important, however, that future control limits should be based on data coming from a controlled process. As a practical rule, therefore, points falling outside the trial control limits are left out and new control limits computed using the remaining points. This procedure may be repeated until all points lie within the control limits.

If this is done, then a possible difficulty has to be kept in view. For in future the control chart may indicate a false lack of control, in the sense that the process may be in control at some other level, although it may be out of control at the aimed-at level.

There are some who advocate that in the case of R -charts, s -charts, p -charts or c -charts, lower limits are of interest only as an indication of improvement which is welcome. Steps should be taken to preserve the improvement. For them, it is usual to plot only the upper control limits on these charts.

According to others, this is not so. They contend that when on these charts points fall below the LCL , the assignable cause is just as important to find out as in the case of a point above the UCL . It may be that inspection personnel are not alert. If this is a real process improvement, this should be maintained in future. So to them both UCL and LCL are important.

27.9 Natural tolerance limits and specification limits

The control chart may show that the process is in control at a particular level. But it may also be of interest to know whether the process can meet the specification limits set for the item. A decision on this point may be made by comparing what are called the 'natural tolerance limits' of the process with the specification limits. If μ and σ are the process average and process s.d., respectively, then the limits $\mu \pm 3\sigma$, which include on the average 9,973 out of 10,000 items, will be called the natural tolerances of the process. \bar{x}' and σ' will be estimated by \bar{x} and s/c_4 or R/d_2 , and in this way we shall get estimates of the tolerances, which will be compared with the specification limits.

If the estimated natural tolerances are not included within the specification limits, then a readjustment of the process will be advisable, with respect to either the process average or the process dispersion or both ; or else a revision of specification limits will be called for.

If the estimated tolerances lie well within the specification limits, this will signify that the process is too good. Then too a revision of the specification limits may be called for, or else it may mean that some relaxation of the conditions of production may be allowed, leading perhaps to lower costs.

The ideal situation will be attained when the tolerance limits are approximately coincident with the specification limits.

Ex. 27.1 The following data relate to the life (in hours) of 15 samples of 6 electric bulbs each, drawn at intervals of one hour from a production process. Draw the \bar{x} and R charts and comment.

Sample No.	Life-time (in hours)					
1	620	687	666	769	839	686
2	501	585	524	585	655	668
3	673	701	696	567	622	660
4	646	626	572	628	632	743
5	494	984	659	643	660	640
6	634	755	625	582	685	555
7	619	710	664	693	773	534
8	631	723	614	535	551	570
9	482	791	533	612	497	499
10	706	524	626	503	662	754
11	530	432	379	690	724	536
12	485	497	608	593	648	729
13	585	535	762	588	625	737
14	462	490	635	587	554	673
15	722	608	665	587	531	705

To draw the control charts for mean and range, we have to calculate the mean and range for each of the samples. The sample totals, means and ranges are shown below :

Sample No.	Total	Mean	Range
1	4,267	711.17	219
2	5,518	586.33	167
3	3,909	651.50	134
4	3,847	641.17	171
5	4,080	680.00	490
6	3,836	639.33	200
7	3,993	665.50	176
8	3,620	603.33	188
9	3,414	569.00	309
10	3,775	629.17	251
11	3,291	548.50	345
12	3,560	593.33	244
13	3,832	638.67	227
14	3,401	566.83	211
15	3,818	636.33	191
Total	—	9,360.16	3,523

The mean of sample means and that of sample ranges are

$$\bar{x} = \frac{9,360.16}{15} = 624.01$$

and $R = \frac{3,523}{15} = 234.87.$

From Table VII of Appendix B, we get for $n=6$, $A_3 = .483$. Thus for the mean-chart

$$\begin{aligned} LCL &= \bar{x} - A_3 R \\ &= 624.01 - .483 \times 234.87 \\ &= 624.01 - 113.44 = 510.57, \end{aligned}$$

$$\text{Central line} = \bar{x} = 624.01$$

and $UCL = \bar{x} + A_3 R = 624.01 + 113.44 = 737.45.$

The mean-chart, showing the control limits, the central line and the sample means plotted against the sample numbers, appears in Fig. 27.2.

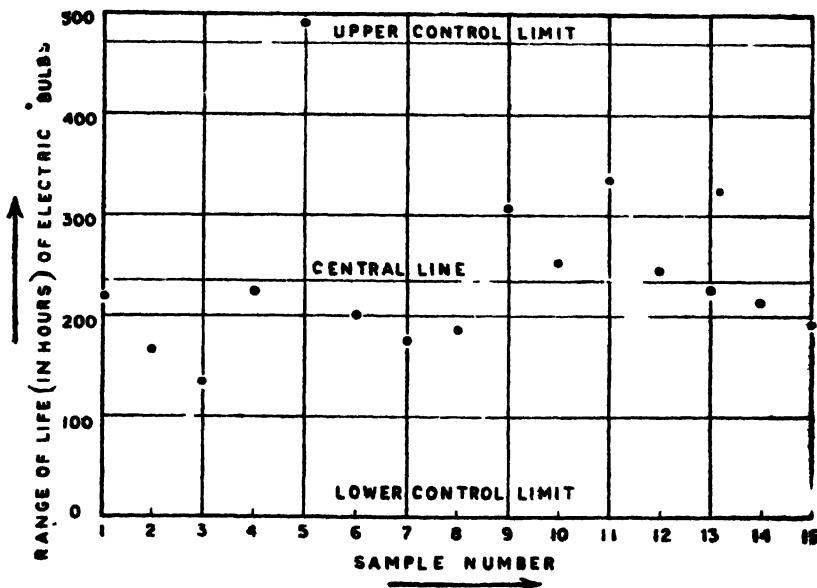


Fig. 27.2 Mean-chart for life (in hours) of electric bulbs.

From the chart we see that all the sample values are well within the control limits. Thus, during the period under consideration,

the process has been in a state of control so far as average life is concerned.

To see whether the process dispersion has also been under control or not, we draw the range-chart. From Table VII of Appendix B, we find, for $n=6$, $D_3=0$ and $D_4=2.004$. Thus, for the range-chart,

$$LCL = D_3 \bar{R} = 0,$$

$$\text{Central line} = \bar{R} = 234.87$$

and

$$UCL = D_4 \bar{R} = 2.004 \times 234.87 = 470.68.$$

The range-chart is shown in Fig. 27.3.

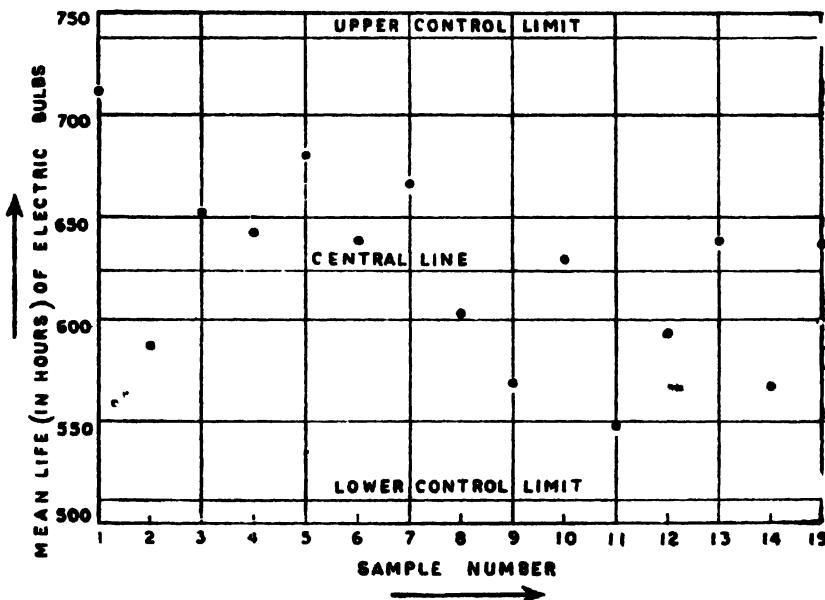


Fig. 27.3 Range-chart for life (in hours) of electric bulbs.

From the chart we find that all the sample ranges are within the control limits, except the range for the fifth sample. Thus we may say that the process dispersion has also been under control, although there is a slight indication of lack of it in the fifth sub-group.

Ex. 27.2 Following are the figures for the number of defectives in 22 lots, each containing 2,000 rubber belts :

425, 430, 216, 341, 225, 322, 280, 306, 337, 305, 356, 402, 216, 264, 126, 409, 193, 326, 280, 389, 451, 420.

Drawing the control chart for fraction defective, plot the points on it. Comment on the state of control of the process.

To draw the control chart for fraction defective, we find the fraction defectives for all the 22 lots under consideration. These are :

.2125, .2150, .1080, .1705, .1125, .1610, .1400, .1530, .1685, .1525, .1780, .2010, .1080, .1320, .0630, .2045, .0965, .1630, .1400, .1945, .2255, .2100.

The pooled fraction defective is

$$\bar{p} = \sum p_i / 22 = 3.5095 / 22 = .1595.$$

Thus the control limits and the central line are

$$\begin{aligned} LCL &= \bar{p} - 3\sqrt{\bar{p}(1-\bar{p})/n} \\ &= .1595 - 3\sqrt{.1595 \times .8405 / 2000} \\ &= .1595 - 3 \times .0082 = .1349, \end{aligned}$$

$$\text{Central line} = .1595$$

$$\text{and } UCL = .1595 + 3 \times .0082 = .1841.$$

The control chart is drawn in Fig. 27.4.

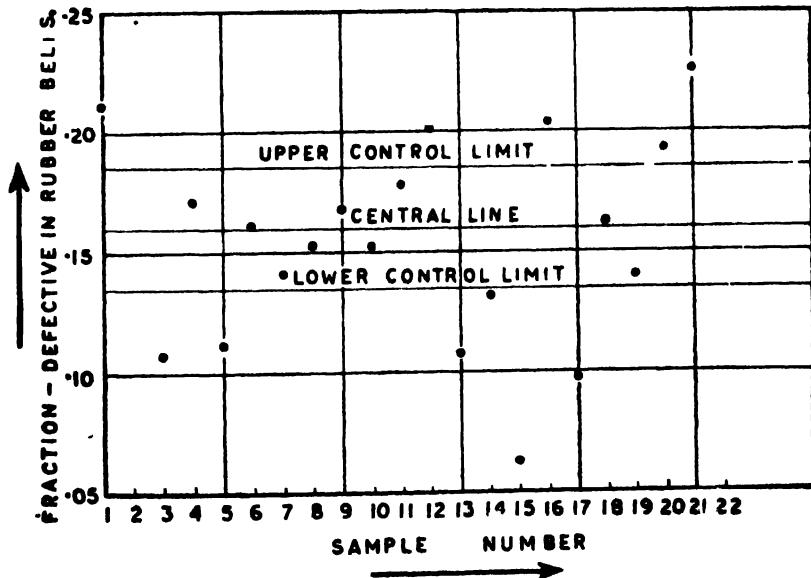


Fig. 27.4 \bar{p} -chart for fraction defective in rubber belts.

From the chart we see that quite a large number of points are outside the control limits. Thus we infer that the production process is completely out of control.

Ex. 27.3 Following are the numbers of defects found in 1,000 items of cotton piece-goods inspected every day in a certain month :

Day	Number of defects	Day	Number of defects
1	1	16	20
2	1	17	1
3	3	18	6
4	7	19	12
5	8	20	4
6	1	21	5
7	2	22	1
8	6	23	8
9	1	24	7
10	1	25	9
11	10	26	2
12	5	27	3
13	0	28	14
14	19	29	6
15	16	30	8

Do these data come from a controlled process ?

Here we have to draw the control chart for the number of defects. If c_i denotes the number of defects in the i th sub-group (here the i th day), we have

$$\bar{c} = \sum_i c_i / 30 = 187 / 30 = 6.23.$$

The control limits and the central line, therefore, are as follows :

$$LCL = \bar{c} - 3\sqrt{\bar{c}} = 6.23 - 3 \times 2.50$$

is negative, so we should take

$$LCL = 0$$

$$Central\ line = \bar{c} = 6.23$$

$$and \quad UCL = 6.23 + 3 \times 2.50 = 13.73.$$

Fig. 27.5 shows the control chart.

The chart indicates that the process is not under control. The sub-groups 14, 15 and 16 and again the 28th sub-group give evidence of lack of control.

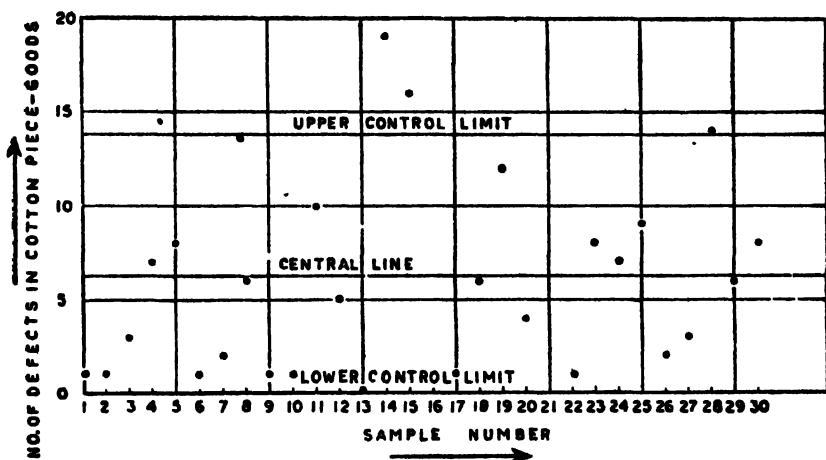


Fig. 27.5 c-chart for defects in cotton piece-goods.

27.10 Advantages of process control

Process control ensures that the quality of a manufacturing process is satisfactory. It may mean a very great saving in industry in addition to the enhanced reputation that comes from the merchandising of a uniformly good product. When a trouble starts in the process, it is of great economic importance to detect it immediately.

Process control provides us with a sound basis for making specifications. There is no point in having specifications which cannot be reached economically. On the other hand, if the inherent tolerances of the process are far inside the specification limits, these limits may well be revised.

Again, lot control would be far more economical if the process were under control, because in that case rejections and the amount of sampling necessary in coming to decisions would be minimised.

27.11 Sampling inspection by attributes

From economic considerations, it is not practicable to inspect fully in lot control; one has to take recourse to *sampling inspection*. For simplicity, we shall mainly deal here with sampling inspection for attributes; i.e., the items are judged good or defective by inspection and the quality of the lot adjudged from the sample fraction defective.

A sampling plan may be of either the *acceptance-rejection* or the *acceptance-rectification* type. In the former, the lot is either accepted or rejected in the light of the sample. In the latter, if the sample does not straightway lead to the acceptance of the lot, it is subjected to cent per cent inspection and, in either case, all defective items encountered are replaced by non-defectives. Although we shall here be concerned mainly with the second type, most of the discussion will be relevant to both the types if we consider the words 'accept' and 'reject' to be inter-changeable with, respectively, the phrases 'accept and replace all defective items in the sample' and 'inspect the lot fully and replace all defectives therein'.

We shall first introduce several concepts which are of importance in deriving optimum sampling inspection plans.

Producer's risk : By 'producer' we shall mean any person, firm or department that prepares goods to be supplied to another person or firm or another department of the same firm. Any sampling inspection plan for acceptance or rejection of a lot possesses the disadvantage of occasionally rejecting a lot of satisfactory quality. Suppose the producer claims that he has standardised the quality at a level of fraction defective \bar{p} , called the *producer's process average*. The probability of rejecting a lot under the sampling inspection plan when the fraction defective is actually \bar{p} is called the *producer's risk* and is denoted by P_p . Clearly, P_p can be kept small by making \bar{p} sufficiently small. But the producer may find it more economical to allow a fairly high risk than to try to reduce \bar{p} .

Consumer's risk : By 'consumer' we shall mean the person or firm or department that receives the articles from the producer. The consumer has to face the risk of accepting a lot of unsatisfactory quality on the basis of sampling inspection. If p_t be the *lot tolerance fraction defective*,* i.e. the maximum fraction defective in the lot that he will tolerate, then the probability of accepting a lot with fraction defective p_t , under the sampling inspection plan, is called the *consumer's risk* and is denoted by P_c .

Average outgoing quality limit (AOQL) : The expected fraction defective remaining in the lot after the application of the sampling plan is called the *average-outgoing quality (AOQ)*. This is naturally a

* 100 p_t is the lot tolerance per cent defective (*LTPD*).

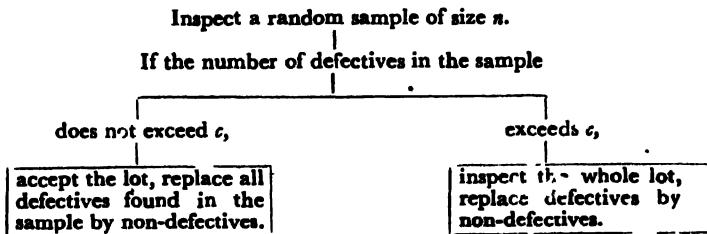
function of p , the actual fraction defective in the lot. The maximum value of the average outgoing quality, the maximum being taken with respect to p , is known as the *average outgoing quality limit* or, briefly, the *AOQL*.

Average sample number (ASN) : The expected value of the sample size required for coming to a decision, i.e. for acceptance or rejection of a lot, under the sampling inspection plan, is called the *average sample number (ASN)*. This is naturally a function of p , the actual fraction defective of the lot. The curve obtained by plotting *ASN* against p is called the *ASN* curve. Obviously, other factors remaining the same, the lower the *ASN* curve, the better is the sampling inspection plan.

Operating characteristic (OC) : The *operating characteristic (OC)* is the mathematical expression, $L(p)$, stating the probability of accepting a lot as a function of p , the fraction defective of the lot. The curve obtained by plotting the operating characteristic against p is called the *OC* curve. Naturally, the steeper the *OC* curve, the greater is the protection to the consumer. An ideal plan, of course, would be one which rejects all lots which are of worse quality than some predetermined value of the fraction defective p and accepts all lots which are equal to or better than that quality. Such a plan, however, can never be attained.

27.11.1 Single sampling plans

A single sampling plan for attributes may be described as follows :



Thus there are two unknown quantities to be determined in this sampling plan, viz. n and c . There are two approaches for determining these quantities :

Lot quality protection

In this approach, the values of n and c are determined from specified values of N , the lot size, p_i , the lot tolerance fraction defective, \bar{p} , the process average, and P_c , the consumer's risk.

If p_i be the lot-tolerance fraction defective, the expression for P_c will be

$$P_c = \sum_{x=0}^c \binom{N-Np_i}{n-x} \binom{Np_i}{x} / \binom{N}{n}, \quad \dots \quad (27.21)$$

and if \bar{p} be the producer's process average, the expression for P_c will be

$$P_c = 1 - \sum_{x=0}^c \binom{N-N\bar{p}}{n-x} \binom{N\bar{p}}{x} / \binom{N}{n}. \quad \dots \quad (27.22)$$

If the actual fraction defective in the lot is \bar{p} , as claimed by the producer, then the expected number I of items to be inspected is

$$I = n + (N-n)P_c, \quad \dots \quad (27.23)$$

since n items are inspected in any case and the remaining $N-n$ are inspected if the number of defectives in the sample exceeds c .

The lot-size N will be specified in any given case, while the consumer's requirement will fix the values of p_i and P_c . Hence expression (27.21) gives an equation in the two unknowns n and c . This equation is satisfied for various combinations of values of n and c . To safeguard the producer's interests too, one would select that pair of n and c for which the expected number of items to be inspected, given by (27.23), is a minimum. The solution, however, is theoretically very difficult to obtain. Extensive tables have been prepared by Dodge and Romig, who obtained the solution by numerical methods.

Average quality protection

In this approach, the problem of protecting the consumer from an inferior product is solved by ensuring him a certain quality level of the product after inspection, regardless of what quality level is being maintained by the producer. This is done by specifying the value of the average outgoing quality limit.

If p be the actual fraction defective in the lot of size N , the average outgoing quality under the single sampling scheme is

given by

$$AOQ = \sum_{x=0}^c \left(\frac{Np-x}{N} \right) \binom{N-Np}{n-x} \binom{Np}{x} / \binom{N}{n}, \quad \dots \quad (27.24)$$

since the fraction defective in the lot after inspection is $(Np-x)/N$, where x is the number of defectives found in the sample, provided x does not exceed c , and it is zero if x exceeds c .

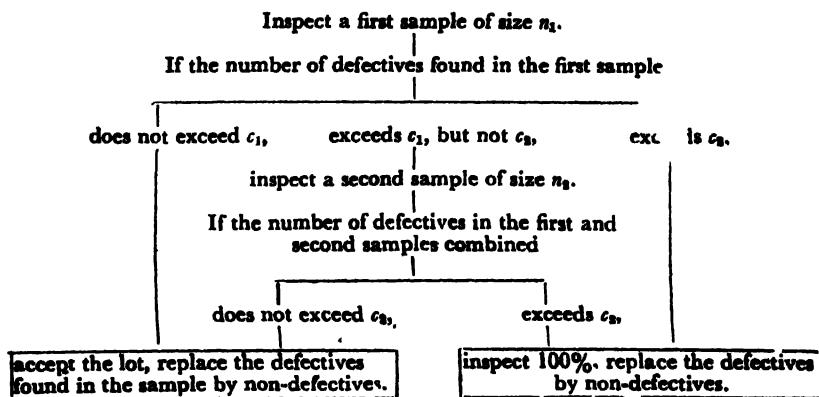
The maximum value of the expression in (27.24) with respect to p is the $AOQL$.

The consumer's interests are taken care of by specifying the $AOQL$. Given the values of N and the $AOQL$, expression (27.24) gives an equation in n and c . In order to safeguard the producer's interests, that pair of n and c satisfying (27.24) is selected for which the expected number of items to be inspected, for the specified value of p , is a minimum.

Extensive tables for sampling plans under this approach are also provided by Dodge and Romig.

27.11.2 Double sampling plans

In the double sampling inspection plan for sampling from a given lot of size N , the procedure for taking action regarding the given lot is as follows :



The values to be determined here are n_1 , n_2 , c_1 and c_2 . As in single sampling, here also there are two approaches for determining these values, viz. approach of lot quality protection and that of average quality protection. The various expressions will, however, be different. The expressions are given below.

Let us denote by $P_{x_1, n_1; Np, N}$ the quantity

$$\binom{N-Np}{n-x} \binom{Np}{x} / \binom{N}{n}.$$

The expression for the consumer's risk P_c is then

$$P_c = \sum_{x=0}^{e_1} P_{x_1, n_1; Np, N} \\ + \sum_{i=1}^{e_2 - e_1} \sum_{x=0}^{e_2 - e_1 - i} P_{e_1+i, n_1; Np, N} \times P_{x_2, n_2; Np - e_1 - i, N - n_1}, \dots \quad (27.25)$$

while the producer's risk P_p is

$$P_p = 1 - \sum_{x=0}^{e_1} P_{x_1, n_1; Np, N} \\ - \sum_{i=1}^{e_2 - e_1} \sum_{x=0}^{e_2 - e_1 - i} P_{e_1+i, n_1; Np, N} \times P_{x_2, n_2; Np - e_1 - i, N - n_1}. \dots \quad (27.26)$$

The expected number I of items inspected per lot for lots with fraction defective \hat{p} is

$$I = n_1 + n_2 (1 - \sum_{x=0}^{e_1} P_{x_1, n_1; N\hat{p}, N}) + (N - n_1 - n_2) P_p, \dots \quad (27.27)$$

while the AQ for lots having fraction defective p is

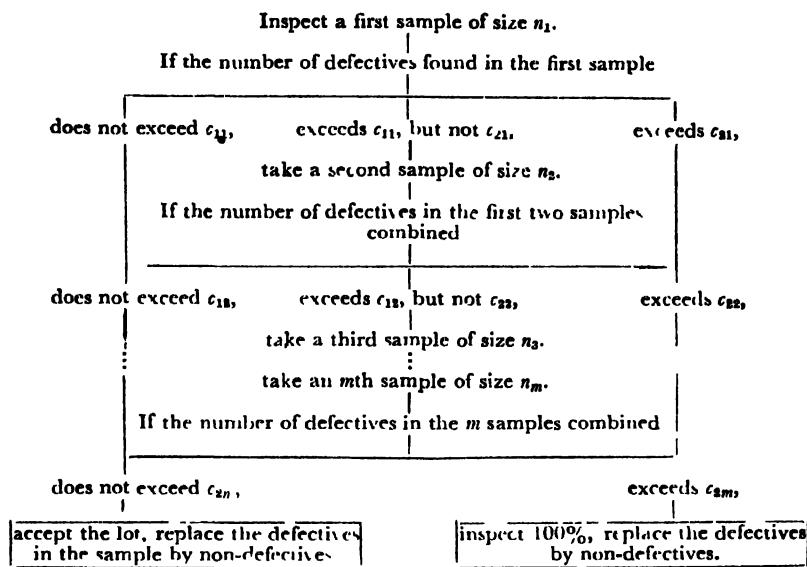
$$AOQ = \sum_{x=0}^{e_1} \left(\frac{Np - x}{N} \right) P_{x_1, n_1; Np, N} \\ + \sum_{i=1}^{e_2 - e_1} \sum_{x=0}^{e_2 - e_1 - i} \left(\frac{Np - (e_1 + i + x)}{N} \right) P_{e_1+i, n_1; Np, N} \\ \times P_{x_2, n_2; Np - e_1 - i, N - n_1}. \dots \quad (27.28)$$

The maximum value of this AQ with respect to p is the AQ_L in the double sampling plan.

Extensive tables for both the approaches are provided by Dodge and Romig.

27.11.3 Multiple sampling plans

The multiple sampling inspection plan is an extension of the double sampling plan, in which the decision to accept the lot or to inspect fully is reached in m or fewer samples, where m is greater than two. In an m -ple sampling plan, the scheme is as follows :



We are not, however, going to discuss here the details of multiple sampling.

27.11.4 Sequential sampling inspection plans

We have seen that a double sampling plan provides for two stages of sampling while a multiple sampling plan provides for more than two. A sequential sampling plan may be said to be the most extreme type of plan in the sense that it provides an infinite number of stages, at each stage there being three possible courses of action, viz. acceptance of lot, rejection of lot and suspension of judgment till the next stage of sampling.

Suppose p is the (unknown) proportion of defectives in the lot. It would be possible to specify two values of p , say p_0 and p_1 ($p_0 < p_1$), such that from the producer's point of view, it will be a serious error to reject the lot when $p \leq p_0$ and from the consumer's point of view, it will be a serious error to accept the lot when $p \geq p_1$. On the other hand, for $p_0 < p < p_1$, both may be indifferent to whether the lot is accepted or rejected. It may also be possible to decide upon two values α and β such that the producer wants the probability of rejection of lot to be less than or equal to α when $p \leq p_0$ and such

that the consumer wants the probability of acceptance to be less than or equal to β when $p \geq p_1$. Thus it is desired that

$$\begin{aligned} L(p) &\geq 1-\alpha \text{ for } p \leq p_0 \\ \text{and} \quad L(p) &\leq \beta \quad \text{for } p \geq p_1. \end{aligned} \quad \} \quad \dots \quad (27.29)$$

The sequential sampling plan for this problem (assuming just one item is taken at each stage) will then be based on the ratio

$$\frac{p_{1m}}{p_{0m}} = \frac{\prod_{i=1}^m p_1^{x_i} (1-p_1)^{1-x_i}}{\prod_{i=1}^m p_0^{x_i} (1-p_0)^{1-x_i}},$$

where x_i ($i=1, 2, \dots, m$) is the sample observation taken at the i th stage, p_{1m} =joint p.m.f. of x_1, \dots, x_m under $H_1: p=p_1$ and p_{0m} = joint p.m.f. under $H_0: p=p_0$. Here it is supposed that x_i , the observation corresponding to the i th item drawn, has the p.m.f.

$$p^{x_i} (1-p)^{1-x_i},$$

with $x_i=1$ if the i th item is defective and $x_i=0$ otherwise.

Thus

$$\frac{p_{1m}}{p_{0m}} = \frac{p_1^{d_m} (1-p_1)^{m-d_m}}{p_0^{d_m} (1-p_0)^{m-d_m}}, \quad \dots \quad (27.30)$$

d_m being the number of defectives found up to the m th stage.

The sequential plan gives a rule for the course of action to be followed at each stage. Thus at the m th stage (which is reached only if no decision has been reached beforehand) one is to

accept the lot if

$$\frac{p_{1m}}{p_{0m}} \leq \frac{\beta}{1-\alpha},$$

reject the lot if

$$\frac{p_{1m}}{p_{0m}} \geq \frac{1-\beta}{\alpha},$$

and suspend judgment till one more unit is taken if

$$\frac{\beta}{1-\alpha} < \frac{p_{1m}}{p_{0m}} < \frac{1-\beta}{\alpha}.$$

On simplification, it will be found to reduce to the form :

accept the lot if $d_m \leq a_m$,

reject the lot if $d_m \geq r_m$

and suspend judgment till one more unit is taken if $a_m < d_m < r_m$,

where
$$a_m = \frac{\log \frac{\beta}{1-\alpha}}{\log \frac{p_1(1-p_0)}{p_0(1-p_1)}} + m \frac{\log \frac{1-p_0}{1-p_1}}{\log \frac{p_1(1-p_0)}{p_0(1-p_1)}} \dots (27.31)$$

and
$$r_m = \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{p_1(1-p_0)}{p_0(1-p_1)}} + m \frac{\log \frac{1-p_0}{1-p_1}}{\log \frac{p_1(1-p_0)}{p_0(1-p_1)}} \dots (27.32)$$

The sequential plan has a two fold merit :

(1) It involves no distribution problem, so that the acceptance number a_m and the rejection number r_m at each stage can be determined simply in terms of p_0 , p_1 , α and β .

(2) Although the sequential plan provides for an infinite number of stages, in any particular case the sampling process will necessarily terminate after a finite number of stages. Its principal merit lies in the fact that the ASN for this plan is found to be almost half as much as the sample size required for a single sampling plan that provides the same type of control on the probabilities of error I and error II.

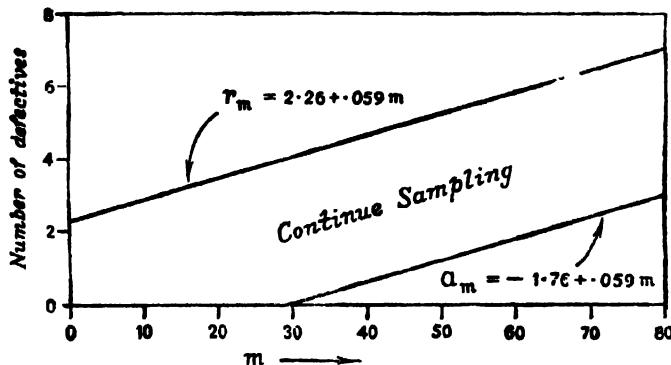


Fig. 27.1 Sequential sampling plan i.e. proportion of defectives under a manufacturing process (with $p_0=.04$, $p_1=.08$, $\alpha=.15$, $\beta=.25$).

The sequential sampling scheme can be performed graphically,

by taking two axes of co-ordinates for m and d_m , and by drawing the acceptance line $d_m = a_m$ and the rejection line $d_m = r_m$ (see Fig. 27.1). As long as the points (m, d_m) lie between these lines, sampling is to be continued. However, whenever a point lies on or below the acceptance line, sampling is to be stopped and the lot accepted. On the other hand, whenever a point lies on or above the rejection line, sampling is to be terminated and the lot rejected.

27.11.5 Comparison of the three types of plans

The two main considerations on the basis of which the three types of plans may be compared are the operating characteristic and the average sample number. Let us consider three equivalent sampling inspection plans, single, double and multiple, for which the *OC* curves are practically the same. The three plans are equivalent in the sense that they give the same amount of protection against rejection (or 100% inspection) of good lots or acceptance of bad lots. The average amount of inspection required per lot is a maximum for single sampling and a minimum for multiple sampling. The exact amount of saving depends on the lot-quality and the particular plan under consideration. Speaking generally, double sampling often requires 25%—33% less inspection on the average and multiple sampling 33%—50% less on the average than single sampling.

Two other factors may also be considered :

- (a) The training of inspectors to use sampling plans is easiest for single sampling and is most difficult for multiple sampling.
- (b) The psychological satisfaction gained from giving the inspection lot more than one chance is absent in single sampling and is a maximum in multiple sampling.

27.12 Sampling inspection by variables

In sampling inspection by variables, each item of a sample taken from the lot of manufactured products is not simply classified as defective or non-defective. But, for each item, measurements are taken on a quality characteristic along a continuous scale in terms of inches, cm., lb., gm., seconds or some such units.

For quality characteristics that can be measured, it is generally true that cost of inspection per item is smaller by attributes than

by variables. Moreover, in sampling inspection by variables the acceptance criteria must be applied separately to each variable. E.g., if 15 different variables are of importance, 15 sets of criteria must be used in inspection by variables, whereas a single set of criteria will be needed in attributes inspection.

However, despite these limitations, variables inspection may be preferred, for this makes a greater amount of information available regarding the lot than does attributes inspection. Put in a different way, for a given quality protection against various types of error, as shown by the *OC* curve, sampling inspection by variables requires smaller samples than attributes inspection. Further, it may be found that, of the 15 variables mentioned earlier, only one or two are troublesome and of real importance. As such, this type of inspection is expected to be more profitable.

27.12.1 Underlying principle

Let x be the quality characteristic in question. It will be assumed that x has the normal distribution in the lot. Associated with x , there will be the specification limits. If only the upper specification limit U is given, then an item is considered defective if, and only if, $x > U$. If only a lower specification limit L is given, then an item is considered defective if, and only if, $x < L$. Whenever both limits are specified, we have, on the other hand, to consider an item defective if, and only if, $x < L$ or $x > U$. If μ and σ are the mean and s.d. of x in the lot, then the lot is to be considered good or bad keeping in view the proportion defective

$$\begin{aligned} p'_U &= \frac{1}{\sigma\sqrt{2\pi}} \int_U^{\infty} \exp[-(x-\mu)^2/2\sigma^2] dx \\ &= 1 - \Phi\left(\frac{U-\mu}{\sigma}\right) \end{aligned} \quad \dots \quad (27.33)$$

in the first case,

$$\begin{aligned} p'_L &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^L \exp[-(x-\mu)^2/2\sigma^2] dx \\ &= \Phi\left(\frac{L-\mu}{\sigma}\right) = \Phi\left[-\left(\frac{\mu-L}{\sigma}\right)\right] \end{aligned} \quad \dots \quad (27.34)$$

in the second case

and

$$p'_L + p'_U$$

in the third.

However, p'_L and p'_U are unknown quantities, because μ and σ are unknown. Sampling inspection provides us with estimates of p'_L and p'_U or, in other words, of μ and σ , in the light of which the lot is to be accepted or rejected.

We shall consider separately the cases where (a) μ is unknown but σ is known and (b) both μ and σ are unknown. It may then be said that in (a) sampling inspection by variables is based on the sample mean \bar{x} and in (b) it is based on the sample mean \bar{x} and the sample s.d. $s' = \sqrt{\sum (x_i - \bar{x})^2 / (n-1)}$.

27.12.2 Variables inspection with known s.d.

When σ is known, there exist minimum-variance unbiased estimates of p'_U and p'_L , viz.

$$p_U = 1 - \Phi \left[\sqrt{\frac{n}{n-1}} \left(\frac{U - \bar{x}}{\sigma} \right) \right] \quad \dots \quad (27.35)$$

$$\text{and} \quad p_L = \Phi \left[-\sqrt{\frac{n}{n-1}} \left(\frac{\bar{x} - L}{\sigma} \right) \right]. \quad \dots \quad (27.36)$$

A sampling plan should naturally lead to the acceptance of the lot if, and only if, the sample proportion of defectives is small. Thus, for an upper specification limit, the lot is to be accepted if

$$p_U < M \text{ (say)}$$

or, equivalently, if

$$\frac{U - \bar{x}}{\sigma} \geq k, \quad \text{or} \quad \bar{x} + k\sigma \leq U,$$

where M is a quantity determined in accordance with the specified probability of error I and

$$k = \sqrt{\frac{n-1}{n}} \cdot \tau_M, \quad \dots \quad (27.37)$$

τ_M being the upper $100 M\%$ point of the normal deviate.

For a lower specification limit, on the other hand, the lot is to be accepted if

$$p_L < M,$$

$$\text{i.e. if} \quad \frac{\bar{x} - L}{\sigma} \geq k \quad \text{or} \quad \bar{x} - k\sigma \geq L.$$

Lastly, if both specification limits are given, then the lot is accepted whenever

$$p_U + p_L \leq M.$$

The values of k , corresponding to the lot size, the sample size and the specified acceptable quality level (with probability of wrong rejection $\alpha = .05$), are given in Tables A and K of Bowker and Goode's *Sampling Inspection by Variables*.

27.12.3 Variables inspection with unknown s.d.

Here also we first look for minimum-variance unbiased estimates of p'_U and p'_L . The minimum-variance unbiased estimate of p'_U is p_U , a function of $\frac{U - \bar{x}}{s'}$ and that of p'_L is p_L , a function of $\frac{\bar{x} - L}{s'}$.

It can be shown that $p_U \leq M$ if, and only if, $\frac{U - \bar{x}}{s'} \geq k$ say. Hence, for an upper specification limit, the lot is accepted if

$$\frac{U - \bar{x}}{s'} \geq k, \quad \text{or} \quad \bar{x} + ks' \leq U,$$

the value of k now being of a more complicated form than in section 27.12.2.

Similarly, $p_L \leq M$ if, and only if, $\frac{\bar{x} - L}{s'} \geq k$. Hence for a lower specification limit, the lot is accepted if

$$\frac{\bar{x} - L}{s'} \geq k \quad \text{or} \quad \bar{x} - ks' \geq L.$$

For two-sided specification limits, the lot is to be accepted if, and only if, $p_U + p_L \leq M$. It is not possible to give an exact equivalent procedure in terms of \bar{x} and s' .

(An approximation would be to accept the lot if the following three criteria are all satisfied :

$$\frac{U - \bar{x}}{s'} \geq k, \quad \frac{\bar{x} - L}{s'} \geq k, \quad s' \leq \text{maximum standard deviation (MSD)},$$

where MSD is a constant depending on M .)

The value of k , for given lot size, sample size and acceptable quality level (with $\alpha = .05$) is again obtainable from Tables A and B of Bowker and Goode's *Sampling Inspection by Variables*.

Questions and exercises

27.1 Explain the theoretical basis of control charts. Explain the construction of various types of control charts for variables and attributes for detection of lack of control in a continuous flow of manufactured products.

27.2 Describe single, double and multiple sampling inspection plans. Give a general outline of methods for determining the constants involved in single and double sampling plans.

27.3 Discuss the following concepts in connection with sampling inspection plans :

Consumer's risk, Producer's risk, $AOQL$, OC curve and ASN curve.

27.4 (a) For single sampling plan, obtain the expressions for the OC and ASN functions. Hence show that $P_c = L(p_i)$, $P_p = 1 - L(\bar{p})$ and that I is the value of the ASN function at $p = \bar{p}$.

(b) Do the same for a double sampling plan.

27.5 Describe the technique of sampling inspection by variables for the normal distribution case.

27.6 A machine is manufacturing mica discs with specified thickness between 0.008" and 0.015". Samples of size 4 are drawn every hour and their thickness in inches recorded as follows :

<i>Sample</i>	<i>Thickness of mica discs (in units of 0.001")</i>			
1	14	8	12	12
2	11	10	13	8
3	11	12	16	14
4	17	12	17	16
5	15	12	14	10
6	13	8	15	15
7	14	12	13	10
8	11	10	8	16
9	14	10	12	9
10	12	10	12	14
11	10	12	8	10
12	10	10	8	8
13	8	12	10	8

<i>Sample</i>	<i>Thickness of mica discs (in units of 0.001")</i>			
14	13	8	11	14
15	7	8	14	13
16	8	10	9	13
17	7	8	16	10
18	7	10	12	10
19	10	12	2	13
20	12	8	10	14

Draw control charts for the mean and the range, and comment on the following points :

- (a) Is the process under control ?
- (b) If so, can it meet the specifications ?
- (c) If the answer to (b) is "no", what percent of articles will fail to meet specifications in the long run ?

Partial ans. $\bar{x} = 11.19$, $\bar{R} = 5.65$.

27.7 The following table gives the results of daily inspection of dowel pin plates for picking up plates with surface defects.

Construct the control chart for fraction defective taking the standardised values z , of (27.16), and comment.

	<i>Date</i>	<i>Number inspected</i>	<i>Number of defectives</i>
January	2	502	18
	4	530	13
	5	480	13
	6	510	15
	7	540	21
	8	520	17
	9	580	28
	11	476	10
	12	570	23
	13	520	10
	14	510	15
	15	536	22

Partial ans. $\hat{p} = 0.3267$.

27.8 The following table gives the number of defects noted at final inspection of aircraft. Find c and the control limits and plot a control chart for c . Comment on the state of control.

Aircraft No.	Number of defects	Aircraft No.	Number of defects
1	7	9	20
2	15	10	11
3	13	11	22
4	18	12	15
5	10	13	8
6	14	14	24
7	7	15	14
8	10	16	8

Partial ans. $\bar{c} = 13.5$.

27.9 The products of a manufacturing industry are submitted for acceptance in lots of 1,000. From past experience, the fraction defective is known to be $p=0.01$. Samples of size n are inspected, and if the number of defectives exceeds c , the remaining articles of the lot are also inspected ; otherwise, the whole lot is accepted. From the following plans, choose the one which involves minimum inspection on the average :

- (i) $n=50, c=0$;
- (ii) $n=80, c=1$;
- (iii) $n=100, c=2$.

(The Poisson approximation may be used).

27.10 (a) For lots of size 3,000, if the process average is 1% and the consumer's lot tolerance percent defective is 3%, what would be the recommended single sampling plan ? What is its $AOQL$?

(b) What would be the corresponding double sampling plan ? What is its $AOQL$? Ans. (a) $n=550, c=11$; $AOQL=1.2\%$

$$(b) n_1=250, n_2=575, c_1=3, c_2=17; AOQL=1.3\%.$$

27.11 (a) Determine from the $AOQL$ Tables, for lots of size 1,000 and a process average of 1.5%, the single sampling plan for which the $AOQL$ will be 2%.

(b) What is the corresponding double sampling plan ?

Ans. From Dodge & Romig's tables, (a) $n=65, c=2$;
 * (b) $n_1=70, n_2=100, c_1=1, c_2=6$.

27.12 Determine, for a sequential sampling plan involving item-by-item inspection, for which $p_0=0.03$, $p_1=0.05$, $\alpha=0.05$ and $\beta=0.10$, the acceptance and rejection lines.

27.13 Determine the *OC* and *ASN* functions of the following three sampling inspection plans and discuss their relative merits. (The lot size may be supposed to be large.)

Sampling plan	Sample	Sample size	Combined sample		
			Size	Acceptance number of defectives	Rejection number of defectives
I	1st	5	5	1	2
II	1st 2nd	6	3 9	0 1	2 2
III	1st 2nd 3rd	3 2 2	3 5 7	0 0 2	2 2 3

SUGGESTED READING

- [1] Bowker, A. H. and Goode, H. P. *Sampling Inspection by Variables* (Chs. 1, 2, 6, 7, 11). McGraw-Hill, 1952.
- [2] Bowker, A. H. and Lieberman, G. J. *Engineering Statistics* (Chs. 12, 13). Prentice-Hall, 1959, and Asia Publishing House, 1962.
- [3] Burr, I. W. *Engineering Statistics and Quality Control*. McGraw-Hill, 1953.
- [4] Cowden, D. J. *Statistical Methods in Quality Control* (Chs. 1, 12, 16, 17, 26, 33, 34, 37, 39). Prentice-Hall, 1957, and Asia Publishing House, 1960.
- [5] Dodge, H. F. and Romig, H. G. *Sampling Inspection Tables*. John Wiley, 1959.
- [6] Duncan, A. J. *Quality Control and Industrial Statistics* (Parts II & IV). Richard D. Irwin, 1953.
- [7] Ekambaram, S. K. *The Statistical Basis of Quality Control Charts*. Asia Publishing House, 1960.
- [8] Grant, E. L. *Statistical Quality Control* (Parts I—IV). McGraw-Hill, 1964.
- [9] Shewhart, W. A. *Economic Control of Quality of Manufactured Product* (Chs. 1, 3, 11, 19, 20). Van Nostrand, 1931.
- [10] Tippett, L. H. C. *Technological Applications of Statistics* (Chs. 1—5, 7). John Wiley, 1950.

APPENDICES

A

INDIAN OFFICIAL STATISTICS

A1 Introduction

By *official statistics* we shall mean numerical data that are collected, compiled and published by different Government bodies. These data appear in different official publications. Most of these publications are issued by various statistical units of the Union Government and deal with data of an all-India nature. There are of course, State publications, from the Indian States, but they are outside the scope of the present discussion.

In India, at present, we have a decentralised statistical system with the Central Statistical Organisation (CSO) as the advisory and co-ordinating body. The collection of data at the Central level is done by various statistical units attached to different Central Ministries and at the State level, by various statistical units in State Ministries under the leadership and guidance of the State Statistical Bureaux. For Central subjects like Railways, Posts and Telegraphs, Foreign Trade, Banking and Currency, Population, etc., the Central Government bears the full responsibility and cost of collection of data. For State subjects like Agriculture, Education, Public Health, etc., the State Governments bear the responsibility for collection of data, although the Central Government may issue directives to bring about uniformity in the nature of the data. There are, however, some concurrent subjects like Industries, Trade Union and Labour, Relief and Rehabilitation, etc. The National Sample Survey (NSS) unit of the Central Government conducts a sample survey on an all-India basis every year to fill in the gaps of the existing statistics and to provide the Planning Commission with data which it may require from time to time. The Indian Statistical Institute (ISI) at Calcutta, besides providing training and carrying out research in theoretical and applied statistics, used to give technical help to the NSS and to do the computational work of the NSS data. However, recently the NSS organisation has set up its own data processing unit and the ISI is no longer associated with its work.

We shall now discuss briefly the various official publications,

their contents and the method of collection of the data under the following heads :

- (1) Population.
- (2) Agriculture.
- (3) Prices.
- (4) Industrial Production
- (5) Trade and Commerce.
- (6) Labour.
- (7) Transport and Communications.
- (8) Miscellaneous.

A2 Population statistics

The main bulk of population statistics emerges out of decennial population censuses. The first Indian census took place in 1872, but it is often left out of account, being too limited in coverage and scope. The next census was in 1881, and since then we are having our population censuses every tenth year.

Up to 1931 the *de facto canvasser* method of census taking used to be employed, in which persons would be counted wherever they were found on the census night and the required information would be collected by enumerators by direct interviews. Since 1941 we have switched over to the *de jure canvasser* method, in which persons are counted at their normal places of residence, and the census is spread over a number of days, usually from two to three weeks. Under the period system, a person is counted at his normal residence provided he is present there at any time during the reference period ; otherwise he will be counted wherever he is found on the day to which the census relates.

Indian censuses are now conducted under the Census Act of 1948, by virtue of which individuals are legally bound to supply data sought for during a census. The census machinery was made permanent in 1949 with the establishment of the office of the Census Commissioner and Registrar-General of India with headquarters at New Delhi.

The census data are collected by about two million honorary workers selected from amongst school teachers, social workers and low-paid Government servants. Each investigator collects data for his Block, which consists of a number of census houses specially

defined for the purpose. A number of Blocks make a Circle, and a Circle-Supervisor is in charge of a Circle. A number of Circles constitute a Charge under a Charge-Superintendent. Other officers in the hierarchy are the District Census Officers, the State Superintendents of Census Operations and, above all, the Registrar-General (and ex-officio Census Commissioner) of India. The census takes place over a period of two to three weeks, the last three days being earmarked for verification.

Hand tabulation of data takes about three to four years. Census reports are published in a number of volumes, there being a general report and special reports dealing with specific topics. Each State also has its own volumes.

The first two regular censuses conducted in 1881 and 1891 were concerned with the following topics :

- (1) Distribution of population, including density per square mile, urban and rural population, housing conditions in towns and average number of persons per house.
- (2) Movement of population, including internal migration.
- (3) Sex.
- (4) Age.
- (5) Occupation of the population, including general village industries and urban occupations.
- (6) Ethnographic distribution, including elaborate discussion of the races in India.
- (7) Religion.
- (8) Literacy and level of education.
- (9) Special physical infirmities like deafness and muteness, leprosy, blindness, etc., according to age and sex, caste and race.
- (10) Civil conditions.

In the censuses of 1901, 1911, 1921 and 1931, the above items of information were more and more extended, the scope and classification of occupational and ethnographic data were made more and more elaborate, while emphasis gradually shifted to the industrial and economic sides of the life of the people. In the 1941 census, although elaborate questions were asked of every citizen, tabulation was confined to a few main items due to paper economy during the Second World War.

The first census after independence was held in 1951. The enumeration acquired added dignity and importance as it was held on the eve of the First Five-year Plan and made many important changes in the traditional procedure. One of the important departures from usual practice was that, apart from the usual statistical data relating to race, language, infirmities, etc., detailed information was gathered in the course of the census on economic characteristics of the population, together with material relating to backward classes and to 'special groups' of communities.

In the 1961 census, there were further changes in the nature of the data collected. A need was felt to collect some data on a household basis, viz. information concerning economic activities of the households—cultivation of land and household industries. The number of family members and that of hired workers participating in household industry or cultivation were also recorded. There were five questions for collection of economic data from every individual—four for persons working and one for persons not working. Working persons were classified as—cultivators, agricultural labourers, those working in household industry and others. For persons other than cultivators or agricultural labourers, the nature of industry, trade or service, was recorded. It was also investigated whether a working person was an employer, employee, single worker or family worker. For persons not doing any gainful work, it was recorded whether he/she was engaged in household duties or was a full-time student, infant, pensioner, beggar, convict in jail or unemployed and seeking employment.

Lastly, there was a complete count of technical and scientific personnel in India, and certain particulars regarding them were collected in a separate schedule meant to be filled up by the persons themselves.

In the 1971 census, the household schedule was replaced by an establishment schedule. Industrial establishments were classified as (i) manufacturing, (ii) processing, (iii) servicing and (iv) household industries. Establishments other than household industries were classified as registered factories or unregistered workshops and by size of employment. They were also classified by industry and by fuel, power or manual labour used. Household industries were classified

by industry, by fuel, power or manual labour used and by size of employment. Trade and commercial establishments were classified by nature of business or trade and by size of employment.

As in 1961, individuals were classified as working and non-working. For working persons, both the main activity and secondary activities were noted. Working persons were classified as workers in manufacturing, processing, servicing or repairing—in household industries or in non-household industries, trade, business, profession or service, or workers in cultivation, or agricultural labourers. Non-working persons were also classified into different categories as in the 1961 census.

The schedule for scientific and technical graduates was extended to graduates in arts and other fields in the 1971 census.

The data collected in our censuses are not as accurate and reliable as they may seem. One of the main defects of our census data arises from the faulty age-returns. The Indian people have shown a peculiar indifference to supplying age data. Very few seem to know precisely their own ages—the enumerator has to make guesses on many occasions. There are preferences for certain ages in our age-returns, e.g. even numbers, numbers which are multiples of 5, etc. Statistics of infirmities have been found huge under-estimates, first because the definitions are ambiguous and secondly because there is deliberate concealment. Literacy is an item in respect of which there has been changes in definition from census to census thus making the data non-comparable over time. The literacy figures seem to be slightly over-estimates. The occupational classification has also changed from census to census. Data on castes are also not very accurate. Caste-prejudices are still not infrequent in our country, specially in rural areas, and many people may falsely register their castes. Some people may falsely register themselves under backward classes for enjoying certain privileges from the Government.

However, Indian censuses have improved a lot both in regard to the procedure and in regard to the nature of the data collected. We have now a permanent census department and we have continuity in our census operations. The experience gained in one census can now be utilised for improving the future censuses. In conducting our future censuses, we should remember that our census data should

be comparable over time. The schedules, definitions and concepts and classifications should not be changed too often. Since sample verification of census data has proved satisfactory, henceforth we may well consider whether certain items may be dropped during the census and may be enumerated later by sampling.

Although our Government feels proud of the fact that our census system is the cheapest in the world since we get the co-operation and devotion of a vast army of honorary workers, this should not continue too long. To get more accurate and reliable data, the enumerators should be paid and given thorough training for the job. For quickness of publication of census data, the processing of data should be centrally managed and electric or electronic machines should be used wherever possible.

Now a word about our vital statistics registers. Data regarding migration and registration of births and deaths in urban areas are more or less reliable. But even now registration of births and deaths in rural areas is incomplete. So any inter-census estimate on the basis of vital statistics and migration statistics is bound to be unreliable.

A3 Agricultural statistics

The Directorate of Economics and Statistics (DES), Ministry of Food and Agriculture, Government of India, has been responsible for the compilation and publication of agricultural statistics in India since 1948. It has brought about a lot of improvement both in the quality and in the quantity of the data collected. The data are collected mainly by the State Governments and are supplied to the DES for compilation and publication.

The available data may be classified into the following groups : (1) land utilisation statistics, (2) area and yield statistics, including crop forecasts, (3) agricultural wages and prices, and (4) miscellaneous statistics regarding livestock and poultry, forestry and fisheries, etc.

Before 1943, even the coverage of primary agricultural statistics, viz. those relating to acreage and yield of crops, was extremely limited. During 1943-44, land utilisation statistics were available for 69% of the total geographical area of the Indian Union. The reliability of the data varied from State to State. In particular, statistics of acreage were known to be unreliable in the permanently settled

areas, being based on estimates given by the village *chowkidars*. Yield statistics were available in respect of only 10 crops through forecasts, while post-harvest estimates were available for a few more crops. The reliability of the estimates, being based on the condition of the crop in relation to the normal crop, was also questionable. Statistics of land utilisation suffered from inadequate classification and absence of uniform definitions in different States. The position in other sectors of agriculture was even worse. The method of collection of data on harvest prices was defective. Statistics on livestock and poultry were collected only during the quinquennial livestock censuses. No regular data were available on livestock products. Forest statistics were collected only for State-owned forests, while statistics on fisheries were not collected at all.

The DES has made improvements in almost all the sectors. The geographical coverage of land utilisation statistics is now almost complete (was 94% in 1967-68, most of the non-reporting area being accounted for by Jammu & Kashmir). The coverage of crop forecasts for foodgrains has been complete since 1948-49. For commercial crops also, the coverage is much larger than before and fairly complete for important crops like cotton, jute, oilseed and sugar cane. The coverage in other sectors of agriculture has also gone up.

Steps have been taken to improve the scope of existing statistics by amplifying the classification or collecting information on more crops. The land utilisation statistics are now available for 9 classes instead of 5 and uniform definitions have been adopted in all the States. With respect to areas under different crops, the old classification has been amplified. Before 1943-44, in all 54 forecasts on 10 crops were available. The system of forecasting has gradually been extended to all principal crops. Now in all 70 forecasts are prepared. Besides, *ad hoc* estimates are available for some minor crops of commercial importance.

The DES has also been able to improve the accuracy and reliability of existing data. Our acreage and yield statistics were subject to large observational errors. The primary reporter, viz. the *patwari* (a village revenue agent), was over-burdened with other work and could devote very little time to collect reliable data. Moreover, he was not trained for the job. Besides, very little supervision

and checking was done of the work performed by the *patwari*. In temporarily settled areas, where the land utilisation statistics and acreage statistics are based on land records maintained by the *patwari*, arrangements have been made to provide him with some training in the technique of observation. Provision has also been made for adequate supervision and checking at the district level. In permanently settled areas, viz. Bihar, Orissa, Kerala and West Bengal, where there is no *patwari*, the task of crop reporting was formerly entrusted to the village headman or *chowkidar*. Recently, however, West Bengal and Kerala have adopted a random sampling method for estimating acreage and Bihar the complete enumeration method with the help of specially appointed staff. In Orissa also, the State Statistical Bureau has been conducting agricultural sample surveys since 1959-60. A working group set up for the improvement of agricultural statistics during the Fourth Plan has, however, recommended complete enumeration in Orissa, Kerala and West Bengal, too.

Yield statistics were formerly based on the formula :

total yield = area \times normal yield \times condition factor.

The concept of 'normal yield' was rather vague and ambiguous whereas the 'condition factor' was based on reports obtained from the *patwaris* or *chowkidars*. At present, for most food crops and some cash crops, these are based on average yield per acre determined from crop-cutting experiments in randomly selected areas. For pre-harvest forecasts, however, the old formula has to be continued. Attention has, therefore, been paid to the improvement of reporting of 'condition factor' and 'normal yield'. The normal yield is now defined as the average of actual yield per acre as determined by crop-cutting surveys over the period of the last 10 years. The procedure of arriving at the condition factor on the basis of the rate of germination of seeds, weather conditions and crop conditions has been standardised and recommended for adoption in all the States.

The defects in the system of reporting harvest prices have also been studied and measures for improvements have been adopted.

Detailed statistics of agricultural wages are now regularly collected in all the States on a monthly basis, giving separately the most commonly prevailing rates for field-labour, other agricultural labour and skilled labour.

Collection of livestock statistics and statistics of livestock products including wool has now been undertaken on a regular annual basis.

Statistics on forests, forest products, employment in forestry and forest industry are now fairly complete for Government-owned forests, but for privately-owned forests the data are not yet complete.

Statistics on fisheries are based on *ad hoc* surveys conducted from time to time.

The following are the important publications that provide data on agricultural statistics :

Indian Agricultural Statistics (annual), Vol. 1 giving Summary Tables (State-wise) and Vol. 2 giving Detailed Tables (district-wise), presents the land utilisation statistics on total area, classification of area into 9 classes, area under crops, area irrigated and crops irrigated.

Estimates of Area and Yield of Principal Crops of India (annual), Vol. 1 giving Summary Tables and Vol. 2 giving Detailed Tables, gives the acreage and yield statistics.

The forecasts are published in the *Agricultural Situation in India* (monthly), which is brought out by the DES.

Special annual reports for some commodities, like tea, rubber, coffee, oilseeds, cotton, sugar, lac, etc., are published by the DES in its so-called Commodity Series.

Indian Livestock Statistics (annual) and *Report of the Indian Livestock Census* (quinquennial) give the data regarding livestock and poultry population and products.

Bulletin of Agricultural Prices (weekly), *Agricultural Prices in India* (annual) and *Agricultural Situation in India* (monthly) give data on agricultural prices, viz. farm (harvest) prices, procurement prices, wholesale prices, retail prices, etc. *Agricultural Wages in India* (annual) supplies detailed statistics of wages of different classes of agricultural labourers. *Indian Forest Statistics* (annual) is the publication giving forestry data, whereas no information except for reports of *ad hoc* surveys is available for fisheries in India.

Besides the publications mentioned, the DES brings out an annual *Abstract of Agricultural Statistics*. The *Abstract* and the monthly journal, *Agricultural Situation in India*, publish an annual index number of agricultural production in India. It covers 19 commodities divided into two groups—food and non-food. The weighted

arithmetic mean of production relatives is used, the weights being the total values of production in the base period. The year ending June, 1950 is the base period.

∴ The data relating to agriculture have already improved with respect to both quality and quantity. But there is still much scope for improvement. There are two methods now followed in different States for finding out the acreage—the complete enumeration method and the random sampling method. The two methods should be compared with respect to cost and accuracy and a uniform method should be adopted. Experiments should be conducted to evaluate the accuracy of crop forecasts based on 'normal yield' and 'crop condition' factors and if possible, a correction formula may be obtained. Alternative estimates based on meteorological factors (through the multiple regression method) may also be considered in this connection.

Statistics for yields separately for irrigated and non-irrigated areas are not collected in most States. Statistics of areas and yields under improved agricultural practices are also not collected. The method of crop-cutting experiments, which is now employed for determining yield rates of principal crops, has not yet been extended to commercial crops and minor crops in most States. Another important set of statistics to be built up is that of land utilisation according to land use potentialities. Data on costs and returns have also to be collected.

There are other fields in which considerable improvement is yet to be made, e.g. the statistics of fisheries and livestock products and statistics of privately-owned forests.

A4 Price statistics

Price statistics can be discussed under the following heads :

- (a) wholesale prices and wholesale price index numbers and
- (b) retail or consumer prices and consumer price index numbers (which are the same as cost of living index numbers).

Index Numbers of Wholesale Prices in India (weekly)—Old Series and Revised Series—are published by the Economic Adviser to the Ministry of Commerce and Industry, Government of India. The old series is based on 230 quotations on 78 items divided into 5 major

groups and 18 sub-groups. With the year ending August, 1939 as base, the index uses a weighted geometric mean of price-relatives, the weights being the values of quantities marketed in the base period. The revised series comprises 555 quotations on 112 items, divided into 6 major groups and 21 sub-groups. With the financial year 1952-53 as base, the new index uses a weighted arithmetic mean of price-relatives, the weights being the values of quantities marketed in the base-period. The publications give, besides the index numbers for all commodities and for each group and sub-group, the individual price-quotations and price-relatives. Recently, the base period of the index has been shifted to the financial year 1961-62, and the revised index number comprises 139 items with 774 quotations. There are 7 main groups, the two new groups being Chemicals, and Machinery and Transport Equipment.

In order to secure a representative character for the index numbers, particularly from the point of view of the markets covered, several varieties are included for many commodities, but the quotations are first averaged by simple geometric mean (for old series) or simple arithmetic mean (for new series), so that at the time of the actual compilation of the index number each commodity has only one price relative.

Besides these, the Economic Adviser also publishes, on weekly, monthly and annual bases, index numbers of wholesale prices of 28 important commodities with the financial year 1952-53 as base.

The statistics of retail prices are highly inadequate in our country. Weekly price quotations of some important commodities in selected centres are available in the *Indian Labour Journal* (monthly), published by the Labour Bureau, Ministry of Labour, Government of India. Various State statistical organisations compile their own cost of living index numbers (or consumer price index numbers), and some of these are available in the *Indian Labour Journal*. The Labour Bureau also compiles and publishes in the journal working-class cost of living index numbers in two separate series : (a) Labour Bureau series covering 21 centres and (b) State series covering 18 centres. The year 1949 is the base. A revised series of working-class consumer price index numbers (base period = the year 1960) for 41 centres is also published in the journal.

The CSO is compiling consumer price index numbers for urban non-agricultural non-manual workers since January, 1961 for 45 centres. The index for each centre includes approximately 180 items of goods and services classified into 5 main groups and 23 sub-groups. The base year of these index numbers is also 1960. During 1958-59, middle-class family living surveys were conducted with the help of the NSS to provide weights needed for this series of index numbers. The centres were selected on the basis of administrative importance, middle-class concentration and regional representation.

The Bureau of Applied Economics and Statistics, West Bengal, compiles and publishes comprehensive data on cost of living index numbers for 25 towns of West Bengal, including Calcutta, separately for people in 5 different expenditure levels, viz. Re. 1—Rs. 100, Rs. 101—Rs. 200, Rs. 201—Rs. 350, Rs. 351—Rs. 700, and Rs. 701 and above. The base period of the series is the year 1960. The index numbers are published on a monthly basis in the *Monthly Statistical Digest, West Bengal*.

The weekly journal *Capital* also publishes a cost of living index number for Calcutta and the neighbouring industrial area.

The cost of living index numbers for different centres in India compiled by different State Governments and the Central Government suffer from the defect of non-comparability, due to diverse methods being adopted in their computation. Besides, most of the index numbers relate to working class people. Steps should be taken to extend the scope of the data to middle class and other classes of people as well.

The Reserve Bank compiles an index number of security prices with the financial year 1961-62 as base. It is published in the (monthly) *Reserve Bank of India Bulletin*.

A5 Industrial statistics

Statistics of industrial production may be discussed under two heads : (a) statistics relating to large-scale manufacturing industries and (b) statistics relating to small-scale cottage industries. Statistics of the second type are lacking in India, except for reports on some *ad hoc* surveys conducted from time to time.

As regards manufacturing industries, the Directorate of Industrial

Statistics (DIS) is responsible for the compilation and publication of data collected by the State Governments under its directives. An annual census of manufacturing industries is conducted under statutory power provided by the Collection of Statistics Act, 1953. The census covers 29 industry groups out of 63 groups into which Indian industries are divided.

The publication of the DIS are : (1) *Annual Report of the Census of Manufacturing Industries* (giving data on the number of factories, capital, employment, wages, value of materials consumed, value of products manufactured and value added by manufacture, etc.) ; and (2) *Monthly Statistics of Production of Selected Industries in India*. Statistics of cotton production may be obtained from *Monthly Statistics of Cotton Production*, issued by the office of the Textile Commissioner. The DIS also compiles a monthly index number of industrial production. It covers 201 items of manufacture and is calculated by the weighted arithmetic mean of production relatives, the weights being the values added by manufacture. The indices for different months are adjusted for varying number of days in a month and for seasonal variation. It is published in the *Monthly Journal of the DIS*.

As part of the National Sample Survey, a sample survey of manufacturing industries is being conducted since 1950, covering all the industries registered under the Indian Factories Act, 1948. The data are available in the reports of the sample survey.

Since 1955 a more comprehensive survey, known as the *Annual Survey of Industries*, covering the entire industries sector and all factories under the Factories Act, 1948, has replaced the census and the sample survey. Factories employing 50 or more workers if using power and 100 or more workers if not using power are completely enumerated. The remaining factories, namely, those employing 10 to 49 workers if using power and 20 to 99 workers if not using power, are covered on the basis of a sample survey. The report is published in 10 volumes by the CSO.

A6 Trade statistics

Trade statistics of India may be classified into two groups corresponding to : (a) foreign trade and (b) inland trade—wholesale and retail.

Statistics of trade are compiled and published by the Directorate-General of Commercial Intelligence & Statistics (DGCIS).

The statistics of foreign trade are adequate and reliable. These are available in the following publications :

(1) *Monthly Statistics of Foreign Trade in India*, Vol. 1 (exports) and Vol. 2 (imports). It contains combined figures of trade by sea, air and land, under exports, re-exports and imports, giving particulars of quantity and values of articles.

(2) *Indian Trade Journal* (weekly), which gives, among other things, statistics of land-borne trade with Nepal, Sikkim and Bhutan.

(3) *Supplement to the Monthly Statistics of Foreign Trade*, which gives data relating to the over-all balance of trade, foreign trade by customs zones, etc.

(4) *Customs and Excise Revenue Statement of the Indian Union*.

(5) *Monthly Bulletin of the Reserve Bank of India*, which also gives some data relating to foreign trade, shipping, balance of payments, etc.

Statistics relating to inland retail trade in India are not available. Data regarding inland wholesale trade are available in :

(1) *Accounts Relating to the Inland (Rail & River-borne) Trade of India* (monthly), giving information on trade movements of 63 selected articles between 36 trade blocks into which the country is divided upto April 1956. A number of revisions were necessitated by the reorganisation of States, and at present the number of trade blocks is 34.

(2) *Statistics of Coasting Trade of India* (monthly), giving total exports and imports to and from each of 12 maritime blocks.

(3) *Statistics of Maritime Navigation of India*, giving information on shipping in foreign and coasting trade of India.

A7 Labour statistics

The Labour Bureau compiles and publishes data under various Acts, like the Factories Act (1948), the Payment of Wages Act (1923), Trades Unions Act (1926), Workmen's Compensation Act (1923), etc. The *Indian Labour Journal* (monthly) publishes various data regarding employment, wages and earnings, trade unions, industrial disputes, absenteeism, accidents, etc. The Bureau also publishes the *Indian Labour Statistics* (annual) and the *Indian Labour Yearbook*. The

Chief Inspector of Mines is responsible for the compilation and presentation of annual data relating to labour employed in mines, in a publication entitled *Annual Report of the Chief Inspector of Mines in India. Statistics of Mines in India*, issued by the Directorate-General of Mines Safety, also gives valuable information. *The Monthly Coal Bulletin* gives data relating to labourers in coal mines.

A8 Transport and communications statistics

The statistics on railway transport are published by the Railway Board in the following publications :

(1) *Annual Report of the Railway Board on Indian Railways*, Vol. 1 and Vol. 2 (Vol. 1 gives a description of the general position and points out the special features of the statistical tables which are published in Vol. 2), and (2) *Monthly Railway Statistics*.

Since April 1952, the Railways have been classified as

(A) Government Railways
and (B) Non-Government Railways.

Government Railways are again classified into different zones, while Non-Government Railways are classified by the names of the railways. The railways are also classified by gauges : Broad—5'6", Metre—3'3 $\frac{1}{2}$ " and Narrow—2' and 2'6".

The following details about the railways are available :

- (1) Mileage—figures are available for both route mileage and track mileage.
- (2) Number of passengers carried and earnings therefrom.
- (3) Tonnage of goods carried and revenue thereon.
- (4) Total capital-at-charge.
- (5) Gross earnings.
- (6) Working expenses.
- (7) Net earnings.
- (8) Information regarding number of locomotives, wagons of various kinds, etc.

Various other statements of economic and general interest are also published.

Statistics of road transport are compiled by the Ministry of Transport, Government of India, and given in the publication *Basic Road Statistics*. Information on civil aviation are compiled by the Director-General of Civil Aviation, Government of India, and

published in the *Monthly Newsletter of Civil Aviation*. Statistics on posts and telegraphs are compiled by the Director-General of Posts and Telegraphs and are presented in the *Statistical Abstract* (annual) published by the CSO.

A9 Miscellaneous statistics

Mineral production : Statistics regarding mineral production and distribution are available in :

- (1) *Annual Report of the Chief Inspector of Mines*,
 - (2) *Indian Minerals Year Book*,
 - (3) *Mineral Production of India* (monthly as well as quarterly),
 - (4) *Review of Mineral Production in India* (quinquennial)
- and (5) *Monthly Coal Bulletin*.

The Indian Bureau of Mines is mainly responsible for the compilation of mineral statistics in India.

Educational statistics : Data regarding education may be available from the following publications issued by the Ministry of Education, Government of India :

- (1) *Education in India*,
 - (2) *Education in Universities in India*
- and (3) *Education in the States in India*.

Financial and banking statistics : The following publications from the Reserve Bank of India give various data on this subject :

- (1) *The Reserve Bank of India Bulletin* (monthly),
 - (2) *Report on Trend and Progress of Banking in India* (annual)
- and (3) *Statistical Tables Relating to Banks in India* (annual).

National income : The National Income Unit under the CSO is responsible for the preparation of official estimates of the annual national income and social accounts of India. The office works under the guidance of a committee set up by the Government and advised by international experts. It publishes the *Annual Report on the National Income of India*.

Besides the publications mentioned earlier, we should mention the Abstracts published by the CSO, which give summary data regarding all topics of interest. The publications are :

- (1) *Statistical Abstract, India* (annual)
- and (2) *Monthly Abstract of Statistics*.

SUGGESTED READING

- [1] Asthana, B. N. and Srivastava, S. S. *Applied Statistics of India*. Chaitanya Publishing House, 1965.
- [2] *Guide to Current Agricultural Statistics*. Economic & Statistical Adviser, Ministry of Food & Agriculture, Govt. of India, 1954.
- [3] Gupta, D. B. and Premi, M. K. *Sources and Nature of the Official Statistics of the Indian Union*. Ranjit Printers & Publishers, 1970.
- [4] Saluja, M. R. *Indian Official Statistical Systems*. Statistical Publishing Society, Calcutta. and Indian Economic Society, Hyderabad, 1972.
- [5] *Statistical System in India*. Central Statistical Organisation, Govt of India, 1966

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 21.6 (for Table VI).
4. Section 27.5 (for Table VII).

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	.3229724	.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	.1670930	.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	.2636830	.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	.8364.9	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	$\phi(\tau)$	$\Phi(\tau)$
1.51	.1275830	.9344783	2.01	.0529192	.9777844	2.51	.0170947	.9939634
1.52	.1256646	.9357445	2.02	.0518636	.9783083	2.52	.0166701	.9941323
1.53	.1237628	.9369916	2.03	.0508239	.9788217	2.53	.0162545	.9942969
1.54	.1218775	.9382198	2.04	.0498001	.9793248	2.54	.0158476	.9944574
1.55	.1200090	.9394292	2.05	.0487920	.9798178	2.55	.0154493	.9946139
1.56	.1181573	.9406201	2.06	.0477996	.9803007	2.56	.0150596	.9947664
1.57	.1163225	.9417924	2.07	.0468226	.9807738	2.57	.0146782	.9949151
1.58	.1145048	.9429466	2.08	.0458611	.9812372	2.58	.0143051	.9950600
1.59	.1127042	.9440826	2.09	.0449148	.9816911	2.59	.0139401	.9952012
1.60	.1109208	.9452007	2.10	.0439836	.9821356	2.60	.0135830	.9953388
1.61	.1091548	.9463011	2.11	.0430674	.9825708	2.61	.0132337	.9954729
1.62	.1074061	.9473839	2.12	.0421661	.9829970	2.62	.0128921	.9956035
1.63	.1056748	.9484493	2.13	.0412795	.9834142	2.63	.0125581	.9957308
1.64	.1039611	.9494974	2.14	.0404076	.9838226	2.64	.0122315	.9958547
1.65	.1022649	.9505285	2.15	.0395500	.9842224	2.65	.0119122	.9959754
1.66	.1005864	.9515428	2.16	.0387069	.9846137	2.66	.0116001	.9960930
1.67	.0989255	.9525403	2.17	.0378779	.9849966	2.67	.0112951	.9962074
1.68	.0972823	.9535213	2.18	.0370629	.9853713	2.68	.0109969	.9963189
1.69	.0956568	.9544860	2.19	.0362619	.9857379	2.69	.0107056	.9964274
1.70	.0940491	.9554345	2.20	.0354746	.9860966	2.70	.0104209	.9965330
1.71	.0924591	.9563671	2.21	.0347009	.9864474	2.71	.0101428	.9966358
1.72	.0908870	.9572838	2.22	.0339408	.9867906	2.72	.0098712	.9967359
1.73	.0893326	.9581849	2.23	.0331939	.9871263	2.73	.0096058	.9968333
1.74	.0877961	.9590705	2.24	.0324603	.9874545	2.74	.0093466	.9969280
1.75	.0862773	.9599408	2.25	.0317397	.9877755	2.75	.0090936	.9970202
1.76	.0847764	.9607961	2.26	.0310319	.9880894	2.76	.0088465	.9971099
1.77	.0832932	.9616364	2.27	.0303370	.9883962	2.77	.0086052	.9971972
1.78	.0818278	.9624620	2.28	.0296546	.9886962	2.78	.0083697	.9972821
1.79	.0803801	.9632730	2.29	.0289847	.9888993	2.79	.0081398	.9973646
1.80	.0789502	.9640697	2.30	.0283270	.9892759	2.80	.0079155	.9974449
1.81	.0775379	.9648521	2.31	.0276816	.9895559	2.81	.0076965	.9975229
1.82	.0761433	.9656205	2.32	.0270481	.9898296	2.82	.0074829	.9975988
1.83	.0747663	.9663750	2.33	.0264265	.9900969	2.83	.0072744	.9976726
1.84	.0734068	.9671159	2.34	.0258166	.9903581	2.84	.0070711	.9977443
1.85	.0720649	.9678432	2.35	.0252182	.9906133	2.85	.0068728	.9978140
1.86	.0707404	.9685572	2.36	.0246313	.9908625	2.86	.0066793	.9978818
1.87	.0694333	.9692581	2.37	.0240556	.9911060	2.87	.0064907	.9979476
1.88	.0681436	.9699460	2.38	.0234910	.9913437	2.88	.0063067	.9980116
1.89	.0668711	.9706210	2.39	.0229374	.9915758	2.89	.0061274	.9980738
1.90	.0656158	.9712834	2.40	.0223945	.9918025	2.90	.0059525	.9981342
1.91	.0643777	.9719334	2.41	.0218624	.9920237	2.91	.0057821	.9981929
1.92	.0631566	.9725711	2.42	.0213407	.9922397	2.92	.0056160	.9982498
1.93	.0619524	.9731966	2.43	.0208294	.9924506	2.93	.0054541	.9983052
1.94	.0607652	.9738102	2.44	.0203284	.9926564	2.94	.0052963	.9983589
1.95	.0595947	.9744119	2.45	.0198374	.9928572	2.95	.0051426	.9984111
1.96	.0584409	.9750021	2.46	.0193563	.9930531	2.96	.0049929	.9984618
1.97	.0573038	.9755808	2.47	.0188850	.9932443	2.97	.0048470	.9985110
1.98	.0561831	.9761482	2.48	.0184233	.9934309	2.98	.0047050	.9985589
1.99	.0550789	.9767045	2.49	.0179711	.9936128	2.99	.0045666	.9986051
2.00	.0539910	.9772499	2.50	.0175283	.9937903	3.00	.0044318	.9986501

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 I of *Biomstrika 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 $\chi_{\alpha, v}^2$

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 $t_{\alpha, v}$

$\frac{\alpha}{v}$	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*

*V*alues of $F_{0.05; v_1, v_2}$

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	236.8	238.9	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	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.47	19.48	19.49	19.50		
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	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	5.75	5.72	5.69	5.66	5.63	
5	6.61	5.79	5.41	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.46	4.43	4.40	4.36		
6	5.99	5.14	4.76	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.77	3.74	3.70	3.67		
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.22	4.07	3.84	3.69	3.58	3.50	3.49	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	
9	5.12	4.26	3.86	3.63	3.48	3.33	3.27	3.22	3.14	3.07	3.02	2.95	2.90	2.85	2.79	2.74	2.66	2.58	
10	4.96	4.10	3.71	3.48	3.36	3.20	3.19	3.10	3.01	2.91	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.54	
11	4.84	3.98	3.59	3.36	3.20	3.19	3.19	3.10	3.01	2.93	2.85	2.80	2.75	2.69	2.62	2.54	2.49	2.45	
12	4.73	3.89	3.49	3.26	3.11	3.00	2.92	2.82	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	
13	4.67	3.81	3.41	3.18	3.03	2.92	2.82	2.76	2.70	2.65	2.60	2.53	2.46	2.40	2.35	2.31	2.27	2.22	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.79	2.71	2.64	2.59	2.54	2.48	2.43	2.38	2.33	2.29	2.25	2.21	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.49	2.43	2.38	2.33	2.28	2.23	2.19	2.15	
16	4.49	3.63	3.24	3.01	2.85	2.70	2.66	2.59	2.54	2.49	2.45	2.38	2.33	2.28	2.23	2.19	2.15	2.11	
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.50	2.45	2.41	2.34	2.27	2.21	2.15	2.10	2.06	2.01	
18	4.41	3.55	3.16	2.93	2.77	2.68	2.62	2.58	2.51	2.46	2.41	2.34	2.27	2.21	2.15	2.11	2.06	2.01	
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	
22	4.30	3.44	3.05	2.82	2.68	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	2.01	1.99	1.95	1.90	1.85	1.80	
28	4.20	3.34	2.95	2.71	2.55	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.62	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.90	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	
280	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	

*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 $F_{0.1: v_1, v_2}$

v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366			
2	99.17	99.25	99.30	99.33	99.36	99.37	99.40	99.42	99.45	99.46	99.47	99.47	99.48	99.49	99.49	99.50			
3	29.46	28.71	28.24	27.91	27.49	27.35	27.05	26.87	26.69	26.50	26.41	26.41	26.41	26.41	26.41	26.42	26.50		
4	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46		
5	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				
6	12.06	11.39	9.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88		
7	9.78	9.15	8.75	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	
8	8.45	7.85	7.51	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	4.86	
9	6.02	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	4.31		
10	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.81	4.71	4.56	4.41	4.33	4.25	4.17	4.08	3.91		
11	6.22	5.32	5.07	4.89	4.74	4.63	4.54	4.49	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60		
12	9.33	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	3.36	
13	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	3.17	
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.67	3.56	3.43	3.35	3.27	3.18	3.09	
15	6.21	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	2.87	
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	3.51	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	
18	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	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	2.49	
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	2.42	
21	6.63	4.62	3.99	3.67	3.36	3.07	2.87	2.67	2.45	2.26	2.05	1.86	1.67	1.48	1.29	1.10	0.91	0.72	
22	5.72	4.82	4.31	3.99	3.76	3.45	3.15	2.89	2.63	2.36	2.17	1.93	1.69	1.45	1.21	1.01	0.79	0.57	
23	5.61	4.72	4.22	3.90	3.67	3.36	3.07	2.79	2.56	2.30	2.06	1.81	1.56	1.31	1.06	0.81	0.56	0.31	
24	5.53	4.64	4.14	3.82	3.59	3.29	2.98	2.63	2.33	2.03	1.73	1.43	1.13	0.83	0.53	0.23	0.03	0.00	
25	7.64	5.45	4.57	4.07	3.75	3.33	3.00	2.67	2.33	2.00	1.67	1.33	1.00	0.67	0.33	0.06	0.00		
26	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.87	2.56	2.23	1.87	1.53	1.17	0.82	0.47	0.17	0.00		
27	7.31	5.18	4.31	3.83	3.51	3.20	2.89	2.60	2.30	1.99	1.67	1.33	1.00	0.67	0.33	0.17	0.00		
28	7.08	4.98	4.13	3.65	3.34	3.12	2.85	2.56	2.27	1.96	1.63	1.30	1.00	0.67	0.33	0.17	0.00		
29	6.85	4.79	3.95	3.48	3.17	2.96	2.70	2.41	2.14	1.84	1.51	1.18	0.85	0.52	0.29	0.15	0.00		
30	6.63	4.61	3.78	3.32	2.90	2.64	2.31	2.04	1.71	1.41	1.08	0.75	0.42	0.19	0.06	0.00	0.00		

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. RANDOM SAMPLING NUMBERS*

4652	3819	8431	2150	2352	2472	0043	3488
9031	7617	1220	4129	7148	1943	4890	1749
2030	2327	7353	6007	9410	9179	2722	8445
0641	1489	0828	0385	8488	0422	7209	4950
8479	6062	5593	6322	9439	4996	1322	4918
9917	3490	5533	2577	4348	0971	2580	1943
6376	9899	9259	5117	1336	0146	0680	4052
7287	0983	3236	3252	0277	8001	6058	4501
0592	4912	3457	8773	5146	2519	3931	6794
6499	9118	3711	8838	0691	1425	7768	9544
0769	1109	7909	4528	8772	1876	2113	4781
8678	4873	2061	1835	0954	5026	2967	6560
0178	7794	6488	7364	4094	1649	2284	7753
3392	0963	6364	5762	0322	2592	3452	9002
0264	6009	1311	5873	5926	8597	9051	8995
4039	7732	8163	2798	1984	1292	0041	2500
9376	7365	7987	1937	2251	3411	6737	0367
3039	3780	2137	7641	4030	1604	2517	9211
8971	8653	1855	5285	5631	2649	6696	5475
0373	4153	5199	5765	2067	6627	3100	5716
9092	4773	0002	7000	7800	2292	2933	6125
2464	1038	3163	3569	7155	2029	2538	7080
3027	6215	3125	5856	9543	3660	0255	5544
5754	9247	1164	3283	1865	5274	5471	1346
4358	3716	6949	8502	1573	5763	5046	7135
7178	8324	8379	7365	4577	4864	0629	5100
5035	5939	3665	2160	6700	7249	1738	2721
3318	0220	3611	9887	4608	8664	2185	7290
9058	1735	7435	6822	6622	8286	8901	5534
7886	5182	7595	0305	4903	3306	8088	3899
3354	8454	7386	1333	5345	6565	3159	3991
3415	7671	0846	7100	1790	9449	6285	2525
3918	5872	7898	6125	2268	1898	0755	6034
6138	9045	6950	8843	6533	0917	6673	5721
3825	1704	2835	4677	4637	7329	3156	3291
1349	0417	9311	9787	1284	0769	8422	1077
4234	0248	7760	6.04	2754	4044	0842	9080
6880	3201	7044	3657	5263	0374	7563	6599
0714	5008	5076	1134	5342	1608	5179	0967
3448	6421	3304	0583	1260	0662	7257	0766
5711	7343	7539	3684	9397	5335	4031	1486
2588	3301	0553	2427	3598	2580	7017	9176
8581	4253	7404	5264	5411	3431	3092	8573
8475	6322	3949	9675	6533	1133	8776	2216
0272	5624	8549	5552	7469	2799	2822	9620
7383	7795	7939	2652	4456	6993	2950	8573
5126	2089	7729	0945	3901	4445	7117	8186
2064	3760	0939	7319	5939	3432	2030	4752
9315	8185	7805	6294	7072	6491	4012	1016
6814	8752	3462	6001	3302	3895	7371	3432

TABLE VI (Contd.)

4433	0247	9747	0412	3893	2503	2972	4154
9193	7314	1501	4702	7030	9601	0630	3727
4246	0693	6041	0931	2952	4968	8239	7729
6974	1051	8966	5157	2154	9558	7646	3043
5673	1602	8741	0513	8713	6108	7329	7698
7370	7319	4104	6025	4209	5042	4501	7824
6934	0165	3319	6222	4129	6524	4322	9422
1592	6953	7868	5874	0805	1138	9428	0189
4683	7249	1998	0956	8325	4001	2261	8844
4206	3295	1732	6780	8409	6957	5292	5041
5885	3316	1187	1217	3912	1107	7220	0035
2584	4222	9438	9652	0338	9712	8715	9587
1275	5976	4273	4895	5751	3112	5082	6050
6801	1709	0038	1231	5222	2473	8909	9970
6853	9282	1196	0347	3135	5902	2384	7929
3210	4345	4448	0229	0371	8269	4448	3348
1684	5742	1897	2503	1656	5702	4613	4108
2391	2897	3406	4844	8756	8011	0246	3663
2543	3913	1429	6379	3369	9040	5983	0436
6793	5986	8153	0769	3347	4014	7007	9018
8118	4615	9668	3408	8878	3534	5549	6929
4970	2717	9943	1136	9504	0519	5240	0991
4496	1109	8238	9173	6244	7230	0991	1463
9022	5050	5383	9582	1326	2516	5589	4051
4816	1007	1067	2866	7916	2674	5578	1675
8897	4869	3221	3266	3567	3365	3675	2195
4234	7491	8194	5072	6555	0799	1940	1232
6933	5786	6675	7853	8325	9408	3252	6799
0502	3633	7793	1529	4067	5459	8641	3247
6440	9450	8896	1441	7718	4849	3192	5958
1248	0405	4572	6861	3737	9558	1025	8707
3110	1168	6046	5837	6243	6745	2362	7710
8822	3604	7844	2085	7923	7979	0648	9003
8680	1201	2536	0308	8733	9722	4556	4684
5327	1250	9502	0340	9894	0438	2677	9200
3798	0805	8037	7474	0516	8715	8398	5552
2688	7601	3408	6525	2710	4547	9156	1623
8552	8348	7934	1530	3523	6882	4334	7237
8713	5638	7620	3148	4508	3123	4023	4560
2104	4716	7582	4576	8105	7527	9082	2426
6503	8499	3100	2209	3406	6314	6910	8051
0085	0711	9557	8428	4332	9685	6492	7422
3822	3407	5603	5431	0083	7074	6929	7054
2193	9184	4815	0566	1214	8483	2282	0916
5392	1390	7100	4578	5107	7946	4502	2765
4635	6166	3085	4297	8619	0912	6917	5364
0495	3715	6053	1723	0114	8257	4650	9901
3296	3067	3040	0852	2939	4015	6927	7710
1348	5573	7270	6840	745u	5933	6472	3750
3132	2603	5574	1528	8104	5520	7279	7940

*Reproduced from *Tracts for Computers*, No. XV (*Random Sampling Numbers*, arranged by L. H. C. Tippett), pp. 12-13, with the kind permission of the Department of Statistics, University College, London.

TABLE VII. FACTORS USEFUL IN THE CONSTRUCTION OF CONTROL CHARTS*

Sample size <i>n</i>	Mean chart		Standard deviation chart		Factor for central line		Factor for control limits		Range chart		
	<i>A₁</i>	<i>A₂</i>	<i>c_a</i>	<i>B₁</i>	<i>B₂</i>	<i>B₃</i>	<i>B₄</i>	<i>d₁</i>	<i>D₁</i>	<i>D₂</i>	
2	2.121	3.760	1.880	0.5642	0	1.843	0	3.267	1.128	0	3.267
3	1.732	2.394	1.023	0.7236	0	1.858	0	2.568	1.693	0	2.575
4	1.500	1.880	0.729	0.7979	0	1.808	0	2.266	2.059	0	2.282
5	1.342	1.596	0.577	0.8407	0	1.756	0	2.089	2.326	0	2.115
6	1.225	1.410	0.483	0.8686	0.026	1.711	0.090	1.970	2.534	0	5.078
7	1.134	1.277	0.419	0.8982	0.105	1.672	0.118	1.882	2.704	0.205	5.203
8	1.061	1.175	0.373	0.9027	0.167	1.638	0.185	1.815	2.847	0.387	5.307
9	1.000	1.094	0.337	0.9139	0.219	1.609	0.239	1.761	2.970	0.546	5.394
10	0.949	1.028	0.308	0.9227	0.262	1.584	0.284	1.716	3.078	0.687	5.469
11	0.905	0.973	0.285	0.9300	0.299	1.561	0.321	1.679	3.173	0.812	5.534
12	0.866	0.925	0.266	0.9359	0.331	1.541	0.354	1.646	3.258	0.924	5.592
13	0.832	0.884	0.249	0.9410	0.359	1.523	0.382	1.618	3.336	1.026	5.646
14	0.802	0.848	0.235	0.9453	0.384	1.507	0.406	1.594	3.407	1.121	5.693
15	0.775	0.816	0.223	0.9490	0.406	1.492	0.428	1.572	3.472	1.207	5.737
16	0.750	0.788	0.212	0.9523	0.427	1.478	0.448	1.552	3.532	1.285	5.779
17	0.728	0.762	0.203	0.9551	0.445	1.465	0.466	1.534	3.583	1.359	5.817
18	0.707	0.738	0.194	0.9576	0.461	1.454	0.482	1.518	3.640	1.426	5.854
19	0.688	0.717	0.187	0.9599	0.477	1.443	0.497	1.503	3.689	1.490	5.888
20	0.671	0.697	0.180	0.9619	0.491	1.433	0.510	1.490	3.735	1.548	5.922
21	0.655	0.679	0.173	0.9638	0.501	1.424	0.523	1.477	3.778	1.606	5.950
22	0.640	0.662	0.167	0.9655	0.516	1.415	0.534	1.466	3.819	1.659	5.979
23	0.626	0.647	0.162	0.9670	0.527	1.407	0.545	1.455	3.858	1.710	6.006
24	0.612	0.632	0.157	0.9684	0.538	1.399	0.555	1.445	3.895	1.759	6.031
25	0.600	0.619	0.153	0.9696	0.548	1.392	0.565	1.435	3.931	1.804	6.058

*Reproduced from Table B2, ASTM SPT-15C, *Manual on Quality Control of Materials*, with the kind permission of the American Society for Testing and Materials.

INDEX

- Adjusted death rate (*see* standardised death rate)
- Aggregative index, simple, 284
 - weighted, 286
- Alpha test of intelligence, 274
- Amount of information, 88-90
- Analysis of covariance, 35, 107-118
 - for a one-way layout, 109-111
 - for an RBD, 111-113
 - for any complete block design, 113
 - some facts about, 118
- Analysis of variance, 3-44
 - effects of the violation of the assumptions, 43-44
 - for testing linearity of regression, 36-38
 - in the study of relationship, 35-43
 - one-way classification, 6-14
 - two-way classification, 14-34
- AOQL, 380-381
- ASN, 381
- Autocorrelation, 334-335
- Autoregression equations, 334
- Beta test of intelligence, 274
- Bias, 135-137
 - due to defective sampling technique, 136
 - due to faulty demarcation of sampling units, 137
 - due to non-response, 136
 - due to substitution, 137
 - due to wrong choice of statistic, 137
 - in index numbers, 295-297
 - interviewer, 136
 - observational, 136
 - prestige, 136
 - procedural, 135-136
 - response, 136
 - sampling, 136-137
- c-chart, 370-372
- Census, complete, 129, 131-132
 - data, 165-166, 181
- Chain index, 289-290
- Change-over design, 67-68
- Code numbers, 135
- Cohort, 198
- Comparative mortality index, 190-191
- Completely randomised design, 56-58
- Component method, 236
- Confounding, 82-97
 - complete, 85
 - partial, 85, 89-95
- Consumer price index number, 280, 291-292, 295-297, 298-300
- Consumer's risk, 380
- Control charts, 362-373
 - for fraction defective, 368-370
 - for mean, 363-365
 - for number of defectives, 367-368
 - for number of defects, 370-372
 - for range, 366-367
 - for s.d., 365-366
- Control limits, lower, 360
 - upper, 360
- Correlation between two time series, 338-339
- Correlogram, 334-338
- Cost function, 131, 148, 153-154
- Cost of living index number (*see* consumer price index number)
- Critical difference, 11
- Cross-over design (*see* change-over design)
- Cyclical fluctuations, 309, 329-333
- Demand curve, 342-343, 344-348
- Edgeworth-Marshall formula, 286
- Effect of test-length on test-parameters, 264-265, 272-273

- Elasticity of demand**
- income-elasticity, 350
 - price-elasticity, 343-344
- Engel curve**, 348-349, 350-353
- Engel's law**, 349
- Equilibrium price**, 343
- Error control** (*see* local control)
- Error score**, definition of, 260
- Error variance**, 263
- Errors in index number**, 287-288
- formula error, 288
 - homogeneity error, 288
 - sampling error, 288
- Errors in measurement**, some mathematical methods for, 169-170
- Expectation of life**, 197-198
- complete, 197
 - curtate, 198
- Experiment**, 49
- Experimental error**, 49-50
- Experimental unit**, 49
- Exploratory survey** (*see* pilot survey)
- Factor**, 273-274
- general, 273-274
 - group, 274
 - specific, 274
- Factorial experiment**, 68-103
- in a single replicate, 98
 - 2ⁿ-experiment, 69-78
 - 2ⁿ-experiment, 78-95
 - 2ⁿ-experiment, 95-97
- Family budget enquiry**, 292
- Fisher's diagram**, 51
- Fisher's ideal index number**, 287
- Fisher, Irving**, 287, 288
- Fisher, R.A.**, 52
- Free-hand curve-fitting**, 310
- g-factor**, 274
- Gompertz curve**, 319
- Graduation formulæ**, 228-236, 313-319
- Graeco-Latin square**, 66-67
- Group average method**, 318-319
- Group factor theory**, 274
- Group test of intelligence**, 274
- Guard areas**, 54
- Index number**, 280
- of wholesale prices in India (revised series), 297-298
- Inductive inference**, 129
- Intelligence quotient (IQ)**, 275
- Intelligence tests**, 273-276
- Interaction effects**, 71, 79
- generalised, 96
- Interpenetrating subsamples**, 172
- Interval scale**, 245
- Interview method**, 133
- Intrablock subgroup**, 96
- Irregular fluctuations**, 309-310, 333, 334
- Kuczynski, R.R.**, 199
- Lag correlation**, 339
- Lahiri, D.B.**, 167
- Laspeyres' formula**, 286
- Latin square design**, 61-66
- orthogonal Latin squares, 63
 - standard squares, 62
 - transformation set, 62
- Least significant difference** (*see* critical difference)
- Life table**, 195-212
- abridged, 200, 205-211
 - complete, 196-200
 - Greville's method, 206-209
 - King's method, 200, 205-206
 - Reed and Merrell's method, 209-210
- Linear hypothesis**, 5
- Link index**, 290
- Link relatives method**, 326-327
- Local control**, 50, 53-54
- Logistic curve**, 224-231, 319
- fitting of, 226-231
 - Pearl and Reed's method, 226-228
 - Rhode's method, 229-231
- Loss function**, 130
- Mail questionnaire method**, 133
- Main effects**, 70-71, 78-79
- Makcham's formula**, 233-236
- fitting of, 234-236

- Mean chart, 363-365
- Mental age, 275
- Mental ratio, 275
- Migration, 182, 237
- Missing-plot technique, 119-120
 - in RBD, 124
- Model, linear
 - fixed effects, 4, 6-11, 14-18, 24-27
 - linear hypothesis (*see* fixed effects)
 - mixed effects, 4, 19-20, 28-30
 - of analysis of variance, 4
 - of test theory, 259-260
 - random effects, 4, 11-13, 18-19, 27-28
 - variance components (*see* random effects)
- Modified exponential curve, 318-319
- Monthly averages method, 320
- Moving averages, 310-313, 321-323, 331-333, 335-336
- National Sample Surveys (NSS), 170-172
- Norm, 274
- OC* curves, 381
- Official statistics, Indian, 399-414
 - agricultural, 404-408
 - industrial, 410-411
 - labour, 412-413
 - miscellaneous, 414
 - population, 400-404
 - price, 408-410
 - trade, 411-412
 - transport and communications, 413-414
- Orthogonality of a design, 82-83
- Paasche's formula, 286
- Parallel tests, 261-262
- p*-chart, 368-370
- Periodogram analysis, 329-331
- Pilot survey, 134, 149, 154
- Polynomial fitting, 313-317
- Population, 129, 131, 141-142
 - existent, 142
 - finite, 141
 - hypothetical, 142
 - infinite, 141
- Population projection, 236-238
- Power test, 267
- Precision (*see* amount of information)
- Price relative, 280
- Primary table, 135
- Principles of designs, 50-54
- Principles of sample surveys, 130-131
- Producer's risk, 380
- Questionnaire, 133
- Radix (*see* cohort)
- Random sampling numbers series, 138-141
 - "A Million Random Digits", 139
 - advantages of, 138-139
 - Fisher and Yates' 139
 - Kendall and Smith's, 139
 - Tippett's, 139
- Randomisation, 52
- Randomised block design, 58-61
- Range chart, 366-367
- Rates of vital events, 182-183
 - age-specific fertility rate, 214
 - case fatality rate, 195
 - cause of death rate, 191-192
 - crude birth rate, 212-213
 - crude death rate, 183-184
 - crude rate of natural increase, 216-217
 - general fertility rate, 213
 - gross reproduction rate, 217-218
 - infant mortality rate, 193-194
 - maternal mortality rate, 192
 - morbidity incidence rate, 222-223
 - morbidity prevalence rate, 223
 - net reproduction rate, 218-221
 - specific death rate, 184-186
 - standardised death rates, 187-190
 - total fertility rate, 215
- Ratio estimates, 160-162, 164-165
- Ratio scale, 245
- Ratio-to-moving average method, 322-323
- Ratio-to-trend method, 329-325
- Rational sub-groups, 358-359

- Regression analysis**, 35-43, 116-117
- homogeneity of a group of regression coefficients, 116-117
 - test for multiple linear regression model, 40-43
- Regression estimates**, 162-165
- Reliability**, definition of, 263-264
- Kuder-Richardson method, 268-270
 - methods of estimation of, 265-270
 - parallel test method, 266
 - rational equivalence method, 270
 - split-half method, 267-268
 - test-retest method, 266-267
- Replication**, 50, 52-53
- Reporting**, 135
- Sample survey**, 129
- Sampling**, different types of, 142-169
- circular systematic, 158
 - double, 160-165
 - mixed, 142
 - multiphase, 159-160
 - multistage, 151-157
 - non-probabilistic, 142
 - objective, 142
 - probabilistic, 142
 - purposive, 165-166
 - quota, 169
 - simple random, 142-145
 - stratified random, 145-151
 - subjective, 142
 - systematic, 157-159
 - with probability proportional to size, 166-168
- Sampling enquiries**, 129-130
- Sampling frame**, 134
- Sampling inspection by attributes**, 379-388
- double, 383-384
 - multiple, 384-385
 - sequential, 385-387
 - single, 381-383
- Sampling inspection by variables**, 388-391
- with known s.d., 390
 - with unknown s.d., 390-391
- Sampling unit**, 134
- Scaling procedures**, 245-259
- for qualitative answers, 256
 - for rankings or ratings, 254-255
 - Likert's method, 254
 - product scale, 256-259
 - test items, 246-247
 - test scores, 247-253
 - equivalent scores, 249-250
 - linear derived scores, 248
 - percentile scaling, 247
 - σ -scaling, 248
 - standard scores, 248
 - T -scaling, 248-249
 - Z -scaling, 248
- Schedule of enquiries** (*see* questionnaire)
- Scrutiny of data**, 135
- Seasonal fluctuations**, 308, 320-327
- changing pattern, 328
- Secular trend**, 307-308, 310-320
- Semi-average method**, 320
- Serial correlation** (*see* autocorrelation)
- Series of experiments**, 120-122
- Simple index numbers**, 284
- Size and shape of plots and blocks**, 54-55
- Slutzky-Yule effect**, 333
- Spearman-Brown formula**, 265
- Specification limits**, 373
- Split-plot design**, 98-107
- s.d. chart**, 365-366
- Standard error of measurement**, 263
- Stationary population**, 197
- Stationary time series**, 333
- different schemes for oscillations in, 333, 334
- Storing of information**, 135
- Supply curve**, 343-344
- Survivorship**, 236-237
- Tabulation of data**, 135
- Technique of random sampling**, 137-141
- Tests for index numbers**, 288-290
- circular test, 290
 - factor-reversal test, 289
 - time-reversal test, 288-289
- Tests for random sampling numbers**
- series, 139-140
 - frequency test, 140
 - gap test, 140

- Tests (*contd.*)
— poker test, 140
— serial test, 140
- Time series, definition of, 305
— components of, 306-310
— preliminary adjustments of data, 305-306
- Tolerance limits, 373
- Treatment, 49
- Trend (*see* secular trend)
- True score, 260, 262
- Two-factor theory, 273-274
- Unbiased estimator, best linear, 149, 146, 153
- Unweighted index numbers (*see* simple index numbers)
- Uses of index numbers, 300-301
- Validity, definition of, 270
— different concepts of, 271-272
— concurrent validity, 271
— construct validity, 272
— content validity, 271
— predictive validity, 271
— estimation of, 270-272
- Variance function, 131, 147, 153
- Vital events, 181
- Vital index, 216-217
- Vital statistics, 181
- Vital statistics registers, 181
- Wholesale price index number, 280
- Yates' method, 75-76, 82