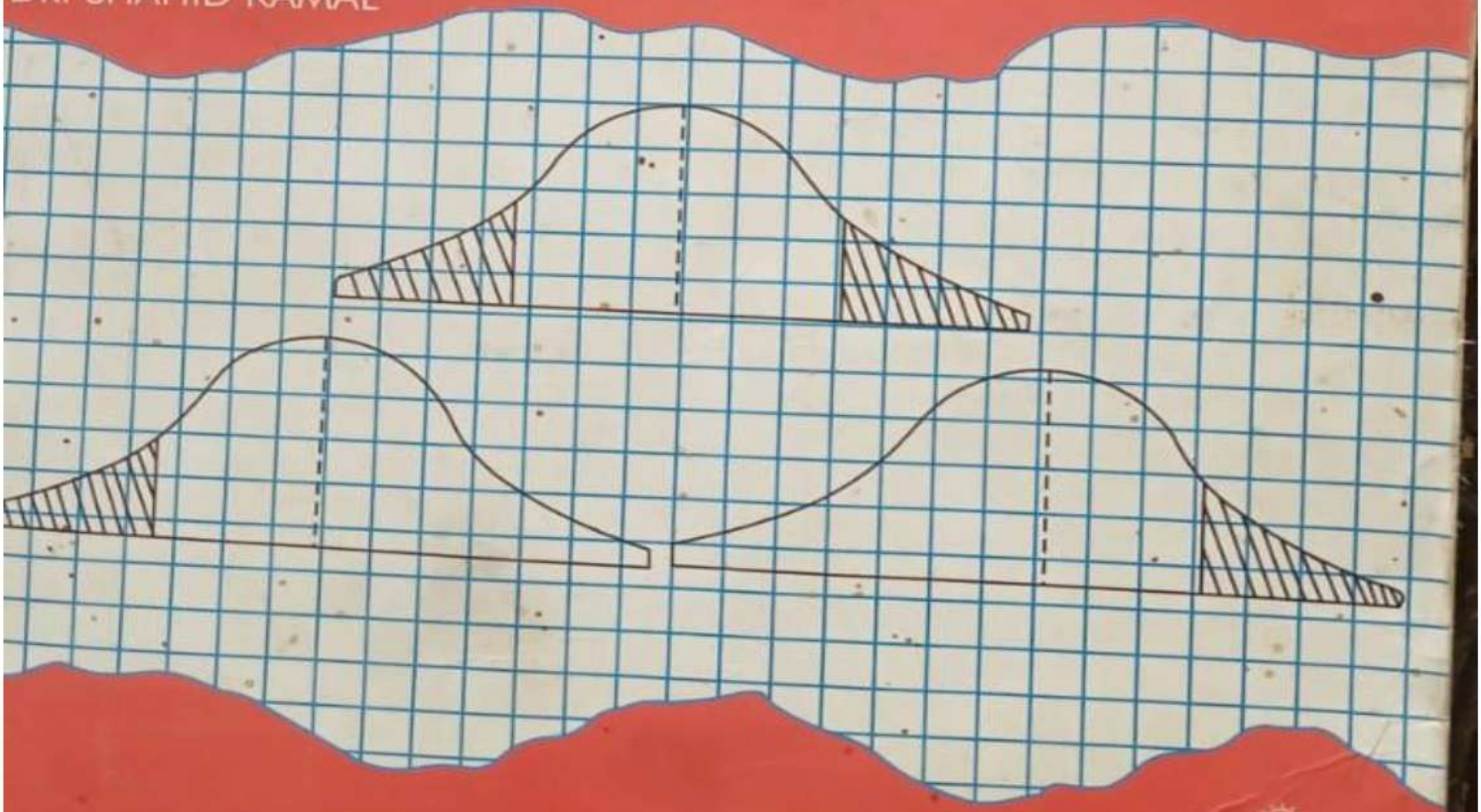
ILMI®

INTRODUCTION TO STATISTICAL THEORY PART II

PROF. SHER MUHAMMAD CHAUDHRY

DR. SHAHID KAMAL



ILMI

INTRODUCTION TO STATISTICAL THEORY

PART 2

(A text book for Degree and Post-Graduate Students)

By

Prof. Sher Muhammad Chaudhry
B. Sc. (Hons.) M.A. (Gold Medalist)
F.S.S. (London)
Formerly, Head of Statistics Department,
Government College, Lahore

Dr. Shahid Kamal
M. Sc (Pb), Ph. D. (U.K)
Institute of Statistics,
University of the Punjab, Lahore

Masha Allah
BOOK STORE
S.S, P.C.S
Near Old Campus
12-3467513

CONTENTS

Preface

14 SURVEY SAMPLING AND SAMPLING DISTRIBUTIONS

14.1. Introduction ... 1
14.1.1. Statistical Populations ... 1
14.1.2. Advantages of Sampling ... 3
14.1.3. Sample Design and Sample Survey ... 4
14.1.4. Sampling Frame ... 4
14.1.5. Probability and Non-probability Sampling ... 5
14.1.6. Sampling With and Without Replacement ... 5
14.1.7. Sampling and Non-sampling Errors ... 5
14.1.8. Sampling Bias ... 6
14.1.9. Random Number Table ... 7
14.2. Probability or Random Samples ... 9
14.2.1. Simple Random Sample ... 9
14.2.2. Stratified Random Sample ... 17
14.2.3. Systematic Random Sample ... 21
14.2.4. Cluster Sample ... 22
14.2.5. Multistage Sample ... 22
14.2.6. Multiphase Sample ... 23
14.2.7. Sequential Sampling ... 23
14.3. Non-Probability Samples ... 23
14.3.1. Purposive Sample ... 23
14.3.2. Quota Sample ... 24
14.4. Sampling Distributions ... 24
14.4.1. Sampling Distribution of the Mean ... 25
14.4.2. Central Limit Theorem ... 41
14.4.3. Sampling Distribution of Differences between Means ... 44
14.4.4. Sampling Distribution of a Sample Proportion ... 49
14.4.5. Sampling Distribution of Differences between Proportions ... 54
14.4.6. Sampling Distribution of Variances ... 55
Exercises ... 55

15 STATISTICAL INFERENCE: ESTIMATION

15.1. Introduction ... 67
15.2. Estimates and Estimators ... 67
15.3. Point Estimation ... 68

15.3.1. Critaria for Good Point Estimators ... 69
15.3.2. Pooled Estimators from Two or More Samples ... 82
15.4. Methods of Point Estimation ... 85
15.4.1. The Method of Maximum Likelihood ... 85
15.4.2. The Method of Moments ... 92
15.4.3. The Method of Least-Squares ... 93
15.5. Estimation by Confidence Interval ... 94
15.5.1. Confidence Interval Estimate of a Population Mean ... 95
15.5.2. Interpretation of a Confidence Interval ... 101
15.5.3. Confidence Interval for Difference of Means ... 103
15.5.4. Confidence Interval for Population Proportion ... 107
15.5.5. Confidence Interval for the Differences between Proportions ... 108
15.5.6. One-sided Confidence Interval ... 110
15.5.7. Sample size for Estimating Population Mean ... 111
15.5.8. Sample size for Estimating Population Proportion ... 113
Exercises ... 113

16 STATISTICAL INFERENCE: HYPOTHESIS TESTING ... 123

16.1. Introduction ... 123
16.1.1. Null and Alternative Hypotheses ... 124
16.1.2. Simple and Composite Hypotheses ... 124
16.1.3. Test-statistic ... 125
16.1.4. Acceptance and Rejection Regions ... 125
16.1.5. Type I and Type II Errors ... 128
16.1.6. The Power of a Test ... 129
16.1.7. The Significance Level ... 129
16.1.8. Test of Significance ... 130
16.1.9. One-tailed and Two-tailed Tests ... 133
16.1.10. Sample size when α and β are specified ... 135
16.1.11. Formulation of Hypotheses ... 137
16.1.12. General Procedure for Testing Hypotheses ... 137
16.2. Tests based on Normal Distribution ... 139
16.2.1. Testing Hypothesis about Mean of a Normal Population when σ is known ... 141
16.2.2. Testing Hypothesis about Mean of a Normal Population when σ is unknown and $n > 30$ ... 143
16.2.3. Testing Hypothesis about Mean of a Non-Normal Population when sample size is large ... 145
16.2.4. Testing Hypotheses about Difference between Two Population Means ...

16.2.5. Testing Hypotheses about a Population Proportion when sample size is large ... 149
16.2.6. Testing Hypothesis about Difference between Two Proportions ... 152
16.2.7. Testing Hypotheses about Standard Deviation: Large Samples ... 156
16.2.8. Relationship between Confidence Interval and Tests of Hypotheses ... 158
16.3. Tests based on Binomial Distribution ... 159
Exercises ... 161

17 THE CHI-SQUARE DISTRIBUTION AND STATISTICAL INFERENCE

17.1. Introduction ... 169
17.2. The Chi-Square Distribution ... 169
17.2.1. Properties of the Chi-Square Distribution ... 171
17.2.2. The χ^2-table ... 175
17.3. Confidence Interval Estimate of Variance of a Normal Population ... 178
17.3.1. Confidence Interval Estimate of σ^2 from a Sample Variance ... 178
17.3.2. Confidence Interval of σ^2 from Several Sample Variances ... 180
17.4. Tests based on Chi-Square Distribution ... 181
17.4.1. Testing Hypothesis about Variance of a Normal Population ... 181
17.4.2. Testing Hypothesis about the Equality of Variances of $k(k>2)$ Normal Populations ... 184
17.5. Karl Pearson's Approximation ... 187
17.5.1. Testing Hypothesis about p's of the Multinomial Distribution ... 190
17.5.2. Pearson's Chi-square Test of Goodness-of-Fit ... 194
17.5.3. Testing Hypothesis about Independence of two Variables ... 200
17.6. An Aside -- Attributes ... 200
17.6.1. Consistence ... 203
17.6.2. Independence ... 204
17.6.3. Association of Attributes ... 205
17.6.4. Measures of Association ... 205
17.6.5. Contingency Tables ... 207

17.7. Testing Hypothesis of Independence in Contingency Tables ... 207
17.7.1. Coefficient of Contingency for an $r \times c$ contingency table ... 213
17.7.2. Yates' correction for Continuity ... 214
17.7.3. An Exact Test for a 2×2 Contingency Table ... 215
17.8. Testing Hypothesis abut Equality of Several Proportions ... 216
17.9. The Chi-Square Test as a Test of Homogeneity ... 219
Exercises ... 222

18 THE STUDENT'S t-DISTRIBUTION AND STATISTICAL INFERENCE ... 239
18.1. Introduction ... 240
18.2. The Student's t-distribution ... 242
18.2.1. Properties of Student's t-distribution ... 243
18.2.2. The t-tables ... 245
18.2.3. Distribution of Difference of Sample Means: Small Samples and $\sigma_1 = \sigma_2$ ... 245
18.2.4. Assumptions in Using t-distribution ... 246
18.3. Confidence Interval Estimates of Mean from Small Sample ... 249
18.4 Small Sample Tests of Means
18.4.1. Testing Hypothesis about Mean of a Normal Population when σ is unknown and $n < 30$ ... 249
18.4.2. Testing Hypotheses about Difference of Means of Two Normal Populations when $\sigma_1 \neq \sigma_2$ but unknown ... 252
18.4.3. Testing Hypotheses about Difference of Means of Two Normal Populations when $\sigma_1 = \sigma_2$ and unknown ... 257
18.4.4. Testing Hypotheses about Two Means with Paired Observations ... 259
Exercises ... 265

19 THE F-DISTRIBUTION AND STATISTICAL INFERENCE
19.1. Introduction ... 273
19.2. The F-Distribution ... 274
19.2.1. Properties of the F-Distribution ... 276

19.2.2. The F-Tables of Areas ... 278
19.2.3. Assumptions in Using F-Distribution ... 278
19.3. Confidence Interval for the Variance Ratio ... **282**
19.4. Tests based on F-Distribution ... **284**
19.4.1. Testing Hypothesis about the Equality of Two Variances ... 284
Exercises ... 290

20 THE ANALYSIS OF VARIANCE

20.1. Introduction ... **295**
20.2. One-Way Analysis of Variance ... **295**
20.2.1. Partitioning the Sum of Squares ... 296
20.2.2. Partitioning the Degrees of Freedom ... 298
20.2.3. The Analysis of Variance Table ... 300
20.2.4. Alternative Computing Formulas ... 301
20.2.5. One-Way Analysis of Variance: Unequal Sample Sizes ... 305
20.2.6. Assumptions of One-Way Analysis of Variance ... 310
20.3. Two-Way Analysis of Variance ... **311**
20.3.1. Two-Way Analysis of Variance without Interaction ... 312
20.3.2. Two-Way Analysis of Variance with Interaction ... 318
20.4. Multiple Comparisons Tests ... 324
20.4.1. The Least Significant Difference Test ... 324
20.4.2. The Student-Newman Keul's Multiple Range Test ... 327
20.4.3. Duncan's Multiple Range Test ... 328
20.4.4. Contrasts - Scheffe's Method ... 330
20.5. The Analysis of Variance Models ... **331**
20.5.1. Least-Squares Estimates of Effects In One-Way ANOVA ... 332
20.5.2. Least-Square Estimates of Effects in a Two-Way ANOVA ... 334
Exercises ... 336

21 STATISTICAL INFERENCE IN REGRESSION AND CORRELATION

21.1. Introduction ... **347**
21.2. Interval Estimation in the Simple Linear Regression ... **350**
21.2.1. Confidence Interval Estimate of Population Regression Coefficient ... 350

21.2.2. Confidence Interval Estimate of α, the Intercept of Regression Line ... 352

21.2.3. Confidence Interval Estimate of Mean Value $\mu_{Y.X}$ for a Given value of X ... 352

21.2.4. Prediction Interval of an Individual Y value for a Given value of X ... 355

21.3. Hypothesis Testing in the Regression Model ... 357

21.3.1. Testing Hypothesis about β, the Population Regression Coefficient ... 357

21.3.2. Testing Hypothesis about α, the Intercept of Population Regression ... 361

21.3.3. Testing Hypothesis about Mean Value $\mu_{Y.X}$ ... 362

21.3.4. Testing Hypothesis about Population Variance ... 362

21.3.5. Testing Hypothesis about Equality of Regression Coefficients of Two Regression Lines ... 362

21.3.6. Testing Hypothesis about the Linearity of Regression ... 365

21.4. Confidence Interval Estimate for Population Correlation Coefficient ... 368

21.5 Hypothesis Testing about Correlation Coefficient ... 370

21.5.1. Testing the Hypothesis that $\rho = \rho_0 (\neq 0)$ ... 371

21.5.2. Testing Hypothesis about Equality of Two Correlations ... 372

21.5.3. Testing Hypothesis about $\rho = 0$ ... 374

21.5.4. Testing Hypothesis about the Equality of Several Correlations ... 376

21.6 Inference in Partial, Multiple Correlation and Regression ... 377

21.6.1. Testing Hypothesis about Partial Correlation ... 377

21.6.2. Testing Hypothesis that a Multiple Correlation Coefficient is zero ... 379

21.7 Analysis of Variance for Regression ... 380

21.7.1. ANOVA for Simple Linear Regression and Test of $H_0 : \beta = 0$ ... 380

21.7.2. ANOVA for Multiple Regression and Testing Hypothesis about the β Parameters ... 384

Exercises ... 388

22 THE ANALYSIS OF COVARIANCE

22.1. Introduction ... 397

22.2. One-Way Analysis of Covariance and Partitioning the Sum of Products ... 398

22.2.1. Alternative Computing Formulas for Sum of Products ... 402

22.3. Two-Way Analysis of Covariance ... 405

22.4. Analysis of Covariance Models. One-Way Classification ... 411

22.4.1. Assumptions Made in Analysis of Covariance ... 413

22.4.2. Uses of Covariance Analysis ... 414

Exercises ... 414

23 EXPERIMENTAL DESIGNS

23.1. Introduction ... 419

23.2. Basic Principles of Experimental Designs ... 420

23.2.1. Randomization ... 420

23.2.2. Replication ... 420

23.2.3. Local Control ... 421

23.3. The Completely Randomized Design ... 421

23.3.1. Experimental Layout ... 421

23.3.2. Statistical Model and Analysis ... 422

23.3.3. Advantages and Disadvantages ... 424

23.4. The Randomized Complete Block Design ... 426

23.4.1. Experimental Layout ... 426

23.4.2. Statistical Model and Analysis ... 427

23.4.3. Advantages and Disadvantages ... 429

23.4.4. Randomized Complete Block Design with Interaction within Blocks ... 433

23.4.5. Missing Observations in *RCB* Design ... 434

23.4.6. Estimation of Missing Observations by Covariance ... 438

23.4.7. Efficiency of a *RCB* Design Relative to a *CR* Design ... 440

23.5. The Latin Square Design ... 441

23.5.1. Construction and Layout ... 442

23.5.2. Statistical Model and Analysis ... 443

23.5.3. Advantages and Disadvantages ... 446

23.5.4. Missing Observations in a Latin Square ... 449

23.5.5. Efficiency of Latin Squares ... 451

23.5.6. Orthogonal Latin Squares and Greeco Latin Square ... 451

23.6. Single Degree of Freedom Contrasts ... 453

23.7. Factorial Experiments ... 454

23.7.1. Main Effects and Interaction Effects ... 455
23.7.2. Effects in a 2^2-Factorial Experiment ... 455
23.7.3. Effects in a 2^3-Factorial Experiment ... 457
23.7.4. Design and Analysis for Factorial Experiments ... 458
23.7.5. Yates' Technique for Computing Contrasts ... 462
23.7.6. Advantages and Disadvantages ... 463
Exercises ... 464

24 NONPARAMETRIC TESTS

24.1. Introduction ... 477
24.2. The Sign Test ... 478
24.3 The Wilcoxon Signed-Rank Test for the Paired Observations ... 483
24.4 The Wilcoxon Rank Sum Test for Independent Samples ... 487
24.5 The Mann-Whitney U Test ... 491
24.6 The Median Test (Two or More Samples) ... 497
24.7 The Runs Test for Randomness ... 498
24.8 The Kolmogorov-Smirnov Tests ... 502
24.8.1. The Kolmogorov-Smirnov One-Sample Test ... 502
24.8.2. The Kolmogorov-Smirnov Two Sample Test ... 503
24.9 The Kruskal-Wallis H Test ... 506
Exercises ... 510
Appendix A — Vital Statistics ... 517
Appendix B — Statistical Tables ... 542
References ... 548
Answers to Exercises ... 550
Index ... 572

14

Survey Sampling and Sampling Distributions

14.1 INTRODUCTION

Sampling is a statistical technique which is used in almost every field in order to collect information and on the basis of this information inferences about the characteristics of a population are made. The values of the population characteristics are summarized by certain numerical descriptive measures, called *parameters*. The values of the population parameters, which are in most situations unknown, would have to be estimated and to get estimates, we resort to sampling. The observations composing a sample are used to calculate a corresponding numerical descriptive measure, called a *statistic*. Thus we use statistics to estimate parameters. Considerations of time and cost are other reasons for sampling. Prior to introducing some of the most commonly used sampling methods, we proceed to some definitions and to a brief description of the basic concepts involved in sampling.

14.1.1 Statistical Populations. A *statistical population* (*or universe*) is defined as the aggregate or totality of all individual members or objects, whether animate or inanimate, concrete or abstract, of some characteristics of interest. The individual members of the population are called *sampling units* or simply *units*. A sampling unit from which information is required, may be a college student, an animal, a tree, a household, a block, a town, a small area, a field, a business firm, etc. A set of n sampling units selected from a given population is called a *sample* of size n and the process of selecting a sample is known as *sampling*. The numerical values assigned to units of interest are treated as values of a random variable X, and the distribution of X is called the *population distribution*.

A population can be either finite or infinite depending upon whether it contains a countable or an uncountable number of units. The population of all students in a college, the population of all licensed motor drivers, the population of all houses in a country, etc. are examples of finite population. The total number of units in a finite population is called the *size* of the population and is denoted by N. The examples of an infinite population are the population of all points on a line, the population of pressures at various points in the atmosphere, etc. A population of concrete units such as trees, households, students, etc. is called an *existent* population while a *hypothetical* population consists of all conceivable ways in which an event can occur, *e.g.* all possible throws of a die. Such a population does not exist in a concrete manner but is only to be thought of.

Furthermore, a *sampled population* is that from which a sample is chosen, whereas a population about which we wish to draw inferences, is called a *target population*. The following two examples may suffice to illustrate the difference between a *sampled* population and a *target* population.

Suppose we desire to know the opinions of college students in the province of the Punjab with regard to the present examination system. Then our population will consist of the total number of students in all the colleges in the province. Suppose on account of shortage of resources or time, we are able to conduct such a survey only on five colleges scattered throughout the province, say, situated in large urban areas. In such a case, the target population consists of the students of all the colleges in the province, while on the other hand, the sampled population consists of the students of five college, from which the sample of students will be selected. As long as the students of these five colleges are representative of the students of all the colleges, the results would be applicable to all the colleges. Similarly, the sampled population may consist of patients in district hospitals and the target population may consist of the total number of patients in the province. It is of some importance to emphasize that the sampled population should be such that its results are applicable to the target population. In case, these results cannot be extended to the target population, they hold good for the sampled population.

A population is *discrete* when the number of units comprising the population is countable, otherwise it is *continuous*.

A complete coverage of the N sampling units will yield numerical values $x_1, x_2, \ldots, x_N$ for X, where X will refer to some *characteristic* of interest associated with each unit in a population. If, for example, the sampling unit is a student, then X might refer to age, height, weight,

marks obtained, attitude towards present system of examination, and the like.

The two *basic purposes* of sampling are (i) to provide sufficient information about the characteristics of a population without examining every unit of the population, and (ii) to find the reliability of the estimates derived from the sample. We find the reliability by computing the standard error of a statistic and if possible, its exact sampling distribution. The definitions of these terms appear in later sections.

14.1.2 Advantages of Sampling. The important advantages of sampling over complete enumeration are briefly stated below:

(i) Sampling saves money as it is much cheaper to collect the desired information from a small sample than from the whole population.

(ii) Sampling saves a lot of time and energy as the needed data are collected and processed much faster than census information. And this is a very important consideration in all types of investigations or surveys.

(iii) Sampling provides information that is almost as accurate as that obtained from a complete census; rather a properly designed and carefully executed sample survey will provide more accurate results. Moreover, owing to the reduced volume of work, persons of higher calibre and properly trained can be employed to analyse the data.

(iv) Sampling makes it possible to obtain more detailed information from each unit of the sample as collecting data from a few units of the population (*i.e. sample*) can be more complete and thorough.

(v) Sampling is essential to obtaining the data when the measurement process physically damages or destroys the sampling unit under investigation. For example, in order to measure the average lifetime of light bulbs, the measurement process destroys the sampling units, *i.e.* the bulbs, as they are used until they burn out. A manufacturer will therefore use only a sample of light bulbs for this purpose and will not burn out all the bulbs produced. Similarly, the whole pot of soup cannot be tasted to determine if it has an acceptable flavour.

(vi) Sampling may be the only means available for obtaining the needed information when the population appears to be infinite or is inaccessible such as the population of mountainous or thickly forested areas. In such cases, taking a complete census to collect data would neither be physically possible nor practically feasible.

(vii) Sampling has much smaller "*non-response*", following up of which is much easier. The term *non-response* means the non-availability of information from some sampling units included in the sample for any reason such as failure to locate or measure some of the units, refusals, not-at-home, etc.

(viii) Sampling is extensively used to obtain some of the census information.

(ix) The most important advantage of sampling is that it provides a valid measure of reliability for the sample estimates and this is one of the two basic purposes of sampling.

14.1.3 Sample Design and Sample Survey. A *sample design* is a definite statistical plan concerned with all principal steps taken in the selection of a sample and the estimation procedure. These steps are formulated in advance of conducting the sample. The term *survey* has been defined as a means of collecting information to meet a definite need. When a survey is carried out by a sampling method, it is called a *sample survey*. The *main steps* in a sample survey are to:

(i) clearly state the objectives of the survey;

(ii) define the population we wish to study as clearly as possible;

(iii) construct the sampling frame by clearly defining the sampling units;

(iv) choose an appropriate sample design and proper sample size;

(v) organize a reliable field work to achieve the objectives of the survey;

(vi) summarize and analyse the data.

14.1.4 Sampling Frame. A *sampling frame* is a complete list or a map that contains all the N sampling units in a population. A complete list of the names of all the students in the Government College, Lahore on, say, March 15 is the frame. A list of all households in a city, a map of a village showing all fields, etc. are other examples of the frame. The requirements of a reasonably good frame are that the frame should

(i) not contain inaccurate sampling units,

(ii) be complete and exhaustive, *i.e.* should contain all units and should cover the whole of the population,

(iii) be free from errors of omission and duplication of sampling units; and

(iv) be as *up-to-date* as possible at the time of use.

Most of the frames used for sample surveys do not meet all these requirements.

14.1.5 Probability and Non-probability Sampling. Sampling methods are broadly classified as: *Probability Sampling* and *Non-probability Sampling*. When each unit in a population has a known non-zero (not necessarily equal) probability of its being included in the sample, the sampling is said to be *probability sampling*. A probability sampling is also called *random sampling*. The major types of probability sampling are Simple random sampling, Stratified random sampling, Systematic sampling, Cluster sampling, etc. The advantage of probability sampling is that it provides a valid estimate of sampling error. Probability sampling is widely used in various areas such as industry, agriculture, business, etc.

A *non-probability* sampling, also called *non-random sampling*, is a process in which the personal judgement determines which units of the population are selected for a sample. The disadvantage of non-probability sampling is that the reliability of the sample results cannot be determined in terms of probability. Non-probability sampling techniques include Purposive sampling and Quota sampling.

14.1.6 Sampling With and Without Replacement. Samples may be selected with replacement or without replacement. Sampling is said to be *with replacement* when from a finite population a sampling unit is drawn, observed and then returned to the population before another unit is drawn. The population in this case remains the same and a sampling unit might be selected more than once. If, on the other hand, a sampling unit is chosen and not returned to the population after it has been observed, the sampling is said to be *without replacement*. Here the sampling units cannot be selected again for that sample as the units drawn are not replaced. Though the successive drawings become dependent, but for all practical purposes, they are considered as independent drawings. When sampling is performed with replacement, a finite population can theoretically be considered as an infinite population and a sample of *any size* can be drawn because the population is not exhausted. But when sampling is done without replacement, the sample size cannot be greater than the population size.

14.1.7 Sampling and Non-sampling Errors. A sample being only a part of a population cannot perfectly represent the population, no matter how carefully the sample is selected. This results in a difference between the value of sample statistic and the true value of the corresponding population parameter. Such a difference is called *sampling error* for that sample. If, for example, $\bar{x}$ is the mean obtained

from a sample of size n and μ is the corresponding population parameter, then the difference between $\bar{x}$ and μ is sampling error, that is

$$\text{sampling error} = \bar{x} - \mu$$

As the sample size increases, the sampling error is reduced, and in a complete enumeration (census), there is no sampling error as $\bar{x}$ becomes equal to μ. Sampling error is measured by what is known as *reliability* *"which is related to the variance of the sample statistic. The smaller the variance, the greater the reliability of the sample results will be".*

Aside from sampling errors which arise because a sample comprises only a portion of the population, there are erro which occur at the stages of gathering and processing of data, dless of whether a sample or a complete census is taken. These errors are called *non-sampling errors*. *Non sampling errors* include all kinds of *human errors*.- faulty sampling frame, biased method of selection of units, bias in response, non-response to mail questionnaires, errors of observation and measurement, processing errors such as errors in editing and coding, missclassification of observations, etc. These errors can be avoided through the proper selection of questionnaires, following up the non-response, proper training of the investigators, correct manipulation of the collected information, etc.

14.1.8 Sampling Bias. In survey sampling, the word *bias* means a systematic component of error which deprives a survey result of its representativeness. Bias is different from a random error in the sense that the random errors balance out in the long run while bias is cumulative and does not become less as the sample size increases. Bias is introduced by the following methods of selection:

(i) **Deliberate Selection.** Bias originates from deliberate selection which is based on personal judgement of what is representative.

(ii) **Substitution.** Sometimes it becomes difficult to make contact with certain members or information is not obtained from certain units, then we substitute members or units that are conveniently available. Such substitutions introduce bias.

(iii) **Incomplete Coverage.** Bias also emerges when we fail to cover the whole of the selected sample. For example, we select a sample through mail questionnaire. We get a biased sample when there are a large number of non-respondents to a survey.

(iv) **Haphazard Selection.** Haphazard human selection can also introduce bias as every human being has a tendency away from randomness in his choices.

(v) **Inadequate Interviewing.** Bias also enters when the interviewing is hasty, incomplete and misleading.

In order to draw valid conclusions, all possible sources of sampling bias must be avoided. This end is achieved if the sample is drawn entirely *at random*—a well defined concept in Statistics, meaning that every unit has a known non-zero probability of being included in the sample. A sample that is free from selection and procedural bias, is called a *good* or an *unbiased* sample. It is interesting to note that in some types of investigations or surveys, a certain amount of sampling bias is, however, tolerated.

14.1.9 Random Number Table. A table of random numbers contains a series of the digits 0, 1, 2,...,9, produced by a mechanical device which ensures that

(i) each of these digits has an equal probability of $\frac{1}{10}$ of occurrence,

(ii) each of the pairs of the digits 00,01, ..., 10, 11, ..., 99 has the same probability of $\frac{1}{100}$ of occurrence and so on.

The digits are then written in the order in which they occur and are grouped for convenience in blocks of two, four, five, etc. The widely used random number tables are:

(i) *Random Sampling Numbers*, constructed by Tippett, comprise 41 600 digits combined in groups of four, (Tracts for Computers, No. *XV*);

(ii) *Random Sampling Numbers*, published by Kendall and Smith, comprise 100,000 digits printed in twos and fours, (Tracts for Computers, No. *XXIV*);

(iii) *Random Numbers*, published by Fisher and Yates, comprise 15,000 digits arranged in twos, (Statistical Tables for Use in Biological, Agricultural and Medical Research);

(iv) *Random Numbers*, published by the *RAND* Corporation, (A Million Random Digits).

All these tables have been adequately tested to ensure that the basic requirements of equal probability and independence are satisfied, and are considered sufficiently random for selecting samples.

These tables are used by arbitrarily selecting a page, proceeding across a row, down a column or in any other direction, and taking digits in pairs, in groups of three, four or five as the size of the population may suggest. An extract from tables prepared by Fisher and Yates is given as Table 14.1 on page 8 as an illustration of their general appearance.

Table 14.1 Random Numbers

53 74 23 99 67	61 32 28 69 84	94 62 67 86 24	98 33 41 19 95
63 38 06 86 54	99 00 65 26 94	02 82 90 23 07	79 62 67 80 60
35 30 58 21 46	06 72 17 10 94	25 21 31 75 96	49 28 24 00 49
63 43 36 82 69	65 51 18 37 88	61 38 44 12 45	32 92 85 88 65
98 25 37 55 26	01 91 82 81 46	74 71 12 94 97	24 02 71 37 07
02 63 21 17 69	71 50 80 89 56	38 15 70 11 48	43 40 45 86 98
64 55 22 21 82	48 22 28 06 00	61 54 13 43 91	82 78 12 23 29
85 07 26 13 89	01 10 07 82 04	59 63 69 36 03	69 11 15 83 80
58 54 16 24 15	51 54 44 82 00	62 61 65 04 69	38 18 65 18 97
34 85 27 84 87	61 48 64 56 26	90 18 48 13 26	37 70 15 42 57
03 92 18 27 46	57 99 16 96 56	30 33 72 85 [illegible]	84 64 38 56 98
62 95 30 27 59	37 75 41 66 48	86 77 80 61 [illegible]	23 53 04 01 63
08 45 93 15 22	60 21 75 46 91	98 97 27 85 42	28 88 61 08 84
07 08 55 18 40	45 44 75 13 90	24 94 96 61 02	57 55 66 83 15
01 85 89 95 66	51 10 19 34 88	15 84 97 19 75	12 76 39 43 78
72 84 71 14 35	19 11 58 49 26	50 11 17 17 76	86 31 57 20 18
88 78 28 16 84	13 52 53 94 53	75 45 69 30 96	73 89 65 70 31
45 17 75 65 57	28 40 19 72 12	25 12 74 75 67	60 40 60 81 19
96 76 28 12 54	22 01 11 94 25	71 96 16 16 88	68 64 36 74 45
43 31 67 72 30	24 02 94 08 63	38 32 36 66 02	69 36 38 25 39
50 44 66 44 21	66 06 58 05 62	68 15 54 35 02	42 35 48 96 32
22 66 22 15 86	26 63 75 41 99	58 42 36 72 24	58 37 52 18 51
96 24 40 14 51	23 22 30 88 57	95 67 47 29 83	94 69 40 06 07
31 73 91 61 19	60 20 72 93 48	98 57 07 23 69	65 95 39 69 58
78 60 73 99 84	43 89 94 36 45	56 69 47 07 41	90 22 91 07 12
84 37 90 61 56	70 10 23 98 05	85 11 34 76 60	76 48 45 34 60
36 67 10 08 23	98 93 35 08 86	99 29 76 29 81	33 34 91 58 93
07 28 59 07 48	89 64 58 89 75	83 85 62 27 89	30 14 78 56 27
10 15 83 87 60	79 24 31 66 56	21 48 24 06 93	91 98 94 05 49
55 19 68 97 65	03 73 52 16 56	00 53 55 90 27	33 42 29 38 87
53 81 29 13 39	35 01 20 71 34	62 33 74 82 14	53 73 19 09 03
51 86 32 68 92	33 98 74 66 99	40 14 71 94 58	45 94 19 38 81
35 91 70 29 13	80 03 54 07 27	96 94 78 32 66	50 95 52 74 33
37 71 67 95 13	20 02 44 95 94	64 85 04 05 72	01 32 90 76 14
93 66 13 83 27	92 79 64 64 72	28 54 96 53 84	48 14 52 98 94
02 96 08 45 65	13 05 00 41 84	93 07 54 72 59	21 45 57 09 77
49 83 43 48 35	82 88 33 69 96	72 36 04 19 76	47 45 15 18 60
84 60 71 62 46	40 80 81 30 37	34 39 23 05 38	25 15 35 71 30
18 17 30 88 71	44 91 14 88 47	89 23 30 63 15	56 34 20 47 89
79 69 10 61 78	71 32 76 95 62	87 00 22 58 40	92 54 01 75 25

Table 14.1 is taken from Table XXXIII, Random Numbers (11) of Fisher and Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, published by Oliver & Boyd, Edinburgh and reproduced by permission of the authors and publishers.

These days, the Computers are programmed to generate random numbers. Such random numbers are called *pseudorandom numbers*.

14.2 PROBABILITY OR RANDOM SAMPLES

A sample is called a *random sample* if the probability of selection for each unit in the population is known prior to sample selection. The important kinds of random samples which differ in the manner in which the sampling units are selected, are discussed in the subsections that follow:

14.2.1 Simple Random Sample. A sample is defined to be a *simple random sample* (*SRS*) if it is selected in such a manner that (i) each unit in the population has an equal probability of being included in the sample and (ii) each possible sample of the same size has an equal probability of being the sample selected.

Suppose a finite population contains N units and a sample of n units is to be selected. If we sample with replacement, the number of all *possible* samples of size n that could be selected is N^n, as the first unit of the sample can be selected in N different ways, the second unit can also be selected in N ways and so on. When we sample without replacement, then the number of all possible samples when the *order* of the units is considered, is the number of *permutations* of n units from N, *i.e.* ${}^NP_n = N(N-1) \ldots (N-n+1)$. But in practical problems, we ignore the *order* in which the n units are drawn. Then the number of different samples of n units that can be selected when the *order* is disregarded, is the number of *combinations* of n units from a finite population of N units, *i.e.*, $\binom{N}{n} = \dfrac{N!}{n!\,(N-n)!}$. Thus there are $\binom{N}{n}$ samples that could be selected and these samples occur with equal probabilities.

As an illustration, suppose we wish to select random samples of size 2 from a population of say, 5 students, identified as A, B, C, D and E. If we sample *with replacement*, then there are $(5)^2 = 25$ possible samples, which are listed below:

AA	*BA*	*CA*	*DA*	*EA*
AB	*BB*	*CB*	*DB*	*EB*
AC	*BC*	*CC*	*DC*	*EC*
AD	*BD*	*CD*	*DD*	*ED*
AE	*BE*	*CE*	*DE*	*EE*

where the first letter corresponds to the student selected on the first draw and the second letter corresponds to the student chosen on the second draw.

We see that A appears in 9 of the 25 samples, so $P(A) = \frac{9}{25}$.

Similarly, $P(B) = P(C) = P(D) = P(E) = \frac{9}{25}$, showing that each unit of the population has an equal probability of being in the sample. Furthermore, each of the 25 samples has an equal probability of $\frac{1}{25}$ of being selected.

When the sampling is done *without replacement* and the *order* is disregarded, there are $\binom{5}{2} = 10$ possible distinct samples, which are listed below:

$$AB, AC, AD, AE, BC, BD, BE, CD, CE, DE$$

Now A appears in 4 of the 10 samples, so $P(A) = \frac{4}{10}$

Similarly, $P(B) = P(C) = P(D) = P(E) = \frac{4}{10}$, showing that each unit in the population has an equal probability $\left(\frac{4}{10}\right)$ and each of the 10 distinct samples has the same probability $\left(\frac{1}{10}\right)$ of being selected. A simple random sample is also known as *unrestricted random sample*. An important advantage of simple random sampling is that it provides unbiased estimates of the population mean, population totals and of sampling variance of the estimates.

Selection of Simple Random Sample. A simple random sample can be selected by the following methods:

(i) *Goldfish Bowl Procedure.* Allot to each unit in the population a different serial number from 1 to N and record each number on a card or a slip of paper. Place these numbered cards or the folded slips of paper in a bowl or a basket and mix them thoroughly. Then draw out blindly the desired number of cards or the folded slips of paper one by one for the sample, mixing thoroughly after each drawing. The population units corresponding to the numbers appearing on the selected cards or slips of paper are then included in the sample and the desired information is obtained. This method of random selection works well for small

populations but becomes practically impossible when the population size is quite large or infinite. This difficulty is overcome by a procedure, similar to drawing slips from a basket, that uses a random number table.

(ii) *Using a random Number Table.* Assign a number from 1 to N to each of the N units in the population. Consult a table of random numbers and select randomly a starting point in the table. Read digits in groups of two, three or more according to the largest number assigned to a unit in the population, from the table vertically, horizontally or diagonally. Record the number, discarding a number that is greater than N and that appears a second time if sampling is without replacement. Continue this process of selection until the desired sample size is reached.

For instance, if our population consists of 100 units, we may assign two-digit numbers from the range 00 to 99. (We start with 00 so that the last unit is assigned 99, a two-digit number). As the last number assigned to the population has two digits, so we read the random numbers consisting of two digits. If, for example, we read the numbers 53, 63, 35, 98, 02, etc. in the random number table, (Table 14.1; top left hand corner, reading down the column) we would include in the random sample those units of the population whose numbers are 2, 35, 53, 63, 98, etc.

Sampling from a given frequency distribution is performed by assigning a sampling number to each frequency and then drawing a random sample with the help of random numbers. Example 14.2 illustrates the procedure. In case of sampling from a given probability distribution, we construct a *cumulative probability distribution either* by calculating the probabilities associated with each value (or each class) of the variable *or* from the given probabilities. We then need to assign sampling numbers. To do this, we ignore the decimal point from the cumulative probabilities in order to have all (N) members to choose from. The range of sampling numbers to be used depends upon the number of decimal places used in the calculation of probabilities. The technique is demonstrated by Examples 14.3 and 14.4.

(iii) *Using a Computer.* Computer programs are available that provide random numbers. The sampling units corresponding to the random numbers are included in the sample.

Example 14.1. Assume that a population consists of 5 students and the marks obtained by them in a certain statistics class are 20, 15, 12, 16 and 18. Draw all possible random samples of two students when sampling is performed (i) with replacement, (ii) without replacement. Calculate the mean marks for each sample.

Let the five students be identified as A, B, C, D and E. Then (i) the number of possible random samples of 2 students which can be selected *with replacement* from this population is $(5)^2 = 25$. Let X_1 denote the marks of the student selected first and X_2, the marks of the student selected on the second draw. Then the possible random samples of size $n = 2$ with values of $\bar{X}$ are given below:

Sample No.	Sample Students	Sample Marks (X_1, X_2)	Sample Mean Marks $(\bar{X})$
1	A, A	20, 20	20
2	A, B	20, 15	17.5
3	A, C	20, 12	16
4	A, D	20, 16	18
5	A, E	20, 18	19
6	B, A	15, 20	17.5
7	B, B	15, 15	15
8	B, C	15, 12	13.5
9	B, D	15, 16	15.5
10	B, E	15, 18	16.5
11	C, A	12, 20	16
12	C, B	12, 15	13.5
13	C, C	12, 12	12
14	C, D	12, 16	14
15	C, E	12, 18	15
16	D, A	16, 20	18
17	D, B	16, 15	15.5
18	D, C	16, 12	14
19	D, D	16, 16	16
20	D, E	16, 18	17
21	E, A	18, 20	19
22	E, B	18, 15	16.5
23	E, C	18, 12	15
24	E, D	18, 16	17
25	E, E	18, 18	18

(ii) the number of random samples of 2 students that can be drawn *without replacement* is $\binom{5}{2} = 10$. These samples with values of mean marks are given below:

Sample No.	Sample Students	Sample Marks (X_1, X_2)	Sample Mean Marks $(\bar{X})$
1	*A, B*	20, 15	17.5
2	*A, C*	20, 12	16
3	*A, D*	20, 16	18
4	*A, E*	20, 18	19
5	*B, C*	15, 12	13.5
6	*B, D*	15, 16	15.5
7	*B, E*	15, 18	16.5
8	*C, D*	12, 16	14
9	*C, E*	12, 18	15
10	*D, E*	16, 18	17

Example 14.2. The following frequency distribution gives the ages of a population of 1,000 college students:

Age (X_i)	14	15	16	17	18	19	20	Total
No. of Students	6	61	270	491	153	15	4	1,000

Using a random number table, select a simple random sample of 20 students. Find the sample mean age and compare it with the population mean age.

First of all we assign a number to each of the 1,000 students from the range 000 to 999. (We start with 000 so that the last student is assigned 999, a three digit number). The 6 students corresponding to the first class ($x = 14$) are assigned the numbers 000, 001, 002, 003, 004 and 005. The next class ($x = 15$) has 61 students to whom we assign the next 61 numbers 006 to 066 inclusive, and so forth. The sampling numbers can conveniently be assigned by compiling a cumulative frequency column. The assigned numbers are shown in column 3 in the following table:

Age (X_i)	No. of Students	Assigned Numbers	f_iX_i
14	6	000 – 005	84
15	61	006 – 066	915
16	270	067 – 336	4320
17	491	337 – 827	8347
18	153	828 – 980	2754
19	15	981 – 995	285
20	4	996 – 999	80
Σ	1,000	----	16,785

Next, we consult a table of random numbers to select a sample of 20 college students at random by finding 20 three-digit numbers between 000 and 999. Starting from the top left hand corner of Table 14.1, we find, going down the columns, the following 20 numbers: 537, 633, 353, 634, 982, 026, 645, 850, 585, 348, 039, 629, 084, 070, 018, 728, 887, 451, 967 and 433. Our sample students are those who correspond to these numbers. Thus the corresponding ages are: 17, 17, 17, 17, 19, 15, 17, 18, 17, 17, 15, 17, 16, 16, 15, 17, 18, 17, 18 and 17. These ages are grouped in the following frequency distribution.

Frequency Distribution of Sample Data

Age (x_i)	Tally	f_i	f_ix_i
15	III	3	45
16	II	2	32
17	~~IIII~~ ~~IIII~~ I	11	187
18	III	3	54
19	I	1	19
Σ	---	20	337

Hence the sample mean age, $\bar{x} = \frac{1}{n} \Sigma f_i x_i = \frac{337}{20} = 16.85$ years, and the population mean age, $\mu = \frac{1}{N} \Sigma f_i X_i = \frac{16785}{1000} = 16.78$ years.

Example 14.3. Select a random sample of size 15, using a random number table, from a Poisson distribution with parameter $\mu = 3$.

Given X has a Poisson distribution with

$$P(X = x) = \frac{e^{-3}(3)^x}{x!}, \text{ for } x = 0, 1, 2, 3, \ldots$$

We first calculate the probabilities associated with each value of x.

Now $P(X = 0) = e^{-3} = 0.0498$, and using the Poisson recurrence formula, we obtain probabilities for other values of x. We find the cumulative probabilities $P(X \leq x)$ and ignoring the decimal point, we assign 3-digit sampling numbers from the range 000 to 999, as 3 decimal places are used in calculating probabilities. The probabilities and the assigned numbers are shown below:

x	$P(X = x)$	Cumulative $P(X \leq x)$	Assigned Numbers
0	0.050	0.050	000 – 049
1	0.149	0.199	050 – 198
2	0.224	0.423	199 – 422
3	0.224	0.647	423 – 646
4	0.168	0.815	647 – 814
5	0.101	0.916	815 – 915
6	0.050	0.966	916 – 965
7	0.022	0.988	966 – 987
8+	0.012	1.000	988 – 999

We now consult a table of random numbers to select a sample of 15 by finding 15-three digit numbers. Let us select three columns, say columns 21, 22 and 23 of Table 14.1, page 8. Then going down the three columns we select the first 15 numbers. These numbers and the x-values (shown in brackets) corresponding to them, are listed below:

946 (6)	381 (2)	303 (2)
028 (0)	615 (3)	869 (5)
252 (2)	596 (3)	987 (7)
613 (3)	626 (3)	249 (2)
747 (4)	901 (5)	158 (1)

The sample results are shown in the following table:

x	0	1	2	3	4	5	6	7
$f(x)$	1	1	4	4	1	2	1	1

Example 14.4. With the help of random numbers, draw a sample of 25 from a Normal distribution with $\mu = 60$ and $\sigma = 2.5$.

For a normal distribution, the range of classes is $\mu \pm 3\sigma$. Here $\mu = 60$ and $\sigma = 2.5$, therefore the range for the classes is $60 \pm 3(2.5)$, *i.e.* 52.5 to 67.5. Using 8 classes, each with a class-interval $h = 2$, the probabilities have been calculated (see Example 9.21 page 407 of Part I).

We assign 4-digit sampling numbers from the range 0000 to 9999 as the probabilities have 4 decimal places (the decimal point ignored while assigning numbers). The classes, the cumulative probabilities and the assigned numbers are shown below:

Classes	*ucb*	z	$P(Z<z)\ \Phi(z)$	Assigned Numbers
Upto 52.5	52.5	−3.0	0.0013	0000 – 0012
52.5–54.5	54.5	−2.2	0.0139	0013 – 0138
54.5–56.5	56.5	−1.4	0.0808	0139 – 0807
56.5–58.5	58.5	−0.6	0.2743	0808 – 2742
58.5–60.5	60.5	+0.2	0.5793	2743 – 5792
60.5–62.5	62.5	1.0	0.8413	5793 – 8412
62.5–64.5	64.5	1.8	0.9641	8413 – 9640
64.5–66.5	66.5	2.6	0.9953	9641 – 9952
Over 66.5	∞	∞	1.0000	9953 – 9999

We now select a random sample of size 25 by finding 25 four-digit numbers. For this purpose we use the four columns 11–14 of Table 14.1, page 8. The random numbers and the classes to which each sample value corresponds, are listed below:

6132 (60.5–)	7150 (60.5–)	5799 (60.5–)	1911 (56.5–)	6606 (60.5–)
9900 (64.5–)	4822 (58.5–)	3775 (58.5–)	1352 (56.5–)	2663 (56.5–)
0672 (54.5–)	0110 (52.5–)	6021 (60.5–)	2840 (58.5–)	2322 (56.5–)
6551 (60.5–)	5154 (58.5–)	4544 (58.5–)	2201 (56.5–)	6020 (60.5–)
0191 (54.5–)	6148 (60.5–)	5110 (58.5–)	2402 (56.5–)	4389 (58.5–)

Finally, we arrange this information in a table and get the sample given by:

Classes	52.5–,	54.5–,	56.5–,	58.5–,	60.5–,	62.5–,	64.5–,
f	1	2	6	7	8	0	1

14.2.2. Stratified Random Sample. A sample of size n is defined to be a *stratified random sample* if it is selected from a population which has been divided into a number of non-overlapping groups or sub-populations, called *strata*, such that part of the sample is drawn at random from *each stratum*. Stated differently, let a heterogeneous population containing N units be divided into k subpopulations or strata of sizes $N_1, N_2, ..., N_k$ in such a manner that all units in each stratum are believed to be very similar with respect to the measurements of interest. Then a *stratified random sample* of size n will be composed of the simple random samples of predetermined sizes $n_1, n_2, ..., n_k$ ($\Sigma n_i = n$) drawn independently from the strata 1, 2, ..., k respectively. It is to be emphasized that good stratification requires that each of these strata should be internally homogeneous but externally the strata should differ from one another.

A population may be stratified according to the size such as large, medium and small; to the administrative grouping such as provinces and districts; to the geographic area such as urban and rural; to the natural characteristics such as age-group, sex, family size, occupation, education, tastes of the consumers, etc. The advantages of stratified random sampling are low cost, greater accuracy and a better coverage. Stratified random sampling is used when (i) the variations among strata are greater than the variations within strata, (ii) information about some parts of the population is desired.

The purpose of stratified random sampling is three-fold. *Firstly*, the strata obtained by subdividing the heterogeneous population into homogeneous groups, adequately represent the population so the information concerning individual stratum is gathered. *Secondly*, it provides improved estimates of the population characteristics. *Thirdly*, it reduces the variance of the estimator.

Allocation of Sample Sizes. By *allocation* of a sample we mean the way the total sample size n is distributed among the various strata into which the population has been divided. Four methods of allocating the sample numbers are available. They are:

(a) *Equal Allocation*. The allocation is called *equal* when from each stratum, equal number of sampling units is selected. That is the total sample size n is distributed equally among all the k strata. Thus the stratum sample size n_i for equal allocation is

$$n_i = \frac{n}{k}, \text{ for } i = 1, 2, ..., k$$

This is the simplest method of allocation.

(b) *Proportional Allocation*. The allocation is said to be *proportional* when the total sample size n is distributed among the different strata in proportion to the sizes of strata. In other words, the allocation is proportional if

$$n_i = n \cdot \frac{N_i}{N}, \text{ for } i = 1, 2, \ldots, k$$

where N_i is the population size of the *ith* stratum, n_i is the *ith* stratum sample size and N is the total size of the population. A sample of size n_i from stratum i is drawn by random numbers and investigated. This way of allocation is the next simplest method and it is the most frequently used method. The advantage of proportional allocation is that it does not require information either on the stratum variance or on the costs of sampling units in different strata.

(c) *Optimum Allocation*. The allocation is called *optimum* when the total sample size n is allocated among the different strata in such a way that for a given cost of selecting the sample, the variance of the estimated mean $\bar{X}_{st}$, *i.e.* $\text{Var}(\bar{X}_{st})$ is minimized. The stratum sample size n_i for this method of allocation is

$$n_i = n \cdot \frac{N_i\sigma_i / \sqrt{c_i}}{\sum N_i\sigma_i / \sqrt{c_i}},$$

where N_i is the population of the *ith* stratum,

σ_i is the stratum standard deviation, and

c_i is the cost of surveying one unit in the *ith* stratum.

This method of allocation tells us that we should select a larger size sample from a given stratum if

(i) the stratum comprises a larger number of units, *i.e.* N_i is larger,

(ii) the variation within the stratum is greater, *i.e.* σ_i is bigger,

(iii) the sampling units in the stratum are less costly to measure, *i.e.* c_i is smaller.

When information on the stratum standard deviation, σ_i is not available, we either estimate it by S_i or we may use the proportional allocation.

(d) *Neyman Allocation*. This method of allocation was proposed by *J.* Neyman (1894–1981) in 1934 and it consists of finding n_i which minimizes the variance of the stratified sample mean for a fixed total sample size n, assuming the costs of surveying the units, *i.e.* c_i to be the same in all strata. The stratum sample size n_i is given by the relation

$$n_i = n \cdot \frac{N_i\sigma_i}{\sum N_i\sigma_i}, \text{ for } i = 1, 2, \ldots, k.$$

Neyman allocation becomes exactly the same as the proportional allocation when stratum standard deviations are equal.

Example 14.5. Suppose a population of $N=9$ is stratified into 3 strata with the following measurements:

Stratum I	$X_{11} = 1, X_{12} = 2, X_{13} = 4$
Stratum II	$X_{21} = 6, X_{22} = 8$
Stratum III	$X_{31} = 11, X_{32} = 15, X_{33} = 16, X_{34} = 19$

If two measurements are drawn from each stratum for the sample, state how many samples of size 6 could be chosen from this population? List these samples and compute the mean for each sample.

Here the population consists of 3 strata and from each stratum 2 units are to be selected to make up a sample of $n=6$. Assuming the sampling without replacement, we can choose $\binom{3}{2} = 3$ possible subsamples from the first stratum, $\binom{2}{2} = 1$ possible subsample from the second stratum and $\binom{4}{2} = 6$ possible subsamples from the third stratum. Each of the possible subsamples from stratum I is to be associated with the subsample from stratum II and then these combinations are further associated with each of the possible subsamples from stratum III. Hence, in all there are $3 \times 1 \times 6$, *i.e.* 18 possible samples of size 6 with 2 measurements from each stratum. The first sample consists of (1, 2, 6, 8, 11, 15) of which (1, 2) is from stratum I, (6, 8) is from stratum II and (11, 15) is from stratum III. The 18 possible samples are listed below and the sample means appear in the last column.

Sample No.	Sample data from stratum			Total	Sample Means
	I	II	III		
1	1, 2	6, 8	11, 15	43	7.17
2	1, 2	6, 8	11, 16	44	7.33
3	1, 2	6, 8	11, 19	47	7.83
4	1, 2	6, 8	15, 16	48	8.00
5	1, 2	6, 8	15, 19	51	8.50
6	1, 2	6, 8	16, 19	52	8.67
7	1, 4	6, 8	11, 15	45	7.50
8			11, 16	46	7.67
9			11, 19	49	8.17
10			15, 16	50	8.33
11			15, 19	53	8.83
12			16, 19	54	9.00
13	2, 4	6, 8	11, 15	46	7.67
14			11, 16	47	7.83
15			11, 19	50	8.33
16			15, 16	51	8.50
17			15, 19	54	9.00
18			16, 19	55	9.17

Example 14.6. Select a stratified random sample of size $n=5$ by proportional allocation from the following population. Find the sample mean and the estimate of the population mean.

Stratum I	$X_{11} = 12, X_{12} = 14, X_{13} = 13, X_{14} = 8.$
Stratum II	$X_{21}=25, X_{22}=30, X_{23}=40, X_{24}=35, X_{25}=24, X_{26}=28$

To select a stratified random sample of $n=5$ by using proportional allocation, *i.e.* $n_i = n \cdot \frac{N_i}{N}$, the sample size is allocated as

$$n_1 = n \cdot \frac{N_1}{N} = 5 \times \frac{4}{10} = 2, \text{ and}$$

$$n_2 = n \cdot \frac{N_2}{N} = 5 \times \frac{6}{10} = 3.$$

Using a table of random numbers, we select the following subsamples from

Stratum I: $X_{11} = 12,\ X_{12} = 14,$

Stratum II: $X_{22} = 30,\ X_{23} = 40,\ X_{25} = 24.$

Now the sample mean of sample n is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{x}_i = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$$

$$= \frac{1}{5}[12 + 14 + 30 + 40 + 24] = \frac{120}{5} = 24$$

An *estimator* (a term to be defined later) of the population mean by the stratified sampling, denoted by $\bar{X}_{st}$, is given by

$$\bar{X}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{X}_i = \sum_{i=1}^{k} W_i \bar{x}_i,$$ where $\bar{x}_i$ is the *ith* sub-sample mean and $W_i (= N_i/N)$ is termed the *weight* of the *ith* stratum.

Hence the *estimate* of the population mean is

$$\bar{x}_{st} = \frac{N_1}{N}\bar{x}_1 + \frac{N_2}{N}\cdot\bar{x}_2 \text{ where } \bar{x}_1 = \frac{12 + 14}{2} = 13, \text{ etc.}$$

$$= \frac{4}{10}(13) + \frac{6}{10}\left(\frac{94}{3}\right) = \frac{240}{10} = 24.$$

It should be noted that the estimate of the population mean, $\bar{x}_{st}$, is not generally the same as the sample mean of sample n, *i.e.* the overall sample mean, $\bar{x} = \frac{1}{n}\sum n_i \bar{x}_i$ of the stratified random sample, except when $\frac{n_i}{n} = \frac{N_i}{N}$.

14.2.3 Systematic Random Sample. A sample of size n is defined to be a *systematic random sample* if it is obtained by choosing one unit at random from the first k units and thereafter selecting every kth unit after the N units in the population have been serially numbered from 1 to N or arranged in a systematic fashion. The letter k, called the *sampling interval*, stands for some integer nearest to $\frac{N}{n} = \frac{\text{Population size}}{\text{Sample size}}$ and the sample is generally expressed by saying "a 1 in k sample". For example, when the number of the unit selected at random from the first k units is i ($i = 1, 2, \ldots, k$), the systematic sample of size n will contain the units with numbers $i, i + k, i + 2k, \ldots, i + (n - 1)k$. Suppose that $i = 7$ and $k = 20$, then a systematic sample consisting of every 20th unit, will be composed of the units numbered 7, 27, 47, 67 and so on. Thus in systematic sampling, sampling units are selected at uniform interval after a random start.

When the sampling interval corresponds to some periodic or cyclic characteristic in the population, the systematic sampling will result in a non-representative sample. For example, suppose every 20th shop in a big bazaar is a corner shop and the sampling interval is also 20. If the random start coincides with a corner shop, then the sample will include all the corner shops, and the sample will be highly non-representative as different characteristics (say, dealing in beverages) are associated with the corner shops. One way of avoiding of this sort of periodicity is to take a *fractional* sampling interval and then to round off the numbers obtained, *e.g.* if sampling interval is 20.7 and we start at 7, then the subsequent numbers are 27.7, 48.4, 69.1, 89.8, etc. and the rounding off results in 28, 48, 69, 90, etc. The sample will then consist of the units with serial numbers 7, 28, 48, 69, 90, etc.

If we think of the population as being divided into n strata, each consisting of k units, then a systematic sample resembles a stratified random sample with one unit from each stratum. The advantages of systematic sampling are that it saves much time and effort, it is economical, it is easily selected and conveniently worked out. Furthermore, a systematic sample is more representative of the population sampled as the sample is more evenly distributed across the population.

14.2.4. Cluster Sample. A random sample is said to be a *cluster sample* if it consists of *first* selecting at random groups of individual units, called *clusters* (treated as sampling units) into which a population can be divided and *then* including in the sample either all the units from each of the chosen clusters, or selecting a random sample of the units which the cluster comprises. In other words, suppose that a population is first divided into M smaller groups (equal or unequal in size), called *clusters*, such as the blocks of the cities, households, classes, etc. A random sample consisting of a number of clusters (say, m) is then selected from all these clusters, where each chosen cluster is either subsampled or all the subunits in the sampled clusters are included in the sample. Such a sample is called a *cluster sample*. Cluster sampling requires that the clusters should be as internally dissimilar as possible and different clusters should be very similar.

The procedure is called *one-stage cluster sampling* when all the units which each of the sampled clusters comprises, are included. If each of the sampled clusters is subsampled, then the sampling plan is called *two-stage cluster sampling* or *subsampling*. The plan is called *multistage cluster sampling* when more than two stages are involved in taking the sample. When the clusters relate to geographical areas, the sampling is known as *area sampling*.

As each cluster is treated as a single sampling unit in the selection process, the clusters are therefore called the *primary sampling units* (*psu*), while the subunits composing a cluster, are called the *secondary sampling units* (*ssu*).

The advantage of cluster sampling is savings in cost and time, *i.e.* the cost of sample selection and travel expenses of interviewers are considerably low. Cluster sampling is used when (i) the sampling frames of adequate coverage are not available, (ii) the variations among clusters are smaller than the variations within clusters. It is to be noted that cluster sampling is mostly used in statistical quality control.

14.2.5. Multistage Sample. A sample is called a *multistage sample* when it is selected in stages, the sampling units at each stage being subsampled from the larger units selected at the previous stage. Here a population is divided into a number of units, called *first-stage units*, which are subsampled. Each of the selected second stage units, is further divided into *third-stage units*, from which a subsample is again selected and so on. For example, in a sample survey, we select a random sample of n_1 districts (first-stage units), then we take a subsample of n_2 villages (second-stage units) from each of the selected districts, again, we select a sub-sample of n_3 households (third stage units) from each of the selected villages and so on. In a multistage sample, the sample size is the

number of units included in the sample at the final stage in the sampling. The advantage of multistage sampling is that only the parts of the population chosen at any stage need to be listed for sampling at the next stage. This technique is commonly used in large-scale surveys.

14.2.6. Multi-phase Sample. A sample is said to be a *two-phase sample* when certain items of information are collected from all the units in a sample and other items of usually more detailed information are collected from a subsample of the units composing the original sample. Two -phase sampling, also called *Double Sampling*, was proposed in 1938 for the first time by Neyman. When more phases are added, it becomes *multi-phase sampling*. It is important to note that in multi-phase sampling, the *same* units are used at each phase, whereas in multi-stage sampling, the units are different at different stages of sampling. The advantages of multi-phase sampling are that it can prove cheaper and it reduces the burden on respondents.

14.2.7. Sequential Sampling. This is another method of sampling where the sample size is not fixed in advance but sampling units are drawn on by one or in lots, and the decision is based on a definite rule relating to the sampling units themselves. That is, we draw one unit at a time and after each drawing, we make a decision whether to accept the lot or group, whether to reject it or whether to continue sampling. A graphic or tabular procedure is generally used to find when sampling should terminate. This technique was developed during World War II by Abraham Wald (1902–50) and is good for reaching decisions rapidly.

14.3 NON-PROBABILITY SAMPLE

A non-probability sample or a non-random sample is also called a *judgement sample*. The important types of a non-random sample are the *purposive sample* and the *quota sample*. They are briefly described in the subsections that follow.

14.3.1. Purposive Sample. A purposive sample is a non-random sample in which the selection of the sampling units is based on a person's *expertise* about the population. A purposive sample is liable to bias to be introduced by the deliberate subjective choice of the person who selects the sample. As the purposive sampling is not based on probability theory, there is therefore no objective method for measuring the reliability of the sample results, and hence the information gathered from such a sample cannot be made a basis for statistical inference.

The purposive sampling in spite of these obvious drawbacks is in several situations preferred over probability sampling and gives quite satisfactory results. For instance, when taking a sample of melons from a

truck-load, the sampler inspects the whole load and then selects according to his expert subjective judgement, those melons which he considers to be representative. Purposive sampling can also be appropriate when a population contains relatively few large units whose characteristics are known. Its main use, however, is in Economic and Business statistics.

14.3.2. Quota Sample. A quota sample is a type of judgement sample. It is a sample, usually of human being, in which the information is collected purposively from the segments of a population (*the quotas*), *e.g.* the quotas of men and women; urban and rural; upper, middle and lower income groups; etc. These factors are termed *quota controls*. They are intended to make the sample as representative as possible and to reduce sampling bias that creeps in because the selection of respondents within the quotas depends on the personal choice of the interviewers. Interviewers being human, are likely to look for persons who either share similar opinions or are personally known to them or are conveniently located.

Quota sampling may be considered as stratified sampling in which the selection of units within strata is non-random. The advantages of quota sampling are that it is cheaper, it is easy administratively and it is a very quick form of investigation. Quota sampling is widely used in *public opinion polls* and *market research surveys*.

14.4 SAMPLING DISTRIBUTIONS

A sampling distribution is defined as a probability distribution of the values of a *statistic* such as a mean, a standard deviation, a proportion, etc, computed from all possible samples of the same size, which might be selected with or without replacement from a population. As a sampling distribution of a statistic is a probability distribution, therefore the sum of all probabilities in it is always equal to one; and the distribution has its own mean and its own standard deviation. The values of the statistic computed from *one or more* samples actually selected from the population and the sampling distribution of the statistic provide all the information one needs in making decisions about the values of the population parameters. There are many types of sampling distributions but the most frequently used types in statistical inference are the binomial, the normal, the t-distribution, the chi-square distribution, and the F distribution. A sampling distribution should not be confused with a *sample distribution* which is the distribution of individual values of a single sample.

Standard Error. The standard deviation of a sampling distribution of a sample statistic is called the *standard error* (abbreviated to *S.E.*) *of the statistic*. The standard error thus measures the dispersion of the values of a statistic, that might be computed from all possible samples, whereas the standard deviation of a population (or sample) measures the dispersion of the values of the population (sample) units about the population (sample) mean.

14.4.1. Sampling Distribution of the Mean. The sampling distribution of the mean is the probability distribution or the relative frequency distribution of the means $\bar{X}$ of all possible random samples of the same size that *could* be selected from a given population. The mean of this distribution is represented by $\mu_{\bar{x}}$ and the standard deviation, which is called the *standard error of the mean*, by $\sigma_{\bar{x}}$ or *S.E.* $(\bar{X})$. The value $\sigma_{\bar{x}}$ indicates the spread in the distribution of all possible sample means.

The sampling distribution of $\bar{X}$ has the following properties.

(i) The mean of the sampling distribution of the mean (equivalently, the mean of all possible sample means) is equal to the population mean, that is $\mu_{\bar{x}} = \mu$, regardless of whether sampling is done with replacement or without replacement.

Proof. Let us first consider sampling *without replacement* from a finite population of size N. The number of distinct simple random samples of size n that can be selected without replacement from a population of size N is $\binom{N}{n} = k$, say. Let $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_i, \ldots, \bar{X}_k$ be the means of $k = \binom{N}{n}$ possible random samples of size n, where $\bar{X}_i$ is the mean of the *ith* sample. Then the mean of the sampling distribution of $\bar{X}$ (equivalently, the mean of all possible sample means), denoted by $\mu_{\bar{x}}$, is

$$\mu_{\bar{x}} = \frac{\bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_k}{k}$$

$$= \frac{1}{k}\left[\left(\frac{X_1+X_2+\ldots}{n}\right) + \left(\frac{X_1+X_3+\ldots}{n}\right) + \ldots + \left(\frac{X_2+X_3+\ldots}{n}\right) + \ldots\right]$$

In order to simplify the expression on the right, we find out the number of samples that contain any specified value X_i. The number of such samples is $\binom{N-1}{n-1}$, that is, the number of ways in which the $(n-1)$ other units in the sample are to be selected from the remaining $(N-1)$ units.

Next, we determine the co-efficient of the value X_i by collecting all the terms in the expression containing X_i. Thus the co-efficient of X_i is

$$\frac{\binom{N-1}{n-1}}{k} \cdot \frac{1}{n} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \frac{1}{n} = \frac{(N-1)!\,(N-n)!\,n!}{(n-1)!\,(N-n)!\,N!} \cdot \frac{1}{n} = \frac{1}{N}$$

$$\text{Hence } \mu_{\bar{x}} = \frac{X_1}{N} + \frac{X_2}{N} + \ldots + \frac{X_i}{N} + \ldots + \frac{X_N}{N}$$

$$= \frac{X_1 + X_2 + \ldots + X_i + \ldots + X_N}{N}$$

$= \mu$, mean of the population.

Sampling with replacement. Let $X_1, X_2, \ldots, X_n$ be the observations of a simple random sample of size n from a population having N observations. Then a specified X_i taken from the population, could be any one of the N values with an equal probability of $\frac{1}{N}$ as all the values are equally likely. Thus X_i is a random variable and therefore

$$E(X_i) = \frac{1}{N} \sum_{k=1}^{N} X_k = \mu.$$

For repeated sampling, the mean of a sample $\bar{X} = \frac{1}{n} \sum X_i$ varies from sample to sample, therefore

$$E(\bar{X}) = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n}\left[E(X_1) + E(X_2) + \ldots + E(X_n)\right]$$

$$= \frac{1}{n}\left[\mu + \mu + \ldots + \mu\right]$$

$$= \frac{1}{n}\left[n\mu\right] = \mu, \text{ the population mean.}$$

(ii) The standard deviation of the sampling distribution of the mean is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}, \quad (\sigma = \text{population } s.d.)$$

when sampling is performed without replacement from a finite population of size N, *or*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

when sampling is done with replacement from a finite population or sampling from an infinite population.

Proof. The variance of $\bar{X}$, denoted by $\sigma^2_{\bar{x}}$ or $\text{Var}(\bar{X})$ is defined as

$$\text{Var}(\bar{X}) = E\,[\bar{X} - E(\bar{X})]^2 = E\,[\bar{X} - \mu]^2$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right]^2$$

$$= \frac{1}{n^2} \cdot E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i \neq j}^{n}(X_i - \mu)(X_j - \mu)\right]$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] + \frac{1}{n^2} E\left[\sum_{i \neq j}^{n}(X_i - \mu)(X_j - \mu)\right]$$

The simplification depends on whether the sampling is performed without replacement from a finite population of size N or sampling is done with replacement. The two cases are treated separately.

First case: *Sampling without replacement.*

Since the probability of obtaining $(X_i - \mu)^2$ on the *ith* draw is equal to the probability of obtaining X_i on the *ith* draw which is $\frac{1}{N}$, therefore the expected value of $(X_i - \mu)^2$ becomes σ^2, *i.e.*

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

Again, since the sampling is without replacement, the probability of selecting $(X_i - \mu)(X_j - \mu)$ on the *ith* and *jth* draw is $\frac{1}{N} \cdot \frac{1}{N-1}$, because they are not independent on account of the reduction in size from N to $N-1$. Thus

$$E(X_i-\mu)(X_j-\mu) = \frac{1}{N}\cdot\frac{1}{N-1}\sum_{i\neq j}^{N}(X_i-\mu)(X_j-\mu)$$

$$= \frac{1}{N(N-1)}\left\{\left[\sum_{i=1}^{N}(X_i-\mu)\right]^2 - \sum_{i=1}^{N}(X_i-\mu)^2\right\}$$

$$= \frac{1}{N(N-1)}\left\{0 - \sum_{i=1}^{N}(X_i-\mu)^2\right\}$$

$$= \frac{-\sigma^2}{N-1}$$

Substituting these values, we get

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + \frac{1}{n^2}\sum_{i\neq j}^{n}\frac{-\sigma^2}{N-1}$$

$$= \frac{1}{n^2}(n\sigma^2) - \frac{1}{n^2}\cdot\frac{1}{N-1}\sum_{i\neq j}^{n}\sigma^2$$

$$= \frac{\sigma^2}{n} - \frac{1}{n^2}\frac{1}{N-1}\cdot n(n-1)\sigma^2$$

$$= \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}.$$

Hence $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$.

The factor $\frac{N-n}{N-1}$ is usually called the *finite population correction* (*fpc*) or *finite correction factor* (*fcf*) for the variance because in sampling from finite population, the variance of the mean is reduced by this amount. It is important to note that in sampling without replacement from a finite population of size N, *fpc* is dropped from the formula whenever n, the sample size, is less than 5% of N; and *fpc* is used when n is 5% or greater than 5% of N.

Second case : *Sampling with replacement*

When sampling is done with replacement or sampling from an infinite population, the X_i and X_j are statistically independent. Therefore

$$E(X_i - \mu)(X_j - \mu) = 0. \text{ Hence we get}$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n},$$

or $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\text{population standard deviation}}{\text{square root of sample size}}.$$

It is to be noted that the standard error of the mean is always less than the standard deviation of the population. This means that the sampling distribution of the mean has less variability than the population from which the samples were taken. If the value of σ is not known and if the sample size is *large* (as a rule of thumb adopted by many authors, a sample containing 30 or more observations constitutes a *large* or *sufficiently large* sample), it is replaced by s, the standard deviation of the sample. The *S.E.* of the mean then becomes

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

(iii) **Shape of the distribution.** (a) If the population sampled is normally distributed, then the sampling distribution of the mean $\bar{X}$, will also be normal regardless of sample size.

To prove this, we proceed as follows:

By definition, the moment generating function of $\bar{X}$ is

$$M_{\bar{X}}(t) = E(e^{t\bar{X}}) = E(e^{\Sigma tX_i/n})$$

$$= E\left[\prod_{i=1}^{n} e^{tX_i/n}\right] = \prod_{i=1}^{n} E(e^{tX_i/n}).$$

But $$E(e^{tX_i/n}) = M_X\left(\frac{t}{n}\right).$$

If X is $N(\mu, \sigma^2)$, then

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}, \text{ and}$$

$$M_X\left(\frac{t}{n}\right) = e^{\mu t/n + \frac{1}{2}\sigma^2 (t/n)^2}$$

Since $X_1, X_2, \ldots, X_n$ is a random sample, therefore

$$M_{\bar{X}}(t) = E\left(e^{t\bar{X}}\right) = \left(E\left(e^{tX_i/n}\right)\right)^n$$

$$= \left[M_X\left(\frac{t}{n}\right)\right]^n = e^{n\left(\frac{\mu t}{n} + \frac{1}{2}\sigma 2t2/n2\right)}$$

$$= e^{\mu t + \sigma^2 t^2/2n}$$

But this is the *m.g.f.* of a normal distribution with mean = μ and variance = $\frac{\sigma^2}{n}$. Thus $\bar{X}$ is normally distributed variable with mean μ and variance σ^2/n where μ and σ^2 are the mean and variance of the population.

(b) If the population sampled is non-normal, then for sufficiently large sample size, the sampling distribution of $\bar{X}$ will approximate the normal distribution.

This is a special case of the most important statistical theorem, known as the *Central Limit Theorem*, which is stated and proved in the next section.

We know that the standardized form of a random variable is obtained by subtracting its mean from it and dividing the difference by its standard deviation, that is

$$Z = \frac{\text{value of random variable} - \text{mean of random variable}}{\text{standard deviation of random variable}}.$$

We have proved above that the sample mean $\bar{X}$ is normally distributed random variable with mean equal to population mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. The standard normal variable then becomes

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

If sampling is without replacement and sample size n is 5% or greater than 5 per cent of the population size N, then Z values are obtained by the formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$

The sampling distribution of $\bar{X}$ thus offers solutions to probability questions about the values of the sample means.

Example 14.7. Assume that a population consists of 7 similar containers having the following weights (kilograms):

9.8, 10.2, 10.4, 9.8, 10.0, 10.2 and 9.6.

(a) Find the mean μ and the standard deviation σ of the given population.

(b) Draw random samples of 2 containers without replacement and calculate the mean weight $\bar{X}$ of each sample.

(c) Form a frequency distribution of $\bar{X}$ and a sampling distribution of $\bar{X}$.

(d) Find the mean and the standard deviation of the sampling distribution of $\bar{X}$.

(a) The population mean μ and standard deviation σ are

$$\mu = \frac{\Sigma X}{N} = \frac{9.8 + 10.2 + \ldots + 9.6}{7} = \frac{70.0}{7} = 10.0 \text{ kg; and}$$

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}} = \sqrt{\frac{(9.8-10)^2+(10.2-10)^2+\ldots+(9.6-10)^2}{7}}$$

$$= \sqrt{\frac{0.48}{7}} = \sqrt{0.0686} = 0.262 \text{ kg.}$$

(b) Let the containers be identified as A, B, C, D, E, F and G. Now the number of possible random samples of $n = 2$ containers without replacement is $\binom{7}{2} = 21$. The 21 possible random samples with the values of their mean weights are given on page 32:

(c) The frequency distribution of $\bar{X}$ and the sampling distribution of the mean $\bar{X}$, which is just the relative frequency distribution of $\bar{X}$, are obtained below:

(i) Frequency Distribution of $\bar{X}$

Sample Mean $\bar{x}$	Tally	f
9.7	II	2
9.8	II	2
9.9	IIII	4
10.0	~~IIII~~	5
10.1	IIII	4
10.2	II	2
10.3	II	2
Σ	--	21

(ii) Sampling Distribution of $\bar{X}$

Sample Mean $\bar{x}$	Probability $f(\bar{x})$
9.7	2/21
9.8	2/21
9.9	4/21
10.0	5/21
10.1	4/21
10.2	2/21
10.3	2/21
Σ	1

Sample No.	Sample Combination	Weights in Samples X_1	X_2	Sample Mean weight ($\bar{X}$)
1	A, B	9.8,	10.2	10.0
2	A, C	9.8,	10.4	10.1
3	A, D	9.8,	9.8	9.8
4	A, E	9.8,	10.0	9.9
5	A, F	9.8,	10.2	10.0
6	A, G	9.8,	9.6	9.7
7	B, C	10.2,	10.4	10.3
8	B, D	10.2,	9.8	10.0
9	B, E	10.2,	10.0	10.1
10	B, F	10.2,	10.2	10.2
11	B, G	10.2,	9.6	9.9
12	C, D	10.4,	9.8	10.1
13	C, E	10.4,	10.0	10.2
14	C, F	10.4,	10.2	10.3
15	C, G	10.4,	9.6	10.0
16	D, E	9.8,	10.0	9.9
17	D, F	9.8,	10.2	10.0
18	D, G	9.8,	9.6	9.7
19	E, F	10.0,	10.2	10.1
20	E, G	10.0,	9.6	9.8
21	F, G	10.2,	9.6	9.9

(d) The mean and standard deviation of sampling distribution of $\bar{X}$, are computed below:

Sample Mean $\bar{x}$	Probability $f(\bar{x})$	$\bar{x}\, f(\bar{x})$	$\bar{x}-\mu_{\bar{x}}$	$(\bar{x}-\mu_{\bar{x}})^2$	$(\bar{x}-\mu_{\bar{x}})^2 f(\bar{x})$
9.7	2/21	19.4/21	−0.3	0.09	0.18/21
9.8	2/21	19.6/21	−0.2	0.04	0.08/21
9.9	4/21	39.6/21	−0.1	0.01	0.04/21
10.0	5/21	50.0/21	0	0	0
10.1	4/21	40.4/21	−0.1	0.01	0.04/21
10.2	2/21	20.4/21	0.2	0.04	0.08/21
10.3	2/21	20.6/21	0.3	0.09	0.18/21
Σ	1	10.0	--	--	0.6/21

$$\mu_{\bar{x}} = \Sigma \bar{x}\, f(\bar{x}) = 10.0\ kg, \text{ and}$$

$$\sigma_{\bar{x}} = \sqrt{\Sigma (\bar{x} - \mu_{\bar{x}})^2 f(\bar{x})} = \sqrt{\frac{0.6}{21}} = \sqrt{0.0286} = 0.17\ kg,$$

which is a smaller value indicating that the sampling distribution of the mean is more concentrated about the population mean.

Example 14.8. A sample of size $n=3$ is to be randomly selected without replacement from a population that has $N=5$ items whose values are 0, 3, 6, 9 and 12.

(a) Find the sampling distribution of the sample mean, $\bar{X}$.

(b) Calculate the mean and the standard deviation of $\bar{X}$, and verify that

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}.$$

Let the items be designated by the letters A, B, C, D and E.

(a) The number of samples of size $n=3$ that could be drawn without replacement from a population of size $N=5$ is

$$\binom{5}{3} = \frac{5!}{2!\,3!} = 10.$$

The 10 possible samples and their means are given below:

Sample No.	Sample Combinations	Sample Values	Sample Mean ($\bar{X}$)
1	*A, B, C*	0, 3, 6	3
2	*A, B, D*	0, 3, 9	4
3	*A, B, E*	0, 3, 12	5
4	*A, C, D*	0, 6, 9	5
5	*A, C, E*	0, 6, 12	6
6	*A, D, E*	0, 9, 12	7
7	*B, C, D*	3, 6, 9	6
8	*B, C, E*	3, 6, 12	7
9	*B, D, E*	3, 9, 12	8
10	*C, D, E*	6, 9, 12	9

The sampling distribution is obtained by listing all possible means and their probabilities (relative frequencies) as below:

Sampling Distribution of $\bar{X}$

Sample Mean $\bar{X}$	Number of sample means (f)	Probability $f(\bar{x})$
3	1	1/10
4	1	1/10
5	2	2/10
6	2	2/10
7	2	2/10
8	1	1/10
9	1	1/10
Σ	10	1

(b) Next, we calculate the mean and the standard deviation (the *standard error*) of the sampling distribution of the mean as follows:

Calculation of Mean and $S.D.$ of Sampling Distribution of $\bar{X}$.

Sample Mean $\bar{x}$	Probability $f(\bar{x})$	$\bar{x}\, f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
3	1/10	3/10	9/10
4	1/10	4/10	16/10
5	2/10	10/10	50/10
6	2/10	12/10	72/10
7	2/10	14/10	98/10
8	1/10	8/10	64/10
9	1/10	9/10	81/10
Σ	1	60/10	390/10

Now $\mu_{\bar{x}} = \Sigma \bar{x}\, f(\bar{x}) = \dfrac{60}{10} = 6$, and

$$\sigma_{\bar{x}} = \sqrt{[\Sigma \bar{x}^2 f(\bar{x})] - [\Sigma \bar{x} f(\bar{x})]^2}$$

$$= \sqrt{\frac{390}{10} - \left(\frac{60}{10}\right)^2} = \sqrt{39 - 36} = \sqrt{3} = 1.732$$

In order to verify the given result, we first calculate the mean μ and the variance σ^2 of the given population. Thus

$$\mu = \frac{1}{5}[0 + 3 + 6 + 9 + 12] = \frac{1}{5}[30] = 6, \text{ and}$$

$$\sigma^2 = \frac{1}{5}[(0-6)^2 + (3-6)^2 + (6-6)^2 + (9-6)^2 + (12-6)^2] = 18$$

Verification: Now $\dfrac{\sigma^2}{n} \cdot \dfrac{N-n}{N-1} = \dfrac{18}{3} \cdot \dfrac{5-3}{5-1} = \dfrac{18 \times 2}{3 \times 4} = 3 = \sigma_{\bar{x}}^2$

Hence the result.

Example 14.9. Suppose that a random variable X has the following population distribution:

x	3	6	9
$f(x)$	1/3	1/3	1/3

If a sample of three numbers is taken with replacement, obtain the sampling distribution of the sample mean and verify that $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

The population distribution of the r.v. X may be written as

$$f(x) = P(X=x) = \begin{cases} 1/3 & \text{for } x = 3 \\ 1/3 & \text{for } x = 6 \\ 1/3 & \text{for } x = 9 \end{cases}$$

implying that the members of the population have the numerical values of 3, 6 and 9.

The number of possible samples of size $n=3$, which could be selected with replacement from this population is $3^3=27$. The 27 random samples with their means are given below:

Sample No.	Sample Values X_1	X_2	X_3	Sample Mean $\bar{X}$
1	3	3	3	3
2	3	3	6	4
3	3	3	9	5
4	3	6	3	4
5	3	6	6	5
6	3	6	9	6
7	3	9	3	5
8	3	9	6	6
9	3	9	9	7
10	6	3	3	4
11	6	3	6	5
12	6	3	9	6
13	6	6	3	5
14	6	6	6	6
15	6	6	9	7
16	6	9	3	6
17	6	9	6	7
18	6	9	9	8
19	9	3	3	5
20	9	3	6	6
21	9	3	9	7
22	9	6	3	6
23	9	6	6	7
24	9	6	9	8
25	9	9	3	7
26	9	9	6	8
27	9	9	9	9

The sampling distribution of the sample mean $\bar{X}$ is obtained below, together with two columns needed for the calculation of the *S.E.* of the mean:

Sampling Distribution of $\bar{X}$ and Calculation of *S.E.* $(\bar{X})$

$\bar{X}$	No. of sample means	Probability $f(\bar{x})$	$\bar{x}\, f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
3	1	1/27	3/27	9/27
4	3	3/27	12/27	48/27
5	6	6/27	30/27	150/27
6	7	7/27	42/27	252/27
7	6	6/27	42/27	294/27
8	3	3/27	24/27	192/27
9	1	1/27	9/27	81/27
Σ	27	1	162/27	1026/27

Now, $\mu_{\bar{x}} = \Sigma \bar{x}\, f(\bar{x}) = \dfrac{162}{27} = 6$

$$\sigma^2_{\bar{x}} = \Sigma \bar{x}^2 f(\bar{x}) - [\Sigma \bar{x}\, f(\bar{x})]^2$$

$$= \frac{1026}{27} - (6)^2 = 38 - 36 = 2.$$

And, $\mu = \Sigma x\, f(x) = 3 \times \frac{1}{3} + 6 \times \frac{1}{3} + 9 \times \frac{1}{3} = 6,$

$$\sigma^2 = \Sigma x^2 f(x) - \mu^2 = \left[9 \times \frac{1}{3} + 36 \times \frac{1}{3} + 81 \times \frac{1}{3}\right] - (6)^2$$

$$= (3 + 12 + 27) - 36 = 6.$$

Verification: $\dfrac{\sigma^2}{n} = \dfrac{6}{3} = 2 = \sigma^2_{\bar{x}}$

Hence, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Example 14.10. The weights of 1500 ball bearings are normally distributed with a mean of 22.40 ounces and a standard deviation of 0.048 ounces. If 300 random samples of size 36 are drawn from this population, (a) determine the expected mean and standard deviation of the sampling distribution of mean if sampling is done (i) with replacement, (ii) without replacement; (P.U. B.A./B.Sc., 1971)

(b) How many of the random samples would have their means between 22.39 and 22.42 oz?

(a) There would be $(1500)^{36}$ and $\binom{1500}{36}$ possible samples of size 36, that could be obtained theoretically from a population of weights of 1500 ball bearings with and without replacement respectively. Obviously the number of theoretically possible samples is much larger than 300. Therefore the sampling distribution of the mean will not be a true sampling distribution. (Such a distribution is called *experimental* sampling distribution). But 300 being a large number, there should be a close agreement between the experimental sampling distribution of 300 sample means and the true sampling distribution of mean. Hence the expected mean and standard deviation are found to be as:

(i) *Sampling with replacement:*

$$\mu_{\bar{x}} = \mu = 22.40 \text{ oz.},$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.048}{\sqrt{36}} = 0.008 \text{ oz.}$$

(ii) *Sampling without replacement:*

$$\mu_{\bar{x}} = \mu = 22.40 \text{ oz.},$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Since sample size $n=36$ is less than 5% of the population size $N=1500$, therefore according to the generally accepted rule for the use of *fpc*, the factor $\sqrt{\frac{N-n}{N-1}}$ is dropped. Thus

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.048}{\sqrt{36}} = 0.008 \text{ oz.}$$

(b) The sampling distribution of the mean $\bar{X}$ is normal because the population sampled is normally distributed. Thus

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - 22.40}{0.008}$$ is a standard normal variable.

To find the expected number of samples that would have their means between 22.39 and 22.42 *oz*, we will transform these values to the corresponding *z*-values. Thus

$$\text{at } \bar{x} = 22.39, \text{ we find } z_1 = \frac{22.39 - 22.40}{0.008} = -1.25, \text{ and}$$

$$\text{at } \bar{x} = 22.42, \text{ we find } z_2 = \frac{22.42 - 22.40}{0.008} = 2.50.$$

Using the **Table** of areas under the normal curve, we find

$$\begin{aligned} P(22.39 \le \bar{X} \le 22.42) &= P(-1.25 \le Z \le 2.50) \\ &= P(-1.25 \le Z \le 0) + P(0 \le Z \le 2.50) \\ &= 0.3944 + 0.4938 = 0.8882. \end{aligned}$$

Hence the expected number of samples = (300) (0.8882) = 267.

Example 14.11. A construction company has 310 employees who have an average annual salary of Rs. 24,000. The standard deviation of annual salaries is Rs. 5,000. In a random sample of 100 employees, what is the probability that the average salary will exceed Rs. 24,500?

The sample size ($n = 100$) is large enough to assume that the sampling distribution of $\bar{X}$ is approximately normally distributed with mean

$$\mu_{\bar{x}} = \mu = \text{Rs. } 24{,}000.$$

and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{5000}{\sqrt{100}} \sqrt{\frac{310-100}{310-1}}$$
$$= \text{Rs. } 412.20,$$

where we have used *fpc*, because the sample size $n = 100$ is greater than 5 per cent of the population size $N = 310$.

Equivalently, $Z = \dfrac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \dfrac{\bar{X} - 24000}{412.20}$ is approximately $N(0, 1)$.

We are required to evaluate $P(\bar{X} > 24{,}500)$.

At $\bar{x} = 24{,}500$, we find that $z = \dfrac{24500 - 24000}{412.20} = 1.21$

Hence using *Table* of areas under normal curve, we get

$$\begin{aligned} P(\bar{X} > 24{,}500) &= P(Z > 1.21) \\ &= 0.5 - P(0 \le Z \le 1.21) \\ &= 0.5 - 0.3869 = 0.1131. \end{aligned}$$

Example 14.12. Calculate the standard error of the mean from the following data collected in one of the many random sample inquiries to find average earning of a particular class.

Earning (Rs.)	1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80
Number of persons	20	98	150	218	200	164	110	40

Since the population standard deviation σ is not known and the sample size ($n = 1000$) is large enough to replace it with the sample standard deviation s, we therefore first calculate the sample standard deviation as below:

Earning (Rs.)	f	u	fu	fu^2
1–10	20	−3	−60	180
11–20	98	−2	−196	392
21–30	150	−1	−150	150
31–40	218	0	0	0
41–50	200	1	200	200
51–60	164	2	328	656
61–70	110	3	330	990
71–80	40	4	160	640
Σ	1000	--	612	3208

The sample mean and the sample standard deviation are;

$$\bar{x} = a + \frac{\Sigma fu}{n} \times h = 35.5 + \frac{612}{1000} \times 10 = \text{Rs. } 41.62, \text{ and}$$

$$s = h\sqrt{\frac{\Sigma fu^2}{n} - \left(\frac{\Sigma fu}{n}\right)^2}$$

$$= 10\sqrt{\frac{3208}{1000} - \left(\frac{612}{1000}\right)^2} = 10\sqrt{2.833456} = \text{Rs. } 16.83.$$

Hence the standard error of the sample mean is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{16.83}{\sqrt{1000}} = \frac{16.83}{31.62} = \text{Rs. } 0.53.$$

14.4.2 Central Limit Theorem. The *central limit theorem* is perhaps the most important theorem in all of statistical inference. It is concerned with the means of large samples and provides solutions when the shape of the population distribution is unknown or highly skewed.

The theorem states that "If a variable X from a population has mean μ and *finite* variance σ^2, then the sampling distribution of the sample mean $\bar{X}$ approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ as the sample size n approaches infinity."

To prove this theorem, we use the moment generating function about the mean. By definition, we have

$$M_x(t) = E\left[e^{(X-\mu)t}\right]$$

$$= 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \ldots$$

$$= 1 + 0 + \frac{\sigma^2 t^2}{2!} + \frac{\mu_3 t^3}{3!} + \ldots$$

Let us define a random variable Y as $Y = \frac{X-\mu}{\sigma\sqrt{n}}$.

Then the *m.g.f.* of Y is

$$M_y(t) = E\left[e^{t(X-\mu)/\sigma\sqrt{n}}\right] = M_x(t/\sigma\sqrt{n})$$

$$= 1 + \frac{\sigma^2}{2!}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \frac{\mu_3}{3!}\left(\frac{t}{\sigma\sqrt{n}}\right)^3 + \frac{\mu_4}{4!}\left(\frac{t}{\sigma\sqrt{n}}\right)^4 + \ldots$$

[Replacing t by $\frac{t}{\sigma\sqrt{n}}$ in $M_x(t)$]

$$= 1 + \frac{t^2}{2n} + \frac{\mu_3 t^3}{3!\,\sigma^3 n\sqrt{n}} + \frac{\mu_4 t^4}{4!\,\sigma^4 n^2} + \ldots$$

Let us define another variable Z as a linear function of $\bar{X}$ as

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma/\sqrt{n}} = \sum_{i=1}^{n}\frac{X_i-\mu}{\sigma\sqrt{n}} = \sum_{i=1}^{n} Y_i$$

We know that the *m.g.f.* of a sum of identically distributed random variables is the *nth* power of their own *m.g.f.* Thus

$$M_z(t) = [M_y(t)]^n$$

$$= \left[1 + \frac{t^2}{2n} + \frac{\mu_3 t^3}{3!\,\sigma^3 n\sqrt{n}} + \ldots\right]^n = e^{t^2/2} \text{ as } n \to \infty.$$

But this is the *m.g.f.* of a normal random variable with zero mean and unit variance. Hence Z has in the limit a *standard normal distribution*.

Now $\bar{X} = \mu + \dfrac{Z\sigma}{\sqrt{n}}$

Since a linear function (here $\bar{X}$) of a normal random variable (Z) is a normal random variable, therefore $\bar{X}$ is in the limit normally distributed with mean μ and variance σ^2/n.

It is interesting to note that we have neither assumed that the distribution of X is continuous, nor we have said anything about the shape of the distribution of X, whereas the limiting distribution of X is continuous and normal. Thus the distribution of the sample mean regardless of the shape of the population distribution but having a finite variance, is approximately normal with mean μ and variance σ^2/n.

Example 14.13 Given the population 1, 1, 1, 3, 4, 5, 6, 6, 6 and 7.

(a) Find the probability that a random sample of size 36 selected with replacement will yield a sample mean between 3.26 and 4.74.

(b) Find the mean and standard deviation for the sampling distribution of means for a sample of size 4 selected at random without replacement. Between what two values would you expect at least $\frac{3}{4}$ of the sample means to fall? (P.U., B.A./B.Sc. 1986)

The mean and standard deviation of the population are:

$$\mu = \frac{\Sigma X}{N} = \frac{40}{10} = 4, \text{ and}$$

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}} = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

$$= \sqrt{\frac{210}{10} - \left(\frac{40}{10}\right)^2} = \sqrt{5} = 2.236$$

To calculate mean and standard deviation, we may describe the population by the following probability distribution:

x	1	3	4	5	6	7
$P(X=x)$	3/10	1/10	1/10	1/10	3/10	1/10

(a) As the sampling is performed with replacement, therefore a sample of any size can be selected. A sample of size $n=36$ is large enough for the *central limit theorem* to apply. The sampling distribution of $\bar{X}$ is therefore approximately normal with mean $\mu_{\bar{X}} = \mu = 4$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.236}{6} = 0.373$, that is

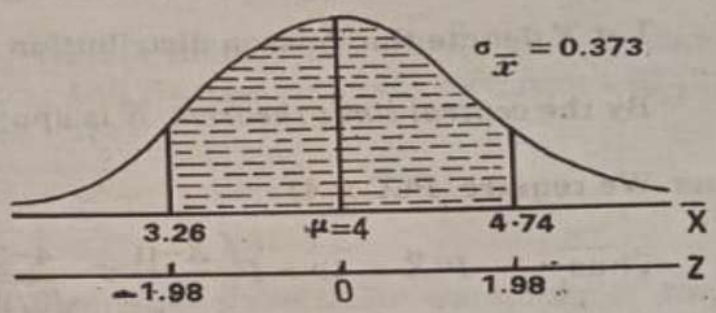
$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - 4}{0.373} \text{ is approximately } N(0, 1).$$

To find the probability that the mean of a random sample of size $n=36$ will fall between 3.26 and 4.74, we transform 3.26 and 4.74 to z values. Thus at $\bar{x} = 3.26$, we find

$$z = \frac{3.26-4}{0.373} = -1.98,$$

and at $\bar{x} = 4.74$, we find

$$z = \frac{4.74-4}{0.373} = 1.98.$$



Hence using **Table** of areas under normal curve, we find

$$P(3.26 \le \bar{X} \le 4.74) = P(-1.98 \le Z \le 1.98)$$
$$= P(-1.98 \le Z \le 0) + P(0 \le Z \le 1.98)$$
$$= 0.4762 + 0.4762 = 0.9524.$$

(b) As the sample is without replacement and sample size $n=4$ is greater than 5% of the population size $N=10$, therefore the mean and standard deviation of the sampling distribution of $\bar{X}$, are

$$\mu_{\bar{X}} = \mu = 4, \text{ and}$$

$$\sigma_{\bar{X}} = \frac{4}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{2.236}{\sqrt{4}} \sqrt{\frac{10-4}{10-1}} = (1.118)(0.816) = 0.912$$

The *Chebyshev's inequality* says "*at least* $\left(1 - \frac{1}{k^2}\right)$ fraction of the data lies in the interval mean $\pm\, k(s.d.)$" and the problem says "*at least* $\frac{3}{4}$

of the sample means should fall in the same interval," so $\frac{3}{4}$ is $1-\frac{1}{k^2}$ that is

$$1-\frac{1}{k^2}=\frac{3}{4} \text{ or } \frac{1}{k^2}=\frac{1}{4}$$

or $\quad k^2=4$ or $k=2$.

Hence we would expect *at least* $\frac{3}{4}$ of the sample means to fall in the interval $\mu_{\bar{X}} \pm 2\sigma_{\bar{X}}$, that is between $4-2(0.912)$ and $4+2(0.912)$ or between 2.2 and 5.8.

Example 14.14 A random sample of size 25 is selected from a Poisson distribution with $\mu=3$. Find, using the *central limit theorem*, the probability that the sample mean will be greater than 4.

Let X denote the Poisson distribution with $\mu=3$. Then $\text{Var}(X)=3$.

By the *central limit theorem*, $\bar{X}$ is approximately $N\left(3, \frac{3}{25}\right)$.

We require $P(\bar{X}>4)$

Thus $$P(\bar{X}>4)=P\left(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}}>\frac{4-3}{\sqrt{3/25}}\right)$$

$$=P(Z>2.89)=0.0019$$

14.4.3 Sampling Distribution of Differences between Means. Suppose we have two large or infinite populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2 respectively. Let independent random samples of sizes n_1 and n_2 be selected from the respective populations, and the differences $\bar{x}_1-\bar{x}_2$ between the means of all possible pairs of samples be computed. Then a probability distribution of the differences $\bar{X}_1-\bar{X}_2$ can be obtained. Such a distribution is called the *sampling distribution of the differences of sample mean* $\bar{X}_1-\bar{X}_2$. The sampling distribution of the differences $\bar{X}_1-\bar{X}_2$ has the following properties:

(i) The mean of the sampling distribution of $\bar{X}_1-\bar{X}_2$, denoted by $\mu_{\bar{X}_1-\bar{X}_2}$, is equal to the difference between population means, that is

$$\mu_{\bar{X}_1-\bar{X}_2}=\mu_1-\mu_2 \quad [\because E(\bar{X}_1-\bar{X}_2)=E(\bar{X}_1)-E(\bar{X}_2)=\mu_1-\mu_2]$$

(ii) The standard deviation of the sampling distribution of $\bar{X}_1-\bar{X}_2$ (*standard error of* $\bar{X}_1-\bar{X}_2$), denoted by $\sigma_{\bar{X}_1-\bar{X}_2}$, is given by

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \quad \left[\because \text{Var}(\bar{X}_1-\bar{X}_2) = \text{Var}(\bar{X}_1)+\text{Var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right] \text{(Samples are independent)}$$

This expression for the *S.E.* of $\bar{X}_1-\bar{X}_2$ also holds for finite populations when sampling is performed with replacement. When population standard deviations are equal or both the samples come from the same population, the expression for the *S.E.* becomes

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{\frac{2\sigma^2}{n}}, \text{ when } n_1 = n_2 = n.$$

If the values of σ_1 and σ_2 are not known and if both sample sizes are large, they are replaced by S_1 and S_2, the standard deviations of the respective samples. The *S.E.* becomes

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

If, on the other hand, the populations are finite, sampling is done without replacement and the sample sizes are larger than 5 per cent of the population sizes, then the *S.E.* is

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1}\cdot\frac{N_1-n_1}{N_1-1} + \frac{\sigma_2^2}{n_2}\cdot\frac{N_2-n_2}{N_2-1}}.$$

(iii) *Shape of the distribution.* If the populations are normally distributed, the sampling distribution of $\bar{X}_1-\bar{X}_2$, regardless of sample sizes, will be normal with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. In other words, the variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is normally distributed with zero mean and unit variance. If the populations are non-normal and if both sample sizes are large, (≥ 30),

then the sampling distribution of differences between means is approximately a normal by the central limit theorem.

Example 14.15. Draw all possible random samples of size $n_1=2$ with replacement from a finite population consisting of 4, 6, 8. Similarly draw all possible random samples of size $n=2$ with replacement from another finite population consisting of 1, 2, 3.

(a) Find the possible differences between the sample means of the two populations.

(b) Construct the sampling distribution of $\bar{X}_1-\bar{X}_2$ and compute its mean and variance.

(c) Verify that $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1-\mu_2$ and $\sigma^2_{\bar{x}_1-\bar{x}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_1}$.

There are $(3)^2 = 9$ possible samples which can be drawn with replacement from each population. These two sets of samples and their means are given below:

From population 1			From population 2		
Sample No.	Sample values	$\bar{x}_1$	Sample No.	Sample values	$\bar{x}_2$
1	4,4	4	1	1,1	1.0
2	4,6	5	2	1,2	1.5
3	4,8	6	3	1,3	2.0
4	6,4	5	4	2,1	1.5
5	6,6	6	5	2,2	2.0
6	6,8	7	6	2,3	2.5
7	8,4	6	7	3,1	2.0
8	8,6	7	8	3,2	2.5
9	8,8	8	9	3,3	3.0

(a) The 81 possible differences $\bar{x}_1-\bar{x}_2$ are presented in the following table.

Differences of Independent Means

$\bar{x}_2$	$\bar{x}_1$								
	4	5	6	5	6	7	6	7	8
1.0	3.0	4.0	5.0	4.0	5.0	6.0	5.0	6.0	7.0
1.5	2.5	3.5	4.5	3.5	4.5	5.5	4.5	5.5	6.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
1.5	2.5	3.5	4.5	3.5	4.5	5.5	4.5	5.5	6.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
2.5	1.5	2.5	3.5	2.5	3.5	4.5	3.5	4.5	5.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
2.5	1.5	2.5	3.5	2.5	3.5	4.5	3.5	4.5	5.5
3.0	1.0	2.0	3.0	2.0	3.0	4.0	3.0	4.0	5.0

(b) The sampling distribution of $\bar{X}_1-\bar{X}_2$ (*i.e.*, the relative frequency distribution of the possible differences $\bar{x}_1-\bar{x}_2$) is constructed below and the mean and variance of this distribution are also computed below.

$\bar{x}_1-\bar{x}_2$ $(=d)$	Tally	f	Probability $f(\bar{x}_1-\bar{x}_2)$	$df(d)$	$d^2 f(d)$
1.0	I	1	1/81	1/81	1.0/81
1.5	II	2	2/81	3/81	4.5/81
2.0	~~IIII~~	5	5/81	10/81	20.0/81
2.5	~~IIII~~ I	6	6/81	15/81	37.5/81
3.0	~~IIII~~ ~~IIII~~	10	10/81	30/81	90.0/81
3.5	~~IIII~~ ~~IIII~~	10	10/81	35/81	122.5/81
4.0	~~IIII~~ ~~IIII~~ III	13	13/81	52/81	208.0/81
4.5	~~IIII~~ ~~IIII~~	10	10/81	45/81	202.5/81
5.0	~~IIII~~ ~~IIII~~	10	10/81	50/81	250.0/81
5.5	~~IIII~~ I	6	6/81	33/81	181.5/81
6.0	~~IIII~~	5	5/81	30/81	180.0/81
6.5	II	2	2/81	13/81	84.5/81
7.0	I	1	1/81	7/81	49.0/81
Total	---	81	1	324/81	1431/81

Thus the mean and the variance are

$$\mu_{\bar{x}_1-\bar{x}_2} = \sum (\bar{x}_1 - \bar{x}_2)\, f(\bar{x}_1 - \bar{x}_2)$$

$$= \sum df(d) = \frac{324}{81} = 4, \text{ and}$$

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \sum (d - \mu_{\bar{x}_1-\bar{x}_2})^2 f(d) = \sum d^2 f(d) - [\sum df(d)]^2$$

$$= \frac{1431}{81} - \left(\frac{324}{81}\right)^2 = \frac{53}{3} - 16 = \frac{5}{3}$$

(c) The mean and variance of the first population are

$$\mu_1 = \frac{4+6+8}{3} = 6, \text{ and}$$

$$\sigma_1^2 = \frac{(4-6)^2 + (6-6)^2 + (8-6)^2}{3} = \frac{8}{3}.$$

The mean and variance of the second population are

$$\mu_2 = \frac{1+2+3}{3} = 2, \text{ and}$$

$$\sigma_2^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}.$$

Now $\mu_{\bar{x}_1-\bar{x}_2} = 4 = 6 - 2 = \mu_1 - \mu_2$, and

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{8}{3}\cdot\frac{1}{2} + \frac{2}{3}\cdot\frac{1}{2}$$

$$= \frac{4}{3} + \frac{1}{3} = \frac{5}{3} = \sigma^2_{\bar{x}_1-\bar{x}_2}$$

Hence the result.

Example 14.16. Car batteries produced by company A have a mean life of 4.3 years with a standard deviation of 0.6 years. A similar battery produced by company B has a mean life of 4.0 years and a standard deviation of 0.4 years. What is the probability that a random sample of 49 batteries from company A will have a mean life of *at least* 0.5 years more than the mean life of a sample of 36 batteries from company B?

We are given the following data:

Population A: $\mu_1 = 4.3$ years, $\sigma_1 = 0.6$ years, $n_1 = 49$

Population B: $\mu_2 = 4.0$ years, $\sigma_2 = 0.4$ years, $n_2 = 36$

Both sample sizes ($n_1=49$, $n_2=36$) are large enough to assume that the sampling distribution of the differences $\bar{X}_1-\bar{X}_2$ is approximately a normal with mean

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 = 4.3 - 4.0 = 0.3 \text{ years and standard deviation}$$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{0.36}{49} + \frac{0.16}{36}} = 0.1086 \text{ years.}$$

Thus the variable

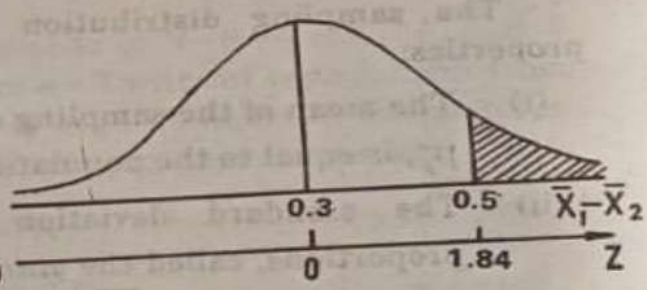
$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(\bar{X}_1 - \bar{X}_2) - 0.3}{0.1086} \text{ is approximately } N(0, 1).$$

We are required to find the probability that the mean life of 49 batteries produced by company A will have a mean life of *at least* 0.5 years longer than the mean life of 36 batteries produced by company B, that is we want $P(\bar{X}_1 - \bar{X}_2 \geq 0.5)$.

Transforming $\bar{X}_1 - \bar{X}_2 = 0.5$ to z value, we find that

$$z = \frac{0.5 - 0.3}{0.1086} = 1.84$$



Hence using *Table* of areas under normal curve, we find

$$P(\bar{X}_1-\bar{X}_2 \geq 0.5) = P(Z \geq 1.84)$$

$$= 0.5 - P(0 < Z < 1.84)$$

$$= 0.5 - 0.4671$$

$$= 0.0329$$

14.4.4. Sampling Distribution of Sample Proportion. A population proportion p may be identified with the population mean, where the mean is obtained from the units whose possible values are either 0's or 1's. In other words, let

$Y_i = 1$, if the *ith* unit possesses the characteristic of interest,

$= 0$, if the *ith* unit does not possess the characteristic of interest,

Then the mean is

$$\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$= \frac{\text{Number of units having the characteristic of interest}}{\text{Total number of units in the population}}$$

$$= \frac{X}{N},$$ where X represents the number of units having the characteristics of interest.

Thus the mean is simply the proportion of 1's in the population and we write p for μ, meaning *proportion* (usually called the proportion of success).

Similarly, the sample proportion $\hat{P}$ is defined as

$$\hat{P} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{X}{n}.$$

It is interesting to note that $X = \sum Y_i$ is a binomial random variable and the binomial parameter p is being called a proportion of success here. The sample proportion $\hat{P}$ has different values in different samples. It is obviously a random variable and has a probability distribution. This probability distribution of the proportions of all possible random samples of size n, is called the *sampling distribution of* $\hat{P}$.

The sampling distribution of $\hat{P}$ has the following important properties:

(i) The mean of the sampling distribution of proportions, denoted by $\mu_{\hat{p}}$, is equal to the population proportion p, that is $\mu_{\hat{p}} = p$.

(ii) The standard deviation of the sampling distribution of proportions, called the *standard error of* $\hat{P}$ and denoted by $\sigma_{\hat{p}}$, is given as $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$, when the sampling is performed with replacement or

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}\cdot\frac{N-n}{N-1}},$$ when sampling is done without replacement from a finite population, and where $q = 1-p$. It is of importance to remember that $\frac{N-n}{N-1}$ is to be used when the sample size n is 5% or more than 5% of the population size N.

When the population proportion p is not known and both the population and the sample sizes are large, then the sample proportion $\hat{P}$

obtained from sample data is used in place of p in the expression for the $S.E.$ of P, getting

$$S_{\hat{p}} = \sqrt{\frac{\hat{p}\,\hat{q}}{n}}, \text{ where } \hat{q} = 1 - \hat{p}.$$

When the sample is selected without replacement from a finite population of size N, the $S.E.$ becomes

$$S_{\hat{p}} = \sqrt{\frac{\hat{p}\,\hat{q}}{n} \cdot \frac{N-n}{N-1}}.$$

(iii) *Shape of the distribution.* The sampling distribution of $\hat{P}$ is the binomial distribution. However, for sufficiently large sample sizes, the sampling distribution of P is approximately normal. As a rule of thumb, the sampling distribution of P will be approximately normal whenever both np and nq are equal to or greater than 5.

It helps to remember that we use a continuity correction of $\pm \frac{1}{2}$, whenever we consider the normal approximation to the binomial distribution. Now, we need to use a continuity correction of $\pm\frac{1}{2n}$ as $\hat{P} = \frac{X}{n}$.

Example 14.17. A population consists of $N=6$ values 1, 3, 6, 8, 9 and 12. Draw all possible samples of size $n=3$ without replacement from the population and find the proportion of even numbers in the samples. Construct the sampling distribution of sample proportions and verify that

(i) $\mu_{\hat{p}} = p$, (ii) $\text{Var}(\hat{P}) = \frac{pq}{n} \cdot \frac{N-n}{N-1}$.

where $q = 1 - p$; $\hat{P}$ and p are sample and population proportions respectively.

The number of possible samples of size $n=3$ that could be selected without replacement is $\binom{6}{3} = 20$. Let $\hat{P}$ represent the proportion of even numbers in the sample. then the 20 possible samples and the proportion of even numbers are given as follows:

Sample No.	Sample Data	Sample Proportion ($\hat{P}$)
1	1, 3, 6	1/3
2	1, 3, 8	1/3
3	1, 3, 9	0
4	1, 3, 12	1/3
5	1, 6, 8	2/3
6	1, 6, 9	1/3
7	1, 6, 12	2/3
8	1, 8, 9	1/3
9	1, 8, 12	2/3
10	1, 9, 12	1/3
11	3, 6, 8	2/3
12	3, 6, 9	1/3
13	3, 6, 12	2/3
14	3, 8, 9	1/3
15	3, 8, 12	2/3
16	3, 9, 12	1/3
17	6, 8, 9	2/3
18	6, 8, 12	1
19	6, 9, 12	2/3
20	8, 9, 12	2/3

The sampling distribution of sample proportion is given below:

$\hat{P}$	No. of samples	Probability $f(P)$	$\hat{p}\, f(\hat{p})$	$\hat{p}^2 f(\hat{p})$
0	1	1/20	0	0
1/3	9	9/20	3/20	1/20
2/3	9	9/20	6/20	4/20
1	1	1/20	1/20	1/20
Σ	20	1	10/20	6/20

Now $\mu_{\hat{p}} = \Sigma \hat{p}\, f(\hat{p}) = \frac{10}{20} = 0.5$, and

$$\sigma^2_{\hat{p}} = \Sigma \hat{p}^2 f(\hat{p}) - [\Sigma \hat{p}\, f(\hat{p})]^2 = \frac{6}{20} - \left(\frac{10}{20}\right)^2 = \frac{1}{20} = 0.05.$$

To verify the given relations, we first calculate the population proportion p and the population variance pq. Thus

$$p = \frac{X}{N}, \text{ where } X \text{ represents the number of even numbers.}$$

$$= \frac{3}{6} = 0.5, \text{ and}$$

$$\sigma^2 = pq = (0.5)\,(0.5) = 0.25$$

Therefore $\mu_{\hat{p}} = 0.5 = p$, and

$$\frac{pq}{n} \cdot \frac{N-n}{N-1} = \frac{0.25}{3} \cdot \frac{6-3}{6-1} = \frac{0.25}{5} = 0.05 = \text{Var}(\hat{P})$$

Hence the result.

Example 14.18. Ten percent of the 1-kilogram boxes of sugar in a large warehouse are underweight. Suppose a retailer buys a random sample of 144 of these boxes. What is the probability that at least 5 per cent of the sample boxes will be underweight?

Here the statistic is the sample proportion $\hat{P}$.

The sample size ($n = 144$) is large enough to assume that the sample proportion P is approximately normally distributed with mean

$$\mu_{\hat{p}} = p = 0.10, \text{ and standard error}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.10)\,(0.9)}{144}} = \frac{0.3}{12} = 0.025.$$

Therefore $Z = \dfrac{\hat{P} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \dfrac{\hat{P} - p}{\sqrt{pq/n}} = \dfrac{\hat{P} - p}{0.025}$ is approximately $N(0, 1)$.

We are asked to find the probability that the sample proportion of the underweight boxes is equal to or greater than 5% *i.e.*, we require $P(\hat{p} \geq 0.05)$.

Thus $P(\hat{p} \geq 0.05) \Rightarrow P\left(\hat{p} \geq 0.05 - \dfrac{1}{(2)\,(144)}\right)$ [Continuity correction]

$$= P\left(\frac{\hat{p} - 0.10}{0.025} \geq \frac{(0.05 - 1/288) - 0.10}{0.025}\right)$$

$$= P\,(Z \geq -2.14)$$

$$= P(-2.14 \leq Z \leq 0) + P(0 \leq Z \leq \infty)$$

$$= 0.4838 + 0.5 = 0.9838 \text{ (From area table)}$$

14.4.5. Sampling Distribution of Differences between Proportions. Suppose there are two *binomial* populations with proportions of successes p_1 and p_2 respectively. Let independent random

samples of sizes n_1 and n_2 be drawn from the respective populations, and the differences $\hat{p}_1 - \hat{p}_2$ between the proportions of all possible pairs of samples be computed. Then a probability distribution of the differences $\hat{P}_1 - \hat{P}_2$ can be obtained. Such a probability distribution is called the *sampling distribution of the differences between the proportions* $\hat{P}_1 - \hat{P}_2$, which has the following important properties:

(i) The mean of the sampling distribution of $\hat{P}_1 - \hat{P}_2$, denoted by $\mu_{\hat{P}_1 - \hat{P}_2}$, is equal to the difference between the population proportions, that is $\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$.

(ii) The standard deviation of the sampling distribution of $\hat{P}_1 - \hat{P}_2$ (*i.e. the standard error of* $\hat{P}_1 - \hat{P}_2$) denoted by $\sigma_{\hat{P}_1 - \hat{P}_2}$ is given by

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}, \text{ where } q = 1 - p.$$

If both populations have the same proportion of successes, i.e. $p_1 = p_2 = p$ or if both the samples have been drawn from a common binomial population, then

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Whenever the value of the common proportion p is not known, then for sufficiently large sample sizes, it is replaced with its estimate $\hat{p}_c$ which is computed by taking a weighted mean of the two observed sample proportions $\hat{p}_1$ and $\hat{p}_2$ as follows:

$$\hat{p}_c = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{\text{Sum of successes in the two samples}}{\text{Total sample size}}$$

The standard error of $\hat{P}_1 - \hat{P}_2$ then becomes

$$S_{\hat{P}_1 - \hat{P}_2} = \sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{ where } \hat{q}_c = 1 - \hat{p}_c.$$

If, on the other hand, $p_1 \neq p_2$ and also are not known, then for large sample sizes, they are replaced with the sample proportions $\hat{p}_1$ and $\hat{p}_2$ respectively. The *S.E.* of $\hat{P}_1 - \hat{P}_2$ then becomes

$$S_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.$$

(iii) *Shape of the distribution.* The sampling distribution of $\hat{P}_1 - \hat{P}_2$ is approximately normal for sufficiently large sample sizes.

Example 14.19. Two random samples of sizes $n_1=40$ and $n_2=45$ are drawn from a binomial population with $p=0.60$. What is the probability that $-0.15<\hat{p}_1-\hat{p}_2<+0.15$?

Both sample sizes ($n_1=40$ and $n_2=45$) are large enough to assume that the sampling distribution of $\hat{P}_1-\hat{P}_2$ is approximately a normal with mean

$$\mu_{\hat{p}_1-\hat{p}_2}=p_1-p_2=0, \text{ and standard deviation}$$

$$\sigma_{\hat{P}_1-\hat{P}_2}=\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=\sqrt{(0.60)\,(0.40)\left(\frac{1}{40}+\frac{1}{45}\right)}$$

$$=\sqrt{(0.24)\,(0.0472)}=0.106.$$

Thus the variable

$$Z=\frac{(\hat{P}_1-\hat{P}_2)-0}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$=\frac{\hat{P}_1-\hat{P}_2}{0.106} \text{ is approximately } N(0,1).$$

Now, at $\hat{p}_1-\hat{p}_2=-0.15$, we find that $z=\dfrac{-0.15}{0.106}=-1.42$,

Hence using Table of areas under normal curve, we find

$$P(-0.15<\hat{p}_1-\hat{p}_2<0.15)=P(-1.42<Z<1.42)$$

$$=P(-1.42<Z<0)+P(0<Z<1.42)$$

$$=0.4222+0.4222=0.8444.$$

The desired probability is therefore 0.8444.

14.4.6. Sampling Distribution of Variances. The sampling distribution of the sample variances calculated from all possible random samples of size n from a normal population with variance σ^2, is the so called *Chi-Square Distribution*, which is discussed in chapter 17. The sampling distribution followed by the ratio of two sample variances is called the *F-distribution* to be introduced in chapter 19.

EXERCISES

14.1 (a) Explain the following terms:

Population; Sample; Sampling Frame; Parameter; Statistic.

(b) What is meant by sampling? What is the object of sampling?

14.2 (a) Describe the advantages of sampling over complete enumeration. (P.U., B.A/B.Sc., 1982)

(b) Explain the term "non-response" in sample surveys.

14.3 Define and distinguish between:

(a) Target and Sampled Populations;

(b) Probability and Nonprobability Sampling;

(c) Sampling With and Without Replacement;

(d) Sampling and Non-sampling Errors;

(e) Random Sampling and Simple Random Sampling. (P.U., B.A./B.Sc., 1993)

14.4 (a) Explain sampling and non-sampling errors. What methods would you suggest to control each type of error?

(b) What is a biased sample? Explain the different ways in which bias may arise in sample surveys. How bias can be eliminated? (P.U., B.A./B.Sc. 1989, 93)

14.5 What are two broad categories of errors in data collected by sample surveys? Describe some of the methods for reducing these errors. (P.U., B.A/B.Sc. 1960)

14.6 Discuss the common types of sampling techniques used to gather information and the advantages and disadvantages of each of these techniques.

14.7 Discuss (i) Simple random sample, (ii) Stratified random sample, (iii) Systematic random sample, (iv) Cluster sample. Give an example of study in which each of these sampling plans could be used appropriately.

14.8 (a) Explain what you understand by probability sampling and non-probability sampling. What are their relative advantages and disadvantages?

(b) Write a short account of stratified random sampling. (P.U., B.A./B.Sc. 1983)

14.9 (a) Define a simple random sample. Discuss how a simple random sample could be selected.

(b) Discuss the following:

(i) The "Goldfish bowl" method.

(ii) The random number table.

14.10 (a) What is a stratified random sample? In what way does it differ from a simple random sample, and what are the advantages and disadvantages of using this sampling technique?

(b) What is meant by allocation of sample size? Explain how a sample is allocated in stratified sampling.

14.11 Describe, in detail, various probability and non-probability sampling techniques. (P.U., B.A./B.Sc. 1981)

14.12 (a) Distinguish between

(i) Probability and Non-probability Sampling

(ii) Sampling and Non-sampling Errors

(iii) Multistage and Multiphase Sampling.

(P.U., B.A/B.Sc. 1974, 92)

(b) What is a cluster sample? Why are cluster samples used?

14.13 (a) Describe the necessity of sampling and sample surveys.

(b) What is a Systematic sample? Describe the procedure of drawing a systematic sample of n units from a population of N units.

14.14 Describe the following types of Samples:

(a) Random Sample (b) Quota Sample

(c) Systematic Random Sample (d) Multi-stage Random Sample

State the conditions under which each would be used and the advantages to be gained. (P.U., M.A. Stats. 1964)

14.15 Explain each of the following:

(i) Area Sampling (ii) Optimum Allocation

(iii) Sequential Sampling (iv) Purposive Sampling

(v) Quota Sampling.

14.16 Draw all possible distinct samples of size two from the following population: 2, 4, 6, 8, 10.

Calculate the means and the variances of the samples and of the population. Discuss the results. (P.U., B.A/B.Sc. 1972)

14.17 Explain how you would select a random sample of 10 households from a list of 250 households, by using a table of random numbers. (P.U.M.A. Stats. 1967)

*14.18 Using a random number table, select 30 samples of size 1 with replacement from the following population distribution heights. Find the mean of sample means.

Height (inches)	No. of students
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8

(P.U. B.Sc. Hons. Part I, 19

14.19 Draw, with the help of random numbers, a random sample size 10 from a

(i) Binomial distribution with parameters $p = 0.4$ and $n = 5$;

(ii) Poisson distribution with the parameter $\mu = 4$.

14.20 Using a random number table, draw a sample of size 30 from Normal distribution with $\mu = 100$ and $\sigma^2 = 64$.

14.21 (a) Describe stratified random sampling, explaining in detail the following types of allocation of sample sizes:

(i) Proportional Allocation (ii) Optimum Allocation.

(P.U., B.A./B.Sc., 199

(b) Select a stratified random sample of size $n = 8$ by proportional allocation from the following population and find the sample mean and the estimate of the population mean.

Stratum I	$X_{11} = 3, X_{12} = 6, X_{13} = 4, X_{14} = 7,$
Stratum II	$X_{21} = 10, X_{22} = 12, X_{23} = 15, X_{24} = 16, X_{25} = 16, X_{26} = 20,$
Stratum III	$X_{31} = 16, X_{32} = 18, X_{33} = 21, X_{34} = 22, X_{35} = 26, X_{36} = 23$

(P.U., B.A/B.Sc. 1975)

14.22 (a) At a small private college, the students are classified as follows:

Classification	B.Sc.	B.A.	F.Sc.	F.A.
No. of students	150	163	195	220

If we wish to select a stratified random sample of size $n = 40$ by proportional allocation, how large a sample must we take from each stratum?

(P.U., B.A/B.Sc. 1987)

(b) A large company has 300,000 employees, the age distribution of whom is shown as follows:

Age (years)	Percentages
25 or younger	15
26 – 35	30
36 – 45	25
46 – 55	20
56 or older	10

A sample of 2 per cent of all the employees is desired. Design a sampling plan such that each age-group is proportionally represented. (P.U., B.A/B.Sc., 1988)

14.23 (a) What is a sampling distribution? Describe the properties of the sampling distribution of the means.

(b) What is the finite-correction factor? When is it appropriately used in sampling applications and when can it, without too great an undesirable consequence, be ignored? (P.U., B.A./B.Sc., 1996)

14.24 (a) Explain the difference between the population distribution, the sample distribution and sampling distribution. (P.U., B.A./B.Sc. 1989, 93)

(b) What is meant by standard error and what are its practical uses? Derive a formula for the standard error of the mean. (P.U., B.A. Hons. Part I; 1970)

14.25 (a) Explain the difference between (i) a sample distribution and a sampling distribution (ii) a standard deviation and a standard error.

(b) Suppose a friend says, "I know the formula for computing the standard error of the mean, but I do not understand what the standard error really is." Write a note to your friend explaining what the standard error really is.

14.26 (a) Distinguish between a parameter and a statistic. What is meant by Standard Error and what are its practical uses? (P.U. B.A./B.Sc. 1993)

(b) Assume that simple random samples of two children are selected with replacement from a population of five children with ages 4, 5, 6, 7 and 8. Let X be the age of any child, find the following:

(i) The theoretical sampling distribution of $\bar{X}$, the mean of two children in any sample.

(ii) The mean and the standard error of $\bar{X}$.

(B.Z.U. B.A/B.Sc. 20[illegible])

14.27 Given the population 2, 4, 8, 8, 10, 10.

(i) How many samples of size $n=2$ can be drawn with replacement from this population?

(ii) Compute and tabulate the sampling distribution of the mean for samples of size $n=2$.

14.28 A finite population consists of the numbers 2, 4 and 6.

(a) Form the sampling distribution of $\bar{X}$, when random samples of size 4 are drawn, with replacement.

(b) Verify that $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$. (P.U., B.A/B.Sc. 19[illegible])

14.29 A population consists of 2, 2, 4, 4, 6, 8 and 10.

(i) Calculate the sample means for all possible random samples of size $n=2$, that can be drawn from this population, without replacement.

(ii) Verify that $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

(iii) Between what two values would you expect *at least* $\frac{8}{9}$ of the sample means to fall?

(P.U., B.A./B.Sc. 1999)

14.30 Given the six-element population 0, 3, 6, 12, 15, and 18. How many samples of size $n=3$ can be drawn, without replacement from this population? Compute the sampling distribution of the mean for samples of size 3. Compute the mean and standard deviation of this distribution.

14.31 Draw all possible samples of size $n=3$ with replacement from the population 3, 6, 9 and 12. Form a sampling distribution of the sample means. Hence state and verify the relation between

(i) mean of the sampling distribution of the mean and the population mean;

(ii) variance of the sampling distribution of the mean and the population variance.

(P.U., B.A./B.Sc. 1982)

14.32 A population of $N=5$ has the following values:

4, 5, 7, 9, 10.

(a) Find the population mean and variance.

(b) Suppose samples of size $n=3$ are selected. Find $\sigma^2_{\bar{x}}$ when sampling is done. (i) without replacement, (ii) with replacement.

(c) Select all possible samples of size $n=3$ without replacement and calculate $\sigma^2_{\bar{x}}$ directly.

(d) Select all possible samples of size $n=3$ with replacement and calculate $\sigma^2_{\bar{x}}$ directly.

14.33 A population consists of four numbers 2, 4, 6, 8. Draw all possible sample of size $n=3$ with replacement. Find the mean and the median for each sample. Form the sampling distribution of means and the sampling distribution of medians. Which of these distributions has the smaller variance? How did the means of these two distributions compare with the population mean?

14.34 Given the following population distribution:

x	1	2	3	4
$f(x)$	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{2}{7}$	$\frac{1}{7}$

Find the sampling distribution of the mean if a sample of three numbers is taken without replacement. How does the variance of the sampling distribution compare with the population variance?

14.35 A random sample of size $n=100$ is taken from a population having a mean of 20 and a standard deviation of 5. The shape of the population distribution is unknown.

(a) What can you say about the sampling distribution of the sample mean $\bar{X}$?

(b) Find the probability that $\bar{X}$ will exceed 20.75.

(P.U., B.A/B.Sc. 1996)

14.36 In a local agriculture reporting area, the average wheat yield is known to be 60 bushels per acre with a standard deviation of 10 bushels. If a random sample of 64 acres is selected and the wheat yield recorded, what is the probability that the sample mean will lie between 59 and 61 bushels?

14.37 The heights of 1000 students are approximately normally distributed with a mean of 68.5 inches and a standard deviation

of 2.7 inches. If 200 random samples of size 25 are drawn from this population and the means recorded to the nearest tenth of an inch, determine

(a) The expected mean and standard deviation of the sampling distribution of the mean.

(b) The number of sample means that fall between 67.9 and 69.2 inclusive.

(P.U., B.A./B.Sc. 1977-S)

14.38 The heights of a large number of shrubs of the same kind produced for sale by a horticultural nursery are normally distributed with mean 1.14m and standard deviation 0.25m. Fifty samples, each consisting of 100 shrubs, are selected. In how many of these samples would you expect to find the mean sample being to be (i) greater than 1.16m; (ii) between 1.13m and 1.18m?

14.39 (a) The following table shows the distribution of 14-year-old schoolboy intelligence test markings:

I.Q.	80–89	90–99	100–109	110–119	120–129	130–139	140–149
Number	30	52	75	109	65	42	27

On the assumption that this group is a random sample, estimate the standard error of the mean and explain its usefulness.

(b) The random variable X has the following probability distribution:

x	4	5	6	7
$P(X=x)$	0.2	0.4	0.3	0.1

(i) Find the mean $\mu_{\bar{x}}$ and variance $\sigma^2_{\bar{x}}$ of the mean $\bar{X}$ for a random sample of 36.

(ii) Find the probability that the mean of 36 items will be less than 5.5.

(P.U., B.A./B.Sc. 1987)

14.40 (a) The mean of a certain normal distribution is equal to $S.E.$ of the mean of samples of 100 from that distribution. Find the probability that the mean of a sample of 25 from the distribution will be negative.

(b) A normal population has a mean of 0.1 and a standard deviation of 2.1. Find the probability that the mean of a simple random sample of 900 members will be negative.

Solution. (a) The variable X is $N\left(\frac{\sigma}{10}, \sigma\right)$ as

$\mu = \frac{\sigma}{10}$, the *S.E.* of the mean of samples of size $n = 100$.

And the *S.E.* of the mean of samples of size $25 = \frac{\sigma}{5}$.

From the standard normal variable $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$, we get

$$\bar{X} = \mu + Z\frac{\sigma}{\sqrt{n}} = \frac{\sigma}{10} + Z \cdot \frac{\sigma}{5}$$

Now $\bar{X}$ will be negative if $\left(\frac{\sigma}{10} + Z\frac{\sigma}{5}\right) < 0$,

$$i.e. \text{ if } Z < -\frac{\sigma}{10} \times \frac{5}{\sigma} \text{ or if } Z < -0.5.$$

Hence using the *Table* of areas under normal curve, we get

$$P(Z < -0.5) = 0.5 - P(-0.5 < Z < 0)$$

$$= 0.5 - 0.1915 = 0.3085$$

14.41 (a) A random sample of size 100 is taken from a Binomial distribution with parameters $p = 0.5$ and $n = 40$. Find, using the central limit theorem, the approximate probability that $\bar{X}$ is (i) greater than 20.5; (ii) less than 19.3; and hence (iii) between 19.3 and 20.5.

(b) A sample of 36 cases is drawn from a negatively skewed population with a mean of 2 and a standard deviation of 3. What is the probability that the sample mean obtained will be negative? How many points must we go from the mean to include 50 percent of all sample means? (P.U., B.A/B.Sc., 1988)

14.42 (a) Describe the properties of the sampling distribution of the differences between two means. (P.U., B.A./B.Sc. 1983,87)

(b) Random samples of size 100 are drawn, with replacement, from two populations and their means $\bar{X}_1$ and $\bar{X}_2$ computed. If $\mu_1 = 10$, $\sigma_1 = 2$, $\mu_2 = 8$ and $\sigma_2 = 1$, find the probability that the difference between a given pair of sample means is (i) less than 1.5, and (ii) greater than 1.75 but less than 2.5.

14.43 Let $\bar{X}_1$ represent the mean of a sample of size $n_1=2$, with replacement, from the finite population 3, 4, 5. Similarly, let $\bar{X}_2$ represent the mean of a sample of size $n_2 = 2$, with replacement, from the population 1, 1, 3.

(a) Find the possible differences between the sample means of the two populations.

(b) Construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$ and compute its mean and variance.

(c) Verify that $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$ and $\sigma^2_{\bar{x}_1-\bar{x}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$.

(P.U., B.A./B.Sc. 1985)

14.44 The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 years, while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 years. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B? (P.U., B.A./B.Sc. 1980)

14.45 A random sample of size 25 is taken from a normal population having a mean of 80 and a standard deviation of 5. A second random sample of size 36 is taken from a different normal population having a mean of 75 and a standard deviation of 3. Find the probability that the sample mean computed from the 25 measurements will exceed the sample mean computed from the 36 measurements by at least 3.4 but less than 5.9. Assume the means to be measured to the nearest tenth. (P.U. B.A./B.Sc. 1986-S)

14.46 What is meant by the term sampling distribution of sample proportion P? Describe its important properties and explain its usefulness in statistical inference.

14.47 A population consists of $N=6$ numbers 0, 3, 4, 6, 9 and 15. Draw all possible samples of size $n=3$, without replacement, from the population and find the sample proportion of even numbers in the samples. Construct the sampling distribution of sample proportions and verify that

$$\mu_{\hat{p}} = p \text{ and } \mathrm{Var}(\hat{P}) = \frac{pq}{n} \cdot \frac{N-n}{N-1}.$$

14.48 A population consists of $N=7$ numbers 1, 1, 2, 3, 4, 4, 5. Draw all possible samples of size $n=3$ without replacement from this population and find the sample proportion of odd numbers in the samples. Construct the sampling distribution of sample proportion and verify

$$\text{(i) } \mu_{\hat{p}} = p, \text{ (ii) } \sigma^2_{\hat{p}} = \frac{pq}{n} \cdot \left(\frac{N-n}{N-1}\right).$$ (P.U., B.A./B.Sc., 1989)

14.49 (a) Two per cent of the trees in a plantation are known to have a certain disease. What is the probability that, in a sample of 250 trees, (i) less than 1%, (ii) more than 4% are diseased?

(b) Suppose that 60% of a city population favours public finding for a proposed recreational facility. If 150 persons are to be randomly selected and interviewed, what is the probability that the sample proportion favouring this issue will be less than 0.52?

14.50 A small, professional society has $N=4500$ members. The president has mailed $n=400$ questionnaires to a random sample of members asking whether they wish to affiliate with a larger group. Assuming that the proportion of the *entire* membership favouring consolidation is $p=0.7$, find the probability that the sample proportion P differs from this by no more than 0.05.

14.51 (a) Describe the sampling distribution of differences between proportions and explain its usefulness in statistical inference.

(b) Two random samples of sizes $n_1 = 40$ and $n_2 = 45$ are drawn from a binomial population with $p=0.70$. What is the probability that $-0.1 < \hat{P}_1 - \hat{P}_2 < 0.1$?

14.52 A population consists of five observations 1, 2, 3, 4, 5. Draw all possible samples of size 2 with replacement. Find the mean of the sampling distribution of the variances. Compare it with the variance of the population. (P.U., B.A./B.Sc. 1990)

14.53 Show that the variance of the sample mean, $\bar{Y}$ from a simple random sample of size n drawn without replacement from a population of size N is given by

$$V(\bar{Y}) = \frac{\sigma^2}{n} \cdot \frac{(N-n)}{(N-1)}.$$

(P.U., M.A. Stats. 1966)

14.54 (a) State and prove the central limit theorem.

(P.U. B.A. Hons. Part II, 1970)

(b) Explain why the central limit theorem is so important in Statistical Analysis.

❖❖❖❖❖❖❖❖❖❖

15

Statistical Inference: Estimation

15.1 INTRODUCTION

The process of drawing inferences about a population on the basis of information contained in a sample taken from the population is called *Statistical Inference*. Statistical inference is traditionally divided into two major areas: *estimation of parameters* and *testing of hypothesis*.

Estimation is a procedure by which we obtain an estimate of the true but unknown value of a population parameter by using the sample observations $X_1, X_2, ..., X_n$ from the population. For example, we may estimate the mean and the variance of a population by computing the mean and the variance of a sample drawn form the population.

Testing of hypothesis is a procedure which enables us to decide on the basis of information obtained by sampling whether to accept or reject any specified statement or hypothesis regarding the value of the parameter in a statistical problem.

We shall discuss estimation in this chapter and we shall deal with testing of hypothesis in the next chapter.

15.2 ESTIMATES AND ESTIMATORS

An *estimate* is a numerical value of the unknown parameter obtained by applying a rule or a formula, called an *estimator*, to a sample $X_1, X_2, ..., X_n$ of size n, taken from the population. In other words, an *estimator* stands for the rule or method that is used to estimate a parameter whereas an *estimate* stands for the numerical value obtained by substituting the sample observations in the rule or the formula. For instance, if $X_1, X_2, ..., X_n$ is a random sample of size n from a population

with mean μ, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of μ and $\bar{x}$, t numerical value of $\bar{X}$, is an estimate of μ.

The symbol θ (the Greek letter *theta*) is customarily used to den an unknown parameter that could be a mean, median, proportion standard deviation, while an estimator of θ is commonly denoted by (read theta hat), *i.e.* by placing the hat (^) over the symbol represent the parameter or sometimes by T. It is of interest to note that estimator is always a statistic which is a function of the samp observations and hence is a random variable as the sample observatio are likely to vary from sample to sample.

There are two categories of estimates: point estimates and inter estimates. When an estimate for the unknown population parameter expressed by a sirgle value, it is called a *point estimate*. An estim expressed by a range of values within which the true value of t population parameter is believed to lie, is referred to as an *inter estimate*. Suppose we wish to estimate the average height of a very la group of students on the basis of a sample. If we find the sample aver height to be 64", then 64" is a point estimate of the unknown populat mean. If, on the other hand, we state that the true average height i value between 62" and 66", then the range of values from 62" to 66", is interval estimate.

15.3 POINT ESTIMATION

Point estimation of a population parameter provides as an estim a single value calculated from the sample that is likely to be close value to the unknown parameter. It is to be noted that a point estim will not, in general, be equal to the population parameter as the rand sample used is one of the many possible samples which could be cho from the population.

The value $\bar{x}$ (if it exists) of the statistic $\bar{X}$ computed from a sampl size n, is a point estimate of the population parameter μ. Here t parameter θ is the population mean μ and the estimator $\hat{\theta}$ is the samp mean $\bar{X}$.

If an estimate can be expressed as a sum of the weight observations (*i.e.* as a linear combination), it is said to be a *line estimate*. For example, $\bar{x}$ is a linear estimate of the parameter μ beca it can be expressed as

$$\bar{x} = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \dots + \frac{1}{n}x_n,$$

which is a linear combination of the values of X's and in terms of weights, each observation is given a weight equal to $\frac{1}{n}$.

Example 15.1. A random sample of $n = 6$ has the elements 6, 10, 13, 14, 18 and 20. Compute a point estimate of (i) the population mean, (ii) the population standard deviation, and (iii) the standard error of the mean.

(i) The sample mean is

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{6 + 10 + 13 + 14 + 18 + 20}{6} = \frac{81}{6} = 13.5$$

Thus the point estimate of the population mean μ is 13.5 and $\bar{X}$ is the estimator.

(ii) The sample standard deviation is

$$S = \sqrt{\frac{1}{n}\Sigma(X_i - \bar{X})^2}$$

estimate

$$= \sqrt{\frac{\Sigma X_i^2}{n} - \left(\frac{\Sigma X}{n}\right)^2} = \sqrt{\frac{1225}{6} - \left(\frac{81}{6}\right)^2}$$

$$= \sqrt{204.1667 - 182.25} = \sqrt{21.9167} = 4.68$$

estimate

Thus the point estimate of the population standard deviation σ is 4.68 and S is the estimator.

(iii) When the sample size is less than 5% of the population size, the standard error of the mean is

sampling chapter

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

We use the sample standard deviation S to estimate the $S.E.\ \sigma_{\bar{x}}$.

Standard deviation $\bar{x}$

Thus $$S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{4.68}{\sqrt{6}} = \frac{4.68}{2.45} = \boxed{1.91.} \rightarrow S.E(\bar{x})$$

where $S_{\bar{x}}$ is the estimator for $\sigma_{\bar{x}}$ and 1.91 is the point estimate of the standard error of the mean. → standard error of $\bar{x}$ (S.E $\bar{x}$)

15.3.1 Criteria for Good Point Estimators: A point estimator is considered a *good estimator* if it satisfies various criteria. Four of these criteria are:

(i) unbiasedness, (ii) consistency, (iii) efficiency, and (iv) sufficiency. Now we discuss these properties in turn.

(i) **Unbiasedness.** An estimator is defined to be *unbiased* if the statistic used as an estimator has its expected value equal to the true value of the population parameter being estimated. In other words, let $\hat{\theta}$ be an estimator of a parameter θ, then $\hat{\theta}$ will be called an unbiased estimator if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$, the statistic is said to be a *biased estimator*. The estimator is defined to be positively biased when $E(\hat{\theta}) > \theta$ and it is said to be negatively biased when $E(\hat{\theta}) < \theta$. Unbiasedness is a property that requires that the probability distribution of $\hat{\theta}$ be necessarily *centred* at the parameter θ, irrespective of the value of n.

Let us consider the sample mean $\bar{X}$ as an estimator of the population mean μ. Then we have $\theta = \mu$ and $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$\text{Now} \quad E(\hat{\theta}) = E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n} E[X_1 + X_2 + \ldots + X_n]$$

$$= \frac{1}{n}[E(X_1) + E(X_2) + \ldots + E(X_n)]$$

$$= \frac{1}{n}(n\mu) = \mu \qquad (\because X_i\text{'s are drawn from a population with mean } \mu).$$

Thus we see that the sample mean $\bar{X}$ is an unbiased estimator of the mean μ of any population.

The sample proportion $\hat{P}$ is also an unbiased estimator of the population parameter p as

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X)$$

$$= \frac{np}{n} = p \quad (X \text{ is the number of successes in } n \text{ trials})$$

Similarly, the sample median is also an unbiased estimator of μ when the population is normally distributed. But the sample variance S^2 is a biased estimator of σ^2 as $E(S^2) \neq \sigma^2$.

Example 15.2. Let $X_1, X_2, \ldots, X_n$ be a random sample from the population with a mean of μ and a variance of σ^2. Then show that the sample variance $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimator of σ^2.

The sample variance is defined by the relation

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= \frac{1}{n}\sum[(X_i - \mu) - (\bar{X} - \mu)]^2$$

$$= \frac{1}{n}[\sum(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum(X_i - \mu) + n(\bar{X} - \mu)^2]$$

$$= \frac{1}{n}[\sum(X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2]$$

$$[\because \sum(X_i - \mu) = n(\bar{X} - \mu)]$$

$$= \frac{1}{n}\sum(X_i - \mu)^2 - (\bar{X} - \mu)^2$$

Taking expected values of both sides, we have

$$E(S^2) = E\left[\frac{1}{n}\sum(X_i - \mu)^2 - (\bar{X} - \mu)^2\right]$$

$$= \frac{1}{n}\sum E(X_i - \mu)^2 - E(\bar{X} - \mu)^2$$

$$= \frac{1}{n}\sum \mathrm{Var}(X_i) - \mathrm{Var}(\bar{X}) = \frac{1}{n}\sum \sigma_i^2 - \frac{\sigma^2}{n}$$

$$= \sigma^2 - \frac{\sigma^2}{n} = \left(\frac{n-1}{n}\right)\sigma^2.$$

This shows that S^2 is a biased estimator of the variance σ^2. To get an unbiased estimator of σ^2, we should multiply the sample variance S^2 by $\frac{n}{n-1}$. Thus, writing $s^2 = \frac{n}{n-1}S^2$, we find

$$E(s^2) = E\left[\frac{n}{n-1}S^2\right] = \frac{n}{n-1}E(S^2) = \sigma^2$$

It then follows from this result that

$$s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

is an unbiased estimator of the population variance σ^2.

Alternatively,

$$E(s^2) = E\left[\frac{1}{n-1}\sum (X_i - \bar{X})^2\right]$$

$$= \frac{1}{n-1} E\left[\sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right]$$

$$= \frac{1}{n-1}\left[\sum E(X_i - \mu)^2 - n\,E(\bar{X} - \mu)^2\right]$$

$$= \frac{1}{n-1}\left[\sum \mathrm{Var}(X_i) - n\,\mathrm{Var}(\bar{X})\right]$$

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

It is to be noted that, for large samples, S^2 becomes an unbiased estimator of the population variance σ^2 as the amount of bias, which is $-\sigma^2/n$, becomes negligible.

Moreover, the quantity $\frac{1}{n}\sum (X_i - \mu)^2$ is an unbiased estimator of σ^2 but μ is usually not known.

Example 15.3. If random samples of size 2 are drawn with replacement from a population consisting of the five numbers 2, 3, 6, 8 and 11; show by finding all possible samples, that $\bar{X}$ and s^2 are unbiased estimators of μ and σ^2.

Population consists of 2, 3, 6, 8 and 11.

$$\therefore \quad \mu = \frac{\sum X_i}{N} = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6, \text{ and}$$

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

$$= \frac{(2-6)^2 + (3-6)^2 + \ldots + (11-6)^2}{5} = \frac{54}{5} = 10.8$$

The possible samples of size 2 which can be drawn with replacement together with the values of $\bar{X}$ and s^2 are given as follows:

No.	Sample	$\bar{X}_i$	s^2	No.	Sample	$\bar{X}_i$	s^2
1	2, 2	2	0	13	6, 6	6	0
2	2, 3	2.5	0.5	14	6, 8	7	2
3	2, 6	4	8	15	6, 11	8.5	12.5
4	2, 8	5	18	16	8, 2	5	18
5	2, 11	6.5	40.5	17	8, 3	5.5	12.5
6	3, 2	2.5	0.5	18	8, 6	7	2
7	3, 3	3	0	19	8, 8	8	0
8	3, 6	4.5	4.5	20	8, 11	9.5	4.5
9	3, 8	5.5	12.5	21	11, 2	6.5	40.5
10	3, 11	7	32	22	11, 3	7	32
11	6, 2	4	8	23	11, 6	8.5	12.5
12	6, 3	4.5	4.5	24	11, 8	9.5	4.5
				25	11, 11	11	0

Now $\bar{X}$ is an unbiased estimator of μ if $E(\bar{X}) = \mu$. To find $E(\bar{X})$, we obtain the sampling distribution of $\bar{X}$ as below:

$\bar{X}$	Tally	f_i	Probability $f(\bar{x})$	$\bar{x}f(\bar{x})$
2	I	1	1/25	2/25
2.5	II	2	2/25	5/25
3	I	1	1/25	3/25
4	II	2	2/25	8/25
4.5	II	2	2/25	9/25
5	II	2	2/25	10/25
5.5	II	2	2/25	11/25
6	I	1	1/25	6/25
6.5	II	2	2/25	13/25
7	IIII	4	4/25	28/25
8	I	1	1/25	8/25
8.5	II	2	2/25	17/25
9.5	II	2	2/25	19/25
11	I	1	1/25	11/25
Total	---	25	1	150/25

Thus $\quad E(\bar{X}) = \sum \bar{x} f(\bar{x}) = \frac{150}{25} = 6 = \mu$

Again s^2 is an unbiased estimator of the population variance σ^2 if $E(s^2) = \sigma^2$. To find $E(s^2)$, we have the sampling distribution of s^2 as follows:

s^2	f	fs^2
0	5	0
0.5	2	1
2	2	4
4.5	4	18
8	2	16
12.5	4	50
18	2	36
32	2	64
40.5	2	81
$\sum$	25	270

$$\therefore \quad E(s^2) = \frac{1}{\sum f}(\sum fs^2) = \frac{270}{25} = 10.8 = \sigma^2$$

Hence the result.

Example 15.4 A random sample of size n is drawn from the rectangular distribution

$$f(x) = \frac{1}{\theta}, \; 0 \le x \le \theta$$

Show that the statistic $2\bar{X}$ is an unbiased estimator for θ.

The statistic $2\bar{X}$ is an unbiased estimator for θ if $E(2\bar{X}) = \theta$. We know that

$$E(2\bar{X}) = 2E(\bar{X}) = 2E(X) \qquad (\because \text{ both } \bar{X} \text{ and } X_i \text{ are unbiased estimator of } \mu)$$

Now $\quad E(X) = \int_0^\theta x \cdot \frac{1}{\theta} dx = \frac{1}{\theta}\left[\frac{x^2}{2}\right]_0^\theta = \frac{\theta}{2}$

$$\therefore \quad E(2\bar{X}) = 2\left(\frac{\theta}{2}\right) = \theta$$

Hence, $2\bar{X}$ is an unbiased estimator for θ.

Example 15.5. Find an unbiased estimator for θ^2, where θ is the parameter in a binomial distribution.

Let the random variable X have the binomial distribution with parameters n and $\theta(=p)$,

Then $E(X) = n\theta$ and $\text{Var}(X) = n\theta(1-\theta) = n\theta - n\theta^2$.

But $\text{Var}(X) = E(X^2) - [E(X)]^2$, so that

$$E(X^2) = \text{Var}(X) + [E(X)]^2$$
$$= n\theta - n\theta^2 + n^2\theta^2$$

or $$E(X^2) - n\theta = n^2\theta^2 - n\theta^2 = n(n-1)\,\theta^2$$

$$\therefore \quad \theta^2 = \frac{E(X^2) - E(X)}{n\,(n-1)} = \frac{E\,[X(X-1)]}{n\,(n-1)} \qquad [n\theta = E(X)]$$

which is the unbiased estimator of θ^2.

Example 15.6. Let the sample linear regression be $Y_i = \hat{\alpha} + \hat{\beta}X_i + \varepsilon_i$, where each Y_i is normally distributed and X_i's are fixed. Then show that $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of the parameters α and β.

The sample regression coefficient may be expressed as

$$\hat{\beta} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{n\sum X\,(\alpha + \beta X + \varepsilon) - \sum X \sum(\alpha + \beta X + \varepsilon)}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{n\alpha\sum X + n\beta\sum X^2 + n\sum X\varepsilon - n\alpha\sum X - \beta(\sum X)^2 - \sum X\sum\varepsilon}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{\beta\,[n\sum X^2 - (\sum X)^2] + n\sum X\varepsilon - n\sum X\varepsilon}{n\sum X^2 - (\sum X)^2}$$

$$= \beta + \frac{\sum (X - \bar{X})\,\varepsilon}{\sum (X - \bar{X})^2}$$

Taking expected values of both sides, we get

$$E(\hat{\beta}) = E\left[\beta + \frac{\sum (X - \bar{X})\,\varepsilon}{\sum (X - \bar{X})^2}\right]$$

$$= \beta + \frac{\sum (X - \bar{X})\,E(\varepsilon)}{\sum (X - \bar{X})^2} \quad (\because X_i\text{'s are treated as constant w.r.t. this expectation})$$

$$= \beta \qquad [E(\varepsilon) = 0]$$

Thus $\hat{\beta}$ is an unbiased estimator for β.

Now $E(\hat{\alpha}) = E(\bar{Y} - \hat{\beta}\bar{X})$ $\quad (\because \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X})$

$$= E\left[\frac{\sum Y_i}{n} - \hat{\beta}\bar{X}\right] = \frac{1}{n}\sum E(Y_i) - \bar{X}E(\hat{\beta})$$

$$= \frac{\sum(\alpha + \beta X_i)}{n} - \bar{X}\beta \qquad [\because E(Y_i) = \alpha + \beta X_i]$$

$$= \alpha + \beta\bar{X} - \beta\bar{X} = \alpha.$$

Hence, $\hat{\alpha}$ is also an unbiased estimator for α.

(ii) Consistency. An estimator is said to be *consistent* if the statistic to be used as estimator becomes closer and closer to the population parameter being estimated as the sample size n increases. Stated a little differently, an estimator $\hat{\theta}$ is called a consistent estimator of θ if the probability that $\hat{\theta}$ becomes closer and closer to θ, approaches unity with the increasing sample size. Symbolically, $\hat{\theta}$ is a consistent estimator of the parameter θ if, for any arbitrarily small positive quantity e,

$$\lim_{n \to \infty} P[\,|\hat{\theta} - \theta| \le e\,] = 1.$$

A consistent estimator may or may not be unbiased. The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is an unbiased estimator of μ, is a consistent estimator of the mean μ. The sample proportion $\hat{P}$ is also a consistent estimator of the parameter p of a population that has a binomial distribution. The median is not a consistent estimator of μ when the population has a skewed distribution. The sample variance $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X - \bar{X})^2$, though a biased estimator, is a consistent estimator of the population variance σ^2. It may be shown that a statistic whose standard error decreases with the increasing sample size, will be consistent. It should be noted that consistency is a large sample property.

To prove that an estimator is consistent, we may state a criterion that is sometimes quite useful, as follows:

"Let $\hat{\theta}$ be an estimator of θ based on a sample of size n. Then $\hat{\theta}$ is a consistent estimator of θ, if $\mathrm{Var}(\hat{\theta}) \to 0$ as $n \to \infty$."

To illustrate, let us consider $\bar{X}$, the sample mean, based on a random sample of size n. We know that

$$E(\bar{X}) = \mu \text{ and } \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Now as $n \to \infty$, $\frac{\sigma^2}{n}$ approaches 0, *i.e.* $\lim_{n\to\infty} \mathrm{Var}(\bar{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$.

Hence $\bar{X}$ is an unbiased and a consistent estimator of μ.

Also, for the sample proportion $\hat{P}\left(=\frac{X}{n}\right)$, we have seen that $E(\hat{P}) = E\left(\frac{X}{n}\right) = p$, *i.e.* $\hat{P}$ is an unbiased estimator for p.

$$\text{Now} \quad \mathrm{Var}(\hat{P}) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X)$$

$$= \frac{1}{n^2}(npq) \qquad (\because X \text{ is binomially distributed})$$

$$= \frac{pq}{n}$$

Since the value of $\frac{pq}{n}$ approaches zero as n tends to ∞, therefore $\hat{P}\left(=\frac{X}{n}\right)$ is a consistent estimator.

Hence, the sample proportion $\hat{P}\left(=\frac{X}{n}\right)$ is an unbiased and a consistent estimator for the population proportion p.

(iii) **Efficiency.** An unbiased estimator is defined to be *efficient* if the variance of its sampling distribution is smaller than that of the sampling distribution of any other unbiased estimator of the same parameter. In other words, suppose there are two unbiased estimators T_1 and T_2 of the same parameter θ, then T_1 will be said to be *more efficient* estimator than T_2 if $\mathrm{Var}(T_1) < \mathrm{Var}(T_2)$. The relative efficiency of T_1 compared to T_2 is given by the ratio $E_f = \frac{\mathrm{Var}(T_2)}{\mathrm{Var}(T_1)}$, which is greater than 1. It thus provides a criterion for comparing different unbiased estimators of a parameter.

Both the sample mean and the sample median for a population that has a normal distribution, are unbiased and consistent estimators of μ, but the variance of the sampling distribution of sample means $\left(\frac{\sigma^2}{n}\right)$ is

smaller than that of the sampling distribution of sample medians, $\left(\frac{\pi\sigma^2}{2n}\right)$

i.e. $\frac{\text{Var (median)}}{\text{Var }(\bar{X})} = \frac{\pi\sigma^2}{2n} \times \frac{n}{\sigma^2} = \frac{\pi}{2} = 1.57 > 1$. Hence the sample mean is more *efficient* than the sample median as an estimator of μ, the sample mean may therefore be preferred as an estimator. The efficiency of the sample mean relative to sample median is 1.57 or 157%, which means that a sample mean calculated from a sample of size 100 can do the same job as the sample median calculated from a sample of size 157.

If an unbiased estimator $\hat{\theta}$ has smaller variance than *any other* unbiased estimator, it is called the *minimum variance unbiased estimator* of θ. We generally prefer the unbiased estimator which has the minimum variance. An unbiased estimator having the minimum variance, is called the *best* or *most efficient* estimator for θ. It is of interest to note that the variance of $\hat{\theta}$ cannot becomes smaller than a certain lower bound and a variance equal to this lower bound is called the minimum variance. This result is given by the Cramer-Rao inequality. The sample mean $\bar{X}$ is the minimum variance unbiased estimator of μ for a normal distribution with known variance.

An estimator $\hat{\theta}$ that is linear, is unbiased and has minimum variance among all linear unbiased estimators of θ, is called a *best linear unbiased estimator* or *BLUE* for short.

The efficiency of biased estimators of θ, is compared on the basis of *Mean Square Error*, (abbreviated *MSE*) which is defined by the expected value of the squared differences between the estimator and the true value, *i.e.* the parameter. Mathematically speaking, if $\hat{\theta}$ is any estimator, then the *MSE* about the true value θ would be

$$\begin{aligned} MSE\,(\hat{\theta}) &= E\,(\hat{\theta} - \theta)^2 \\ &= E\,[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E\,[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2\,[E(\hat{\theta}) - \theta]E[\hat{\theta} - E(\hat{\theta})] \\ &= \text{Var}\,(\hat{\theta}) + (\text{Bias})^2 \qquad (\because E[\hat{\theta} - E(\hat{\theta})] = 0) \end{aligned}$$

That is, the mean square error of $\hat{\theta}$ is equal to the variance of the estimator plus the squared bias. If $E(\hat{\theta}) = \theta$, then $E(\hat{\theta} - \theta)^2$ is equal to variance of $\hat{\theta}$, that is in case of unbiased estimator, the variance and *MSE* would coincide. It is to be noted that *MSE* is an important criterion for comparing two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$. We shall prefer an estimator which gives the smaller *MSE* about the parameter to be estimated.

Example 15.7 Let X_1, X_2, X_3 and X_4 be a random sample of size $n = 4$ from a $N(\mu, \sigma^2)$. A statistician wishes to estimate the mean by using either of the following two estimators of the mean μ

$$T_1 = \frac{X_1 + X_2 + X_3 + X_4}{4}, \quad \text{(the sample mean, } \bar{X}\text{)}$$

$$T_2 = \frac{X_1 + 2X_2 + 3X_3 + X_4}{7}, \quad \text{(a weighted mean, } \bar{X}_w\text{)}$$

Which estimator should be preferred?

To answer this question correctly, we need to compare the expected values and the variances of these two statistics.

Let us first see if the two estimators are unbiased.

$$\text{Now} \quad E(T_1) = E\left[\frac{X_1 + X_2 + X_3 + X_4}{4}\right]$$

$$= \frac{1}{4} E\,[X_1 + X_2 + X_3 + X_4] = \frac{1}{4}(4\mu) = \mu.$$

$\therefore$ T_1, the sample mean $\bar{X}$, is an unbiased estimator.

$$\text{Again} \quad E(T_2) = E\left[\frac{X_1 + 2X_2 + 3X_3 + X_4}{7}\right]$$

$$= \frac{1}{7}[E(X_1) + 2E(X_2) + 3E(X_3) + E(X_4)]$$

$$= \frac{1}{7}[\mu + 2\mu + 3\mu + \mu] = \mu, \text{ i.e.}$$

T_2, the weighted mean of the observations, is also an unbiased estimator of μ. Thus we see that both means are unbiased estimators of μ.

Next, we find their variances.

$$\text{Var}(T_1) = \text{Var}\left[\frac{X_1 + X_2 + X_3 + X_4}{4}\right] = \frac{\sigma^2}{4}.$$

$$\text{Var}(T_2) = \text{Var}\left[\frac{X_1 + 2X_2 + 3X_3 + X_4}{7}\right]$$

$$= \text{Var}\left(\frac{X_1}{7}\right) + \text{Var}\left(\frac{2X_2}{7}\right) + \text{Var}\left(\frac{3X_3}{7}\right) + \text{Var}\left(\frac{X_4}{7}\right)$$

$$= \frac{1}{49}\text{Var}(X_1) + \frac{4}{49}\text{Var}(X_2) + \frac{9}{49}\text{Var}(X_3) + \frac{1}{49}\text{Var}(X_4)$$

$$= \frac{1}{49}\sigma^2 + \frac{4}{49}\sigma^2 + \frac{9}{49}\sigma^2 + \frac{1}{49}\sigma^2 = \frac{15}{49}\sigma^2.$$

Now $\frac{\text{Var}(T_2)}{\text{Var}(T_1)} = \frac{15}{49}\sigma^2 \div \frac{\sigma^2}{4} = \frac{60}{49}$, which is greater than 1, showing that $\text{Var}(T_1) < \text{Var}(T_2)$. Thus T_1 is more efficient than T_2 as an estimator for μ.

Hence $T_1 = \bar{X}$ is a *better* estimator of μ than $T_2 = \bar{X}_w$ and should be preferred.

(iv) **Sufficiency.** An estimator is defined to be *sufficient*, if the statistic used as estimator uses all the information that is contained in the sample. Any statistic that is not computed from all values in the sample is not a sufficient estimator. The sample mean $\bar{X}$ is a sufficient estimator of μ. This implies that $\bar{X}$ contains all the information in the sample relative to the estimation of the population parameter μ and no other estimator such as the sample median, etc. calculated from same sample can add any useful information concerning μ.

The sample proportion $\hat{P}$ is also a sufficient estimator of the population proportion p. The concept of a sufficient estimator introduced by Sir R.A. Fisher (1890–1962), is of paramount importance in Statistics.

Practical Criteria for Sufficiency. A necessary and sufficient condition for a statistic $\hat{\theta}$ to be a sufficient estimator for an unknown parameter θ is that the joint distribution of the sample of n values can be factored into two non-negative functions (factors), one depending on $\hat{\theta}$ and θ, and the other depending on X's but not on θ.

Mathematically, let $X_1, X_2, \ldots, X_n$ be a random sample from a probability distribution $f(x; \theta)$ and let $f(x_1, x_2, \ldots, x_n; \theta)$ denote its joint distribution. Then $\hat{\theta}$ $[= h(x_1, x_2, \ldots, x_n)]$ is a sufficient estimator for θ if and only if

$$f(x_1, x_2, \ldots, x_n; \theta) = g(\hat{\theta}; \theta)\, h(x_1, x_2, \ldots, x_n)$$

where $g(\hat{\theta}; \theta)$ = probability function of the estimator $\hat{\theta}$, and

$h(x_1, \ldots, x_n)$ = a function of $(x_1, x_2, \ldots, x_n)$ not depending on θ.

This is known as the Fisher-Neyman *Factorization criterion* for sufficiency.

We demonstrate the technique by the following examples:

Example 15.8 Let $X_1, X_2, \ldots, X_n$ be a random sample from the Poisson distribution $f(x; \theta) = \frac{e^{-\theta}\theta^x}{x!}$, $(x = 0, 1, 2, \ldots)$. Then show that $\bar{X}$ is a sufficient estimator for θ.

Given the Poisson distribution

$$f(x;\theta) = \frac{e^{-\theta}\,\theta^x}{x!} \quad \text{for } x = 0, 1, 2, \ldots \qquad (\theta \text{ is used for } \mu)$$

The joint distribution of the sample values is

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta)\, f(x_2;\theta) \ldots f(x_n;\theta)$$

$$= \prod_{i=1}^{n} \frac{e^{-\theta}\,\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\,\theta^{\Sigma x}}{x_1!\,x_2! \ldots x_n!}$$

$$= \frac{e^{-n\theta}\,(n\theta)^{\Sigma x}}{(\Sigma x_i)!} \cdot \frac{(\Sigma x_i)!}{x_1!\,x_2! \ldots x_n!\; n^{\Sigma x_i}}$$

Now if we write

$g(\hat{\theta};\theta) = \dfrac{e^{-n\theta}\,(n\theta)^{\Sigma x_i}}{(\Sigma x_i)!}$, a distribution of Σx_i that depends on θ, and

$h(x_1, \ldots, x_n) = \dfrac{(\Sigma x_i)!}{x_1!\,x_2! \ldots x_n!}\,\dfrac{1}{n^{\Sigma x_i}}$, a function not depending on θ;

we see that the joint distribution has been split up into two functions, which satisfy the factorization criterion. This shows that ΣX_i is a sufficient estimator for θ.

Hence, $\bar{X}$ is a sufficient estimator for θ as any one-to-one function of $\hat{\theta}$ is also sufficient.

Example 15.9. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean $=\mu$ and variance $=1$. Show that $\bar{X} = \dfrac{\Sigma x_i}{n}$, the mean of the sample is a sufficient statistic for the parameter μ.

The normal distribution with mean $=\mu$ and $\sigma^2 = 1$, is given by

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}}\, e^{-(x-\mu)^2/2}$$

The joint distribution for the sample from the population is

$$f(x_1, x_2, \ldots, x_n; \mu) = \frac{1}{\sqrt{2\pi}} e^{-(x_1-\mu)^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(x_2-\mu)^2/2} \ldots \frac{1}{\sqrt{2\pi}} e^{-(x_n-\mu)^2/2}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2}$$

$$= \frac{1}{(2\pi)^{n/2}}\, e^{-\Sigma(x_i-\mu)^2/2}$$

We can write

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-\mu)]^2$$

$$= \sum(x_i-\bar{x})^2 + \sum(\bar{x}-\mu)^2$$

$$= \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$$

$$\therefore f(x_1, x_2, \ldots, x_n; \mu) = \frac{1}{(2\pi)^{n/2}} e^{-n(\bar{x}-\mu)^2/2} \cdot e^{-\Sigma(x-\bar{x})^2/2}$$

$$= g(\bar{x}, \mu)\, h(x_1, x_2, \ldots, x_n),$$

where $g(\bar{x}, \mu)$ depends on $\bar{x}$ (estimator) and μ, (parameter) a $h(x_1, \ldots, x_n)$ depends on x and $\bar{x}$ but not on μ.

Thus the joint distribution has been factored into two fac satisfying the factorisation criterion.

Hence, $\bar{X}$ is a sufficient statistic for the parameter μ.

15.3.2. Pooled Estimators from Two or More Samples. Ot we need to estimate certain parameters by pooling (combining) values from two or more random samples taken usually from the sa population.

Let us first consider two random samples of sizes n_1 and n_2 fro population with unknown mean μ and unknown variance σ^2. If $\bar{X}_1$ a $\bar{X}_2$ are the two sample means, then the combined mean given by

$$\bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1+n_2}$$

is an unbiased estimator for μ, since

$$E(\bar{X}_c) = E\left[\frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1+n_2}\right] = \frac{1}{n_1+n_2}E[n_1\bar{X}_1 + n_2\bar{X}_2]$$

$$= \frac{1}{n_1+n_2}[n_1E(\bar{X}_1) + n_2E(\bar{X}_2)]$$

$$= \frac{1}{n_1+n_2}[n_1\mu + n_2\mu] = \mu$$

In order to find an unbiased estimator for the population var σ^2, based on the two samples with variances S_1^2 and S_2^2, we consid following:

We have already seen that the statistic $\left(\frac{n}{n-1}\right) S^2$ is an unbiased estimator for σ^2. This implies that

$$E\left[\frac{n_1 S_1^2}{n_1 - 1}\right] = \sigma^2 \text{ or } E(n_1 S_1^2) = (n_1 - 1)\sigma^2, \text{ and}$$

$$E\left[\frac{n_2 S_2^2}{n_2 - 1}\right] = \sigma^2 \text{ or } E(n_2 S_2^2) = (n_2 - 1)\sigma^2$$

Adding these two results, we get

$$E(n_1 S_1^2) + E(n_2 S_2^2) = (n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2$$

or $$E[n_1 S_1^2 + n_2 S_2^2] = (n_1 + n_2 - 2)\sigma^2$$

so $$E\left[\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\right] = \sigma^2$$

Hence, on the basis of two samples, the statistic $(n_1 S_1^2 + n_2 S_2^2)/(n_1 + n_2 - 2)$ is an unbiased estimator for σ^2. This statistic is generally denoted by S_p^2 and is called the *pooled unbiased estimator* for the variance σ^2.

These results can be easily extended to three or more random samples taken from the same population. For example, if we take three random samples, then the unbiased estimator for μ is

$$\bar{X}_c = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + n_3 \bar{X}_3}{n_1 + n_2 + n_3},$$

and the pooled unbiased estimator for the variance σ^2, is

$$S_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2 + n_3 S_3^2}{n_1 + n_2 + n_3 - 3}$$

Likewise, if we take two random samples of sizes n_1 and n_2 and sample proportions P_1 and P_2, from a binomial population with unknown proportion p of successes, then the unbiased estimator for p, based on pooled data, is given by

$$\hat{P}_c = \frac{n_1 \hat{P}_1 + n_2 \hat{P}_2}{n_1 + n_2}, \text{ as}$$

$$E\left[\frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1 + n_2}\right] = \frac{1}{n_1 + n_2}E[n_1\hat{P}_1 + n_2\hat{P}_2]$$

$$= \frac{1}{n_1 + n_2}[n_1E(\hat{P}_1) + n_2E(\hat{P}_2)]$$

$$= \frac{1}{n_1 + n_2}[n_1p + n_2p] = p.$$

We now give an illustrative example.

Example 5.10. A sample of 40 observations from a population with unknown mean μ and unknown variance σ^2 gave $\sum X = 800$ and $\sum X^2 = 16052$.

A second sample taken from the same population gave

x	18	19	20	21	22	23
f	12	23	10	7	5	3

Using the data given by the two samples, find unbiased estimates of μ and σ^2.

We first compute the sample means and sample variances.

For sample 1:

$$\bar{x}_1 = \frac{\sum X_i}{n_1} = \frac{800}{40} = 20.0$$

$$S_1^2 = \frac{\sum X_i^2}{n_1} - \left(\frac{\sum X_i}{n_1}\right)^2 = \frac{16052}{40} - \left(\frac{800}{40}\right)^2$$

$$= 401.3 - 400.0 = 1.3$$

For sample 2:

$$\bar{x}_2 = \frac{\sum fx_i}{n_2} = \frac{1179}{60} = 19.65$$

$$S_2^2 = \frac{\sum fx_i^2}{n_2} - \left(\frac{\sum fx_i}{n_2}\right)^2 = \frac{23285}{60} - \left(\frac{1179}{60}\right)^2$$

$$= 388.0833 - 386.1225 = 1.96$$

Now, we calculate, using the information given by the two samples, unbiased estimates of μ and σ^2.

An unbiased estimate fo μ is the combined mean $\bar{x}_c$, where

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{(40)\,(20.0) + (60)\,(19.65)}{40 + 60} = 19.79,$$

An unbiased estimate for σ^2 is s_p^2, where

$$s_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \frac{(40)\,(1.3) + (60)\,(1.96)}{40 + 60 - 2} = 1.73$$

Hence, on the basis of two samples, an unbiased estimate of μ is 19.79, and an unbiased estimate of σ^2 is 1.73.

15.4 METHODS OF POINT ESTIMATION

A point estimator of a parameter can be obtained by several methods but we shall consider the following three methods only.

(i) The Method of Maximum Likelihood.

(ii) The Method of Moments.

(iii) The Method of Least-Squares.

These methods give estimates which may differ as the methods are based on different theories of estimation.

15.4.1. The Method of Maximum Likelihood. The *method or the principle of maximum likelihood*, abbreviated *ML*, which is a very useful method of estimation, was introduced in 1922 by Sir Ronald A. Fisher (1890–1962). The principle and the underlying logic is "*to consider every possible value that the parameter might have, and for each value, compute the probability that the given sample would have occurred if that were the true value of the parameter. That value of the parameter for which the probability of a given sample is greatest, is chosen as an estimate.*" Estimates obtained by this method are called the *maximum likelihood estimates* (*MLE*). The *ML* estimators are consistent, efficient and sufficient but not necessarily unbiased. The method of maximum likelihood is applicable to both discrete and continuous random variables.

To illustrate, let us consider a sample of $n = 10$ rocks from a binomial population with unknown parameter p, where p denotes the proportion of rocks in a riverbed that are sedimentary in type. Suppose that $X = 4$ rocks are found to be sedimentary in type. Now the problem is to estimate the unknown parameter p from the sample data.

For the given sample, both $n(=10)$ and $X(=4)$ are fixed and the only variable is the value of p. Using the principle of *ML*, we consider all possible value of $p(p = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)$ and

calculate the probability for each possible value in order to choose as an estimate that value of p for which the probability of the given sample is greatest.

By the binomial formula, the probability of $X=4$ when $n=10$, given p, is

$$f(4, p) = \binom{10}{4} p^4 (1-p)^6$$

Evaluating the expression for all possible values of p, we obtain

p	$f(4, p)$	
0.0	0.000	
0.1	0.011	
0.2	0.088	
0.3	0.200	
0.4	0.251	← maximum likelihood
0.5	0.205	
0.6	0.111	
0.7	0.036	
0.8	0.005	
0.9	0.000	
1.0	0.000	

Of all possible values of p, the probability of the actual sample drawn corresponding to the value $p=0.4$ is greatest. In other words, $p=0.4$ makes the sample result *most likely*. Hence $\hat{p} = 0.4$ is the maximum likelihood estimate.

Mathematically, the position of the maximum value can be found by equating the derivative of

$$f(4, p) = \binom{10}{4} p^4 (1-p)^6$$

w.r.t. p to zero and solving for p. Thus

$$\frac{\partial}{\partial p} f(4, p) = \binom{10}{4} [4p^3 (1-p)^6 - 6p^4 (1-p)^5] = 0$$

Or $\quad 2p^3 (1-p)^5 [2(1-p) - 3p] = 0$

w either $p^3 = 0$, which gives $p = 0$;

$(1-p)^5 = 0$, which gives $p = 1$;

or $\quad 2(1-p) - 3p = 0$, which gives $p = \frac{2}{5}$.

We see that the first two values give a minimum, and therefore our estimate is $\hat{p} = \frac{2}{5} = 0.4$.

The mathematical technique of finding *ML* estimators is presented as below:

Let $X_1, X_2, \ldots, X_n$ be a random sample from a probability distribution $f(x; \theta)$, where θ is a single unknown parameter. Then the joint distribution for $X_1, X_2, \ldots, X_n$ is

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta)\, f(x_2; \theta) \ldots f(x_n; \theta)$$

$$= \prod_{i=1}^{n} f(x_i, \theta),$$

as the observations in a random sample are statistically independent.

This joint distribution regarded as a function of the parameter θ, is called the *Likelihood Function* of the sample and is usually denoted by the symbol $L(\theta)$, *i.e.* $L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$

Now we wish to estimate the unknown parameter θ by that value which maximizes the *likelihood function* $L(\theta)$. Such a value, called the *ML estimate*, is obtained by solving for θ the equation $\frac{\partial L(\theta)}{\partial \theta} = 0$ for which $\frac{\partial^2 L(\theta)}{\partial \theta^2} < 0$, provided that $L(\theta)$ is a differentiable function of θ.

In practice, it is often more convenient to find the maximum of the natural logarithm (ln) of $L(\theta)$ instead of $L(\theta)$ because of the fact that the maximum of $ln\ L(\theta)$ occurs at the same point (value) as the maximum of $L(\theta)$. Then the *ML* estimator is the solution of $\frac{\partial}{\partial \theta}[ln\ L(\theta)] = 0$ for which $\frac{\partial^2}{\partial \theta^2}[ln\ L(\theta)] < 0$.

We have considered only one parameter θ. This method may be extended to two or more unknown parameters.

The following examples illustrate the application of the method of *ML*.

Example 15.11 Consider the geometric distribution

$$f(x) = pq^{x-1} \text{ for } x = 1, 2, 3, \ldots$$

Find the maximum likelihood estimator for p when

(i) only one experiment is performed,

(ii) n experiments are performed.

We find the maximum likelihood estimators for p in both situations as follows:

(i) Since only one experiment is being performed so the *likelihood function* would be

$$L(p) = f(x; p) = p(1-p)^{x-1}$$

The natural logarithm (ln) of the likelihood function is

$$ln\, L(p) = ln\, p + (x-1)\, ln\, (1-p)$$

Differentiating w.r.t. p, we obtain

$$\frac{\partial}{\partial p}[ln\, L(p)] = \frac{1}{p} - \frac{x-1}{1-p}$$

Equating to zero, we get

$$\frac{1}{p} - \frac{x-1}{1-p} = 0 \text{ or } \frac{(1-p) - p(x-1)}{p\,(1-p)} = 0$$

or $\quad 1 - px = 0$ gives $\hat{p} = \frac{1}{x}$.

(ii) The *likelihood function* for n experiments would be

$$L(p) = f(x_1; p)\, f(x_2, p) \ldots f(x_n; p)$$

$$= \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\Sigma x_i - n}$$

The natural logarithm (ln) of the likelihood function is

$$ln\, L(p) = n\, ln\, p + (\Sigma x_i - n)\, ln\, (1-p)$$

Differentiating w.r.t. p, we get

$$\frac{\partial}{\partial p}[ln\, L(p)] = \frac{n}{p} - \frac{\Sigma x_i - n}{1-p}$$

Equating to zero and solving for p, we obtain

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

Hence, the *MLE* of p is the reciprocal of the mean of the number of trials

Example 15.12. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n taken from the Bernoulli distribution

$$f(x) = p^x q^{1-x}, \quad x = 0, 1; 0 \le p \le 1$$

Find the maximum likelihood estimator of p.

The *likelihood function* for a sample of size n with values $X_1, X_2, \ldots, X_n$ is

$$L(p) = \prod_{i=1}^{n} p^{x_i} q^{1-x_i} \qquad \text{(Here } \theta = p\text{)}$$

$$= p^{\Sigma x_i} (1-p)^{n-\Sigma x_i}$$

The natural logarithm of the likelihood function is

$$ln\, L(p) = \Sigma x\, (ln\, p) + (n - \Sigma x)\, [ln\, (1-p)]$$

Differentiating *w.r.t.* p and equating to zero, we get

$$\frac{d}{dp}[ln\, L(p)] = \frac{\Sigma x}{p} + (n - \Sigma x)\left(\frac{-1}{1-p}\right) = 0$$

Solving for p, we get

$$(1-p)\,\Sigma x - p(n - \Sigma x) = 0$$

or $$\hat{p} = \frac{\Sigma x}{n} = \bar{x}$$

which is the *MLE* of p.

Example 15.13. Let $X_1, X_2, \ldots, X_n$ denote a random sample from Poisson distribution, $f(x; \lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$ Find the maximum likelihood estimator of λ. (P.U. B.A./B.Sc., 1987, 93)

The probability function of the Poisson distribution is

$$f(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \text{ for } x = 0, 1, 2, \ldots$$

The *likelihood function* for a sample of size n with values $X_1, X_2, \ldots, X_n$ is

$$L(\theta) = f_1(x_1; \lambda)\, f_2(x_2; \lambda) \ldots f_n(x_n; \lambda)$$

$$= \frac{e^{-n\lambda}(\lambda)^{\Sigma x}}{x_1!\, x_2! \ldots x_n!}$$

Taking natural logarithm (ln) of the likelihood function, we obtain

$$ln\, L(\theta) = -n\lambda + \Sigma x\, (ln\, \lambda) - ln\, (x_1!) - ln\, (x_2!) \ldots$$

Differentiating w.r.t. λ, we get

$$\frac{\partial\,[ln\,L(\theta)]}{\partial\lambda} = -n + \frac{\sum x}{\lambda}$$

Equating to zero and solving for λ, we obtain

$$\hat{\lambda} = \frac{\sum x}{n} = \bar{x}, \text{ the sample mean.}$$

Example 15.14. Assume that the random variable X has the exponential distribution

$$f(x;\theta) = \theta\, e^{-\theta x}, \quad x > 0, \theta > 0.$$

where θ is the parameter of the distribution. Use the method of *ML* to estimate θ if five observations of X were $x_1 = 0.9$, $x_2 = 1.7$, $x_3 = 0.4$, $x_4 = 0.3$ and $x_5 = 2.4$.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n taken from the exponential distribution

$$f(x,\theta) = \theta\, e^{-\theta x}, \quad x > 0, \theta > 0.$$

Then the *likelihood function* for the given sample is

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \theta^n\, e^{-\theta \sum x_i}$$

The natural logarithm (ln) of the likelihood function is

$$ln\,L(\theta) = n\,ln\,\theta - \theta \sum x_i$$

Differentiating *w.r.t.* θ and equating to zero, we get

$$\frac{d}{d\theta}[ln\,L(\theta)] = \frac{n}{\theta} - \sum x_i = 0$$

Solving, we get $\hat{\theta} = \dfrac{n}{\sum x_i}$ as the *MLE* of θ.

Now consider the given sample data.

$$n = 5 \text{ and } \sum x_i = 0.9 + 1.7 + 0.4 + 0.3 + 2.4 = 5.7$$

Substituting these values, we obtain

$$\hat{\theta} = \frac{5}{5.7} = 0.88 \text{ as the desired estimate of } \theta.$$

Example 15.15 Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with parameters μ and σ^2, find *MLE* for

(i) μ, when population variance σ^2 is known,

(ii) σ^2, when population mean μ is known,

(iii) μ and σ^2 simultaneously.

The density function of a normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

The *likelihood function* for a sample of size n with values $X_1, X_2, \ldots, X_n$ is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} . e^{-\Sigma(x_i-\mu)^2/2\sigma^2}$$

The natural logarithm of the likelihood function is

$$ln\, L(\theta) = -\frac{n}{2} ln\,(2\pi) - \frac{n}{2} ln\,(\sigma^2) - \frac{1}{2\sigma^2}\Sigma(x_i - \mu)^2$$

(i) Here $\theta = \mu$ as σ^2 is known. Therefore differentiating it *w.r.t.* μ and equating to zero, we get

$$\frac{\partial}{\partial\mu}[ln\, L(\mu)] = 0 - \frac{1}{2\sigma^2}[2\,\Sigma(x_i - \mu)\,(-1)] = 0$$

which gives $\hat{\mu} = \frac{\Sigma x_i}{n}$.

Again $\frac{\partial^2}{\partial\mu^2}[ln\, L(\mu)] = -\frac{n}{\sigma^2} < 0.$

Thus the *MLE* of μ is $\bar{X} = \frac{\Sigma x_i}{n}$, the sample mean.

(ii) Here $\theta = \sigma^2$ as μ is known.

$\therefore$ Differentiating *w.r.t* σ^2 and equating to zero, we get

$$\frac{\partial}{\partial\sigma^2}[ln\, L(\sigma^2)] = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma(x_i - \mu)^2 = 0$$

Solving for σ^2, we find

$$\hat{\sigma}^2 = \frac{1}{n}\Sigma(x_i - \mu)^2.$$

(iii) Here $\theta = (\mu, \sigma^2)$. To get joint *ML* estimators for μ and σ^2, we have to solve together the two equations

$$\frac{\partial}{\partial \mu}[\ln L(\mu, \sigma^2)] = 0 \text{ and } \frac{\partial}{\partial \sigma^2}[\ln L(\mu, \sigma^2)] = 0$$

Now $\frac{\partial}{\partial \mu}[\ln L(\mu, \sigma^2)] = 0$ gives $-\frac{1}{2\sigma^2}[2\sum(x_i - \mu)(-1)] = 0$,

and $\frac{\partial}{\partial \sigma^2}[\ln L(\mu, \sigma^2)] = 0$ gives $-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i - \mu)^2 = 0$.

Solving these equations for μ and σ^2, we get

$$\hat{\mu} = \frac{\sum x}{n} = \bar{X}, \text{ and}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2, \text{ where } \mu = \bar{x} \text{ is substituted.}$$

Thus the joint *ML* estimators for μ and σ^2 are the sample mean $\bar{X}$, and the sample variance S^2, which is not an unbiased estimator of σ^2.

15.4.2. The Method of Moments. The *method of moments* which is due to Karl Pearson (1857–1936), consists of calculating a few moments of the sample values and equating them to the corresponding moments of a population, thus getting as many equations as are needed to solve for the unknown parameters. The moment estimators are, in general, inefficient.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n. Then the *rth* sample moment about zero is

$$m'_r = \frac{\sum X_i^r}{n}, \; r = 1, 2, \ldots$$

and the corresponding *rth* population moment is μ'_r. We then match these moments and get as many equations as we need to solve for the unknown parameters.

The following examples illustrate the method:

Example 15.16. Let X be uniformly distributed on the interval $(0, \theta)$. Find an estimator of θ by the method of moments.

Since the distribution has only one parameter, we therefore need to take only one equation.

The first sample moment about zero is $m'_1 = \frac{\Sigma X_i}{n}$. The first population moment about zero is

$$\mu'_1 = \int_0^\theta x \cdot \frac{1}{\theta} dx = \frac{1}{\theta}\left[\frac{x^2}{2}\right]_0^\theta = \frac{\theta}{2}$$

Matching these moments, we get

$$\frac{\Sigma X_i}{n} = \frac{\theta}{2} \text{ or } \hat{\theta} = 2\bar{X}$$

Hence, the moment estimator of θ is just twice the sample mean.

Example 15.17 Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal population with parameters μ and σ^2. Find these parameters by the method of moments.

Here we need two equations as there are two unknown parameters. The first two sample moments about zero are

$$m'_1 = \frac{1}{n}\Sigma X_i = \bar{X} \text{ and } m'_2 = \frac{1}{n}\Sigma X_i^2$$

The corresponding two moments of a normal distribution are

$$\mu'_1 = \mu \text{ and } \mu'_2 = \sigma^2 + \mu^2 \qquad (\because \sigma^2 = \mu'_2 - \mu'^2_1)$$

To get the desired estimators by the method of moments, we match them. Thus

$$\mu = \frac{1}{n}\Sigma X_i \text{ and } \sigma^2 + \mu^2 = \frac{1}{n}\Sigma X_i^2$$

Solving, we get

$$\hat{\mu} = \frac{1}{n}\Sigma X_i = \bar{X}, \text{ and}$$

$$\hat{\sigma}^2 = \frac{\Sigma X^2}{n} - \bar{X}^2 = \frac{1}{n}\Sigma(X - \bar{X})^2 = S^2$$

as the moment estimators for μ and σ^2.

15.4.3. The Method of Least-Squares. The method of *Least Squares*, abbreviated *LS*, which is due to Gauss (1777–1855) and Markov (1856-1922), is based on the theory of *linear estimation*. An estimator found by minimizing the sum of squares of deviations of the sample values from some function that has been hypothesized as a fit for the data, is called the *least squares estimator*.

For our purposes, it will suffice to consider the linear relation $E(Y) = \alpha + \beta X$, where α and β are the unknown parameters. Let

(x_i, Y_i); $i = 1, 2, \ldots, n$, be a random sample of the random variables with their associated values. Then the values of α and β that minimize the sum of squares

$$\sum [Y_i - (\alpha + \beta x_i)]^2$$

are the *least-squares estimators* of the parameters α and β. The method of least-squares has already been discussed in connection with regression analysis. (See article 10.4.2 and chapter 10 of Part I).

15.5 ESTIMATION BY CONFIDENCE INTERVAL

A confidence interval estimate of the unknown parameter θ is an interval computed from a random sample of n values $x_1, x_2, \ldots, x_n$ with a statement of how *confident* (*e.g.* 90 per cent, 95 per cent or 99 per cent) we are that the interval contains the unknown parameter θ. A confidence interval estimate is in the from $(L < \theta < U)$, where L, and U depend upon the value of the statistic $\hat{\theta}$ of a random sample selected from the population and the sampling distribution of the statistic. To make an assertion that θ lies in the interval (L, U), we may determine from the sampling distribution of $\hat{\theta}$ two values L and U such that $P(L<\theta<U)$ is equal to any specified probability, conventionally denoted by $1-\alpha$. If, of instance, L and U are two statistics such that for all θ

$$P(L < \theta < U) = 1 - \alpha, \text{ for } 0 < \alpha < 1,$$

then the probability of the interval (L, U) containing the population parameter θ is $1 - \alpha$. The interval (L, U) is called a $100\,(1 - \alpha)$ percent *confidence interval* for the unknown parameter θ, the probability $(1-\alpha)$ associated with interval estimate is called the *confidence co-efficient* or the *confidence level*, and α is the probability that the parameter θ will lie outside the interval (L, U). Thus an interval has a specified probability $(1-\alpha)$ of containing the true value of the parameter. For example, if $\alpha = 0.05$, then the probability that the interval (L, U) contains θ, is 0.95.

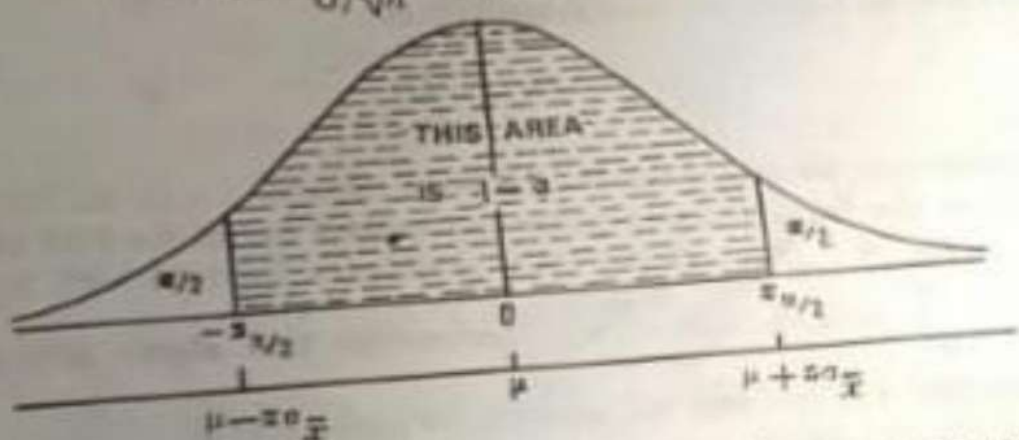
The endpoints that bound the confidence interval, are called the *lower* and *upper confidence limits* for θ. These limits are random variables as they can be different for different samples. The width of the confidence interval, *i.e.* the difference $U - L$, is called the *precision* of the estimate. The precision may be increased either by increasing the sample size or by decreasing the *confidence level*. The concept of a confidence interval was introduced in 1937 by the Polish-English-American statistician Jerzy Neyman (1894–1981).

Some of the most commonly used confidence intervals for population parameters are discussed in the sections that

15.5.1. Confidence Interval Estimate of a Population Mean. To compute a confidence interval for the population mean μ, we have to see whether or not the population is normal, whether or not the population standard deviation is known, and whether the sample size is large or small. We discuss these different cases below.

(i) **Normal Population with σ known.** Let a random sample $X_1, X_2, \ldots, X_n$ of size n be drawn from a normal population with an unknown mean μ and a known standard deviation σ. Then the sampling distribution of the mean $\bar{X}$ will be normal with a mean μ and a standard deviation $\sigma/\sqrt{n}$, and the variable $Z = \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ will be *exactly* standard normal, no matter how small the sample size will be. The normal distribution tells us that the probability that a value of Z will fall in the interval from $-z_{\alpha/2}$ to $z_{\alpha/2}$ is equal to $1-\alpha$, where $z_{\alpha/2}$ is equal to $\alpha/2$. That is we can make the following probability statement

$$P\left[-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right] = 1-\alpha.$$



To put μ inside inequalities within the brackets, we proceed as below:

(a) We multiply all terms inside the brackets by $\dfrac{\sigma}{\sqrt{n}}$ and get

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(b) We subtract $\bar{X}$ from each term and have

$$-\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(c) We multiply all terms by -1 (remember, we inverse the direction of the inequality sign when we multiply both sides of the inequality by a negative number) and obtain

$$\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

which is equivalent to

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

(d) We substitute this result in the probability statement and get

$$P\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

Hence, for a particular sample of size n, a $100(1-\alpha)$ per cent confidence interval for μ is given by

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

which may be expressed more compactly as

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If, for instance, we desire a 95% confidence interval, *i.e.* $(1-\alpha)\,100 = 95\%$, then from the table of areas under the normal curve, we find that the value $z_{0.025}$ is 1.96, and the 95% confidence interval will be from $\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}}$ to $\bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}$. This means that about 95% of the intervals found in this way will contain the parameter μ.

Example 15.18. The standard deviation of the amounts poured into bottles by an automatic filling machine is 1.8 ml (millileter). The amounts of fill in a random sample of bottles, in ml, were 451, 479, 482, 480, 477, 478, 481 and 482. Suppose the population of amounts of fill is normal. Construct a 90% confidence interval for the mean amount in all bottles filled by the machine.

The 90% confidence interval for the mean amount in all bottles, μ, given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Here $\bar{x} = \frac{\Sigma x}{n} = \frac{3840}{8} = 480$, $\sigma = 1.5$, $n = 8$ and $z_{0.05} = 1.645$.

Substituting these values, we get

$$480 \pm 1.645 \left(\frac{1.5}{\sqrt{8}}\right)$$

or $\quad 480 \pm (1.645)(0.636)$

or $\quad 480 \pm 1.05$ or 478.95 to 481.05

Hence the 90% confidence interval for μ calculated from the given sample is (478.95, 481.05).

Example 15.19. A confidence interval is constructed from a sample of size 25, for the mean of a normal population which has $\sigma = 50$. The limits for the interval are 110.2 and 135.8. What confidence co-efficient was used?

The $100(1-\alpha)\%$ confidence limits are given by

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ and } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Substituting the given values, we get

$$\bar{X} - z_{\alpha/2}\frac{50}{\sqrt{25}} = 110.2$$

$$\bar{X} + z_{\alpha/2}\frac{50}{\sqrt{25}} = 135.8$$

Subtracting, we get

$$2z_{\alpha/2}(10) = 135.8 - 110.2$$

$$z_{\alpha/2} = \frac{25.6}{20} = 1.28$$

We know that $z_{\alpha/2}$ denotes that the area to the right of $z_{\alpha/2}$ is $\alpha/2$. From area tables, we find that the area to the right of the value $z_{\alpha/2} = 0.5 - 0.4 = 0.1$, implying that $\alpha/2 = 0.1$ or $\alpha = 0.2$

Thus $1 - \alpha = 1 - 0.2 = 0.8$

Hence an 80% confidence co-efficient was used.

(ii) **Normal Population with σ unknown.** When a random sample $X_1, X_2, \ldots, X_n$ of size n is drawn from a normal population with σ unknown, we estimate σ by the sample standard deviation, which is then

used in place of σ. If the sample size is sufficiently large ($n \geq 30$), then the *Central Limit Theorem* allows us to assume that the sampling distribution of $\bar{X}$ is approximately normal with a mean of μ and a standard deviation of $\frac{S}{\sqrt{n}}$, where S is the sample standard deviation.

The probability expression for estimating μ then becomes

$$P\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha,2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Thus a 100 $(1-\alpha)$ per cent confidence interval for μ is given by

$$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}.$$

When σ is unknown and sample size is small ($n < 30$), the sampling distribution of $\bar{X}$ will not be normally distributed. The sampling distribution of $\bar{X}$ then follows a distribution, known as Student's-t distribution. We shall discuss this case in chapter 18.

Example 15.20. The Punjab Highway Department is studying traffic pattern on the G.T.Road near Lahore. As part of the study, the department needs to estimate the average number of vehicles that pass the Ravi bridge each day. A random sample of 64 days gives $\bar{x} = 5410$ and $S = 680$. Find the 90 per cent confidence interval estimate for μ, the average number of vehicles per day.

The 90% confidence interval for μ is

$$\bar{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}},$$

where $\bar{x} = 5{,}410$, $S = 680$, $n = 64$ and $z_{0.05} = 1.645$

Substituting these values, we get

$$5{,}410 \pm (1.645)\left(\frac{680}{\sqrt{64}}\right)$$

or $\quad 5{,}410 \pm (1.645)\,(85)$

or $\quad 5{,}410 \pm 139.8$ or 5270.2 to 5549.8

Thus the 90% confidence interval estimate for μ is (5270, 5550).

(iii) **Non-normal Population with known or unknown σ. (Large samples).** The *central limit theorem* tells us that for large sample sizes, the sampling distribution of the mean $\bar{X}$ is approximately a normal, even

though the population sampled is non-normal. That is, the random variable $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ is approximately standard normal and consequently

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Therefore an approximate 100 $(1 - \alpha)$ per cent confidence interval for μ, the mean of a non-normal population with σ known is given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

In case σ is unknown and is estimated by the sample standard deviation S, the confidence interval estimate for μ becomes

$$\bar{x} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}.$$

If we sample without replacement from a finite population of size N and sample size n is greater than 5% of population size, then the confidence interval estimate for μ is given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

Example 15.21 A sample of 100 observations from a population known to be non-normal yielded the sample values $\bar{x} = 182$ and $S^2 = 299$. Find an approximate 99% confidence interval for μ.

The sample size $(n = 100)$ is large enough to allow us to assume that the sampling distribution of $\bar{X}$ is approximately normal with mean $= \mu$ and standard deviation $= \dfrac{S}{\sqrt{n}}$. Therefore the approximate 100 $(1-\alpha)$ per cent confidence interval for μ is

$$\bar{x} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}.$$

Now $\bar{x} = 182$, $S = \sqrt{299} = 17.29$, $n = 100$ and $z_{0.005} = 2.58$ as the confidence co-efficient is 0.99.

Thus the 99% confidence interval for μ is

$$182 \pm (2.58) \left(\frac{17.29}{\sqrt{100}}\right)$$

or $$182 \pm (2.58)\,(1.729)$$

or 182 ± 4.46 or 177.54 to 186.46.

Hence an approximate 99% confidence interval for μ is (177.54, 186.46).

Example 15.22. A random sample of size $n = 200$, selected without replacement from a population of size $N = 1000$ with $\sigma = 1.08$, showed that $\bar{x} = 69.2$. Construct a 95 percent confidence interval for the true mean of the population.

As the sample is selected without replacement and its size $n = 200$ is greater than 5% of population size $N = 1{,}000$, we therefore use the following expression to calculate the desired 95% confidence interval for μ:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Substituting the values, we get

$$69.2 \pm (1.96) \frac{1.08}{\sqrt{200}} \sqrt{\frac{1000-200}{1000-1}}$$

or $$69.2 \pm (1.96) \left(\frac{1.08}{14.14}\right) (0.8947)$$

or $$69.2 \pm (1.96)(0.068)$$

or $$69.2 \pm 0.13 \text{ or } 69.07 \text{ to } 69.33$$

Hence the 95% confidence interval for μ is (69.07, 69.33).

Example 15.23. A sample of readings from a normal population with unknown mean μ and unknown variance σ^2 gave the following data:

x	17.4	17.5	17.6	17.7	17.8
f	12	16	19	23	10

A second sample of readings taken from the same population gave $n_2 = 72$, $\Sigma X = 1267.2$, $\Sigma X^2 = 22536$. Combine the two samples to give estimates of μ and σ^2, and give the approximate 90% confidence interval for μ.

(I.U., M.Sc. 1995)

First we calculate the sample means and the sample variances.

For sample 1: $\bar{x}_1 = \frac{\Sigma fx}{n_1} = \frac{1408.3}{80} = 17.60$

$$S_1^2 = \frac{\sum fx^2}{n_1} - \left(\frac{\sum fx}{n_1}\right)^2 = \frac{24792.63}{80} - \left(\frac{1408.3}{80}\right)^2$$

$$= 309.9079 - 309.8920 = 0.0159$$

For sample 2: $\bar{x}_2 = \frac{\sum X}{n_2} = \frac{1267.2}{72} = 17.60$

$$S_2^2 = \frac{\sum X^2}{n_2} - \left(\frac{\sum X}{n_2}\right)^2 = \frac{22536}{72} - \left(\frac{1267.2}{72}\right)^2$$

$$= 313 - 309.76 = 3.24$$

Next we calculate the *pooled* estimates $\bar{X}_c$ and s_p^2 as below:

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{80(17.60) + 72\,(17.60)}{80 + 72} = 17.60$$

$$s_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \frac{80\,(0.0159) + 72(3.24)}{80 + 72 - 2} = \frac{234.552}{150} = 1.5637$$

A 90% confidence interval for μ, based on the combined samples, is

$$\bar{x}_c \pm 1.645 \cdot \frac{s_p}{\sqrt{n}}$$

or $\quad 17.60 \pm 1.645 \frac{\sqrt{1.5637}}{\sqrt{152}}$

or $\quad 17.60 \pm 1.645\,(0.1014)$

or $\quad 17.60 \pm 0.17$ or 17.43 to 17.77

Hence, the 90% confidence interval for μ, on the basis of two samples, is (17.43, 17.77).

15.5.2. Interpretation of a Confidence Interval. A 100 $(1-\alpha)$ per cent confidence interval for μ is

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

It is to be emphasized that in this expression, μ is constant and it is the endpoints (*i.e.* limits) of the interval which are random variables. Therefore after computing the confidence interval for a particular sample, it is erroneous to say that

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

because in this expression, no random variable appears, whereas a probability statement is made about random variables. This means that the probability measure cannot be attached to the stated interval. If the statement is correct, in the sense that it includes μ, the probability is 1 and if it is incorrect, the probability is zero. In neither case, the probability is $(1-\alpha)$. However, of all possible intervals, $100\,(1-\alpha)$ per cent will include μ and α per cent of the intervals will not include μ in the long run.

To be specific, suppose an actual sample of size $n = 16$ is selected from a normal population with an unknown mean μ and a known standard deviation $\sigma = 2$. Let the sample mean $\bar{x} = 6.2$ and the confidence co-efficient $(1-\alpha) = 0.95$. Substituting these values, we get

$$P\left[\,6.2 - 1.96\left(\frac{2}{\sqrt{16}}\right) < \mu < 6.2 + 1.96\left(\frac{2}{\sqrt{16}}\right)\right] = 0.95$$

or $$P(5.22 < \mu < 7.18) = 0.95.$$

This probability statement is erroneous because μ is not a random variable. The parameter μ is either in the interval or it is not. If μ lies in the interval (5.22, 7.18), then $P(5.22 < \mu < 7.18) = 1$ and if it does not lie in the interval (5.22, 7.18), then $P(5.22 < \mu < 7.18) = 0$. We can say that

$$P\left[\,\bar{X} - 1.96\left(\frac{2}{\sqrt{16}}\right) < \mu < \bar{X} + 1.96\left(\frac{2}{\sqrt{16}}\right)\right] = 0.95$$

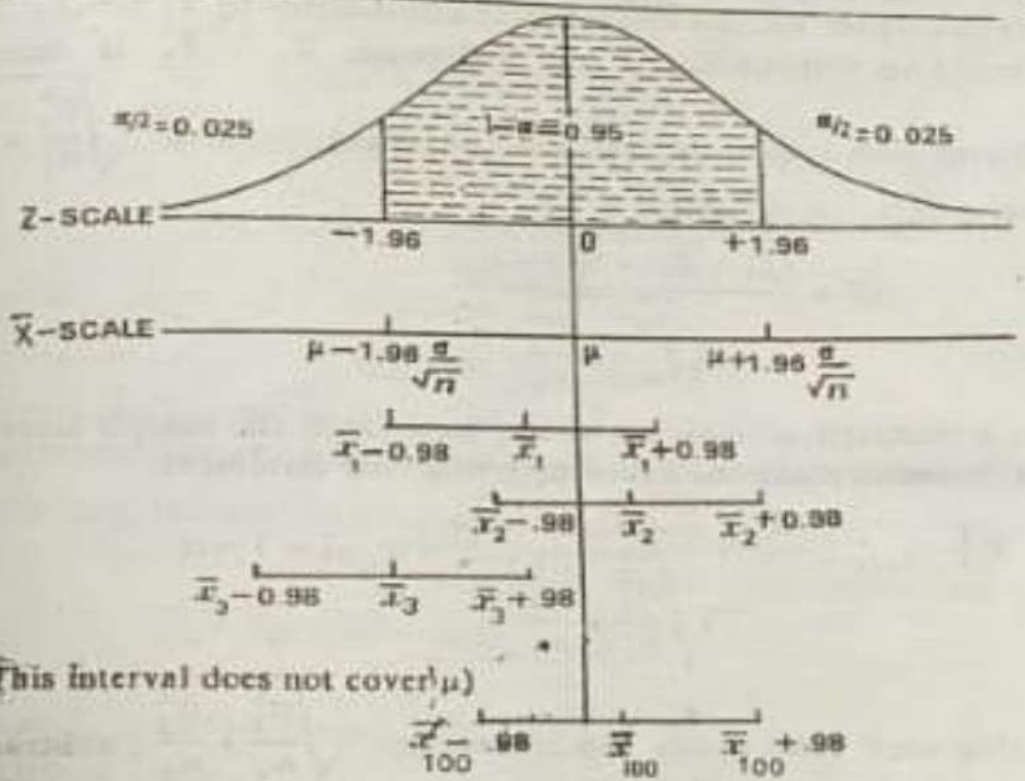
or $$P(\bar{X} - 0.98 < \mu < \bar{X} + 0.98) = 0.95.$$

The interval $\bar{X} \pm 0.98$ is a random variable because $\bar{X}$ does not have any particular numerical value but takes different values in different samples. It is therefore correct to say that

$$P(\bar{X} - 0.98 < \mu < \bar{X} + 0.98) = 0.95,$$

meaning thereby that "the probability that the random interval $(\bar{X}-0.98, \bar{X} + 0.98)$ *covers* the true value of μ is 0.95". In other words, in repeated samples of size 16 from a normal population with standard deviation 2, the interval $(\bar{X} - 0.98, \bar{X} + 0.98)$ *will contain* the true unknown value of μ about 95 per cent of time.

To illustrate, let us draw 100 samples of 16 observations each, calculate $\bar{X}$ for each sample and hence find the interval $(\bar{x}_i - 0.98, \bar{x}_i + 0.98)$ for each sample. These interval estimates based on 100 possible values of the random variable $\bar{X}$ are shown in the figure on next



On the average, about 95 of these 100 intervals will contain the true value of μ. Thus we see that having taken our sample and found $\bar{x} = 6.2$, we cannot say that

$$P[6.2 - 0.98 \leq \mu \leq 6.2 + 0.98] = 0.95.$$

Rather, we say that we are 95 per cent confident that the true population mean μ will be in the interval $(6.2 - 0.98, 6.2 + 0.98)$.

15.5.3 Confidence Interval for Difference of Means. To construct the confidence interval for the difference between two means, $\mu_1 - \mu_2$, the following three cases are to be considered:

(i) Both the populations are normal with known standard deviations.

(ii) Both the populations are normal with unknown standard deviations.

(iii) Both the populations are non-normal, in which case, both sample sizes are necessarily large.

(i) **Normal Populations with known Standard Deviations.** Suppose we have two normal populations. Population 1 has an unknown mean μ_1 and a known standard deviation σ_1; and Population 2 has an unknown mean μ_2 and a known standard deviation σ_2. Independent

means calculated. Let the sample means be denoted by $\bar{X}_1$ and $\bar{X}_2$. Then the sampling distribution of the difference, $\bar{X}_1 - \bar{X}_2$, is normally distributed with a mean $\mu_1 - \mu_2$, and a standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. In other words, the variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is *exactly* standard normal, no matter how small the sample sizes are. We can therefore make the following probability statement

$$P\left[-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}\right] = 1 - \alpha.$$

Multiplying each term inside the bracket by $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, subtracting $(\bar{X}_1 - \bar{X}_2)$ and then multiplying by -1 (inequality signs reversed), we get

$$P\left[(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < (\mu_1 - \mu_2) < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$

Hence the 100 $(1-\alpha)$ per cent confidence interval for particular samples obtained, for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Example 15.24. Two independent samples of 100 mechanists and 100 carpenters are taken to estimate the difference between the weekly wages of the two categories of workmen. The relevant data are given below:

	Sample mean wage	Population variance
Mechanists	345	196
Carpenters	340	204

Determine the 95 and the 99 per cent confidence limits for the true difference between the average wages for mechanists and carpenters.

(P.U., B.A./B.Sc. 1986)

The 95% confidence limits for $\mu_1 - \mu_2$ are given by

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Substituting the values, we get

$$(345 - 340) \pm 1.96 \sqrt{\frac{196}{100} + \frac{204}{100}}$$

or $\quad 5 \pm (1.96)(2)$

or $\quad 5 \pm 3.92$ or 1.08 to 8.92

Hence the 95% confidence limits for the true difference between the average weekly wages for mechanists and carpenters are 1.08 to 8.92.

Similarly, the 99% confidence limits for $\mu_1 - \mu_2$ are

$$(\bar{x}_1 - \bar{x}_2) \pm 2.58 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

or $\quad (345 - 340) \pm 2.58 \sqrt{\frac{196}{100} + \frac{204}{100}}$

or $\quad 5 \pm 2.58(2)$

or $\quad 5 \pm 5.16$ or -0.16 to 10.16.

Thus the 99% confidence limits for the true difference between the average weekly wages are -0.16 to 10.16.

(ii) **Normal Populations with unknown Standard Deviations.** When the independent samples of sizes n_1 and n_2 are drawn from normal populations with unknown standard deviations, we estimate them by the respective sample standard deviations. If sample sizes are sufficiently large, then we can assume that the sampling distribution of the difference, $\bar{X}_1 - \bar{X}_2$, is approximately normal with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$. Hence the 100 $(1-\alpha)$ per cent confidence interval estimate for $\mu_1 - \mu_2$, for particular samples obtained, is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

If, on the other hand, sample sizes are small and the populations have unknown equal standard deviations, then we use Student's

t-distribution to construct the confidence interval. The t-distribution shall be discussed in chapter 18.

Example 15.25. A test in statistics was given to 50 girls and 75 boys. The girls made an average grade of 76 with a standard deviation of 6, while the boys made an average grade of 82 with a standard deviation of 8. Find a 96% confidence interval for the difference $\mu_1 - \mu_2$, where μ_1 is the mean score of all boys and μ_2 is the mean score of all girls who might take this test. (P.U., B.A/B.Sc. 1984)

As the sample sizes are sufficiently large ($n_1, n_2 > 30$), we can therefore use the sample standard deviations S_1 and S_2 in place of population standard deviations σ_1 and σ_2. Assuming the populations to be normally distributed, the 96% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm 2.054 \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where $z_{\alpha/2}$, i.e. $z_{0.02} = 2.054$

Substituting the given values, we get

$$(82-76) \pm 2.054 \sqrt{\frac{(8)^2}{75} + \frac{(6)^2}{50}}$$

or $$6 \pm 2.054 \sqrt{\frac{64}{75} + \frac{36}{50}}$$

or $$6 \pm (2.054)(1.254)$$

or $$6 \pm (2.58) \text{ or } 3.42 \text{ to } 8.58.$$

Hence the desired 96% confidence interval for $\mu_1 - \mu_2$ calculated from the given values, is (3.42, 8.58).

(iii) **Non-normal Populations.** If the sample sizes are sufficiently large, then the *central limit theorem* tells us *that the* sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ will be approximately normal even though the populations are non-normal. An approximate $100(1-\alpha)$ per cent confidence interval for $\mu_1 - \mu_2$ when the population standard deviations are known, would be

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two particular samples.

If the population standard deviations are unknown, then they are estimated by the sample standard deviations. The approximate $100(1-\alpha)$ per cent confidence interval for $\mu_1 - \mu_2$ is then given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

where S_1 and S_2 are the sample standard deviations.

15.5.4. Confidence Interval for Population Proportion. (Large sample). Let a random sample of size $n (n > 30)$ be drawn from a binomial population with an unknown proportion of successes p and let the sample proportion be $\hat{P} = \frac{X}{n}$. Now, we wish to estimate p by an interval to be computed from the sample data.

We know that the sampling distribution of a sample proportion $\hat{P}$ is approximately normal with a mean of p and standard deviation of $\sqrt{\frac{pq}{n}}$ if the sample size is sufficiently large and p is not too close to zero or 1. Thus the random variable $Z = \frac{\hat{P} - p}{\sqrt{\frac{pq}{n}}}$ will be approximately $N(0, 1)$ and hence we can state that

$$P\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Multiplying all terms of the inequality by $\sqrt{pq/n}$, subtracting $\hat{P}$ from each term and then multiplying by -1 (reversing the direction of inequality signs), we get

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{pq}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

Thus, for a particular sample of size n $(n > 30)$, an approximate $100(1-\alpha)$ per cent confidence interval for p is given by

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{pq}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right).$$

But we are faced with the difficulty that the standard error of the sample proportion involves the unknown p. For large samples, this difficulty is overcome by using the sample proportion $\hat{p}$ in place of p. Hence for a particular sample, the approximate $100(1-\alpha)$ per cent

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Example 15.26 In a certain large city, a random sample of 400 families contacted by a local TV station showed that 275 owned colour TV sets. Find an approximate 90 per cent confidence interval on the true proportion of all families living in the city who own colour TV sets.

An approximate 90% confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Now, sample proportion $\hat{p} = \frac{275}{400} = 0.69$, so $\hat{q} = 1 - 0.69 = 0.31$.

The degree of confidence is 90%, therefore $z_{0.05} = 1.645$

Substituting these values, we get

$$0.69 \pm (1.645)\sqrt{\frac{(0.69)(0.31)}{400}}$$

or $\quad 0.69 \pm (1.645)(0.023)$

or $\quad 0.69 \pm 0.038$ or 0.652 to 0.728.

Hence the approximate 90% confidence interval for the true proportion that owns colour TV sets is (0.652, 0.728).

15.5.5. Confidence Interval for the Difference between Proportions. (Large samples). Suppose there are two binomial populations with unknown proportions of successes p_1 and p_2 respectively. Let $\hat{P}_1$ be the proportion of successes based on a random sample of size n_1 drawn from first population, and $\hat{P}_2$ be the proportion of successes based on a random sample of size n_2 drawn from second population. Then the sampling distribution of the difference, $\hat{P}_1 - \hat{P}_2$ will be approximately normal with a mean of $p_1 - p_2$ and a standard deviation of $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$, when the sample sizes are sufficiently large. That is, for sufficiently large samples, the random variable

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \text{ is approximately } N(0, 1),$$

and we can make the following probability statement

$$P\left(-z_{\alpha/2} < \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha.$$

We see that the standard error of $\hat{P}_1 - \hat{P}_2$ involves the unknown parameters p_1 and p_2, we therefore replace p_1 and p_2 with their sample estimates $\hat{p}_1$ and $\hat{p}_2$. Hence for sufficiently large samples, an approximate $100(1-\alpha)$ per cent confidence interval for $p_1 - p_2$ is given by

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.$$

Example 15.27. In a poll of college students in a large state university, 300 of 400 students living in dormitories approved a certain course of action, whereas 200 of 300 students not living in dormitories approved it. Estimate the difference in the proportions favouring the course of action and compute 90% confidence interval for it.

(P.U., B.A./B.Sc. 1982)

Let $\hat{p}_1$ and $\hat{p}_2$ be the observed proportions in the first and second sample respectively. Then

$$\hat{p}_1 = \frac{300}{400} = 0.75, \text{ and } \hat{p}_2 = \frac{200}{300} = 0.67.$$

$\therefore$ Difference in proportions $= \hat{p}_1 - \hat{p}_2 = 0.75 - 0.67 = 0.08$

The degree of confidence is 0.90, therefore $z_{0.05} = 1.645$.

The 90% confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm (1.645)\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

or $\quad 0.08 \pm (1.645)\sqrt{\frac{(0.75)(0.25)}{400} + \frac{(0.67)(0.33)}{300}}$

or $\quad 0.08 \pm (1.645)\sqrt{0.000469 + 0.000737}$

or $\quad 0.08 \pm (1.645)(0.0347)$

or $\quad 0.08 \pm 0.057$ or 0.023 to 0.137

Hence the 90 per cent confidence interval for $p_1 - p_2$ is $(0.023, 0.137)$.

15.5.6 One-sided Confidence Interval: Since the interval $L \le \theta \le U$ for the parameter θ specifies both a lower and an upper limit, it may therefore be more appropriate to call it a *two-sided* $100(1-\alpha)$ per cent confidence interval. Occasionally, we may wish to find only an upper or a lower limit but not both, for the parameter θ, that is we want a one-

sided interval. In such a case, the entire α area will be located at one end of the sampling distribution. Thus a *one-sided* $100(1-\alpha)$ per cent lower confidence interval for θ is given by the interval $L \le \theta$ where we choose the lower confidence limit L such that $P(L \le \theta) = 1-\alpha$.

Similarly, a *one-sided* $100(1-\alpha)$ percent upper confidence interval for θ is given by the interval $\theta \le U$ where we choose the upper confidence limit U such that $P(\theta \le U) = 1-\alpha$.

In case of a normal distribution with σ known the probability statement for the *one-sided* $100(1-\alpha)$ per cent confidence interval for μ would be

$$P\left(\mu > \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \text{ or } P\left(\mu < \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Thus in *one-sided confidence interval*, a $100\,(1-\alpha)$ per cent lower confidence limit would be $\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}$, and a $100(1-\alpha)$ per cent upper confidence limit would be $\bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$.

One-sided confidence intervals for the difference between two means $\mu_1-\mu_2$ may also be obtained. Thus for example, when populations are normal with known standard deviations, a $100(1-\alpha)$ per cent upper confidence interval $\mu_1-\mu_2$ is

$$\mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2) + z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

and a $100(1-\alpha)$ per cent lower confidence interval is

$$(\bar{X}_1-\bar{X}_2) - z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2$$

Example 15.28 For estimating the average weight of college men in the city of Lahore, a sample of 100 college men is randomly selected and a sample mean of 120 pounds is obtained. Assume that the variance of the population is 1600. Determine the lower limit of the 95 per cent confidence interval with 280 pounds as the upper limit.

(P.U., B.A/B.Sc. 1980)

This is an example of one-sided confidence interval with upper limit as a constant (280 pounds). Therefore the lower limit of the 95% confidence interval estimate for μ is

$$\bar{x} - (1.645)\frac{\sigma}{\sqrt{n}} = 120 - (1.645)\frac{40}{\sqrt{100}}$$

$$= 120 - (1.645)(4)$$

$$= 120 - 6.58 = 113.42 \text{ pounds.}$$

17.7.7. **Sample size for Estimating Population Mean.** The $100(1-\alpha)$ per cent confidence interval μ is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

which may be written as

$$|\bar{x} - \mu| = z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where $\frac{\sigma}{\sqrt{n}}$ is the standard error of $\bar{X}$ when sampling is performed with replacement or population is very large (infinite). The quantity $|\bar{X} - \mu|$ is also called the *error of the estimator* $\bar{X}$ and is denoted by e. Thus a $100(1-\alpha)$ per cent error bound for estimating μ is given by $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. In other words, in order to have a $100(1-\alpha)$ per cent confidence that the error in estimating μ with $\bar{x}$ to be less than e, we need n such that

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or $$\sqrt{n} = z_{\alpha/2}\frac{\sigma}{e}$$

or $$n = \left(\frac{z_{\alpha/2}\,\sigma}{e}\right)^2$$

Hence the desired sample size for being $100(1-\alpha)\%$ confident that the error in estimating μ will be less than e, when sampling is with replacement or population is very large, is given by

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{e}\right)^2.$$

It is important to note that the population standard deviation σ is generally not known, its estimate is found either from past experience or from a pilot sample of size $n > 30$. In case of fractional result, it is always to be rounded to the next higher integer for the sample size.

The standard error of $\bar{X}$, when sampling is performed without replacement from a small population of size N, is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

In this case, the $100(1-\alpha)$ per cent error bound for estimating μ becomes

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

Solving for n, we get

$$n = \frac{N (z_{\alpha/2}\, \sigma)^2}{(N-1)\, e^2 + (z_{\alpha/2}\, \sigma)^2}$$

as the desired sample size when sampling is without replacement from a small population of size N.

Example 15.29. A research worker wishes to estimate the mean of a population using a sample sufficiently large that the probability will be 0.95 that the sample mean will not differ from the true mean by more than 25 per cent of the standard deviation. How large a sample should be taken?

If the sample mean will not differ from the true mean by more than 25% of σ with a probability of 0.95, then

$$e = |\bar{x} - \mu| = \frac{25\sigma}{100} = \frac{\sigma}{4}, \text{ and } z_{\alpha/2} = 1.96.$$

Substituting these values in the formula

$$n = \left(\frac{z_{\alpha/2}\, \sigma}{e}\right)^2, \text{ we get}$$

$$n = \left(\frac{1.96 \times \sigma}{\sigma/4}\right)^2 = 61.4656$$

Hence the required sample size should be 62, the next higher integer as the sample size cannot be fractional.

15.5.8. Sample size for Estimating Population Proportion. The large sample confidence interval for p is given by

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

This implies that $e = z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$

Therefore solving for n, we obtain

$$n = \frac{(z_{\alpha/2})^2\,\hat{p}\,\hat{q}}{e^2}$$

Since the values of $\hat{p}$ and $\hat{q}$ are not known as the sample has not yet been selected, we therefore use an estimate $\hat{p}$ obtained from pilot sample information.

Example 15.30 In a random sample of 75 axle shafts, 12 have a surface finish that is rougher than the specification will allow. How large a sample is required if we want to be 95% confident that the error in using $\hat{p}$ to estimate p is less than 0.05?

Here $e = |p - \hat{p}| = 0.05$, $\hat{p} = \frac{12}{75} = 0.16$,

$\hat{q} = 1 - \hat{p} = 0.84$ and $z_{0.025} = 1.96$ $(\alpha/2 = 0.025)$

Substituting these values in the formula

$$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 \hat{p}\,\hat{q}, \quad \text{we obtain}$$

$$n = \left(\frac{1.96}{0.05}\right)^2 (0.16)\,(0.84) = 207$$

as the required sample size.

EXERCISES

15.1 (a) Explain what is meant by statistical inference?

(b) Distinguish between

(i) Estimation and Testing of Hypothesis.

(ii) Estimates and Estimators.

(iii) Point estimate and Interval estimate.

15.2 What do you mean by Point Estimation? Explain the desirable properties of a good point estimator. (P.U., B.A/B.Sc. 1980, 91, 93, 96)

15.3 (a) Explain with examples the following properties of a point estimator:

(i) Unbiasedness, (ii) Consistency, (iii) Efficiency

(b) Explain what is meant by the Mean Square Error of an estimator. Prove that $MSE(T) = \text{Var}(T) + (\text{Bias})^2$.

(P.U., B.A./B.Sc. 1992)

(c) Is an estimator a random variable? Why or why not?

15.4 (a) A sample of size 10 yields values 8, 4, 10, 5, 5, 4, 9, 4, 3, 7. Estimate the mean, the variance and the third mean moment of the random variable X which models these values.

(b) Find a point estimate of μ and the estimated standard error in each of the following cases:

(i) $n = 70$, $\sum x_i = 852$, $\sum (x_i - \bar{x})^2 = 215$.

(ii) $n = 160$, $\sum x_i = 1985$, $\sum (x_i - \bar{x})^2 = 475$.

15.5 (a) Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with a mean of μ and a variance of σ^2. Then prove that

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator of σ^2. (P.U., B.A./B.Sc. 1993)

(b) If random samples of size $n = 2$ are drawn with replacement from a population consisting of the six members 10, 12, 14, 16, 18 and 20, show by finding all possible samples, that $\bar{x}$ and s^2 are unbiased estimators of μ and σ^2.

15.6 (a) Explain what you understand by

(i) an unbiased estimator (ii) a consistent estimator, (iii) the relative efficiency of two estimators, and (iv) a sufficient estimator.

(b) If X_1, X_2 and X_3 are a random sample from a normal population with the mean μ and the variance σ^2, what is the relative efficiency of the estimtor $T_1 = \dfrac{X_1 + 2X_2 + X_3}{4}$ with respect to $T_2 = \bar{X}$? (P.U., B.A./B.Sc. 19[illegible])

15.7 (c) Explain why the sample mean $\bar{X}$, as well as the sample proportion P, is

(i) an unbiased estimator, (ii) a consistent estimator, (iii) an efficient estimator, (iv) a sufficient estimator.

(b) Taking all possible samples of size 2 with replacement from finite population 3, 5 and 2, show that

$E(S^2) \neq \sigma^2$, where $S^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$. (P.U., B.A/B.Sc. 1987)

15.8 (a) Show that (i) $\bar{X}$ is a consistent estimator of μ, (ii) $\bar{X}_1 - \bar{X}_2$ is a consistent estimator of $\mu_1 - \mu_2$, (iii) $\hat{P}\left(= \frac{X}{n}\right)$ is a consistent estimator of P, if the random sample is obtained under sampling with replacement.

(b) If $X_1, X_2, \ldots, X_n$ is a random sample taken from a population with mean μ and variance σ^2, show that $\frac{X_1 + 2X_2 + \ldots + nX_n}{n(n+1)/2}$ is an unbiased and consistent estimator for μ.

15.9 (a) Based on a random sample of 3 observations, consider 3 possible estimators of μ:

$$\bar{X}_1 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3;\ \bar{X}_2 = \frac{5}{8}X_1 + \frac{1}{8}X_2 + \frac{1}{4}X_3;$$

$$\bar{X}_3 = 0.2X_1 + 0.3X_2 + 0.4X_3$$

Find (i) which are unbiased, (ii) the efficiency of each unbiased estimator relative to other. (P.U., B.A/B.Sc. 1989)

(b) A random sample $X_1, X_2, \ldots, X_8$ is to be taken from a population with mean μ and variance σ^2. The following statistics are to be taken for μ:

$$T_1 = \frac{X_1 + X_2 + X_3}{3};\ T_2 = \frac{2X_2 - X_1 + 4X_6 + X_8}{6};$$

$$T_3 = \frac{X_8 - X_1}{8};\ T_4 = X_5$$

(i) Which of the above are unbiased?

(ii) Which of the above is most efficient?

15.10 (a) What is meant by 'sufficiency? Explain the Fisher-Neyman factorization criterion for sufficiency.

(b) Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution $f(x; p) = p^x q$, $x = 0, 1, 2, \ldots$ Show that $\sum X_i$ is a sufficient statistic.

(c) Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. Show that $\bar{X}$ is a sufficient estimator for μ.

*15.11 A random sample of n_1 observations, taken from a population with unknown mean μ and unknown variance σ^2, has mean $\bar{X}_1$ and variance S_1^2. A second sample of n_2 observations, has mean $\bar{X}_2$ and variance S_2^2. Show that an unbiased estimator of the population mean μ is given by $\dfrac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$ and an unbiased estimator of the population variance σ^2 is given by $(n_1S_1^2 + n_2S_2^2)/(n_1 + n_2 - 2)$.

15.12 Two samples of sizes n_1 and n_2 respectively are drawn from an infinite population having mean μ and variance σ^2. The two sample means are denoted by $\bar{X}_1$ and $\bar{X}_2$ respectively. Two estimators are defined as follows:

$T_1 = (n_1\bar{X}_1 + n_2\bar{X}_2) / (n_1 + n_2)$ and $T_2 = (\bar{X}_1 + \bar{X}_2)/2$

(a) Show that, as $n_1, n_2 \to \infty$, $\text{Var}(T_1) \to 0$ and hence T_1 can be thought of as a consistent pooled estimator for μ.

(b) Show that (i) T_2 is also an unbiased estimator for μ, (ii) $\text{Var}(T_2) = \dfrac{\sigma^2}{4}\left[\dfrac{1}{n_1} + \dfrac{1}{n_2}\right]$ and hence that T_2 is also a consistent estimator for μ.

(c) Show that $\dfrac{\text{Var}(T_1)}{\text{Var}(T_2)} = \dfrac{4n_1n_2}{(n_1 + n_2)^2}$ and thus, in general, $\text{Var}(T_1) < \text{Var}(T_2)$.

15.13 (a) What is the basic criterion of estimation by the method of maximum likelihood?

(b) Define the term *Likelihood Function* and describe the mathematical technique of finding maximum likelihood estimators. (P.U., B.A/B.Sc. 1993)

15.14 (a) What do you understand by a maximum likelihood estimator?

(b) If $X_1, X_2, \ldots, X_n$ be a random sample of size n taken from the Bernoulli distribution

$$f(x) = p^x q^{1-x}, \quad x = 0, 1; \; 0 \le p \le 1,$$

find the maximum likelihood estimator of p and show that it is a sufficient estimator. (P.U., B.A/B.Sc. 1986-S)

15.15 Consider the binomial distribution

$$f(x; p) = \binom{n}{x} p^x q^{n-x} \text{ for } x = 0, 1, 2, \ldots, n.$$

Find the maximum likelihood estimator of p when

(i) a single observation is taken;

(ii) a sample of m observations $X_1, X_2, \ldots, X_m$ is taken.

(P.U., B.A/B.Sc. 1992)

15.16 (a) Find the ML estimates of μ and σ^2 from a sample of size n independent observations from a normal distribution.

(b) Find the maximum likelihood estimates for $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ if a random sample of size 15 from $N(\mu, \sigma^2)$ yielded the following values:

31.5	36.9	33.8	30.1	33.9
35.2	29.6	34.4	30.5	34.2
31.6	36.7	35.8	34.5	32.7

15.17 (a) Let $X_1, X_2, \ldots, X_n$ be a random sample taken from the negative exponential distribution

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x > 0$$

Find the maximum likelihood estimate for λ.

(b) Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population $N(0, \sigma^2)$. Find the maximum likelihood estimator (MLE) of σ^2. (P.U., B.A/B.Sc. 1988)

15.18 If $x_1, x_2, \ldots, x_n$ are the values of a random sample of size n from a Poisson population with parameter λ, find an estimate of λ using

(i) the method of moments;

(ii) the method of maximum likelihood.

15.19 (a) Explain what is meant by (i) Confidence Interval, (ii) Confidence Limits, and (iii) Confidence Co-efficient.

(b) Describe the procedure followed in establishing an interval estimate for a population mean.

(c) We know that the statement $P(\bar{X}-1.96\sigma_{\bar{x}} < \mu < \bar{X} + 1.96\sigma_{\bar{x}}) = 0.95$ is correct, while the statement $P(140 < \mu < 160) = 0.95$ is not. Explain why the latter is erroneous.

15.20 (a) Explain the meaning of the following terms:

(i) Random Interval, (ii) Confidence Co-efficient, (iii) Interval Estimation.

(b) Let the observed value of the mean $\bar{X}$ of a random sample of size $n=20$ from a normal distribution with mean μ and variance $=80$ be 81.2. Find a 95% confidence interval for μ.

(c) Explain what is meant by the statement. "We are 95% confident that our interval estimate contains μ".

15.21 (a) How will you determine *confidence interval* for mean of a normal distribution?

(b) Find a 95 per cent confidence interval for μ, the true mean of a normal population which has $\sigma = 10$. Consider a sample of size $n=25$ with a mean of 67.53.

15.22 A school wishes to estimate the average weight of students in the sixth grade. A random sample of size $n=25$ is selected and the sample mean is found to be $\bar{x}=100$ lbs. The standard deviation of the population is known to be 15 lbs. Compute the 90% confidence interval for the population mean. Assume the population of weights to be normal. (P.U., B.A./B.Sc. 1991)

15.23 (a) What do you mean by *Confidence Interval*? Find 98% confidence interval for mean of a normal distribution when sample is large.

(b) Find a 90% confidence interval for the mean of a normal distribution with $\sigma=3$, given the sample (2.3, −0.2, −0.4, −0.9). (P.U., B.A/B.Sc. 1976, 87, 92)

15.24 (a) A confidence interval is constructed, from a random sample of size $n=50$, for the mean yield of a normal population which has $\sigma=21$ tons. The limits for the interval are 866.11 and 875.89 tons. What confidence co-efficient was used?

(b) The 95% confidence interval for the mean length of life of a particular brand of light bulbs is (1023.3 h, 1101.7 h). This interval is based on results from a random sample of 36 light bulbs. Find the 90% confidence interval for the mean length of life of this brand of light bulbs, assuming that the length of life is normally distributed.

15.25 (a) Explain the concept of Interval Estimation.

(b) A random sample of size $n=50$ from a normal population yielded the sample values $\bar{x} = 190$ and $S^2 = 800$. Find a 95% confidence interval for μ.

(c) A random sample of 100 values was taken from a normal population with mean μ and the following results were obtained: $\sum X = 3978.7$ and $\sum X^2 = 1583098.3$. Find 98% confidence interval for μ. (P.U., B.A/B.Sc. 1996)

15.26 (a) A restaurant wishes to estimate the average amount of money a customer spends for lunch. A random sample of size $n=36$ is selected and the sample mean is found to be $\bar{x}$ = Rs. 35.00. Assuming σ = Rs. 13.77, find 95% confidence limits for μ.

(b) A random sample of $n=100$ widths for two-by-fours was selected without replacement from a shipment of $N=500$ boards. The results show that $\bar{x} = 3.5$ inches and $S=0.1$ inches. Construct a 90% confidence interval estimate of the boards in the entire shipment.

15.27 In July 1969, the first man walked on the moon. Armstrong, Aldren and Collins brought back 64 rock samples. The rocks had an average Earth weight of 172 ounces. The sample variance S^2, was 299 (ounces)^2. The moon rock population is known, however, to follow a distribution which is not normal. Find a 99% confidence interval estimate for the mean weight of rocks on the lunar surface.

15.28 A sample of 64 readings from a normal population with mean μ and variance σ^2 gave $\sum X = 5452.8$ and $\sum(X-\bar{X})^2 = 973.44$. A second sample of readings from the same population gave

x	82	83	84	85	86	87
f	6	9	19	27	22	17

Combine the two samples to give estimates of μ and σ^2 and give the approximate 97% confidence interval for μ.

5.29 (a) Discuss the problem of finding a confidence interval for the difference $(\mu_1-\mu_2)$ between the two means of two independent normal distributions if the variances σ_1^2 and σ_2^2 are known but not necessarily equal.

(b) Let two independent random samples, each of size 100, from two independent normal distributions $N(x;\mu_1;\sigma^2)$ and $N(y;\mu_2;\sigma^2)$ yield $\bar{x} = 4.8$, $S_1^2 = 8.64$, $\bar{y} = 5.6$, $S_2^2 = 7.88$. Find 95% confidence interval for $(\mu_1 - \mu_2)$.

(P.U., B.A/B.Sc. Hons. Part III, 19[illegible])

15.30 (a) If $\bar{x}_1, \bar{x}_2$ are the sample means of two independent samples of sizes n_1, n_2 from normal populations with known variances σ_1^2, σ_2^2 respectively, show that the 95% confidence limits of the difference between population means are

$$(\bar{x}_1 - \bar{x}_2) \pm 2.326\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \quad \text{(P.U., B.A/B.Sc. 1962)}$$

(b) How will these limits be affected, if

(i) the populations are not known to be normal,

(ii) the variances are not known,

(iii) the confidence co-efficient is 95%.

15.31 (a) Explain the difference between a point estimate and an interval estimate. Why is an interval estimate more useful?

(b) A manufacturing company consists of two departments producing identical products. It is suspected that the hourly outputs in the two departments are different. Two random samples of production hours are respectively selected, and the following data are obtained:

	Department 1	Department 2
Sample size	64	49
Sample mean	100	90

The variances of the hourly outputs for the two departments are known to be $\sigma_1^2 = 256$ and $\sigma_2^2 = 196$ respectively. What is the point estimate for the true difference between the mean outputs of the two departments? Find the 95 per cent confidence limits for the true difference. (P.U., B.A/B.Sc. 1986-S)

15.32 In order to ascertain the age distribution of operatives in a certain industry, random samples of 1720 males and 1230 females are drawn. The sample means and standard deviations were 33.93 years and 14.20 years for the males, and 27.44 years

and 10.79 years for the females. Calculate the 95% confidence interval estimates for

(a) the mean age of all male operatives,

(b) the mean age of all female operatives, and

(c) the differences between their mean ages. (B.Z.U., B.A/B.Sc. 1985)

15.33 The means and variances of the weekly incomes in rupees of the workers employed in the different factories, from the samples are given below:

Factory	Sample Size	Mean	Variance
A	160	12.80	64
B	220	11.25	47

(a) What is the maximum likelihood estimate of the difference in mean incomes?

(b) Compute the 90 per cent confidence interval estimate for the real differences in the incomes of the workers from the two factories.

15.34 A random sample of size $n_1 = 100$ yielded the sample value $\bar{x}_1 = 509$, $S_1^2 = 950$. A random sample of size $n_2 = 100$ from another population yielded $\bar{x}_2 = 447$ and $S_2^2 = 876$. Find a 95 per cent confidence interval for $\mu_1 - \mu_2$.

15.35 (a) If $\hat{p}$ is the observed proportion in n independent Bernoullian trials, show that the 95% confidence limits for a population proportion p are, for large samples, given by

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}\,\hat{q}}{n}}.$$

(b) A random sample of 400 members of the labour force in a certain region showed that 32 were unemployed. Construct the 95% confidence interval for the proportion unemployed in the region.

(c) A random sample of 75 college students is selected and 16 are found to have cars on campus. Use a 95% confidence interval to estimate the fraction of students who have cars on campus. (P.U., B.A./B.Sc. 1990)

15.36 (a) Given $X = 60$, construct a 95% confidence interval for the parameter p of a binomial distribution for which $n = 100$.

(b) In a random sample of 1000 homes in a certain city, it is found that 638 are heated by natural gas. Find 98% confidence interval for the fraction of homes in this city that are heated by natural gas.

(c) A random sample of size $n=144$ gave $\hat{p} = 0.76$. Construct a 90% confidence interval for p. Interpret the 90% confidence interval.

15.37 (a) Determine a method for constructing a confidence interval for $p_1 - p_2$, the difference of two population proportions.

(b) Find a 95% confidence interval for $p_1 - p_2$ if a sample of size $n_1 = 100$ yielded $\hat{p}_1 = 0.54$ and a sample of size $n_2 = 100$ yielded $\hat{p}_2 = 0.49$.

15.38 A poll is taken among residents of a city and its suburbs to determine the feasibility of a proposal to construct a civic centre. If 2400 of 5000 city residents favour the proposal and 1200 of 2000 suburban residents favour it, find a 90% confidence interval for the true difference between the proportion of city and suburban residents who favour the proposal to construct the civic centre.

(P.U., B.A/B.Sc. 1988)

15.39 (a) A random sample of 65 bolts from a shipment had a mean diameter of 0.51 inches and a standard deviation of 0.16 inches. Construct a 99% one-sided confidence interval for the maximum mean diameter of the population.

(b) The life in hours of a 75-watt light bulb is known to be approximately normally distributed with standard deviation $\sigma = 25$ hours. A random sample of 20 bulbs has a mean life of $\bar{x} = 1014$ hours. Construct a 95% lower confidence interval on the mean life.

15.40 (a) If you wish to estimate the mean of a normal population whose variance is 10, how large a sample should you take so that the probability is 0.90 that your estimate will not be in error by more than 0.4 units?

(b) If the standard deviation of the lifetimes of television tubes is estimated as 100 hours, how large a sample must we take in order to be 95% confident that the error in the estimated mean lifetime will not exceed 20 hours?

(c) Find the approximate sample size necessary to estimate a binomial proportion p correct to within a bound of 0.06 with probability equal to 0.95. Assume p is near 0.1.

❖❖❖❖❖❖❖❖❖❖❖

16

Statistical Inference: Hypothesis Testing

16.1 INTRODUCTION

Hypothesis testing is a very important phase of statistical inference. It is a procedure which enables us to decide on the basis of information obtained from sample data whether to accept or reject a statement or an assumption about the value of a population parameter. Such a statement or assumption which may or may not be true, is called a *statistical hypothesis*. We *accept* the hypothesis as being true, when it is supported by the sample data. We *reject* the hypothesis when the sample data fail to support it.

It is important to understand what we mean by the terms *reject* and *accept* in hypothesis testing. The *rejection* of a hypothesis is to declare it false. The *acceptance* of a hypothesis is to conclude that there is not sufficient evidence to reject it. Acceptance does not necessarily mean that the hypothesis is true.

The basic concepts associated with hypothesis testing are discussed below:

16.1.1. Null and Alternative Hypothesis. A *null hypothesis*, generally denoted by the symbol H_0, is any hypothesis which is to be tested for possible rejection under the assumption that it is true. Today the term is used for any hypothesis that is being tested. The word *null* in the term *null hypothesis* implies that usually H_0 is the hypothesis of no effect. A null hypothesis should always be precise such as "the given coin is unbiased" or "a drug is ineffective in curing a particular disease" or "the difference between the two teaching methods is null or zero." The hypothesis is usually assigned a numerical value. For example, suppose we think that the average height of students in all colleges is 62". This

statement is taken as a hypothesis and is written symbolically as $H_0 : \mu = 62''$. In other words, we hypothesize that $\mu = 62''$.

An *alternative hypothesis* is any other hypothesis which we accept when the null hypothesis H_0 is rejected. It is customarily denoted by H_1 or H_A. A null hypothesis H_0 is thus tested against an alternative hypothesis H_1. For example, if our null hypothesis is $H_0 : \mu = 62''$, then our alternative hypothesis may be $H_1 : \mu \neq 62''$ or $H_1 : \mu > 62''$ or $H_1 : \mu < 62''$.

16.1.2. Simple and Composite Hypotheses. *A simple hypothesis* is one in which *all* parameters of the distribution are specified. For example, if the heights of college students are normally distributed with $\sigma^2 = 4$, the hypothesis that its mean μ is, say, 62'', that is $H : \mu = 62$, we have stated a simple hypothesis, as the mean and variance together specify a normal distribution completely. A simple hypothesis, in general, states that $\theta = \theta_0$ where θ_0 is the specified value of a parameter θ, (θ may represent μ, p, $\mu_1 - \mu_2$, etc.).

A hypothesis which is not simple (i.e. in which not all of the parameters are specified) is called a *composite hypothesis*. For instance, if we hypothesize that $H : \mu > 62$ (and $\sigma^2 = 4$) or $H : \mu = 62$ and $\sigma^2 > 4$, the hypothesis becomes a composite hypothesis because we cannot know the exact distribution of the population in either case. Obviously, the parameters $\mu > 62''$ and $\sigma^2 < 4$ have more than one value and no specified values are being assigned. The general form of a composite hypothesis is $\theta \leq \theta_0$ or $\theta \geq \theta_0$, that is the parameter θ does not exceed or does not fall short of a specified value θ_0. The concept of simple and composite hypotheses applies to both null hypothesis and alternative hypothesis.

Hypotheses may also be classified as *exact* and *inexact*. A hypothesis is said to be an *exact hypothesis* if it selects a unique value for the parameter such as $H : \mu = 62$ or $p = 0.5$. A hypothesis is called an *inexact hypothesis* when it indicates more than one possible values for the parameter such as $H : \mu \neq 62$ or $H : p > 0.5$. A simple hypothesis must be an exact one while an exact hypothesis is not necessarily a simple hypothesis. An inexact hypothesis is a composite hypothesis.

16.1.3. Test-statistic. A sample statistic which provides a basis for testing a null hypothesis, is called a *test-statistic*. Every test-statistic has a probability (sampling) distribution which gives the probability of obtaining a specified value of the test-statistic when the null hypothesis is true. It is important to remember that a test-statistic does not *prove* the hypothesis to be correct but it furnishes an *evidence* against the

hypothesis. The sampling distributions of the most commonly used test-statistics are normal, t, chi-square or F.

16.1.4. Acceptance and Rejection Regions. All possible values which a test-statistic may assume can be divided into two mutually exclusive groups: one group consisting of values which appear to be consistent with the null hypothesis, and the other having values which are unlikely to occur if H_0 is true. The first group is called the *acceptance region* and the second set of values is known as the *rejection region* for a test. The rejection region is also called the *critical region*. The value(s) that separates the critical region from the acceptance region, is called the *critical value*(s). The critical value which can be in the same units as the parameter or in the standardized units, is to be decided by the experimenter keeping in view the degree of confidence he (she) is willing to have in the null hypothesis.

16.1.5. Type I and Type II Errors. When we perform a hypothesis test, we derive the evidence from the sample in the form of a test-statistic. There is a possibility that the sample evidence may lead us to make a wrong decision. We may reject a null hypothesis H_0, when it is, in fact, true or we may accept a null hypothesis H_0, when it is actually false. The former type is called an *error of the first kind* or a *Type I-error*, while the latter, an *error of the second kind* or a *Type II-error*. The decision and the corresponding two types of error may be displayed in a tabular form as below:

True Situation	DECISION	
	Accept H_0	Reject H_0 (or accept H_1)
H_0 is true	Correct decision (No error)	Wrong decision (Type-I error)
H_0 is false	Wrong decision (Type-II error)	Correct decision (No error)

A legal analogy will help in understanding the difference between Type I and Type II errors. In a court trial, the supposition of Law is that the accused (the defendant) is innocent. This supposition of innocence may be regarded as a kind of null hypothesis H_0 that is to be rejected or accepted. After having heard the evidence presented during the trial, the judge arrives at a decision. Suppose the accused is, in fact, innocent (*i.e.* H_0 is true), but the finding of the judge is guilty. The judge has rejected a true null hypothesis and in so doing, has made a Type I error. If, on the other hand, the accused is, in fact, guilty (*i.e.* H_0 is false) and the finding of the judge is innocent, the judge has accepted a false null

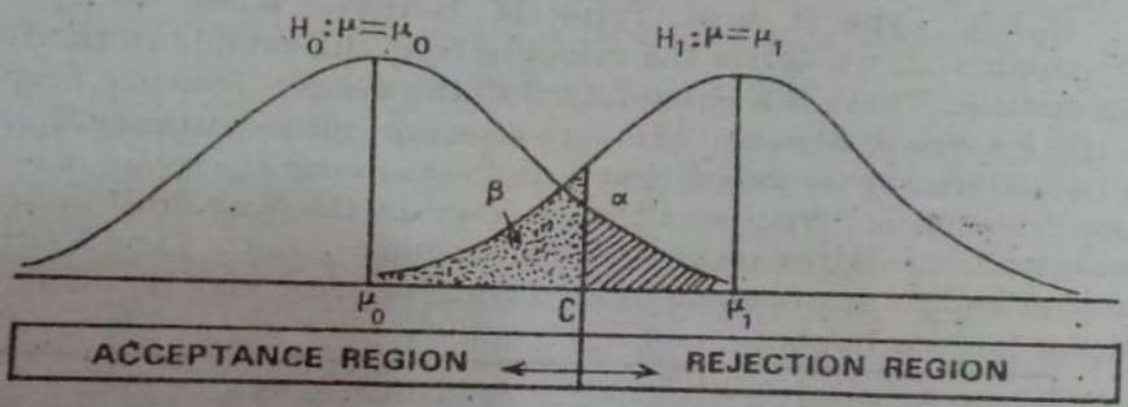
hypothesis and by accepting a false hypothesis, he has committed a Type II error.

The probability of making a Type I error is conventionally denoted by α (*alpha*) and that of committing a Type II error is indicated by β (*beta*). Thus α is the probability of rejecting H_0 when H_0 is true and β is the probability of accepting H_0 when H_0 is false (*i.e.* H_1 is true). In symbols, we may write

$$\alpha = P \text{ (Type I error)} = P \text{ (reject } H_0 / H_0 \text{ is true)},$$

$$\beta = P \text{ (Type II error)} = P \text{ (accept } H_0 / H_0 \text{ is false)}.$$

Let us consider two distributions: one under the null hypothesis $H_0 : \mu = \mu_0$ (*i.e.* distribution assuming H_0 is true) and the other under alternative hypothesis $H_1 : \mu = \mu_1$ (*i.e.* distribution assuming H_1 is true).



The probabilities of α and β are the shaded and dotted areas respectively of the distributions under the null hypothesis and under the alternative hypothesis. When our null hypothesis H_0 is true, then any value greater than or equal to C (the critical point) constitutes the rejection region equal to α (one–sided). That is α is associated with extreme values of the μ_0-distribution. The commonly used values of α are 0.05 and 0.01. On the other hand, β is associated with the area under the μ_1-distribution in the *acceptance region* established from μ_0-distribution. The probability of accepting H_0 when H_1 is true, *i.e.* β, thus depends both on the null hypothesis H_0 and on the alternative hypothesis H_1. In order to determine β (the probability of Type II error) we require α (the probability of Type I error) and the values of both μ_0 and μ_1. When α becomes smaller, β tends to become larger and when α becomes larger, β tends to become smaller. Thus there is an inverse relationship between α and β. We can decrease both α and β by increasing the sample size.

Example 16.1. The proportion of adults living in a small town who are matriculates is estimated to be $p=0.3$. To test this hypothesis a random sample of 15 adults is selected. If the number of matriculates in our sample is anywhere from 2 to 7, we shall accept the null hypothesis that $p=0.3$; otherwise we shall conclude that $p \neq 0.3$. Evaluate α assuming $p=0.3$. Evaluate β for the alternatives $p=0.2$ and $p=0.4$.

(P.U.,B.A/B.Sc.1986)

The null and alternative hypotheses are given as

$$H_0 : p = 0.3 \text{ and } H_1 : p \neq 0.3.$$

Let X denote the number of adults who are matriculates. Then the test-statistic has the binomial distribution with $p=0.3$ and $n=15$. The acceptance region, as given, consists of all values from $X=2$ to $X=7$. Then the critical region is composed of two parts: all values less than 2 and all values greater than 7. Thus the probability of making Type I error, *i.e.* α consists of $P(X<2)$ and $P(X>7)$.

Hence $\alpha = P(X<2 \text{ when } p=0.3) + P(X>7 \text{ when } p=0.3)$

$$= \sum_{x=0}^{1} b(x\,;\,15,\,0.3) + \sum_{x=8}^{15} b(x;\,15,\,0.3)$$

$$= \sum_{x=0}^{1} b(x\,;\,15,\,0.3) + [1 - \sum_{x=0}^{7} b(x;\,15,\,0.3)]$$

$$= 0.0353 + [1 - 0.9500] \quad \text{(From Binomial probability tables)}$$

$$= 0.0853$$

To compute β, the probability of Type II error, we need a specific alternative hypothesis. Now, we are given $H_0 : p=0.3$; and $H_1 : p=0.2$. A Type II error results when a *false* null hypothesis is accepted. That is a Type II error occurs if any value of the distribution under $H_1 : p=0.2$ falls in the region $X=2$ to $X=7$, the acceptance region of the distribution under null hypothesis $H_0 : p=0.3$.

Hence $\beta = P(2 \leq X \leq 7 \text{ when } H_1 : p = 0.2)$

$$= \sum_{x=2}^{7} b(x;\,15,\,0.2)$$

$$= \sum_{x=0}^{7} b(x;\,15,\,0.2) - \sum_{x=0}^{1} b(x;\,15,\,0.2)$$

$$= 0.9958 - 0.1671 \quad \text{(From Binomial probability tables)}$$

$$= 0.8287$$

Similarly, when $H_1 : p = 0.4$, we have

$$\beta = P(2 \le X \le 7, \text{ when } p = 0.4)$$

$$= \sum_{x=2}^{7} b(x; 15, 0.4)$$

$$= \sum_{x=0}^{7} b(x; 15, 0.4) - \sum_{x=0}^{1} b(x; 15, 0.4)$$

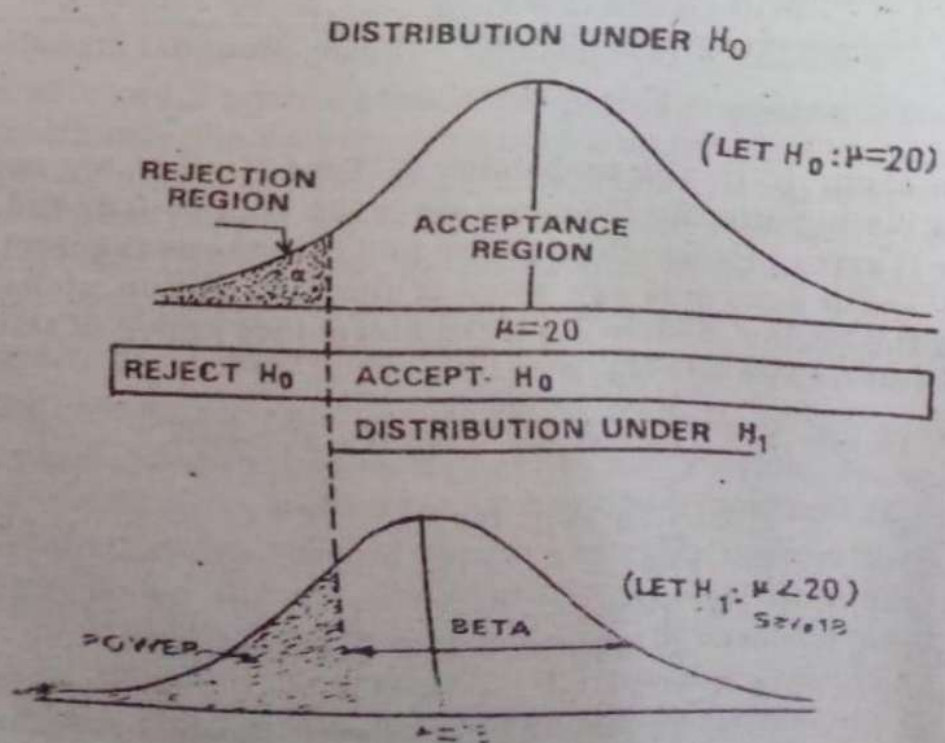
$$= 0.7869 - 0.0052 = 0.7817$$

16.1.6. The Power of a Test with respect to a specified alternative hypothesis, is the probability of rejecting a null hypothesis when it is actually false. The *power* is the complement of β, the probability of committing a Type II error. It is therefore numerically equivalent to one minus β. Symbolically,

$$\text{Power} = P(\text{reject } H_0 / H_0 \text{ is false})$$

$$= 1 - \beta$$

To represent α, β and power of a test graphically, we show the distributions of the test-statistic under both hypotheses H_0 and H_1 as below:



The shaded area in the lower diagram represents *power*. This probability corresponds to the rejection region of the distribution under H_0. The power generally increases with an increase in the sample size. A test for which β is small, is defined to be a *powerful test*.

A curve giving the probabilities of making Type II errors for various parameteric values under alternative hypotheses, is called an *Operating Characteristic Curve* or simply the *OC* curve. The *Power curve* which may be regarded as the complement of the *OC* curve, shows the probabilities of rejecting the null hypothesis H_0 for various values of the parameter θ.

16.1.7. The Significance Level of a test is the probability used as a standard for rejecting a null hypothesis H_0 when H_0 is assumed to be true. This probability is equal to some small pre-assigned value, conventionally denoted by α. The value α is also known as the *size of the critical region*. It is note-worthy that the significance level and the probability of Type I error are equivalent. The most frequently used values of α, the significance level, are 0.05 and 0.01, *i.e.* 5 percent and 1 percent but occasionally 0.10 or 0.001 is used. By $\alpha = 5\%$, we mean that there are about 5 chances in 100 of incorrectly rejecting a true null hypothesis. To put it in another way, we say that we are 95% confident in making the correct decision.

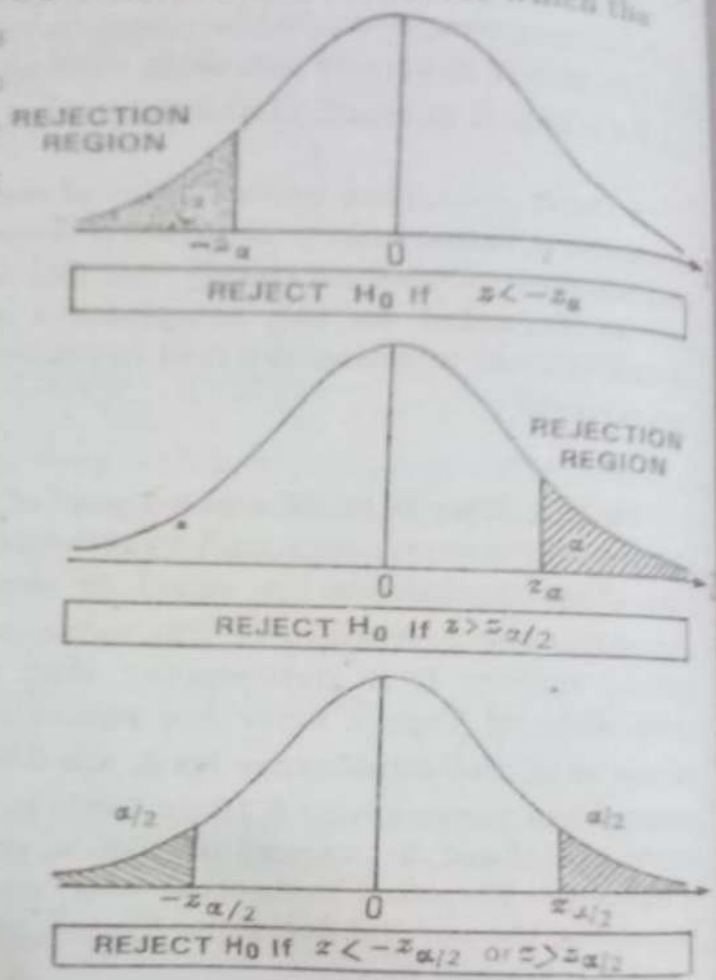
16.1.8. Test of Significance. A *test of significance* is a rule or procedure by which sample results are used to decide whether to accept or reject a null hypothesis. Such a procedure is usually based on a test-statistic and the sampling distribution of such a statistic under H_0. A value of the statistic is said to be *statistically significant* when the probability of its occurrence under H_0 is equal to or less than the significance level α, that is the value falls in the rejection region, H_0 in this case is rejected. If, on the other hand, the value falls in the acceptance region, it is said to be *statistically insignificant*. In this case, H_0 may be accepted. There are two desirable qualities for a test of significance. First, when the null hypothesis is actually true, it must have a low probability of rejecting H_0, and secondly, when H_0 is actually false, it must have a high probability of rejecting H_0. It is to be noted that the word *significant* is used in a special sense.

16.1.9. One-tailed and Two-tailed Tests. A test for which the entire rejection region is located in only one of the two tails—either in the right tail or in the left tail—of the sampling distribution of the test-statistic, is called a *One-tailed test* or *One-sided test*. For example, if Z is a test-statistic, then the rejection region consists of all z-values which are greater than $+z_\alpha$ or less than $-z_\alpha$ where α is the size of critical region. A one-tailed test is used when the alternative hypothesis H_1 is formulated in the following form:

$$H_1 : \theta > \theta_0 \text{ or}$$

$$H_1 : \theta < \theta_0 \text{ i.e.}$$

H_1 is composite hypothesis.



If, on the other hand, the rejection region is divided equally between the two tails of the sampling distribution of the test-statistic, the test is referred to as a *Two-tailed test* or *Two-sided test*. In this case, the alternative hypothesis H_1 is set up as:

$$H_1 : \theta \neq \theta_0, \text{ (i.e. } H_1 \text{ is two-sided composite hypothesis)}$$

meaning thereby that values both larger and smaller than θ_0 are to be included. In case of test-statistic being normal distribution, the rejection (critical) regions are shown by shading the appropriate portions of area under the sampling distribution in the figures shown above. The location of critical region can be determined only after the alternative hypothesis H_1 has been stated. It is important to note that the one-tailed and the two-tailed tests differ only in location of the critical region, not in the size.

Example 16.2. We wish to test the hypothesis that the mean weight of a population of people is 140 lbs. Using $\sigma = 15$ lb, $\alpha = 0.05$ and a sample of 36 people, find

(a) the values of $\bar{x}$ which would lead to rejection of the hypothesis, and (b) β, the probability of Type II error, if $\mu = 150$ lb. Use a two sided test.

(P.U., M.Sc., 1970, 86)

We are given the following information:

$$H_0 : \mu = 140 \text{ lb}, \ \sigma = 15 \text{ lb}, \ \alpha = 0.05 \text{ and } n = 36.$$

(a) To find the values of $\bar{x}$ (the critical point) which would lead to rejection of the hypothesis $H_0 : \mu = 140$ lb, we use the test statistic (assuming normal population) given by

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}.$$

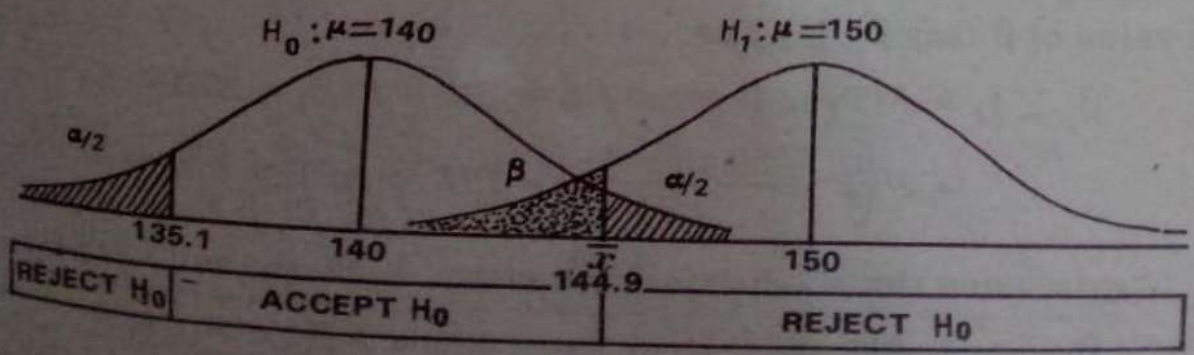
Since the test is two-sided, so there would be two critical values. Corresponding to the significance level $\alpha = 0.05$, the critical values of Z from the table of normal curve are -1.96 and 1.96. Thus

$$\pm 1.96 = \frac{\bar{x} - 140}{15 / \sqrt{36}}.$$

Simplifying, we get $\bar{x} = 135.1$ and 144.9 as the two critical values.

Hence the hypothesis $H_0 : \mu = 140$ lb will be rejected if $\bar{x} < 135.1$ lb or $\bar{x} > 144.9$ lb.

(b) A Type-II error can be committed only by accepting a false H_0. The hypothesis $H_0 : \mu = 140$ lb will be false if μ takes a value greater than 140 lb. Given $H_1 : \mu = 150$ lb so that $H_0 : \mu = 140$ lb becomes false. Therefore, the probability of accepting $H_0 : \mu = 140$ lb (false) when $H_1 : \mu = 150$, *i.e.* the probability of a Type-II error is indicated by the dotted area β in the figure shown below. To compute this area, we use the distribution under the alternative hypothesis $H_1 : \mu = 150$ lb.



at $\bar{x} = 135.1$, we find $z = \dfrac{135.1 - 150}{15/\sqrt{36}} = -5.96$.

Thus $\beta =$ Area between $z = -5.96$ and $z = -2.04$, *i.e.* area in the acceptance region of the distribution under H_0, when H_1 is $\mu = 150 = 0.0207$

This is the probability of accepting the null hypothesis $H_0: \mu = 140$ lb, when, in fact, the alternative hypothesis $H_1: \mu = 150$ lb is true.

Example 16.3. A random sample of size 4 is drawn from a normal population with known variance 15. A one-tailed test of the form $H_0: \mu \leq 30$ against $H_1: \mu > 30$ at the 5% level of significance is performed. Calculate the probabilties of Type-II error (β) for the values of $\mu = 31, 32, 34$ and 36 in the alternative hypothesis. Also calculate the powers of the test and hence sketch the power curve for this test.

Given $H_0: \mu \leq 30$ and $H_1: \mu > 30$; $\alpha = 0.05$, $n = 4$ and $\sigma^2 = 15$.

Under $H_0: \mu \leq 30$, the population under consideration is $N(30, 15/4)$.

To find the values of β, the probability of Type-II error, we first need to calculate the critical values. For a one-tailed test (the upper tailed) with $\alpha = 0.05$, the critical value is given by

$$c = \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

i.e.
$$c = 30 + (1.645)(\sqrt{15/4}) = 30 + (1.645)(\sqrt{3.75})$$
$$= 30 + (1.645)(1.94) = 33.19$$

Since 4 values of μ in the alternative hypothesis are specified, so we associate the variables $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ and $\bar{X}_4$ with each of the four alternative (H_1) distributions.

Using the alternative (H_1) distribution with $\mu = 31$, we calculate the value of β (say $\beta_{\mu = 31}$) as

$$\beta_{\mu = 31} = P(\text{Type-II error} / \mu = 31) = P(\bar{X}_1 < 33.19)$$
$$= P\left(Z < \frac{33.19 - 31}{1.94}\right) = P(Z < 1.13) = 0.8708$$

Again using the H_1-distribution with $\mu = 32$, the value of β is

$$\beta_{\mu = 32} = P(\text{Type-II error} / \mu = 32) = P(\bar{X}_2 < 33.19)$$
$$= P\left(Z < \frac{33.19 - 32}{1.94}\right) = P(Z < 0.61) = 0.7291$$

Similarly,

$$\beta_{\mu=34} = P(\text{Type-II error} / \mu = 34) = P(\bar{X}_3 < 33.19)$$
$$= P\left(Z < \frac{33.19 - 34}{1.94}\right) = P(Z < -0.42) = 0.3372$$

$$\beta_{\mu=36} = P(\text{Type-II error} / \mu = 36) = P(\bar{X}_4 < 33.19)$$
$$= P\left(Z < \frac{33.19 - 36}{1.94}\right) = P(Z < -1.45) = 0.0735$$

The power of the test for $\mu = \mu_1$, say $P_w(\mu_1)$ is given by $1-\beta_{\mu=\mu_1}$. Thus the required powers are:

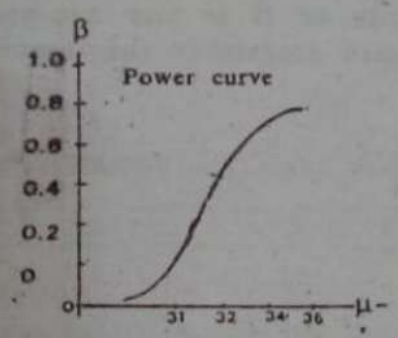
$P_w(31) = 1-\beta_{\mu=31} = 1 - 0.8708 = 0.1292,$

$P_w(32) = 1-\beta_{\mu=32} = 1 - 0.7291 = 0.2709,$

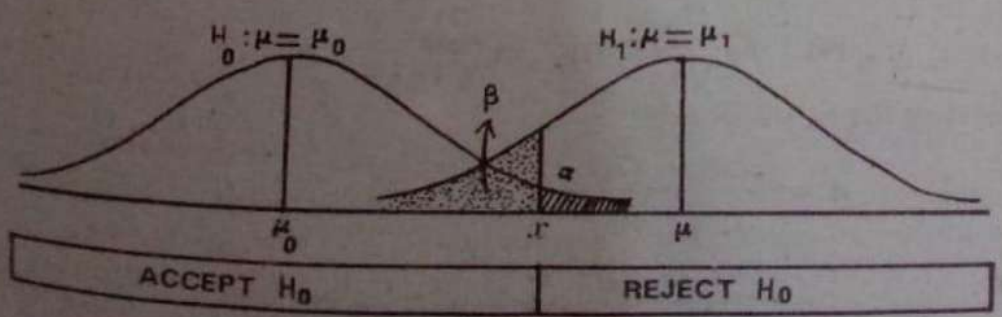
$P_w(34) = 1-\beta_{\mu=34} = 1 - 0.3372 = 0.6628,$ and

$P_w(36) = 1-\beta_{\mu=36} = 1 - 0.0735 = 0.9265.$

The sketch of the power curve for the test is shown below:



16.1.10. Sample size when α and β are specified. We may determine the sample size necessary for discriminating between the two hypotheses when the two Types of error are specified. Let us state our null hypothesis as $H_0 : \mu = \mu_0$ and the alternative hypothesis as $H_1 : \mu = \mu_1$. There are then two sampling distributions from two normal populations $N(\mu_0, \sigma_0^2)$ and $N(\mu_1, \sigma_1^2)$.



The rejection region under the μ_0-distribution is the area α (one-tailed test) and it lies to the *right* of the point x. The type II error under the μ_1-distribution is represented by the area β and it lies to the *left* of the point x, *i.e.* it is associated with the area under the μ_1-distribution in the *acceptance region* established from the μ_0-distribution. We know that every value to the right of x, *i.e.* falling in the critical region, calls for the rejection of the null hypothesis H_0 and every value to the left of x, *i.e.* falling in the acceptance region, calls for the acceptance of the null hypothesis H_0. When H_0 is true, the upper limit of the acceptance region (the point x) is determined by the expression

$$\frac{x-\mu_0}{\sigma_0/\sqrt{n}} = z_0 \text{ or } x = \mu_0 + z_0\frac{\sigma_0}{\sqrt{n}},$$

where z_0 is the normal deviate corresponding to the lower limit of the critical region.

Under alternative hypothesis H_1, corresponding to an area equal to the specified magnitude of β is the acceptance region under H_0. Considering H_1, we again determine the upper limit of the acceptance region by

$$\frac{x-\mu_1}{\sigma_1/\sqrt{n}} = -z_1 \qquad (-ve \text{ sign as the critical point } x \text{ lies to the left of } \mu_1)$$

$$\text{or} \qquad x = \mu_1 - z_1\frac{\sigma_1}{\sqrt{n}}$$

where z_1 is the normal deviate corresponding to the upper limit of the area representing Type-II error.

Hence the upper limit of the acceptance region can be represented by two relations, viz.,

$$x = \mu_0 + z_0\frac{\sigma_0}{\sqrt{n}}, \text{ and } x = \mu_1 - z_1\frac{\sigma_1}{\sqrt{n}}$$

Equating these two values of x, we get

$$\mu_0 + z_0\frac{\sigma_0}{\sqrt{n}} = \mu_1 - z_1\frac{\sigma_1}{\sqrt{n}}$$

Solving for n, we obtain

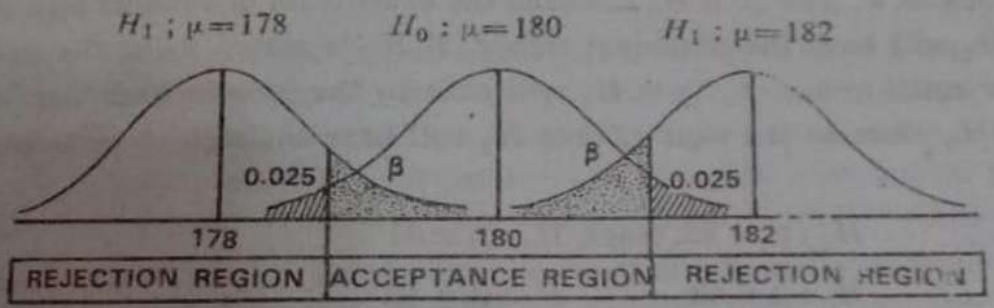
$$n = \frac{(\sigma_0 z_0 + \sigma_1 z_1)^2}{(\mu_1-\mu_0)^2}$$

$$= \frac{\sigma^2(z_0+z_1)^2}{(\mu_1-\mu_0)^2}, \text{ when } \sigma_0 = \sigma_1 = \sigma$$

The required sample size for a two-tailed test can be determined in a similar way.

Example 16.4. A firm wishes to test the hypothesis that the average monthly wage of its shop employees is Rs. 180 with a standard deviation of Rs. 4.50. They wish to run a risk of only 0.05 of rejecting the null hypothesis when it is true, and a risk of 0.05 of accepting the null hypothesis if the average is as high as Rs. 182 or as low as Rs. 178. How large a sample should be taken? (P.U., M.Sc. 1986)

A risk of only 0.05 of rejecting the null hypothesis when it is true, implies that the probability of a type I error, *i.e.* $\alpha = 0.05$ and a risk of accepting the null hypothesis if the average is as high as Rs. 182 or as low as Rs. 178, implies that the probability of a type II error, *i.e.* $\beta = 0.05$. The distributions of average monthly wage for alternative values of μ and the acceptance and rejection regions at the specified levels of α and β are represented by normal distributions in the following figure:



The value of the variable Z for two tailed test at $\alpha = 0.05$ under the μ_0-distribution from area tables is $z_0 = 1.96$ and the value of the variable Z at $\beta = 0.05$ under the μ_1-distribution from area tables is $z_1 = 1.645$.

Substituting these values in the formula

$$n = \frac{(z_0 + z_1)^2 \sigma^2}{(\mu_1 - \mu_0)^2}, \text{ we get}$$

$$n = \frac{(1.96 + 1.645)^2 (4.5)^2}{(182 - 180)^2}$$

$$= \frac{(3.605)^2 (4.5)^2}{4} = 65.79$$

Hence the sample size must be at least 66, the next higher integer.

16.1.11. Formulation of Hypotheses. To formulate the null (H_0) and alternative (H_1) hypotheses which are statements about parameters, not statistics, is perhaps the most difficult task. The hypotheses must be formulated in such a way that when one is true, the

other is false, *i.e.* H_0 and H_1 are opposites. The basic rule in formulating hypotheses is *to make H_1 the hypothesis that the experimenter thinks is true or the hypothesis that he(she) wants to establish as true.*

Sometimes hypotheses are formulated from an experimenter's attitude toward a claim. If the experimenter wishes to establish a certain claim with substantive support of sample information, then the claim is taken as the alternative hypothesis H_1 and its negation becomes the null hypothesis H_0. If the experimenter wants to disprove or refute the claim, then claim is made H_0. If the problem simply says to test a claim, it would be interpreted to mean that the claim is to be disproved (and would be made H_0).

In situations where we want to test for a change in the value of a parameter, the old or accepted value of parameter is used for H_0 and H_1 includes the new value.

The null hypothesis H_0 will always contain some form of an equality sign such as $=$, $\leq$ or $\geq$. If H_0 contains the exact form of equality sign $=$, then H_1 will have the not-equal sign $\neq$. If H_0 is stated using the *less-than-or-equal-to* sign $\leq$, then H_1 will contain the *greater than* sign $>$, and if H_0 contains the sign $\geq$, then H_1 will have the sign $<$. Examples are

$$H_0 : \mu = 62, \text{ (say)}, \quad H_1 : \mu \neq 62$$

$$H_0 : \mu \leq 62 \qquad H_1 : \mu > 62$$

$$H_0 : \mu \geq 62 \qquad H_1 : \mu < 62$$

It is of importance to note that by rejecting the null hypothesis $H_0 : \mu = 62$ (and accepting the alternative $H_1 : \mu > 62$), we are automatically rejecting all values of μ that are less than 62, because the null and the alternative hypotheses being opposites, cover all possible values for the parameter μ.

In general, if θ_0 is a specified value of a parameter θ, then the null and alternative hypotheses in case of *two-tailed test*, take the form

$$H_0 : \theta = \theta_0 \text{ and } H_1 : \theta \neq \theta_0,$$

and in case of *one-tailed test*, are stated in either of the following two forms

(i) $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$,

(ii) $H_0 : \theta \geq \theta_0$ and $H_1 : \theta < \theta_0$.

The sampling distribution of $\hat{\theta}$ in case of an inexact null hypothesis ($H_0 : \theta \le \theta_0$ or $H_0 : \theta \ge \theta_0$) is not defined and hence we cannot set up the acceptance and rejection regions. In such a case, we would take the null hypothesis *as if* it is an exact one, *i.e.* $H_0 : \theta = \theta_0$.

16.1.12. General Procedure for Testing Hypotheses. The procedure for testing a hypothesis about a population parameter involves the following six steps:

(i) State your problem and formulate an appropriate null hypothesis H_0 with an alternative hypothesis H_1, which is to be accepted when H_0 is rejected.

(ii) Decide upon a significance level, α of the test, which is the probability of rejecting the null Hypothesis if it is true.

(iii) Choose an appropriate test-statistic, determine and sketch the sampling distribution of the test-statistic, assuming H_0 is true.

(iv) Determine the rejection or critical region in such a way that the probability of rejecting the null hypothesis H_0, if it is true, is equal to the significance level, α. The location of the critical region depends upon the form of H_1. The significance level will separate the acceptance region from the rejection region.

(v) Compute the value of the test-statistic from the sample data in order to decide whether to accept or reject the null hypothesis H_0.

(vi) Formulate the decision rule as below:

(a) Reject the null hypothesis H_0, if the computed value of the test-statistic falls in the rejection region and conclude that H_1 is true.

(b) Accept the null hypothesis H_0, otherwise.

When a hypothesis is rejected, we can give a measure of the strength of the rejection by giving the *P-value*, the smallest significance level at which the null hypothesis is being rejected.

16.2 TESTS BASED ON NORMAL DISTRIBUTION

Suppose we wish to test a hypothesis that a parameter θ of a normal distribution has some specified value θ_0. We draw a random sample of size n from the population and calculate $\hat{\theta}$ as an estimate of θ. It has been shown in the previous chapter that the sampling distribution of $\hat{\theta}$ is a normal or approximately normal with mean θ and standard deviation

σ_θ. Then the variable $Z = \dfrac{\hat{\theta} - \theta}{\sigma_\theta}$ is $N(0, 1)$. If the null hypothesis H_0 : $\theta = \theta_0$ is true, then $\dfrac{\hat{\theta} - \theta_0}{\sigma_\theta}$ is $N(0, 1)$ and is used as the test-statistic for testing the hypothesis $H_0 : \theta = \theta_0$.

If the significance level is α, then the critical region will consist of all values of Z which are

(i) less than $-z_{\alpha/2}$ *and* greater than $z_{\alpha/2}$ in case of two-tailed test,

(ii) less than $-z_\alpha$ *or* greater than z_α in case of one-tailed test.

Critical values of Z for the most frequently used values of α, are given below:

Significance level (α)	Two-tailed test	One-tailed test
0.10	$\pm 1.645 = z_{\alpha/2}$	$\pm 1.28 = z_\alpha$
0.05	$\pm 1.96 = z_{\alpha/2}$	$\pm 1.645 = z_\alpha$
0.01	$\pm 2.58 = z_{\alpha/2}$	$\pm 2.33 = z_\alpha$

The decision rule is then formulated as below:

Reject the null hypothesis H_0, when the z value of the statistic exceeds these values. Accept the null hypothesis H_0, otherwise.

In this chapter, we deal with the following tests of hypotheses:

(i) To test whether the mean μ of a normal population, is equal to a specified value μ_0, when the population standard deviation σ is known. Symbolically, $H_0 : \mu = \mu_0$, when σ is known.

(ii) To test whether the mean μ of a normal population is equal to a specified value μ_0, when the population standard deviation is not known and sample size is large.

(iii) To test whether the mean μ of a non-normal population is equal to a specified value μ_0, when the sample size is large.

(iv) To test whether the difference between means of two normal distributions $(\mu_1 - \mu_2)$ is equal to a specified value Δ_0 (Δ_0 may be zero), when σ_1 and σ_2 are known, *i.e.* $H_0 : \mu_1 - \mu_2 = \Delta_0$ when σ_1 and σ_2 are known.

(v) To test whether the difference between means of two normal distributions is equal to specified value, Δ_0 when σ_1 and σ_2 are not known and sample sizes are large.

(vi) To test whether the difference between means of two non-normal distributions is equal to a specified value Δ_0, when sample sizes are large.

(vii) Other tests based on the normal distribution for *large* sample size such as tests of standard deviation, proportions, etc.

16.2.1. Testing Hypothesis about Mean of a Normal Population when σ is known. Suppose a random sample of size n is drawn from a normal population with mean having a specified value μ_0 and a known standard deviation σ. The sample mean is given by $\bar{x}$. We wish to determine whether the sample accords with the hypothesis that the population mean μ has the specified value μ_0. For this purpose, we employ the normal distribution test $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$, and the *procedure* is outlined below:

(i) Formulate the null and alternative hypotheses about μ. Three possible forms are

(a) $H_0 : \mu = \mu_0$, and $H_1 : \mu \neq \mu_0$

(b) $H_0 : \mu \leq \mu_0$, and $H_1 : \mu > \mu_0$

(c) $H_0 : \mu \geq \mu_0$, and $H_1 : \mu < \mu_0$.

(ii) Decide on significance level α, *i.e.* take $\alpha = 0.05$ or 0.01.

(iii) The test-statistic in this case will be $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$. Under the null hypothesis, Z has a standard normal distribution.

(iv) Determine the rejection region, which actually depends on the alternative hypothesis, from the table of areas under the normal curve by finding areas exactly equal to α. The rejection regions for H_0 corresponding to different alternative hypotheses are given below:

When the alternative hypothesis is	the rejection region will be
(a) $H_1 : \mu \neq \mu_0$ (two-sided)	$z < -z_{\alpha/2}$ and $z > z_{\alpha/2}$
(b) $H_1 : \mu > \mu_0$ (one-sided)	$z > z_\alpha$
(c) $H_1 : \mu < \mu_0$ (one-sided)	$z < -z_\alpha$

For example, when H_1 is $\mu \neq \mu_0$, the areas in the two tails for $\alpha = 0.05$ would be 0.05 if $z < -1.96$ and $z > 1.96$, as the critical values of Z are $-z_{0.025} = -1.96$ and $z_{0.025} = 1.96$. In this text, where the alternative hypothesis has not been stated explicitly or implicitly, a two-sided alternative hypothesis has been assumed.

(v) Calculate the value of Z from the sample data.

(vi) Decide as below:

Reject H_0, when the calculated value of Z falls in the rejection region, otherwise, accept it.

In case of rejection, the decision would be that μ differs from μ_0.

Example 16.5. A random sample of $n=25$ values gives $\bar{x}=83$. Can this sample be regarded as drawn from a normal population with mean $\mu=80$ and $\sigma=7$?

(i) We formulate our null and alternative hypotheses as

$$H_0 : \mu = 80 \text{ and } H_1 : \mu \neq 80 \quad \text{(two-sided)}$$

(ii) We set the significance level at $\alpha = 0.05$.

(iii) The test-statistic to be used is $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$, which under the null hypothesis is a standard normal variable.

(iv) The critical region for $\alpha = 0.05$ is $|Z| \geq 1.96$. The hypothesis will be rejected if, for the sample, $|Z| \geq 1.96$.

(v) We calculate the value of Z from the sample data as

$$z = \frac{83 - 80}{7 / \sqrt{25}} = \frac{3 \times 5}{7} = 2.14.$$

(vi) **Conclusion.** Since our calculated value $z = 2.14$ falls in the critical region, so we *reject* our null hypothesis $H_0 : \mu = 80$ and accept $H_1 : \mu \neq 80$. We may conclude that the sample with $\bar{x} = 83$ cannot be regarded as drawn from the population with $\mu=80$.

Example 16.6. Test the hypothesis that the mean of a normal population with known variance 70 is 31, if a sample of size 13 gave $\bar{x} = 34$. Let the alternative hypothesis be $H_1 : \mu > 31$, and let $\alpha = 0.10$.

(i) We formulate our hypotheses as

$$H_0 : \mu = 31, \text{ and } H_1 : \mu > 31 \quad \text{(one-sided)}$$

(ii) We are given the significance level as $\alpha = 0.10$.

(iii) The test-statistic to be used is $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$, which under null hypothesis has a standard normal distribution.

(iv) The critical region for $\alpha = 0.10$ is $Z > 1.28$

(v) We calculate the value of Z from the sample data as

$$z = \frac{34 - 31}{\sqrt{70} / \sqrt{13}} = \frac{3 \times 3.606}{8.367} = 1.29$$

(vi) **Conclusion.** Since our calculated value $z = 1.29$ falls in the critical region, so we reject our null hypothesis $H_0 : \mu = 31$. We may conclude that there is evidence at the 10% level that sample does not come from the given population.

16.2.2. Testing Hypothesis about Mean of a Normal Population when σ is unknown and $n > 30$. When the population standard deviation σ is not known, we use the sample standard deviation S as an estimate of the true but unknown population standard deviation. For large sample size ($n > 30$), the *central limit theorem* allows us to assume that the sampling distribution of $\bar{X}$ is approximately normal with a mean of μ and a standard deviation of $\dfrac{S}{\sqrt{n}}$. In other words, when σ is unknown but n is large, we replace $\dfrac{\sigma}{\sqrt{n}}$ with $\dfrac{S}{\sqrt{n}}$, and the variable $Z = \dfrac{\bar{X} - \mu_0}{S / \sqrt{n}}$, which is approximately $N(0, 1)$, is then used as the test-statistic to test the hypothesis $H_0 : \mu = \mu_0$. The rest of the procedure is the same.

Example 16.7. The marks obtained by students at a large number of colleges are known to be normally distributed with a mean of 25. A random sample of 36 students showed an average number of marks of 27 with a standard deviation of 5. What conclusion should be drawn?

(i) Since it is given that the average number of marks obtained by students at a large number of colleges is 25, we therefore have

$$H_0 : \mu = 25 \text{ and } H_1 : \mu \neq 25 \quad \text{(two-tailed)}$$

(ii) Let us specify the significance level at $\alpha = 0.05$.

(iii) Because σ is not known and the sample size $n > 30$, we use S in place of σ. Thus the test-statistic is $Z = \dfrac{\bar{X} - \mu_0}{S / \sqrt{n}}$, which has an approximate standard normal distribution under the given null hypothesis.

(iv) The rejection region is $|Z| \geq 1.96$.

(v) Computing the value of Z, we find

$$z = \frac{27 - 25}{5 / \sqrt{36}} = \frac{2 \times 6}{5} = 2.40.$$

(vi). **Conclusion.** Since the calculated value $z = 2.40$ falls in the rejection region, we therefore reject $H_0 : \mu = 25$ and accept $H_1 : \mu \neq 25$. On the basis of the evidence, we may conclude that this sample of students appears to be superior.

16.2.3. Testing Hypothesis about Mean of a Non-Normal Population when sample size is large. The *central limit theorem* tells us that for large sample sizes, the sampling distribution of $\bar{X}$ is approximately a normal even though the population sampled is non-normal. That is, the random variable $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ or $Z = \dfrac{\bar{X} - \mu_0}{S / \sqrt{n}}$ according as σ is known or not known, is approximately standard normal and is used as the test-statistic to test the hypothesis $H_0 ; \mu = \mu_0$. The rest of the procedure is the same.

Example 16.8. A random sample of 100 workers with children in day care shows a mean day-care cost of Rs. 2,600 and a standard deviation of Rs. 500. Verify the department's claim that the mean exceeds Rs. 2,500 at the 0.05 level with this information.

We make H_1, what the department claims, that the mean exceeds Rs. 2,500, and take the negation of its claim as H_0. Thus, we have

(i) $H_0 : \mu \leq 2{,}500$

$H_1 : \mu > 2{,}500$ (exceeds 2,500)

(ii) We are given the significance level at $\alpha = 0.05$.

(iii) The test-statistic, under H_0 is

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}},$$

which is approximately normal as $n = 100$ is large enough to make use of the central limit th

(iv) The rejection region is $Z \geq Z_{0.05} = 1.645$

(v) Computing the value of Z from sample information, we find

$$z = \frac{2600 - 2500}{500/\sqrt{100}} = \frac{100}{50} = 2$$

(vi) **Conclusion.** Since the calculated value $z = 2$ falls in the rejection region, we therefore reject H_0, and may conclude that the department's claim is supported by the sample evidence.

Example 16.9. A random sample of 100 observations from a population known to be non-normal yielded the sample values $\bar{x} = 182$ and $S^2 = 299$. Test the hypothesis $H_0 : \mu \leq 180$ against $H_1 : \mu > 180$. Let $\alpha = 0.05$.

(i) We are given our hypotheses as

$$H_0 : \mu \leq 180, \text{ and}$$

$$H_1 : \mu_1 > 180. \quad \text{(one-sided)}$$

(ii) The significance level is $\alpha = 0.05$.

(iii) Since the sample size ($n = 100$) is large enough to allow to assume that the sampling distribution of $\bar{X}$ is approximately normal with mean $= \mu$ and standard deviation $= \frac{S}{\sqrt{n}}$, we therefore use the variable $Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ as the test-statistic under the given hypothesis.

(iv) The critical region is $Z > z_{0.05} = 1.645$.

(v) Here $\mu_0 = 180$, $\bar{x} = 182$, $n = 100$ and $S = \sqrt{299} = 17.29$

$$\therefore \quad z = \frac{\bar{x} - \mu_0}{S / \sqrt{n}} = \frac{182 - 180}{17.29 / \sqrt{100}} = \frac{20}{17.29} = 1.16$$

(vi) **Conclusions.** Since our calculated value $z = 1.16$ falls in the acceptance region, so we accept $H_0 : \mu \leq 180$ and reject $H_1 : \mu > 180$.

16.2.4. Testing Hypotheses about Difference between Two Population Means. To test hypotheses about the difference between two population means, we deal with the following three cases:

(1) Both the populations are normal with known standard deviations.

(2) Both the populations are normal with unknown standard deviations.

(3) Both the populations are non-normal, in which case, both sample sizes are necessarily large.

Case 1. Let $\bar{X}_1$ be the mean of the first random sample of size n_1 from a normal population with a mean of μ_1 and a known standard deviation σ_1, and $\bar{X}_2$ be the mean of the second random sample of size n_2 from another normal population with a mean of μ_2 and a known standard deviation σ_2. Then the sampling distribution of the difference $\bar{X}_1-\bar{X}_2$ is normally distributed with a mean of $\mu_1-\mu_2$ and a standard deviation of $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$. In other words, the variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

is *exactly* standard normal, no matter how small the sample sizes are. Hence it is used as the test-statistic for testing hypotheses about the difference between two population means. The difference is equal to, less than, or greater than a specified value Δ_0 (Δ_0 may equal zero). Then the hypotheses are

(i) $H_0 : \mu_1 - \mu_2 = \Delta_0$, and $H_1 : \mu_1 - \mu_2 \neq \Delta_0$,

(ii) $H_0 : \mu_1 - \mu_2 \leq \Delta_0$, and $H_1 : \mu_1 - \mu_2 > \Delta_0$,

(iii) $H_0 : \mu_1 - \mu_2 \geq \Delta_0$, and $H_1 : \mu_1 - \mu_2 < \Delta_0$,

When the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ is true, the test-statistic becomes

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \text{ which is } \textit{exactly} \text{ standard normal.}$$

In case $\Delta_0 = 0$, the test-statistic reduces to

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}.$$

The *procedure* for testing the hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ may be stated as below:

(i) Formulate the null and the alternative hypotheses:

$H_0 : \mu_1 - \mu_2 = \Delta_0$ against the appropriate alternative.

(ii) Decide on the significance level α.

(iii) The test-statistic Z, which under H_0 becomes

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is *exactly* standard normal.

(iv) The rejection region is

$$Z < -z_{\alpha/2} \text{ and } Z > z_{\alpha/2}, \text{ when } H_1 \text{ is } \mu_1 - \mu_2 \neq \Delta_0,$$

$$Z > z_\alpha, \text{ when } H_1 \text{ is } \mu_1 - \mu_2 > \Delta_0,$$

$$Z < -z_\alpha, \text{ when } H_1 \text{ is } \mu_1 - \mu_2 < \Delta_0,$$

(v) Compute the value of Z from the sample data.

(vi) Decide as below:

Reject H_0 if z falls in the critical region, accept H_0, otherwise.

Example 16.10. A random sample of size 36 from a normal population with variance 24 gave $\bar{x}_1 = 15$. A second sample of size 28 from another normal population with variance 80 gave $\bar{x}_2 = 13$. Test $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$. Let $\alpha = 0.05$.

(i) The null and the alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0, \text{ and} \qquad (\because \Delta_0 = 0)$$

$$H_1 : \mu_1 - \mu_2 \neq 0. \quad \text{(two sided)}$$

(ii) The significance level is $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

which is exactly standard normal.

(iv) The rejection region is $|Z| \geq 1.96$.

(v) We compute the value of Z from the sample data as

$$z = \frac{15 - 13}{\sqrt{\frac{24}{36} + \frac{80}{28}}} = \frac{2}{\sqrt{3.5238}} = \frac{2}{1.88} = 1.06$$

(vi) **Conclusion.** Since the calculated value $z = 1.06$ does not fall in the rejection region, so we do not reject $H_0 : \mu_1 - \mu_2 = 0$.

Example 16.11. The two samples A and B detailed below, were taken from normal populations of standard deviation 0.8. Test whether the difference of means is significant.

A	10.5,	11.6,	12.7,	12.9,	13.5,	13.6,	14.8	
B	11.3,	12.4,	12.4,	13.9,	14.2,	14.7,	14.9	15.6

(P.U., B.A./B.Sc. 1983)

(i) We formulate our null and alternative hypotheses as

$$H_0 : \mu_1 - \mu_2 = 0 \text{ and } H_1 : \mu_1 - \mu_2 \neq 0 \quad \text{(two-sided)}$$

(ii) We set the significance level at $\alpha = 0.05$.

(iii) The test-statistic to be used under H_0 is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \quad (\because \sigma_1 = \sigma_2 = \sigma)$$

which is exactly standard normal.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations: Here $n_1 = 7$, $n_2 = 8$, $\sigma = 0.8$,

$$\bar{X}_1 = \frac{\Sigma X_A}{n_1} = \frac{89.6}{7} = 12.8, \text{ and}$$

$$\bar{X}_2 = \frac{\Sigma X_B}{n_2} = \frac{109.4}{8} = 13.675.$$

$$\therefore \quad z = \frac{12.8 - 13.675}{0.8\sqrt{\dfrac{1}{7} + \dfrac{1}{8}}} = \frac{-0.875}{(0.8)(0.5176)} = \frac{-0.875}{0.414} = -2.11.$$

(vi) **Conclusion.** Since the calculated value $z = -2.11$ falls in the rejection region, so we reject $H_0 : \mu_1 - \mu_2 = 0$. On the basis of the evidence, we may conclude that the difference between means is significant.

Case 2. When independent random samples of sizes n_1 and n_2 are drawn from normal populations with known means μ_1 and μ_2 but unknown standard deviations, the sample standard deviations S_1 and S_2 can be substituted for the population standard deviation σ_1 and σ_2

sample sizes are large ($n_1, n_2 > 30$), we can assume that the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal with a mean $\mu_1 - \mu_2$ and a standard deviation $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$. That is, the variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

is approximately $N(0, 1)$. The test-statistic to be used under the hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ becomes

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

which is approximately standard normal. The rest of the *procedure* is the same.

Example 16.12. A form of intelligence test was given to random samples of soldiers and sailors in a certain country. The following results were recorded:

	Number in Samples	Mean Score	Sample Standard Deviation
Soldiers	332	12.78	2.43
Sailors	615	12.99	2.48

Assume the populations of scores to be normal. What conclusion should be drawn?

(i) We must decide between the hypotheses

$H_0: \mu_1 - \mu_2 = 0$, *i.e.* there is no difference between the mean scores,

$H_1: \mu_1 - \mu_2 \neq 0$, *i.e.* there is a significant difference between means.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) Since the population standard deviations are unknown and the sample sizes are large, we therefore substitute the sample standard deviations for the population standard deviations. Then the test-statistic to be used under H_0 is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

which is approximately standard normal.

(iv) The critical region is $|Z| \geq 1.96$.

(v) We compute the value of Z from the given information as below:

$$z = \frac{12.99 - 12.78}{\sqrt{\frac{(2.43)^2}{332} + \frac{(2.48)^2}{615}}},$$

$$= \frac{0.21}{\sqrt{0.0178 + 0.0100}} = \frac{0.21}{0.17} = 1.24.$$

(vi) **Conclusion**. Since the calculated value $z = 1.24$ does not fall in the rejection region, therefore we accept the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$ at 5% significance level. In other words, on the basis of the evidence, we may conclude that the difference between mean scores is insignificant or merely due to chance.

Case 3. The populations are non-normal and the sample sizes are sufficiently large. It is interesting to note that the *central limit theorem* also applies for the sampling distribution of the difference between two sample means. Thus, if sufficiently large samples are drawn from the non-normal populations, the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ will be *approximately* normal with a mean of $\mu_1 - \mu_2$ and a standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ or $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$, according as the population standard deviations are known or unknown. In other words, if the sample sizes are sufficiently large, the variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S.E.\ (\bar{X}_1 - \bar{X}_2)}$$

is *approximately* standard normal, regardless of the form of the population distributions. The test-statistic under the hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ then becomes

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{S.E.\ (\bar{X}_1 - \bar{X}_2)},$$

which is approximately $N(0, 1)$. The rest of the procedure for testing the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ is the same.

Example 16.13. A random sample of size 40 from a non-normal population yielded the sample values $\bar{x}_1 = 70.4$, $S_1^2 = 31.40$. Another random sample of size 50 from a second non-normal population yielded

the sample values $\bar{x}_2 = 65.3$, $S_2^2 = 44.82$. Test $H_0 : \mu_1 - \mu_2 \leq 2$ against $H_1 : \mu_1 - \mu_2 > 2$. Let $\alpha = 0.05$.

(i) The null and the alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 \leq 2, \text{ and } H_1 : \mu_1 - \mu_2 > 2 \text{ (one-sided)}$$

(ii) The significance level is chosen at $\alpha = 0.05$.

(iii) Since the populations are non-normal and the sample sizes are large, therefore the test-statistic under H_0 is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}},$$

which is *approximately* standard normal.

(iv) The critical region is $Z > 1.645$ as for $\alpha = 0.05$, $z_{0.05} = 1.645$,

(v) Computations: The sample values are given to be

$$n_1 = 40, \bar{x}_1 = 70.4, \; S_1^2 = 31.40,$$

$$n_2 = 50, \bar{x}_2 = 65.3, \; S_2^2 = 44.82.$$

$$\therefore \quad z = \frac{(70.4 - 65.3) - 2}{\sqrt{\dfrac{31.40}{40} + \dfrac{44.82}{50}}} = \frac{3.1}{\sqrt{0.785 + 0.8964}}$$

$$= \frac{3.1}{1.30} = 2.34.$$

(vi) **Conclusion.** We see that z = 2.34 falls in the critical region. Therefore, we reject our null hypothesis $H_0 : \mu_1 - \mu_2 \leq 2$ and accept $H_1 : \mu_1 - \mu_2 > 2$.

16.2.5. Testing Hypothesis about a Population Proportion when sample size is large. Let $\hat{P}$ be the proportion of success in a sample of size n drawn from a binomial population having proportion p. If the sample size is sufficiently large, then $\hat{P}$ will be *approximately* normally distributed with a mean p and a standard deviation $\sqrt{\dfrac{pq}{n}}$, where $q = 1 - p$. In other words, if the sample is large, then the variable

$$Z = \frac{\hat{P} - p}{\sqrt{pq/n}}$$

is approximately standard normal.

When $\hat{P} = \frac{X}{n}$, where X is the actual number of successes in a random sample, the standardised variable becomes

$$Z = \frac{\frac{X}{n} - p}{\sqrt{\frac{pq}{n}}} = \frac{X - np}{\sqrt{npq}}.$$

Suppose we wish to test the null hypothesis that the population proportion p has a specified value p_0. If the hypothesis $H_0 : p = p_0$ is true, then the variable $\frac{\hat{P} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$ or $\frac{X - np_0}{\sqrt{np_0 q_0}}$ is *approximately* standard normal, and is used as a test-statistic for testing $H_0 : p = p_0$. The decision procedure is stated below:

(i) Formulate the null hypothesis and the appropriate alternative hypothesis.

(ii) Choose the significance level of size α.

(iii) The test-statistic, if the null hypothesis $H_0 : p = p_0$ is true, will be

$$Z = \frac{X - np_0}{\sqrt{np_0 q_0}}, \qquad \text{(without continuity correction)}$$

$$= \frac{(X \pm \frac{1}{2}) - np_0}{\sqrt{np_0 q_0}}, \qquad \text{(with continuity correction)}$$

For large n, Z is *approximately* standard normal.

(iv) The critical regions are established as under:

(a) When H_1 is $p \neq p_0$, the critical region is $Z < -z_{\alpha/2}$ and $Z > z_{\alpha/2}$.

(b) When H_1 is $p > p_0$, the critical region is $Z > z_\alpha$.

(c) When H_1 is $p < p_0$, the critical region is $Z < -z_\alpha$.

(v) Compute the value of Z from the sample data.

(vi) Decide as below:

Reject H_0 if z falls in the critical region.

Accept H_0 otherwise.

Example 16.14. A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be an unbiased one.

(i) We consider the null hypothesis that the coin is unbiased, *i.e.*,

$H_0 : p = \frac{1}{2}$ against the alternative hypothesis $H_1 : p \neq 0.5$.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic to be used under H_0, is

$$Z = \frac{(x \pm \frac{1}{2}) - np_0}{\sqrt{np_0 q_0}}, \qquad \text{(with continuity correction)}$$

which is approximately standard normal.

(iv) The critical region is $|Z| > 1.96$.

(v) We compute the value of Z as below:

$$z = \frac{(216 - \frac{1}{2}) - 400 \times \frac{1}{2}}{\sqrt{400 \times \frac{1}{2} \times \frac{1}{2}}} = \frac{15.5}{10} = 1.55$$

We have used $x - 1/2$ as $x > np_0$.

(vi) **Conclusion.** Since the computed value $z = 1.55$ is less than 1.96, we accept the hypothesis, and may conclude that the coin is an unbiased one.

Example 16.15. In an experiment to decide whether butter and margarine can be distinguished, 90 individuals were each given three sandwiches, two containing butter and the other one containing margarine. They were then asked which one of the three contains margarine, and 38 correct identifications were made. Test the significance of the test. (P.U., M.Sc. 1971)

(i) We formulate our hypotheses as

$H_0 : p \leq \frac{1}{3}$ and the individuals have no power of identification, *i.e.*, the results are due to chance.

$H_1 : p > \frac{1}{3}$.

(ii) We choose the level of significance at $\alpha = 0.05$

(iii) The t '-statistic to be sued under H_0, is

$$Z = \frac{X - np_0}{\sqrt{np_0q_0}}, \quad \text{(without continuity correction)}$$

$$= \frac{(X \pm \frac{1}{2}) - np_0}{\sqrt{np_0q_0}}, \quad \text{(with continuity correction)}$$

which is approximately a standard normal.

(iv) The critical region is $Z > 1.645$ because the alternative hypothesis is stated on a *greater than* basis.

(v) We then compute the value of Z as below:

$$z = \frac{38 - 90 \times \frac{1}{3}}{\sqrt{90 \times \frac{1}{3} \times \frac{2}{3}}} = \frac{8}{4.47} = 1.79. \text{ (without continuity correction)}$$

$$\text{or } z = \frac{(38 - \frac{1}{2}) - 90 \times \frac{1}{3}}{\sqrt{90 \times \frac{1}{3} \times \frac{2}{3}}} = 1.68, \text{ (with continuity correction)}$$

(vi) **Conclusion.** Since the calculated value of Z falls in the critical region in either case, so we reject $H_0 : p \leq \frac{1}{3}$ and may conclude that the results are *significantly* different from chance identification.

16.2.6. Testing Hypothesis about Difference between Two Proportions. Suppose we wish to test the hypothesis that the difference between two proportions is equal to a specified value Δ_0 or that the two proportions are equal. The statistic on which we base our decision rule is the variable $\hat{P}_1 - \hat{P}_2$, where $\hat{P}_1$ is the proportion of success in the first sample of size n_1 and $\hat{P}_2$ is the proportion of success in the second sample of size n_2, samples are drawn from two binomial populations with unknown proportion of success p_1 and p_2 respectively. If the samples are sufficiently large, the sampling distribution of the difference $\hat{P}_1 - \hat{P}_2$ is approximately normal with a mean of $p_1 - p_2$ and a standard deviation of $\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$. That is, for sufficiently large sample sizes, the variable

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$$

is *approximately* standard normal.

When the values of p_1 and p_2 are not known, then for large sample sizes, they are replaced with the sample proportions $\hat{p}_1$ and $\hat{p}_2$ respectively. The standardized variable then becomes

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

and the test-statistic, if the hypothesis $H_0 : p_1 - p_2 = \Delta_0$ is true, will be

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

If we wish to test the hypothesis $H_0 : p_1 = p_2 = p$, then p (the common population proportion) is replaced with its estimate $\hat{p}_c$, which is computed by taking a weighted mean of the two observed sample proportions $\hat{p}_1$ and $\hat{p}_2$ as follows:

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

The z value for testing the hypothesis $H_0 : p_1 = p_2$ then becomes

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c\hat{q}_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

The procedure for testing the hypothesis $H_0 : p_1 = p_2$ is given below:

(i) Formulate the null and the appropriate alternative hypotheses.

(ii) Decide upon the significance level of size α.

(iii) The test-statistic under H_0 is

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}_c\hat{q}_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

which, for large sample sizes, is *approximately* standard normal.

(iv) The critical regions are established as under:

(a) When H_1 is $p_1 - p_2 \neq 0$, the critical region is $Z < -z_{\alpha/2}$ and $Z > z_{\alpha/2}$.

(b) When H_1 is $p_1 - p_2 > 0$, the critical region is $Z > z_\alpha$.

(c) When H_1 is $p_1 - p_2 < 0$, the critical region is $Z < -z_\alpha$.

(v) Compute the value of Z from the sample data.

(vi) Decide as below:

Reject H_0, if z falls in the critical region.

Accept H_0, otherwise.

Example 16.16. In a random sample of 500 men from Lahore city, 300 are found to be smokers. In one of 1000 men from Karachi city, 550 are smokers. Do the data indicate that the two cities are significantly different with respect to the prevalence of smoking among men?

(i) We formulate our hypotheses as

$H_0 : p_1 = p_2$, *i.e.* there is no difference between the proportions of smokers;

$H_1 : p_1 \neq p_2$.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic under H_0, is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

where $\hat{p}_1$ = proportion of smokers in the city of Lahore,

$\hat{p}_2$ = proportion of smokers in the city of Karachi,

$\hat{p}_c$ = an estimate of the common population proportion on the assumption that the two cities are alike with respect to the prevalence of smoking among men, *i.e.*

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \text{ and } \hat{q}_c = 1 - \hat{p}_c.$$

The statistic Z, for large sample sizes, is *approximately* standard normal.

(iv) The critical region is $|Z| > 1.96$.

(v) Computations: Here $\hat{p}_1 = \frac{300}{500} = 0.60$, $\hat{p}_2 = \frac{550}{1000} = 0.55$, and

$$\hat{p}_c = \frac{300 + 550}{500 + 1000} = \frac{850}{1500} = 0.567, \text{ so that } \hat{q}_c = 0.433.$$

Thus $$z = \frac{0.60 - 0.55}{\sqrt{(0.567)\,(0.433)\left(\frac{1}{500} + \frac{1}{1000}\right)}}$$

$$= \frac{0.05}{\sqrt{(0.2455)\,(0.003)}} = \frac{0.05}{0.027} = 1.85.$$

(vi) **Conclusion.** Since the calculated value $z = 1.85$ does not fall in the critical region, so we accept the null hypothesis $H_0 : p_1 = p_2$ and conclude that the data do not appear to indicate that the two cities are different with respect to the prevalence of smoking among men.

Example 16.17. A candidate for mayor in a large city believes that he appeals to at least 10 per cent more of the women voters than the men voters. He hires the services of a poll-taking organisation, and they find that 62 of 100 women interviewed support the candidate, and 69 of 150 men support him. At the 0.05 significance level, is the hypothesis accepted or rejected?

Let p_1 = proportion of women voters, and

p_2 = proportion of men voters.

Then we make H_1 the hypothesis what the candidate for mayor believes as true and its negation as H_0. Thus

(i) The null and alternative hypotheses are

$H_0 : p_1 - p_2 \leq 0.10$, and (negation of candidate's claim)

$H_1 : p_1 - p_2 > 0.10$, (candidate believes as true)

(ii) We are given the significance level at $\alpha = 0.05$.

(iii) The test-statistic under H_0, is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0.10}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

which, for large sample sizes, is *approximately* standard normal.

(iv) The critical region is $Z > z_{0.05} = 1.645$

(v) Computations: Here $\hat{p}_1 = \frac{62}{100} = 0.62$, so that $\hat{q}_1 = 0.38$,

$\hat{p}_2 = \frac{69}{150} = 0.46$, so that $\hat{q}_2 = 0.54$,

$$\text{Thus} \quad z = \frac{(0.62 - 0.46) - 0.10}{\sqrt{\frac{(0.62)(0.38)}{100} + \frac{(0.46)(0.54)}{150}}}$$

$$= \frac{0.06}{\sqrt{0.002356 + 0.001656}} = \frac{0.06}{0.063} = 0.95.$$

(vi) **Conclusion.** Since the calculated value $z = 0.95$ does not fall in the critical region, so we accept the null hypothesis $H_0 : p_1 - p_2 \leq 0.10$.

16.2.7. Testing Hypotheses about Standard Deviation: Large Samples. Suppose that we wish to test the null hypothesis that the standard deviation σ of a normal population has a specified value σ_0. For a sufficiently large sample of size n, drawn from a normal distribution with mean μ and standard deviation σ, it is found that the *standard error* of the sample standard deviation S, is $\frac{\sigma}{\sqrt{2n}}$. The hypothesis $H_0 : \sigma = \sigma_0$ can then be tested by computing the following test-statistic

$$Z = \frac{S - \sigma_0}{\sigma_0 / \sqrt{2n}},$$

which, if $H_0 : \sigma = \sigma_0$ is true, is approximately $N(0, 1)$.

In case of two *large* samples of sizes n_1 and n_2, each drawn independently from a normal population, the standard error of the difference $S_1 - S_2$ is

$$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}.$$

It is found that for large sample sizes, the sampling distribution of the difference $S_1 - S_2$ is approximately normal with a mean of $\sigma_1 - \sigma_2$ and a standard deviation of $\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$. Thus the variable

$$Z = \frac{(S_1 - S_2) - (\sigma_1 - \sigma_2)}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

is approximately standard normal. The test-statistic, if $H_0 : \sigma_1 = \sigma_2$ is true, would be

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}},$$

which is approximately $N(0, 1)$.

The sample standard deviations can be substituted for population standard deviations when they are not known.

The credibility of the null hypothesis $H_0 : \sigma_1 = \sigma_2$ may be tested in the usual manner. It should be noted that this formula is not used when the populations are not normal.

Example 16.18. A random sample of size 100 from a normal population showed a standard deviation of 8.9. Test at the 0.05 level of significance the hypothesis that $\sigma = 7.5$ against the alternative that $\sigma \neq 7.5$.

(i) The null and alternative hypotheses are

$$H_0 : \sigma = 7.5 \text{ and } H_1 : \sigma \neq 7.5. \quad \text{(two-tailed)}$$

(ii) We are given the significance level at $\alpha = 0.05$.

(iii) Since the sample size $(n = 100)$ is sufficiently large and the sample is drawn from a normal population, so the test statistic under H_0, would be

$$Z = \frac{S - \sigma_0}{\sigma_0 / \sqrt{2n}},$$

which is approximately $N(0, 1)$.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations: We are given $S = 8.9$, $\sigma_0 = 7.5$ and $n = 100$.

$$\therefore \quad z = \frac{8.9 - 7.5}{7.5 / \sqrt{200}} = \frac{(1.4)(14.14)}{7.5} = 2.64.$$

(vi) **Conclusion.** Since the calculated value $z = 2.64$ falls in the critical region, so we reject the null hypothesis $H_0 : \sigma = 7.5$ in favour of the alternative $H_1 : \sigma \neq 7.5$.

Example 16.19. The following table gives means and standard deviations of marks obtained by the candidates in an examination held at two centres A and B. Assume the marks follow a normal distribution.

Centre	No. of candidates	Mean	Standard Deviation
A	1275	44.8	8.3
B	2346	47.3	6.5

Is the difference between the standard deviations significant?

(P.U., M.A. Econ., 1970)

(i) We set up our hypotheses as

H_0 : $\sigma_1 = \sigma_2$, *i.e.* there is no difference between the standard deviations, and

H_1 : $\sigma_1 \neq \sigma_2$.

(ii) We specify the significance level at $\alpha = 0.05$.

(iii) The test-statistic, if $H_0 : \sigma_1 = \sigma_2$ is true, is

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}},$$

where the sample standard deviations are substituted for population standard deviations as they are not known. The variable Z is approximately standard normal.

(iv) The critical region is $|Z| \geq 1.96$.

(v) We compute the value of Z by substituting the sample values as

$$z = \frac{8.3 - 6.5}{\sqrt{\frac{(8.3)^2}{2(1275)} + \frac{(6.5)^2}{2(2346)}}}$$

$$= \frac{1.8}{\sqrt{0.0270 + 0.0090}} = \frac{1.8}{0.19} = 9.47.$$

(vi) **Conclusion.** Since the calculated value $z = 9.47$ falls in the critical region, so we reject the hypothesis $H_0 : \sigma_1 = \sigma_2$ and conclude that the difference between the standard deviations is *significant*.

16.2.8. Relationship Between Confidence Interval and Tests of Hypothesis. There is a close relationship between the confidence interval for a parameter θ and a test of hypothesis about θ. Let $[L, U]$ be a $100(1-\alpha)\%$ confidence interval for the parameter θ. Then we will accept the null hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at a level of significance α if θ_0 falls inside the confidence interval, but if θ_0 falls outside the interval $[L, U]$, we will reject H_0. In the language of hypothesis testing, the $(1-\alpha)$ 100% confidence interval is known as the *acceptance region* and the region outside the confidence interval is called the *rejection* or *critical* region. The critical values are the end points of the confidence interval.

16.3 TESTS BASED ON BINOMIAL DISTRIBUTION

A binomial population as previously stated, consists of only two discrete classes such as males and females, cured and not-cured, successes and failures, occurrence and non-occurrence, good and defective, and so forth. The units of a sample drawn from such a population will fall into either one class or the other of the two discrete classes. If the proportion of one class (say, successes) is p, then the proportion of the second class will be $1 - p = q$. The sampling distribution of the proportion of successes in case of small sample ($n < 30$) will be *binomial distribution*, which is used to test the hypothesis that the proportion of successes in a binomial population is equal to a specified value, *i.e.* $H_0 : p = p_0$, where p_0 is the specified value of p, the parameter of the binomial distribution.

The procedure for testing $H_0 : p = p_0$, when sample size is small, is as follows:

(i) Formulate the null hypothesis as $H_0 : p = p_0$, with an appropriate alternative hypothesis about p.

(ii) Take the significance level at $\alpha = 0.05$ or 0.01. It may be impossible to set α at exactly 0.05 or 0.01 as the test-statistic is a discrete distribution.

(iii) The test-statistic is X, the number of successes in n trials, *i.e.* the test statistic is the binomial random variable X.

(iv) The rejection region will consist of all values of X whose probabilities (areas) are equal to or less than the significance level α. In case of *one-tailed test*, the probabilities (areas) in the desired tail are added till we reach the significance level α. In case of *two-tailed test*, the probabilities are added from both the tails in such a way that the sum is equal to or less than α and half sum comes from each tail.

(v) Find x, the number of successes. To make the computations easier, the probabilities are shown either graphically or by cumulative or decumulative columns.

(vi) Decide as below:

Reject the hypothesis H_0, if x falls in the critical region.
Accept H_0, otherwise.

Example 16.20. A coin is tossed 8 times and comes up heads 7 times. Can we conclude that the coin is fair at a significance level of 0.05?

(i) Let p denote the probability of heads in a single toss of the coin. Then our null hypothesis that the *coin is fair*, will be formulated as

$$H_0 : p = 0.5$$

and the alternative hypothesis would be $H_1 : p \neq 0.5$.

(ii) The significance level is approximately 0.05.

(iii) The test-statistic to be used is x, the number of heads.

(iv) *Critical Region*. First we compute the probabilities associated with X, the number of heads, by using the binomial distribution

$$P(X=x) = \binom{8}{x} p^x q^{8-x}. \text{ Under } H_0 : p = \frac{1}{2},$$

$$P(X=x) = \binom{8}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{8-x} = \binom{8}{x} \left(\frac{1}{2}\right)^x$$

The probabilities of 0, 1, 2, ..., 7, and 8 heads are given below:

Heads x	Probability $P(X=x)$	Cumulative	Decumulative
0	$\frac{1}{256} = 0.0039$	0.0039	---
1	$\frac{8}{256} = 0.0312$	0.0351	---
2	$\frac{28}{256} = 0.1094$	0.1445	---
3	$\frac{56}{256} = 0.2188$	---	---
4	$\frac{70}{256} = 0.2734$	---	---
5	$\frac{56}{256} = 0.2188$	---	---
6	$\frac{28}{256} = 0.1094$	---	0.1445
7	$\frac{8}{256} = 0.0312$	---	0.0351
8	$\frac{1}{256} = 0.0039$	---	0.0039
Total	$= 1.0000$		

We use the cumulative column and the decumulative column as the critical region is composed of two portions of area (probability), one in each tail of the distribution If $\alpha = 0.05$ then $\alpha/2 = 0.025$ (area in each tail).

We observe that $P(X \leq 1) = 0.0351 > 0.025$, and

$$P(X \geq 7) = 0.0351 > 0.025.$$

Therefore the true significance level is

$$\alpha = P(X \leq 0) + P(X \geq 8) = 0.0039 + 0.0039 = 0.0078.$$

Hence the critical region is $X \leq 0$ and $X \geq 8$.

(v) Computation: $x = 7$

(vi) **Decision.** Since $x = 7$ does not fall in the critical region, so we accept our null hypothesis $H_0 : p = 0.5$, and conclude that the coin is fair.

EXERCISES

16.1 Define the following concepts in your own words as fully as you can:

(i) Hypothesis Testing (ii) Statistical Hypothesis

(iii) Null Hypothesis (iv) Critical Region.

(iv) Level of Significance (P.U., B.A/B.Sc. 1980)

16.2 Explain with examples the difference between

(i) Null Hypothesis and Alternative Hypothesis.

(ii) Simple Hypothesis and Composite Hypothesis.

(iii) Acceptance Region and Rejection Region.

(iv) Type I-Error and Type II-Error.

(v) One-tailed Test and Two-tailed Test.

(P.U., B.A/B.Sc. 1979, 83, 91, 93)

16.3 Explain what is meant by (i) a statistical hypothesis, (ii) test-statistic, (iii) the power of a test, (iv) significance level, (v) test of significance, and (vi) operating characteristic function.

16.4 (a) Distinguish between any two of the following concepts:

(i) Statistical Estimation and Hypothesis Testing.

(ii) Type I and Type II Errors.

(iii) Rejection and Non-rejection Regions.

(iv) A Test at α level of significance and $1-\alpha$ Confidence Interval.

(b) How is the Type I-error related to the Type II error? Are type I and type II errors such that $\alpha + \beta = 1$? (P.U., B.A/B.Sc. 1986-S)

16.5 (a) "A judge can acquit a guilty person or convict an innocent person. An examiner may pass a bad student or may fail a good student." Discuss the relevance of the concept of the two types of errors in these two case. Give other examples also.

(b) The proportion of families buying milk from company A in a certain city is believed to be $p=0.6$. If a random sample of 10 families shows that 3 or less buy milk from company A, we shall reject the hypothesis that $p=0.6$ in favour of the alternative $p<0.6$. Evaluate α if $p=0.6$. Evaluate β for the alternatives $p=0.3$, $p=0.4$ and $p=0.5$. (P.U., B.A/B.Sc. 1993)

16.6 (a) Define Type-I and Type-II errors in testing hypotheses. A normal distribution is known to have a variance of 2.8. A one-tailed (increase) test is proposed of the form $H_0 : \mu \leq 14$ versus $H_1 : \mu > 14$. Find the probability of making a Type-II error (β) with a sample size 2 if the significance level of the test is (i) 0.05, (ii) 0.01, when the true population mean is 16.5. (I.U., M.Sc., 1995)

(b) Given $H_0 : \mu \geq 200$, $H_1 : \mu < 200$, $n = 100$, $\alpha = 0.023$, and $\sigma = 25$.

(i) For what values of the sample mean $\bar{x}$ will H_0 be accepted?

(ii) Compute β if μ is actually 191. (iii) What is the power of the test in (ii)? What does it mean?

16.7 (a) Explain how the null hypothesis and the alternative hypothesis are formulated.

(b) An exercise physiologist wants to demonstrate that the average person walks more than 800 km per year. State the null and alternative hypotheses. What do we use test-statistic?

(c) Describe the general procedure for testing a hypothesis about a population parameter.

16.8 Based on a sample of 25 observations from a normal population with $\sigma = 3$, the hypothesis $H_0 : \mu = 67$ against $H_1 : \mu > 67$ is tested at 5% level of significance. Compute the probabilities of committing type-II errors, β and the powers of the test, when alternative hypotheses of 68.5, 68.0, 67.5 and 66 are used.

16.9 Given $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$ and α and β are probabilities of making type I and type II errors respectively, show that for a one-sided hypothesis test, the required sample size n is given by the expression

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_1 - \mu_0)^2}.$$

Also use this formula to find n when $\sigma = 12$, $\mu_0 = 28$, $\mu_1 = 32$, $\alpha = 0.05$ and $\beta = 0.01$.

16.10 The hypothesis $H_0 : \mu = 100$ is to be tested with $\alpha = 0.05$. The population standard deviation is known to be $\sigma = 10$.

(a) Would a sample of size $n = 100$ result in a value of β less than 0.2 if, in fact, $\mu = 110$?

(b) How large a sample would be required so that $\beta = 0.01$ if, in fact, $\mu = 110$? (P.U., M.Sc., 1972)

16.11 Suppose that $H_0 : \mu = 200$ miles and $H_1 : \mu > 200$ miles. An $\alpha = 0.05$ is required and $\beta = 0.10$ is acceptable when the true mean is 205 miles. Find the optimum sample size. It is estimated that $\sigma = 15$. What decision rule would you establish?

16.12 (a) What statistical hypotheses can be tested by means of the normal distribution?

(b) Past experience has shown that the scores of students who take a certain mathematics test are normally distributed with mean 75 and variance 36. The Mathematics Department members would like to know whether this year's group of 16 students is typical. They decide to test the hypothesis that this year's students are typical against the alternative that they are not typical. When the students take the test, the average score is 82. What conclusion should be drawn?

16.13 (a) What is the difference between a one-sided test and a two-sided test? When should each be used?

(b) A random sample of size 36 is taken from a normal population with a known variance $\sigma^2 = 25$. If the mean of the sample is $\bar{x} = 42.6$, test the null hypothesis $\mu = 45$ against the alternative hypothesis $\mu < 45$ with $\alpha = 0.05$ (α is the probability of committing Type I-error). (P.U., B.A/B.Sc., 1978)

16.14 (a) The heights of college male students are known to be normally distributed with a mean of 67.39 inches and $\sigma = 1.30$ inches. A random sample of 400 students showed a mean height of 67.47 inches. Using a 0.05 significance level, test the hypothesis $H_0 : \mu = 67.39$ against the alternative $H_1 : \mu > 67.39$.

(b) The IQ's of the college students are known to be normally distributed with a mean of 123. A random sample of 49 students showed an average IQ of $\bar{x} = 120.67$ and $S = 8.44$. Test the hypothesis that $\mu = 123$ against the alternative that it is less. Let $\alpha = 0.05$.

16.15 (a) A sample of size 40 from a non-normal population yielded the sample mean $\bar{x} = 71$ and $S^2 = 200$. Test $H_0 : \mu = 72$ against $H_1 : \mu \neq 72$ using a 0.01 significance level.

(b) Suppose that the mean μ of a random variable X is unknown but the variance for X is known to be 144. Should we reject the null hypothesis $H_0 : \mu = 15$ in favour of an alternative hypothesis $H_1 : \mu \neq 15$ at $\alpha = 0.05$, if a random sample of 64 observations yields a mean $\bar{x} = 12$? (P.U., B.A/B.Sc. 1985)

16.16 It is claimed that an automobile is driven on the average more than 20,000 kilometers per year. To test this claim, a random sample of 100 automobile owners are asked to keep a record of the kilometers they travel. Would you agree with the claim if the random sample showed an average of 23,500 kilometers and a standard deviation of 3900 kilometers? Use a 0.01 level of significance. (P.U., B.A/B.Sc. 1979)

16.17 (a) A sample of 900 members has a mean 2.4 inches. Could it be reasonably regarded as being a simple random sample from a large population whose mean is 2.9 inches and standard deviation 3.2 inches?

(b) A sample of size 400 has $\bar{x} = 6.0''$. Can it be regarded as a simple random sample from a large population with mean 6.2" and standard deviation 2.25"?

16.18 (a) A process is in control when the average amount of instant coffee that is packed in a jar, is 6 oz. The standard deviation is 0.2 oz. A sample of 100 jars is selected at random and the sample average is found to be 6.1 oz. Is the process out of control?

(b) Can you reject a claim that the average age of members of Parliament is at least 50, if a random sample of 36 members has a mean age of 48.7 with a standard deviation of 3.1 years. Assume all members' ages are normally distributed; test at the 0.01 level.

16.19 (a) A sample of size 6 from a normal population with variance 24 gave $\bar{x}_1 = 15$. A sample of size 8 from a normal population with variance 80 gave $\bar{x}_2 = 13$. Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$. Let $\alpha = 0.05$.

(b) A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean $\bar{x}_1 = 81$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, has a mean $\bar{x}_2 = 76$. Test the hypothesis at the 0.06 level of significance that $\mu_1 = \mu_2$ against the alternative $\mu_1 \neq \mu_2$.

16.20 The two samples A and B detailed below were taken from normal populations of standard deviation 2.5. Decide whether the difference of sample means is significant at the 0.05 level of significance.

A	16	18	23	26	19	24	25	23	21	22
B	20	21	23	25	25	27	24	26	24	28

16.21 An examination was given to two classes of 40 and 50 students respectively. In the first class, mean grade was 74 with standard deviation of 8, while in the second class the mean grade was 78 with a standard deviation of 7. Is there a significant difference between mean grades (i) at 5% level of significance? (ii) at 1% level of significance? (P.U., B.A/B.Sc. 1992)

16.22 A manufacturer suspects a difference in the quality of the spare parts he receives from two suppliers. He obtains the following data on the service life of random samples of parts from two suppliers.

Supplier	Number in Sample	Means	Standard Deviation
A	50	150	10
B	100	153	5

Test whether the difference between the two sample means is statistically significant at the 1% level of significance.

23 (a) A simple sample of heights of 6400 Englishmen has a mean of 67.85 inches and a standard deviation of 2.56 inches, while a simple sample of heights of 1600 Australians has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that Australians are on the average taller than Englishmen? Use $\alpha = 0.05$. (P.U., B.A/B.Sc. 1996)

(b) A potential buyer of light bulbs bought 50 bulbs of each of 2 brands. Upon testing the bulbs, he found that brand A had a mean life of 1282 hours with a standard deviation of 80 hours, whereas brand B had a mean life of 1208 hours with a standard deviation of 94 hours. Can the buyer be quite certain that the two brands do differ in quality?

24 A random sample of 80 light bulbs manufactured by company A had an average life time of 1258 hours with a standard deviation of 94 hours, while a random sample of 60 light bulbs manufactured by company B had an average lifetime of 1029 hours with a standard deviation of 68 hours. Because of the high cost of bulbs from company A, we are inclined to buy from company B unless the bulbs from company A will last over 200 hours longer on the average than those from company B. Run a test using $\alpha = 0.01$ to determine from whom we should buy our bulbs. (P.U., B.A/B.Sc. 1991)

25 (a) Explain how you test the hypotheses on proportions.

(b) A basketball player has hit on 60% of his shots from the floor. If on the next 100 shots he makes 70 baskets, would you say that his shooting has improved? Use a 0.05 level of significance. (P.U., B.A/B.Sc. 1978, 81)

26 (a) A coin is tossed 900 times and heads appear 490 times. Does this result support the hypothesis that the coin is unbiased?

(b) The sex distribution of 98 births reported in a newspaper was 52 boys and 46 girls. Is this consistent with an equal sex division in the population?

27 In a poll of 10,000 voters selected at random from all the voters in a certain district, it is found that 5,180 voters are in favour of a particular candidate. Test the null hypothesis that the proportion of all the voters in the district, who favour the candidate is equal to or less than 50% against the alternative that it is greater than 50%. Use a 0.05 level of significance.

16.28 (a) In the inspection of a product, it is found that in a random sample of 200 units, 12 are defective. Is this consistent with an average of 5 per cent set as a standard?

(b) A sample of size 78 from a binomial population gave 35 successes. Test the null hypothesis that the true proportion of successes is 0.55 against the alternative that it is less. Let $\alpha = 0.05$.

16.29 (a) The manufacturer of a patent medicine claimed that it was 90% effective in relieving an allergy for a period of 8 hours. In a sample of 200 people who had the allergy, the medicine provided relief for 160 people. Determine whether the manufacturer's claim is legitimate at the $\alpha = 0.01$ level.

(b) It is claimed that 90% of men cannot tell the difference between two different brands of cheese, but of the members of a random sample of 500 men, 72 could distinguish between them. Is the claim justified? (I.U., M.Sc. 1995)

16.30 An electrical company claimed that at least 95% of the parts which they supplied on a government contract conformed to specifications. A sample of 400 parts was tested, and 45 did not meet specifications. Can we accept the company's claim at a 0.05 level of significance?

16.31 (a) A random sample of 150 light bulbs manufactured by a firm X showed 12 defective bulbs while a random sample of 100 light bulbs manufactured by another firm Y showed 4 defective bulbs. Is there a significant difference between the proportions of two firms? (P.U., B.A/B.Sc. 1984)

(b) Random samples of 500 men and 500 women are selected to determine whether the proportions of men and women favouring a political candidate are different. Perform a hypothesis test at 5 percent level if, in the samples, 225 men and 275 women favour the candidate. What is implied by the test result?

16.32 (a) A machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine been improved?

(b) A manufacturer of house-dresses sent out advertising by mail. He sent samples of material to each of 2 groups of 1,000 women. For one group, he enclosed a white return envelope and for the other group, a blue envelope. He received orders from 10% and 13% respectively. Do the data indicate that the colour of the envelope has an effect on the sales? Use 5% level of significance. (P.U., B.A/B.Sc. 1996)

16.33 A civil service examination is given to a group of 200 candidates. On the basis of their total scores, the 200 candidates are divided into two groups, the upper 30 per cent and the remaining 70 per cent. Consider the first question on the examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is no good at discriminating ability of the type being examined here?

Hint. Here $\hat{p}_1 = \frac{40}{60}$, $\hat{p}_2 = \frac{80}{140}$ and $\hat{p}_c = \frac{40 + 80}{60 + 140}$

$$\text{so that } z = \frac{0.67 - 0.57}{\sqrt{(0.6)\,(0.4)\left(\frac{1}{60} + \frac{1}{140}\right)}} = 1.32 < z_{0.025}\,.$$

Accept the hypothesis and conclude that the first question is not satisfactory, etc.,

16.34 The standard deviation of a simple sample of 1,000 members is 5.9 years and that of an independent sample of 900 members is 6.1 years. Show that the samples can be reasonably regarded as drawn from equally variable normal populations. (I.U., M.Sc., 1993)

16.35 A coin is tossed 10 times and comes up heads 8 times. Can we conclude that the coin is fair at a significance level of 0.05.

17

The Chi-Square Distribution and Statistical Inference

17.1 INTRODUCTION

Another distribution that has many important applications in statistical inference, is the χ^2-distribution (χ is the Greek letter *chi*, pronounced *ki* as in kite). The chi-square distribution was first obtained in 1875 by F.R. Helmert (1843–1917), a German Physicist. Later in 1900, Karl Pearson (1857–1936) showed that as n approaches infinity, a discrete multinomial distribution may be transformed and made to approach a chi-square distribution. This approximation has broad applications such as a test of goodness of fit, as a test of independence and as a test of homogeneity. The chi-square distribution contains only one parameter, called the *number of degrees of freedom* (*df*), where the term degree of freedom represents the number of independent random variables that express the chi-square. If the random variables entering a chi-square are subjected to linear restrictions, then the number of degrees of freedom is reduced by the number of restrictions involved. Later we shall find that the number of degrees of freedom is given as the total number of observations in a sample *minus* the number of population parameters that must be estimated from the sample data.

17.2 THE CHI-SQUARE (χ^2) DISTRIBUTION

Let $Z_1, Z_2, \ldots, Z_n$ be normally and independently distributed variables with zero means and unit variances. Then a random variable expressed by the quantity.

$$\chi^2 = \sum_{i=1}^{n} Z_i^2$$

is defined as a chi-square random variable with n degrees of freedom. That is, a χ^2 random variable is defined as the sum of squares of n independent standard normal random variables. Its density function has the following form:

$$f(\chi^2) = \frac{1}{2^{n/2}\,\Gamma\,(n/2)}\,(\chi^2)^{(n/2)-1}\,.\,e^{-\chi^2/2}\,,\quad 0 < \chi^2 < \infty\,.$$

Random variables having the above density function are said to possess the chi-square distribution with n degrees of freedom (denoted by $\chi^2_{(n)}$, where the parameter n, called the degrees of freedom, is a positive integer.

To obtain the distribution of $\chi^2 = \sum_{i=1}^{n} Z_i^2$, we use the moment generating function technique. The *m.g.f.* of the chi-square defined above is

$$\begin{aligned} M_0(t) &= E[e^{t\chi^2}] = E[e^{t\Sigma Z_i^2}] \\ &= E[e^{tZ_1^2}\,.\,e^{tZ_2^2}\ldots e^{tZ_n^2}] \\ &= E\left[\prod_{i=1}^{n} e^{tZ_i^2}\right] = \prod_{i=1}^{n} E[e^{tZ_i^2}] \text{ as } Z_i \text{ are all independent.} \end{aligned}$$

$$\begin{aligned} \text{Now}\quad E[e^{tZ_i^2}] &= \int_{-\infty}^{\infty} e^{tz^2}\,.\,\frac{1}{\sqrt{2\pi}}\,e^{-z^2/2}\,dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\,e^{-(1-2t)z^2/2}\,dz \\ &= \frac{1}{\sqrt{1-2t}}\int_{-\infty}^{\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}}\,e^{-(1-2t)z^2/2}\,dz \end{aligned}$$

But for $t < \frac{1}{2}$, the integral on the right is equal to 1 as it represents the total probability under a normal curve with mean zero and variance $\frac{1}{1-2t}$. Therefore,

$$E[e^{tZ_i^2}] = \frac{1}{\sqrt{1-2t}}\,,\quad \text{for } t < \frac{1}{2}$$

Hence $M_0(t) = \prod_{i=1}^{n} E(e^{tZ_i^2})$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{1-2t}} = \left(\frac{1}{1-2t}\right)^{n/2} \text{ for } t < \frac{1}{2}.$$

We know that the *m.g.f.* of the Gamma (Pearson Type III) distribution, viz.,

$$f(x) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax}, \text{ is } M_0(t) = \left(\frac{1}{1-t/a}\right)^p.$$

Comparing, we find that $p = \frac{n}{2}$ and $a = \frac{1}{2}$. Thus the *p.d.f.* of χ^2 is

$$f(\chi^2) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} (\chi^2)^{(n/2)-1} . e^{-\chi^2/2}$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} (\chi^2)^{(n/2)-1} . e^{-\chi^2/2}, \quad 0 < \chi^2 < \infty$$

If $n = 1$, then $\chi^2 = Z^2$ so that the square of a standard normal random variable is distributed as χ^2 with one degree of freedom. Moreover, the quantity $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ is a χ^2 random variable with one degree of freedom as $\bar{X}$ is normally distributed with mean μ and variance σ^2/n.

17.2.1. Properties of the Chi-Square Distribution. The chi-square distribution has the following properties:

(i) The chi-square is a continuous distribution ranging from zero to plus infinity, *i.e.* $0 < \chi^2 < \infty$.

(ii) The mean of a chi-square distribution is equal to the number of degrees of freedom and its variance is equal to twice the number of degrees of freedom. That is, $E[\chi^2_{(n)}] = n$ and $Var[\chi^2_{(n)}] = 2n$.

(iii) The *moments* of the χ^2-distribution about the origin are found as below:

The *m.g.f.* is given by

$$M_0(t) = \int_0^\infty e^{t\chi^2} \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)} (\chi^2)^{(n/2)-1} . e^{-\chi^2/2}\, d(\chi^2)$$

$$= (1-2t)^{-n/2}, \text{ for } t < \frac{1}{2}.$$

Expanding $M_0(t)$ in power series, we get,

$$M_0(t) = 1 + \frac{n}{2}(2t) + \frac{\left(\frac{n}{2}\right)\left(\frac{n}{2}+1\right)}{2!}(2t)^2 + \dots + \frac{\left(\frac{n}{2}\right)\left(\frac{n}{2}+1\right)\dots\left(\frac{n}{2}+\overline{r-1}\right)}{r!}(2t)^r + \dots$$

Thus $\mu_r' =$ Co-efficient of $\frac{t^r}{r!}$ in the expansion of $M_0(t)$

$$= 2^r\left(\frac{n}{2}\right)\left(\frac{n}{2}+1\right)\left(\frac{n}{2}+2\right)\dots\left(\frac{n}{2}+\overline{r-1}\right)$$

Putting $r = 1, 2, 3$ and 4, we obtain

$$\mu_1' = n,\ \mu_2' = n(n+2),$$

$$\mu_3' = n(n+2)(n+4),$$

$$\mu_4' = n(n+2)(n+4)(n+6).$$

Thus $E(\chi^2) = \mu_1' = n$, $\text{Var}(\chi^2) = \mu_2' - (\mu_1')^2 = 2n$, $\mu_3 = 8n$ and $\mu_4 = 12n^2 + 48n$.

Furthermore, $\beta_1 = \frac{8}{n}$ and $\beta_2 = 3 + \frac{12}{n}$. The cumulants of the distribution also exist for all orders.

(iv) The curve of a χ^2-distribution is positively skewed. The skewness decreases as n increases. For example, when $n = 1$,

$$f(\chi^2) = \frac{1}{\sqrt{2\pi\chi^2}}\, e^{-\chi^2/2},$$

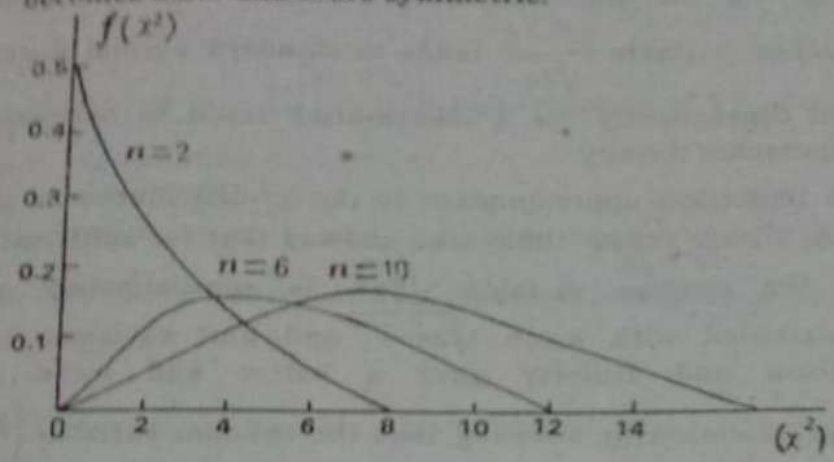
the curve is extremely J-shaped and the skewness is the highest. When $n = 2$, $f(\chi^2) = \frac{1}{2} \cdot e^{-\chi^2/2}$.

The curve becomes a steadily decreasing exponential curve with initial point $(0, \frac{1}{2})$.

When $n = 3$, $f(\chi^2) = \frac{1}{\sqrt{2\pi}}(\chi^2)^{(1/2)}\, e^{-\chi^2/2}$,

The curve originates at the point (0, 0) rises to a maximum at $\chi^2 = n-2$ *i.e.*, 1, then decreases with the X-axis as asymptote, and the curve is unimodal. For $n > 3$, the distribution takes a

curve similar to that for $n=3$. The mode of a chi-square distribution is the value equal to $n-2$. As n increases, the distribution becomes more and more symmetric.



χ^2-distribution for various values of n

v) The χ^2-distribution tends to normal distribution as the number of degrees of freedom approaches infinity.

The *m.g.f.* of $\chi^2_{(n)}$ is $M_0(t) = (1-2t)^{-n/2}$

Let us consider the chi-square standard variable $\dfrac{\chi^2 - n}{\sqrt{2n}}$.

Then its *m.g.f.* about mean is

$$M(t) = e^{-nt/\sqrt{2n}}\, M_0\left(\frac{t}{\sigma}\right) = e^{-nt/\sqrt{2n}}\left(1 - \frac{2t}{\sqrt{2n}}\right)^{-n/2}$$

Taking natural logs, we have

$$ln\ M(t) = -\frac{nt}{\sqrt{2n}} - \frac{n}{2}\, ln\left[1 - \frac{2t}{\sqrt{2n}}\right].$$

$$= \frac{-nt}{\sqrt{2n}} + \frac{n}{2}\left[\frac{2t}{\sqrt{2n}} + \frac{1}{2}\frac{4t^2}{2n} + \text{higher powers of } \frac{1}{n}\right]$$

$$= \frac{-nt}{\sqrt{2n}} + \frac{nt}{\sqrt{2n}} + \frac{t^2}{2} + \text{higher powers of } \frac{1}{n}$$

$$= \frac{t^2}{2} + \text{higher powers of } \frac{1}{n}$$

Thus as $n \to \infty$, $\ln M(t) \to \frac{t^2}{2}$ so that $M(t) \to e^{t^2/2}$, which is the *m.g.f.* of the standard normal random variable. Hence the random veriable $\frac{\chi^2 - n}{\sqrt{2n}}$ tends to standard normal distribution and consequently the χ^2-distribution tends to normality as n approaches infinity.

(vi) An important approximation to the χ^2-distribution is given by R.A. Fisher (1890–1962) who showed that for sufficiently large n, the random variable $\sqrt{2\chi^2}$ is approximately normally distributed with mean $\sqrt{2n-1}$ and unit variance. In 1931, Wilson and Hilferty gave a better and more accurate approximation by showing that the random variable $\left(\frac{\chi^2}{n}\right)^{1/3}$ is approximately normally distributed with mean $1 - \frac{2}{9n}$ and variance $\frac{2}{9n}$.

(vii) **Additive Property.** If X and Y are independent χ^2-random variables with n_1 and n_2 degrees of freedom respectively, then the sum $X + Y$ is a χ^2-random variable with $n_1 + n_2$ degrees of freedom.

$$\begin{aligned} \text{The } m.g.f. \text{ of } X + Y &= [m.g.f. \text{ of } \chi^2_{(n_1)}]\,[m.g.f. \text{ of } \chi^2_{(n_2)}] \\ &= (1-2t)^{-n_1/2} \cdot (1-2t)^{-n_2/2} \\ &= (1-2t)^{-(n_1+n_2)/2} \end{aligned}$$

But this is the *m.g.f.* for a χ^2 random variable with $n_1 + n_2$ degrees of freedom. Thus the sum of two independent chi-squares is itself a chi-square. This property can be extended to any number of χ^2 random variables.

(viii) **Partitioning Property.** A χ^2 random variable can be partitioned into two or more than two parts which are also χ^2 random variables and the sum of their degrees of freedom equals the total degrees of freedom.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$. Then the quantity $\Sigma(X_i - \mu)^2$ may be expressed by the following simple identity:

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \bar{X}) + n(\bar{X} - \mu)^2$$

Dividing all the terms in the above identity by σ^2, we get

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 + n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2$$

It is obvious that $\sum\left(\frac{X_i-\mu}{\sigma}\right)^2$ is distributed as $\chi^2_{(n)}$ being sum of n standard normal random variables.

The term $n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2$ may be written as $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ which is a $\chi^2_{(1)}$ random variable as the square of a standard normal random variable has a $\chi^2_{(1)}$ distribution.

Furthermore, it has also been shown that $\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 = \frac{nS^2}{\sigma^2}$ follows the χ^2-distribution with $(n-1)$ degrees of freedom and that its distribution is independent of $\bar{X}$. Thus we have the identity

$$\chi^2_{(n)} = \chi^2_{(n-1)} + \chi^2_{(1)}.$$

This property forms the basis of the Analysis of Variance (Chapter 20) as this is the distribution of the sums of squares.

17.2.2. The χ^2-table. The areas for the χ^2-distribution have been tabulated for various values of α and n, the degrees of freedom. Table 17.1 on page 176 contains values of $\chi^2_{\alpha,(n)}$, which denotes the values for which the area *to its right* under the chi-square distribution with $n(n=1, 2, \ldots, 30)$ degrees of freedom is equal to α.

Example 17.1 Compute $\chi^2_{0.05}$ for 30 and 50 degrees of freedom by (i) Fisher's approximation, (ii) Wilson-Hilferty approximation, and compare them with the tabulated values.

We can use the facts that

(i) $Z = \sqrt{2\chi^2} - \sqrt{2n-1}$ (Fisher's approximation)

(ii) $Z = \dfrac{(\chi^2/n)^{1/3} - \left(1 - \frac{2}{9n}\right)}{\frac{2}{9n}}$, (Wilson-Hilferty approximation)

are approximately normally distributed with zero mean and unit variance, to compute the values of χ^2_α for n degrees of freedom $(n \geq 30)$ from the tables of the standard normal distribution.

Table 17.1 Distribution of χ^2

The entries in this table are values of $\chi^2_{\alpha,(n)}$ for which the area to their right under the chi-square distribution with n degree of freedom is equal to α.

n	0.99	0.98	0.975	0.95	0.10	0.05	0.025	0.02	0.01
1	0.0002	0.001	0.001	0.004	2.71	3.84	5.02	5.41	6.64
2	.020	.040	0.051	.103	4.61	5.99	7.38	7.82	9.21
3	.115	.185	0.216	.352	6.25	7.82	9.35	9.84	11.34
4	.297	.429	0.484	.711	7.78	9.49	11.14	11.67	13.28
5	.554	.752	0.831	1.145	9.24	11.07	12.83	13.39	15.09
6	.87	1.13	1.24	1.64	10.64	12.59	14.45	15.03	16.81
7	1.24	1.56	1.69	2.17	12.02	14.07	16.01	16.62	18.48
8	1.65	2.03	2.18	2.73	13.36	15.51	17.54	18.17	20.09
9	2.09	2.53	2.70	3.32	14.68	16.92	19.02	19.68	21.67
10	2.56	3.06	3.25	3.94	15.99	18.31	20.48	21.16	23.21
11	3.05	3.61	3.82	4.58	17.28	19.68	21.92	22.62	24.72
12	3.57	4.18	4.40	5.23	18.55	21.03	23.34	24.05	26.22
13	4.11	4.76	5.01	5.89	19.81	22.36	24.74	25.47	27.69
14	4.66	5.37	5.63	6.57	21.06	23.68	26.12	26.87	29.14
15	5.23	5.98	6.26	7.26	22.31	25.00	27.49	28.26	30.58
16	5.81	6.61	6.91	7.96	23.54	26.30	28.84	29.63	32.00
17	6.41	7.26	7.56	8.67	24.77	27.59	30.19	31.00	33.41
18	7.02	7.91	8.23	9.39	25.99	28.87	31.53	32.35	34.81
19	7.63	8.57	8.91	10.12	27.20	30.14	32.85	33.69	36.19
20	8.26	9.24	9.59	10.85	28.41	31.41	34.17	35.02	37.57
21	8.90	9.92	10.28	11.59	29.62	32.67	35.48	36.34	38.93
22	9.54	10.60	10.98	12.34	30.81	33.92	36.78	37.66	40.29
23	10.20	11.29	11.69	13.09	32.01	35.17	38.08	38.97	41.64
24	10.86	11.99	12.40	13.85	33.00	36.42	39.36	40.27	42.92
25	11.52	12.70	13.12	14.61	34.38	37.65	40.65	41.57	44.31
26	12.20	13.41	13.84	15.38	35.56	38.88	41.92	42.86	45.64
27	12.88	14.12	14.57	16.15	36.74	40.11	43.19	44.14	46.96
28	13.56	14.85	15.31	16.93	37.92	41.34	44.46	45.42	48.28
29	14.26	15.57	16.05	17.71	39.09	42.56	45.72	46.69	49.59
30	14.95	16.31	16.79	18.49	40.26	43.77	46.98	47.96	50.89

For $n > 30$, the expression $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be used as a normal variable with zero mean and unit variance, remembering that the probability for χ^2 corresponds with that of a single tail of the normal curve.

"Table 17.1 (except columns 4 and 8) is taken from Table IV of Fisher and Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, published by Oliver & Boyd, Ltd.; Edinburgh, and reproduced by permission of the authors and publishers."

From Fisher's approximation, we get

$$\chi_\alpha^2 = \frac{1}{2}[z_\alpha + \sqrt{2n-1}]^2$$

and from Wilson-Hilferty approximation, we obtain

$$\chi_\alpha^2 = n\left[\left(1 - \frac{2}{9n}\right) + z_\alpha\sqrt{\frac{2}{9n}}\right]^3$$

where z_α is the value of standard normal random variable corresponding to the level α of significance.

Substituting $\alpha = 0.05$, $n = 30$ in Fisher's approximation, we get

$$\chi^2_{0.05,(30)} = \frac{1}{2}[1.645 + \sqrt{2(30)-1}]^2 \qquad (\because\ z_{0.05} = 1.645)$$

$$= \frac{1}{2}[1.645 + 7.681]^2 = \frac{1}{2}[86.974] = 43.49$$

For $n = 50$ and $\alpha = 0.05$,

$$\chi^2_{0.05,(50)} = \frac{1}{2}[1.645 + \sqrt{2(50)-1}]^2$$

$$= \frac{1}{2}[1.645 + 9.950]^2$$

$$= \frac{1}{2}[134.44] = 67.22.$$

Using Wilson-Hilferty approximation, we get for $n = 30$ and $\alpha = 0.05$,

$$\chi^2_{0.05,(30)} = 30\left[1 - \frac{2}{9(30)} + 1.645\sqrt{\frac{2}{9(30)}}\right]^3$$

$$= 30\,[1 - 0.0074 + 1.645(0.086)]^3$$

$$= 30\,[1.134]^3 = 30\,(1.458) = 43.74;$$

and for $n = 50$, we have

$$\chi^2_{0.05,(50)} = 50\left[1 - \frac{2}{9(50)} + 1.645\sqrt{\frac{2}{9(50)}}\right]^3$$

$$= 50\,[1 - 0.0044 + 1.645(0.067)]^3$$

$$= 50\,[1.1058]^3 = 67.6.$$

These values agree very well with the tabulated values of $\chi^2_{0.05,(30)} = 43.8$ and $\chi^2_{0.05,(50)} = 67.5$.

17.3 CONFIDENCE INTERVAL ESTIMATE OF VARIANCE OF A NORMAL POPULATION

The confidence interval estimate of the population variance σ^2 is based on the sampling distribution of S^2, the sample variance, and the sampling distribution of S^2 is the chi-square distribution. We therefore use the χ^2-distribution to obtain confidence interval estimate for σ^2 when we are given (i) one sample variance, (ii) several sample variances.

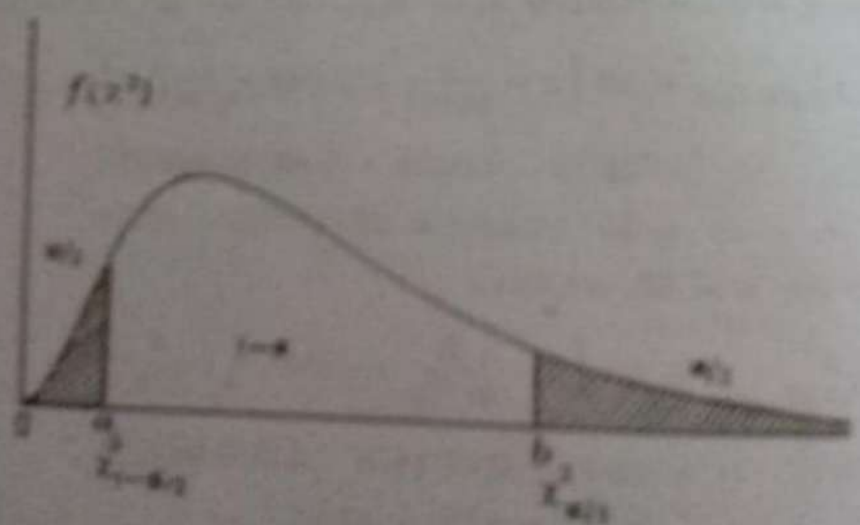
17.3.1. Confidence Interval Estimate of σ^2 from a Sample Variance. Let $\bar{X}$ and S^2 be the mean and variance of a random sample $X_1, X_2, \ldots, X_n$ of size n drawn from a normal population with variance σ^2. Then the statistic

$$\frac{nS^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \quad \left[\text{or } \frac{(n-1)\,s^2}{\sigma^2} \text{ where } s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right]$$

that is the ratio of the sum of squared deviations from the sample mean to the population variance, has a chi-square distribution with $(n-1)$ degrees of freedom.

To construct a two-sided confidence interval for σ^2, we find two values of χ^2-distribution with $(n-1)$ degrees of freedom, say a and b, such that

$$\int_0^a f(\chi^2)\,d(\chi^2) = \frac{\alpha}{2} \text{ and } \int_b^\infty f(\chi^2)\,d(\chi^2) = \frac{\alpha}{2}$$



We then have an interval with an associated probability $1-\alpha$, that contains the variance σ^2, as

$$P\left[a < \frac{nS^2}{\sigma^2} < b\right] = 1 - \alpha.$$

To put σ^2 inside inequalities within the brackets, we proceed as below:

(i) We divide all terms inside the brackets by nS^2 and get

$$\frac{a}{nS^2} < \frac{1}{\sigma^2} < \frac{b}{nS^2}$$

(ii) We replace each term with its reciprocal (remember, we inverse the direction of the inequality signs when we replace a term by its reciprocal) and obtain

$$\frac{nS^2}{a} > \sigma^2 > \frac{nS^2}{b}$$

which is equivalent to

$$\frac{nS^2}{b} < \sigma^2 < \frac{nS^2}{a}$$

(iii) We substitute this result in the probability statement and get

$$P\left[\frac{nS^2}{b} < \sigma^2 < \frac{nS^2}{a}\right] = 1 - \alpha$$

or $$P\left[\frac{\Sigma(X_i - \bar{X})^2}{b} < \sigma^2 < \frac{\Sigma(X_i - \bar{X})^2}{a}\right] = 1 - \alpha.$$

The values of a and b are obtained from the tables of the chi-square distribution having $(n-1)$ degrees of freedom, which are found as

$$a = \chi^2_{1-\alpha/2} \text{ and } b = \chi^2_{\alpha/2}.$$

Hence for a particular sample of size n, the $100(1-\alpha)$ percent confidence interval for σ^2 is given by

$$\frac{\Sigma(X_i - \bar{X})^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{\Sigma(X_i - \bar{X})^2}{\chi^2_{1-\alpha/2}}\Bigg],$$

where $\chi^2_{1-\alpha/2}$ and $b = \chi^2_{\alpha/2}$ are the values of a χ^2-distribution having $(n-1)$ degrees of freedom leaving areas of $1-\alpha/2$ and $\alpha/2$ respectively to the right. Thus a 95% confidence interval for σ^2 would be

$$\frac{(n-1)\,s^2}{\chi^2_{0.025}} < \sigma^2 < \frac{(n-1)\,s^2}{\chi^2_{0.975}}$$

We can obtain a confidence interval for σ by taking the square root of the end points of the interval for σ^2, but experience has shown that σ cannot be estimated with much precision for small sample sizes.

17.3.2. Confidence Interval of σ^2 from Several Sample Variances. Suppose we are given k random samples from different populations having equal variances σ^2 or from the same population with variance σ^2. Let $S_1^2, S_2^2, \ldots, S_k^2$ be the sample variances based on random samples of sizes $n_1, n_2, \ldots, n_k$. The pooled unbiased estimate of σ^2 is then given by

$$s_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2 + \ldots + n_k S_k^2}{n_1 + n_2 + \ldots + n_k - k}.$$

When the population or populations are normal, the sampling distribution of $\dfrac{\sum n_i S_i^2}{\sigma^2}$ or $\dfrac{(\sum n_i - k)\, s_p^2}{\sigma^2}$ would be a chi-square distribution with $(\sum n_i - k)$ degrees of freedom.

Hence the $(1-\alpha)100\%$ confidence interval for σ^2 is given by

$$\frac{\sum n_i S_i^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{\sum n_i S_i^2}{\chi^2_{1-\alpha/2}}$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are the values of a chi-square distribution having $(\sum n_i - k)$ degrees of freedom.

Example 17.2. A random sample of size $n=8$ from a normal population gave the values 9, 14, 10, 12, 7, 13, 11, 12. Find the 90 per cent confidence interval for σ^2.

First we calculate the sample mean $\bar{x}$, which is

$$\bar{x} = \frac{\sum x}{n} = \frac{88}{8} = 11$$

Then $\sum_{i=1}^{8} (X_i - \bar{X})^2 = (9-11)^2 + (14-11)^2 + \ldots + (12-11)^2 = 36$

From the χ^2-table, we find that χ^2 for 7 *d.f.* at $\alpha = 0.10$ level of significance, *i.e.* $\chi^2_{0.95,(7)} = 14.07$, and $\chi^2_{0.05,(7)} = 2.17$.

Hence the 90 per cent confidence interval for σ^2 is

$$\frac{\Sigma(X_i - \bar{X})^2}{\chi^2_{0.05,(7)}} < \sigma^2 < \frac{\Sigma(X_i - \bar{X})}{\chi^2_{0.95,(7)}}$$

or $$\frac{36}{14.07} < \sigma^2 < \frac{36}{2.17}$$

or $$2.56 < \sigma^2 < 16.61$$

Thus the 90% confidence interval for σ^2 is (2.56, 16.61).

17.4 TESTS BASED ON CHI-SQUARE DISTRIBUTION

Some of the most frequently used tests of hypotheses, that are based on χ^2-distribution, are presented in the sections that follow.

17.4.1. Testing Hypothesis about Variance of a Normal Population. Suppose we desire to test a null hypothesis H_0 that the variance σ^2 of a normally distributed population has some specified value, say σ_0^2. To do this, we need to draw a random sample $X_1, X_2, \ldots, X_n$ of size n from the normal population and compute the value of the sample variance S^2. If the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ is true, then the statistic $\chi^2 = \dfrac{nS^2}{\sigma_0^2}$ has a χ^2-distribution with $(n-1)$ degrees of freedom.

The format for the test would be as below:

(i) Formulate the null and alternative hypotheses about σ^2. Three possible forms are

$$H_0 : \sigma^2 = \sigma_0^2 \text{ and } H_1 : \sigma^2 \neq \sigma_0^2$$

$$H_0 : \sigma^2 \leq \sigma_0^2 \text{ and } H_1 : \sigma^2 > \sigma_0^2$$

$$H_0 : \sigma^2 \geq \sigma_0^2 \text{ and } H_1 : \sigma^2 < \sigma_0^2$$

(ii) Decide on significance level α. The commonly used values are at $\alpha = 0.05$ or $\alpha = 0.01$.

(iii) The test-statistic to be used is

$$\chi^2_{(n-1)} = \frac{nS^2}{\sigma_0^2} = \frac{\Sigma(X - \bar{X})^2}{\sigma_0^2}$$

which under H_0, has a chi-square distribution with $(n-1)$ degrees of freedom.

(iv) Determine the critical region which depends on α and the alternative hypothesis H_1.

(a) When H_1 is $\sigma^2 \neq \sigma_0^2$, the critical region is

$\chi^2 < \chi^2_{1-\alpha/2,(n-1)}$ and $\chi^2 > \chi^2_{\alpha/2,(n-1)}$ (Two-tailed test)

(b) When H_1 is $\sigma^2 > \sigma_0^2$, the critical region falls in the right tail and its value is $\chi^2 > \chi^2_{\alpha,(n-1)}$. (One-tailed test)

(c) When H_1 is $\sigma^2 < \sigma_0^2$, the critical region will be entirely in the left tail with the critical value $\chi^2_{1-\alpha,(n-1)}$. (One-tailed test)

(v) Compute the value $\chi^2 = \dfrac{nS^2}{\sigma_0^2}$ from the given data.

(vi) Decide as below:

Reject H_0 if the calculated value of χ^2 falls in the critical region, otherwise accept it.

It is of importance to note that, for small sample sizes, the standard deviation of a normal population can be tested in the same way as the population variance is tested.

Example 17.3. The variability in the tensile strength of a type of steel wire must be controlled carefully. A sample of the wire is subjected to test and it is found that sample variance is $S^2 = 31.5$. The sample size was $n = 16$ observations.

(a) Test the hypothesis that the population variance is 25 against the alternative that the variance is greater than 25. Use a 0.05 level of significance.

(b) Construct the 95 per cent confidence limits on the variance in the tensile strength of the wire. (I.U., M.Sc., 1991)

(a) (i) We have to decide between the hypotheses

$$H_0 : \sigma^2 = 25, \text{ and } H_1 : \sigma^2 > 25$$

(ii) The level of significance is $\alpha = 0.05$.

(iii) The test statistic is $\chi^2 = \dfrac{nS^2}{\sigma_0^2}$, which under H_0, has a χ^2-distribution with $(n-1)$ degrees of freedom, assuming that the population is normal.

(iv) The critical region is $\chi^2 > \chi^2_{0.05,(15)} = 25.0$ (one-tailed test)

(v) We calculate the value of χ^2 from the sample data as

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{16(31.5)}{25} = 20.16.$$

(vi) **Conclusion**. Since the calculated value of χ^2 falls in the acceptance region, so we accept our null hypothesis, *i.e.* we have reasonable evidence to conclude that $\sigma^2 = 25$.

(b) The 95 per cent confidence interval for $\sigma^2 = 25$ is given by

$$\frac{nS^2}{\chi^2_{0.025,(15)}} < \sigma^2 < \frac{nS^2}{\chi^2_{0.975,(15)}}$$

Substituting the values, we get

$$\frac{16\,(31.5)}{27.49} < \sigma^2 < \frac{16\,(31.5)}{6.26}$$

$$\text{or } 18.3 < \sigma^2 < 80.5$$

Hence the desired confidence limits on the population variance are (18.3, 80.5).

Example 17.4. Given that X_i are normally distributed and given the sample values $\bar{x} = 42$, $S = 5$ and $n = 20$. Test the hypothesis that $\sigma = 8$.

(i) Our null hypothesis is $H_0 : \sigma = 8$. Let the alternative hypothesis be $H_1 : \sigma \neq 8$. (Two-tailed Test)

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic is $\frac{nS^2}{\sigma_0^2} = \chi^2_{(n-1)}$, since X_i's are normally distributed, and this statistic under null hypothesis has a chi-square distribution with $(n-1)$ degrees of freedom.

(iv) The critical region is $\chi^2 > \chi^2_{0.025,(19)} = 32.85$ and $\chi^2 < \chi^2_{0.975,(19)} = 8.91$.

(v) Now we compute the value of χ^2 from the given data as

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{20\,(5)^2}{(8)^2} = \frac{20 \times 25}{64} = 7.81$$

(vi) **Conclusion.** Since the computed value of χ^2 falls in the critical region, we therefore reject the hypothesis, and conclude that there is no evidence to accept the hypothesis that $\sigma = 8$.

Example 17.5. The manager of a bottling plant is anxious to reduce the variability in net weight of fruit bottled. Over a long period, the standard deviation has been 15.2 gm. A new machine is introduced and the net weights (in grams) in randomly selected bottles (all of the same nominal weight) are 987, 966, 955, 977, 981, 967, 975, 980, 953, 972. Would you report to the manager that the new machine has a better performance? (M.Sc., P.U., 1989, I.U., 1993, 96)

(i) We have to decide between the hypotheses

$H_0 : \sigma = 15.2$, *i.e.* the standard deviation is 15.2 gm

$H_1 : \sigma < 15.2$ *i.e.* the standard deviation has been reduced.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic is

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{\Sigma (X_i - \bar{X})^2}{\sigma_0^2},$$

which under H_0, has a χ^2-distribution with $(n-1)$ degrees of freedom, assuming that the weights are normally distributed.

(iv) The critical region is $\chi^2 \le \chi^2_{0.95,(9)} = 3.32$ (the lower 5% point)

(v) Computations. $n = 10$, $\Sigma X_i = 9713$, $\Sigma X_i^2 = 9435347$

Now $nS^2 = \Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - (\Sigma X_i)^2/n$

$$= 9435347 - (9713)^2/10 = 1110.1$$

$$\therefore \quad \chi^2 = \frac{1110.1}{(15.2)^2} = \frac{1110.1}{231.04} = 4.81$$

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 4.81$ does not fall in the critical region, we therefore cannot reject the null hypothesis that the standard deviation is 15.2 gm and hence we would not report to the manager that the new machine has a better performance.

17.4.2. Testing Hypothesis about the Equality of Variances of k (k > 2) Normal Populations. Suppose we wish to test the null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \quad (=\sigma^2, \text{ say})$$

against the alternative

$$H_1 : \text{Not all the variances are equal.}$$

For this purpose, several test procedures have been devised. The test procedure due to *M.S.* Bartlett is presented here. This test procedure is based on a statistic whose sampling distribution is approximately a chi-square distribution with $(k-1)$ degrees of freedom, when k random samples of sizes $n_1, n_2, \ldots, n_k$ $(\sum n_i = n)$ are drawn from independent normal populations with variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$. Let $s_1^2, s_2^2, \ldots, s_k^2$ be the unbiased estimates of the populations variances, computed from the k samples, and let these estimates be combined to give the pooled unbiased estimate of variance, if H_0 is true, as

$$s_p^2 = \frac{\sum_{i=1}^{k} (n_i - 1)\, s_i^2}{\sum_{i=1}^{k} (n_i - 1)} = \frac{\sum (n_i - 1)\, s_i^2}{n - k}, \text{ where } s_i^2 = \frac{\sum_{i=1}^{n_i} (x_i - \bar{x}_i)^2}{n_i - 1},$$

Then the test statistic is given by

$$u = 2.3026\, \frac{q}{c},$$

where $q = (n-k) \log s_p^2 - \sum_{i=1}^{k} (n_i - 1) \log s_i^2$, and

$$c = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n-k} \right).$$

The statistic u under H_0, has approximately a chi-square distribution with $(k-1)$ degrees of freedom. The value of q is large when the sample variances s_i^2 differ greatly and is equal to zero when all s_i^2 are equal. Therefore, we reject H_0 when the calculated value of u is greater than or equal to the tabulated value of χ^2 at the desired level of significance for $(k-1)$ degrees of freedom. The rest of the procedure for testing the hypothesis is the same.

This test is generally known as **Bartlett's test** for homogeneity of variances.

Example 17.6. Suppose that four random samples of sizes $n_1 = 11$, $n_2 = 9$, $n_3 = 12$, $n_4 = 15$ are selected from four normal populations and gave $s_1^2 = 392$, $s_2^2 = 427$, $s_3^2 = 620$, $s_4^2 = 667$. Test the hypothesis of equal variances. (I.U., M.Sc., 1995)

(i) The hypotheses would be stated as

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2, \text{ and}$$

$$H_1 : \text{Not all the variances are equal.}$$

(ii) We use a level of significance of $\alpha = 0.05$, and one-sided test.

(iii) The test-statistic would be

$$u = 2.3026 \frac{q}{c},$$

where $q = (n-k) \log s_p^2 - \sum_{i=1}^{k} (n_i - 1) \log s_i^2$, and

$$c = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n-k} \right).$$

The statistic u under H_0, has approximately a chi-square distribution with $(k-1)$ degrees of freedom.

(iv) The critical region is $u \geq \chi^2_{0.05,(3)} = 7.81$

(v) Computations for Bartlett's test for homogeneity of variances.

Sample	s_i^2	$n_i - 1$	$1/(n_i - 1)$	$(n_i - 1)\, s_i^2$	$\log s_i^2$	$(n_i - 1) \log s_i^2$
1	392	10	0.1000	3920	2.5933	25.9330
2	427	8	0.1250	3416	2.6304	21.0432
3	620	11	0.0909	6820	2.7924	30.7164
4	667	14	0.0714	9338	2.8241	39.5374
Σ	---	43	0.3873	23494	--	117.2300

The pooled unbiased estimate of variance is

$$s_p^2 = \frac{\Sigma (n_i - 1)\, s_i^2}{\Sigma (n_i - 1)} = \frac{23494}{43} = 546.3721.$$

Now $q = (n-k)\log s_p^2 - \sum_{i=1}^{k}(n_i - 1)\log s_i^2$

$= (43)(\log 546.3721) - 117.2300$

$= (43)(2.73746) - 117.2300 = 117.71078 - 117.2300$

$= 0.48078$, and

$$c = 1 + \frac{1}{3(k-1)}\left[\sum \frac{1}{n_i - 1} - \frac{1}{n-k}\right]$$

$$= 1 + \frac{1}{9}(0.3873 - 0.02326) = 1 + 0.0404 = 1.0404$$

$$\therefore \quad u = 2.3026\left(\frac{0.48078}{1.0404}\right) = \frac{1.10704}{1.0404} = 1.06$$

(vi) **Conclusion.** Since the computed value of u does not fall in the rejection region, we therefore cannot reject H_0 and may conclude that the variances are homogeneous.

17.5 KARL PEARSON'S APPROXIMATION

Karl Pearson (1857–1936) has established a relationship between the discrete multinomial distribution and the chi-squre distribution by transforming and making the multinomial distribution approach a χ^2-distribution as n approaches infinity. This approximation is widely used to test agreement between the observed data and the expected (or hypothesized) results.

Suppose a random sample of n observations is distributed over k mutually exclusive and exhaustive classes or cells. Let $p_i (i = 1, 2, \ldots, k)$ be the probability that an observation falls in the ith class or cell and n_i be the number of observations falling in that class such that $\sum p_i = 1$ and $\sum n_i = n$; that is, the data have the following multnomial structure:

Class/Cell	1	2	...	k	Total
Probabilities	p_1	p_2	...	p_k	1
Frequencies	n_1	n_2	...	n_k	n

Then the probability function $f(n_1, n_2, \ldots n_k)$ of n_1 belonging to the first cell or class, n_2 belonging to the second cell or class, ... and n_k belonging to the kth cell or class, is given by

$$f(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!}\, p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Assuming that n and all n_i $(i = 1, 2, \ldots, k)$ are sufficiently large, we apply *stirling's* approximation to $n!$ and obtain

$$f(n_1, n_2, \ldots, n_k) = \frac{(2\pi)^{1/2}\, n^{n+1/2}\, e^{-n}\, p_1^{n_1}\, p_2^{n_2} \cdots p_k^{n_k}}{(2\pi)^{k/2}\, n_1^{n_1+1/2} \cdots n_k^{n_k+1/2}\, e^{-(n_1+n_2+\ldots+n_k)}}$$

$$= \frac{\left(\frac{np_1}{n_1}\right)^{n_1+1/2} \cdots \left(\frac{np_k}{n_k}\right)^{n_k+1/2}}{(2\pi n)^{(k-1)/2}\,(p_1 p_2 \cdots p_k)^{1/2}}$$

$$= \frac{\prod_{i=1}^{k}\left(\frac{np_i}{n_i}\right)^{n_i+1/2}}{(2\pi n)^{(k-1)/2}\prod_{i=1}^{k} p_i^{1/2}}$$

Now $E(n_i) = np_i$ meaning that expected cell frequency = n (cell probability) and $\mathrm{Var}(n_i) = np_i\,(1 - p_i)$ for $i = 1, 2, \ldots, k$.

The standard normal variable corresponding to n_i is

$$Z_i = \frac{n_i - np_i}{\sqrt{np_i\,(1-p_i)}} = \frac{n_i - np_i}{\sigma_i},$$

so that $np_i = n_i - Z_i\sigma_i$ and $n_i = Z_i\sigma_i + np_i$

Substituting these transformations in $f(n_1, n_2, \ldots, n_k)$, *i.e.*, replacing n_i by its equivalent, we get

$$f(n_1, n_2, \ldots, n_k) = \frac{\prod_{i=1}^{k}\left(1 + \frac{Z_i\sigma_i}{np_i}\right)^{-Z_i\sigma_i - np_i - 1/2}}{(2\pi n)^{(k-1)/2}\,(p_1 p_2 \cdots p_k)^{1/2}}$$

$$= \frac{M}{C}, \text{ say.}$$

Taking natural logarithm (ln) of M, we get

$$ln\, M = \sum_{i=1}^{k}\left(-Z_i\sigma_i - np_i - \tfrac{1}{2}\right) ln\left(1 + \frac{Z_i\sigma_i}{np_i}\right)$$

$$= -\sum_{i=1}^{k}\left(Z_i\sigma_i + np_i + \tfrac{1}{2}\right)\left[\frac{Z_i\sigma_i}{np_i} - \frac{1}{2}\left(\frac{Z_i\sigma_i}{np_i}\right)^2 + \frac{1}{3}\left(\frac{Z_i\sigma_i}{np_i}\right)^3 - \ldots\right]$$

Replacing σ_i by its value $\sqrt{np_i\,(1-p_i)}$ and arranging the terms in descending order of n, we obtain

$$\ln M = -\sum_{i=1}^{k} Z_i \sqrt{p_i(1-p_i)}\, n^{1/2} + \sum_{i=1}^{k}\left(\frac{1}{2} Z_i^2 q_i - q_i Z_i^2\right) + \text{terms of lower order.}$$

Since $\sum_{i=1}^{k} Z_i \sqrt{p_i(1-p_i)} \cdot n^{1/2} = \sum_{i=1}^{k}(n_i - np_i)$

$$= \sum_{i=1}^{k} n_i - n \sum_{i=1}^{k} p_i = n - n = 0,$$

and $\sum_{i=1}^{k}\left(\frac{1}{2} Z_i^2 q_i - q_i Z_i^2\right) = -\frac{1}{2}\sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i}$

$$\therefore \quad M = e^{-(1/2)\Sigma(n_i - np_i)^2/np_i} = e^{-\chi^2/2},$$

where $\chi^2 = \sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i}$.

Hence, we get the result as

$$f(n_1, n_2, \ldots, n_k) \simeq \frac{e^{-\chi^2/2}}{C} = B\, e^{-\chi^2/2},$$

where $B = (2\pi n)^{(1-k)/2} \cdot (p_1 p_2 \ldots p_k)^{1/2}$

That is $f(n_1, \ldots, n_k)$ varies as that of the sum of squares of k normal variables subject to $\sum_{i=1}^{k} Z_i \sqrt{p_i(1-p_i)}\, n = 0$, but otherwise independent. Thus it is a good approximation if all the expected values np_i are sufficiently large and hence the χ^2 as defined above, conforms to the χ^2-distribution with $(k-1)$ degrees of freedom.

We therefore conclude that, if a sample of n observations is distributed over k classes such that the observed frequency in the *ith* class is n_i and the expected frequency in that class is np_i, the statistic

$$\chi^2 = \sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k}\frac{(\text{Observed } n_i - \text{Expected } n_i)^2}{\text{Expected } n_i}$$

called the χ^2-*statistic*, has a χ^2-distribution with $(k-1)$ degrees of freedom.

It is clear that χ^2 will be small when all n_i are close to their expected values np_i. The χ^2 will become larger when the difference becomes larger. The χ^2 thus measures the amount of deviation (or agreement) between observed and expected results.

It is important to bear in mind that in the light of the proof given above and the discussion by Fisher (1890–1962) on this subject, the following requirements, while applying χ^2-statistic, must be satisfied:

(i) The size of the sample n or the total number of observations should be at least 50, otherwise the Z_i will not be normally distributed.

(ii) The observations in the sample or the frequencies in the classes or cells should be independent.

(iii) The restrictions or constraints, if any, should be linear.

(iv) The expected number, np_i in any of the classes or cells should not be less than 5. So when the expected frequency, np_i in any class is less than 5, we may combine this class with one or more other classes to meet this requirement.

17.5.1. Testing Hypothesis about p's of the Multinomial Distribution. In multinomial type problems, where there are classes or cells and where the cell probabilities p_i's are completely specified, the procedure for testing the hypothesis $H_0 : p_i = p_{i0}$, $i = 1, 2, \ldots, k$, is given below:

(i) Formulate the null and alternative hypotheses about p's as

$H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$, and

$H_1 : p_i \neq p_{i0}$ for at least one value of $i = 1, 2, \ldots, k$,

where $p_{10}, p_{20}, \ldots, p_{k0}$ are specified values.

(ii) Decide on significance level α.

(iii) The test-statistic to use is

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}$$

which, if H_0 is true, has an approximate chi-square distribution with $(k-1)$ degrees of freedom.

(iv) Determine the critical region which depends on α and the degrees of freedom.

(v) Compute the value of χ^2, after having calculated the expected values np_{i0}, from the given data.

(vi) Decide as below:

Reject H_0, if the computed value of $\chi^2 \geq \chi^2_{\alpha,(k-1)}$.

Accept H_0, otherwise.

Example 17.7. In a genetic experiment involving the crossing of two types of peas, Mendel observed 315 round and yellow, 108 round and green, 101 angular and yellow, and 32 angular and green seeds from the plants but his theory of heredity called for a 9 : 3 : 3 : 1 ratio. Using a 5 percent and a 1 percent significance level, do the data support the theory? (P.U., B.A./B.Sc. 1972, 82, 91)

(i) We state our null and alternative hypotheses as

$H_0 : p_1 = \frac{9}{16}, \; p_2 = \frac{3}{16}, \; p_3 = \frac{3}{16}, \; p_4 = \frac{1}{16}$ for a multinomial distribution involving four cells and with $n = 556$; and

$H_1 : p_i \neq p_{i0}$ for at least one value of $i = 1, 2, 3, 4$.

(ii) The significance levels are set at $\alpha = 0.05$ and $\alpha = 0.01$.

(iii) The test-statistic under H_0, is

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - np_{i0})^2}{np_{i0}},$$

which has an approximate chi-square distribution with 3 *d.f.*

(iv) **Computations.** Under H_0, the expected frequencies are

np_{10} = round and yellow seeds $= 556 \times \frac{9}{16} = 312.75$;

np_{20} = round and green seeds $= 556 \times \frac{3}{16} = 104.25$;

np_{30} = angular and yellow seeds $= 556 \times \frac{3}{16} = 104.25$;

np_{40} = angular and green seeds $= 556 \times \frac{1}{16} = 34.75$.

The value of χ^2 is then computed as below:

Cell	n_i	np_{i0}	$n_i - np_{i0}$	$(n_i - np_{i0})^2$	$(n_i - np_{i0})^2/np_{i0}$
1	315	312.75	2.25	5.0625	0.016
2	108	104.25	3.75	14.0625	0.135
3	101	104.25	−3.25	10.5625	0.101
4	32	34.75	−2.75	7.5625	0.218
Σ	556	556.00	---	---	$\chi^2 = 0.470$

(v) The critical regions are $\chi^2 \geq \chi^2_{0.05,(3)} = 7.82$, and

$$\chi^2 \geq \chi^2_{0.01,(3)} = 11.34.$$

(vi) **Conclusion.** Since the computed value of χ^2 does not fall in the critical region at both the significance levels, so we do not reject our null hypothesis. There is sufficient evidence to conclude that the data support the hypothesized theory.

Example 17.8. Two hundred digits were chosen at random from a set of tables. The frequencies of the digits were:

Digit:	0	1	2	3	4	5	6	7	8	9
Frequency:	18	19	23	21	16	25	22	20	21	15

Use the χ^2-test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which these were chosen.

(i) We formulate our hypotheses as

$H_0 : p_1 = p_2 = \ldots = p_{10} = \frac{1}{10}$ for a multinomial distribution involving 10 classes and with $n = 200$, and

$H_1 : p_i \neq p_{i0}$ for at least one value of $i = 1, 2, \ldots, 10$, (or the digits were not distributed in equal numbers).

(ii) We use a significance level of $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \sum \frac{(n_i - np_{i0})^2}{np_{i0}}.$$

When H_0 is true, the statistic has an approximate chi-square distribution with $k-1$, *i.e.* 9 degrees of freedom.

(iv) The rejection region is $\chi^2_{0.05,(9)} \geq 16.92$.

(v) **Computations.** Under H_0, the expected frequency of each of the digits 0, 1, 2, ..., 9 is $np_i = 200\left(\frac{1}{10}\right) = 20$. Therefore

$$\chi^2 = \sum \frac{(n_i - np_{i0})^2}{np_{i0}} = \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \ldots + \frac{(15-20)^2}{20}$$

$$= \frac{1}{20}[4 + 1 + 9 + 1 + 16 + 25 + 4 + 0 + 1 + 25]$$

$$= \frac{86}{20} = 4.3.$$

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 4.3$ does not fall in the rejection region, we are therefore unable to reject our null hypothesis. Hence we may conclude that the digits appear to be distributed in equal numbers in the tables from which they were chosen.

Example 17.9. A coin is thrown 1600 times, and 840 heads are recorded. Test the hypothesis that the coin is unbiased. Use both the chi-square approximation and the normal approximation and compare the results. Use $\alpha = 0.05$.

(a) **Using Chi-square Approximation (to multinomial)**

(i) We state our null and alternative hypotheses in multinomial notation, as

$H_0 : p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{2}$ for a multinomial distribution involving 2 classes and with $n = 1600$; and

H_1: both p's are not equal to 1/2.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0, is

$$\chi^2 = \frac{(n_1 - np_{10})^2}{np_{10}} + \frac{(n_2 - np_{20})^2}{np_{20}},$$

where n_1 and n_2 are the number of heads (840) and number of tails (760) in this case. The test-statistic has an approximate chi-square distribution with 1 *d.f.*

(iv) **Computations.** Under H_0, the expected frequencies are

$$np_{10} = 1600 \times \frac{1}{2} = 800, \quad np_{20} = 1600 \times \frac{1}{2} = 800.$$

$$\therefore \quad \chi^2 = \frac{(840 - 800)^2}{800} + \frac{(760 - 800)^2}{800} = 2 + 2 = 4$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$

(vi) **Conclusion.** Since the computed value of $\chi^2 = 4$ falls in the critical region, we therefore reject our null hypothesis. The data provide evidence to conclude that the coin is biased.

(b) **Using Normal Approximation (to binomial)**

(i) We set up our null and alternative hypotheses in binomial notation, as

$$H_0 : p = \frac{1}{2} \text{ and } H_1 : p \neq \frac{1}{2}.$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$Z = \frac{x - np_0}{\sqrt{np_0q_0}}, \text{ (without continuity correction)}$$

where x is the number of heads and Z is approximately standard normal.

(iv) **Computations.** We compute the value of Z as

$$Z = \frac{840 - 800}{\sqrt{1600\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{40}{20} = 2$$

(v) The critical region is $|Z| \geq 1.96$.

(vi) **Conclusion.** Since the computed value $z = 2$ falls in the critical region, so we reject H_0 and may conclude that the coin is a biased one.

Comparison. The solutions give identical results. It is observed that the chi-square approximation gives a test that is equivalent to the normal approximation when $k = 2$. The chi-square statistic is the square of the normal statistic. Hence with two categories, either procedure can be used.

17.5.2. Pearson's Chi-Square Test For Goodness-of-Fit. A χ^2-statistic can also be applied even when the cell probabilities are not known and they depend upon the unknown parameters of a specified distribution such as the binomial distribution, the Poisson distribution the normal distribution, etc. This test is based on the property that "*the χ^2-test is applicable when the cell probabilities depend upon unknown parameters, provided that the unknown parameters are replaced with their estimates and provided that one degree of freedom is deducted for each parameter estimated.*" When there are k classes/categories and the class probabilities are known, the number of degrees of freedom is $(k-1)$. When the probabilities depend upon m parameters, the degrees of freedom would be $k-1-m$, *i.e. d.f.* = number of classes$-1-$number of parameters estimated from the sample. For example, in a normal

distribution, the cell probabilities depend upon the two parameters and σ, therefore the degree of freedom is $(k-1-2)$, *i.e.* $(k-3)$.

A *goodness-of-fit* test is a hypothesis test that is concerned with the determination whether results of a sample conform to a hypothesized distribution which may be the uniform, binomial, Poisson, Normal or any other distribution. This is a kind of hypothesis test for problems where we do not know the probability distribution of the random variable under consideration, say X, and we wish to test the hypothesis that X follows a particular distribution. In the test procedure, the range of all possible values of the random variable assumed to follow a particular distribution is divided into k mutually exclusive classes and the probabilities p_i's are calculated for each of the classes, using the estimates of the parameters of the probability distribution specified in H_0. The $n\hat{p}_i$ represents the expected number of observations that fall in the *ith* class and n_i represents the observed number of observations in that class. The differences between observed and expected number of observations can arise from sampling error or from H_0 being false. Small differences are generally attributed to sampling error, large differences which are considered to arise from H_0 being false, are unlikely if the hypothesized distribution gives a satisfactory fit to the sample data (H_0 true). To see whether there is evidence of small or large differences, the test-statistic to use is

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = \sum \frac{(o_i - e_i)^2}{e_i},$$

which, when H_0 is true, has an approximate chi-square distribution with $d.f. = k - 1 -$ number of parameters estimated by sample statistics. The symbols o_i and e_i represent the observed and expected values of n_i respectively.

When the observed values are equal to the expected values, the $\chi^2 = 0$. The larger the differences between observed and expected values, the larger will be the value of χ^2. A small computed value of χ^2 indicates a good fit and it leads to the acceptance of the null hypothesis. A large computed value of χ^2 indicates a poor fit and it leads to the rejection of the null hypothesis. Hence the *rejection region in a goodness-of-fit test* (and all tests that compare frequencies) *will fall in the right tail of the chi-square distribution.*

The procedure for a goodness-of-fit test is as follows:

(i) Formulate the null and alternative hypotheses as

H_0: The population has a specified probability distribution, and

H_1: The population does not have the specified distribution.

(ii) Choose the level of significance α. The commonly used value is $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

which, if H_0 is true, has an approximate chi-square distribution with $d.f. = k-1-$number of estimated parameters.

(iv) Determine the critical region, which depends upon α and the degrees of freedom.

(v) Compute the expected values and the value of χ^2.

(vi) Decide as below:

Reject H_0, if the calculated value of χ^2 exceeds the χ^2_α value against the appropriate degrees of freedom from the χ^2-table. Accept H_0, otherwise.

Example 17.10. Five pennies were tossed 1,000 times and the number of heads were observed as given below:

Number of heads	0	1	2	3	4	5
Frequencies	38	144	342	287	164	25

Test whether a binomial distribution gives a satisfactory fit to these data. (P.U., B.A./B.Sc. 1980)

(i) We state our hypotheses as

H_0: The population distribution is a binomial with $n=5$, but with parameter p unspecified, and

H_1: The population distribution is not a binomial with $n=5$.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) We use the test-statistic

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i},$$

which, if H_0 is true, has an approximate χ^2-distribution with degrees of freedom $= k-1-$number of estimated parameters. A binomial distribution has two parameters n and p but n is specified ($n=5$). We have to estimate the value of one parameter

p from the sample data. Therefore the degrees of freedom $= 6-1-1=4$ ($\because k=6$ categories).

(iv) The critical region is $\chi^2 \geq \chi^2_{0.05,(4)} = 9.49$.

(v) **Computations.** To estimate the value of p, we first compute the mean number of heads, $\bar{x}$. Thus

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{2470}{1000} = 2.47.$$

Theoretically, $\bar{x} = np$, so that $\hat{p} = \frac{\bar{x}}{n} = \frac{2.47}{5} = 0.494$.

Hence the expected (fitted) frequencies are the terms in the binomial expansion of $1{,}000(0.506+0.494)^5$, which are given in the column headed e_i. Next, we calculate the value of χ^2 as follows:

Number of heads	Observed f_i (o_i)	Expected f_i (e_i)	$o_i - e_i$	$(o_i - e_i)^2$	$(o_i - e_i)^2/e_i$
0	38	33.2	4.8	23.04	0.69
1	144	161.9	−17.9	320.41	1.98
2	342	316.2	25.8	665.64	2.[illegible]5
3	287	308.7	−21.7	470.89	1.53
4	164	150.7	13.3	176.89	1.17
5	25	29.4	−4.4	19.36	0.66
Total	1,000	1,000.1	---	---	$\chi^2 = 8.18$

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 8.18$ does not fall in the critical region, we therefore are unable to reject our null hypothesis. We may accept the hypothesis that the distribution of the number of heads is a binomial distribution and conclude that the fit of data is good.

Example 17.11. A skilled typist, on routine work, kept a record of mistakes made per day during 300 working days.

Mistakes per day (x)	0	1	2	3	4	5	6
Number of days (f)	143	90	42	12	9	3	1

Test the hypothesis that X has a Poisson distribution by applying the χ^2 goodness-of-fit-test.

(i) We state our null and alternative hypotheses as

H_0: The population has a Poisson distribution with mean μ unspecified, and

H_1: The population does not have a Poisson distribution.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

which, if H_0 is true, has an approximate χ^2-distribution with degrees of freedom $= k-1-$number of estimated parameters.

(iv) **Computations.** A Poisson distribution has one parameter μ. To estimate the value of μ, we compute the mean number of mistakes per day. Therefore

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{267}{300} = 0.89$$

Thus the fitted Poisson distribution is

$$p(x; 0.89) = \frac{e^{-0.89}(0.89)^x}{x!}, \text{ for } x = 0, 1, 2, \ldots$$

We calculate the expected frequencies (given $e^{-0.89} = 0.4107$) which appear in the column headed e_i, and the value of χ^2 as follows:

Mistakes per day (x)	Observed f, (o_i)	Expected f, (e_i)	$o_i - e_i$	$(o_i - e_i)^2/e_i$
0	143	123.2	19.8	3.18
1	90	109.6	−19.6	3.51
2	42	48.8	− 6.8	0.95
3	12	14.5		
4	9	3.2	6.6	2.37
5	25	0.6 } 18.4		
6	1	0.1		
Σ	300	300	—	$\chi^2 = 10.01$

To meet the requirement that the expected frequency in any of the classes or cells should be at least 5

that the new final class corresponds to $x \geq 3$ with expected frequency = 18.4, the corresponding observed frequencies are combined also. The test-statistic is then $\chi^2 = \sum_{i=0}^{3} \frac{(o_i - e_i)^2}{e_i}$. The number of degrees of freedom is $k-1-$number of estimated parameters. Here $k=4$, the number of expected frequencies used in computing χ^2, and we have to estimate the value of one parameter, μ from the sample data. Thus the degrees of freedom $= 4-1-1=2$.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 10.01$ falls in the critical region, so we reject our null hypothesis. We may conclude that the Poisson distribution is not a good fit to the data.

Example 17.12. Test the hypothesis that the following frequency distribution follows a normal distribution at $\alpha = 0.05$.

Intervals	10–,	12–,	14–,	16–,	18–,	20–,	22–	24–,	26–
Frequency	4	30	106	206	272	219	120	37	6

(i) We formulate our null and alternative hypotheses as

H_0: The distribution is normal with mean μ and variance σ^2, and

H_1: The distribution is not normal.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i},$$

which, if H_0 is true, has an approximate χ^2-distribution with $(k-1-m)$ degrees of freedom and where k denotes the number of intervals (after pooling, if necessary) and m represents the number of parameters estimated from the sample data.

(iv) **Computations.** We first need to fit a normal distribution, but neither the mean μ, nor the standard deviation σ, is given. We therefore estimate μ by the sample mean $\bar{x}$ and σ by the sample standard deviation s. Using the data, we find $\sum fx = 19140$ and $\sum fx^2 = 374688$, so that $\bar{x} = 19.14$ and $s = 2.89$.

Next we need to compute the expected frequencies for all classes and the value of χ^2. The necessary calculations for expected

frequencies, e_i's ($e_i = n\hat{p}_i$, where $\hat{p}_i$ is the estimate of p_i) together with the value of χ^2 are shown below:

Upper class boundary	z_i	$P(Z<z)$	$\hat{p}_i$	Expected frequency e_i $(=n\hat{p}_i)$	Observed frequency (o_i)	$(o_i-e_i)^2/e_i$
12	−2.49	0.0068	0.0068	6.8	4	1.147
14	−1.78	0.0375	0.0307	30.7	30	0.016
16	−1.09	0.1379	0.1004	100.4	106	0.312
18	−0.39	0.3483	0.2104	210.4	206	0.092
20	+0.30	0.6179	0.2696	269.6	272	0.021
22	0.99	0.8389	0.2210	221.0	219	0.018
24	1.68	0.9535	0.1146	114.6	120	0.254
26	2.37	0.9911	0.0376	37.6	37	0.010
∞	∞	1.0000	0.0089	8.9	6	0.945
Total	--	--	--	1000.0	1000	2.815

There are 9 calss-intervals (no classes have been combined) and we have used the sample mean, $\bar{x}$ and the sample standard deviation, s to estimate the two parameters μ and σ, so the number of degrees of freedom is $9-1-2=6$.

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(6)} = 12.59$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 2.82$ does not fall in the critical region, so we are unable to reject our null hypothesis and may conclude that the normal distribution provides a good fit for the given frequency distribution.

17.5.3. Testing Hypothesis about Independence of two Variables. The χ^2-statistic can also be used to test the hypothesis about independence of two variables, each of which is classified into a number of categories or attributes. Before we discuss the format of the tests of independence, we give a brief description of the theory of attributes.

17.6 AN ASIDE – ATTRIBUTES

We have been dealing with quantitative data obtained by measuring the actual magnitude of some variable character such as age, height, price, score, etc. for each of the individuals or objects. The quantitative data may a so be obtained if we simply note the *presence* or *absence* of a

qualitative characteristic and count how many do or do not possess it. The qualitatively distinct characteristics such as male or female, tall or short, satisfied or dissatisfied, high or low, healthy or diseased, positive or negative, etc. are called *attributes*. The attributes cannot be measured accurately but they can be divided into classes and their numbers in each class can be counted. If the data (*i.e.*, population) are divided into two distinct and mutually exclusive classes by a single attribute as for instance, the population of human beings is divided into males and females, the process is called *dichotomy* (cutting in two).

The capital letters A, B, C, ... are used to denote the several attributes. The individuals or objectives *possessing* the attributes A, B, C, ... are conventionally designated A, B, C, ... while the absence of these attributes by the Greek letters α, β, γ, ... Thus if A denotes that the individual or object possesses the attribute A, then α will denote that the individual does not possess the attribute A. Hence "α" is equivalent to "not-A". The attributes denoted by the letters A, B, C, ... are called the positive attributes and those denoted by α, β, γ, ..., the negative attributes. The letters are enclosed in brackets to denote the class-frequencies of the corresponding attributes such as (A), (B), etc., and they are combined to denote two or more attributes simultaneously. Thus, if A represents "intelligence" and B, "smoking", AB will represent "intelligence and smoking and (AB) the number of individuals who are intelligent and smokers." Similarly $(A\beta)$ will represent the number of individuals who are "intelligent but non-smokers" and so on. The order of the class-frequencies depends on the number of attributes combined together except N, the population size which is regarded as a frequency of order zero.

It is interesting to note that any class-frequency can always be expressed in terms of class-frequencies of higher order. For example,

$$(A) = (AB) + (A\beta)$$

i.e., the number of A's is equal to the number of A's which are B's plus the number of A's which are β's. The frequencies of the highest order are termed as the *ultimate* class-frequencies. We can also express every class-frequency as the sum of certain of the ultimate class-frequencies. As for instance, we have

$$(A) = (AB) + (A\beta)$$
$$= (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$$

The frequencies on the right hand side are some of the ultimate class-frequencies, if we consider three attributes A, B, and C.

In order to express any class-frequency in terms of the known class frequencies, we can treat the class symbols as *operators* and can multiply them together like algebraical quantities. Writing $A.N$ for the operation of dichotomising N according to attribute A, we can write

$$A \,.\, N = (A);$$

i.e., by dichotomising N according to A, we get (A), the class frequency of A.

Similarly, $\alpha \,.\, N = (\alpha)$

Adding, we get $A.N + \alpha \,.\, N = (A) + (\alpha)$

or $(A + \alpha) \,.\, N = N$ $\quad [\because N = (A) + (\alpha)]$

Omitting N, we get a symbolic relation as $A + \alpha = 1$. Hence we can replace A by $1-\alpha$ and α by $1-A$. It has been noted that these operative symbols obey the ordinary laws of algebra. For example, we have

$$\begin{aligned}(\alpha B) &= \alpha B \,.\, N = (1 - A)\, B \,.\, N \\ &= (B - AB) \,.\, N = (B) - (AB)\end{aligned}$$

Similarly,

$$\begin{aligned}(\alpha\beta\gamma) &= \alpha\beta\gamma \,.\, N = (1 - A)\,(1 - B)\,(1 - C) \,.\, N \\ &= (1 - A - B - C + AB + BC + AC - ABC) \,.\, N \\ &= N-(A)-(B)-(C) + (AB) + (BC) + (AC)-(ABC).\end{aligned}$$

Example 17.13. Given that $(A) = (\alpha) = (B) = (\beta) = (C) = (\gamma) = \frac{N}{2}$ and also that $(ABC) = (\alpha\beta\gamma)$, show that

$$2(ABC) = (AB) + (BC) + (CA) - \frac{N}{2}$$

We are given that $(ABC) = (\alpha\beta\gamma)$

But

$$\begin{aligned}(\alpha\beta\gamma) &= \alpha\beta\gamma \,.\, N = (1-A)\,(1-B)\,(1-C) \,.\, N \\ &= (1-A-B-C + AB + BC + CA - ABC) \,.\, N \\ &= N - (A) - (B) - (C) + (AB) + (BC) + (CA) - (ABC) \\ &= N - \frac{N}{2} - \frac{N}{2} - \frac{N}{2} + (AB) + (BC) + (CA) - (ABC) \\ &= (AB) + (BC) + (CA) - \frac{N}{2} - (ABC).\end{aligned}$$

Substituting these values, we get

$$2(ABC) = (AB) + (BC) + (CA) - \frac{N}{2}$$

Example 17.14. 100 children took three examinations. 40 passed the first, 39 passed the second and 48 passed the third. 10 passed all three, 21 failed all three, 9 passed the first two and failed the third, 19 failed the first two and passed the third. Find how many children passed *at least* two examinations?

Let A, B and C stand for passing the first, second and third examinations respectively. Then we are given

$$N = 100;\ (A) = 40;\ (B) = 39;\ (C) = 48;\ (ABC) = 10;$$
$$(\alpha\beta\gamma) = 21;\ (AB\gamma) = 9;\ (\alpha\beta C) = 19$$

We are required to find the number of students who passed *at least* two (*i.e.*, two or more) examinations; that is we need the value of

$$(AB\gamma) + (A\beta C) + (\alpha BC) + (ABC)$$

Now $\quad (C) = (ABC) + (\alpha BC) + (A\beta C) + (\alpha\beta C)$

$\therefore \quad (ABC) + (\alpha BC) + (A\beta C) = (C) - (\alpha\beta C)$

Adding $(AB\gamma)$ to both sides, we get

$$(ABC) + (\alpha BC) + (A\beta C) + (AB\gamma) = (C) - (\alpha\beta C) + (AB\gamma)$$
$$= 48 - 19 + 9 = 38.$$

17.6.1. Consistence. Any class-frequencies observed within one and the same population are said to be *consistent* with one another, because they conform with one another and do not conflict in any way. The necessary and sufficient condition for the *consistence* is that no ultimate class-frequency should be negative for the obvious reason that they are obtained by counting real attributes. Hence to test the consistence, we calculate the ultimate class-frequencies from the given data. If any ultimate class-frequency turns out to be negative, the given set of class frequencies will be inconsistent. The data are consistent, when all the ultimate class-frequencies are positive.

Example 17.15. A market investigator returns the following data. Of 1,000 people consulted, 811 liked chocolates, 752 liked toffee and 418 liked boiled sweets; 570 liked chocolates and toffee, 356 liked chocolates and boiled sweets and 348 liked toffee and boiled sweets; 297 liked all three. Show that this information as it stands must be incorrect.

Let A, B and C denote the liking of chocolates, toffee and boiled sweets respectively. Then we are given

$$N = 1{,}000;\quad (A) = 811;\quad (B) = 752;\quad (C) = 418;$$
$$(AB) = 570;\quad (AC) = 356;\quad (BC) = 348;\quad (ABC) = 297.$$

The test of consistence is that no ultimate class-frequency should be negative.

$$\text{Now} \quad (\alpha\beta\gamma) = \alpha\beta\gamma \,.\, N = (1-A)\,(1-B)\,(1-C)\,.\,N$$
$$= N-(A)-(B)-(C) + (AB) + (AC) + (BC) - (ABC)$$
$$= 1000 - 811 - 752 - 418 + 570 + 356 + 348 - 297$$
$$= -4, \text{ which is negative.}$$

Hence the information as it stands, is not correct. However, if the data returned are alleged to be the result of an actual enquiry in a definite population, there must have been some misprint or miscount or mis-reporting.

17.6.2. Independence. Suppose in a population of size N, the class frequencies of two attributes A and B are given by (A) and (B). Then, we have

$$\text{the proportion of } (A) = \frac{(A)}{N},$$

$$\text{the proportion of } (B) = \frac{(B)}{N},$$

$$\text{the proportion of } (A) \text{ and } (B) \text{ combined} = \frac{(A)}{N} \cdot \frac{(B)}{N}, \text{ and}$$

$$\text{the expectation of } (A) \text{ and } (B) \text{ combined} = \frac{(A)}{N} \cdot \frac{(B)}{N} \,.\, N = \frac{(A)\,(B)}{N}$$

The two attributes A and B are said to be *independent* if the actual frequency equals the expected one, that is, if

$$(AB) = \frac{(A)\,(B)}{N}.$$

Similarly, α and β will be independent if $(\alpha\beta) = \dfrac{(\alpha)\,(\beta)}{N}$, and so on. In case the ultimate class-frequencies on the two attributes A and B are given, the required criterion of independence for the two attributes A and B will be

$$(AB)\,(\alpha\beta) = \frac{(A)\,(B)}{N} \cdot \frac{(\alpha)\,(\beta)}{N}$$
$$= \frac{(A)\,(\beta)}{N} \cdot \frac{(\alpha)\,(B)}{N} = (A\beta)\,(\alpha B).$$

Suppose the class-frequencies may be grouped into a table with two rows and two columns as follows:

Attribute	B	β	Total
A	(AB)	$(A\beta)$	(A)
α	(αB)	$(\alpha\beta)$	(α)
Total	(B)	(β)	N

When the two attributes are independent, this table must have the following form:

Attribute	B	β	Total
A	$(A)(B)/N$	$(A)\,(\beta)/N$	(A)
α	$(\alpha)(B)/N$	$(\alpha)\,(\beta)/N$	(α)
Total	(B)	(β)	N

17.6.3. Association of Attributes. The word *association* has a technical meaning in Statistics. In ordinary language, if A and B appear together fairly often, we speak of them as being associated. But in statistical usage, they are said to be associated only if they appear together in a larger number of cases than is to be expected if they were independent. Thus the mere fact that some A's are B's, however great the proportion, is not enough to conclude that A and B are associated. This is a fundamental principle and should always be borne in mind. Symbolically, A and B are said to be *positively associated* or simply *associated*, if

$$(AB) > \frac{(A)\,(B)}{N}$$

On the contrary, A and B are said to be *negatively associated* or briefly, *dissociated* if

$$(AB) < \frac{(A)\,(B)}{N}$$

It should be remembered that dissociation does not imply independence.

17.6.4. Measures of Association. The strength of association between two attributes A and B is measured by a co-efficient, called the *co-efficient of association* and defined by the formula

$$Q = \frac{(AB)\,(\alpha\beta) - (A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta) + (A\beta)\,(\alpha B)}.$$

This co-efficient is due to George Udny Yule (1871–1951) and like the co-efficient of correlation, it lies between -1 and $+1$. $Q=0$, when th

As the co-efficient is positive and high, we conclude that coaching helps in success, *i.e.*, the utility of coaching class is very great.

17.6.5. Contingency Tables. A table that consists of two or more rows and two or more columns, into which n observations are classified according to two different criteria (or variables) is commonly called a *contingency table*, a name, probably due to Karl Pearson. For example, if the n observations of a random sample are classified according to two variables, say A and B where there are r distinct classes $A_1, A_2, \ldots, A_r$ in A and c distinct classes $B_1, B_2, \ldots, B_c$ in B and the number of observations (*i.e.* frequencies) belonging to A_i and B_j is o_{ij} with $\sum\sum o_{ij} = n$, then the contingency table will be one of r rows and c columns as shown below:

An $r \times c$ Contingency Table

Classes	B_1	B_2	...	B_j	...	B_c	Total
A_1	o_{11}	o_{12}	...	o_{1j}	...	o_{1c}	(A_1)
A_2	o_{21}	o_{22}	...	o_{2j}	...	o_{2c}	(A_2)
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	o_{i1}	o_{i2}	...	o_{ij}	...	$o_{.c}$	(A_i)
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	o_{r1}	o_{r2}	...	o_{rj}	...	o_{rc}	(A_r)
Total	(B_1)	(B_2)	...	(B_j)	...	(B_c)	n

A contingency table may be extended to higher dimensions. The simplest form of a contingency table is the 2×2 table which is obtained when both criteria are dichotomised. The totals of the frequencies in each of the rows and columns are called the *marginal total* or frequencies. Contingency tables provide a useful method of comparing two variables.

17.7 TESTING HYPOTHESIS OF INDEPENDENCE IN CONTINGENCY TABLES

The data presented in a contingency table can be used to test the hypothesis that the two variables of classification are independent. If this hypothesis is rejected, the two variables of classification are not independent and we say that there is some association (or interaction) between the two variables of classification. To do so, we must calculate the expected frequencies based on this hypothesis, keeping the marginal totals fixed.

Let e_{ij} denote the expected frequency belonging to A_i and B_j. Assuming the hypothesis of independence is true, the proportion of

members belonging to any class A_i should be the same and equal to the proportions in the total. Thus

$$\frac{e_{ij}}{(A_i)} = \frac{\sum_{i=1}^{r} e_{ij}}{\sum_{i=1}^{r} (A_i)} = \frac{(B_j)}{n} \text{ so that } e_{ij} = \frac{(A_i)\,(B_j)}{n},$$

that is, under H_0: the classifications are independent, the expected frequency in any cell is equal to the product of the marginal totals common to that cell divided by the total number of observations.

If our hypothesis of independence is true, the differences between observed and expected frequencies are small and are attributed to sampling error. Large differences arise from H_0 being false. The chi-square statistic provides a means for deciding whether the differences are large or small overall. Hence the statistic to use is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

A large value of χ^2 indicates that the null hypothesis is false. The number of degrees of freedom in a contingency table is obtained as follows:

In an $r \times c$ contingency table, there are in all $r \times c$ cells. If we look at the jth column of the contingency table, we see that when $(r-1)$ of the values are determined, the rth value is automatically obtained as the marginal total is to remain fixed. Similarly, for the ith row, we see that when $(c-1)$ of the values are determined, the cth value is determined from the known fixed marginal total. Thus there are $(r-1) \times (c-1)$ values that may be determined *freely*, whereas the remaining $rc-(r-1)(c-1)$ values will be determined from the marginal totals. Hence the number of degree of freedom in an $r \times c$ contingency table is $(r-1)(c-1)$, *i.e.*, the product of the number of rows minus one and the number of columns minus one.

The procedure for testing the null hypothesis of independence in contingency tables is given below:

(i) Formulate the null and alternative hypotheses as

H_0: the two variables of classification are independent, and

H_1: the two variables of classification are not independent, i.e. they are associated.

(ii) Choose a significance level α. The commonly used levels are at $\alpha = 0.05, 0.01$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

which, if H_0 is true, has an approximate chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

(iv) Compute the expected frequencies under H_0 for each cell by the formula $e_{ij} = \dfrac{(A_i)(B_j)}{n} = \dfrac{(ith \text{ Row Total})(jth \text{ Colum Total})}{\text{Total number of observations}}$.

Also calculate the value of χ^2 and the degrees of freedom.

(v) Determine the critical region which depends on α and the number of degrees of freedom.

(vi) Decide as below:

Reject H_0, if the computed value of $\chi^2 \geq \chi^2_{\alpha,(r-1)(c-1)}$

Accept H_0, otherwise.

Example 17.17. Four hundred and ninety-two candidates for scientific posts gave particulars of their university degrees and their hobbies. The degrees were in either maths, chemistry or physics, and the hobbies could be classified roughly as music, craftswork, reading or drama. The data are presented concisely in the following contingency table:

Categories	Maths.	Chemistry	Physics	Total
Music	24	83	17	124
Craftswork	11	62	28	101
Reading	32	121	34	187
Drama	10	26	44	80
Total	77	292	123	492

Discuss the association between the two criteria of classification, i.e., the degrees and the hobbies. (P.U., B.A./B.Sc., 1969)

(i) We state our null and alternative hypotheses as

H_0: The two criteria of classification are independent, and

H_1: The two criteria of classification are not independent, i.e., they are associated.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

which, if H_0 is true, has an approximate χ^2-distribution with $(4-1)(3-1) = 6$ degrees of freedom.

(iv) We first calculate the expected frequencies under the null hypothesis of independence for each cell by the formula

$$e_{ij} = \frac{(A_i)(B_j)}{n} = \frac{(ith \text{ Row Total})(jth \text{ Column Total})}{\text{Total number of observations}}.$$

The expected frequencies are given in the following table:

Categories		Maths: B_1	Chemistry: B_2	Physics: B_3	Total
Music:	A_1	$\frac{77\times124}{492} = 19.4$	73.6	31.0	124
Craftswork	A_2	$\frac{77\times101}{492} = 15.8$	59.9	25.3	101
Reading:	A_3	29.3	111.0	46.7	187
Drama:	A_4	12.5	47.5	20.0	80
Total		77	292	123	492

Now we calculate the chi-square statistic as below:

	o_{ij}	e_{ij}	$o_{ij}-e_{ij}$	$(o_{ij}-e_{ij})^2$	$(o_{ij}-e_{ij})^2/e_{ij}$
	24	19.4	+4.6	21.16	1.09
	11	15.8	−4.8	23.04	1.46
	32	29.3	2.7	7.29	0.29
	10	12.5	−2.5	6.25	0.50
	83	73.6	9.4	88.36	1.20
	62	59.9	2.1	4.41	0.07
	121	111.0	10.0	100.00	0.90
	26	47.5	−21.5	462.25	9.73
	17	31.0	−14.0	196.00	6.32
	28	25.3	2.7	7.29	0.29
	34	46.7	−12.7	161.29	3.45
	44	20.0	24.0	576.00	28.80
Total	492	492	--	--	$\chi^2 = 54.06$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(6)} = 12.59$

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 54.06$ falls in the critical region, we therefore reject our null hypothesis of independence and conclude that the data provide evidence of a statistical association between the two criteria of classification.

Alternative Procedure. When an automatic desk calculator is available, a much shorter method for computing χ^2, given by *J.* Skory, contains the following three steps:

(i) Compute $\sum_i \frac{O_{ij}^2}{(A_i)}$ for each column. Denote these by T_j.

(ii) Compute $\sum_j \frac{T_j}{(B_i)}$. Denote this by R.

(iii) Then $\chi^2 = (R - 1)\, n$.

Applying this method to our Example 17.17, we get

(i) $T_1 = \frac{(24)^2}{124} + \frac{(11)^2}{101} + \frac{(32)^2}{187} + \frac{(10)^2}{80} = 12.5691,$

$T_2 = \frac{(83)^2}{124} + \frac{(62)^2}{101} + \frac{(121)^2}{187} + \frac{(26)^2}{80} = 180.3600,$

$T_3 = \frac{(17)^2}{124} + \frac{(28)^2}{101} + \frac{(34)^2}{187} + \frac{(44)^2}{80} = 40.4748,$ and

(ii) $R = \frac{12.5691}{77} + \frac{180.3600}{292} + \frac{40.4748}{123} = 1.11.$

Hence (iii) $\chi^2 = 492\,(1.11 - 1) = 54.12.$

Example 17.18. Show that in a 2×2 contingency table wherein the frequencies are $\frac{a \mid b}{c \mid d}$, the value of χ^2 calculated on the hypothesis of independence is given by

$$\chi^2 = \frac{(a + b + c + d)\,(ad - bc)^2}{(a + b)\,(c + d)\,(b + d)\,(a + c)}$$

(P.U., M.A. 1966, B.A. Hons. Econ., 1969)

We are given the following 2×2 contingency table:

Attribute	B_1	B_2	Total
A_1	a	b	$a+b$
A_2	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Under the hypothesis of independence, we calculate the expected frequencies e_{ij} $(i = j = 1, 2)$ as below:

$$e_{11} = \frac{(a+c)(a+b)}{a+b+c+d} \qquad e_{12} = \frac{(b+d)(a+b)}{a+b+c+d}$$

$$e_{21} = \frac{(a+c)(c+d)}{a+b+c+d} \qquad e_{22} = \frac{(b+d)(c+d)}{a+b+c+d}.$$

Hence $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

$$= \frac{\left[a - \frac{(a+c)(a+b)}{a+b+c+d}\right]^2}{\frac{(a+c)(a+b)}{a+b+c+d}} + \frac{\left[b - \frac{(b+d)(a+b)}{a+b+c+d}\right]^2}{\frac{(b+d)(a+b)}{a+b+c+d}}$$

$$+ \frac{\left[c - \frac{(a+c)(c+d)}{a+b+c+d}\right]^2}{\frac{(a+c)(a+b)}{a+b+c+d}} + \frac{\left[d - \frac{(b+d)(c+d)}{a+b+c+d}\right]^2}{\frac{(b+d)(c+d)}{a+b+c+d}}$$

$$= \frac{(ad-bc)^2}{a+b+c+d}\left[\frac{1}{(a+c)(a+b)} + \frac{1}{(b+d)(a+b)} + \frac{1}{(a+c)(c+d)} + \frac{1}{(b+d)(c+d)}\right]$$

$$= \frac{(ad-bc)^2}{a+b+c+d}\left[\frac{(a+b+c+d)^2}{(a+c)(b+d)(a+b)(c+d)}\right]$$

$$= \frac{(ad-bc)^2(a+b+c+d)}{(a+c)(b+d)(a+b)(c+d)}$$

This is a special case of χ^2 given by $\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ in a general contingency table. Hence this is a short-cut method of computing chi-square when a 2×2 contingency table is given.

Example 17.19. A random sample of 250 men and 250 women were polled as to their desire concerning the ownership of television. The following data resulted:

Classification	Men	Women	Total
Want television	80	120	200
Don't want television	170	130	300
Total	250	250	500

Test the hypothesis that desire to own a television set is independent of sex at the 0.05 level of significance.

(i) We state our null and alternative hypotheses as

H_0: The two variables of classification are independent, and

H_1: The two variables of classification are not independent.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

which for a 2×2 table becomes

$$\chi^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + c)(b + d)(a + b)(c + d)}$$

This statistic, if H_0 is true, has an approximate chi-square distribution with 1 degree of freedom.

(iv) **Computations.** Substituting the values in the formula, we get

$$\chi^2 = \frac{500\,(80 \times 130 - 120 \times 170)^2}{250 \times 250 \times 200 \times 300} = \frac{40}{3} = 13.33.$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(1)} = 3.84$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 13.33$ falls in the critical region, so we reject H_0 and conclude that desire to own a television set and sex are associated.

17.7.1. Co-efficient of Contingency for an r×c contingency table. The chi-square statistic shows only whether the sample data do or do not conform to the hypothesis. It does not tell anything about the strength of the association, which we sometimes desire to measure. For this purpose, Karl Pearson (1857–1936) has defined a co-efficient C, known as the *Pearson's co-efficient of mean-square contingency*, by the relation

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

where n indicates sample size.

This co-efficient measures the strength of the association or dependence of two variables of classification of a contingency table. When there is complete independence, $C=0$. When the two classifications are perfectly associated, $C = \sqrt{(k-1)/k}$, where k is the smaller of r and c. Thus C lies between 0 and $\sqrt{(k-1)/k}$. In a 2×3 contingency table, the maximum value of C is $\sqrt{(2-1)/2} = 0.707$. The larger the value of C, the stronger is the association or dependence.

Another measure, known as *Cramer's co-efficient of contingency*, is defined as

$$Q = \frac{\chi^2}{n\,(k-1)},$$

where n denotes the total sample size and k is the smaller of r and c. If the variables are completely independent, $Q=0$ and $Q=1$ when there is perfect relationship.

17.7.2. Yates' Correction for Continuity. In applying χ^2-approximation, we are required to combine the smaller frequencies (less than 5) with larger ones. But in case of two classes only, we cannot pool the smaller frequency into the larger one. For such a situation, Frank Yates in 1934 showed that the χ^2-approximation is markedly improved if we use the following formula:

$$\chi^2 = \sum_{i=1}^{2} \frac{\left(|o_i - e_i| - \frac{1}{2}\right)^2}{e_i}$$

This adjustment is known as *Yates' correction for continuity*. It should be used only when there is one degree of freedom and one frequency is small.

We also know that the distribution in a contingency table is necessarily discrete but the χ^2-distribution is essentially continuous. The approximation to χ^2 is just like the approximation of the discrete binomial distribution to the normal distribution, where a correction for continuity has already been discussed. So in a 2×2 contingency table with small frequencies, the cell frequencies are adjusted by adding $\frac{1}{2}$ to the smaller and subtracting $\frac{1}{2}$ from the larger frequencies and keeping the marginal totals unaltered. With this adjustment, the formula for χ^2 becomes

$$\chi^2 = \frac{n\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

where $n = a + b + c + d$.

This correction should be used if any expected frequency in a 2×2 table is less than 10.

17.7.3 An Exact Test for a 2×2 Contingency Table. When the frequencies in a 2×2 contingency table are fairly small, there will be some doubt about the adequacy of the chi-square approximation. An exact test, often called the *Fisher's exact test* (or sometimes, the *Fisher-Irwin Exact test*) was proposed almost simultaneously in the mid. 1930s by R.A. Fisher, J.O. Irwin and Frank Yates for testing the hypothesis of independence for such a 2×2 contingency table. Instead of comparing the observed and expected cell frequencies, the test is based on calculating the exact probabilities of the cell frequencies for all possible 2×2 tables obtained by varying the smallest cell frequency from observed value down to zero and with marginal frequencies fixed

It was shown that under the null hypothesis of no association between row and column classifications the exact probability of observing a table with frequencies

Attribute	B_1	B_2	Total
A_1	a	b	$a+b$
A_2	c	d	$c+d$
Total	$a+c$	$b+d$	n

where all the marginal frequencies are fixed, is given by

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

Supposing that d is the smallest frequency, we obtain other possible 2×2 tables by decreasing d by unity, changing other cell frequencies and continuing the process till d becomes zero. We calculate the probabilities of the observed table and other possible tables for all values of d from the observed value down to zero. Then the total probability, *i.e.* $P = p_0 + p_1 + \ldots + p_d$ corresponds to one tail of the distribution and is comparable with half the probability calculated from χ^2. Thus for a two-sided test, we double the probability so obtained, *i.e.* $\chi^2 \simeq 2P$. If $2P$ is not negligible, we reject our hypothesis of independence.

Example 17.20. Use the Fisher's exact test to test the hypothesis that inoculation is independent of immunity from attack among a population exposed to a certain disease, given the following data:

Classes	Not Inoculated	Inoculated	Total
Not attacked	3	5	8
Attacked	10	2	12
Total	13	7	20

We calculate the exact probability of observing the above contingency table, using

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

The smallest frequency is $d=2$. Therefore probability for $d=2$, denoted p_2, is

$$p_2 = \frac{8!\,12!\,13!\,7!}{3!\,10!\,5!\,2!\,20!} = 0.0477.$$

As the range of variation of d is from 0 to 2, therefore the other two possible 2×2 tables are

2	6	8
11	1	12
13	7	20

and

1	7	8
12	0	12
13	7	20

Thus the probabilities of these tables for $d = 1$ and $d = 0$ are

$$p_1 = \frac{8!\,12!\,13!\,7!}{2!\,11!\,6!\,1!\,20!} = 0.0043, \text{ and}$$

$$p_0 = \frac{8!\,12!\,13!\,7!}{1!\,12!\,7!\,0!\,20!} = 0.0001.$$

The total probability $P = 0.0477 + 0.0043 + 0.0001 = 0.0521$.

$\therefore \quad 2P = 2(0.052) = 0.104$, which is not negligible.

Hence we reject the hypothesis of independence.

17.8 TESTING HYPOTHESIS ABOUT EQUALITY OF SEVERAL PROPORTIONS

Contingency tables can also be used to test the hypothesis about the equality of several proportions (binomial parameters). Suppose we take $k(k > 2)$ independent random samples from k dichotomous (binomial) populations, where the ith sample contains n_i observations, of which X_i

show a certain characteristic, say A. Let the proportion of characteristic A be $p_i = \frac{X_i}{n_i}$ (cell proportion to column total) for $i = 1, 2, \ldots, k$. The data forming a $2 \times k$ contingency table are shown below:

Sample	1	2	...	i	...	k	Total
A	X_1	X_2	...	X_i	...	X_k	R
Not A	n_1-X_1	n_2-X_2	...	n_i-X_i	...	n_k-X_k	$n-R$
Total	n_1	n_2	...	n_i	...	n_k	n
Proportion of A	p_1	p_2	...	p_i	...	p_k	$P=R/n$

The hypotheses we wish to test, can be stated as

H_0: $p_1 = p_2 = \ldots = p_k$, *i.e.* the proportions p_i in row A are equal (or the k samples are drawn randomly from populations with the *same* proportions); and

H_1: At least two of the proportions are not equal.

To test the equality of proportions (or binomial parameters), we perform the test in exactly the same way as the test for independence in contingency tables. That is, we calculate the expected frequency for each of the observed frequencies by multiplying the row and column totals and dividing this by n, the overall total; and use the test-statistic

$$\chi^2 = \sum \frac{(\text{observed } f - \text{expected } f)^2}{\text{expected } f},$$

the summation being over all the $2k$ cells of the table. Under the null hypothesis, the test statistic is distributed approximately as a χ^2-distribution with $(2-1)(k-1) = (k-1)$ degrees of freedom.

A 2×2 contingency table may also be used to test the equality of two binomial proportions p_1 and p_2, *i.e.* the hypothesis $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ by performing the χ^2-test. An alternative formula is the z-test

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $\hat{p}_1 = \frac{a}{a+b}$ and $\hat{p}_2 = \frac{c}{c+d}$ are the estimates of p_1 and p_2, $\hat{p}$ is the combined estimate, $n_1 = a + b$ and $n_2 = c + d$. It is interesting to note that Z^2 is exactly $\chi^2_{(1)}$. (see Exercise 17.51)

Example 17.21 From the adult male population of seven large cities, random samples of sizes indicated below were taken, and the numbers of married and single men recorded.

City	A	B	C	D	E	F	G	Total
Married	133	164	155	106	153	123	146	980
Single	36	57	40	37	55	33	36	294
Total	169	221	195	143	208	156	182	1274

Test the hypothesis, at the 0.05 level of significance that the proportions of married men are the same in all the 7 cities.

(P.U., B.A. Hons., 1965)

The test is carried out as below:

(i) We set up our null and alternative hypotheses as

H_0: The proportions p_i of married men in all seven cities are the same, and

H_1: The proportions p_i of married men in at least two cities are not the same.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test statistic to be used is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

where the summation is over all the 14 cells of the table and which, if H_0 is true, has an approximate χ^2-distribution with (7–1), *i.e.* 6 *d.f.*

(iv) **Computations.** We calculate the expected frequencies under H_0, for each cell by the formula

$$e_{ij} = \frac{\text{(Row Total) (Column Total)}}{n}$$

Thus the expected frequency for married men in city A is $(169 \times 980)/1274 = 130$. The other expected frequencies are calculated in a similar way. The expected frequencies for the 14 cells in the table are given as follows:

City	A	B	C	D	E	F	G	Total
Married	130	170	150	110	160	120	140	980
Single	39	51	45	33	48	36	42	294
Total	169	221	195	143	208	156	182	1274

Now $\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$, the summation being over the 14 cells.

$$= \frac{(133-130)^2}{130} + \frac{(164-170)^2}{170} + \ldots + \frac{(33-36)^2}{36} + \frac{(36-42)^2}{42}$$

$$= 5.34$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05,(6)} = 12.59$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 5.34$ does not fall in the critical region, so we are unable to reject H_0. In other words, so far as this test is concerned, there is no evidence against the null hypothesis.

17.9 THE CHI-SQUARE TEST AS A TEST OF HOMOGENEITY

The chi-square statistic can also be used when the rows of a table which look like a contingency table, represent each a different sample or set of observations. The hypothesis to be tested is that two or more different random samples come from the same population or that samples are homogeneous as the word *homogeneous* in statistics is often used to indicate *the same* or *equal*. The χ^2-test applied in such a situation is called a *test of homogeneity*.

When there are two random samples, a simpler method proposed by Brandt and Snedecor is used to calculate the value of χ^2. This method is developed as below:

Suppose we draw two independent random samples of size n from a population and we wish to test whether the two samples are homogeneous. Let the values of the two samples be presented in the following $2 \times n$ table:

	1	2	...	i	...	n	Total
Sample I	a_1	a_2	...	a_i	...	a_n	A
Sample II	b_1	b_2	...	b_i	...	b_n	B
Total	c_1	c_2	...	c_i	...	c_n	N

As the samples are independent, therefore the expected values e_i in the first sample are given by $e_i = \frac{A\,c_i}{N}$, and the expected values e_i in the second sample are given by $e_i = \frac{B\,c_i}{N}$.

$$\text{Now} \quad \chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} = \sum_i \frac{o_i^2 + e_i^2 - 2\,e_i o_i}{e_i}$$

$$= \sum_i \frac{o_i^2}{e_i} + \sum_i e_i - 2\sum_i o_i = \sum_i \frac{o_i^2}{e_i} - N \quad [\because \Sigma e_i = \Sigma c_i = N]$$

$$\text{Thus} \quad \chi^2 = \sum_i \left[\frac{a_i^2}{\frac{Ac_i}{N}} + \frac{b_i^2}{\frac{Bc_i}{N}}\right] - N = N\sum_i \left[\frac{a_i^2}{Ac_i} + \frac{b_i^2}{Bc_i}\right] - N$$

$$= N\sum_i \left[\frac{a_i^2}{Ac_i} + \frac{c_i^2 + a_i^2 - 2c_i a_i}{Bc_i}\right] - N, \quad [\because b_i^2 = (c_i - a_i)^2]$$

$$= N\sum_i \frac{a_i^2}{c_i}\left(\frac{1}{A} + \frac{1}{B}\right) + \frac{N}{B}\sum_i c_i - \frac{2N}{B}\sum_i a_i - N$$

$$= N\sum \frac{a_i^2}{c_i}\left(\frac{N}{AB}\right) + \frac{N^2}{B} - \frac{2NA}{B} - N \quad [\because \Sigma c_i = N, \Sigma a_i = A]$$

$$= N\sum \frac{a_i^2}{c_i}\left(\frac{N}{AB}\right) + \frac{N^2 - N(A + B) - NA}{B} \quad (\text{put } N = A + B)$$

$$= \frac{N^2}{AB} \cdot \sum_i \frac{a_i^2}{c_i} - \frac{NA}{B}$$

$$= \frac{N^2}{AB}\left[\sum_i \frac{a_i^2}{c_i} - \frac{A^2}{N}\right],$$

where either row can be chosen as the a_i's.

This is known as Brandt-Snedecor formula with $n-1$ degrees of freedom. The rest of the procedure is the same.

Example 17.22. In a certain community, a random sample of 50 men and another sample of 50 women over 21 years of age were asked about their educational background, classified as junior high, senior high or college. The results are:

	Junior High	Senior High	College
Male	13	25	12
Female	23	20	7

Test whether the two samples are homogeneous in respect of educational levels. Let $\alpha = 0.05$.

(i) We formulate our null and alternative hypotheses as

H_0: The two samples (groups) are homogeneous, and

H_1: The two samples are not homogeneous.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the Brandt-Snedecor formula, *i.e.*

$$\chi^2 = \frac{N^2}{AB}\left[\sum_i \frac{a_i^2}{c_i} - \frac{A^2}{N}\right].$$

which has an approximate χ^2-distribution with $(n-1)$ degrees of freedom.

(iv) The critical region is $\chi^2 \geq \chi^2_{0.05,(2)} = 5.99 \quad (\because \; n=3)$

(v) **Computations.** We calculate the value of χ^2 as follows:

	Junior High	Senior High	College	Total
Male (a_i)	13	25	12	$50=A$
Female	23	20	7	$60=B$
Total (c_i)	36	45	19	$100=N$

Now $\chi^2 = \frac{N^2}{AB}\left[\sum_i \frac{a_i^2}{c_i} - \frac{A^2}{N}\right]$

$$= \frac{(100)^2}{(50)\,(50)}\left[\frac{(13)^2}{36} + \frac{(25)^2}{45} + \frac{(12)^2}{19} - \frac{(50)^2}{100}\right]$$

$$= 4\,(4.69 + 13.89 + 7.58 - 25) = 4 \times 1.16 = 4.64$$

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 4.64$ does not fall in the critical region, so we are unable to reject H_0. We may conclude that the two samples (groups) are homogeneous in respect of educational levels.

EXERCISES

17.1 (a) Define a chi-square random variable and its density function.

(b) Discuss the important properties of χ^2-distribution.

17.2 (a) Find the distribution of the sum of squares of n independent random variables, each of which is distributed normally with zero mean and unit variance.

(b) Show that the χ^2-distribution tends to normal distribution for large degrees of freedom. (P.U., M.Sc 1970)

17.3 Show that $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ is distributed as χ^2 with n degrees of freedom. Explain why $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2$ is distributed as χ^2 with $(n-1)$ degrees of freedom.

17.4 (a) Show that for large n, $\sqrt{2\chi^2}$ is normally distributed about mean $\sqrt{2n-1}$ and with variance unity.

(b) Compute $\chi^2_{0.05}$ for 40, 60 and 105 degrees of freedom by (i) Fisher's approximation (ii) Wilson-Hilferty approximation and compare the values with the table values.

17.5 (a) Explain how you determine a confidence interval estimate of σ^2 of a normal population.

(b) Given that X is normally distributed and given the sample values $\bar{x} = 42$, $S = 5$ and $n = 20$. Find the 98 percent confidence interval for σ^2.

17.6 (a) The following are the volumes, in deciliters, of 10 cans of peaches distributed by a certain company: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2 and 46.0. Find a 95% confidence interval for the variance of all such cans of peaches distributed by this company, assuming volume to be a normally distributed variable. (I.U.M.Sc., 1966)

(b) The contents of 10 similar containers of a commercial soap are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 and 9.8 litres. Find a 95% confidence interval for the variance of all such containers, assuming an approximately normal distribution.

17.7 (a) Given the following sample values from a normal population, find 96 per cent confidence limits for σ^2 based on combining these sample values properly. The sample variances are $S_1^2=25$, $S_2^2=36$, $S_3^2=16$ with $n_1=5$, $n_2=5$, $n_3=10$.

(b) Assume that the random variable X is $N(\mu, \sigma^2)$. Three random samples for X provide the following information:

Sample Size:	3, 5, 7
Sample Mean:	42, 45, 40
Sample Variance:	25, 16, 9

Obtain a 95% confidence interval for σ^2.

(P.U., B.A/B.Sc. 1988)

17.8 Suppose 10 samples of 9 values each, have variances as follows:

23.5, 30.6, 29.3, 27.5, 27.5, 26.3, 29.8, 30.7, 22.3, 26.5.

Obtain a pooled estimate of σ^2 and use it to find 90 per cent confidence limits for σ^2. (M.Sc., P.U., 1989, I.U., 1993)

17.9 (a) Explain how you would test the hypothesis about variance of a normal population.

(b) A sample of 9 parts produced by a certain production process are measured as 5, 7, 2, 4, 8, 9, 8, 6 and 5 inches respectively. Test the hypothesis that the process has the variance equal to 4 (inches)^2 at the 5 per cent level. (P.U., B.A/B.Sc. 1983)

17.10 (a) A sample of 25 observations has $S^2 = 12.6$, would you accept or reject at the 5% level of significance the hypothesis that $\sigma^2=20$? Also compute a 90% confidence interval for σ^2.

(P.U., B.A/B.Sc. 1992)

(b) A standard examination has been given for several years with $\mu=70$ and $\sigma^2=9$. A school using this examination for the first time, gave it to a group of 25 students who obtained a mean $\bar{x}=71$ and a variance of $S^2=12$. Is there reason to doubt that the score of all students in the school would have a variance of 9?

17.11 (a) A random sample of 15 have the following values:

10.21, 9.72, 10.13, 8.89, 10.20, 9.65, 10.02, 10.00, 9.45, 10.11, 8.97, 10.21, 9.36, 9.55, 10.23. Test the hypothesis that $\sigma^2 = 0.12$ against $\sigma^2 > 0.12$ at (i) 5% and (ii) 1% level of significance. (P.U., B.A./B.Sc. 1989)

(b) A height distribution has a variance σ^2 of $(2.792)^2$. Do the following 10 values, selected at random, have a greater variance than what is expected? 67.50, 70.75, 72.00, 63.25, 65.25, 68.75, 69.25, 68.50, 66.50 and 64.75. (I.U., M.Sc., 1986)

17.12 (a) The weights of a random sample of 10 boxes of a particular brand of cereal are 14.2, 13.7, 14.1, 14.3, 14.1, 13.8, 14.4, 14.8, 13.9 and 14.3. Test the hypothesis that $H_0 : \sigma^2 = 0.02$ against the alternative $H_1 : \sigma^2 < 0.02$, using a 0.01 level of significance.

(b) A manufacturer of car batteries claims that the life of his batteries have a standard deviation equal to 0.9 years. If a random sample of 10 of these batteries have a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ years? Use a 0.05 level of significance.

17.13 (a) Describe how you would test the equality of $k(k>2)$ variances of normal populations. (P.U., M.Sc. 1989)

(b) Show that the estimates 3.8, 4.4, 8.1, 6.1 and 9.4 of the population variance, based on 5, 8, 6, 7 and 4 d.f. respectively, may be regarded as homogeneous.

17.14 (a) Describe Bartlett's test for homogeneity of variances.

(b) Three independent samples gave the following results:

Size	Observations
5	34, 40, 47, 60, 84.
9	40, 59, 60, 67, 86, 92, 95, 98, 108.
3	46, 93, 100.

Use Bartlett's test to test the hypothesis of equal variances. Let $\alpha = 0.05$. (I.U., M.Sc. 1994)

17.15 (a) Six samples of size 5 each, have the variances 10.4, 13.8, 11.7, 19.3, 16.4 and 15.8. Test the hypothesis of homogeneity of variances by Bartlett's test. (P.U., M.Sc. 1988)

(b) For the data given below;

Sample 1: 4, 7, 6, 6; Sample 2: 5, 1, 3, 5, 3, 4;
Sample 3: 3, 8, 6, 8, 9, 5.

Use Bartlett's test to test the hypothesis that the variances of three populations are equal. ($\alpha = 0.05$) (P.U., M.Sc. 19[illegible])

17.16 A random selection of nine individuals was made at each of 10 out patients clinics across the country. Pulse rates were recorded and variances of 10 samples were 24, 31, 29, 28, 28, 26, 30, 31, 22 and 26. Test the hypothesis that the variances are equal. (I.U., M.Sc 1992)

17.17 n objects are classified independently into k-groups, the probability for the ith group being p_i, and the number falling in it being n_i $(i = 1, 2, ..., k)$. What is the joint distribution of the n_i? Show that as n tends to infinity, the distribution of $\sum \frac{(n_i - np_i)^2}{np_i}$ tends to a χ^2-distribution with $(k-1)$ degrees of freedom.

(P.U., M.A. Stat., 1964)

17.18 (a) Describe three distinct uses of the chi-square distribution. According to a genetic model, the proportion in three groups should be $p^2 : 2pq : q^2$ where $p + q = 1$. Are the data consistent with the sample 9, 51, 45 if $p = 0.4$?

(P.U., B.A./B.Sc. 1992)

(b) The proportion of individuals possessing the four blood types should be in the proportion $q^2 : p^2 + 2pq : r^2 + 2qr; 2pr$, where $p+q+r=1$. Given the observed frequencies 180, 360, 132, 98, test for compatibility with $p=0.4$, $q=0.4$ and $r=0.2$.

(M.Sc; P.U; 1967; I.U; 1995)

17.19 (a) Genetic theory states that children having one parent of blood type M and the other of blood type N will always be of one of the three types M, MN, N and that the proportion of these types will on the average be 1 : 2 : 1. A report states, "Of 162 children having one M parent and one N parent, 28.4% were found to be of type M, 42% of type MN and the remainder of the type N. The low value of χ^2 demonstrates the truth of the genetic theory." Calculate the value of χ^2, make the appropriate test of significance and comment on the conclusions quoted.

(P.U., B.A./B.Sc 1973)

(b) A machine is supposed to mix peanuts, hazelnuts, cashews, and pecans in the ratio 5 : 2 : 2 : 1. A can containing 500 of these mixed nuts was found to have 269 peanuts, 112 hazelnuts, 74 cashews, and 45 pecans. At the 0.05 level of significance, test the hypothesis that the machine is mixing the nuts in the supposed ratio.

(P.U., B.A/B.Sc. 1988)

17.20 A thousand individuals were classified according to sex and according to whether or not they were colour-blind as follows.

Classes	Male	Female
Normal	452	494
Colour Blind	38	16

According to the genetic theory, the frequencies in four classes should be

45%	50%
4%	1%

Test the hypothesis that the data are consistent with theory.
(P.U., B.A/B.Sc. 1987)

17.21 (a) In 200 tosses of a coin, 115 heads and 85 tails were observed. Test the hypothesis that the coin is fair, using a level of significance of 0.05.

(b) In 360 tosses of a pair of dice, 74 "sevens" and 24 "elevens" are observed. Using a 0.05 level of significance, test the hypothesis that the dice are fair.

17.22 (a) The sex distribution of 98 births reported in a newspaper was 52 boys and 46 girls. Is this consistent with an equal sex division in the population? Use the χ^2-approximation and the normal approximation. (I.U., M.Sc. 1993)

(b) In a certain disease with 40% mortality, of 10 patients given a certain treatment only one dies. Is the treatment effective at 5% level of significance? (P.U., M.Sc. 1989)

17.23 (a) The following table records the observed number of briths at a hospital in four consecutive quarterly periods.

Quarter	Jan.–Mar.,	Apr.–Jun.,	Jul.–Sep.,	Oct.–Dec.
Number of births	110	57	53	80

It is hypothesized that twice as many babies are born during the Jan.–Mar. quarter than are born in any of the other three quarters. At $\alpha = 0.10$, test if these data strongly contradict the stated hypothesis.

(b) The grades in a statistics course were as follows:

Grade	A	B	C	D	E
f	14	18	32	20	16

Test the hypothesis, at the 0.05 level of significance, that the distribution of grade uniform.

17.24 (a) A random number table of 250 digits showed the following distribution of the digits 0, 1, 2, ..., 9.

Digit	: 0	1	2	3	4	5	6	7	8	9
Frequency:	17	31	29	18	14	20	35	30	20	36

Test the hypothesis, at 0.05 level of significance, that the digits were distributed in equal numbers in the table.

(b) The following distribution shows the number of deaths from overdoses of narcotics. Use chi-square statistic to test the hypothesis that equal number die in all age-groups.

Age	15–19,	20–24,	25–29,	30–34,	35–39,	40–44,	45–49
No. of deaths	40	35	32	10	13	13	4

(I.U., M.Sc. 1986)

17.25 (a) A die is tossed 180 times with the following results:

x	1	2	3	4	5	6
f	20	36	46	35	21	22

Is this a balanced die? Use $\alpha = 0.01$. (P.U., B.A./B.Sc. 1990)

(b) The following figures show the number of births in an area over a year by months of occurrence.

January	50759	May	51371	September	52162
February	46472	June	47388	October	50824
March	51419	July	49995	November	47768
April	49670	August	51043	December	51129

Use the χ^2 test to discuss whether there is any seasonality in births revealed by these data. (P.U., B.A.,/B.Sc. 1963)

17.26 (a) Discuss the χ^2-tests of goodness-of-fit. What are the assumptions in the application of these tests to practical problems?

(b) Records taken of the number of male and female births in 800 families having four children, are as follows:

No. of male births	0	1	2	3	4
Families	32	178	290	236	64

Test whether the data are consistent with the hypothesis that the binomial law holds and that the chance of a male birth is equal to that of a female birth, that is $p=q=\frac{1}{2}$.

(P.U., B.A/B.Sc. 1976, 93)

17.27 (a) What is the purpose of the goodness-of-fit-test? Describe three situations where this test might be used appropriately.

(b) Three six-sided dice were thrown 648 times and the number of 5's or 6's noted at each throw.

Number of 5's or 6's	0	1	2	3
Number of throws	179	298	141	30

Test the hypothesis that the data conform to binomial distribution with $p=\frac{1}{3}$ and $n=3$. Let $\alpha=0.05$.

(P.U., B.A./B.Sc. 1983)

17.28 Twelve dice were thrown 4096 times and a throw of 6 was reckoned as a success. The observed frequencies were as given below:

No. of success	0	1	2	3	4	5	6	7&over	Total
Frequency	447	1145	1181	796	380	115	24	8	4096

Find the value of χ^2 on the hypothesis that the dice were unbiased and hence show that the data are consistent with the hypothesis so far as the χ^2-test is concerned. (P.U., B.A./B.Sc. 1975)

17.29 (a) Suppose that 6 coins are tossed simultaneously 640 times and the following frequency distribution is observed:

No. of heads	0	1	2	3	4	5	6
Frequency	13	70	137	210	145	56	9

Test the null hypothesis that the coins are well-balanced. Use $\alpha=0.01$.

(P.U., B.A./B.Sc. 1992)

(b) When the first proof of a book containing 250 pages was read, the following distribution of printing mistakes were found:

No. of mistakes per page	0	1	2	3	4	5
Frequency	139	76	28	4	2	1

Fit an appropriate distribution to the data and test the goodness-of-fit.

17.30 Given the data

x	0	1	2	3	4	5
f	40	32	18	8	2	0

Fit a Poisson distribution and test the goodness-of-fit.

(P.U., B.A/B.Sc. 1982)

17.31 Test whether the data given below may be regarded as conforming to a Poisson Distribution?

x	0	1	2	3	4	5	6	7
f	305	365	210	80	28	9	2	1

17.32 The wages of 1,000 employees range from Rs. 4.50 to Rs. 19.50. They are grouped in 15 classes with a common class interval of Re. 1, and the class-frequencies, from the lowest class to the highest, are 6, 17, 35, 48, 65, 90, 131, 173, 155, 117, 75, 52, 21, 9, 6. Fit a normal distribution and apply the chi-square goodness-of-fit test.

17.33 The heights of 200 employees are distributed as follows:

Heights	58–60,	61–63,	64–66,	67–69,	70–72,	73–75,	76–78
Frequency	9	20	45	55	43	17	11

Test whether a normal distribution gives a satisfactory fit to the data at $\alpha = 0.05$.

17.34 (a) Define attributes and ultimate class-frequencies.

(b) Compute the ultimate class-frequencies from the data given below:

$N = 500$, $(ABC) = 240$, $(\alpha\beta\gamma) = 25$, $(A\beta) = 18$, $(B\gamma) = 98$,

$(\gamma) = 125$, $(A)-(\alpha) = 80$ and $(B) - (\beta) = 200$.

17.35 In a class of statistics, there were 300 students. Their results in the First Terminal, Second Terminal and the Annual Examinations were as follows:

"120 passed the first terminal, 112 passed the second terminal and 144 passed the annual examination, 38 passed all the three, 39 failed in all the three, 44 passed the first two and failed in the annual, 63 failed in the first two but passed the annual examination."

Find how many students passed atleast two examinations?

17.36 (a) Define the term *Consistence*. State the necessary and sufficient conditions for the consistence of a set of independent class frequencies. (P.U., B.A./B.Sc. 1978)

(b) Certain data obtained from a study of a group of 1000 subscribers to a certain magazine relating to their sex, marital status and education were reported as follows: 312 males, 470 married, 525 college graduates, 42 male college graduates, 147 married college graduates, 86 married males and 25 married male college graduates.

Show that the numbers reported in the various groups are not consistent.

17.37 When are two attributes said to be independent, positively associated or negatively associated? Discuss the association when $(AB)=256$, $(\alpha B) = 768$, $(A\beta)=48$, $(\alpha\beta) = 144$. (P.U. B.A/B.Sc. 1977)

17.38 (a) When are two attributes independent? Describe different froms of the Criterion of Independence.

(b) Show as briefly as possible whether A and B are independent, positively associated or negatively associated in the following cases:

(i) $N = 5000$, $(A) = 2350$, $(B) = 3100$, $(AB) = 1600$;

(ii) $(A) = 490$, $(AB) = 294$, $(\alpha) = 570$, $(\alpha B) = 380$;

(iii) $(AB) = 256$, $(\alpha B) = 768$, $(A\beta) = 48$, $(\alpha\beta)=144$.

(P.U., M.A. Econ. 1974)

17.39 (a) What is meant by Association of Attributes? How is it measured?

(b) Prove that $Q = \dfrac{2Y}{1 + Y^2}$, where Q and Y are the co-efficient of association and colligation respectively.

Solution. (b) By definition,

$$Y = \frac{1 - \sqrt{\dfrac{(\alpha B)\,(A\beta)}{(AB)\,(\alpha\beta)}}}{1 + \sqrt{\dfrac{(\alpha B)\,(A\beta)}{(AB)\,(\alpha\beta)}}} = \frac{\sqrt{(AB)\,(\alpha\beta)} - \sqrt{(\alpha B)\,(A\beta)}}{\sqrt{(AB)\,(\alpha\beta)} + \sqrt{(\alpha B)\,(A\beta)}}$$

Inverting, we get

$$\frac{1}{Y} = \frac{\sqrt{(AB)(\alpha\beta)} + \sqrt{(\alpha B)(A\beta)}}{\sqrt{(AB)(\alpha\beta)} - \sqrt{(\alpha B)(A\beta)}}$$

or $$\frac{1+Y}{1-Y} = \frac{\sqrt{(AB)(\alpha\beta)}}{\sqrt{(\alpha B)(A\beta)}}$$

Squaring, we obtain

$$\frac{(1+Y)^2}{(1-Y)^2} = \frac{(AB)(\alpha\beta)}{(\alpha B)(A\beta)}$$

or $$\frac{(1+Y)^2 - (1-Y)^2}{(1+Y)^2 + (1-Y)^2} = \frac{(AB)(\alpha\beta) - (\alpha B)(A\beta)}{(AB)(\alpha\beta) + (\alpha B)(A\beta)}$$

or $$\frac{4Y}{2(1+Y^2)} = Q$$

Hence $Q = \dfrac{2Y}{1+Y^2}$.

17.40 (a) Define a Contingency Table. How do you determine the number of degrees of freedom in an $r \times c$ contingency table?

(b) What is the chi-square test for independence? Describe the situations where this test might be used appropriately.

17.41 (a) State the important applications of χ^2-statistic.

(b) The following table shows the number of recruits taking (i) a preliminary and (ii) a final test in car driving. Use the χ^2 test to discuss whether there is any association between the results of preliminary and those of the final test.

Categories		Preliminary Pass	Preliminary Fail	Total
Final	Pass	605	135	740
	Fail	195	65	260
Total		800	200	1000

17.42 (a) Test the association between injection against typhoid and exemption from attack from the following contingency table:

Classes	Attacked	Not-attacked	Total
Inoculated	528	25	553
Not-inoculated	790	175	965
Total	1318	200	1518

(b) The following table gives the census data of orchards. Test the hypothesis that the two variables of classification are independent.

Classes	Shaded	Unshaded	Total
High Yielders	350	205	555
Low Yielders	250	195	445
Total	600	400	1000

17.43 Find χ^2 and test whether the two attributes are independent. Let $\alpha=0.05$.

Attributes	A_1	A_2	A_3	Total
B_1	215	325	60	600
B_2	135	175	90	400
Total	350	500	150	1000

(P.U., B.A./B.Sc. 1960; Opt. 1969; 72)

17.44 (a) Test the null hypothesis that the two variables of classification are independent, using a 0.05 level of significance.

Classes	A_1	A_2	A_3
B_1	337	291	302
B_2	225	207	238

(P.U., B.A/B.Sc. 1961; M.A. Econ. 1965)

(b) The following is percentage distribution by income level and ownership of a random sample of 400 families in the city of Lahore.

	Income Level		
	Less than Rs. 12,000	Rs. 12,000 to Rs. 60,000	More than Rs. 60,000
Home Owner	5%	35%	10%
Renter	15%	25%	10%

Test the hypothesis that the home ownership is independent of the family income level, using 1% level of significance.

(P.U., M.Sc. 19[illegible])

17.45 A certain drug is claimed to be effective in curing colds. In an experiment on 164 people with colds, half of them were given

drug and half of them were given sugar pills. The patients' reactions to the treatment are recorded in the following table. Test the hypothesis that the drug is no better than sugar pills for curing colds. Let $\alpha = 0.05$.

Category	Helped	Harmed	No Effect
Drug	52	10	20
Sugar	44	12	26

17.46 A thousand households are taken at random and divided into three groups A, B and C, according to the total monthly income. The following table shows the numbers in each group having a colour television receiver, a black and white receiver, or no television at all.

	A	B	C
Colour television	56	51	93
Black and white	118	207	375
None	26	42	32

Test the hypothesis that there is no association between total income and television ownership.

17.47 Calculate χ^2 (chi-square) from the following contingency table of attributes and test for independence at $\alpha = 0.01$.

Attributes	A_1	A_2	A_3
B_1	44	22	4
B_2	265	257	178
B_3	41	91	98

(C.S.S., 1965)

17.48 Gilby classified 1725 school children according to intelligence and apparent family economic level. A condensed classification follows:

Classes	Dull	Intelligent	Very Capable
Very well clothed	81	322	233
Well clothed	141	457	153
Poorly clothed	127	163	48

Test the null hypothesis of independence of the two classifications at the 0.01 level of significance.

17.49 An insurance company wants to determine whether a policy holder's age is independent of whether or not the policy holder has filed an accident claim. A study of 1000 of its policyholders gave the following results:

Age →	Under 25,	25–40,	40–55,	Over 55
Reported claim	93	72	53	63
No claim	115	155	265	184

Test the hypothesis at 0.05 level of significance that the claim status is independent of the policyholder's age.

17.50 A random sample of 200 married men, all retired, were classified according to education and number of children.

Education	Number of children		
	0–1	2–3	Over 3
Elementary	14	37	32
Secondary	19	42	17
College	12	17	10

Test the hypothesis, at the 0.05 level of significance, that the size of a family is independent of the level of education attained by the father. (I.U. M.Sc. 1985, 91)

17.51 Show that a chi-squared test for a 2×2 contingency table is equivalent to testing the difference between the two proportions, using the normal approximation. (I.U., M.Sc. 1993, 94)

Solution. Let the 2×2 contingency table be

	I	II	Total
A	a	b	r_1
B	c	d	r_2
Total	n_1	n_2	n

where

$$n_1 = a + c,\ n_2 = b + d,\ r_1 = a + b,\ r_2 = c + d \text{ and } n = a + b + c + d$$

Suppose the two proportions are denoted by $\hat{p}_1$ and $\hat{p}_2$ and the timate of the population proportion is denoted by $\hat{p}$. Then

$$\hat{p}_1 = \frac{a}{n_1},\ \hat{p}_2 = \frac{b}{n_2},\ 1 - \hat{p}_1 = \frac{c}{n_1},\ 1 - \hat{p}_2 = \frac{d}{n_2};$$

$\hat{p} = \frac{r_1}{n}$. $1 - \hat{p} = \hat{q} = \frac{r_2}{n}$, so that $a = n_1\hat{p}_1$, $b = n_2\hat{p}_2$,

$c = n_1(1 - \hat{p}_1)$, $d = n_2(1 - \hat{p}_2)$, $r_1 = n\hat{p}$ and $r_2 = n\hat{q}$.

Now
$$\chi^2 = \frac{n(ad - bc)^2}{n_1 n_2 r_1 r_2}$$
$$= \frac{n[n_1 n_2 \hat{p}_1(1 - \hat{p}_2) - n_1 n_2 \hat{p}_2(1 - \hat{p}_1)]^2}{n_1 \cdot n_2 \cdot n\hat{p} \cdot n\hat{q}}$$
$$= \frac{n_1 n_2[\hat{p}_1(1 - \hat{p}_2) - \hat{p}_2(1 - \hat{p}_1)]^2}{n\hat{q}\hat{p}} = \frac{n_1 n_2(\hat{p}_1 - \hat{p}_2)^2}{n\hat{q}\hat{p}}$$
$$= \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{ where } n = n_1 + n_2$$
$$= Z^2, \text{ where } Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

That is, χ^2 is a square of Z-statistic for testing the equality of two proportions. Hence, in a 2×2 contingency table, the two test-procedure are equivalent.

17.52 Given the following contingency table for hair colour and eye-colour, calculate the co-efficient of contingency and interpret the result.

Eye Colour	Hair Colour		
	Fair	Grey	Brown
Blue	69	49	28
Black	91	56	27
Dark Blue	57	24	33

(P.U., B.A./B.Sc. 1984)

17.53 (a) Explain the use of Yates' correction for continuity.

(b) Out of a group of 320 people exposed to infection, 255 had not been immunized, and of these 95 contracted the disease. Of those who had been immunized, 15 were infected. Does it seem that treatment gave any protection against infection? What is the difference in the significance of the result of the χ^2 test according as Yates' correction is or is not, applied?

(P.U., B.A/B.Sc. 1983)

17.54 (a) Prove that χ^2 for the table

$a+\frac{1}{2}$	$b-\frac{1}{2}$
$c-\frac{1}{2}$	$d+\frac{1}{2}$

if given by

$$\chi^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)},$$

if $ad < bc$ and where $n = a+b+c+d$.

(b) A random sample of 30 adults is classified according to sex and the number of hours they watch television during a week.

Classes	Male	Female
Over 25 hours	5	9
Under 25 hours	9	7

Using a 0.01 level of significance, test the hypothesis that a person's sex and time watching television are independent.

(P.U., B.A/B.Sc. 1986-S)

17.55 (a) Describe Fisher's exact test for a 2×2 contingency table.

(b) Suppose that a number of patients were treated for cancer with results as in the following table:

Toxity Present	Tumer Regression	
	Yes	No
Yes	5	2
No	1	7

Use Fisher's exact test to test the independence.

(P.U., M.Sc. 1984, 86)

(c) Deaths in 6 months after fractured neck of femur in a specialised orthopaedic ward (A) and a general ward (B) are given below:

		Ward	
		A	B
Deaths	Yes	2	6.
	No	18	14

Test the hypothesis of independence by using the Fisher-Irwin exact test.

17.56 (a) The following data are intended to show dependence of brittleness in polyethylene bars on the duration of heat treatment at a particular phase of the manufacturing process.

	Brittle	Tough
Treatment 1	2	8
Treatment 2	6	3

Use the Fisher's exact test to test the null hypothesis that the brittleness of polyethylene bars does not vary with the two heat treatments.

(P.U., M.Sc. 1989)

(b) Use the Fisher's exact test to test the hypothesis that inoculation is independent of immunity from attack among a population exposed to a certain disease from the following data:

	Inoculated	Not-inoculated
Attacked	9	2
Not attacked	7	6

(I.U., M.Sc. 1995)

17.57 (a) In a study to determine whether or not the proportions of defectives produced by workers was the same for the day, evening, or night shift worked, the following data were collected:

	Shift		
	Day	Evening	Night
Defectives	45	55	70
Non-defectives	905	890	870

Test the hypothesis, at the 0.025 level of significance, that the proportion of defectives is the same for all three shifts.

(b) There are 5 classes, each having 50 students. The results of these 5 classes are given below:

Class	1	2	3	4	5	Total
Pass	42	45	43	45	45	220
Failure	8	5	7	5	5	30
Total	50	50	50	50	50	250

Test the null hypothesis H_0 : 4 pass : 1 failure, using totals.

17.58 The observed row frequencies in a $2 \times k$ contingency table are $a_1, a_2, \ldots, a_k$ and $b_1, b_2, \ldots, b_k$. Using the notation $n_i = a_i + b_i$ for all i $(=1, 2, \ldots, k)$, $\sum a_i = A$, $\sum b_i = B$ and $A + B = N$, show that the χ^2 for testing independence can be expressed as

$$\chi^2 = \frac{N}{AB}\left[\sum_{i=1}^{k}\frac{a_i^2}{n_i} - \frac{A^2}{N}\right].$$

What are the degrees of freedom for the chi-square?

Use this formula to find χ^2 from the following data:

No. of Rows of Kernels	8	10	12	14	Total
Present	18	37	27	0	82
Absent	15	26	43	4	88
Total	33	63	70	4	170

17.59 (a) Prove the Brandt-Snedecor formula for χ^2. (I.U., M.Sc. 1986)

(b) Two groups of freshmen applying to enter a university took the same college aptitude test. The groups (A and B) differed in the type of high school education they had experienced. The frequency distributions of scores for the two groups were as follows:

Score	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–99
Group A	71	68	66	47	51	39	43	39	33	18
Group B	22	8	14	12	3	13	3	14	12	10

Calculate the value of χ^2 and determine whether there is a significant difference in college aptitude test between the groups.

(I.U., M.Sc. 1988)

17.60 In an experiment on the effectiveness of a teaching machine, a machine-instructed group of students was compared with a teacher-instructed group on an achievement test. The following scores were obtained:

	40–49	50–59	60–69	70–79	80–89	90–99	Total
Teacher-instructed	21	40	55	38	10	2	166
Machine-instructed	18	35	42	46	19	4	164
Total	39	75	97	84	29	6	330

Calculate χ^2 and determine whether there is a significant difference in achievements of the two groups.

(M.Sc. P.U., 1989; I.U. 19[illegible])

18

The Student's *t*-Distribution and Statistical Inference

18.1 INTRODUCTION

Earlier it was shown that if we take a random sample from a normal population with mean μ and variance σ^2, the sample mean $\bar{X}$ is normally distributed with mean μ and variance σ^2/n; that is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is a standard normal variable even if n, the sample size is small provided that σ^2 is known. But in actual practice, the population variance σ^2 is usually *not* known and is estimated from the sample data. So when sample size n, is small ($n<30$) and σ^2 is replaced with its unbiased estimate s^2, the statistic

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

is no longer normally distributed. This being the case, how can we test hypotheses about the means and obtain the confidence intervals? This problem was solved by William Sealy Gosset (1876–1937) who in 1908 published a paper titled, "The Probable Error of a Mean", in which he discussed the sampling distribution of $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$. W.S. Gosset, an employee of the Messrs. Guinness, a Dublin brewery, published a series of scientific papers under the pen name "Student", as he was not permitted under the company policy to publish his research under his real name. The distribution of the statistic $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$, after his pen name (Student), is known as *Student's t-distribution*. Although, Gosset

did not complete the mathematical proof, but he reached the correct decision about the sampling distribution of the statistic t. A rigorous proof was provided by Sir R.A. Fisher (1890–1962) in 1925. The t-distribution has a single parameter known as the number of degrees of freedom. The t-distribution is of great importance in the so called *small sample tests* and is profoundly used in statistical inference.

18.2 THE STUDENT'S t-DISTRIBUTION

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and unknown variance σ^2 and let $\bar{x} = \frac{1}{n}\sum x_i$ and $s^2 = \frac{1}{n-1}\sum(x_i-\bar{x})^2$, which is the unbiased estimate of σ^2. Then the sampling distribution of the statistic

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

is called *Student's t-distribution* with $(n-1)$ degrees of freedom. It is interesting to note that the statistic t may be written as

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}\cdot\frac{s}{\sigma}} = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{\sigma}\right)^2}} = \frac{Z}{\sqrt{\frac{U}{n-1}}}$$

where $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is a standard normal random variable and

$U = \sum\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 = \frac{(n-1)\,s^2}{\sigma^2}$ is a χ^2-random variable with $(n-1)$ degrees of freedom. Thus the t-statistic is the quotient of a standard normal variable and the square root of a chi-square random variable divided by its degrees of freedom.

To find the distribution of t, we proceed as follows:

Let Z be a standard normal random variable and U, a χ^2-random variable with v (Greek letter *nu*) degrees of freedom. If Z and U are independent, their joint density will be

$$f(z, u) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}\cdot\frac{1}{2^{v/2}\,\Gamma\left(\frac{v}{2}\right)}u^{(v/2)-1}e^{-u/2},$$

for $-\infty < z < \infty$

Let us put $t = \dfrac{z}{\sqrt{u/v}}$ and $w = u$, so that $z = t\sqrt{w/v}$ and $u = w$.

Then the change of variables technique tells us that

$$f_{x,y}(x, y) = f_{u,v}(x, y) \mid J \mid,$$

where J is the Jacobian of the transformation. Thus

$$J = \begin{vmatrix} \dfrac{\partial z}{\partial t} & \dfrac{\partial z}{\partial w} \\ \dfrac{\partial u}{\partial t} & \dfrac{\partial u}{\partial w} \end{vmatrix} = \begin{vmatrix} \sqrt{\dfrac{w}{v}} & \dfrac{t}{2\sqrt{wv}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{w}{v}}$$

Substituting these values, we obtain the joint density of t and w as

$$f(t, w) = \frac{1}{2^{v/2}\sqrt{2\pi}\,\Gamma\left(\frac{v}{2}\right)} w^{(v/2)-1} e^{-(w+t^2w/v)/2} \sqrt{\frac{w}{v}},$$

for $-\infty < t < \infty$
$0 < w < \infty$

$$= \frac{1}{2^{v/2}\sqrt{2v\pi}\,\Gamma\left(\frac{v}{2}\right)} w^{(v-1)/2} e^{-(w/2)(1+t^2/v)}$$

Integrating out w, we find the density of t as

$$f_v(t) = \int_0^\infty f(t, w)\, dw$$

$$= \frac{1}{2^{v/2}\sqrt{2v\pi}\,\Gamma\left(\frac{v}{2}\right)} \int_0^\infty w^{(v-1)/2} e^{-(w/2)(1+t^2/v)}\, dw$$

Let $\quad y = \dfrac{w}{2}\left(1 + \dfrac{t^2}{v}\right).$

Then $\quad w = 2y\left(1 + \dfrac{t^2}{v}\right)^{-1}$ and $dw = 2\left(1 + \dfrac{t^2}{v}\right)^{-1} dy.$

Substitution gives

$$f_v(t) = \frac{1}{2^{v/2}\sqrt{2v\pi}\,\Gamma\left(\frac{v}{2}\right)} \int_0^\infty \left(\frac{2y}{1 + t^2/v}\right)^{(v-1)/2} e^{-y} \frac{2}{1 + t^2/v}\, dy$$

$$= \frac{1}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} \int_0^\infty y^{(v-1)/2} e^{-y}\, dy$$

$$= \frac{1}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi}}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}\Gamma\left(\frac{v+1}{2}\right)$$

$$= \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi}}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}, \text{ for } -\infty < t < \infty$$

$$= \frac{1}{\sqrt{v}\,\beta\left(\frac{1}{2},\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}, \qquad [\because \sqrt{\pi} = \Gamma(1/2)]$$

which is the so called *Student's t-distribution* with v degrees of freedom.

18.2.1. Properties of Student's t-distribution. The t-distribution has the following properties.

(i) The t-distribution is continuous and symmetric about the value $t=0$, ranging from $-\infty$ to ∞.

(ii) The mean of the t-distribution is $\mu=0$, when $v\geq 2$. The mean is undefined for $v=1$. The variance of the t-distribution is $\sigma^2=\frac{v}{v-2}$ for $v > 2$, and the variance for $v \leq 2$ does not exist. The variance is greater than 1 and approaches 1 as the degrees of freedom increase.

By definition, $t = \frac{Z}{\sqrt{U/v}}$, where Z is $N(0, 1)$, U is a $\chi^2_{(v)}$ random variable and Z and U are independent.

Now $\mu = E(t) = E\left[\frac{Z}{\sqrt{U/v}}\right]$

$$= E\left[Z \cdot \sqrt{v/U}\right] = E[Z]\,E\left[\sqrt{v/U}\right] = 0 \quad [\because E(Z) = 0]$$

and $\sigma^2 = \text{Var}(t) = E(t^2)$ $\quad [\because E(t) = 0]$

$$= E\left(\frac{Z^2}{U/v}\right) = E\left(Z^2 \cdot \frac{v}{U}\right) = E(Z^2)\,E\left(\frac{v}{U}\right)$$

Since Z^2 is a χ^2-random variable with 1 degree of freedom, therefore $E(Z^2) = 1$.

And $$E\left(\frac{v}{U}\right) = \int_0^\infty \frac{v}{u}\,\frac{1}{2^{v/2}\,\Gamma\left(\frac{v}{2}\right)}\,u^{(v/2)-1}\,e^{-u/2}\,du$$

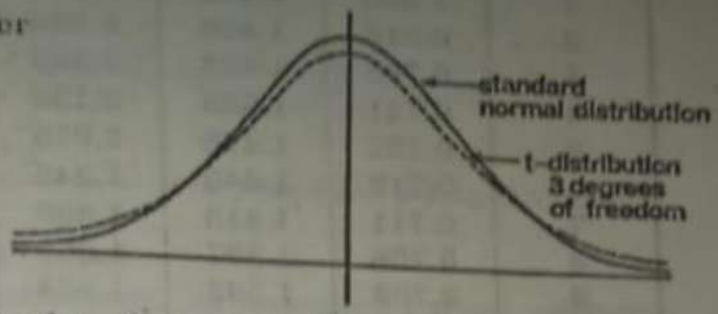
This exists if $v > 2$ and thus is given by

$$\frac{v\,\Gamma[(v-2)/2]}{2\,\Gamma(v/2)} = \frac{v\,\Gamma[(v-2)/2]}{2\,[(v-2)/2]\,\Gamma[(v-2)/2]} = \frac{v}{v-2}.$$

Hence the variance of the t-distribution is $\frac{\nu}{\nu - 2}$.

(iii) The t-distribution is *unimodal* with a bell shape. The density of the distribution reaches its maximum at $t = 0$ and thus the mode of the t-distribution is $t = 0$. The median is also equal to zero.

(iv) The t-distribution, for small values of ν, is flatter than the standard normal distribution which means that the t-distribution is more spread out in the tails than is the standard normal dis tribution.



(v) The shape of the t-distribution changes as the number of degrees of freedom or the sample size changes. Thus there is a different t-distribution for each number of degrees of freedom or sample size.

(vi) The t-distribution approaches the standard normal distribution $N(0, 1)$ as the number of degrees of freedom ν, or the sample size becomes larger.

(vii) The important property of the t-distribution is that it is independent of the unknown population standard deviation σ. It is therefore applied to test hypotheses about the mean of a population irrespective of what the standard deviation may be.

18.2.2. The t-tables. The areas under the t-distribution have been tabulated for various values of t and ν for convenience. The table 18.1 on page 244, contains values of t_α for selected values of α, where

$$\alpha = P(t \geq t_\alpha) = \frac{1}{\sqrt{\nu}\,\beta\left(\frac{1}{2}, \frac{\nu}{2}\right)} \int_{t_\alpha}^{\infty} \frac{dt}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}$$

In other words, these are the values of $t_{\alpha,(\nu)}$ which denote values for which the area *to its right* under the t-distribution with ν degrees of freedom is equal to α. Owing to symmetry of the t-distribution about a mean of zero, it is important to note that $t_{1-\alpha} = -t_\alpha$, meaning thereby that the t values leaving an area of $1-\alpha$ to the right and *therefore an area equal to α to the left*, is equal to the negative t values that leave an area of α in the right tail of the t-distribution. Another point to note is that the entries in the last row of the t-table with $\nu = \infty$ correspond to those for a standard normal variable.

Table 18.1 Student's t-Distribution

The entries in this table are values of $t_{\alpha,(v)}$ for which the area to their *right* under the t-distribution with v degrees of freedom is equal to α. It is necessary to make use of the identity $t_{\alpha,(v)} = -t_{1-\alpha,(v)}$ for $\alpha > 0.50$.

v (d.f.)	0.25	0.10	0.05	0.025	0.01	0.005
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.838	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

"Table 18.1 is taken from Table III of Fisher and Yates: *Statistical Tables for Biological, Agricultural & Medical Research*, published by Oliver & Boyd, Edinburgh, & reproduced by permission of authors & publishers".

18.2.3. Distribution of Difference of Sample Means: Small Samples and $\sigma_1 = \sigma_2$. Let $X_{11}, X_{12}, \ldots, X_{1n_1}$, and $X_{21}, X_{22}, \ldots X_{2n_2}$ be two small independent random samples from two populations $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ with the same but unknown variance σ^2. Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means and

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 \text{ and } s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2 .$$

We know that $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ is a standard normal variable.

As the common variance σ^2 is unknown, its unbiased *pooled* estimate, denoted by s_p^2, obtained from both small samples, is given by

$$s_p^2 = \frac{\sum (X_{1i} - \bar{X}_1)^2 + \sum (X_{2j} - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}.$$

We also note that

$$U = \frac{(n_1 + n_2 - 2)\, s_p^2}{\sigma^2}$$

is a chi-square random variable with $v = n_1 + n_2 - 2$ degrees of freedom and is distributed independent of Z. We therefore construct the statistic t as the ratio of the standard normal variable Z to the square root of the chi-square random variable U, divided by its $d.f. = n_1 + n_2 - 2$. Thus

$$t = \frac{N(0, 1)}{\sqrt{\text{chi-square} / d.f.}} = \frac{Z}{\sqrt{U / n_1 + n_2 - 2}}$$

$$= \frac{\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}{\sqrt{\dfrac{(n_1 + n_2 - 2)\, s_p^2}{\sigma^2 (n_1 + n_2 - 2)}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

which conforms to the Student's t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom.

18.2.4. Assumptions in Using t-distribution. To use the t-distribution, we make the following assumptions:

(i) The sample of n observations $X_1, X_2, \ldots, X_n$ is selected randomly.

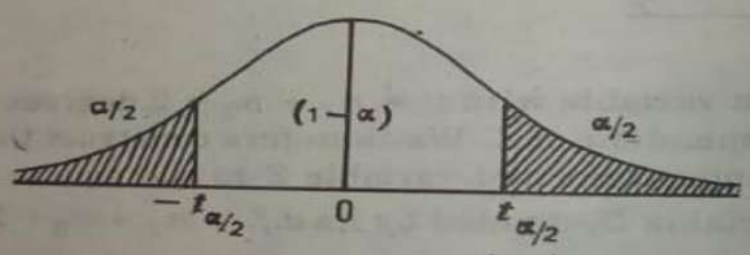
(ii) The population from which the small sample is drawn, is normal. This is essential for $\bar{X}$ and s, the two components of statistic t, to be independent. It has, however, been shown that slight departures from normality do not seriously affectt the tests.

(iii) In case of two small samples, both the samples are selected randomly, both the populations are normal and both the populations have *equal* variances.

18.3 CONFIDENCE INTERVAL ESTIMATE OF MEAN FROM A SMALL SAMPLE

The procedure for constructing the confidence interval for mean from small samples is the same as for mean from large samples except that we use the Student's t-distribution instead of the standard normal distribution. Thus if $-t_{\alpha/2,(v)}$ and $t_{\alpha/2,(v)}$ denote the values of t for which an area equal to $\alpha/2$ lies in each *tail* of the Student's t-distribution with v degrees of freedom, then the probability of t lying between these two values is given by the relation

$$P\left[-t_{\alpha/2,(v)} < t < t_{\alpha/2,(v)}\right] = 1 - \alpha.$$



That is, we have the following probability statement

$$P\left[-t_{\alpha/2,(v)} < \frac{\bar{X} - u}{s / \sqrt{n}} < t_{\alpha/2,(v)}\right] = 1 - \alpha.$$

Multiplying each term inside the bracket by $s / \sqrt{n}$, subtracting $\bar{X}$ and then multiplying by -1 (inequality signs reversed), we get

$$P\left[\bar{X} - t_{\alpha/2,(v)} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,(v)} \frac{s}{\sqrt{n}}\right] = 1 - \alpha$$

Hence the $(1-\alpha)100$ percent confidence interval for μ (when population σ^2 is unknown) for a particular random sample of size $n < 30$, is given by

$$\bar{x} \pm t_{\alpha/2,(v)} \frac{s}{\sqrt{n}}$$

Similarly, the $(1-\alpha)100$ per cent confidence interval for the difference of two means $\mu_1-\mu_2$, when the population variances are unknown but equal, is given by

$$(\bar{x}_1-\bar{x}_2) \pm t_{\alpha/2,(v)}\, s_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of two small random samples of sizes n_1 and n_2 from normal populations, s_p^2 = the pooled estimate of the population common variance, and $v = n_1 + n_2 - 2$ degrees of freedom.

The probability statement has to be interpreted in the same way as done in the case of confidence intervals obtained by using the normal distribution in Chapter 15. A one-sided upper (or lower) confidence limit may be found by replacing the lower (or upper) confidence limit with $-\infty$ $(+\infty)$ and using α instead of $\alpha/2$. For example, the $(1-\alpha)100\%$ upper confidence interval for μ would be

$$\mu \le \bar{x} + t_{\alpha,(v)} \frac{s}{\sqrt{n}}.$$

Example 18.1. The masses, in grams, of thirteen ball bearings taken at random from a batch are

21.4, 23.1, 25.9, 24.7, 23.4, 24.5, 25.0, 22.5, 26.9, 26.4, 25.8, 23.2, 21.9.

Calculate a 95% confidence interval for the mean mass of the population, supposed normal, from which these masses were drawn.

The 95% confidence interval for the mean mass of the population μ, is given by

$$\bar{x} \pm t_{\alpha/2,(v)} \frac{s}{\sqrt{n}}$$

Here $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{314.7}{13} = 24.21,$

$$s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]}$$

$$= \sqrt{\frac{1}{12}[7655.59 - 7618.16]} = \sqrt{\frac{37.43}{12}}$$

$$= \sqrt{3.12} = 1.77,$$

$v = n - 1 = 12$, and

$t_{0.025,(12)} = 2.179,$ (from the t-table)

Subsituting these values, we get

$$24.21 \pm 2.179 \left(\frac{1.77}{\sqrt{13}}\right)$$

or $\quad 24.21 \pm 2.179\ (0.49)$ or 24.21 ± 1.07 or 23.14 to 25.28

Hence the 95% confidence interval for μ calculated from the given sample is (23.1, 25.3) grams.

Example 18.2 Given that

$$\bar{x}_1 = 75,\ n_1 = 9,\ \sum (x_{1i} - \bar{x}_1)^2 = 1482;$$

$$\bar{x}_2 = 60,\ n_2 = 16,\ \sum (x_{2j} - \bar{x}_2)^2 = 1830;$$

and assuming that the two samples were randomly selected from two normal populations in which $\sigma_1^2 = \sigma_2^2$ (but unknown), calculate an 80% confidence interval for the difference between the two population means.

(P.U., M.A. Stat., 1966)

The 80% confidence interval for the difference between the two population means $\mu_1 - \mu_2$, is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,(v)} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $v(d.f.) = n_1 + n_2 - 2$.

Here the difference between sample means, $\bar{x}_1 - \bar{x}_2 = 75 - 60 = 15$.

The pooled estimate of the common variance σ^2 is

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_{1i} - \bar{x}_1)^2 + \sum_j (x_{2j} - \bar{x}_2)^2\right]$$

$$= \frac{1482 + 1830}{9 + 16 - 2} = 144, \text{ so that } s_p = \sqrt{144} = 12;$$

$$v = \text{degrees of freedom} = 9 + 16 - 2 = 23;$$

$t_{0.10,(23)} = 1.32 \qquad$ (from the t-table)

Substituting these values, we obtain

$$15 \pm (1.32)\ (12) \sqrt{\frac{1}{9} + \frac{1}{16}},$$

or $\quad 15 \pm (1.32)\ (5)$ or 15 ± 6.6 or 8.4 to 21.6.

Hence the 80% confidence interval for $\mu_1 - \mu_2$, calculated from the given information is (8.4, 21.6).

18.4 SMALL SAMPLE TESTS OF MEANS

Statistical inference may have to be drawn from small samples when collecting data for large samples is costly and time consuming. So, when small samples are drawn from normally distributed populations with unknown standard deviations, the hypothesis-testing techniques are based on the *Student's t*-distribution. A number of small sample tests of means are discussed in the subsections that follow.

18.4.1. Testing Hypothesis about Mean of a Normal Population when σ is unknown and $n < 30$. Let $X_1, X_2, \ldots, X_n$ be the observations in a small random sample of size n, taken from a normally distributed population whose standard deviation σ is unknown. Then σ is estimated from the sample data. Let $\bar{X}$ be the sample mean and s, be the unbiased estimate of σ, then, if we wish to test the hypothesis that the population mean has a specified value μ_0, the statistic

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

has, when the hypothesis is true, a t-distribution with $v = n - 1$ degrees of freedom. Hence the sample t value is the test-statistic for testing a hypothesis about the mean of a normal population with unknown standard deviation.

The *procedure* for testing the hypothesis $H_0 : \mu = \mu_0$ is given below:

(i) Formulate the null and alternative hypotheses about μ. Three possible forms are

(a) $H_0 : \mu = \mu_0$, and $H_1 : \mu \neq \mu_0$ (two-sided alternative)

(b) $H_0 : \mu \leq \mu_0$, and $H_1 : \mu > \mu_0$ (one-sided alternative)

(c) $H_0 : \mu \geq \mu_0$, and $H_1 : \mu < \mu_0$ (one-sided alternative)

(ii) Decide upon the significance level α.

(iii) The test-statistic to use is

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

which, if H_0 is true, follows a Student's t-distribution with $v = n - 1$ degrees of freedom.

(iv) Determine the critical region which for H_0 corresponding to different alternative hypotheses is given as follows:

When the alternative hypothesis is	the critical region will be
(a) $H_1 : \mu \neq \mu_0$	$t \leq -t_{\alpha/2,(n-1)}$ and $t \geq t_{\alpha/2,(n-1)}$
(b) $H_1 : \mu > \mu_0$	$t \geq t_{\alpha,(n-1)}$
(c) $H_1 : \mu < \mu_0$	$t \leq -t_{\alpha,(n-1)}$

Owing to symmetry, the critical region in case of two-sided alternative, is usually stated as $|t| \geq t_{\alpha/2,(n-1)}$.

(v) Compute the t-value from the given sample data.

(vi) Decide as below:

Reject H_0, when the computed t-value falls in the critical region, otherwise accept it.

Example 18.3. Ten individuals are chosen at random from a normal population and the heights are found to be in inches 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. In the light of these data, discuss the suggestion that mean height in the population is 66 inches.

(P.U., B.A/B.Sc. 1969)

(i) We state our null and alternative hypotheses as

$$H_0 : \mu = 66 \text{ and } H_1 : \mu \neq 66. \qquad \text{(two-sided)}$$

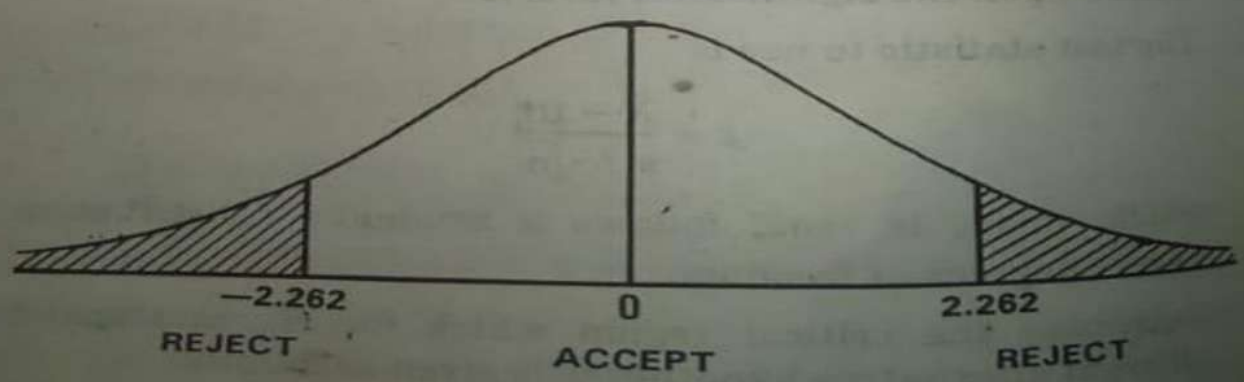
(ii) We set the significance level at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

which, if H_0 is true, has the t-distribution with $n - 1 = 9$ degrees of freedom.

(iv) The critical region is $|t| \geq t_{0.025,(9)} = 2.262$.



(v) We compute the t-value from the sample data as below:

$$\Sigma x_i = 63 + 63 + \dots + 71 = 678,$$

$$\Sigma x_i^2 = (63)^2 + (63)^2 + \dots + (71)^2 = 46050,$$

$$\therefore \quad \bar{x} = \frac{\Sigma x_i}{n} = \frac{678}{10} = 67.8 \text{ inches}$$

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

$$= \frac{1}{9}[46050 - 45968.4] = 9.0667, \text{ so that}$$

$$s = \sqrt{9.0667} = 3.01 \text{ inches.}$$

$$\text{Hence } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{67.8 - 66}{3.01/\sqrt{10}} = \frac{(1.8)(3.1623)}{3.01} = 1.89$$

(vi) **Conclusion.** Since the computed value of $t = 1.89$ does not fall in the critical region, we therefore do not reject H_0 and may conclude that the population mean is 66 inches.

Example 18.4. In a sample survey, six estimates were made of the same mean. When the population mean became known, the following errors were computed: −35, 111, −88, 47, −12, 26. Are these errors consistent with the hypothesis that the population of errors has a zero mean? Assume that the errors are normally distributed.

(i) We formulate our null and alternative hypotheses as

$$H_0 : \mu = 0 \text{ and } H_1 : \mu \neq 0.$$

(ii) We choose the level of significance at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

which, if H_0 is true, has the Student's t-distribution with $n-1=5$ degrees of freedom.

(iv) The critical region is $|t| \geq t_{0.025,(5)} = 2.57$.

(v) We compute the t-value from the sample data as follows:

Estimate No.	x_i	x_i^2
1	−35	1225
2	111	12321
3	−88	7744
4	47	2209
5	−12	144
6	26	676
Total	49	24319

Now, $\bar{x} = \dfrac{\sum x}{n} = \dfrac{49}{6} = 8.17;$

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

$$= \frac{1}{5}\left[24319 - \frac{(49)^2}{6}\right] = \frac{23918.83}{5} = 4783.7667, \text{ so that}$$

$$s = \sqrt{4783.7667} = 69.16$$

Hence $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{8.17 - 0}{69.16/\sqrt{6}} = \dfrac{(8.17)\,(2.4495)}{69.16} = 0.29.$

(vi) **Conclusion.** Since the computed value of $t = 0.29$ does not fall in the critical region, so we do not reject H_0 and may conclude that the data are consistent with the hypothesis that the population of errors has a zero mean.

18.4.2. Testing Hypotheses about Difference of Means of Two Normal Populations when $\sigma_1 = \sigma_2$ but unknown. Let X_{11}, X_{12}, ..., X_{1n_1} and X_{21}, X_{22}, ..., X_{2n_2} be two small independent random samples from two normal populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively. If $\sigma_1 = \sigma_2$ $(=\sigma)$ but unknown, then the unbiased pooled or combined estimate of the common variance σ^2 (The term *common variance* means that each population has the *same variance*), is given by

$$s_p^2 = \frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2}$$

where $s_1^2 = \frac{1}{n_1 - 1} \sum (X_{1i} - \bar{X}_1)^2$ and $s_2^2 = \frac{1}{n_2 - 1} \sum (X_{2j} - \bar{X}_2)^2$.

It has been previously shown that the statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a Student's t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom. Hence for small samples, taken from normal populations with unknown but equal standard deviations, it is used as the test-statistic for testing hypotheses about the difference between two population means.

Suppose we wish to test the hypothesis that the difference between means has a specified value Δ_0, *i.e.*, $H_0 : \mu_1 - \mu_2 = \Delta_0$. If H_0 is true, the test-statistic becomes

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which conforms to the t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom. In case $\Delta_0 = 0$, the test-statistic reduces to

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The *procedure* for testing the hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ in case of small independent samples and when $\sigma_1 = \sigma_2$, may be stated as below:

(i) Formulate the null and the alternative hypotheses; given $\sigma_1 = \sigma_2 = \sigma$ unknown;

$$H_0 : \mu_1 - \mu_2 = \Delta_0 \text{ against the appropriate alternative.}$$

(ii) Decide on the significance level α.

(iii) The test-statistic under H_0 is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has a t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom.

(iv) The critical region is

$t \le -t_{\alpha/2,(v)}$ and $t \ge t_{\alpha/2,(v)}$, when H_1 is $\mu_1 - \mu_2 \ne \Delta_0$;

$t \le -t_{\alpha,(v)}$, when H_1 is $\mu_1 - \mu_2 < \Delta_0$;

$t \ge t_{\alpha,(v)}$, when H_1 is $\mu_1 - \mu_2 > \Delta_0$.

(v) Compute the t-value from the given data.

(vi) Decide as below:

Reject H_0 if t falls in the critical region, accept H_0 otherwise.

It is worth remarking that when the sample size is large, *i.e.* $n > 30$, we generally suggest that the normal distribution be used to approximate the t-distribution because the t and z-values will then be quite close. The t-distribution due to this procedure is sometimes referred to incorrectly as applying only to *small samples*. However, it is emphasized that whenever σ is unknown and the population is normal, the application of the t-distribution is always *correct*. (See Exercise 18.32).

Example 18.5. Given the following samples from two normally distributed populations with the identical standard deviations but unknown, test $H_0 : \mu_1 - \mu_2 \le 3$ against $H_1 : \mu_1 - \mu_2 > 3$. Let $\alpha = 0.10$.

Sample 1	51, 42, 49, 55, 46, 63, 56, 58, 47, 39, 47
Sample 2	38, 49, 45, 29, 31, 35.

(I.U., M.Sc. 1987)

(i) We state our null and alternative hypotheses as

$$H_0 : \mu_1 - \mu_2 \le 3 \text{ and } H_1 : \mu_1 - \mu_2 > 3 \qquad \text{(one-tailed)}$$

(ii) The significance level is set at $\alpha = 0.10$.

(iii) The test-statistic to use is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - 3}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

which, if H_0 is true, has a Student's t-distribution with $n_1 + n_2 - 2$ $(= 11 + 6 - 2 = 15)$ degrees of freedom.

(iv) The critical region consists of all t-values that are greater than or equal to $t_{0.10,(15)} = 1.341$.

(v) Computations.

x_{1i}	x_{1i}^2	x_{2j}	x_{2j}^2
51	2601	38	1444
42	1764	49	2401
49	2401	45	2025
55	3025	29	841
46	2116	31	961
63	3969	35	1225
56	3136		
58	3364		
47	2209		
39	1521		
47	2209		
553	28315	227	8897

Now the sample means are $\bar{x}_1 = \dfrac{\sum x_{1i}}{n_1} = \dfrac{553}{11} = 50.3$,

$$\bar{x}_2 = \frac{\sum x_{2j}}{n_2} = \frac{227}{6} = 37.8, \text{ and}$$

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 = \sum x_{1i}^2 - \frac{(\sum x_{1i})^2}{n_1} = 28315 - \frac{(553)^2}{11}$$

$$= 28315 - 27800.82 = 514.18,$$

$$\sum_{j=1}^{n_2}(x_{2i} - \bar{x}_2)^2 = \sum x_{2j}^2 - \frac{(\sum x_{2j})^2}{n_2} = 8897 - \frac{(227)^2}{6}$$

$$= 8897 - 8588.17 = 308.83$$

$$\therefore\ s_p^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum(x_{1i} - \bar{x}_1)^2 + \sum(x_{2j} - \bar{x}_2)^2\right]$$

$$= \frac{514.18 + 308.83}{11 + 6 - 2} = \frac{823.01}{15} = 54.87, \text{ so that}$$

$$s_p = \sqrt{54.87} = 7.41$$

Thus $t = \dfrac{(50.3 - 37.8) - 3}{7.41\sqrt{\dfrac{1}{11} + \dfrac{1}{6}}} = \dfrac{9.5}{3.76} = 2.53.$

(vi) **Conclusion.** Since the computed value of $t = 2.53$ falls in the critical region, so we reject H_0 and accept H_1.

Example 18.6. From an area planted in one variety of guayule (a rubber producing plant), 54 plants were selected at random. Of these, 15 were offtypes and 12 were aberrant. Rubber percentages for these plants were:

Offtypes	6.21, 5.70, 6.04, 4.47, 5.22, 4.45, 4.84, 5.88, 5.82, 6.09, 5.59, 6.06, 5.59, 6.74, 5.55.
Aberrant	4.28, 7.71, 6.48, 7.71, 7.37, 7.20, 7.06, 6.40, 8.93, 5.91, 5.51, 6.36.

Test the hypothesis of no difference between means of populations of rubber percentages. Assume the populations of rubber percentages are approximately normal and have equal variances. (P.U., B.A/B.Sc. 1983)

(i) We formulate our null and alternative hypotheses as

$H_0 : \mu_1 - \mu_2 = 0$, *i.e.* there is no difference between means; and

$H_1 : \mu_1 - \mu_2 \neq 0$, *i.e.* the two means are different.

(ii) We set the significance level at $\alpha = 0.05$.

(iii) The test-statistic, if H_0 is true, is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

which has a Student's t-distribution with $v = n_1 + n_2 - 2$, *i.e.* 25 degrees of freedom.

(iv) The critical region consists of all t-values which are greater than or equal to $t_{0.025,(25)} = 2.06$ and which are less than or equal to $-t_{0.025,(25)} = -2.06$.

(v) Computations. Let X_{1i} and X_{2j} represent the offtype and aberrant measurements respectively. Then

$$\bar{x}_1 = \frac{\sum x_{1i}}{n_1} = \frac{84.25}{15} = 5.62,$$

$$\bar{x}_2 = \frac{\sum x_{2j}}{n_2} = \frac{80.92}{12} = 6.74, \text{ and}$$

$$\sum_{i=1}^{15}(x_{1i}-\bar{x}_1)^2 = \sum x_{1i}^2 - \frac{(\sum x_{1i})^2}{n_1} = 478.9779 - \frac{(84.25)^2}{15}$$

$$= 478.9779 - 473.2042 = 5.7737;$$

$$\sum_{j=1}^{12}(x_{2j}-\bar{x}_2)^2 = \sum x_{2j}^2 - \frac{(\sum x_{2j})^2}{n_2} = 561.6402 - \frac{(80.92)^2}{12}$$

$$= 561.6402 - 545.6705 = 15.9697$$

$$\therefore \quad s_p^2 = \frac{\sum_i (x_{1i}-\bar{x}_1)^2 + \sum_j (x_{2j}-\bar{x}_2)^2}{n_1+n_2-2} = \frac{5.7737 + 15.9697}{15+12-2}$$

$= 0.8697$, so that

$$s_p = \sqrt{0.8697} = 0.93, \text{ and}$$

$$t = \frac{5.62 - 6.74}{0.93\sqrt{\frac{1}{15}+\frac{1}{12}}} = \frac{-1.12}{0.36} = -3.11.$$

(vi) **Conclusion.** Since the computed value of $t = -3.11$ falls in the critical region, we therefore reject the hypothesis of no difference between the two means. We conclude that there is sufficient evidence to indicate a difference in the means of rubber percentages.

18.4.3. Testing Hypotheses about Difference of Means of Two Normal Populations when $\sigma_1 \neq \sigma_2$ and unknown. Suppose that we are given two small random samples $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ from two normally distributed populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively. If $\sigma_1 \neq \sigma_2$ and unknown, we use their sample estimates s_1 and s_2 to compute the standard error of the difference between means and get

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},$$

as there is no point in combining s_1^2 and s_2^2 to obtain an estimate of a non-existent common population variance. Consequently, for testing the null hypothesis $H_0 : \mu_1 = \mu_2$, the test-statistic, if H_0 is true, is given by the relation

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

which has approximately a Student's t-distribution with v degrees of freedom, where

$$v = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

When the value of v does not happen to be an integer, it is always rounded down for degrees of freedom. This general problem is often called the Behrens-Fisher problem. The rest of the procedure is the same.

Example 18.7. Given two random samples of size $n_1 = 7$ and $n_2 = 6$ from two independent normal populations, with $\bar{x}_1 = 10.91$, $\bar{x}_2 = 4.60$, $s_1 = 6.34$ and $s_2 = 3.09$, test the hypothesis at the 0.05 level of significiance that $\mu_1 = \mu_2$ against the alternative that $\mu_1 \neq \mu_2$. Assume that the population variances are unequal.

(i) We state our null and alternative hypotheses as

$H_0 : \mu_1 = \mu_2$, and $H_1 : \mu_1 \neq \mu_0$.

(ii) The level of significance is set at $\alpha = 0.05$

(iii) Since the populations have *unequal* variances, the test statistic, if H_0 is true, is

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

which has approximately a t-distribution with v degrees of freedom, where

$$v = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

(iv) We then compute the values of t' and v from the data as follows

$$t' = \frac{10.91 - 4.60}{\sqrt{\frac{(6.34)^2}{7} + \frac{(3.09)^2}{6}}} = \frac{6.31}{2.708} = 2.33, \text{ and}$$

$$v = \frac{[(6.34)^2/7 + (3.09)^2/6]^2}{\frac{[(6.34)^2/7]^2}{6} + \frac{[(3.09)^2/6]^2}{5}} = \frac{53.79}{6.00} \cong 8 \text{ (the } d.f. \text{ is}$$

always rounded down as in this calculation)

(v) The critical region is $|t'| \geq t_{0.025,(8)} = 2.306$

(vi) **Conclusion.** Since the calculated value of t' falls in the critical region, so we reject our null hypothesis of equal means.

18.4.4. Testing Hypotheses about Two Means with Paired Observations. In testing hypotheses about two means, we have used independent samples, but there are many situations in which the two samples are *not* independent. This happens when the observations are found a pairs as the two observations of a pair are related to each other. Pairing occurs either naturally or by design. Natural pairing occurs whenever measurement is taken on the same unit or individual at two different times. For examples, suppose 10 young recruits are given a strenuous physical training programme by the Army. Their weights are recorded before they begin and after they complete the training. The two observations obtained for each recruit, *i.e.* the *before-and-after* measurements constitute natural pairing. Observations are also paired to eliminate effects in which there is no interest. For example, suppose we wish to test which of two types (A or B) of fertilizers is the better. The two types of fertilizers are applied to a number of plots and the results are noted. Assuming that the two types are found significantly different, we may find that part of the difference may be due to the different types of soil or different weather conditions, etc. Thus the real difference between the fertilizers can be found only when the plots are paired according to the same types of soil or same weather conditions, etc. We eliminate the undesirable sources of variation to take the observations in pairs. This is pairing by design.

When the observations from two samples are paired either naturally or by design, we find the difference between two observations of each pair. Treating the differences as a random sample from a normal population with mean $\mu_D = \mu_1 - \mu_2$ and unknown standard deviation σ_D, we perform a one-sample t-test on them. This is called a *paired difference t-test* or a *paired t-test*.

Testing the hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ is equivalent to testing $H_0 : \mu_D = 0$ against $H_1 : \mu_D \neq 0$.

Let $d_i = x_{1i} - x_{2i}$ denote the difference between the two sample observations in the *ith* pair. Then the sample mean and standard deviation of the differences are

$$\bar{d} = \frac{\sum d_i}{n} \text{ and } s_d = \frac{\sum (d_i - \bar{d})^2}{n-1},$$

where n represents the number of pairs.

Assuming that (i) $d_1, d_2, \ldots, d_n$ is a random sample of differences and (ii) the differences are normally distributed, the test-statistic

$$t = \frac{\bar{d}}{s_d / \sqrt{n}},$$

follows a t-distribution with $v = n - 1$ degrees of freedom. The rest of the procedure for testing the null hypothesis $H_0 : \mu_D = 0$ is the same.

Example 18.8. Ten young recruits were put through a stenuous physical training programme by the Army. Their weights were recorded before and after the training with the following results:

Recruit	1	2	3	4	5	6	7	8	9	10
Weight *before*	125	195	160	171	140	201	170	176	195	186
Weight *after*	136	201	158	184	145	195	175	190	190	146

Using $\alpha = 0.05$, would you say that the programme affects the average weight of recruits? Assume the distribution of weights before and after to be approximately normal. (P.U., B.A/B.Sc. 1984)

The pairing was natural here, since two observations are made on the same recruit at two different times. The sample consists of 10 recuirts with two measurements on each.

The test is carried out as below:

(i) We state our null and alternative hypotheses as

$H_0 : \mu_D = 0$ and $H_1 : \mu_D \neq 0$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic under H_0 is

$$t = \frac{\bar{d}}{s_d / \sqrt{n}},$$

which has a t-distribution with $n - 1$ degrees of freedom.

(iv) The critical region is $|t| \geq t_{0.025,(9)} = 2.262$.

(v) Computations.

Recruit	Weight		Difference, d_i (after minus before)	d_i^2
	Before	After		
1	125	136	11	121
2	195	201	6	36
3	160	158	−2	4
4	171	184	13	169
5	140	145	5	25
6	201	195	−6	36
7	170	175	5	25
8	176	190	14	196
9	195	190	−5	25
10	139	145	6	36
Σ	1672	1719	47	673

Now $\bar{d} = \dfrac{\Sigma d_i}{n} = \dfrac{47}{10} = 4.7.$

$$s_d^2 = \frac{\Sigma(d_i - \bar{d})^2}{n-1} = \frac{1}{n-1}\left[\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right]$$

$$= \frac{1}{9}\left[673 - \frac{(47)^2}{10}\right] = \frac{673 - 220.9}{9} = 50.23, \text{ so that}$$

$$s_d = \sqrt{50.23} = 7.09.$$

$$\therefore \quad t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{4.7}{7.09 / \sqrt{10}} = \frac{(4.7)(3.16)}{7.09} = 2.09.$$

(vi) **Conclusion.** Since the calculated value of $t = 2.09$ does not fall in the critical region, so we accept H_0 and may conclude that the data do not provide sufficient evidence to indicate that the programme affects average weight.

Example 18.9. The following data give a paired yields of two varieties of wheat. Each pair was planted in a different locality. Test the hypothesis that the mean yields are equal.

Variety *I*	45	32	58	57	60	38	47	51	42	38
Variety *II*	47	34	60	59	63	44	49	53	46	41

(P.U., B.A/B.Sc. 1980)

The pairing was by design here, as the yields are affected by many extraneous factors such as fertility of land, fertilizer applied, weather conditions and so forth. The test is carried out as below:

(i) We state our null and alternative hypotheses as

$H_0 : \mu_D = 0$ (or $\mu_1 = \mu_2$), *i.e.* the mean yields are equal, and

$H_1 : \mu_D \neq 0$, *i.e.* the mean yields differ.

(ii) We select the level of significance at $\alpha = 0.05$.

(iii) The test statistic to be used is

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where $\bar{d} = \bar{x}_1 - \bar{x}_2$ and s_d^2 is the variance of the difference. If the populations are normal, this statistic, when H_0 is true, has a Student's t-distribution with $(n-1)$ *d.f.*

(iv) The critical region is $|t| \geq t_{0.025(9)} = 2.262$ $(\because n = 10)$

(v) Computations. Let X_{1i} and X_{2i} represent the yields of Variety *I* and Variety *II* respectively. Then we have

$$d_i = X_{1i} - X_{2i}$$

$$= -2, -2, -2, -2, -3, -6, -2, -2, -4, -3.$$

$$\Sigma d_i = -28 \text{ and } \Sigma d_i^2 = 94.$$

Now $\bar{d} = \dfrac{\Sigma d_i}{n} = \dfrac{-28}{10} = -2.8$, and

$$s_d^2 = \frac{1}{n-1}\left[\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right] = \frac{1}{9}\left[94 - \frac{(-28)^2}{10}\right]$$

$$= \frac{15.6}{9} = 1.7333, \text{ so that } s_d = 1.32.$$

$$\therefore \quad t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{-2.8}{1.32 / \sqrt{10}} = \frac{(-2.8)\,(3.1623)}{1.32} = -6.71$$

(vi) **Conclusion.** Since the calculated value of $t = -6.71$ falls in the critical region, we therefore reject H_0. The data present sufficient evidence to conclude that the mean yields are not equal.

Example 18.10. An experiment was performed with seven hop plants. One half of each plant was pollinated and the other half was not pollinated. The yield of the seed of each hop plant is tabulated as follows:

Plant Number	1	2	3	4	5	6	7
Pollinated	0.78	0.76	0.43	0.92	0.86	0.59	0.68
Non-pollinated	0.21	0.12	0.32	0.29	0.30	0.20	0.14

(a) Determine at the 5 per cent level whether the pollinated half of the plant gives a higher yield in seed than the non-pollinated half. State the assumptions and hypotheses to be tested and carry through the computations to make decision.

(b) How can the experimenter make a Type-I error? What are the consequences of his doing so?

(c) How can the experimenter make a Type-II error? What are the consequences of his doing so?

(d) Give 90 per cent confidence limits for the difference in mean yields. (P.U., M.Sc. 1971)

(a) (i) We wish to decide between the hypotheses

H_0: The pollinated half *does not give* a higher mean yield than the non-pollinated half. In other words,

$H_0: \mu_P \le \mu_N$ where μ_P denotes the mean yield of pollinated and μ_N that of the non-pollinated; and

$H_1: \mu_P > \mu_N$, that is the pollinated half *does give* a higher yield than the non-pollinated. (This is the claim) (one-tailed test)

(ii) The significance level is set at $\alpha = 0.05$.

(iii) As the observations are paired, therefore the test statistic to be used is

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}.$$

Assuming that (1) the differences in yields are a random sample from the population of differences and (2) the population of

differences is normally distributed, this statistic if H_0 is true, has a Student's t-distribution with $(n-1)$ $d.f.$

(iv) We then compute the value of t as below:

Pollinated	Non-pollinated	d_i	d_i^2
0.78	0.21	0.57	0.3249
0.76	0.12	0.64	0.4096
0.43	0.32	0.11	0.0121
0.92	0.29	0.63	0.3969
0.86	0.30	0.56	0.3136
0.59	0.20	0.39	0.1521
0.68	0.14	0.54	0.2916
Total	--	3.44	1.9008

Now $\bar{d} = \dfrac{\sum d_i}{n} = \dfrac{3.44}{7} = 0.491$, and

$$s_d^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right]$$

$$= \frac{1}{6}\left[1.9008 - \frac{(3.44)^2}{7}\right] = \frac{0.2093}{6} = 0.0349, \text{ so that}$$

$$s_d = \sqrt{0.0349} = 0.1868$$

$$\therefore \quad t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{(0.491)\,(2.646)}{0.1868} = \frac{1.3045}{0.1868} = 6.96,$$

(v) The critical region is $t \geq t_{0.05,(6)} = 1.943$.

(vi) **Conclusion.** Since the calculated value of $t = 6.96$ falls in the rejection region, so we reject our null hypothesis and may conclude that there is evidence that pollination gives a higher mean yield of seed.

(b) The experimenter can make a Type-I error by rejecting a *true* null hypothesis. In this case, the Type-I error is made by rejecting the null hypothesis when the mean yield of seed from pollinated plants is actually not greater than the mean yield of seed from non-pollinated plants. In so doing, the consequences would be that he will use the pollination when it actually does not increase the yield.

(c) The experimenter can make a Type-II error by accepting a *false* null hypothesis. In this case, the Type-II error is made by accepting the null hypothesis when pollination actually *does increase* the yield and the consequence of committing this error would be a loss of potential increased yield.

(d) The 90% confidence limits for the difference in means $\mu_1 - \mu_2$ in case of paired observations, are given by

$$\bar{d} \pm t_{\alpha/2,(n-1)} \cdot \frac{s_d}{\sqrt{n}}$$

Substituting the values, we get

$$0.491 \pm 1.943 \frac{0.1868}{\sqrt{7}}$$

or $\quad 0.491 \pm 0.137$ or 0.354 to 0.628.

Hence the 90% confidence limits for the difference in mean yields, $\mu_1 - \mu_2$, are (0.35, 0.63).

EXERCISES

18.1 (a) Why is the z-test usually inappropriate as a test-statistic when the sample size is small?

(b) Define "Student's t". What are its assumptions? Explain briefly its use and importance in statistics. (P.U., B.A/B.Sc. 1985)

18.2 Derive the distribution of Student's t and discuss its chief properties. (P.U., M.Sc. 1970)

18.3 (a) Is $E(t) = 0$ for all values of v?

(b) With t defined by $t = \dfrac{Z}{\sqrt{\chi^2 / v}}$, show that $\text{Var}(t) = \dfrac{v}{v-2}$ for $v > 2$.

(c) Prove that the t-distribution approaches the standard normal distribution as the number of degrees of freedom v becomes infinite. (P.U., M.Sc. 1970)

18.4 A random sample of size n is drawn from a normal population with mean 5 and variance σ^2. Answer the following:

(i) If $n = 25$, $\bar{x} = 3$ and $s = 2$, what is t?

(ii) If $n = 9$, $\bar{x} = 2$ and $t = -2$, what is s?

(iii) If $n = 25$, $s = 10$ and $t = 2$, what is $\bar{x}$?

(iv) If $s = 15$, $\bar{x} = 14$ and $t = 3$, what is n? (P.U., B.A/B.Sc. 1988)

18.5 (a) Explain how a confidence interval is constructed for the mean of a normal population for small samples.

(b) The breaking strength of 10 specimens of 0.104 inches diameter hard-drawn copper wire are found to be 578, 572, 570, 568, 572, 570, 570, 572, 596, 584 lbs. Calculate the 95% confidence interval for the mean breaking strength of this kind of wire, assuming that the measurements are normally distributed.

18.6 (a) What is meant by the term "Confidence Interval"? Find a 90% confidence interval for the mean of a normal distribution if $\sigma=2$ and if a sample of size 8 gave the values 9, 14, 10, 12, 7, 13, 11, 12. What would be the confidence interval if σ were unknown? (P.U., B.A./B.Sc. 1978)

(b) Find the 90% confidence limits for the mean of a normal distribution with $\sigma=3$, given the sample as 2.3, −0.2, −0.4, and −0.9. How will these limits be affected if σ is not known? (P.U., B.A/B.Sc. 1987, 92)

18.7 (a) A random sample of 10 boys had the following *I.Q's*:

70, 120, 110, 101, 88, 83, 95, 107, 100, 98.

Find the 95% confidence interval for the population mean. Assume that *IQ's* are normally distributed and the variance is not known.

(b) A sample of size 16 from a normal population with unknown standard deviation, gave $\bar{x} = 14.5$ and $s = 5$. Find a 99% confidence interval on the population mean.

18.8 The masses, in grams, of twelve ball bearings taken at random from a batch are 31.4, 33.1, 35.9, 34.7, 33.4, 34.5, 35.0, 32.5, 36.1, 36.4, 35.8, 33.2. Calculate a 90% confidence interval for the mean mass of the population, supposed normal, from which these masses were drawn.

18.9 (a) Give two random samples of size $n_1=9$ and $n_2=16$, from two independent normal populations, with $\bar{x}_1 = 64$, $\bar{x}_2 = 59$, $s_1=4$ and $s_2 = 5$, find a 95% confidence interval for $\mu_1-\mu_2$ assuming that $\sigma_1 = \sigma_2$.

(b) A sample from a normal population with unknown variance consists of the observations 34, 25, 43, 37, 45. A sample from a second normal population with the same unknown variance as the first consists of the observation 20, 31, 23, 35, 41, 39. Find a 95% confidence interval on $\mu_1-\mu_2$.

18.10 A course in mathematics is taught to 10 students by the conventional class room procedure. A second group of 12 students was given the same course by means of programmed materials. At the end of the semester, the same examination was given to each group. Their grades are given as follows:

Group I	70	66	76	77	73	72	38	74	75	69		
Group II	77	83	92	85	82	84	80	78	91	93	80	87

Find a 90% confidence interval for the differences between the average grades of the two groups. Assume the populations to be approximately normal with equal variances. (P.U., B.A/B.Sc. 1981)

18.11 From the area planted in one variety of guayule, 54 plants were selected at random. Of these plants, 15 were "Offtypes" and 12 were "Aberrants". The rubber per centages of these plants were:

Offtypes	4.47, 5.88, 6.21, 5.55, 6.09, 5.70, 5.82, 4.84, 5.59, 5.59, 5.22, 4.45, 6.74, 6.06, 6.04
Aberrants	6.48, 6.36, 4.28, 7.71, 6.40, 7.06, 5.51, 8.93, 7.71, 7.20, 7.37, 5.91

Compute a 90% confidence interval for the difference of two population means. (P.U., B.A/B.Sc. 1990)

18.12 (a) What assumptions are made when Student's t-test is employed to test a hypothesis about a population mean?

(b) Describe the procedure for testing hypothesis about mean of a normal population when standard deviation is unknown and sample size is small.

18.13 Given the following information, what is your conclusion in testing each of the indicated null hypotheses? Assume the populations are normal.

Sample	Sample size (n)	Sample Mean ($\bar{X}$)	Estimate of variance from sample (s^2)	Significance level (α)	Hypotheses H_0	Hypotheses H_1
a	9	12	36	0.05	$\mu=10$	$\mu>10$
b	16	13	64	0.05	$\mu=10$	$\mu\neq10$
c	16	11	81	0.01	$\mu\leq10$	$\mu>10$
d	25	8	64	0.01	$\mu\geq10$	$\mu<10$
e	25	9	49	0.02	$\mu=10$	$\mu\neq10$

(P.U., B.A/B.Sc. 1986)

18.14 A sample of 12 jars of peanut butter was taken from a lot, each jar being labelled "8 ounces net weight." The individual weights in ounces are: 8.2, 8.0, 7.6, 7.6, 7.7, 7.5, 7.3, 7.4, 7.5, 8.0, 7.4, 7.5. Test whether these values are consistent with a population mean of 8. Assume that the weights are normally distributed.

18.15 (a) The nine items of a sample had the following values:

45, 47, 50, 52, 48, 47, 49, 53, 51.

Does the mean of the nine items differ significantly from an assumed normal population mean of 47.5 at $\alpha = 0.05$?

(b) Ten cartons are taken at random from an automatic filling-machine. The mean net weight of the 10 cartons is 15.90 oz., and the sum of squared deviations is 0.276. Does the sample mean differ significantly from the intended weight of 16 oz.?

18.16 A random sample of 16 values from a normal population showed a mean of 41.5 inches and a sum of squares of deviations from this mean equal to 135 (inches)^2. Show that the assumption of a mean of 43.5 inches for the population is not reasonable and that the 95% confidence limits for this mean are 39.9 and 43.1 inches.

18.17 In the past a machine has produced washers having a thickness of 0.050 inches. To determine whether the machine is in proper working order, a sample of 10 washers is chosen for which the mean thickness is 0.053 inches and the standard deviation is 0.003 inches. Test the hypothesis that the machine is in proper working order, using a level of significance of 0.05 and 0.01.

18.18 A manufacturer claims that his light bulbs have an average lifetime of 1500 hours. A purchaser decides to check this claim and finds that for six bulbs the lifetimes are 1472, 1486, 1401, 1350, 1610, 1590 hours. Does this evidence support the manufacturer's claim? Assume that the lifetimes of the light bulbs are normally distributed.

18.19 (a) Describe the procedure for testing hypotheses about the equality of means of two normal populations for small samples.

(b) Two random samples taken independently from normal populations with an identical variance yield the following results:

	n	$\bar{X}$	s^2
Sample I	12	10	1200
Sample II	18	25	900

Test the hypothesis that the true difference between the population means is 10, that is, that $\mu_2-\mu_1=10$ against the alternative that $\mu_2-\mu_1>10$ at the 5 per cent level of significance. (P.U., B.A/B.Sc. 1989)

18.20 The weights in grams of 10 male and 10 female juvenile ring-necked pheasants are:

Males: 1293, 1380, 1614, 1497, 1340, 1643, 1466, 1627, 1383, 1711;

Females: 1061, 1065, 1092, 1017, 1021, 1138, 1143, 1094, 1270, 1028.

Test the hypothesis of a difference of 350 grams between population means in favour of males against the alternative of a greater difference, using a 0.05 level of significance. Assume that the weights are normally distributed. (P.U., B.A/B.Sc. 1987, 92)

18.21 (a) The heights of six randomly selected sailors are in inches: 63, 65, 68, 69, 71 and 72. Those of ten randomly selected soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss in the light of these data that soldiers are on the average taller than sailors. Assume that the heights are normally distributed.

(b) Eight pots, growing three barley plants each, were exposed to a high tension discharge while nine similar pots were enclosed in an earthed wire cage. The number of tillers (shoots) in each pot were as follows:

Caged: 17, 27, 18, 25, 27, 29, 27, 23, 17.

Electrified: 16, 16, 20, 16, 21, 17, 15, 20.

Discuss whether electrification exercises any real effect on tillering.

18.22 Twelve hogs were fed on diet *A*, 15 on diet *B*. The gains in weights for the individual hogs (in pounds) were as shown:

A: 25, 30, 28, 34, 24, 25, 13, 32, 24, 30, 31, 35

B: 44, 34, 22, 8, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22.

What conclusions may be drawn from this experiment?

18.23 (a) What statistical hypotheses can be tested by means of the *t*-distribution?

(b) A group of 12 children are found to have the following intelligence quotients: 112, 109, 125, 113, 116, 131, 112, 123, 108, 113, 132 and 128. Is it reasonable to suppose that these children have come from a large population whose average IQ is 115?

(c) A second group of 10 children is tested, resulting in the following IQ's:

117, 110, 106, 109, 116, 119, 107, 106, 105 and 108.

Is this group significantly different from the first group? (I.U., M.Sc. 1987)

18.24 (a) Two separate groups of subjects were tested. The experimental group (Group E) had 10 subjects; the control group (Group C) had 9 subjects. The data are given below; the scores are assumed to be normally distributed.

Group E: 12, 13, 16, 14, 15, 12, 15, 14, 13 and 16.

Group C: 10, 13, 14, 12, 15, 16, 12, 14 and 11.

Determine whether the means of the two groups differ significantly at the 0.05 level of significance.

(b) The strength of ropes made out of cotton yarn and coir gave on measurement the following values:

Cotton yarn: 7.5, 5.4, 10.6, 9.0, 6.1, 10.2, 7.9, 9.7, 7.1, 8.5

Coir: 8.3, 6.1, 9.6, 10.4, 6.4, 10.0, 7.9, 8.9, 7.5, 9.7

Test whether there is a significant difference in the strength of the two types of ropes at 0.05 level of significance.

18.25 Out of 20 children, 10 selected at random were given a ration of orange juice each day and the other 10, a ration of milk. Their gains in weights after a certain period were found to be as follows:

1st group: 2.6, 1.5, 4, 1, 3.5, 3.4, 2.5, 3, 4, 3.5.

2nd group: 3.5, 2.5, 1.5, 2.5, 3, 2, 3, 2, 1.5, 2.5

Use t-test to determine whether the means differ significantly, assuming the standard deviation for each group is the same.

18.26 Give two random samples of size $n_1=9$ and $n_2=16$, from two independent normal populations, with $\bar{x}_1=75$, $\bar{x}_2=60$, $s_1=13.67$ and $s_2=11.05$, test the hypothesis at the 0.05 level of significance that $\mu_1=\mu_2$ against the alternative that $\mu_1>\mu_2$. Assume that the populations have unequal variances.

18.27 (a) Distinguish between situations requiring a two-sample t-test and a paired-sample t-test. What distributional assumptions are made in each use?

(b) The weights of 4 persons before they stopped smoking and 8 weeks after they stopped smoking are as follows:

Person	1	2	3	4
Before	148	176	153	116
After	154	176	151	121

Use the t-test for paired observations to test the hypothesis at the 0.05 level of significance, that giving up smoking has no effect on a person's weight. (P.U., B.A/B.Sc. 1969, 90)

18.28 To verify whether a course in statistics improved performance, a similar test was given to 12 participants both before and after the course. The original grades recorded in alphabetical order of the participants were 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. After the course, the grades were in the same order 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78.

(a) Was the course useful, as measured by performance on the test? Consider these 12 participants as a sample from a population.

(b) Would the same conclusion be reached if tests were not considered paired? Use 5% level of significance in both cases. (P.U., B.A./B.Sc. 1985)

18.29 In a certain experiment to compare two types of sheep food A and B, the following results of increase in weights were observed:

Sheep No.	1	2	3	4	5	6	7	8
Food A	49	53	51	52	47	50	52	53
Food B	52	55	52	53	50	54	54	53

(a) Assuming that the two samples of sheep are independent, can we conclude that food B is better than food A?

(b) Examine the case when the same set of eight sheep were used in both the foods.

18.30 The government awarded grants to nine different experimental stations of the agricultural department to test the yield capabilities of two varieties of wheat. Five acres of each variety are planted at each station and the yields, in maunds per acre, recorded as follows:

Station	1	2	3	4	5	6	7	8	9
Variety 1	38	23	35	41	44	29	37	31	38
Variety 2	45	25	31	38	50	33	35	40	43

Test the hypothesis, at the 0.05 level of significance, that the average yields of the two varieties of wheat are equal against the alternative hypothesis that they are unequal, assuming the distribution of yields to be approximately normal. Explain why pairing is necessary in this problem.

18.31 A taxi company is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars were then equipped with regular belted tires and driven once again over the test course. The gasoline consumption in km per liter, was recorded as follows:

Radial tires	4.2, 4.7, 6.6, 7.0, 6.7, 4.5, 5.7, 6.0, 7.4, 4.9, 6.1, 5.2
Belted tires	4.1, 4.9, 6.2, 6.9, 6.8, 4.4, 5.7, 5.8, 6.9, 4.7, 6.0, 4.9

At the 0.025 level of significance, can we conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires? Assume the population to be normally distributed. (P.U., B.A/B.Sc. 1988)

18.32 In the experiment on the effectiveness of a teaching machine, a machine instructed group of students was compared with a teacher instructed group on an achievement test. The following scores were obtained:

Score	40–49	50–59	60–69	70–79	80–89	90–99	Total
Teacher instructed	21	40	55	38	10	2	166
Machine instructed	18	35	42	46	19	4	164
Total	39	75	97	84	29	6	330

Apply t-test to determine whether there is significant difference in achievement of the two groups. (M.Sc., P.U., 1988; I.U., 1992)

❖❖❖❖❖❖❖❖❖❖❖

19

The F-Distribution and Statistical Inference

19.1 INTRODUCTION

In the preceding chapter, we used the t-distribution to test the hypothesis about the difference between two means under the *assumption* that the two random samples are drawn independently from two normal populations that have *equal* variances. But in actual practice, the variances may or may not be equal. To check the assumption that the two normally distributed populations have equal variances, we use an important distribution, called the F-distribution, which is the sampling distribution of the ratio of two independent and unbiased estimates of the population variances. If the unbiased estimates denoted by s_1^2 and s_2^2, have been obtained from two random samples of sizes n_1 and n_2, drawn from normal populations having the same variances, then (assuming that s_1^2 is larger than s_2^2) the ratio is given by

$$F = \frac{s_1^2}{s_2^2}.$$

This ratio was named F by G.W. Snedecor (1882–1974) in honour of the great British statistician, Sir R.A. Fisher (1890–1962), who in 1924 developed its distribution as the Z-distribution which was later transformed into the F-distribution, using the relation $F = e^{2z}$.

Dividing the two estimates by the population variance σ^2, the ratio becomes

$$F = \frac{s_1^2 / \sigma^2}{s_2^2 / \sigma^2}.$$

It has already been shown that $\frac{(n_1 - 1)\, s_1^2}{\sigma^2}$, *i.e.* $\frac{\Sigma (X_{1i} - \bar{X}_1)^2}{\sigma^2}$ is distributed as a χ^2-variable with $v_1 = n_1 - 1$ degrees of freedom. Similarly, the quantity $\frac{(n_2 - 1)\, s_2^2}{\sigma^2}$ is distributed as a χ^2-variable with $v_2 = n_2 - 1$ degrees of freedom. Hence, F is a ratio of two independent chi-square random variables, each divided by its respective degrees of freedom. The distribution of this statistic (ratio) is known as the Snedecor's F-distribution or sometimes, the variance ratio distribution. The F-distribution has $v_1 = n_1 - 1$ degrees of freedom for the numerator and $v_2 = n_2 - 1$ degrees of freedom for the denominator. It is interesting to note that the F-distribution has two parameters, namely v_1 and v_2, the degrees of freedom in that order. The F-distribution is extremely important as it has broad applications in modern statistical analysis.

19.2 THE F-DISTRIBUTION

Let s_1^2 and s_2^2 be the unbiased estimated variances of two random samples of sizes n_1 and n_2, drawn from normal populations with same variances. Then the ratio $F = \frac{s_1^2}{s_2^2}$ may be written as

$$F = \frac{s_1^2 / \sigma^2}{s_2^2 / \sigma^2} = \frac{U / v_1}{U / v_2},$$

where $U = \frac{(n_1 - 1)\, s_1^2}{\sigma^2}$ is a χ^2-variable with $v_1 = n_1 - 1$ *d.f.* and

$$V = \frac{(n_2 - 1)\, s_2^2}{\sigma^2} \text{ is a } \chi^2\text{-variable with } v_2 = n_2 - 1 \text{ } d.f.$$

To find the distribution of F, we proceed as follows:

Since U and V are independent χ^2-variables with v_1 and v_2 degrees of freedom respectively, therefore their joint distribution is

$$f(u, v) = \frac{u^{(v_1/2)-1} e^{-u/2}}{2^{v_1/2} \Gamma (v_1/2)} \cdot \frac{v^{(v_2/2)-1} e^{-v/2}}{2^{v_2/2} \Gamma (v_2/2)}, \quad 0 < u, v < \infty$$

$$= \frac{u^{(v_1/2)-1} v^{(v_2/2)-1}}{2^{(v_1+v_2)/2} \Gamma (v_1/2) \Gamma (v_2/2)} e^{-(u+v)/2}$$

To obtain the distribution of F, we make the change of variables as $F = \frac{u/v_1}{v/v_2}$ and $v = v$ so that $u = \frac{vv_1}{v_2} F$, $v = v$ and the Jacobian of transformation is

$$J = \begin{vmatrix} \frac{\partial u}{\partial F} & \frac{\partial u}{\partial v} \\ \frac{\partial v}{\partial F} & \frac{\partial v}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{vv_1}{v_2} & \frac{v_1 F}{v_2} \\ 0 & 1 \end{vmatrix} = \frac{vv_1}{v_2}.$$

Substituting these values, we get

$$f(F, v) = \frac{\left(\frac{vv_1}{v_2} F\right)^{(v_1/2)-1} v^{(v_2/2)-1}}{2^{(v_1+v_2)/2} \Gamma (v_1/2) \Gamma (v_2/2)} e^{-[(vv_1/v_2)F+v]/2} \frac{vv_1}{v_2}$$

Now $f(F) = \int_0^\infty f(F, v)\, dv$

$$= \frac{(v_1/v_2)^{v_1/2} F^{(v_1/2)-1}}{2^{(v_1+v_2)/2} \Gamma(v_1/2) \Gamma(v_2/2)} \int_0^\infty v^{(v_1+v_2)/2-1} e^{-[(v_1/v_2)F+1]v/2} dv$$

Let $y = \frac{v}{2}\left(\frac{v_1}{v_2} F + 1\right)$, so that $v = 2y\left(\frac{v_1}{v_2} F + 1\right)^{-1}$ and

$$dv = 2\left(\frac{v_1}{v_2} F + 1\right)^{-1} dy.$$

Then after simplification, we get

$$f(F) = \frac{\Gamma[(v_1+v_2)/2] (v_1/v_2)^{v_1/2} F^{(v_1/2)-1}}{\Gamma (v_1/2) \Gamma (v_2/2) [1 + v_1 F/v_2]^{(v_1+v_2)/2}}, \quad 0 < F < \infty$$

as the required F-distribution with v_1 degrees of freedom in the numerator and v_2 degrees of freedom in the denominator. It is usually abbreviated as $F(v_1, v_2)$. It is interesting to note that the F-distribution does not depend upon the population variance σ^2 but depends upon the two parameters v_1 and v_2 only.

Fisher's z-distribution can be obtained by writing $F = e^{2z}$ and $dF = 2F\,dz$ in the above distribution. Fisher's z-distribution should not be confused with Fisher's z-transformation of r, the correlation co-efficient. In practice, we generally use F-statistic as it is more easy to compute and more easy to apply.

19.2.1. Properties of the F-Distribution. The F-distribution has the following important properties:

(i) The F-distribution always ranges from zero to infinity.

(ii) The mean and the variance of the F-distribution with v_1 and v_2 degrees of freedom are

$$\mu = \frac{v_2}{v_2 - 2} \text{ for } v_2 > 2, \text{ and}$$

$$\sigma^2 = \frac{2v_2^2\,(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2\,(v_2 - 4)} \text{ for } v_2 > 4.$$

Now the mean of the random variable F, defined as $F = \dfrac{U/v_1}{V/v_2}$, where U and V are independent chi-square variables with v_1 and v_2 degrees of freedom respectively, is given by

$$\mu = E(F) = E\left[\frac{U/v_1}{V/v_2}\right] = \frac{v_2}{v_1} E(U)\, E\left(\frac{1}{V}\right)$$

$$= v_2 \,.\, E\left(\frac{1}{V}\right) \quad \text{as } E(U) = v_1$$

Now $$E\left(\frac{1}{V}\right) = \frac{1}{2^{v_2/2}\,\Gamma\,(v_2/2)} \int_0^\infty \frac{1}{v} \,.\, v^{(v_2/2)-1}\, e^{-v/2}\, dv$$

Let $y = \dfrac{v}{2}$, so that $v = 2y$ and $dv = 2dy$. Then

$$E\left(\frac{1}{V}\right) = \frac{2^{(v_2/2)-1}}{2^{v_2/2}\,\Gamma\,(v_2/2)} \int_0^\infty y^{(v_2-2)/2-1}\, e^{-y}\, dy$$

$$= \frac{2^{(v_2/2)-1}}{2^{v_2/2}\,\Gamma\,(v_2/2)}\, \Gamma\left(\frac{v_2-2}{2}\right)$$

$$= \frac{1}{v_2 - 2} \qquad \left[\because \Gamma\left(\frac{v_2}{2}\right) = \frac{v_2-2}{2}\,\Gamma\left(\frac{v_2-2}{2}\right)\right]$$

Thus $\mu = v_2 E\left(\frac{1}{V}\right) = \frac{v_2}{v_2 - 2}$

Hence, we see that there is no mean for $v_2 \leq 2$. Also the mean is independent of v_1 and is always greater than 1. The result for variance is similarly established.

(iii) The F-distribution for $v_1 > 2$, $v_2 > 2$ is *unimodal* and the mode of the distribution with v_1 (≥ 2) is at $F = \frac{v_2(v_1 - 2)}{v_1(v_2 + 2)}$, which is always less than 1.

(iv) The F-distribution is skewed to the right. But as the degrees of freedom v_1 and v_2 become large, the F-distribution approaches the normal distribution.

(v) If F has an F-distribution with v_1 and v_2 degrees of freedom, then $\frac{1}{F}$ has an F distribution with v_2 and v_1 degrees of freedom. This implies that the critical value of F that cuts off a specified area of α in the lower tail of the distribution, turns out to be the reciprocal of the F value that cuts off the same area in the upper tail of the distribution with the degrees of freedom v_1 and v_2 interchanged, that is the lower and upper tail points are related by

$$F_{1-\alpha}(v_1, v_2) = \frac{1}{F_\alpha(v_2, v_1)}$$

This property is useful for testing a value of $F < 1$, where we take the reciprocal of F and interchange the degrees of freedom v_1, v_2. When $v_1 = v_2$, the distribution $\frac{1}{F}$ is the same as that of F. It is important to note that the degrees of freedom associated with the sample variance in the numerator is always stated first.

(vi) The F-distribution constitutes a wide class of distributions, depending upon v_1 and v_2. For example, a t-variable with n degrees of freedom is the ratio of a standard normal variable and the square root of a chi-square variable divided by its degrees of freedom (n) *i.e.* $t = \frac{Z}{\sqrt{\chi^2 / df}}$. The square of Z is a chi-square variable with 1 degrees of freedom. Thus

$$t^2 = \frac{Z^2}{\chi^2 / df} = \frac{\chi^2/1}{\chi^2 / n},$$

Thus $\mu = v_2 E\left(\frac{1}{V}\right) = \frac{v_2}{v_2 - 2}$

Hence, we see that there is no mean for $v_2 \leq 2$. Also the mean is independent of v_1 and is always greater than 1. The result for variance is similarly established.

(iii) The F-distribution for $v_1 > 2$, $v_2 > 2$ is *unimodal* and the mode of the distribution with v_1 (≥ 2) is at $F = \frac{v_2(v_1 - 2)}{v_1(v_2 + 2)}$, which is always less than 1.

(iv) The F-distribution is skewed to the right. But as the degrees of freedom v_1 and v_2 become large, the F-distribution approaches the normal distribution.

(v) If F has an F-distribution with v_1 and v_2 degrees of freedom, then $\frac{1}{F}$ has an F distribution with v_2 and v_1 degrees of freedom. This implies that the critical value of F that cuts off a specified area of α in the lower tail of the distribution, turns out to be the reciprocal of the F value that cuts off the same area in the upper tail of the distribution with the degrees of freedom v_1 and v_2 interchanged, that is the lower and upper tail points are related by

$$F_{1-\alpha}(v_1, v_2) = \frac{1}{F_\alpha(v_2, v_1)}$$

This property is useful for testing a value of $F<1$, where we take the reciprocal of F and interchange the degrees of freedom v_1, v_2. When $v_1 = v_2$, the distribution $\frac{1}{F}$ is the same as that of F. It is important to note that the degrees of freedom associated with the sample variance in the numerator is always stated first.

(vi) The F-distribution constitutes a wide class of distributions, depending upon v_1 and v_2. For example, a t-variable with n degrees of freedom is the ratio of a standard normal variable and the square root of a chi-square variable divided by its degrees of freedom (n) *i.e.* $t = \frac{Z}{\sqrt{\chi^2 / df}}$. The square of Z is a chi-square variable with 1 degrees of freedom. Thus

$$t^2 = \frac{Z^2}{\chi^2 / df} = \frac{\chi^2/1}{\chi^2 / n},$$

Table 19.1 Percent Points of the F-Distribution

5 Per cent Points of F, i.e. $F_{0.05}(\nu_1, \nu_2)$

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.03	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.73	1.52	1.00

Lower 5 per cent points are found by interchange of ν_1 and ν_2, *i.e.*, ν_1 must always correspond with the greater mean square.

"Table 19.1 is taken from Table V of Fisher and Yates; *Statistical Tables for Biological, Agricultural, and Medical Research*, published by Oliver & Boyd Ltd., Edinburgh, and reproduced by permission of the authors and publishers."

Table 19.2 Percent Points of the F-Distribution

2.5 Per cent Points of F, *i.e.* $F_{0.025}(v_1, v_2)$

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	956.7	976.7	997.2	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.37	39.41	39.46	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.54	14.34	14.12	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	8.98	8.75	8.51	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.76	6.52	6.28	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.60	5.37	5.12	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.90	4.67	4.42	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.43	4.20	3.95	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.10	3.87	3.61	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.85	3.62	3.37	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.66	3.43	3.17	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.51	3.28	3.02	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.39	3.15	2.89	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.29	3.05	2.79	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.20	2.96	2.70	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.12	2.89	2.63	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.06	2.82	2.56	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.01	2.77	2.50	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	2.96	2.72	2.45	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	2.91	2.68	2.41	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.87	2.64	2.37	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.84	2.60	2.33	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.81	2.57	2.30	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.78	2.54	2.27	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.75	2.51	2.24	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.73	2.49	2.22	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.71	2.47	2.19	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.69	2.45	2.17	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.67	2.43	2.15	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.65	2.41	2.14	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.53	2.29	2.01	1.64
60	5.49	3.93	3.34	3.01	2.79	2.63	2.41	2.17	1.88	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.30	2.05	1.76	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.19	1.94	1.64	1.00

Lower 2.5 per cent points are found by interchange of v_1 and v_2, *i.e.*, v_1 must always correspond with the greater mean square.

Table 19.3 Percent Points of the F-Distribution

1 Per cent Points of F, i.e. $F_{0.01}(v_1, v_2)$

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5982	6106	6234	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.61
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.03	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.94	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Lower 1 per cent points are found by interchange of v_1 and v_2, *i.e.*, v_1 must always correspond with the greater mean square.

"Table 19.3 is taken from Table V of Fisher and Yates; *Statistical Tables for Biological, Agricultural, and Medical Research*, published by Oliver & Boyd Ltd., Edinburgh, and reproduced by permission of the authors and publishers."

19.3 CONFIDENCE INTERVAL FOR THE VARIANCE RATIO σ_1^2/σ_2^2

Let two independent random samples of size n_1 and n_2 be taken from two normal populations with variances σ_1^2 and σ_2^2, and let s_1^2 and s_2^2 be the unbiased sample estimates of σ_1^2 and σ_2^2. Then we know that

$$U = \frac{(n_1 - 1)\, s_1^2}{\sigma_1^2} \text{ is a } \chi^2\text{-variable with } v_1 = n_1 - 1 \; d.f. \text{ and}$$

$$V = \frac{(n_2 - 1)\, s_2^2}{\sigma_2^2} \text{ is a } \chi^2\text{-variable with } v_2 = n_2 - 1 \; d.f.$$

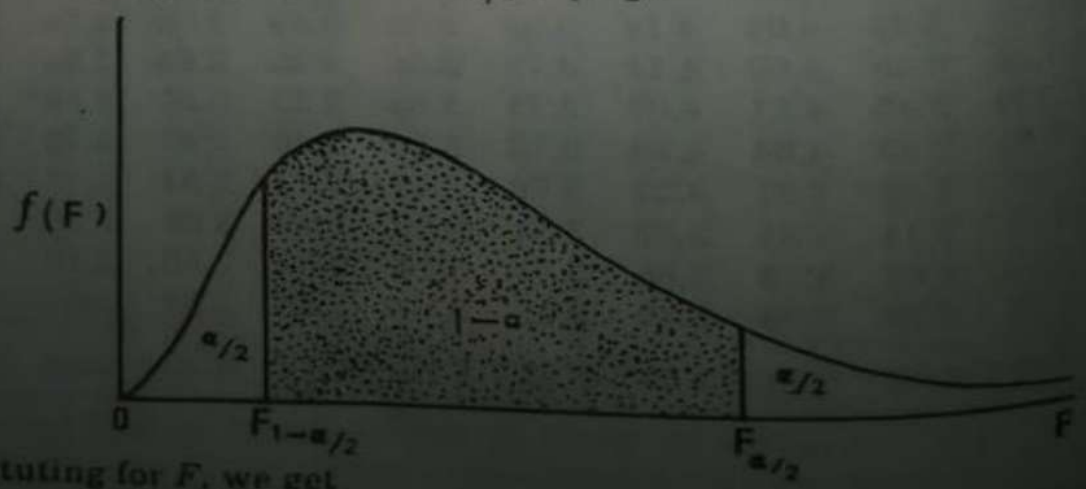
Thus, the ratio

$$F = \frac{U/v_1}{V/v_2} = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{\sigma_2^2\, s_1^2}{\sigma_1^2\, s_2^2}$$

has an F-distribution with v_1 and v_2 degrees of freedom.

To construct a $(1 - \alpha)$ 100 per cent confidence interval for the variance ratio σ_1^2/σ_2^2, we need two critical values that cut off an area of $\alpha/2$ in the lower tail and in the upper tail respectively of the F-distribution with v_1 and v_2 degrees of freedom. If these two values turn out to be $F_{1-\alpha/2}\,(v_1, v_2)$ and $F_{\alpha/2}\,(v_1, v_2)$, then we can make the following probability statement (see figure given below)

$$P\,[F_{1-\alpha/2}\,(v_1, v_2) < F < F_{\alpha/2}\,(v_1, v_2)] = 1-\alpha$$



Substituting for F, we get

$$P\left[F_{1-\alpha/2}(\upsilon_1, \upsilon_2) < \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} < F_{\alpha/2}(\upsilon_1, \upsilon_2)\right] = 1 - \alpha$$

Multiplying each term inside the bracket by s_2^2/s_1^2, and then inverting each term (we inverse the direction of inequality signs when terms are inverted), we obtain

$$P\left[\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2}(\upsilon_1, \upsilon_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\frac{1}{F_{1-\alpha/2}(\upsilon_1, \upsilon_2)}\right] = 1 - \alpha$$

But $\dfrac{1}{F_{1-\alpha/2}(\upsilon_1, \upsilon_2)} = F_{\alpha/2}(\upsilon_2, \upsilon_1)$ [property (v)]

$$\therefore \quad P\left[\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}(\upsilon_1, \upsilon_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2}(\upsilon_2, \upsilon_1)\right] = 1 - \alpha$$

Hence a $(1-\alpha)$ 100 percent confidence interval for σ_1^2/σ_2^2 is given by

$$\left[\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}(\upsilon_1, \upsilon_2)}, \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2}(\upsilon_2, \upsilon_1)\right]$$

We can also find a confidence interval for σ_1/σ_2 by taking the square root of the endpoints of this interval.

Example 19.1. Given two random samples of size $n_1=12$ and $n_2=10$ from two independent normal populations, with $s_1=2.3$ and $s_2=1.5$, find a 90% confidence interval for σ_1^2/σ_2^2 and σ_1/σ_2.

The 90% confidence interval for the ratio σ_1^2/σ_2^2 is

$$\left[\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}(\upsilon_1, \upsilon_2)}, \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2}(\upsilon_2, \upsilon_1)\right]$$

Here $s_1^2 = (2.3)^2 = 5.29$, $s_2^2 = (1.5)^2 = 2.25$, $\alpha=0.10$, $\upsilon_1=12-1=11$ and $\upsilon_2=10-1=9$.

Consulting the F-table, we find that $F_{0.05}(11, 9) = 3.10$ and $F_{0.05}(9, 11) = 2.90$. Substituting these values, we get

$$\left[\frac{5.29}{2.25}\left(\frac{1}{3.10}\right), \frac{5.29}{2.25}(2.90)\right] \text{ or } (0.76, 6.81)$$

Hence the 90% confidence interval for σ_1^2/σ_2^2 is (0.76, 6.81).

Taking the square root of the end points (0.76, 6.81), we get the 90% confidence interval for σ_1/σ_2 as (0.87, 2.61).

19.4 TESTS BASED ON *F*-DISTRIBUTION

The following tests of hypotheses are based on the *F*-distribution:

(i) Testing a hypothesis about the equality of two variances.

(ii) Testing a hypothesis about the equality of $k(k>2)$ population means.

(iii) Testing a hypothesis about linearity of regression.

(iv) Testing hypotheses about various correlation co-efficients.

The discussion of the hypotheses (a) stated at (ii), *i.e.* the hypotheses about the equality of three or more population means, is postponed until the next chapter, where we discuss one of the most important techniques in statistical analysis, known as *analysis of variance*, and (b) stated at (iii) and (iv), will be considered in Chapter 21.

19.4.1. Testing Hypothesis about the Equality of Two Variances. Suppose that we have two independent random samples of size n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2, and we wish to test the hypothesis that the two variances are equal (that is, $H_0 : \sigma_1^2/\sigma_2^2 = 1$ or equivalently $H_0 : \sigma_1^2 = \sigma_2^2$). Let s_1^2 and s_2^2 denote the unbiased estimates, based on $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Then

$$U = \frac{(n_1 - 1)\, s_1^2}{\sigma_1^2} \text{ is distributed as a } \chi^2\text{-variable with } v_1 \text{ } d.f., \text{ and}$$

$$V = \frac{(n_2 - 1)\, s_2^2}{\sigma_2^2} \text{ is distributed as a } \chi^2\text{-variable with } v_2 \text{ } d.f.$$

By definition $F = \dfrac{U/v_1}{V/v_2} = \dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}, \qquad (s_1^2 > s_2^2).$

Assuming that our null hypothesis $H_0 : \sigma_1^2/\sigma_2^2 = 1$ (that is, $H_0 : \sigma_1^2 = \sigma_2^2$) is true, the test-statistic

OPPO A1k · ©Xayn Latif

$$F = \frac{s_1^2}{s_2^2}$$

which has an F-distribution with v_1 and v_2 degrees of freedom. The computed value of F, if H_0 is true, will be relatively close to 1. If it turns out to be considerably larger than 1 (or considerably smaller than 1 if larger s^2 is not placed in the numerator), it will suggest that $\sigma_1^2/\sigma_2^2 \neq 1$ that is, $\sigma_1^2 \neq \sigma_2^2$.

The *procedure* for testing a hypothesis that the population variances σ_1^2 and σ_2^2 are equal, consists of the following steps:

(i) Formulate the null hypothesis as $H_0 : \sigma_1^2/\sigma_2^2 = 1$ (that is $H_0 : \sigma_1^2 = \sigma_2^2$). The alternative hypothesis may be

(a) $H_1 : \sigma_1^2/\sigma_2^2 > 1$, or (b) $H_1 : \sigma_1^2/\sigma_2^2 < 1$ or (c) $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$.

(ii) Decide on the significance level α.

(iii) The test-statistic to use is

$$F = \frac{s_1^2}{s_2^2}, \text{ where } s_1^2 \text{ is larger than } s_2^2.$$

which, if H_0 is true, has an F-distribution with v_1 and v_2 degrees of freedom.

(iv) Calculate the value of F from the sample data.

(v) Determine the critical region of size α from the right tail of F-distribution with v_1 and v_2 degrees of freedom.

(a) When H_1 is $\sigma_1^2/\sigma_2^2 > 1$ (*i.e.* $H_1 : \sigma_1^2 > \sigma_2^2$), the critical region will be $F \geq F_\alpha(v_1, v_2)$.

(b) When H_1 is $\sigma_1^2/\sigma_2^2 < 1$ (*i.e.* $H_1 : \sigma_1^2 < \sigma_2^2$), we interchange the role of two samples and use $F = s_2^2/s_1^2$, then the critical region will be $F \geq F_\alpha(v_2, v_1)$.

(c) When H_1 is $\sigma_1^2/\sigma_2^2 \neq 1$ (*i.e.* $H_1 : \sigma_1^2 \neq \sigma_2^2$), the critical region will be $F \geq F_{\alpha/2}(v_1, v_2)$, when $s_1^2 > s_2^2$, or

$$F \geq F_{\alpha/2}(v_2, v_1), \text{ when } s_2^2 > s_1^2.$$

This procedure avoids the use of the left-hand tail. However, if one wishes to use the left hand tail (two-sided test) also, then the critical region will be $F \geq F_{\alpha/2}(v_1, v_2)$, and $F \leq 1/F_{\alpha/2}(v_2, v_1)$.

(vi) Decide as below:

Reject H_0 if the computed value of F falls in the critical region, accept H_0 otherwise.

Example 19.2. Given two random samples of size $n_1=12$ and $n_2=10$ from two independent normal populations, with $s_1=2.3$ and $s_2=1.5$, test at 0.05 level of significance, the hypothesis $H_0:\sigma_1^2/\sigma_2^2=1$ against the alternative $H_1:\sigma_1^2/\sigma_2^2>1$.

(i) We state our null and alternative hypotheses as

$H_0 : \sigma_1^2 / \sigma_2^2 = 1$ (that is, $H_0 : \sigma_1^2 = \sigma_2^2$), and

$H_1 : \sigma_1^2 / \sigma_2^2 > 1$ (that is, $H_1 : \sigma_1^2 > \sigma_2^2$).

(ii) The level of significance is set at $\alpha=0.05$.

(iii) The test-statistic to use is

$$F = \frac{s_1^2}{s_2^2}, \text{ where } s_1^2 \text{ is larger than } s_2^2,$$

which, if H_0 is true, has an F-distribution with $v_1=11$ and $v_2=9$ degrees of freedom.

(iv) Computations. Substituting the values, we get

$$F = \frac{(2.3)^2}{(1.5)^2} = \frac{5.29}{2.25} = 2.35.$$

(v) The critical region is $F > F_{0.05}(11, 9) = 3.10$.

(vi) **Conclusion.** Since the computed value of F does not fall in the critical region, we therefore do not reject H_0 at $\alpha=0.05$ and may conclude that there is sufficient evidence to indicate that the two variances are equal.

Example 19.3. Two random samples drawn from two normal populations are:

Sample *I*: 20, 16, 26, 27, 23, 22, 18, 24, 25 and 19.

Sample *II*: 27, 33, 42, 35, 32, 34, 38, 28, 41, 43, 30, and 37.

Obtain the estimates of variances of the populations and test whether the two populations have the same variance.

(P.U., B.A./B.Sc. 1974)

(i) We state our null and alternative hypotheses as

$H_0 : \sigma_1^2 / \sigma_2^2 = 1$ (that is, $H_0 : \sigma_1^2 = \sigma_2^2$), and

$H_1 : \sigma_1^2 / \sigma_2^2 \neq 1$ (that is, $H_1 : \sigma_1^2 \neq \sigma_2^2$).

(ii) We choose the level of significance at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{s_1^2}{s_2^2} \quad (s_1^2 > s_2^2)$$

which, if H_0 is true, has an F-distribution with v_1 and v_2 *d.f.*

(iv) Computations. The two sums of squares are

$$\Sigma(X_{1i} - \bar{X}_1)^2 = \Sigma X_{1i}^2 - \frac{(\Sigma X_{1i})^2}{n_1}$$

$$= 4960 - \frac{(220)^2}{10} = 4960 - 4840 = 120, \text{ and}$$

$$\Sigma(X_{2j} - \bar{X}_2)^2 = \Sigma X_{2j}^2 - \frac{(\Sigma X_{2j})^2}{n_2}$$

$$= 15014 - \frac{(420)^2}{12} = 15014 - 14700 = 314.$$

Now we find the two estimates as

$$s_1^2 = \frac{\Sigma(X_{1i} - \bar{X}_1)^2}{n_1 - 1} = \frac{120}{9} = 13.33, \text{ and}$$

$$s_2^2 = \frac{\Sigma(X_{2j} - \bar{X}_2)^2}{n_2 - 1} = \frac{314}{11} = 28.55.$$

Since s_2^2 is larger than s_1^2, we therefore interchange the roles of the two samples and use the test statistic $F = \frac{s_2^2}{s_1^2}$. Substituting the values, we get $F = \frac{28.55}{13.33} = 2.14$.

(v) The critical region is $F \geq F_{0.025}(11, 9) = 3.92$ $(\because v_2 = 11, v_1 = 9)$

(vi) **Conclusion.** Since the computed value of F does not fall in critical region, so we do not reject H_0 and may conclude that the two populations have the same variance.

Alternatively. If we wish to use a two sided test, then the critical region will be

$$F \geq F_{0.025}(9, 11) = 3.59, \text{ and}$$

$$F \leq \frac{1}{F_{0.025}(11, 9)} = \frac{1}{3.92} = 0.26.$$

$$\text{Now} \quad F = \frac{s_1^2}{s_2^2} = \frac{13.33}{28.55} = 0.47$$

Decision. The computed value $F = 0.47$ falls in the acceptance region so we do not reject H_0 and conclude that the two populations have equal variances.

Example 19.4. In an experiment on reaction times in seconds of two individuals A and B, measured under identical conditions, the following results were obtained:

A	0.41, 0.38, 0.37, 0.42, 0.35, 0.38
B	0.32, 0.36, 0.38, 0.33, 0.38

(a) Test the hypothesis at 0.05 level of significance that $H_0 : \sigma_A^2 = \sigma_B^2$ against $H_1 : \sigma_A^2 \neq \sigma_B^2$.

(b) If $H_0 : \sigma_A^2 = \sigma_B^2$ is accepted in part(a), then test the hypothesis at 0.05 level of significance that $H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A \neq \mu_B$.

Assuming the reaction times as two random samples from two normal populations with means μ_A and μ_B and variances σ_A^2 and σ_B^2, we perform the tests as below:

(a) (i) We state our null and alternative hypothesis as

$$H_0 : \sigma_A^2 = \sigma_B^2 \text{ and } H_1 : \sigma_A^2 \neq \sigma_B^2.$$

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{s_A^2}{s_B^2},$$

which, if H_0 is true, follows an F-distribution with $v_1 = 5$ and $v_2 = 4$ $d.f.$

(iv) Computations:

A	A^2	B	B^2
0.41	0.1681	0.32	0.1024
0.38	0.1444	0.36	0.1296
0.37	0.1369	0.38	0.1444
0.42	0.1764	0.33	0.1089
0.35	0.1225	0.38	0.1444
0.38	0.1444		
2.31	0.8927	1.77	0.6297

$$\bar{A} = \frac{\sum A}{n_1} = \frac{2.31}{6} = 0.385, \quad \bar{B} = \frac{\sum B}{n_2} = \frac{1.77}{5} = 0.354.$$

$$s_A^2 = \frac{1}{n_1 - 1}\left[\sum A^2 - \frac{(\sum A)^2}{n_1}\right] = \frac{1}{5}\left[0.8927 - \frac{(2.31)^2}{6}\right] = 0.00067$$

$$s_B^2 = \frac{1}{n_2 - 1}\left[\sum B^2 - \frac{(\sum B)^2}{n_2}\right] = \frac{1}{4}\left[0.6297 - \frac{(1.77)^2}{5}\right] = 0.00078$$

$$\therefore \quad F = \frac{s_A^2}{s_B^2} = \frac{0.00067}{0.00078} = 0.86$$

(v) The critical region is $F \geq F_{0.025}(5, 4) = 9.36$, and

$$F \leq \frac{1}{F_{0.025}(4, 5)} = \frac{1}{7.39} = 0.14.$$

(vi) **Conclusion.** Since the calculated value $F = 0.86$ does not fall in the critical region, so we accept H_0, and conclude that the two population variances are equal.

(b) Since the hypothesis $H_0 : \sigma_A^2 = \sigma_B^2$ is accepted, it would, therefore, now be appropriate to test $H_0 : \mu_A = \mu_B$, using the two sample t-test.

(i) We state our hypotheses as

$$H_0 : \mu_A = \mu_B \text{ against } H_1 : \mu_A \neq \mu_B$$

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$t = \frac{\bar{A} - \bar{B}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which, if H_0 is true, follows a t-distribution with $v = 9$ $d.f.$

(iv) **Computations:** To calculate t, we need to compute s_p^2, which is

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_A^2 + (n_2 - 1)\, s_B^2}{n_1 + n_2 - 2}} = \sqrt{\frac{0.00335 + 0.00[illegible]}{6 + 5 - 2}}$$

$$= 0.0268$$

$$\therefore \quad t = \frac{0.385 - 0.354}{0.0268\sqrt{1/6 + 1/5}} = \frac{0.031}{0.016} = 1.94$$

(v) The critical region is $|t| \geq t_{0.025,(9)} = 2.26$

(vi) **Conclusion:** Since the calculated value $t = 1.94$ does not fall in the critical region, so we accept H_0 and conclude that the two means are equal.

EXERCISES

19.1 Define the F-statistic and F-distribution. Mention what assumptions we make in using the F-distribution. What hypotheses do we test with an F-distribution? Discuss them briefly.

(P.U., B.A/B.Sc. 1991)

19.2 (a) Define the variance-ratio or the F-distribution and sketch out its derivation.

(b) Describe some of the important properties of the F-distribution.

19.3 (a) Find the mean and the variance of an F-random variable with v_1 and v_2 degrees of freedom.

(b) Let F have an F-distribution with parameters v_1 and v_2. Then prove that $1/F$ has an F-distribution with parameters v_2 and v_1.

OPPO A1k · ©Xayn Latif

19.4 (a) Given two random samples from two normal populations with variances σ_1^2 and σ_2^2, explain how you will find the confidence interval for σ_1^2/σ_2^2.

(b) Given $n_1=n_2=16$, $s_1^2=50$, and $s_2^2=16$, construct a 90% confidence interval for the variance ratio σ_1^2/σ_2^2.

(c) Given $n_1=41$, $n_2=13$, $s_1^2=15.6$, and $s_2^2=6.3$, construct a 98 per cent confidence interval for σ_1^2/σ_2^2.

19.5 (a) Under what conditions is the sampling distribution of s_1^2/s_2^2 an F distribution ? Explain the relationship between the F and t distributions, between the F and χ^2 distributions.

(b) Given two random samples of size $n_1=9$ and $n_2=16$ from two independent normal populations, with $s_1=6$ and $s_2=5$, find 98% confidence intervals for σ_1^2/σ_2^2 and σ_1/σ_2.

(c) A random sample of 10 salt-water fish had a variance, s_1^2 in girth of 7.2 (inches)^2, while a random sample of 8 fresh-water fish had a variance, s_2^2 in girth of 3.6 (inches)^2. Find a 90 per cent confidence interval for the ratio between the two variances σ_2^2/σ_1^2. Assume normal populations.

19.6 (a) Describe how you would test the equality of two variances.

(b) Given the following information, what is your conclusion in testing each of the indicated null hypotheses?

	n_1	n_2	s_1^2	s_2^2	α	H_0	H_1
(i)	16	16	50	16	0.05	$\sigma_1^2/\sigma_2^2=1$	$\sigma_1^2/\sigma_2^2>1$
(ii)	13	41	6.3	15.6	0.01	$\sigma_1^2/\sigma_2^2=1$	$\sigma_1^2/\sigma_2^2<1$
(iii)	60	120	8.0	17.0	0.02	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$

19.7 (a) Two independent random samples of size $n_1=10$ and $n_2=7$ were observed to have sample variances of $s_1^2=16$ and $s_2^2=3$. Using a 10% level of significance, test $H_0 : \sigma_1^2 = \sigma_2^2$ against

$H_1 : \sigma_1^2 \neq \sigma_2^2$. Then using a 5% level of significance, test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$ and $H_1 : \sigma_1^2 < \sigma_2^2$.

(P.U., B.A/B.Sc. 1996, I.U., M.Sc. 1996)

(b) Two samples are randomly selected from two classes of students who have been taught by different methods. An examination is given and the results are shown as follows:

	Class I	Class II
Sample Size:	8	10
Means:	95	97
Unbiased sample variances:	47	50

Test whether the two different methods of teaching are equally variable. (P.U., B.A/B.Sc. 1998)

19.8 A standardized placement test in mathematics was given to 20 boys and 16 girls. The boys made an average grade of 82 with a standard deviation of 8, while the girls made an average grade of 78 with a standard deviation of 7. Test the hypothesis that $H_0 : \sigma_1^2 = \sigma_2^2$, against the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$ where σ_1^2 and σ_2^2 are the variances of the populations of grades for all boys and all girls respectively. Use a 0.02 level of significance. (P.U., B.A/B.Sc. 1986)

19.9 Independent random samples were selected from each of two normally distributed populations, $n_1 = 6$ from population 1 and $n_2 = 5$ from population 2. The data are shown below:

Sample 1: 3.1, 4.4, 1.2, 1.7, 0.7, 3.4

Sample 2: 2.3, 1.4, 3.7, 8.9, 5.5.

Do these data provide sufficient evidence to indicate a difference between the population variances? Use $\alpha = 0.05$.

19.10 The following data give the percentage extension under a given load of two independent random samples of yarn, the first before washing, the second after six washings.

Unwashed yarn: 12.3, 13.7, 10.4, 11.4, 14.9, 12.6

Washed yarn: 15.7, 10.3, 12.6, 14.5, 12.6, 13.8, 11.9.

Assuming that both samples come from normal distributions test whether there is a significant difference between the two samples:

OPPO A1k · ©Xayn Latif

(a) as regards variability,

(b) as regards the mean percentage extension.

(I.U, M.Sc. 1988; P.U., B.A/B.Sc. 1993)

19.11 The percent moisture content in a puffed cereal where samples are taken from two different "guns" showed

Gun I: 3.6, 3.8, 3.6, 3.3, 3.7, 3.4

Gun II: 3.7, 3.9, 4.2, 4.2, 4.9, 3.6, 3.5, 4.0.

Test the hypothesis of equal variances and equal means. Use any assumptions you believe appropriate. (P.U., B.A/B.Sc. 1991)

19.12 Two methods of determining moisture content of samples of canned corn have been proposed and both have been used to make determinations on proportions taken from each of 21 cans. Method I is easier to apply but appears to be more variable than method II. If the variability of Method I were not more than 25 per cent greater than that of Method II, we would prefer method I. Based on the following sample results, which method would you recommend?

$$n_1 = n_2 = 21; \quad \bar{y}_1 = 50; \quad \bar{y}_2 = 53$$

$$\Sigma(y_{1i} - \bar{y}_1)^2 = 720; \Sigma(y_{2j} - \bar{y}_2)^2 = 340 \quad \text{(P.U., M.Sc. 1970)}$$

Hint: Test $H_0: \sigma_1^2 = 1.25\sigma_2^2$ against the alternative

$$H_1: \sigma_1^2 > 1.25\ \sigma_2^2.$$

Under the null hypothesis, $s_1^2/1.25\ s_2^2$ is an F-distribution with $\nu_1 = \nu_2 = 20$. [$\because$ If $H_0: \sigma_1^2 / \sigma_2^2 = k$, then $F = (s_1^2/s_2^2)\ (1/k)$].

❖❖❖❖❖❖❖❖❖❖❖

20

The Analysis of Variance

20.1 INTRODUCTION

Earlier, we compared two population means by using a two-sample t-test. However, we are often required to compare more than two population means simultaneously. We might be tempted to apply the two-sample t-test to all possible pairwise comparisons of means. For example, if we wish to compare 4 population means, there will be $\binom{4}{2} = 6$ separate pairs and to test the null hypothesis that all four population means are equal, would require six two-sample t-tests. Similarly, to test the null hypothesis that 10 population means are equal, we would need $\binom{10}{2} = 45$ separate two-sample t-tests. This sort of running multiple two-sample t-tests for comparing means has two disadvantages. *First*, the procedure is tedious and time consuming, and *secondly*, the overall level of significance greatly increases as the number of t-tests increases. Thus a series of two-sample t-tests is not an appropriate procedure to test the equality of several means simultaneously.

Evidently, we require a procedure for carrying out a test on several means simultaneously. One such procedure is the *analysis of variance*, introduced by Sir R.A. Fisher (1890–1962) in 1923. The *analysis of variance* (abbreviated as *ANOVA*) is a technique that partitions the total variation—a term distinct from variance and measured by the sum of squares of deviations from the mean—into its component parts, each of which is associated with a different source of variation. These component parts of variance are then analysed (hence the name analysis of variance) in such a manner that certain hypotheses can be tested. This technique is based on the facts that (i) the more the sample means differ the larger the variance becomes, and (ii) the separate components provide independent and unbiased estimates of the common population

variance. The anlaysis of variance procedure therefore compares two different estimates of variance by using F-distribution to determine whether the population means are equal. The analysis of variance has been shown the most powerful and useful technique whenever the statistical data can be categorised in groups.

When each observation is classified into one sample or another according to a single criterion, we have a *one-way classification* while the classification of each observation on the basis of two criteria of classification, is called a *two-way classification*. In a similar way, a *multi-way classification* is defined. We discuss the analysis of variance procedures for the first two classifications only as a **one-way analysis of variance** and a **two-way analysis of variance** respectively.

20.2 ONE-WAY ANALYSIS OF VARIANCE

The *one-way analysis of variance* is also called the *one-variable-of-classification analysis of variance*. The data are classified into k classes or groups, etc. on the basis of a single criterion. The technical term for a class, a group or a variable of classification is *treatment*.

Suppose we have k samples of equal size r (the case of unequal sample sizes will be discussed later), selected randomly and independently, one from each of k normal populations with means $\mu_1, \mu_2, \ldots, \mu_k$ and common variance σ^2; and we wish to test the null hypothesis that all the k-population means are equal, *i.e.*

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

against the alternative hypothesis

$$H_1 : \text{Not all means are equal.}$$

Let X_{ij} denote the ith observation of the jth sample (or treatment). Then the data can be arranged as in table below:

Observation	Samples (or Treatments)						Total
	1	2	...	j	...	k	
1	X_{11}	X_{12}		X_{1j}		X_{1k}	
2	X_{21}	X_{22}		X_{2j}		X_{2k}	
.	.	.		.		.	
.	.	.		.		.	
.	.	.		.		.	
i	X_{i1}	X_{i2}		X_{ij}		X_{ik}	
.	.	.		.		.	
.	.	.		.		.	
.	.	.		.		.	
r	X_{r1}	X_{r2}		X_{rj}		X_{rk}	
Totals	$T_{.1}$	$T_{.2}$	...	$T_{.j}$	...	$T_{.k}$	$T_{..}$
Means	$\bar{X}_{.1}$	$\bar{X}_{.2}$	...	$\bar{X}_{.j}$	...	$\bar{X}_{.k}$	$\bar{X}_{..}$

Here $\bar{X}_{.j}$, $\bar{X}_{..}$, $T_{.j}$ and $T_{..}$ represent the mean of the jth sample, the overall or grand mean, total of observations in the jth sample and the total of all $rk=n$ observations respectively, where the dot replaces the subscript over which we have summed.

We test the hypothesis by comparing two independent estimates of the common population variance σ^2. The estimates of the variance can be obtained in various ways.

(i) The *first* estimate of the common population variance σ^2 is evidently obtained by pooling the k sample variances. Thus the pooled estimate of σ^2, denoted by s_p^2, is given by

$$s_p^2 = \frac{(r-1)\,s_1^2 + (r-1)\,s_2^2 + \ldots + (r-1)\,s_k^2}{n-k}, \text{ where } n=rk$$

$$= \frac{1}{n-k}\left[\sum_{i=1}^{r}(X_{i1}-\bar{X}_{.1})^2 + \sum_{i=1}^{r}(X_{i2}-\bar{X}_{.2})^2 + \ldots + \sum_{i=1}^{r}(X_{ik}-\bar{X}_{.k})^2\right]$$

$$= \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{.j})^2$$

Later, this estimate will be referred to as the **within samples** estimate of variance σ^2. This is an unbiased estimate regardless of the fact whether or not the null hypothesis is true.

(ii) The *second* estimate is based on the variation among the sample means assuming that all the population means are equal. Theoretically, the variance of the mean of a sample of size r is σ^2/r. Therefore, using the relation $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{r}$, we get $\sigma^2 = r \cdot \sigma_{\bar{x}}^2$.

If $s_{\bar{x}}^2$ is an unbiased estimate of $\sigma_{\bar{x}}^2$, then an estimate of σ^2 will be

$$rs_{\bar{x}}^2 = r\frac{1}{k-1}\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X}_{..})^2$$

Let us denote this estimate by s_b^2 as later we will call it the **between sample** estimate of variance σ^2. This estimate is independent of the *within sample* estimate as it is obtained using the means of the samples. However, this estimate will become greater than the estimate

obtained by pooling the sample variances when the sample means differ *not* true.

A *third* estimate denoted by s_T^2 can also be obtained by treating the data as one large sample consisting of n observations by the relation

$$s_T^2 = \frac{1}{n-1} \sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{..})^2$$

This is also an unbiased estimate of σ^2 when H_0 is true. This estimate is not of use in analysing the results but can be used to simplify the computations.

As s_p^2 and s_b^2 are the independent unbiased estimates of σ^2, so their values should not differ greatly. To detect this, *i.e.* to test the hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, we use the ratio

$$F = \frac{s_b^2}{s_p^2} = \frac{r \sum_{j=1}^{k} (\bar{X}_{.j} - \bar{X}_{..})^2/(k-1)}{\sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{.j})^2/(n-k)},$$

which, if H_0 is true, has an F-distribution with $\upsilon_1 = k-1$ and $\upsilon_2 = n-k$ degrees of freedom. We will reject H_0 when $F \geq F_\alpha(\upsilon_1, \upsilon_2)$ and conclude that the population means are *not* equal. It is important to note that the analysis of variance F-test is always a one-tailed test with rejection region located in the right tail of the F-distribution.

20.2.1 Partitioning the Sum of Squares, Equal Sample Sizes. The estimates of the common population variance σ^2 may be obtained by partitioning the total variation present in the k-samples (of equal size) taken from k normal populations. The variation of all $n = kr$ observations about the grand mean is measured by the expression

$$\sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{..})^2 .$$

and is called the **total sum of squares.**

To partition this total variation, let us construct the following identity:

$$X_{ij} - \bar{X}_{..} = (X_{ij} - \bar{X}_{.j}) + (\bar{X}_{.j} - \bar{X}_{..}).$$

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

OPPO A1k · ©Xayn Latif

$$\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{..})^2 = \sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{.j})^2 + \sum_{j=1}^{k}\sum_{i=1}^{r}(\bar{X}_{.j}-\bar{X}_{..})^2$$
$$+ 2\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{.j})(\bar{X}_{.j}-\bar{X}_{..}).$$

The cross-product term vanishes, because for each j

$$\sum_{i=1}^{r}(X_{ij}-\bar{X}_{.j}) = \sum_{i=1}^{r}X_{ij} - r\,\bar{X}_{.j} = \sum_{i=1}^{r}X_{ij} - r\left(\frac{\sum_{i=1}^{r}X_{ij}}{r}\right) = 0$$

The second term may be written as

$$\sum_{j=1}^{k}\sum_{i=1}^{r}(\bar{X}_{.j}-\bar{X}_{..})^2 = r\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X}_{..})^2$$

because the summation does not have i as a subscript and the factor not containing an i is considered as constant.

Hence we get the following **sum of squares** identity

$$\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{..})^2 = \sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{.j})^2 + r\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X}_{..})^2$$

which indicates that the total variation present in the samples can be partitioned into two parts. The first part is the sum of squares of deviations of the observations from the sample mean and is called the **within (samples) sum of squares**. It is also known as the **error sum of squares**. The second part is the weighted sum of the squares of deviations of the sample means from the grand mean and is called the **between sum of squares**. We can thus represent the **sum of squares** identity symbolically by the equation

$$\text{Total } SS = \text{Within } SS + \text{Between } SS$$

It is important to note that this identity holds irrespective of the fact whether or not the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is true.

Now, we obtain an unbiased estimate of σ^2 by dividing the **Between SS** by its degrees of freedom, *i.e.* by $k-1$ as there are k samples. A second unbiased estimate of σ^2 is obtained by dividing the **within SS** by an appropriate number of degrees of freedom which is $k(r-1)$ or $n-k$ as there are k samples, each containing r observations. These two quantities are known as **mean squares** and are denoted by MSB and MSW or MSE respectively. Therefore, to test the null

©Xayn Latif

hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against the alternative H_1 : not all means are equal, we form the ratio

$$F = \frac{MSB}{MSW} = \frac{\text{estimated variance from between SS}}{\text{estimated variance from within SS}}$$

which, if H_0 is true, has an F-distribution with $v_1 = k-1$ and $v_2 = n-k$ degrees of freedom. It has already been stated that, when H_0 is *not* true, MSB will be larger than σ^2, we will therefore reject H_0 at the α level of significance, if $F \geq F_\alpha(v_1, v_2)$.

20.2.2 Partitioning the Degrees of Freedom. It is interesting to note that the sum of squares identity also partitions the total number of degrees of freedom. To prove this, we take the expected values of both sides of the identity. Thus

$$E\left[\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{..})^2\right] = E\left[(n-1)\frac{\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{..})^2}{n-1}\right] \quad (\because n = rk)$$

$$= E[(n-1)\,s^2] = (n-1)\,\sigma^2,$$

$$E\left[\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{.j})^2\right] = \sum_{j=1}^{k} E\left[(r-1)\frac{\sum_{i}(X_{ij}-\bar{X}_{.j})^2}{r-1}\right]$$

$$= \sum_{j=1}^{k} E\,[(r-1)\,s_w^2]$$

$$= \sum_{j=1}^{k}(r-1)\,\sigma^2 = (n-k)\,\sigma^2, \text{ and}$$

$$E\left[r\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X}_{..})^2\right] = E\left[r\sum_{j=1}^{k}\{(\bar{X}_j-\mu)-(\bar{X}_{..}-\mu)\}^2\right]$$

$$= E\left[r\sum_{j=1}^{k}\{(\bar{X}_{.j}-\mu)^2+(\bar{X}_{..}-\mu)^2-2\,(\bar{X}_{.j}-\mu)\,(\bar{X}_{..}-\mu)\}\right]$$

$$= E\left[\sum_{j=1}^{k} r(\bar{X}_{.j}-\mu)^2 + n(\bar{X}_{..}-\mu)^2 - 2n(\bar{X}_{..}-\mu)^2\right]$$

OPPO A1k · ©Xayn Latif

$$= \sum_{j=1}^{k} r\, E(\bar{X}_{.j} - \mu)^2 - n\, E(\bar{X}_{..} - \mu)^2$$

$$= \sum_{j=1}^{k} r \frac{\sigma^2}{r} - n \frac{\sigma^2}{n}$$

$$= k\sigma^2 - \sigma^2 = (k-1)\,\sigma^2.$$

Substituting these values in the sum of squares identity, we get

$$(n-1)\,\sigma^2 = (n-k)\,\sigma^2 + (k-1)\,\sigma^2$$

Dividing both sides by σ^2, we obtain

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

Clearly the total number of degrees of freedom is $n-1$ as there is only one restriction of computing the grand mean. The *d.f.* for k samples is $k-1$, because the mean of the sample means must equal the grand mean. Similarly, the *d.f.* for **within SS** is $n-k$, due to the k restrictions of computing the k-sample means. Hence we find that

$$\text{Total } df = \text{Within } df + \text{Between } df$$

20.2.3. The Analysis of Variance Table. The various sources of variation, degrees of freedom, the sums of squares and mean squares associated with the sources are generally shown in a table, called an **analysis of variance table** or *ANOVA* table. This table is used in testing the hypothesis that the population means differ. For one-way analysis of variance, with k samples of r observations each, the analysis of variance table is shown below:

Analysis of Variance Table

Source of Variation	df	Sum of Squares (SS)	Mean Square (MS)	Computed F
Between Samples	$k-1$	$r\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X}_{..})^2 = SSB$	$s_b^2 = \frac{SSB}{k-1}$	$F = \frac{s_b^2}{s_w^2}$
Within Samples (Error)	$n-k$	$\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{.j})^2 = SSE$	$s_w^2 = \frac{SSE}{n-k}$	---
Total	$n-1$	$\sum_{j=1}^{k}\sum_{i=1}^{r}(X_{ij}-\bar{X}_{..})^2 = SST$	$s_T^2 = \frac{SST}{n-1}$	---

OPPO A1k · ©Xayn Latif

The procedure for testing the hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ using one-way analysis of variance (samples of equal sizes r) is as below:

(i) Formulate the null and alternative hypotheses as

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, and

H_1 : Not all k means are equal.

(ii) Decide upon a significance level α.

(iii) The test-statistic is

$$F = \frac{s_b^2}{s_w^2},$$

where $s_b^2 = \frac{1}{k-1}\left[r \sum_{j=1}^{k} (\bar{X}_{.j} - \bar{X}_{..})^2\right]$, and

$$s_w^2 = \frac{1}{n-k} \sum_{i=1}^{r} \sum_{j=1}^{k} (X_{ij} - \bar{X}_{.j})^2,$$

are the two estimates of the common variance σ^2, if H_0 is true. The F-statistic, if H_0 is true, has an F-distribution with $v_1 = k-1$, and $v_2 = n-k$ degrees of freedom.

(iv) Compute the necessary **Sums of Squares** and complete the analysis of variance table. Also compute F-ratio.

(v) Determine the critical region, which will consist of all values greater than or equal to $F_\alpha(k-1, n-k)$.

(vi) Decide as below:

Reject H_0 if F falls in the critical region, accept H_0 otherwise.

20.2.4. Alternative Computing Formulas. The computations of the sums of squares can be simplified as below:

$$\text{Total } SS = \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2$$

$$= \sum_i \sum_j (X_{ij}^2 + \bar{X}_{..}^2 - 2X_{ij}\bar{X}_{..})$$

$$= \sum_i \sum_j X_{ij}^2 + n\bar{X}_{..}^2 - 2\bar{X}_{..} \sum_i \sum_j X_{ij}$$

OPPO A1k · ©Xayn Latif

$$= \sum_i \sum_j X_{ij}^2 - n\bar{X}_{..}^2 = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n}$$

$$= \text{sum of squares of all values} - \frac{(\text{sum of all values})^2}{\text{number of values}}$$

Between $$SS = \sum_{j=1}^{k} r\,(\bar{X}_{.j} - \bar{X}_{..})^2$$

$$= r \sum_j \left(\frac{T_{.j}}{r} - \frac{T_{..}}{rk}\right)^2 = \frac{\sum_j T_{.j}^2}{r} - \frac{T_{..}^2}{n} \quad (\because n = rk)$$

The **Within SS** or *SSE* is usually obtained by subtracting the **Between SS** from **Total SS**. The term $\frac{T_{..}^2}{n}$ is generally called a *correction factor* (abbreviated as *CF*) as the deviations are taken from the grand mean. The arithmetic can further be simplified by choosing a convenient origin as all the *SS* are independent of origin.

Example 20.1 Given the data below, test the hypothesis that the means of the three populations are equal. Let $\alpha = 0.05$.

Sample 1	Sample 2	Sample 3
40	70	45
50	65	38
60	66	60
65	50	42

(i) We state our null and alternative hypotheses as

$H_0 : \mu_1 = \mu_2 = \mu_3$, *i.e.* all the three means are equal, and

H_1 : Not all three means are equal.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{s_b^2}{s_w^2},$$

which, if H_0 is true, has an F-distribution with $v_1 = k-1$ and $v_2 = n-k$ degrees of freedom.

1k · ©Xayn Latif

(iv) The computations are carried out as below:

	Sample 1 X_{i1} (X_{i1}^2)	Sample 2 X_{i2} (X_{i2}^2)	Sample 3 X_{i3} (X_{i3}^2)	Total	$\sum_j X_{ij}^2$
	40 (1600) 50 (2500) 60 (3600) 65 (4225)	70 (4900) 65 (4225) 66 (4356) 50 (2500)	45 (2025) 38 (1444) 60 (3600) 42 (1764)	--- --- --- ---	8525 8169 11556 8489
$T_{\cdot j}$	215	251	185	651	36739
$T_{\cdot j}^2$	46225	63001	34225	143451	↑
$\sum_i X_{ij}^2$	11925	15981	8833	36739	← check

$$\text{Correction Factor } (C.F.) = \frac{T_{..}^2}{n} = \frac{(651)^2}{12} = 35316.75$$

$$\text{Total} \quad SS = \sum_i \sum_j X_{ij}^2 - C.F.$$

$$= 36739 - 35316.75 = 1422.25$$

$$\text{Between} \quad SS = \frac{\sum_j T_{\cdot j}^2}{r} - C.F.$$

$$= \frac{143451}{4} - 35316.75 = 546.00, \text{ and}$$

Within SS = Total SS − Between SS = 1422.25 − 546.00 = 876.25.

The Analysis of Variance table is:

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between Samples	2	546.00	273.00	$\frac{273.00}{97.36} = 2.80$
Within Samples	9	876.25	97.36	----
Total Variation	11	1422.25	---	---

(v) The critical region is $F \geq F_{0\cdot05}(2, 9) = 4.26$

(vi) **Conclusion.** Since the calculated value of $F = 2.80$ does not fall in the critical region, so we accept our null hypothesis and conclude that all the three means are equal.

OPPO A1k · © Xayn Latif

20.2.5. One-Way Analysis of Variance. Unequal Sample Sizes. In the preceding sections, we discussed the one-way analysis of variance for the situation in which the k samples were all of the same size r. But, generally, the sizes of the samples are not equal. Let the k random samples be of sizes $r_1, r_2, \ldots, r_k$ respectively with $\sum_{j=1}^{k} r_j = n$. The sum of squares identity would then be written with a slight modification as

$$\sum_{i=1}^{r_j}\sum_{j=1}^{k}(X_{ij}-\bar{X}_{..})^2 = \sum_{i=1}^{r_j}\sum_{j=1}^{k}(X_{ij}-\bar{X}_{.j})^2 + \sum_{j=1}^{k} r_j(\bar{X}_{.j}-\bar{X}_{..})^2$$

The formulas for computing the Total SS and Between SS are given below:

$$\text{Total} \quad SS = \sum_{j=1}^{k}\sum_{i=1}^{r_j} X_{ij}^2 - \frac{T_{..}^2}{n}$$

$$\text{Between} \quad SS = \sum_{j=1}^{k}\frac{T_{.j}^2}{r_j} - \frac{T_{..}^2}{n}.$$

The **Within SS** is obtained by subtraction as before. For degrees of freedom, we replace rk by n, therefore the respective $d.f.$ are $n-1$, $k-1$ and $n-k$. The rest of the analysis is the same.

Example 20.2. Suppose a company makes four kinds of light bulbs and it is desired to test whether there are any differences in the durabilities of the bulbs. Random samples of sizes $n_1=5$, $n_2=10$, $n_3=7$ and $n_4=5$ are selected and the following results are obtained:

$$\bar{x}_1 = 14,\ \bar{x}_2 = 26,\ \bar{x}_3 = 17,\ \bar{x}_4 = 22,\ s_1^2 = 10,\ s_2^2 = 33\tfrac{1}{3},\ s_3^2 = 28,\ s_4^2 = 54,$$

where s^2 is the estimate of the population variance from the sample. Perform the analysis of variance to determine whether the service lives of the four kinds of bulbs do not differ from one another at $\alpha = 0.01$.

(P.U., B.A./B.Sc. 1986)

(i) The null and the alternative hypotheses corresponding to the problem that the service lives of the bulbs do not differ from one another, are formulated as

A1k · ©Xayn Latif

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4, \text{ and}$$

$$H_1 : \text{Not all four means are equal.}$$

(ii) The significance level is set at $\alpha = 0.01$.

(iii) The test-statistic to use is

$$F = \frac{s_b^2}{s_w^2},$$

which, if H_0 is true, has an F-distribution with $v_1 = k-1$ and $v_2 = n-k$ degrees of freedom.

(iv) Computations of sums of squares. To compute the necessary sums of squares, we have to first compute the grand (overall) mean $\bar{X}_{..}$ as

$$\bar{X}_{..} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3 + n_4\bar{X}_4}{n_1+n_2+n_3+n_4}$$

$$= \frac{5(14) + 10(26) + 7(17) + 5(22)}{5 + 10 + 7 + 5} = \frac{559}{27} = 20.7$$

Now

$$\text{Between } SS = \sum_{j=1}^{k} n_j (\bar{X}_{.j} - \bar{X}_{..})^2 \quad [\text{or } \sum n_j \bar{X}_{.j}^2 - (\sum n_j \bar{X}_j)^2/n]$$

$$= [5(14-20.7)^2 + 10(26-20.7)^2 + 7(17-20.7)^2 + 5(22-20.7)^2]$$

$$= 224.45 + 280.90 + 95.83 + 8.45 = 609.63, \text{ and}$$

$$\text{Within } SS = (n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + (n_3 - 1)\, s_3^2 + (n_4 - 1)\, s_4^2$$

$$= 4(10) + 9(33\tfrac{1}{3}) + 6(28) + 4(54)$$

$$= 40 + 300 + 168 + 216 = 724$$

The analysis of variance table becomes

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between bulbs	3	609.63	203.21	6.46
Within bulbs	23	724.00	31.48	---
Total variation	26	1333.63	---	---

PPO A1k · ©Xayn Latif

(v) The critical region is $F \geq F_{0.01}(3, 23) = 4.76$

(vi) **Conclusion.** Since the computed value of $F=6.46$ falls in the critical region, we therefore reject H_0. This implies that the data present sufficient evidence to indicate that the service lives of the four kinds of bulbs do differ from one another at 0.01 level of significance.

Example 20.3. The students in 3 classes in an elementary statistics course obtained total scores as in the table:

8-o'clock: 121, 117, 145, 108, 142, 154, 115, 81, 122, 127, 122

10-o'clock: 97, 145, 119, 139, 143, 133, 149, 107, 154.

2-o'clock: 134, 89, 108, 88, 146, 153, 130, 144, 125, 111, 87, 162.

Is there a significant difference in the scores received by the students meeting at different times of day? State completely the hypothesis you are testing and your conclusions.

(P.U., B.A./B.Sc. Hons., 1961)

(i) The *null* and *alternative hypotheses* corresponding to the problem that there is no significant difference in the scores received by students meeting at different times of day, are stated as

$$H_0 : \mu_1 = \mu_2 = \mu_3, \text{ and}$$

$$H_1 : \text{Not all the means are equal.}$$

(ii) Let us choose the level of significance at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{s_b^2}{s_w^2},$$

which, if H_0 is true, has an F-distribution with $\nu_1 = 2$, $\nu_2 = 23$ degrees of freedom.

(iv) Computations of sums of squares. To make the computational work easier, we choose our origin at $X=100$. Then the computations are given as follows:

O A1k · ©Xayn Latif

	Times of the Day				
	8-o'clock	10-o'clock	2-o'clock	Total	$\sum_j X_{ij}^2$
	21 (441)	−3 (9)	34 (1156)	---	1606
	17 (289)	45 (2025)	−11 (121)	---	2435
	45 (2025)	19 (361)	8 (64)	---	2450
	8 (64)	39 (1521)	−12 (144)	---	1729
	42 (1764)	43 (1849)	46 (2116)	---	5729
	54 (2916)	33 (1089)	53 (2809)	---	6814
	15 (225)	49 (2401)	30 (900)	---	3526
	−19 (361)	7 (49)	44 (1936)	---	2346
	22 (484)	54 (2916)	25 (625)	---	4025
	27 (729)	--- ---	11 (121)	---	850
	22 (484)	--- ---	−13 (169)	---	653
	-- --	--- ---	62 (3844)	---	3844
$T_{\cdot j}$	254	286	277	817	36007
$T_{\cdot j}^2$	64516	81796	76729	--	↑
$\sum_i X_{ij}^2$	9782	12220	14005	36007	←check

$$C.F. = \frac{T_{..}^2}{n} = \frac{(817)^2}{32} = 20859$$

Total $SS = \sum_i \sum_j X_{ij}^2 - C.F. = 36007 - 20859 = 15148$

Between $SS = \sum_j \frac{T_{\cdot j}^2}{n_j} - C.F. = \left(\frac{64516}{11} + \frac{81796}{9} + \frac{76729}{12}\right) - C.F.$

$= (5865 + 9088 + 6394) - 20859 = 488.$

$\therefore$ Within SS = Total SS − Between SS = 14660.

The Analysis of Variance table is:

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Between 'times'	2	488	244.0	0.48
Within 'times'	29	14,660	505.5	---
Total variation	31	15,148	---	---

(v) The critical region is $F \geq F_{0\cdot05}(2, 29) = 3.33$.

(vi) **Conclusion.** Since the computed value of $F = 0.48$ does not fall in the critical region, so we accept H_0 and may conclude that there is no significant difference in the scores received at different times of day.

OPPO A1k · © Xayn Latif

Example 20.4. The following table contains the scores obtained by students in three sections of statistics class. Apply the Analysis of Variance technique to test the homogeneity of their achievements.

Score-Interval	Sections 1	2	3	Total
0 – 4	3	--	10	13
5 – 9	8	7	12	27
10 – 14	17	26	8	51
15 – 19	12	4	5	21
20 – 24	6	3	--	9
Total	46	40	35	121

(i) The hypotheses corresponding to the problem that the sections are equal in achievement, are formulated as

$H_0 : \mu_1 = \mu_2 = \mu_3$, and

H_1 : Not all three means are equal.

(ii) The significance level is chosen at $\alpha = 0.05$.

(iii) The test statistic to use is

$$F = \frac{\text{Between Mean Square}}{\text{Within Mean Squre}},$$

which, if H_0 is true, has an F-distribution with $v_1 = k-1$, $v_2 = n-k$ degrees of freedom.

(iv) Computations of sums of squares. For this purpose, we first calculate $f_j x_i$ and $f_j x_i^2$ (j = 1, 2, 3) for the three sections as:

Mid Values	Section 1			Section 2			Section 3		
x_i	f_1	f_1x_i	$f_1x_i^2$	f_2	f_2x_i	$f_2x_i^2$	f_3	f_3x_i	$f_3x_i^2$
2	3	6	12	0	0	0	10	20	40
7	8	56	392	7	49	343	12	84	588
12	17	204	2448	26	312	3744	8	96	1152
17	12	204	3468	4	68	1156	5	85	1445
22	6	132	2904	3	66	1452	0	0	0
Total	46	602	9224	40	495	6695	35	285	3225

k · ©Xayn Latif

The sums and sums of squares of scores obtained from each section are tabulated below to facilitate computational work:

Section	No. of Students ($\sum f$)	Sums of Scores ($\sum fx$)	SS of Scores ($\sum fx^2$)
1	46	602	9224
2	40	495	6695
3	35	285	3225
Total	121	1382	19144

$$\text{Now Total } SS = \sum_i \sum_j X_{ij}^2 - \frac{T_{..}^2}{n} = 19144 - \frac{(1382)^2}{121} = 3360.33,$$

$$\text{Between } SS = \sum_j \frac{T_{.j}^2}{r_j} - \frac{T_{..}^2}{n} = \frac{(602)^2}{46} + \frac{(495)^2}{40} + \frac{(285)^2}{35} - \frac{(1382)^2}{121}$$

$$= 16324.68 - 15783.67 = 541.01, \text{ and}$$

Within SS is obtained by subtraction.

The analysis of variance table is therefore:

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between Sections	2	541.01	270.50	11.74
Within Sections	118	2819.32	23.89	---
Total	120	3360.33	---	---

(v) The critical region is $F \geq F_{0\cdot05}(2, 118) = 3.07$

(vi) **Conclusion.** Since the calculated value of $F = 11.74$ falls in the critical region, so we reject our null hypothesis. The data present sufficient evidence to indicate that the sections differ in their achievement.

20.2.6. Assumptions of One-Way Analysis of Variance. The one-way analysis of variance test is based on the following assumptions:

(i) The k-samples are selected randomly and independently from the respective populations.

(ii) All the k populations from which the samples are drawn, are normally distributed with means $\mu_1, \mu_2, \ldots, \mu_k$.

OPPO A1k · ©Xayn Latif

(iii) The normal populations all have equal variances (that is, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$). The technical tern. for this assumption is *Homoscedasticity*.

(iv) The effects are additive. This means that X_{ij}, the ith observation in the jth sample, is made up of three component quantities as follows:

$$X_{ij} = \mu + \tau_j + \varepsilon_{ij},$$

where μ is the overall mean, τ_j is the sample or treatment effect for jth population and ε_{ij} is the random error, usually considered a normally and independently distributed variable with zero mean and common variance σ^2.

In practice, these assumptions must be checked before proceeding with the "F-test" for the equality of means. Failure of any assumption will impair the technique. However, investigations have shown that minor deviations from normality and equality of variances are to be tolerated. Sometimes, an appropriate transformation such as using the square root or logs, etc., is made to satisfy these assumptions.

20.3 TWO-WAY ANALYSIS OF VARIANCE

When each observation is classified according to two criteria (or variables)of classification simultaneously, we use the two-way analysis of variance technique. The classified data are recorded in a table, in which the columns represent one criterion (or variable) of the classification and the rows represent the other criterion. If there are c-columns and r-rows in the table, then there will be altogether rc cells. Each cell may contain a single observation or several observations.

There are two basic forms of two-way analysis of variance, depending upon whether the two variables of classification are *independent* or whether they *interact*. Two variables (or criteria) of classification are said to *interact* when they *together* have an added effect that they do not have individually. For example, suppose that we classify salesmen according to, say, age and educational attainment with a view to determining whether age and education have significant effect on the volume of sales. Suppose we observe that the two variable individuall, do not produce significant effects but certain age-groups when *combined* with particular educational attainment, produce significant effects. We then say that there is an *interaction* between age and educational attainment. On the other hand, if a particular combination of certain age-group and educational attainment does not produce any significant

O A1k · ©Xayn Latif

effect, the variables are *independent*. When variables of classification are independent, one observation per cell is recorded. In case of interaction, several observations are made for each cell. We consider both the cases in the following subsections.

20.3.1. Two-Way Analysis of Variance without Interaction. Let X_{ij} denote an observation in the ith row and the jth column in a table consisting of r rows and c columns and containing sample data from normal populations with means μ_{ij} and the common variance σ^2, classified according to two criteria of classification simultaneously. Let $\mu_{i\cdot}$ represent the population mean of the *ith* row and $\mu_{\cdot j}$, that of the *jth* column. Denoting the total and mean of c-values in the *ith* row by $T_{i\cdot}$ and $\bar{X}_{i\cdot}$, the total and mean of r-values in the *jth* column by $T_{\cdot j}$ and $\bar{X}_{\cdot j}$ and the grand total and grand mean by $T_{\cdot\cdot}$ and $\bar{X}_{\cdot\cdot}$, the results are shown in the following form:

Rows	Columns						Total	Mean
	1	2	...	j	...	c		
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1c}	$T_{1\cdot}$	$\bar{X}_{1\cdot}$
2	X_{21}	X_{22}	...	X_{2j}	...	X_{2c}	$T_{2\cdot}$	$\bar{X}_{2\cdot}$
.	.	.		.		.	.	.
.	.	.		.		.	.	.
.	.	.		.		.	.	.
i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ic}	$T_{i\cdot}$	$\bar{X}_{i\cdot}$
.	.	.		.		.	.	.
.	.	.		.		.	.	.
.	.	.		.		.	.	.
r	X_{r1}	X_{r2}		X_{rj}	...	X_{rc}	$T_{r\cdot}$	$\bar{X}_{r\cdot}$
Totals	$T_{\cdot 1}$	$T_{\cdot 2}$	...	$T_{\cdot j}$	...	$T_{\cdot c}$	$T_{\cdot\cdot}$	
Means	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$	...	$\bar{X}_{\cdot j}$	...	$\bar{X}_{\cdot c}$		$\bar{X}_{\cdot\cdot}$

There are now two null hypotheses, one corresponding to the problem that all the r-row-means are equal; and the other corresponding to the problem that all the c-column-means are equal. Thus the two null hypotheses are

$$H_0' : \mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{r\cdot},$$

$$H_0'' : \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot c},$$

and the alternative hypotheses to be considered, are

H_1': Not all $\mu_{i\cdot}$ are equal,

H_1'': Not all $\mu_{\cdot j}$ are equal.

We test these hypotheses by comparing independent estimates of the common population variance σ^2. The estimates of the variance can be obtained by partitioning the total sum of squares into three components corresponding to the three possible sources of variation, viz; **Between Rows, Between Columns** and **Within Samples or Error.** For this purpose, we construct the following identity:

$$X_{ij} - \bar{X}_{..} = (\bar{X}_{i\cdot} - \bar{X}_{..}) + (\bar{X}_{\cdot j} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})$$

Squaring both sides and summing over all values, we get

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{X}_{i\cdot} - \bar{X}_{..})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{X}_{\cdot j} - \bar{X}_{..})^2$$

$$+ \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})^2 + \text{three cross-product}$$

terms, which all reduce to zero.

Since $(\bar{X}_{i\cdot} - \bar{X}_{..})^2$ is identically the same for each row of the c values and $(\bar{X}_{\cdot j} - \bar{X}_{..})^2$ is the same for each observation in the jth column of r values, the **sum-of-squares identity** becomes

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X}_{..})^2 = c\sum_{i=1}^{r}(\bar{X}_{i\cdot} - \bar{X}_{..})^2 + r\sum_{j=1}^{c}(\bar{X}_{\cdot j} - \bar{X}_{..})^2 +$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})^2, \text{ i.e.}$$

Total SS = Between Row-means SS + Between Column-means SS + Within or Error SS.

We may write this identity briefly in the following symbolic from

$$SST = SSR + SSC + SSE.$$

Here we get four estimates of the population variance σ^2. The first estimate is based on $r-1$ *d.f.* and is given by

$$s_1^2 = \frac{SSR}{r-1}$$

When the hypothesis that row means are equal, is true, s_1^2 is an unbiased estimate of σ^2, on the contrary, it will have a larger value.

The second estimate based on $c-1$ *d.f.*, is given by

$$s_2^2 = \frac{SSC}{c-1}$$

If the hypothesis that the column means are equal, is true, s_2^2 is an unbiased estimate of σ^2, otherwise it will also have a larger value.

The third estimate based on $(r-1)(c-1)$ *d.f.*, is given by

$$s_3^2 = \frac{SSE}{(r-1)(c-1)}$$

which is an unbiased estimate of σ^2 irrespective of the fact whether the hypotheses are true or false.

It has been shown by a theorem due to Cochran that the two estimates s_1^2 and s_2^2, derived from the "Between row-means SS" and the "Between column-mean SS" are independent of s_3^2, the estimate derived from the error sum of squares. Hence to test the hypothesis that the row-means are equal, we compute the statistic

$$F_1 = \frac{s_1^2}{s_3^2},$$

which, when the hypothesis is true, has an F-distribution with $\upsilon_1 = r-1$, $\upsilon_2 = (r-1)(c-1)$ *d.f.* We will reject the hypothesis at the α level of significance, when $F_1 \geq F_\alpha; [(r-1), (r-1)(c-1)]$.

Similarly, to test the null hypothesis that the column-means are equal, we compute the ratio

$$F_2 = \frac{s_2^2}{s_3^2},$$

which has an F-distribution with $\upsilon_1 = c-1$, $\upsilon_2 = (r-1)(c-1)$ *d.f.*, if the null hypothesis is true. Hence we reject the null hypothesis at the α level of significance, when

$$F_2 \geq F_\alpha; [(c-1), (r-1)(c-1)].$$

OPPO A1k · ©Xayn Latif

It has also been verified that the degrees of freedom associated with different sums of squares, are additive. That is

$$(rc-1) = (r-1) + (c-1) + (r-1)(c-1)$$

In other words,

Total *d.f.* = Rows *d.f.* + Columns *d.f.* + Error *d.f.*

These results are summarised in the following ANOVA Table:

ANOVA-Table for a Two-Way Analysis of Variance without Interaction

Source of Variation	*df*	Sum of Squares	Mean Square	*F*
Between Rows	$r-1$	$SSR = c\sum_{i=1}^{r}(\bar{X}_{i.}-\bar{X}_{..})^2$	$s_1^2 = \frac{SSR}{r-1}$	$F_1 = \frac{s_1^2}{s_3^2}$
Between Columns	$c-1$	$SSC = r\sum_{j=1}^{c}(\bar{X}_{.j}-\bar{X}_{..})^2$	$s_2^2 = \frac{SSC}{c-1}$	$F_2 = \frac{s_1^2}{s_3^2}$
Error (Within)	$(r-1)\times(c-1)$	SSE = By subtraction	$s_3^2 = \frac{SSE}{(r-1)(c-1)}$	--
Total	$rc-1$	$SST = \sum_i\sum_j(X_{ij}-\bar{X}_{..})^2$	--	---

In practice, we use the following short-cut methods for computing the **Total SS, Between Row-means SS and Between Column-means SS.**

$$\text{Total } SS = \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\bar{X}_{..})^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}X_{ij}^2 - \frac{T_{..}^2}{rc};$$

$$SSR = c\sum_{i=1}^{r}(\bar{X}_{i.}-\bar{X}_{..})^2 = \sum_{i=1}^{r}\frac{T_{i.}^2}{c} - \frac{T_{..}^2}{rc}; \text{ and}$$

$$SSC = r\sum_{j=1}^{c}(\bar{X}_{.j}-\bar{X}_{..})^2 = \sum_{j=1}^{c}\frac{T_{.j}^2}{r} - \frac{T_{..}^2}{rc}.$$

The Error *SS* is usually obtained by subtraction.

The *procedure* for testing the null hypotheses is similar to that used for one-way analysis of variance.

PPO A1k · ©Xayn Latif

Example 20.5. Four experimenters determine the moisture content of samples of a powder, each man taking a sample of six consignments. Their assessments are:

Observers	Consignments					
	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Perform a two-way analysis of variance on these data and discuss whether there is any significant difference between consignments or between observers. (P.U., B.A/B.Sc. (Hons.) 1970)

(i) We set up the two null hypotheses corresponding to the problems that

(a) there is no significant difference between consignments, and

(b) there is no significant difference between observers, as

$$H_0' : \mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3} = \mu_{\cdot 4} = \mu_{\cdot 5} = \mu_{\cdot 6}, \text{ and}$$

$$H_0'' : \mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot} = \mu_{4\cdot}$$

The corresponding alternative hypotheses would be

$$H_1' : \text{Not all } \mu_{\cdot j} \text{ are equal,}$$

$$H_1'' : \text{Not all } \mu_{i\cdot} \text{ are equal,}$$

(ii) We choose the level of significance at $\alpha = 0.05$.

(iii) The test-statistics to use are

$$F_1 = \frac{\text{estimated variance from "Between Consignments } SS\text{"}}{\text{estimated variance from "Error } SS\text{"}} = \frac{s_1^2}{s_3^2},$$

and $$F_2 = \frac{\text{estimated variance from "Between Observers } SS\text{"}}{\text{estimated variance from "Error } SS\text{"}} = \frac{s_2^2}{s_3^2}.$$

OPPO A1k · ©Xayn Latif

which have F-distributions with $\nu_1=5$, $\nu_2=15$ and $\nu_1=3$ $\nu_2=15$ d.f. respectively, when the null hypotheses are true.

(iv) **Computations.** The necessary sums of squares are computed as shown in the table below:

Observers	Consignments (Figures in brackets are the squares of X_{ij}) 1	2	3	4	5	6	$T_{i\cdot}$	$T_{i\cdot}^2$	$\sum_j X_{ij}^2$
1	9 (81)	10 (100)	9 (81)	10 (100)	11 (121)	11 (121)	60	3600	604
2	12 (144)	11 (121)	9 (81)	11 (121)	10 (100)	10 (100)	63	3969	667
3	11 (121)	10 (100)	10 (100)	12 (144)	11 (121)	10 (100)	64	4096	686
4	12 (144)	13 (169)	11 (121)	14 (196)	12 (144)	10 (100)	72	5184	874
$T_{\cdot j}$	44	44	39	47	44	41	259	16849	2831
$T_{\cdot j}^2$	1936	1936	1521	2209	1936	1681	11219		↑
$\sum_i X_{ij}^2$	490	490	383	561	486	421	2831	←	Check

Now Total SS

$$= \sum_i \sum_j X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{rc}$$

$$= 2831 - \frac{(259)^2}{24}$$

$$= 2831 - 2795.04 = 35.96,$$

Between Consignments SS

$$= \sum_j \frac{T_{\cdot j}^2}{r} - \frac{T_{\cdot\cdot}^2}{rc}$$

$$= \frac{11219}{4} - 2795.05 = 9.71,$$

Between Observers SS

$$= \sum_i \frac{T_{i\cdot}^2}{c} - \frac{T_{\cdot\cdot}^2}{rc}$$

$$= \frac{16849}{6} - 2795.04 = 13.13, \text{ and}$$

Error $SS = 35.96 - (9.71 + 13.13) = 13.12$

The ANOVA - Table is

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between Consignments	5	9.71	1.94	$F_1 = \frac{1.94}{0.87} = 2.23$
Between Observers	3	13.13	4.38	$F_2 = \frac{4.38}{0.87} = 5.03$
Error	15	13.12	0.87	---
Total	23	35.96	---	---

(v) The critical regions are (a) $F \geq F_{0\cdot05}(5, 15) = 2.90$,
(b) $F \geq F_{0\cdot05}(3, 15) = 3.29$.

(vi) **Conclusion.** Since the computed value of $F_1 = 2.23$ does not fall in the critical region but the computed value of $F_2 = 5.03$ falls in the critical region, so we accept the hypothesis relating to the consignments, and we reject the hypothesis corresponding to the fact that there is no significant difference between observers.

20.3.2. Two-Way Analysis of Variance with Interaction. In the preceding sub-section, we arranged the data classified on the basis of two variables of classification simultaneously in a rectangular form with r-rows and c-columns, thus having rc cells. Supposing that each cell contains n observations, the rcn observations can then be displayed as in the following table, where X_{ijk} will denote the kth observation in the ith row and the jth column.

Rows	Columns 1	2	...	j	...	c	Total	Mean
1	X_{111} X_{112} ⋮ X_{11n}	X_{121} X_{122} ⋮ X_{12n}	...	X_{1j1} X_{1j2} ⋮ X_{1jn}	...	X_{1c1} X_{1c2} ⋮ X_{1cn}	$T_{1\cdot\cdot}$	$\bar{X}_{1\cdot\cdot}$
...	...	...	...	...	...	...	...	...
i	X_{i11} X_{i12} ⋮ X_{i1n}	X_{i21} X_{i22} ⋮ X_{i2n}	...	X_{ij1} X_{ij2} ⋮ X_{ijn}	...	X_{ic1} X_{ic2} ⋮ X_{icn}	$T_{i\cdot\cdot}$	$\bar{X}_{i\cdot\cdot}$
...	...	...	...	...	...	...	...	...
r	X_{r11} X_{r12} ⋮ X_{r1n}	X_{r21} X_{r22} ⋮ X_{r2n}	...	X_{rj1} X_{rj2} ⋮ X_{rjn}	...	X_{rc1} X_{rc2} ⋮ X_{rcn}	$T_{r\cdot\cdot}$	$\bar{X}_{r\cdot\cdot}$
Total	$T_{\cdot1\cdot}$	$T_{\cdot2\cdot}$	...	$T_{\cdot j\cdot}$	...	$T_{\cdot c\cdot}$	$T_{\cdots}$	--
Mean	$\bar{X}_{\cdot1\cdot}$	$\bar{X}_{\cdot2\cdot}$	...	$\bar{X}_{\cdot j\cdot}$	...	$\bar{X}_{\cdot c\cdot}$	--	$\bar{X}_{\cdots}$

A1k · ©Xayn Latif

The totals of the observations in the ith row, in the jth column, in the (i, j)th cell and the total of all rcn observations are denoted by $T_{i..}$, $T_{.j.}$, $T_{ij.}$ and $T_{...}$ respectively.

Similarly $\bar{X}_{i..}$, $\bar{X}_{.j.}$, $\bar{X}_{ij.}$ and $\bar{X}_{...}$ denote the means of the observations in the ith row, in the jth column, in the (i, j)th cell and the mean of all rcn observations respectively.

We further assume that the observations in the (i, j)th cell are a random sample of size n from a normal population with mean μ_{ij} and variance σ^2 and that all rc populations have the same variance σ^2.

Then there are three hypotheses to be tested, namely

(i) H_0' : that the row-means are equal, against

H_1' : Not all row-means are equal.

(ii) H_0'' : that the column-means are equal, against

H_1'' : Not all column-means are equal, and

(iii) H_0''' : that there is no interaction, against

H_1''' : that the interaction effects are not all equal to zero.

We test these hypotheses by comparing independent estimates of the common variance σ^2. The estimates of the variance are obtained by partitioning the total sum of squares into four components, corresponding to the four possible sources of variation, viz., **Between Columns, Between Rows, an Interaction and an additional part** involving the variation **Within cells** about the cell means.

For this purpose, we construct the following identity:

$$(X_{ijk}-\bar{X}_{...}) = (\bar{X}_{.j.}-\bar{X}_{...}) + (\bar{X}_{i..}-\bar{X}_{...}) + (\bar{X}_{ij.}-\bar{X}_{.j.}-\bar{X}_{i..}+\bar{X}_{...}) + (X_{ijk}-\bar{X}_{ij.})$$

Squaring both sides and summing over all values, we get

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(X_{ijk}-\bar{X}_{...})^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(\bar{X}_{.j.}-\bar{X}_{...})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(\bar{X}_{i..}-\bar{X}_{...})^2 +$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(\bar{X}_{ij.}-\bar{X}_{.j.}-\bar{X}_{i..}+\bar{X}_{...})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(X_{ijk}-\bar{X}_{ij.})^2 + 6 \text{ cross-product}$$

terms that vanish when summed. Hence we are left with after simplification

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(X_{ijk}-\bar{X}_{...})^2 = rn\sum_{j=1}^{c}(\bar{X}_{.j.}-\bar{X}_{...})^2 + cn\sum_{i=1}^{r}(\bar{X}_{i..}-\bar{X}_{...})^2$$

$$+ n\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{X}_{ij.}-\bar{X}_{.j.}-\bar{X}_{i..}+\bar{X}_{...})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(X_{ijk}-\bar{X}_{ij.})^2.$$

k · ©Xayn Latif

We may write this sum-of-squares identity symbolically as

Total SS = Between Columns SS + Between Rows SS + Interaction SS + Error SS

or more briefly as $SST = SSC + SSR + SS(RC) + SSE$.

The corresponding breakdown of the degrees of freedom is given by the relation

$$(rcn - 1) = (c - 1) + (r - 1) + (c - 1)(r - 1) + rc(n - 1)$$

Dividing these sums of squares by their corresponding number of degrees of freedom, we get the variance estimates as

$$s_1^2 = \frac{SSC}{c-1},\ s_2^2 = \frac{SSR}{r-1},\ r_3^2 = \frac{SS(RC)}{(c-1)(r-1)},\ \text{and}\ s_4^2 = \frac{SSE}{rc(n-1)}$$

All these estimates, if all the three null hypotheses are true, provide unbiased estimates of σ^2. Thus the corresponding variance ratios that we compute, are

$$F_1 = \frac{s_1^2}{s_4^2},\ F_2 = \frac{s_2^2}{s_4^2}\ \text{and}\ F_3 = \frac{s_3^2}{s_4^2},$$

which have F-distributions with $v_1 = c-1, v_2 = rc(n-1)$; $v_1 = r-1, v_2 = rc(n-1)$ and $v_1 = (c-1)(r-1)$, $v_2 = rc(n-1)$ *d.f.* respectively when the null hypotheses are true.

We reject the null hypotheses as usual when $F \geq F_\alpha(v_1, v_2)$.

These results are summarised in the following ANOVA Table:

ANOVA-Table for a Two-Way Analysis of Variance with Interaction

Source of Variation	*d.f.*	Sum of Squares	Mean Square	F
Between Column means	$c-1$	SSC	$s_1^2 = \frac{SSC}{c-1}$	$F_1 = \frac{s_1^2}{s_4^2}$
Between Row means	$r-1$	SSR	$s_2^2 = \frac{SSR}{c-1}$	$F_2 = \frac{s_2^2}{s_4^2}$
Interaction (between rows and columns)	$(c-1) \times (r-1)$	$SS(RC)$	$s_3^2 = \frac{SS(RC)}{(c-1)(r-1)}$	$F_3 = \frac{s_3^2}{s_4^2}$
Error	$rc(n-1)$	SSE	$s_4^2 = \frac{SSE}{rc(n-1)}$	----
Total	$rcn - 1$	SST	--	--

OPPO A1k · ©Xayn Latif

When interactions are zero, we may pool the Interaction SS with the Error SS. This increases the number of df for error and hence increases the precision of further testing.

In practice, the sums of squares are computed by the following formulas:

$$\text{Total } SS = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n} X_{ijk}^2 - \frac{T_{...}^2}{rcn},$$

where $\frac{T_{...}^2}{rcn}$ is the correction factor (C.F.)

$$\text{Between Column-means } SS = \sum_{j=1}^{c}\frac{T_{.j.}^2}{rn} - C.F.,$$

$$\text{Between Row-means } SS = \sum_{i=1}^{r}\frac{T_{i..}^2}{cn} - C.F.,$$

$$\text{Interaction } SS = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{T_{ij.}^2}{n} - \sum_{i=1}^{r}\frac{T_{i..}^2}{cn} - \sum_{j=1}^{c}\frac{T_{.j.}^2}{rn} + C.F.,$$

and Error SS as before, is obtained by subtraction.

Example 20.6. The following data represent the result of 3 questions obtained by 3 students in three subjects:

Student	Subjects		
	English	Mathematics	Statistics
1	13	23	22
	18	20	23
	15	13	20
2	21	20	20
	16	14	15
	24	24	22
3	18	17	19
	15	13	21
	12	16	18

Perform an anlaysis of variance upon these data and test the hypothesis that (a) the subjects are of equal difficulty.

(b) the students are of equal ability, and

(c) the students and subjects do not interact.

O A1k · ©Xayn Latif

(i) We state the null and alternative hypotheses as

(a) H_0' : The subjects are of equal difficulty, *i.e.* the difference between column means is zero:

H_1' : Not all subjects are of equal difficulty:

(b) H_0'' : The students have equal ability or the row-means are equal:

H_1'' : Not all students have equal ability: and

(c) H_0''' : There is no interaction between students and subjects:

H_1''' : The interaction effects are not all equal to zero.

(ii) We choose the level of significance at $\alpha = 0.05$.

(iii) The test statistic to use is the usual variance ratio, *i.e.*

$$F = \frac{\text{estimated variance from "Between } SS\text{"}}{\text{estimated variance from "Error } SS\text{"}}$$

which, if the null hypotheses are true, has F-distribution with v_1, v_2 *d.f.*

(iv) Computations. To compute the sums of squares, we first construct a table containing the totals of the observations in the (i, j)th cell, *i.e.*, T_{ij}. The figures within brackets are the squares of T_{ij}. The row and column-totals are also given.

Table of Cell-Totals, Squares, etc.

Student	English	Mathematics	Statistics	$T_{i\cdot\cdot}$	$T_{i\cdot\cdot}^2$	$\sum_j T_{ij\cdot}^2$
1	46 (2116)	59 (3481)	65 (4225)	170	28900	9822
2	61 (3721)	58 (3364)	57 (3249)	176	30976	10334
3	45 (2025)	46 (2116)	58 (3364)	149	22201	7503
$T_{\cdot j\cdot}$	152	163	180	495	82077	--
$T_{\cdot j\cdot}^2$	23104	26569	32400	82073	--	27661
$\sum_i T_{ij\cdot}^2$	7862	8961	10838	--	27661	← Check

PO A1k · ©Xayn Latif

Now correction factor $(C.F.) = \frac{T_{...}^2}{rcn} = \frac{(495)^2}{27} = 9075$

Total $SS = \sum_i \sum_j \sum_k X_{ijk}^2 - C.F. = 13^2 + 18^2 + 15^2 + \ldots + 21^2 + 18^2 - 9075$

$= 9363 - 9075 = 288.$

Between Subjects $SS = \sum_j \frac{T_{.j.}^2}{rn} - C.F. = \frac{82073}{9} - 9075 = 44.22$

Between Subjects $SS = \sum_i \frac{T_{i..}^2}{cn} - C.F. = \frac{82077}{9} - 9075 = 44.67$

Interaction $SS = \sum_i \sum_j \frac{T_{ij.}^2}{n} - \sum_i \frac{T_{i..}^2}{cn} - \sum_j \frac{T_{.j.}^2}{rn} + C.F.$

$= \frac{27661}{3} - \frac{82077}{9} - \frac{82073}{9} + 9075$

$= 9220.33 - 9119.67 - 9119.22 + 9075.00 = 56.44$

Error $SS = SST - SSC - SSR - SS\,(RC)$

$= 228 - (44.22 + 44.67 + 56.44) = 142.67.$

Hence the analysis of variance table is set up as below:

Source of Variation	*d.f.*	Sum of Squares	Mean Square	*F*
Between Subjects	2	44.22	22.11	$F_1 = \frac{22.11}{7.93} = 2.79$
Between Students	2	44.67	22.34	$F_2 = 2.82$
Interaction	4	56.44	14.11	$F_3 = 1.78$
Error	18	142.67	7.93	----
Total	26	288.00	--	--

(v) The critical regions are (a) $F_1 \geq F_{0.05}(2, 18) = 3.55$

(b) $F_2 \geq F_{0.05}(2, 18) = 3.55$

(c) $F_3 \geq F_{0.05}(4, 18) = 2.93$

OPPO A1k · ©Xayn Latif

(vi) **Conclusion.** Since the computed values of $F_1=2.79$, $F_2=2.32$ and $F_3=1.78$ do not fall in the critical regions, therefore we accept all three hypotheses.

20.4 MULTIPLE COMPARISON TESTS

If F-test after the analysis of variance rejects our null hypothesis, all we can conclude is that the *k population means are not all equal*. This conclusion might not be sufficient to satisfy the experimenter, rather he would like to know which means (or sets of means) might differ significantly from each other. For this purpose, several tests based on different view points, have been developed to make comparisons between pairs of means. Such tests, used as a follow-up to F-tests, are called the *Multiple Comparison tests*. The commonly used tests include the Fisher's Least Significant Difference (or *LSD*) test, the Student-Newman-Keul's Multiple Range test, the Duncan's Multiple Range test, the Scheffe's test, the Tukey's T-method, etc. We discuss the first four tests only.

20.4.1 The Least Significant Difference Test. When the null hypothesis of equal means is rejected by F-test after the ANOVA, we can test the significance of differences between means of k samples (or treatments) by using the ordinary two-sample t-test on every pair of the $\frac{1}{2}k(k-1)$ possible pairs of $\bar{X}_i$ and $\bar{X}_j$ $(i \neq j)$ at significance level α. But this procedure involves a large number of decisions. An alternative method of dealing with such a situation is to compute the smallest difference that would be judged significant and to compare the absolute values of all differences of means with it. This smallest difference is called the *least significant difference* or *LSD* and is given by

$$LSD = t_{\alpha/2,(v)} \sqrt{\frac{2\,(MSE)}{r}},$$

where MSE is the Error (Within) Mean Square, r is the size of equal samples, and $t_{\alpha/2,(v)}$ is the value of t at $\alpha/2$ level taken against the error degrees of freedom (v). The test criterion that uses the *least significant difference* is called the *LSD* test. Two means are declared to come from populations with significantly different means, when the absolute value of their difference exceeds the *LSD*.

It is interesting to note that the *LSD* test is actually a two-sample t-test. Two means $\bar{X}_i$ and $\bar{X}_j$ $(i \neq j)$ differ significantly when the value of t computed by

PPO A1k · ©Xayn Latif

$$t = \frac{|\bar{X}_i - \bar{X}_j|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

is greater than or equal to $t_{\alpha/2}$ for $v = n_1 + n_2 - 2$ degrees of freedom. In other words, two means $\bar{X}_i$ and $\bar{X}_j$ are judged significantly different,

if
$$\frac{|\bar{X}_i - \bar{X}_j|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{\alpha/2,(v)}$$

or if
$$|\bar{X}_i - \bar{X}_j| \geq t_{\alpha/2,(v)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\geq t_{\alpha/2,(v)}\, s_p \sqrt{\frac{2}{r}}\,,$$ when $n_1 = n_2 = r$, size of equal samples. Since in the *ANOVA*, the Error (Within) Means Square provides the pooled unbiased estimate (s_p^2) of the common variance σ^2, the relation becomes

$$|\bar{X}_i - \bar{X}_j| \geq t_{\alpha/2,(v)} \sqrt{\frac{2\,(MSE)}{r}}.$$

The *LSD* was defined by the right hand term.

It is customary to arrange the sample means in ascending order of magnitude and to draw a line under any pair of adjacent means (or set of means) that are not significantly different, *e.g.*,

$$\underline{\bar{X}_1, \bar{X}_2}, \bar{X}_3, \underline{\bar{X}_4, \bar{X}_5}, \bar{X}_6$$

indicates that $\bar{X}_1$ and $\bar{X}_2$, $\bar{X}_4$ and $\bar{X}_5$, etc. are not significantly different. It also implies that there exists a significant difference between the groups ($\bar{X}_1$, $\bar{X}_2$) and ($\bar{X}_4$, $\bar{X}_5$, $\bar{X}_6$). Since $\bar{X}_3$ is not connected by a line, it differs significantly. The *LSD* test is only applied after the null hypothesis is rejected in the *ANOVA*.

Example 20.7. Perform the anlaysis of variance on the following data and anaiyse the treatment means using the *LSD* test with a 0.05 level of significance.

	Treatments					
	1	2	3	4	5	6
Blocks 1	1	3	6	4	3	2
2	1	4	4	8	5	1
3	3	6	7	8	4	3
4	1	3	2	3	2	1

A1k · ©Xayn Latif

(i) We formulate our null and alternative hypotheses as

$H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3} = \mu_{\cdot 4} = \mu_{\cdot 5} = \mu_{\cdot 6}$, *i.e.* the six treatment means are equal, and

H_1 : Not all six treatment means are equal.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) We use the F-test to accept or reject H_0. If H_0 is rejected, the LSD-test is used to analyse the treatment means.

(iv) The computations are presented in the following *ANOVA*-Table:

Source of Variation	*d.f.*	Sum of Squares	Mean Square	F
Between Treatments	5	51.83	10.37	$F_1 = \frac{10.37}{1.57} = 6.61$
Between Blocks	3	28.50	9.50	--
Error	15	23.50	$s^2 = 1.57$	----
Total	23	103.83	--	--

(v) The critical region for treatments is $F_1 \geq F_{0.05}(5, 15) = 2.90$.

(vi) **Conclusion.** We reject the null hypothesis of equal treatment means as the computed value of $F_1 = 6.61$ exceeds the table value of $F = 2.90$. As H_0 is rejected, using F-test, we therefore apply the LSD test to find out which means differ from each other.

The *Least-Significant Difference* is given by the relation

$$LSD = t_{0.025,(15)} \sqrt{\frac{2s^2}{r}}, \text{ where } s^2 = 1.57 \text{ and } r = 4.$$

$$= 2.13 \sqrt{\frac{2 \times 1.57}{4}} = 1.89$$

Arranging the treatment means ($\bar{x}_i$) in ascending order of magnitude, and drawing a line under pair of adjacent means (or sets of means) that are not significantly different, we have

$\bar{x}_1$	$\bar{x}_6$	$\bar{x}_5$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
1.75	1.75	3.50	4.00	4.75	5.75

The significantly different pairs are immediately observed from this presentation.

OPPO A1k · ©Xayn Latif

20.4.2. The Student-Newman-Keul's Multiple Range Test or S-N-K Test. The *S-N-K* multiple range test compares the observed range of sample means in the subset with the calculated critical range at the α per cent level of significance. The test procedure uses a modification of the *Studentized range* statistic q defined by

$$q = \frac{\bar{X}_{max} - \bar{X}_{min}}{S.E. \text{ of } \bar{X}} = \frac{\text{Range}}{\sqrt{s^2/r}},$$

where s^2 (the error mean square in the *ANOVA* table) is the unbiased estimate of σ^2, and r is the number of observations in each mean. The sampling distribution of the q statistic is approximated by the Studentized range distribution having parameters k, the number of sample means and v, the number of degrees of freedom for *MSE*.

The critical value for the difference between two means which are $p(p=k, k-1, \ldots, 3, 2)$ steps apart on an ordered scale, is

$$W_p = q_\alpha (p, v) \sqrt{s^2/r}.$$

The value of q_α (a modified q statistic in the *S-N-K* test) for appropriate value of p and v are obtained from the Studentized range table of significant ranges at the α level desired. It is interesting to note that

$$W_2 = q_\alpha (2, v) \sqrt{s^2/r}, \qquad (p=2)$$
$$= t_{\alpha,v} \sqrt{2s^2/r} = LSD$$

To carry out the test, arrange the k sample means in increasing order of magnitude so that $\bar{X}_1 < \bar{X}_2 < \ldots < \bar{X}_k$ and calculate the critical range for $p = k$, *i.e.* W_p by multiplying the studentized range q by $\sqrt{s^2/r}$ $(=\sqrt{MSE/r}$ in an *ANOVA*). Compare the observed range of sample means $\bar{X}_k - \bar{X}_1$ with W_k. If $\bar{X}_k - \bar{X}_1$ does not exceed W_k, conclude that the means are all equal and the test ends there. If $\bar{X}_k - \bar{X}_1$ exceed W_k, subdivide the sample means into two groups of $k-1$ means each, $\bar{X}_k - \bar{X}_2$ and $\bar{X}_{k-1} - \bar{X}_1$. Compare the two ranges $\bar{X}_k - \bar{X}_2$ and $\bar{X}_{k-1} - \bar{X}_1$ with the calculated range for $p=k-1$, *i.e.* W_{k-1}. If either range does not exceed W_{k-1}, then declare the means in the groups to be equal. In case, either range exceeds W_{k-1}, divide the means in the group concerned into two groups of $k-2$ means each and compare with the calculated W_{k-2} and so on until a subset of means is found which does not exceed the calculated value W_i. The process ends whenever the observed range of a subset does not exceed the calculated critical range. This multiple comparison test is so called because each contributed to its formation. It is also called a *step-by-step* or *stepwise procedure* and is perhaps the best.

A1k · ©Xayn Latif

20.4.3. Duncan's Multiple Range Test. D.B. Duncan in 1955 introduced a multiple range test for comparing all pairs of sample means. This test, which has different significant ranges, is based on the Studentized-range statistic $q(p, v)$, where p is the number of steps between two means being compared, but the values of $q(p, v)$ for appropriate values of p and v, the degrees of freedom of error mean square in the *ANOVA*, are obtained from Duncan's table (see Appendix) for significant ranges. The α per cent critical ranges for the p means, called the *least significant range* (*lsr*), and denoted by R_p are given by

$$R_p = q_\alpha(p, v)\sqrt{s^2/r}, \qquad p = 2, 3, \ldots, k-1, k$$
$$= q_\alpha(p, v)\sqrt{MSE/r}.$$

To perform the test, arrange the sample means in increasing order of magnitude and compute a set of $k-1$ *least significant ranges* by multiplying the standard error of mean, *i.e.* $\sqrt{MSE/r}$ by the tabulated values of $q_\alpha(p, v)$. Compare the observed difference between the largest mean and the smallest mean with the least significant range R_k; the difference between the largest mean and the second smallest mean with *lsr* R_{k-1}, and so on until all means have been compared with the largest mean. Next, compare the difference of the second largest mean and the smallest mean with *lsr* R_{k-1}, the difference of the second largest mean with the second smallest mean with *lsr* R_{k-2}. The process continues until the differences of all possible $k(k-1)/2$ pairs of means are considered. A pair of means is declared to be significantly different when the observed difference exceeds the corresponding least significant range. Duncan's test is quite powerful and is very popular.

It is of importance to note that in case of unequal sample sizes r_i not differing too greatly, the harmonic mean, r_h of all r_i's, *i.e.* $r_h = \dfrac{k}{\sum (1/r_i)}$ is used in place of r, the number of observations in equal sample sizes.

Example 20.8. Use Duncan's multiple range test for the data in Example 20.7 to compare all pairs of treatment means. Assume that $\alpha = 0.05$.

In Example 20.7, we found that

$$s^2 = MSE = 1.57,$$

r (the number of observations in each treatment) = 4,

v (*df* for *MSE*) = 15 and k (number of treatments) = 6

We now find

$$\sqrt{MSE/r} = \sqrt{1.57/4} = 0.6265$$

Arranging the treatment means $(\bar{x}_i)$ in increasing order of magnitude, we get

$\bar{x}_1$	$\bar{x}_6$	$\bar{x}_5$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
1.75	1.75	3.50	4.00	4.75	5.75

The values of $q_{0.05}(p, 15)$ for $p = 2, 3, 4, 5, 6$ taken from Duncan's table of significant ranges and the least significant ranges, R_p, obtained by multiplying $q_{0.05}(p, 15)$ by $\sqrt{MSE/r}$, are shown below:

p	$q_{0.05}(p, 15)$	R_p
2	3.01	1.89 (R_2)
3	3.16	1.98 (R_3)
4	3.25	2.04 (R_4)
5	3.31	2.07 (R_5)
6	3.36	2.11 (R_6)

Comparing the differences between all pairs of means with the least significant ranges, R_p, beginning with the largest $(\bar{x}_4)$ against the smallest $(\bar{x}_1)$, we have the following results:

4 versus 1 : $5.75 - 1.75 = 4.00 > 2.11\ (R_6)$

4 versus 6 : $5.75 - 1.75 = 4.00 > 2.07\ (R_5)$

4 versus 5 : $5.75 - 3.50 = 2.25 > 2.04\ (R_4)$

4 versus 2 : $5.75 - 4.00 = 1.75 < 1.98\ (R_3)$

4 versus 3 : $5.75 - 4.75 = 1.00 < 1.89\ (R_2)$

3 versus 1 : $4.75 - 1.75 = 3.00 > 2.07\ (R_5)$

3 versus 6 : $4.75 - 1.75 = 3.00 > 2.04\ (R_4)$

3 versus 5 : $4.75 - 3.50 = 1.25 < 1.98\ (R_3)$

3 versus 2 : $4.75 - 4.00 = 0.75 < 1.89\ (R_2)$

2 versus 1 : $4.00 - 1.75 = 2.25 > 2.04\ (R_4)$

2 versus 6 : $4.00 - 1.75 = 2.25 > 1.98\ (R_3)$

2 versus 5 : $4.00 - 3.50 = 0.50 < 1.89\ (R_2)$

5 versus 1 : $3.50 - 1.75 = 1.75 < 1.98\ (R_3)$

5 versus 6 : $3.50 - 1.75 = 1.75 < 1.98\ (R_2)$

6 versus 1 : $1.75 - 1.75 = 0 < 1.89\ (R_2)$

The pairs of means whose differences are greater than the corresponding least significant ranges, R_p, are significantly different. Drawing a line under means which are not significantly different, we have

$\bar{x}_1$	$\bar{x}_6$	$\bar{x}_5$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
1.75	1.75	3.50	4.00	4.75	5.75

From the anlaysis, we see that the *LSD* method and Duncan's multiple range test produce the same conclusions.

20.4.4. Contrasts – Scheffe's Method. A *contrast* is defined as a linear combination of means with an essential condition that the sum of their constant co-efficients must equal zero. For example, the difference between two means $\mu_i - \mu_j$ or the expressions $\mu_1 + \mu_2 - 2\mu_3$, $\tau_1 - \tau_2 - \tau_3 + \tau_4$ are contrasts as the sums of their co-efficients are zero,. It is a common practice to define a contrast in terms of treatment totals. Thus, if $T_1, T_2, \ldots, T_k$ denote the treatment totals based on the same number of observations, the linear function

$$Q = c_1T_1 + c_2T_2 + \ldots + c_kT_k\ , \text{ where } c_1 + c_2 + \ldots + c_k = 0,$$

or more compactly,

$$Q = \sum_{j=1}^{k} c_jT_j \text{ with } \sum_{j=1}^{k} c_j = 0$$

would be a *contrast* (or comparison) among treatments $\tau_1, \tau_2, \ldots, \tau_k$. Two contrasts are said to be *orthogonal* if the sum of the product of the corresponding co-efficients is equal to zero. That is, two contrasts $Q_1 = \sum c_{1j}T_j$ and $Q_2 = \sum c_{2j}T_j$ will be orthogonal or independent if $\sum c_{1j}c_{2j} = 0$. Non-orthogonal contrasts are sometimes used, though orthogonal contrasts are better.

The sum of squares for a contrast, Q is given by

$$SSQ = \frac{Q^2}{\sum_j r_jc_j^2},$$

where r_j denotes the number of observations in T_j. It is important to note that the number of contrasts should not exceed the number of degrees of freedom associated with treatment means. Furthermore, a contrast always has a single degree of freedom and a t-test is appropriate.

PO A1k · ©Xayn Latif

Scheffe's method is used to test the significance of all possible contrasts. This method requires the computation of

(i) $S = \sqrt{(k-1)\, F_\alpha(v_1, v_2)}$. $\quad (v_1 = k-1)$

where v_1 and v_2 are treatment and error degrees of freedom and F is the tabulated value for a significance level α in the F-test of $H_0 : \tau_1 = \tau_2 = \ldots = \tau_k$;

(ii) the standard error of each contrast to be tested as

$$s_Q = \sqrt{(MSE) \sum r_j f_j^2}, \text{ and then}$$

(iii) the critical value as

$$\text{Scheffe's value} = S(s_Q).$$

A contrast, Q is said to differ significantly from zero when the absolute value of Q exceeds $S(s_Q)$ i.e., Scheffe's critical value. We can also directly test the hypothesis $H_0 : \mu_i - \mu_j$ by the criterion

$$|\bar{X}_i - \bar{X}_j| > \sqrt{(k-1)\, F} \cdot \sqrt{(MSE) \sum c_j^2 / r_j}$$

20.5 THE ANALYSIS OF VARIANCE MODELS

A statistical *model* is a mathematical structure which fully describes how the data could be generated. That is, a model describes what an observation is composed of. In statistical models, it comprises a mean and a random error component, where the mean may involve a single parameter μ or a sum of parameters. The parameter μ is constant while the random error components are assumed to be independently and normally distributed with zero mean and a common variance.

An example of the widely used linear additive model in the ANOVA is the expression

$$X_{ij} = \mu + \tau_j + e_{ij},$$

where X_{ij} is the ith observation for the jth sample or treatment, μ is the overall average effect, τ_j is the treatment effect undergone by the jth sample and e_{ij} are the error terms due to unspecified causes. To specify a model completely, some assumptions are made about the treatments. For example, when we assume that

(i) the treatment effects τ_j's are fixed and $\sum \tau_j = \sum(\mu_j - \mu) = 0$, the model is called a *fixed effects model* or Model 1:

(ii) the treatment effects τ_j's are a random sample from a normal population of treatments τ's with mean zero and variance σ^2_τ, the model is known as a *random effects model* or Model II.

It is important to note the difference between these models. In a *fixed effects model*, we are concerned with the k treatments present in the experiment whereas in a *random effects model*, we are concerned with a population of τ_j's of which the k treatments in the experiment are a random sample. The null and alternative hypotheses under Model I are stated as

H_0: There are no differences among the effects of k treatments present in the experiment, and

H_1: Not all treatment effects are equal.

In other words, we may state

$H_0 : \tau_j = 0$, (for all j) and H_1 : Some $\tau_j \neq 0$.

The hypotheses under Model II are formulated as

H_0: There are no differences among the effects of all the treatments in the population from which the k treatments in the experiment, are a random sample,

implying that

$$H_0 : \sigma^2_\tau = 0 \text{ and } H_1 : \sigma^2_\tau \neq 0.$$

A fixed effects model is sometimes known as the *analysis of variance model*, while a random effects model is called the *component of variance model*. Sometimes one effect is treated as fixed and the other as random, *e.g.*, in a two-way classification, we may assume blocks as random and treatments as fixed. Such a model is called a *mixed model*.

20.5.1. Least-Squares Estimates of Effects in One-Way ANOVA. Let X_{ij} represent the ith observation of the jth sample or treatment and μ_j be mean of jth treatment. Then we can represent this observation by the model (fixed effects).

$$X_{ij} = \mu_j + e_{ij}, \qquad i = 1, 2, \ldots r ; j = 1, 2, \ldots, k$$

$$= \text{Treatment mean} + \text{a random deviation}$$

$$= \mu + (\mu_j - \mu) + e_{ij} ;$$

$$= \mu + \tau_j + e_{ij},$$

where μ is the population general mean, $\tau_j = \mu_j - \mu$ is the additive effect of the jth treatment and e_{ij} denote the random deviations and are assumed to be independently and normally distributed with mean zero and variance σ^2. Thus the observation X_{ij} is composed of three

OPPO A1k · ©Xayn Latif

components which are added together. The null hypothesis that the k-population means are equal, may now be replaced by the hypothesis

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0.$$

The two independent estimates of σ^2, on which the statistic F is based, are determined by means of the *least-squares principle*. That is, we need to find estimates of $\mu = \hat{\mu}$, $\tau_j = \hat{\tau}_j$ which will minimize the sum of squares of the errors. Thus to minimize $\sum_i \sum_j e_{ij}^2$ subject to the restriction that $\sum_j \tau_j = 0$, we put

$$Q = \sum_i \sum_j e_{ij}^2 + \lambda \left[\sum_j \tau_j - 0\right]; \text{ where } \lambda \text{ is Lagrange's multiplier}$$

$$= \sum_i \sum_j [X_{ij} - \mu - \tau_j]^2 + \lambda \sum_j \tau_j.$$

Differentiating Q partially *w.r.t.* μ, τ_j and λ, and equating to zero, we get

$$\frac{\partial Q}{\partial \mu} = 0 = -2 \sum_i \sum_j (X_{ij} - \mu - \tau_j) \qquad \ldots (1)$$

$$\frac{\partial Q}{\partial \tau_j} = 0 = -2 \sum_i (X_{ij} - \mu - \tau_j) + \lambda \qquad \ldots (2)$$

$$\frac{\partial Q}{\partial \lambda} = 0 = \sum_j \tau_j \qquad \ldots (3)$$

From equation (1), we get $\hat{\mu} = \bar{X}_{..}$ $\qquad (\because \sum_j \tau_j = 0)$

Summing over i's in equation (2), we get

$$\bar{X}_{.j} - \bar{X}_{..} - \hat{\tau}_j - \frac{\lambda}{2k} = 0$$

To find λ, we sum over j's and have

$$\bar{X}_{..} - \bar{X}_{..} - 0 - \frac{\lambda}{2k} = 0, \text{ or } \lambda = 0$$

Hence $\qquad \hat{\mu} = \bar{X}_{..}$, and $\hat{\tau}_j = \bar{X}_{.j} - \bar{X}_{..}$

Now **Treatment SS** $= \sum_i \sum_j \hat{\tau}_j^2 = \sum_i \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2$

$$= r \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2, \text{ and}$$

Error SS $= \sum_i \sum_j (X_{ij} - \hat{\mu} - \hat{\tau}_j)^2$

$$= \sum_i \sum_j (X_{ij} - \bar{X}_{..} - \bar{X}_{.j} + \bar{X}_{..})^2$$

$$= \sum_i \sum_j (X_{ij} - \bar{X}_{.j})^2$$

A1k · ©Xayn Latif

These sums of squares are the same as before and are orthogonal as shown by Cochran's theorem.

Expected Mean Squares. From our model $X_{ij} = \mu + \tau_j + e_{ij}$ we have

$$\bar{X}_{.j} = \mu + \tau_j + \bar{e}_{.j} \text{ and } \bar{X}_{..} = \mu + \bar{e}_{..}, \text{ so that}$$

$$\bar{X}_{.j} - \bar{X}_{..} = \tau_j + \bar{e}_{.j} - \bar{e}_{..}$$

Taking expected values, we get

$$E[\text{Treatment } SS] = r\,E\left[\sum_j (\tau_j + \bar{e}_{.j} - \bar{e}_{..})^2\right]$$

$$= r\,E\left[\sum_j \tau_j^2\right] + r\,E\left[\sum_j (\bar{e}_{.j} - \bar{e}_{..})^2\right] + \text{the cross product term which vanishes as } \tau_j \text{ is fixed}$$

$$= r\sum_j \tau_j^2 + r\frac{\sigma^2}{r}(k-1), \text{ so that}$$

$$E[\text{Treatment Mean Square}] = \frac{E[\text{Treatment } SS]}{k-1}$$

$$= \sigma^2 + \frac{r}{k-1}\sum_j \tau_j^2; \text{ and}$$

$$\text{Error } SS = \sum_i \sum_j (X_{ij} - \bar{X}_{.j})^2,$$

$$= \sum_i \sum_j (e_{ij} - \bar{e}_{.j})^2, \text{ so that}$$

$$E[\text{Error Mean Square}] = \sigma^2.$$

These results can be put in a tabular form as below:

Source of Variation	*d.f.*	Expected *MS*	*F*
Treatments (τ_j)	$k-1$	$\sigma^2 + \frac{r}{k-1}\sum \tau_j^2$	
Error (e_{ij})	$k(r-1)$	σ^2	

20.5.2. Least Squares Estimates of Effects in a Two-Way ANOVA. The model (fixed effects) for a two-way analysis of variance with no interaction, is given by

$$X_{ij} = \mu + \beta_i + \tau_j + e_{ij}, \qquad i = 1, 2, \ldots, r; j = 1, 2, \ldots, c$$

where X_{ij} denotes the observation in the ith row and jth column,

μ is the overall mean for all observations,

β_i denotes the effect of the ith Row or Block,

PPO A1k · © Xayn Latif

τ_j denotes the effect of jth Treatment or Column, and

e_{ij}, the error terms which are assumed to be independently and $N(0, \sigma^2)$.

The sum of squares to be minimized subject to $\sum_i \beta_i = 0$ and $\sum_j \tau_j = 0$ is

$$Q = \sum_i \sum_j (X_{ij} - \mu - \beta_i - \tau_j)^2.$$

Differentiating Q *w.r.t.* μ and setting equal to zero, we get

$$\frac{\partial Q}{\partial \mu} = 0 = -2 \sum_i \sum_j (X_{ij} - \mu - \beta_i - \tau_j)$$

As $\sum_i \beta_i = 0$ and $\sum_j \tau_j = 0$, we therefore get $\hat{\mu} = \bar{X}_{..}$

Again $\frac{\partial Q}{\partial \beta_i} = 0 = -2 \sum_j (X_{ij} - \mu - \beta_i - \tau_j)$, so that

$$\hat{\mu} + \hat{\beta}_i = \bar{X}_{i.} \text{ or } \hat{\beta}_i = \bar{X}_{i.} - \bar{X}_{..}$$

Similarly, we get $\hat{\tau}_j = \bar{X}_{.j} - \bar{X}_{..}$

Thus
$$\begin{aligned} SSE &= \sum_i \sum_j (X_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j)^2 \\ &= \sum_i \sum_j (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2, \text{ as before.} \end{aligned}$$

The null hypotheses that (i) the column means (or treatment means) are equal, and (ii) the row means are equal, are replaced by

$$H_0' : \tau_1 = \tau_2 = \ldots = \tau_c = 0,$$

$$H_0'' : \beta_1 = \beta_2 = \ldots = \beta_r = 0.$$

These results are subject to the assumptions of normality, of independence and of equal variance.

Now Treatment SS or $SSC = r \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2$, so that

$$MSC = \frac{SSC}{c-1}, \text{ and}$$

Block (Row) SS or $SSB = c \sum_i (\bar{X}_{i.} - \bar{X}_{..})^2$, so that

$$MSB = \frac{SSB}{r-1}.$$

Similarly,
$$MSE = \frac{SSE}{(r-1)(c-1)}$$

k · ©Xayn Latif

Hence, we reject H_0', if $F_1 = \frac{MSC}{MSE} \geq F_\alpha$; $[(c-1), (c-1)(r-1)]$

Similarly, we reject H_0'', if $F_2 = \frac{MSB}{MSE} \geq F_\alpha$; $[(r-1), (c-1)(r-1)]$.

The corresponding model for the two-way analysis of variance with interaction, is given by

$$X_{ijk} = \mu + \beta_i + \tau_j + (\beta\tau)_{ij} + e_{ijk} \qquad \begin{array}{l} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots, c \\ k = 1, 2, \ldots, n \end{array}$$

where $(\beta\tau)_{ij}$ denotes the interaction effect of the ith row and jth column. To find the LS estimates, the quantity $\sum_i \sum_j \sum_k e_{ijk}^2$ is to be minimized subject to the restrictions

$$\sum_i \beta_i = 0, \sum_j \tau_j = 0, \sum_i (\beta\tau)_{ij} = 0 \text{ and } \sum_j (\beta\tau)_{ij} = 0.$$

The original null hypotheses are in this case replaced by the following hypotheses

$$H_0' : \beta_1 = \beta_2 = \ldots = \beta_r = 0,$$

$$H_0'' : \tau_1 = \tau_2 = \ldots = \tau_c = 0, \text{ and}$$

$$H_0''' : (\beta\tau)_{11} = (\beta\tau)_{12} = \ldots = (\beta\tau)_{rc} = 0.$$

The F-test is carried out exactly in the same way.

EXERCISES

20.1 (a) Discuss why using multiple two-sample t-tests is not an appropriate alternative to anlaysis of variance.

(b) What is meant by Analysis of Variance and degrees of freedom? What are the assumptions underlying a one-way analysis of variance? (P.U., B.A/B.Sc. 1996)

20.2 What is the rationale of one-way analysis of variance? Discuss the procedure of its analysis. Sketch a table of one-way analysis of variance and define the statistic used to test the significance of differences. (P.U., B.A/B.Sc. 1985)

20.3 Describe what is meant by "partitioning the total sum of squares". Partition the total sum of squares $\sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{..})^2$ into

OPPO A1k · ©Xayn Latif

the Error sum of squares and Treatment sum of squares. Find the number of degrees of freedom associated with each of these. How is the partition if the sample sizes are unequal?

20.4 (a) Explain in detail why it is not a good statistical procedure to perform several t-tests on pairs of means, when several means are to be compared. Suggest the alternative and give its assumptions.

(b) Derive the partitioning of the degrees of freedom for anlaysis of variance in one-way classification.

(c) Describe the LSD test in anlaysis of variance.

(P.U., B.A./B.Sc. 1993)

20.5 (a) What is meant by a Two-Way Analysis of Variance and an Interaction?

(b) Derive the partition of the total sum of squares for a two-way analysis of variance without interaction. Find the number of degrees of freedom associated with each of the sources of variation.

20.6 Show that (i) $\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{X}_{i.}-\bar{X}_{..})(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..})=0,$

(ii) $\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{X}_{.j}-\bar{X}_{..})(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..})=0,$

(iii) $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(X_{ijk}-\bar{X}_{ij.})(\bar{X}_{i..}-\bar{X}_{...})=0.$

20.7 (a) Describe the method of Anlaysis of Variance for the one-way classification and apply it to the following data:

Sample Number			
1	2	3	4
11	13	21	10
4	9	18	4
6	14	15	19

(b) Use Bartlett's test to determine if the assumption of equal variances is satisfied.

20.8 (a) Twenty men are used in an experiment, five being assigned at random to each of the four machines. The observations are the amounts produced by the machines in one day. Test the hypothesis at $\alpha=0.05$, that the machines are not different with respect to the number of items produced.

A1k · ©Xayn Latif

Machine Number			
1	2	3	4
64	41	65	45
39	48	57	51
65	41	76	55
46	49	72	48
63	57	64	47

(b) Discuss the assumptions involved in the above analysis.

20.9 (a) The following are three consecutive weeks' earnings of three salesmen employed by a given firm:

Salesmen		
A	*B*	*C*
152	181	160
175	171	130
180	203	124

Calculate F, and assuming that the necessary assumptions can be met, test at a level of 5% significance, whether differences between salesmen are significant.

(b) Determinations of yields of a process with four treatments are given:

	Treatments			
	1	2	3	4
Yields	11	6	8	14
	4	4	6	27
	4	3	4	8
	5	6	11	18

Test the hypothesis that no differences exist among the four treatments at $\alpha = 0.05$.

20.10 The cathode warm-up time in seconds was determined for three different tube types using eight observations on each type of tube. The results are:

OPPO A1k · ©Xayn Latif

Tube Type					
A		B		C	
19	20	20	40	16	19
23	20	20	24	15	17
26	18	32	22	18	19
18	35	27	18	26	18

Do an analysis of variance on these data and test the hypothesis that the three tube types require the same average warm-up time.

20.11 Consider the following 5 random samples:

Samples	A	26,	40,	37,	33,	19,	52,	30,	43
	B	40,	17,	32,	46,	24,	28,	36,	25
	C	50,	65,	38,	41,	36,	45,	48,	53
	D	37,	58,	34,	47,	45,	28,	40,	47
	E	50,	34,	36,	27,	41,	34,	47,	35

Perform the analysis of variance and test the hypothesis at the 0.05 level of significance, that the samples come from populations having the same means.

20.12 Describe a statistical procedure for comparing the means of k groups of observations, possibly of different sizes. State any assumptions that you make, and demonstrate the calculations using the following observations:

Group	Observations						
A	4,	9,	10,	11,	17,	19	
B	6,	8,	10,	11,	12,	12,	15
C	9,	13,	15,	20,	23		

20.13 Determinations are made on the yield using three methods of catalyzing a chemical process.

Method	Measurements				
1	47.2,	49.8,	48.5,	48.7;	
2	50.1,	49.3,	51.5,	50.9;	
3	49.1,	53.2,	51.2,	52.8,	52.3

Do the methods differ significantly at the 5% level of significance? (P.U., B.A./B.Sc. 1972)

20.14 Three sections of the same elementary mathematics course are taught by 3 teachers. The final grades were recorded as follows.

Teacher	A	75, 91, 84, 45, 82, 75, 68, 47, 95, 38.
	B	59, 83, 99, 77, 65, 81, 34, 81, 77, 88, 94, 51, 82.
	C	66, 77, 51, 90, 73, 90, 71, 68, 69.

Test the hypothesis that the average grades given by the three teachers are equal. Use a 0.05 level of significance.

(P.U., B.A/B.Sc. 1975, 86)

20.15 Each of the sets of observations given below is a random sample drawn from a normal population.

A	B	C	D
49	49	44	58
42	44	57	54
47	50	34	64
76	58	48	60
69	70	50	53
58			64
			52
			42

(a) Use Bartlett's test to check the equality of variances.

(b) Perform the analysis of variance to test for the equality of means. State the hypotheses and assumptions.

20.16 (a) Given that the sample size, the arithmetic mean and the variance, s^2 for each of k independent samples are known, how do you compute the SSW and SSB to carry out an analysis of variance for these k samples?

(b) Given the following information:

	Samples			
	1	2	3	4
No. of values (n_i)	4	6	7	3
Sample means ($\bar{X}_i$)	58	57	43	42
Estimate of variance (s_i^2)	10	30.4	5.67	9

Construct an analysis of variance table and test the hypothesis that the population means are equal at $\alpha = 0.05$.

(P.U., B.A./B.Sc. 1996)

20.17 A random sample of 54 plants was selected from a field containing one-year old plants of a variety of guayule, a plant species yielding rubber. Of these plants, 27 were normal, 15 offtypes and 12 aberrants. The percentage rubber content was determined from each plant, and the data are given as follows:

OPPO A1k · ©Xayn Latif

Type	Percentage Rubber Content								
Normal	6.97	7.11	7.26	6.80	7.01	7.00	6.35	6.37	7.29
	7.31	6.86	6.81	6.43	7.43	6.68	7.29	7.12	6.68
	7.34	5.15	6.41	6.45	6.32	6.82	6.86	6.48	7.28
Offtype	6.21	5.70	6.04	4.47	5.22	5.55	4.45	4.84	5.88
	5.82	6.09	5.59	6.06	5.59	6.74			
Aberrant	4.28	7.71	6.48	7.71	7.37	7.20	7.06	6.40	8.93
	5.91	5.51	6.36						

(a) Make an analysis of variance to test whether the data are consistent with the hypothesis that there is no difference in mean rubber content for the three types of plants. Use $\alpha = 0.01$.

(b) If the means of the distributions of each type of plant are denoted by μ_N, μ_0 and μ_A respectively, use Student's t-statistic to test the following particular hypotheses: (i) $\mu_N = \mu_A$, (ii) $\mu_0 = \frac{1}{2}(\mu_N + \mu_A)$.

20.18 Test the hypothesis that the difference between the means of students in arithmetic computation in the different type of schools, Grade 4 are equal, using 0.05 level of significance.

Marks	Frequency		
	Residential	Non-residential	Mission
0 – 9	1	4	---
10 – 19	3	7	4
20 – 29	10	25	8
30 – 39	63	37	17
40 – 49	38	13	29
50 – 59	5	4	7

20.19 In a feeding experiment of some animals, the following results were obtained, the numbers in the table being the gains in weight in pounds. The animals were in groups of 3 each.

Groups	I	II	III	IV
A	7.0	16.0	10.5	13.5
Rations B	14.0	15.5	15.0	21.0
C	8.5	16.5	9.5	13.5

Test the hypothesis of no difference in rations, at $\alpha = 0.05$.

(P.U., B.A/B.Sc. 1968)

A1k · ©Xayn Latif

20.20 Construct the analysis of variance table for the following data:

		Factor B			
		1	2	3	4
Factor A	1	15	31	20	30
	2	22	11	45	26
	3	33	37	30	44
	4	18	31	49	34
	5	37	30	36	21

Test the hypotheses that

(i) Factor A has no influence on yields, and

(ii) Factor B has no influence on yields.

20.21 Four breeds of cattle B_1, B_2, B_3, B_4 were fed on three different rations R_1, R_2, R_3. Gains in weight in pounds over a given period were recorded as given below:

	B_1	B_2	B_3	B_4
R_1	46.5	62	41	45
R_2	47.5	41.5	22	31.5
R_3	50	40	25.5	28.5

At $\alpha = 0.05$, test the null hypothesis that

(a) There are no differences in the mean gains in weight for the different breeds of cattle.

(b) There are no differences between rations. (P.U., B.A/B.Sc. 1969)

20.22 Four varieties of potato are planted each on five plots of ground of the same size and type, and each variety is treated with five different fertilizers. The yields in tons are as follows:

Variety	Fertilizer				
	1	2	3	4	5
1	1.9	2.2	2.6	1.8	2.1
2	2.5	1.9	2.3	2.6	2.2
3	1.7	1.9	2.2	2.0	2.1
4	2.1	1.8	2.5	2.3	2.4

Show how analysis of variance is applied to find out whether there is evidence that (i) any difference exists between the yields of varieties independently of the fertilizer and (ii) any differential

OPPO A1k · ©Xayn Latif

effect is exerted by the fertilizer independently of the variety.

(P.U., B.A/B.Sc. 1963; D.St., 1960; C.S.S. 1961)

20.23 An experiment is conducted in which 4 treatments are to be compared using five subjects. The following data are generated:

Treatment	Subject 1	2	3	4	5
1	12.8	10.6	11.7	10.7	11.0
2	11.7	14.2	11.8	9.9	13.8
3	11.5	14.7	13.6	10.7	15.9
4	12.6	16.5	15.4	9.6	17.1

Perform the analysis of variance, separating out the treatment, subject and error sums of squares. Use a 0.05 level of significance to test the hypothesis that there is no difference between the treatment means.

(P.U., B.A/B.Sc. 1984)

20.24 A certain company had four salesmen A, B, C and D, each of whom was sent for a week into three types of area, country area K, outskirts of a city O and shopping centre of a city S. The sales in pounds per week are shown below:

		Salesman A	B	C	D
District	K	30	70	30	30
	O	80	50	40	70
	S	100	60	80	80

Carry out analysis of variance and interpret the results stating the assumptions under which your results are valid.

20.25 The following data represent the marks obtained by five students in three subjects:

Student	Subjects: English	Statistics	Economics
1	47	73	61
2	80	88	86
3	71	77	59
4	62	60	66
5	56	95	87

Use a 0.05 level of significance to test the hypotheses that

(a) The courses are of equal difficulty.

(b) The students have equal ability.

(P.U., B.A./B.Sc. 1987)

A1k · ©Xayn Latif

20.26 Carry out an appropriate analysis of variance on the following data:

Rows	Columns 1	2	3
1	4	2	5
	7	3	6
	5	2	4
2	9	8	10
	8	7	8
	8	5	7

Use a 0.05 level of significance to test the hypotheses that

(a) the column-means are equal;

(b) the row-means are equal;

(c) the columns and rows do not interact. (P.U., B.A/B.Sc. 1976)

20.27 Perform an analysis of variance upon the following data and test whether there is any significant evidence of differences due to the drugs or the interaction of drugs and diets:

Drug	Diet 1	2	3	4
A	6.0	9.0	5.2	4.9
	6.8	10.1	5.6	4.5
B	4.5	7.9	4.6	2.8
	5.2	8.4	4.4	3.1
C	6.7	9.9	5.9	4.5
	7.1	9.2	6.1	4.7

(P.U., M.Sc. 1972)

20.28 Discuss each of the following:

(a) The Fisher's Least Significant Different Test.

(b) The Student-Newman-Keul's Multiple Range Test.

(c) Duncan's Mu tiple Range Test.

(d) The Scheffe's Method.

20.29 Use Duncan's multiple range test, with a 0.05 level of significance, to test all pairs of sample means in Exercise 20.11.

20.30 Each of five varieties of corn is planted in three plots in a large field. The respective yields, in bushels per acre, are indicated below:

Var 1	Var 2	Var 3	Var 4	Var 5
46.2	49.2	60.3	48.9	52.5
51.9	58.6	58.7	51.4	54.0
48.7	57.4	60.4	44.6	49.3

(a) Test whether the differences among the average yields are statistically significant. Let $\alpha = 0.05$.

(b) Use Duncan's multiple range test to make comparisons between pairs of means.

20.31 (a) What is the least significant difference and how is it used in interpreting the result of an experiment?

(b) The effects of four types of graphite coaters on light box readings are to be studied. As these readings might differ from day to day, observations are to be taken on each of the four types every day for three days. The results are:

Day	Graphite Coater Type			
	M	*A*	*K*	*L*
1	4.0	4.8	5.0	4.6
2	4.8	5.0	5.2	4.6
3	4.0	4.8	5.6	5.0

Construct an analysis of variance table. Determine at 0.05 level whether there is significant difference between Coater Types. Apply the *LSD* test to locate significant difference between the various pairs of Coater types.

20.32 Discuss the difference between *fixed effects* and *random effects models* in the analysis of variance. Write down a model, analysis of variance table and expected mean square for a one-way classification, assuming fixed effects.

20.33 (a) Explain what is meant by the terms *fixed-effects model*, *random effects model* and *mixed model* in the analysis of variance.

(b) Minimize the following function:

D A1k · ©Xayn Latif

$$S = \sum_{i=1}^{r} \sum_{j=1}^{c} (X_{ij} - \mu - \alpha_i - \beta_j)^2,$$

with the help of partial derivatives.

Evaluate the expression for μ, α_i and β_j with the conditions that

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{c} \beta_j = 0.$$

(P.U., B.A/B.Sc. 1983)

20.34 T treatments are applied one to each of T independent samples of size n. Show how to partition the total sum of squares into a sum of squares due to treatments and a residual sum of squares. Find the expected values of your component sums of squares under the following assumptions:

(a) that the T treatments constitute the only treatments of interest:

(b) that the T treatments are a random sample from a large population of treatments. (P.U., B.A./B.Sc. Hons, 1972)

20.35 A highway research engineer wishes to determine the effect of four types of subgrade soil on the moisture content in the top soil. He takes five samples of each type of subgrade soil and the total sum of squares is computed as 280, whereas the sum of squares among the four types of subgrade soil is 120.

(i) Set up an analysis of variance table for these data.

(ii) Set up a statistical model to describe this problem, define each term in the model, and state the assumptions made on each term.

(iii) Set up a test of the hypothesis that the four types of subgrade soil have the same effect on moisture content in the top soil.

(iv) Set up a set of orthogonal contrasts for this problem.

(v) Explain briefly how to set up a test on means after the analysis of variance for these data.

❖❖❖❖❖❖❖❖❖❖❖

21

Statistical Inference in Regression and Correlation

21.1 INTRODUCTION

In Chapters 10 and 11, we have introduced some of the basic concepts and techniques for linear regression models and correlation. In this Chapter, we consider the inferential procedures associated with linear regressions and correlation coefficients.

A simple linear regression model that describes the relationship between X and Y takes the form

$$Y_i = \alpha + \beta X_i + \varepsilon_i \text{ or } Y_i = \mu_{Y.X} + \varepsilon_i$$

where ε_i's are random errors. The random errors ε_i's are assumed to be independent of X_i and normally distributed with $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2_{Y.X}$, a constant for all X_i. These assumptions imply that Y_i also have common variance $\sigma^2_{Y.X}$ as the only random element in the model is ε_i.

The regression line $E(Y_i) = \mu_{Y.X} = \alpha + \beta X_i$ is estimated from the sample data by $\hat{Y} = a + bX$, where $b = \dfrac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$, $a = \bar{Y} - b\bar{X}$ and $\hat{Y}$ is the sample estimate of the population mean $\mu_{Y.X}$.

It is evident that in a linear regression, the quantities b, a, $\hat{Y}$ and $\bar{Y}$ will vary from one sample of data to another. They are thus random variables and hence have sampling distributions. For the purposes of statistical inference, we must know the means, variances and the shapes of these sampling distributions.

We derive the means and variances as below:

(i) **Mean and Variance of the Sampling Distribution of b.** The mean of the sampling distribution of b is $\mu_b = \beta$ and the variance is $\sigma_b^2 = \sigma_{Y.X}^2 / \sum (X_i - \bar{X})^2$.

In Example 15.6 on page 75, we found that

$$\mu_b = E(b) = \beta,$$

$$\text{Now} \quad \sigma_b^2 = \text{Var}(b) = \text{Var}\left[\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}\right]$$

$$= \text{Var}\left[\frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2}\right] \qquad [\because \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X}) Y_i]$$

Since terms involving X_i are constant in a regression model, therefore

$$\sigma_b^2 = \frac{1}{[\sum (X_i - \bar{X})^2]^2} \sum [(X_i - \bar{X})^2 \text{Var}(Y_i)]$$

$$= \frac{1}{[\sum (X_i - \bar{X})^2]^2} \sum (X_i - \bar{X})^2 \sigma_{Y.X}^2 = \frac{\sigma_{Y.X}^2}{\sum (X_i - \bar{X})^2}$$

Since the random errors ε_i in the regression model are assumed to be normally distributed, therefore the distribution of b is normal with a mean of $\mu_b = \beta$ and a standard deviation (or *standard error of b*) of $\sigma_b = \sigma_{Y.X} / \sqrt{\sum (X_i - \bar{X})^2}$.

Generally, $\sigma_{Y.X}^2$ will be unknown, we therefore require an estimate of $\sigma_{Y.X}^2$ from the sample data. The unbiased estimator is given by

$$s_{Y.X}^2 = \frac{\sum (Y_i - \hat{Y})^2}{n - 2}$$

Thus the estimate of σ_b^2, denoted by s_b^2, may be taken as

$$s_b^2 = \frac{s_{Y.X}^2}{\sum (X_i - \bar{X})^2}.$$

(ii) **Mean and Variance of the Sampling Distribution of a.** The mean of the sampling distribution of a, is $\mu_a = \alpha$ and the variance is $\sigma_a^2 = \sigma_{Y.X}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right]$.

In the estimated regression line $\hat{Y} = a + bX$, a is given as $a = \bar{Y} - b\bar{X}$. Therefore

$$\mu_a = E(a) = E(\bar{Y} - b\bar{X})$$

$$= E\left[\frac{\sum Y_i}{n} - b\bar{X}\right] = \frac{1}{n}\sum E(Y_i) - \bar{X}\,E(b)$$

But $\quad E(Y_i) = \alpha + \beta X_i$ and $E(b) = \beta$.

$$\therefore \quad \mu_a = \frac{\sum(\alpha + \beta X_i)}{n} - \bar{X}\beta = \alpha + \beta\bar{X} - \beta\bar{X} = \alpha.$$

And $\quad \sigma_a^2 = \text{Var}(a) = \text{Var}\left[\frac{\sum Y_i}{n} - b\bar{X}\right]$

$$= \text{Var}\left[\frac{\sum Y_i}{n}\right] + \text{Var}\left[b\bar{X}\right]$$

($\because$ b and $\bar{Y}$ are independent)

$$= \frac{1}{n^2}\sum \text{Var}(Y_i) + \bar{X}^2\,\text{Var}(b)$$

$$= \frac{n\sigma_{Y.X}^2}{n^2} + \frac{\bar{X}^2\sigma_{Y.X}^2}{\sum(X - \bar{X})^2}$$

$$= \sigma_{Y.X}^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X - \bar{X})^2}\right].$$

When $\sigma_{Y.X}^2$ is unknown, we use $s_{Y.X}^2$ in place of $\sigma_{Y.X}^2$.

The assumption that the random errors ε_i are normally distributed, leads to the fact that the distribution of a is normal with mean $\mu_a = \alpha$ and standard deviation

$$\sigma_a = \sigma_{Y.X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}}.$$

(iii) **Mean and Variance of the Sampling Distribution of $\hat{Y}$.** We can also find the mean and variance of $\hat{Y}$ when it is used as an estimate of the mean $\mu_{Y.X}$. Let $\hat{Y}$ represent the estimated value of Y_i in a linear regression, for a given value X_0. Then we have

$$\hat{Y} = a + bX_0$$

where X_0 is the value of X, on which the estimate is based.

Now $\mu_{\hat{Y}} = E(\hat{Y}) = E(a + bX_0)$

$= \alpha + \beta X_0$, and ($\because X_0$ is constant *w.r.t.* expectation)

$$\sigma^2_{\hat{Y}} = \text{Var}(\hat{Y}) = \text{Var}(a + bX_0)$$

$$= \text{Var}[\bar{Y} + b(X_0 - \bar{X})] \qquad (\because a = \bar{Y} - b\bar{X})$$

$$= \text{Var}(\bar{Y}) + (X_0 - \bar{X})^2 \text{Var}(b) \quad (\because \bar{Y} \text{ \& } b \text{ are independent})$$

$$= \frac{\sigma^2_{Y.X}}{n} + \frac{(X_0 - \bar{X})^2 \sigma^2_{Y.X}}{\Sigma(X - \bar{X})^2}$$

$$= \sigma^2_{Y.X}\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right].$$

We use $s^2_{Y.X}$ in place of $\sigma^2_{Y.X}$ when $\sigma^2_{Y.X}$ is not known.

The assumption that the ε_i's are normally distributed, leads to the distributional fact that the distribution of Y is normal with mean $\mu_{\hat{Y}} = \alpha + \beta X_0$ and standard deviation

$$\sigma_{\hat{Y}} = \sigma_{Y.X}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X - \bar{X})^2}}.$$

With the information on the means and standard deviations of these statistics, we may construct confidence intervals of, or test hypotheses about, various unknown parameters in a regression model.

21.2 INTERVAL ESTIMATION IN THE SIMPLE LINEAR REGRESSION

In this section, we consider how to construct confidence intervals and make predictions in the context of a simple linear regression model.

21.2.1. Confidence Interval Estimate of Population Regression Co-efficient β. To construct a confidence interval for β, the population regression co-efficient, we use b, the sample estimate of β. The sampling distribution of b is normally distributed with a mean β and a standard deviation $\sigma_{Y.X} / \sqrt{\Sigma(X - \bar{X})^2}$. That is, the variable

$$Z = \frac{b - \beta}{\sigma_{Y.X} / \sqrt{\Sigma(X - \bar{X})^2}}$$

But $\sigma_{Y.X}$ is generally not known, we therefore estimate it from the sample data by $s_{Y.X} = \sqrt{\dfrac{\sum (Y_i - \hat{Y})^2}{n-2}}$. We shall then use the Student's t distribution rather than the normal distribution. In other words, the statistic

$$t = \frac{b - \beta}{s_{Y.X} / \sqrt{\sum (X - \bar{X})^2}} = \frac{b - \beta}{s_b}, \quad \left[\because s_b = \frac{s_{Y.X}}{\sqrt{\sum (X - \bar{X})^2}}\right]$$

which follows a Student's t-distribution with $v = n-2$ degrees of freedom, is used.

Consequently, we make the following probability statement

$$P\left[-t_{\alpha/2,(v)} < \frac{b - \beta}{s_b} < t_{\alpha/2,(v)}\right] = 1 - \alpha,$$

where $t_{\alpha/2,(v)}$ is the value of the t-distribution with v degrees of freedom, leaving an area equal to $\alpha/2$ to the right. The probability statement after simplification becomes

$$P\left[b - t_{\alpha/2,(v)}\, s_b < \beta < b + t_{\alpha/2,(v)}\, s_b\right] = 1 - \alpha.$$

Hence a $100(1 - \alpha)$ percent confidence interval for β, the population regression co-efficient, for a particular sample of size $n (n < 30)$ is given by

$$b \pm t_{\alpha/2,(n-2)}\, s_b,$$

where $$s_b^2 = \frac{s_{Y.X}^2}{\sum (X - \bar{X})^2} = \frac{\sum (Y - \hat{Y})^2}{(n-2) \sum (X - \bar{X})^2}$$

Example 21.1. In a linear regression problem, the following results are obtained:

$$\hat{Y} = 12.13 - 0.1608\, X, \ \sum (Y_i - \hat{Y})^2 = 1160, \ \sum (X_i - \bar{X})^2 = 1300, \text{ and}$$

n, the number of pairs of the values of X and Y = 24.

Assuming normality, compute a 95% confidence interval for β, the population regression co-efficient.

The 95% confidence interval for β is given by

$$b \pm t_{\alpha/2,(n-2)}\, s_b.$$

Here $b = -0.1608$, $t_{0.025,(24-2)} = 2.074$, and (from t-tables)

$$s_b^2 = \frac{s_{Y.X}^2}{\sum (X - \bar{X})^2} = \frac{\sum (Y - \hat{Y})^2}{(n-2) \sum (X - \bar{X})^2}$$

$$= \frac{1160}{(22)\,(1300)} = 0.0406, \text{ so that}$$

$$s_b = \sqrt{0.0406} = 0.20.$$

Substituting these values, we get

$$-0.1608 \pm (2.074)\,(0.20)$$

or $\quad -0.16\,08 \pm 0.4148$ or -0.5856 to 0.2540

Hence the desired 95% confidence interval for population regression coefficient β is $(-0.586, 0.254)$.

21.2.2. Confidence Interval Estimate of α, the Intercept of Regression Line. To construct a confidence interval for α we use a, the sample estimate of α. We have already observed that a is distributed normally with mean $\mu_a = \alpha$ and standard deviation $\sigma_a = \sigma_{Y.X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X-\bar{X})^2}}$

Since $\sigma_{Y.X}$ is usually unknown, we use its unbiased sample estimate $s_{Y.X}$.

It can be shown, for the normal regression model, that the statistic

$$t = \frac{a - \alpha}{s_a},$$

where $s_a = s_{Y.X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X-\bar{X})^2}}$, has a Student's t-distribution with $v = n-2$ degrees of freedom. Using this statistic, a confidence interval for α, can be constructed in the same way as the confidence interval for β, with a replacing b and s_a replacing s_b.

Hence a $(1-\alpha)100$ per cent confidence interval for α is given by

$$a \pm t_{\alpha/2,(n-2)}\, s_a.$$

21.2.3. Confidence Interval Estimate of Mean Value $\mu_{Y.X}$ for a Given Value X_0. Perhaps the major goal of a regression study is to use the estimated regression equation $Y = a + bX$ to estimate the mean $\mu_{Y.X}$ of the Y values, when $X = X_0$. We have observed that the sampling distribution of $\hat{Y}$ is normal with $\mu_{\hat{Y}} = \mu_{Y.X}$ $(=\alpha+\beta X)$ and standard deviation $\sigma_{\hat{Y}} = \sigma_{Y.X}\sqrt{\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}}$. Since $\sigma_{Y.X}$ is usually unknown, we therefore find the sample estimate of $\sigma_{\hat{Y}}$, denoted by $s_{\hat{Y}}$, as

$$s_{\hat{Y}} = s_{Y.X}\sqrt{\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}}.$$

It can be shown that the statistic

$$t = \frac{\hat{Y} - \mu_{Y.X}}{s_{\hat{Y}}}$$

has a Student's t-distribution with $v = n-2$ degrees of freedom. Using t-statistic, we may construct a confidence interval for $\mu_{Y.X}$ for a given value X_0, in the usual manner.

Hence a $(1-\alpha)100$ per cent confidence interval for the mean value $\mu_{Y.X}$, when $X = X_0$, is given by

$$\hat{Y}_0 \pm t_{\alpha/2,(n-2)}\, s_{\hat{Y}},$$

where $\hat{Y}_0 = a + bX_0$.

Example 21.2. Given the data

Fertilizer (X)	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4
Corn Yield (Y)	10	15	30	35	25	30	50	45.

Assuming normality, calculate the 95% confidence intervals for (i) the value of α, (ii) the value of β, (iii) the true value of Y when $X = 3.0$.

To find sample estimates of various parameters in the regression model, the necessary computations are given below:

X	Y	XY	X^2	Y^2
0.3	10	3.0	0.09	100
0.6	15	9.0	0.36	225
0.9	30	27.0	0.81	900
1.2	35	42.0	1.44	1225
1.5	25	37.5	2.25	625
1.8	30	54.0	3.24	900
2.1	50	105.0	4.41	2500
2.4	45	108.0	5.76	2025
10.8	240	385.5	18.36	8500

Now $$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{(8)(385.5) - (10.8)(240)}{(8)(18.36) - (10.8)^2}$$

$$= \frac{3084 - 2592}{146.88 - 116.64} = \frac{492}{30.24} = 16.27, \text{ and}$$

$$a = \bar{Y} - b\bar{X} = 30 - (16.27)(1.35) = 8.04.$$

The estimated regression line is

$$\hat{Y} = 8.04 + 16.27X.$$

Further, $\sum(Y - \hat{Y})^2 = \sum Y^2 - a\sum Y - b\sum XY$

$$= 8500 - (8.04)(240) - (16.27)(385.5)$$

$$= 8500 - 1929.6 - 6272.08 = 298.32, \text{ and}$$

$$s_{Y.X} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{298.32}{6}} = 7.05$$

And $\sum(X_i - \bar{X})^2 = \sum X^2 - (\sum X)^2/n = 18.36 - 14.56 = 3.78$

(i) The 95% confidence interval for α is given by

$$a \pm t_{\alpha/2,(n-2)}\, s_a$$

Now $a = 8.04$, $t_{0.025,(6)} = 2.447$, and (from t-tables)

$$s_a = s_{Y.X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X - \bar{X})^2}}.$$

$$= 7.05\sqrt{\frac{1}{8} + \frac{(1.35)^2}{3.78}} = 7.05\sqrt{0.125 + 0.482}$$

$$= (7.05)(0.779) = 5.492$$

Substituting these values, we get

$$8.04 \pm (2.447)(5.492)$$

or 8.04 ± 13.44 or -5.40 to 21.48.

Hence the 95% confidence interval for α is $(-5.40, 21.48)$.

(ii) The 95% confidence interval for β is given by

$$b \pm t_{\alpha/2,(n-2)}\, s_b.$$

Now $b = 16.27$, $t_{0.025,(6)} = 2.447$, and

$$s_b = \frac{s_{Y.X}}{\sqrt{\sum(X - \bar{X})^2}} = \frac{7.05}{\sqrt{3.78}} = \frac{7.05}{1.94} = 3.63$$

Substituting these values, we get

$$16.27 \pm (2.447)(3.63)$$

or 16.27 ± 8.88, or 7.39 to 25.15.

Hence the 95% confidence interval for β is $(7.39, 25.15)$.

(iii) The 95% confidence interval for the true value of Y, *i.e.* $\mu_{Y.X}$ when $X = 3.0$, is given by

$$\hat{Y} \pm t_{\alpha/2,(n-2)}\, s_{\hat{Y}}.$$

Now the estimated value of Y for $X = 3.0$ is

$$\hat{Y} = a + bX_0 = 8.04 + (16.27)(3.0) = 56.85,$$

$$t_{0.025,(6)} = 2.447, \text{ and}$$

$$s_{\hat{Y}} = s_{Y.X}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X - \bar{X})^2}}.$$

$$= 7.05\sqrt{\frac{1}{8} + \frac{(3.0 - 1.35)^2}{3.78}} \qquad (\because X_0 = 3.0)$$

$$= 7.05\sqrt{0.125 + 0.7202}$$

$$= (7.05)(0.919) = 6.48$$

Substituting these values, we obtain

$$56.85 \pm (2.447)(6.48)$$

or $\quad 56.85 \pm 15.86$ or 40.99 to 72.71.

Hence the desired 95% confidence interval for the true value of Y, *i.e.*, $\mu_{Y.X}$ corresponding to $X = 3.0$ is (40.99, 72.71).

21.2.4. Prediction Interval of an individual Y value for a given value X_0. The estimated regression line $Y = a + bX$ is also a predictor of $Y_i = \alpha + \beta X_i + \varepsilon_i$. That is, $\hat{Y}$ can also be used to predict an *individual* value Y_0 of Y_i rather than a mean value, corresponding to a given X_0. To draw inference about Y_0, we need to know its mean and variance, etc. Accordingly, writing

$$Y_0 = \hat{Y} + \varepsilon$$

where ε is the deviation of Y_0 from the estimated value of $\hat{Y}$, we have

$$E(Y_0) = E(\hat{Y} + \varepsilon) = \alpha + \beta X_0, \text{ and}$$

$$\text{Var}(Y_0) = \text{Var}(\hat{Y}) + \text{Var}(\varepsilon)$$

$$= \sigma^2_{Y.X}\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right] + \sigma^2_{Y.X}$$

$$= \sigma^2_{Y.X}\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$$

It has been shown that Y_0 follows the normal distribution with mean $\alpha + \beta X_0$ and variance $\sigma^2_{Y.X}\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$.

Since $\sigma^2_{Y.X}$ is usually unknown, it is estimated by $s^2_{Y.X}$ from sample data. It has been further shown that the statistic

$$t = \frac{Y_0 - \hat{Y}_0}{S.E.\ (\hat{Y}_0)}$$

follows a Student's t-distribution with $v = n-2$ *d.f.* For the purposes of inference about Y_0, we therefore use the t-distribution.

To consider the construction of a confidence interval about Y_0, we see that the quantity to be predicted is a value taken by a random variable, not a parameter of a distribution. The appropriate interval is thus not *strictly* a confidence interval, it is usually referred to as the *prediction interval*. A prediction interval is one that contains Y with a desired probability. Hence a $(1-\alpha)100$ per cent prediction interval for an individual value Y_0 that corresponds to a given X_0, is given by

$$\hat{Y}_0 \pm t_{\alpha/2,(n-2)}\, s_{Y.X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X - \bar{X})^2}}.$$

A prediction interval for an individual value of Y is always wider than the confidence interval for a mean value of Y.

Example 21.3. Using the data of Example 21.2, construct a 95% prediction interval for Y when $X = 3.0$.

In Example 21.2, we found the estimated regression line as

$$\hat{Y} = 8.04 + 16.27\,X.$$

When $X = 3.0$, $\hat{Y}_0 = 8.04 + 16.27\,(3.0) = 56.85$.

Further, $n = 8$, $\bar{X} = 1.35$, $X_0 = 3.0$, $s_{Y.X} = 7.05$ and $t_{0.025,(6)} = 2.447$.

Now, we find the 95% prediction interval for a Y-value when $X = 3.0$ by

$$\hat{Y} \pm t_{\alpha/2,(n-2)}\, s_{Y.X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X - \bar{X})^2}}.$$

Substituting the values, we get the 95% prediction interval as

$$56.85 \pm (2.447)\,(7.05)\sqrt{1 + \frac{1}{8} + \frac{(3.0 - 1.35)^2}{3.78}}.$$

or $$56.85 \pm (2.447)\,(7.05)\sqrt{1 + 0.125 + 0.7202}$$

or $56.85 \pm (2.447)(7.05)(1.358)$

or 56.85 ± 23.43 or 33.42 to 80.28.

Hence the desired 95% prediction interval for an individual Y value at $X=3.0$ is (33.42, 80.28).

21.3 HYPOTHESIS TESTING IN THE REGRESSION MODEL

In this section, we present several procedures for testing hypotheses about unknown parameters in the linear regression and also about the linearity of regression.

21.3.1. Testing Hypotheses about β, the Population Regression Co-efficient. Suppose that we wish to test the hypothesis that the population regression co-efficient, β has some specified value β_0. We draw a random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ of n pairs of observations from a bivariate normal population and find the estimated regression equation $\hat{Y} = a + bX$, where b, the sample estimate of β, is normally distributed with mean β and standard deviation $\sigma_{Y.X}/\sqrt{\Sigma(X-\bar{X})^2}$ It has been shown earlier that, when $\sigma_{Y.X}$ is estimated from sample data, the statistic

$$t = \frac{b-\beta}{s_b},$$

where $s_b = \dfrac{s_{Y.X}}{\sqrt{\Sigma(X-\bar{X})^2}}$, has a Student's t-distribution with $v=n-2$ degrees of freedom. This statistic is therefore used to test the hypothesis $H_0: \beta=\beta_0$ and the procedure is given below:

(i) Formulate the null and alternative hypotheses about β. Three possible forms are

$$H_0: \beta = \beta_0, \text{ and } H_1: \beta \neq \beta_0.$$

$$H_0: \beta \leq \beta_0, \text{ and } H_1: \beta > \beta_0.$$

$$H_0: \beta \geq \beta_0, \text{ and } H_1: \beta < \beta_0.$$

(ii) Decide on the significance level α.

(iii) The test-statistic to use is

$$t = \frac{b-\beta_0}{s_b},$$

which, if H_0 is true, has a t-distribution with $v=n-2$ degrees of freedom.

(iv) The critical region is

$$|t| \geq t_{\alpha/2,(n-2)}, \text{ when } H_1 \text{ is } \beta \neq \beta_0$$

$$t \geq t_{\alpha,(n-2)}, \text{ when } H_1 \text{ is } \beta > \beta_0$$

$$t \leq -t_{\alpha,(n-2)}, \text{ when } H_1 \text{ is } \beta < \beta_0.$$

(v) Compute the regression equation $\hat{Y} = a + bX$, $s_{Y.X}$, s_b and $t = \dfrac{b-\beta_0}{s_b}$ from the sample data.

(vi) Decide as below:

Reject H_0 if t falls in the critical region, accept H_0 otherwise.

Most frequently, we are interested in testing the hypothesis that $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. It is important to note that testing the hypothesis that $\beta=0$ is equivalent to testing the hypothesis that the *variable Y is independent of the variable X* (in a linear sense). The test-statistic then becomes $t=b/s_b$. If we reject $H_0 : \beta = 0$, we conclude that the two variables are linearly related. If we accept $H_0 : \beta = 0$, we conclude that the two variables are *not* linearly related.

Example 21.4. In a linear regression problem, the following sums were computed from a random sample of size 10.

$$\sum X = 320, \sum Y = 250, \sum X^2 = 12400, \sum XY = 9415 \text{ and } \sum Y^2 = 7230.$$

Using 5 per cent significance level, test the hypothesis that the population regression co-efficient, β is greater than 0.5.

(i) We state our null and alternative hypotheses as

$$H_0 : \beta \leq 0.5, \text{ and } H_1 : \beta > 0.5.$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic, under H_0, is

$$t = \frac{b-\beta_0}{s_b} = \frac{b-0.5}{s_b}$$

which has a Student's t-distribution with $v=10-2$, *i.e.* 8 degrees of freedom.

(iv) Computations. Now $b = \dfrac{n\sum XY - \sum X \sum X}{n\sum X^2 - (\sum X)^2}$

$$= \frac{(10)\,(9415) - (320)\,(250)}{(10)\,(12400) - (320)^2} = \frac{14150}{21600} = 0.655;$$

$$a = \bar{Y} - b\bar{X} = \left(\frac{250}{10}\right) - (0.655)\left(\frac{320}{10}\right) = 4.04;$$

$$s^2_{Y.X} = \frac{\sum(Y - \hat{Y})^2}{n-2} = \frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}$$

$$= \frac{7230 - (4.04)\,(250) - (0.655)\,(9415)}{10-2} = \frac{53.175}{8}$$

$$= 6.647 \text{ so that } s_{Y.X} = \sqrt{6.647} = 2.578,$$

$$\sum(X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$= 12400 - \frac{(320)^2}{10} = 2160,$$

and $$s_b = \frac{s_{Y \cdot X}}{\sqrt{\sum(X - \bar{X})^2}} = \frac{2.578}{\sqrt{2160}} = \frac{2.578}{46.476} = 0.055$$

$$\therefore \quad t = \frac{0.655 - 0.5}{0.055} = 2.82$$

(v) The critical region is $t \geq t_{0.05,(8)} = 1.86$.

(vi) **Conclusion.** Since the calculated value of $t = 2.82$ falls in the critical region, so we reject H_0. We may conclude that there is sufficient evidence to indicate that the population regression coefficient is greater than 0.5.

Example 21.5. Estimate a regression line from the following data of height (X) and weight (Y) of 12 persons:

Height (X)	Weight (Y)
60	110, 135, 120
62	120, 140, 130, 135
64	150, 145
70	170, 185, 160

Test the hypothesis that the population regression coefficient $\beta = 0$, i.e., height and weight are independent. Use a 0.05 level of significance.

(i) We set up our hypotheses as

$H_0 : \beta = 0$, *i.e.*, the two variables X and Y are not related.

$H_1 : \beta \neq 0$, *i.e.*, the two variables are related.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic, if H_0 is true, is

$$t = \frac{b}{s_b},$$

where $s_b^2 = \dfrac{s_{Y.X}^2}{\sum(X-\bar{X})^2} = \dfrac{\sum(Y-\hat{Y})^2}{(n-2)\sum(X-\bar{X})^2}$

Assuming that the distribution of Y_i for each X, is normal with the same mean and the same standard deviation, the statistic t conforms to the Student's t-distribution with $n-2=10$ degrees of freedom.

(iv) Computations. The necessary computations are given below:

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
60	110	3600	12100	6600
60	135	3600	18225	8100
60	120	3600	14400	7200
62	120	3844	14400	7440
62	140	3844	19600	8680
62	130	3844	16900	8060
62	135	3844	18225	8370
64	150	4096	22500	9600
64	145	4096	21025	9280
70	170	4900	28900	11900
70	185	4900	34225	12950
70	160	4900	25600	11200
766	1700	49,068	246,100	109,380

The estimated regression line is $\hat{Y} = a + bX$, where

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{(12)(109380) - (766)(1700)}{(12)(49068) - (766)^2}$$

$$= \frac{10360}{2060} = 5.03, \text{ and}$$

$$a = \bar{Y} - b\bar{X} = \frac{1700}{12} - (5.03)\left(\frac{766}{12}\right)$$

$$= 141.6667 - 321.0816 = -179.41.$$

$$\therefore \quad \hat{Y} = -179.41 + 5.03X$$

Now $\Sigma(Y - \hat{Y})^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY$

$$= 246100 + (179.41)(1700) - (5.03)(109380)$$

$$= 246100 + 304997 - 550181.40 = 915.60,$$

$$\Sigma(X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 49068 - \frac{(766)^2}{12} = 171.67,$$

$$s_{Y.X} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{915.60}{10}} = 9.5687, \text{ and}$$

$$s_b = \frac{s_{Y.X}}{\sqrt{\Sigma(X - \bar{X})^2}} = \frac{9.5687}{\sqrt{171.67}} = \frac{9.57}{13.10} = 0.73.$$

$$\therefore \quad t = \frac{b}{s_b} = \frac{5.03}{0.73} = 6.89$$

(v) The critical region is $|t| \geq t_{0.025,(10)} = 2.23$.

(vi) **Conclusion.** Since the computed value of $t = 6.89$ falls in the critical region, so we reject the null hypothesis and may conclude that there is sufficient reason to say at the 5% level of significance that heights and weights are related.

21.3.2. Testing Hypotheses about α, the Intercept of Population Regression. Suppose that we wish to test the hypothesis that α, the population intercept, has some specified value α_0. The sample estimate of α, is a which has been shown to be normally distributed with a mean of α and a standard deviation of $\sigma_{Y.X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X - \bar{X})^2}}$. It has already been observed that, when $\sigma_{Y.X}$ is estimated from sample data, the statistic

$$t = \frac{a - \alpha}{s_a},$$

where $s_a = s_{Y.X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X - \bar{X})^2}}$, has a Student's t-distribution with $\nu = n-2$ degrees of freedom. Hence we use t-distribution to test the null

hypothesis $H_0 : \alpha = \alpha_0$ against an appropriate alternative hypothesis. The rest of the procedure is the same.

21.3.3. Testing Hypothesis about Mean Value $\mu_{Y.X}$ for a Given Value X_0. Quite frequently, we are interested in testing the hypothesis that the mean of a population of Y_i's when $X=X_0$, has some specified value μ_0. That is, we wish to test $H_0 : \mu_{Y.X_0} = \mu_0$ against a suitable alternative hypothesis. We obtain an unbiased estimate of $\mu_{Y.X}$ for $X=X_0$ from its sample counterpart $Y=a+bX_0$. It can be shown that, if H_0 is true, the statistic

$$t = \frac{(a + bX_0) - \mu_0}{s_{\hat{Y}}},$$

where $s^2_{\hat{Y}} = s^2_{Y.X}\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X - \bar{X})^2}\right]$, has a Student's t-distribution with $v=n-2$ degrees of freedom. Consequently, t-distribution is used to test a hypothesis about $\mu_{Y.X}$ for a given value of X. The test procedure is carried out in the usual manner.

21.3.4. Testing Hypothesis about Population Variance $\sigma^2_{Y.X}$. If we wish to test the null hypothesis $H_0 : \sigma^2_{Y.X} = \sigma^2_0$ where σ^2_0 is some specified value of $\sigma^2_{Y.X}$, against a suitable alternative hypothesis, then the statistic

$$\chi^2 = \frac{(n-2)\, s^2_{Y.X}}{\sigma^2_0}$$

has, if H_0 is true, a chi-square distribution with $(n-2)$ degrees of freedom. We reject H_0 if the calculated value of χ^2 falls in the critical region, otherwise we accept it.

21.3.5. Testing Hypothesis about Equality of Regression Co-efficient of Two Regression Lines. Suppose that there are two linear regression lines

$$\mu_{Y_1.X} = \alpha_1 + \beta_1 X_1 \text{ and } \mu_{Y_2.X} = \alpha_2 + \beta_2 X_2$$

and we wish to test the hypothesis that the two regression lines have equal regression co-efficients, that is, we wish to test $H_0 : \beta_1 = \beta_2$.

Let b_1 and b_2, the least-squares estimates of β_1 and β_2 respectively, be obtained from two random samples of sizes, n_1 and n_2, the first sample is taken from one population and the second sample is drawn

from another population. Then the statistic $b_1 - b_2$ is normally distributed with mean $= \beta_1 - \beta_2$ and variance $= \frac{\sigma^2_{Y_1.X}}{\Sigma(X_{1i}-\bar{X}_1)^2} + \frac{\sigma^2_{Y_2.X}}{\Sigma(X_{2i}-\bar{X}_2)^2}$ as the populations are assumed to be normal for each X. If the two populations have equal variances, then the variance becomes

$$\sigma^2_{Y.X}\left[\frac{1}{\Sigma(X_{1i}-\bar{X}_1)^2} + \frac{1}{\Sigma(X_{2i}-\bar{X}_2)^2}\right].$$

As $\sigma^2_{Y.X}$ is generally not known, its pooled estimate, denoted by $s^2_{Y.X.p}$ is obtained by the relation

$$s^2_{Y.X.p} = \frac{\Sigma(Y_{1i}-\hat{Y}_1)^2 + \Sigma(Y_{2i}-\hat{Y}_2)^2}{n_1 + n_2 - 4}$$

It can be shown that the statistic

$$t = \frac{(b_1 - b_2) - (\beta_1 - \beta_2)}{s_{Y.X.p.}\left[\frac{1}{\Sigma(X_{1i}-\bar{X}_1)^2} + \frac{1}{\Sigma(X_{2i}-\bar{X}_2)^2}\right]^{1/2}}$$

conforms to a t-distribution with $v = n_1 + n_2 - 4$ degrees of freedom. Hence, for testing the hypothesis that $\beta_1 = \beta_2$ given equal variances, we use the statistic

$$t = \frac{b_1 - b_2}{s_{Y.X.p.}\left[\frac{1}{\Sigma(X_{1i}-\bar{X}_1)^2} + \frac{1}{\Sigma(X_{2i}-\bar{X}_2)^2}\right]^{1/2}}$$

which, if H_0 is true, has a t-distribution with $n_1 + n_2 - 4$ degrees of freedom. The rest of the procedure for testing the hypothesis is the same.

A $(1-\alpha)100$ per cent confidence interval for $\beta_1 - \beta_2$ is given by

$$(b_1 - b_2) \pm t_{\alpha/2,(n_1+n_2-4)}\, s_{Y.X.p.}\sqrt{\left[\frac{1}{\Sigma(X_{1i}-\bar{X}_1)^2} + \frac{1}{\Sigma(X_{2i}-\bar{X}_2)^2}\right]}$$

Example 21.6. Two random samples as detailed below have been drawn from two populations with equal variances:

Sample I	X_{1i}	4	3	3	5
	Y_{1i}	12	12	10	13
Sample II	X_{2i}	1	2	3	1
	Y_{2i}	6	5	8	7

(a) Find the estimates of β_1 and β_2, the regression coefficients of the two regression lines.

(b) Test the hypothesis $H_0: \beta_1 = \beta_2$ against the alternative $H_1: \beta_1 \neq \beta_2$. Use $\alpha = 0.05$. (M.Sc., P.U. 1988, 90; I.U., 1993)

(a) To estimate β_1 and β_2, we carry out the necessary calculations as below:

Sample - I

$$\Sigma X_{1i} = 15, \Sigma Y_{1i} = 47, \Sigma X_{1i}^2 = 59, \Sigma X_{1i} Y_{1i} = 179, \text{ and } \Sigma Y_{1i}^2 = 557.$$

$$\therefore \quad b_1 = \frac{(n_1)(\Sigma X_{1i} Y_{1i}) - (\Sigma X_{1i})(\Sigma Y_{1i})}{(n_1)(\Sigma X_{1i}^2) - (\Sigma X_{1i})^2} = \frac{(4)(179) - (15)(17)}{(4)(59) - (15)^2} = 1$$

Sample - II

$$\Sigma X_{2i} = 7, \Sigma Y_{2i} = 26, \Sigma X_{2i}^2 = 15, \Sigma X_{2i} Y_{2i} = 47, \text{ and } \Sigma Y_{2i}^2 = 174.$$

$$\therefore \quad b_2 = \frac{(n_2)(\Sigma X_{2i} Y_{2i}) - (\Sigma X_{2i})(\Sigma Y_{2i})}{(n_2)(\Sigma X_{2i}^2) - (\Sigma X_{2i})^2} = \frac{(4)(47) - (7)(26)}{(4)(15) - (7)^2} = 0.545$$

Thus the estimates of β_1 and β_2, the two regression coefficients of the two regression lines are $b_1 = 1$ and $b_2 = 0.545$.

(b) (i) We state our null and alternative hypotheses as

$$H_0 : \beta_1 = \beta_2, \text{ and } H_1 : \beta_1 \neq \beta_2$$

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic under H_0 is

$$t = \frac{b_1 - b_2}{s_{Y.X.p}\sqrt{\left[\frac{1}{\Sigma(X_{1i} - \bar{X}_1)^2} + \frac{1}{\Sigma(X_{2i} - \bar{X}_2)^2}\right]}}$$

which has a t-distribution with $n_1 + n_2 - 4$ $(=4)$ degrees of freedom.

(iv) Computations. The necessary computations are given below:

Sample I: $a_1 = \bar{Y}_1 - b_1\bar{X}_1 = 11.75 - (1)(3.75) = 8,$

$$\Sigma(X_{1i} - \bar{X}_1)^2 = \Sigma X_{1i}^2 - (\Sigma X_{1i})^2/n_1 = 59 - (15)^2/4 = 2.75;$$

$$\Sigma(Y_{1i} - \hat{Y}_1)^2 = \Sigma Y_{1i}^2 - a\Sigma Y_{1i} - b\Sigma X_{1i} Y_{1i}$$

$$= 557 - (8)(47) - (1)(179) = 2.00$$

Sample II: $a_2 = \bar{Y}_2 - b_2\bar{X}_2 = 6.50 - (0.545)(1.75) = 5.546;$

$$\Sigma(X_{2i}-\bar{X}_2)^2 = \Sigma X_{2i}^2 - (\Sigma X_{2i})^2/n_2 = 15-(7)^2/4 = 2.75;$$

$$\Sigma(Y_{2i} - \hat{Y}_{2i})^2 = \Sigma Y_{2i}^2 - a_2\Sigma Y_{2i} - b_2\Sigma X_{2i}Y_{2i}$$

$$= 174-(5.546)(26)-(0.545)(47) = 4.189$$

Now $$s^2_{Y.X.p} = \frac{\Sigma(Y_{1i} - \hat{Y})^2 + \Sigma(Y_{2i} - \hat{Y}_2)^2}{n_1 + n_2 - 4} = \frac{2.00 + 4.189}{4 + 4 - 4}$$

$= 1.5473$ so that $s_{Y.X.p} = \sqrt{1.5473} = 1.2439$, and

$$t = \frac{1 - 0.545}{1.2439\sqrt{\frac{1}{2.75} + \frac{1}{2.75}}} = \frac{0.455}{(1.2439)(0.8528)}$$

$$= \frac{0.455}{1.0608} = 0.429$$

(v) The critical region is $|t| \geq t_{0.025,(4)} = 2.776$

(vi) Conclusion. Since the calculated value of $t = 0.429$ does not fall in the critical region, we therefore cannot reject H_0 and conclude that the two regression coefficients are equal. The acceptance of $H_0 : \beta_1 = \beta_2$ means that the two regression lines are parallel.

21.3.6. Testing Hypothesis about the Linearity of Regression. Frequently, we are interested in testing the null hypothesis that the regression model is linear, *i.e.* we wish to test $H_0 : \mu_{Y.X} = \alpha + \hat{\beta}X$. For this purpose, we select a random sample of n observations with k distinct values of X and for each (or at least one) distinct values of X, the Y observations are repeated $n_1, n_2, \ldots, n_k$ $(\Sigma n_i = n)$ times respectively, as indicated in the following table;

X	Y-values				Sum
x_1	y_{11},	y_{12},	...,	y_{1n_1}	$y_{1\cdot}$
x_2	y_{21},	y_{22},	...,	y_{2n_2}	$y_{2\cdot}$
...	...	...	...	...	...
x_k	y_{k1},	y_{k2},	...,	y_{kn_k}	$y_{k\cdot}$

where $y_{i\cdot}$ represents the sum of y-values corresponding to x_i. It has been shown that, when H_0 is true, the statistic

$$F = \frac{\chi_1^2/(k-2)}{\chi_2^2/(n-k)}$$

where $\chi_1^2 = \sum \frac{y_{i\cdot}^2}{n_i} - \frac{(\sum y_{ij})^2}{n} - b^2\sum(x-\bar{x})^2$, and

$$\chi_2^2 = \sum y_{ij}^2 - \sum \frac{y_{i\cdot}^2}{n_i}$$

has an F-distribution with $v_1 = k-2$ and $v_2 = n-k$ degrees of freedom. To determine a, b and $\sum(x-\bar{x})^2$, we fit the assumed regression line, taking the n data points.

We reject the null hypothesis of linearity of regression when the computed value of F falls in the critical region located in the right tail of the F-distribution with $v_1 = k-2$ and $v_2 = n-k$ degrees of freedom at the significance level of α. The rest of the *procedure* is the same.

Example 21.7. The following data show the heights (X) and weights (Y) of twelve men. We selected the heights in advance and then observed the weights of a random group of men having the selected heights.

X:	60,	60,	60,	62,	62,	62,	62,	64,	64,	70,	70,	70,
Y:	110,	135,	120,	120,	140,	130,	135,	150,	145,	170,	185,	160

Test the hypothesis at the 0.05 level of significance, that the regression is linear.

Using all the twelve values, we find that the estimated regression line is

$$\hat{Y} = -179.42 + 5.03X, \text{ and}$$

$$\sum(x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{12} = 49068 - \frac{(766)^2}{12} = 171.67.$$

To test the hypothesis, we proceed as below:

(i) We state our null and alternative hypotheses as

H_0 : the regression is linear, *i.e.* $\mu_{Y.X} = \alpha + \beta X$, and

H_1 : the regression is non-linear.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{\chi_1^2/(k-2)}{\chi_2^2/(n-k)},$$

where $\chi_1^2 = \sum \frac{y_{i\cdot}^2}{n_i} - \frac{(\sum y_{ij})^2}{n} - b^2 \sum (x - \bar{x})^2$, and

$$\chi_2^2 = \sum y_{ij}^2 - \sum \frac{y_{i\cdot}^2}{n_i};$$

which, if H_0 is true, has an F-distribution with $v_1 = k-2$ and $v_2 = n-k$ degrees of freedom.

(iv) Computations. To find the necessary calculations, we re-arrange the data as follows:

X	Y-values				$Y_{i\cdot}$
60	110,	135,	120		365
62	120,	140,	130,	135	525
64	150,	145			295
70	170,	185,	160		515
				$\sum y_{ij}$	1700

Now $\chi_1^2 = \sum \frac{y_{i\cdot}^2}{n_i} - \frac{(\sum y_{ij})^2}{n} - b^2 \sum (x - \bar{x})^2$

$$= \frac{(365)^2}{3} + \frac{(525)^2}{4} + \frac{(295)^2}{2} + \frac{(515)^2}{3} - \frac{(1700)^2}{12} - (5.03)^2 (171.67)$$

$$= 245{,}235.41 - 240{,}833.33 - 4{,}343.41 = 58.67, \text{ and}$$

$$\chi_2^2 = \sum y_{ij}^2 - \sum \frac{y_{i\cdot}^2}{n_i};$$

$$= (110)^2 + (135)^2 + \ldots + (160)^2 - \left[\frac{(365)^2}{3} + \frac{(525)^2}{4} + \frac{(295)^2}{2} + \frac{(515)^2}{3}\right]$$

$$= 246{,}100 - 245{,}235.41 = 864.59$$

$$\therefore \quad F = \frac{\chi_1^2 / (k-2)}{\chi_2^2 / (n-k)} = \frac{58.67 / (4-2)}{864.59 / (12-4)} = \frac{29.335}{108.074} = 0.27$$

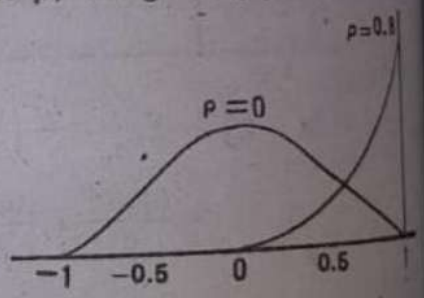
and $v_1 = 2$ and $v_2 = 8$.

(v) The critical region is $F \geq F_{0.05}(2, 8) = 4.46$

(vi) Conclusion. Since the computed value of $F = 0.27$ does not fall in the critical region, we therefore, cannot reject H_0 and may conclude that there is sufficient evidence to indicate that the regression is a straight line.

21.4 CONFIDENCE INTERVAL ESTIMATE FOR POPULATION CORRELATION CO-EFFICIENT.

Let r be the sample correlation co-efficient obtained from a random sample of n pairs of values from a bivariate normal population having a linear correlation ρ. As a sample mean $\bar{x}$ is used to estimate the population mean μ, in the same way r is used to estimate the value of ρ. The sampling distribution of r is shown by Sir *R.A.* Fisher to depend only on ρ and n. The standard deviation of the distribution of r is approximately equal to $\frac{1-\rho^2}{\sqrt{n}}$. It is also shown that the distribution of r is far from normal for large value of ρ, being sharply skewed in the neighbourhood of $\rho = \pm 1$. If the sample is large enough ($n > 400$) and if ρ is only moderately large, then r is approximately normally distributed with mean ρ and standard deviation $\frac{1-\rho^2}{\sqrt{n}}$.



(Distribution of r for $\rho=0$ and $\rho=.8$ when $n=9$)

Thus the standard error of r is $\frac{1-\rho^2}{\sqrt{n}}$, but it is customary to take it as $\frac{1-r^2}{\sqrt{n}}$ inspite of the fact that r is a biased estimate of ρ. Thus the statistic $Z = \frac{r-\rho}{\frac{1-r^2}{\sqrt{n}}}$ may be regarded as a standard normal variable.

But this is not recommended for use because it is not valid if n is small and ρ is large. It should be remembered that the distribution of r for small samples is skew and the skewness increases with ρ.

We can change the *non-normal* distribution of r by a simple transformation into an approximately normal distribution. This change of variable, known as "Fisher's z-transformation", is from r to Z_ρ where

$$Z_f = \frac{1}{2} ln \frac{1+r}{1-r} = 1.1513 \log \frac{1+r}{1-r}.$$

Fisher (1890–1962) showed that the random variable Z_f is approximately normally distributed with a mean of $\mu_z = \frac{1}{2} ln \frac{1+\rho}{1-\rho} = 1.1513 \log \frac{1+\rho}{1-\rho}$ and a standard deviation approximately of $\frac{1}{\sqrt{n-3}}$. Hence the statistic

$$Z = \frac{\left[1.1513 \log \frac{1+r}{1-r}\right] - \left[1.1513 \log \frac{1+\rho}{1-\rho}\right]}{1/\sqrt{n-3}} = \frac{Z_f - \mu_z}{1/\sqrt{n-3}}$$

is approximately $N(0, 1)$, and we can assert that

$$P\left[-z_{\alpha/2} < \frac{Z_f - \mu_z}{1/\sqrt{n-3}} < z_{\alpha/2}\right] \simeq 1 - \alpha$$

This statement after simplification becomes

$$P\left[Z_f - \frac{z_{\alpha/2}}{\sqrt{n-3}} < \mu_z < Z_f + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right] \simeq 1 - \alpha$$

As corresponding to a particular value of r obtained from a particular sample, we have a particular value of z_f, therefore an approximate $100(1-\alpha)$ per cent confidence interval for μ_z is given by

$$z_f \pm \frac{z_{\alpha/2}}{\sqrt{n-3}}.$$

Thus we see that, in order to compute a confidence interval for the population correlation coefficient ρ, we have to first construct a confidence interval for μ_z and then have to transform it into an interval for ρ.

Using Fisher's z-table, we find values, denoted by ρ_L and ρ_U, that correspond to Fisher's z-values equal to $z_f - \frac{z_{\alpha/2}}{\sqrt{n-3}}$ and $z_f + \frac{z_{\alpha/2}}{\sqrt{n-3}}$.

Hence an approximate $100(1-\alpha)$ per cent confidence interval for ρ is (ρ_L, ρ_U).

Example 21.8 A random sample of size 28 pairs from a bivariate normal population showed a correlation co-efficient of 0.7. Find a 95% confidence interval for the population correlation coefficient ρ.

(I.U., M.Sc. 1990)

The degree of confidence is 0.95, therefore $z_{0.025} = 1.96$.

The 95% confidence interval for $\mu_z = 1.1513 \log \frac{1+\rho}{1-\rho}$, is

$$z_f - \frac{1.96}{\sqrt{n-3}} < \mu_z < z_f + \frac{1.96}{\sqrt{n-3}}$$

Now $\quad z_f = 1.1513 \log \frac{1+r}{1-r} = 1.1513 \log \frac{1+0.7}{1-0.7} = 0.87$, and

$$n - 3 = 25.$$

Substituting, we get

$$0.87 - \frac{1.96}{\sqrt{25}} < \mu_z < 0.87 + \frac{1.96}{\sqrt{25}}$$

or $\quad 0.87 - 0.392 < \mu_z < 0.87 + 0.392$

or $\quad 0.48 < \mu_z < 1.26.$

Using Fisher's z-table, we find values, ρ_L and ρ_U, that correspond to Fisher's z-values equal to 0.48 and 1.26 respectively.

Thus $\quad \rho_L = 0.446$ and $\rho_U = 0.851$.

Hence the approximate 95% confidence interval for ρ is (0.45, 0.85).

21.5 HYPOTHESIS TESTING ABOUT CORRELATION COEFFICIENT

As stated earlier, the sampling distribution of r, the sample estimate of the population correlation co-efficient ρ, is neither a normal distribution nor a distribution that becomes approximately normal as the sample size increases. However, we change r into another random variable, denoted by Z_f and defined as $Z_f = 1.1513 \log \frac{1+r}{1-r}$, which is approximately normal with a mean of $\mu_z = 1.1513 \log \frac{1+\rho}{1-\rho}$ and a standard deviation approximately of $\frac{1}{\sqrt{n-3}}$. Thus the statistic

$$Z = \frac{Z_f - \mu_z}{1/\sqrt{n-3}}$$

is approximately standard normal regardless of the value of ρ and hence provides a method to test the hypotheses about ρ. It is important to note that the standard error of Z_f is independent of ρ and that it is Z_f, rather than r, that is used for testing hypotheses about ρ.

The random variable Z_f is used to test the hypotheses that

(a) population correlation co-efficient ρ is equal to a specified value ρ_0, where ρ_0 is not equal to zero, *i.e.* $H_0 : \rho = \rho_0 \ (\neq 0)$,

(b) correlation coefficients of two populations are equal, *i.e.*, $H_0 : \rho_1 = \rho_2$.

To test the hypothesis H_0: $\rho = 0$, the test-statistic to be used, is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, which conforms to *Student's t*-distribution.

21.5.1. Testing the Hypothesis that ρ, the Population Correlation Coefficient equals Some Specified Value other than Zero. The procedure for testing the null hypothesis $H_0 : \rho = \rho_0$ where $\rho_0 \neq 0$, is outlined below:

(i) Set up the null hypothesis $H_0 : \rho = \rho_0$ and formulate an appropriate alternative hypothesis.

(ii) Decide on the significance level of size α.

(iii) The test-statistic to be used is

$$Z = \frac{Z_f - \mu_z}{1/\sqrt{n-3}} = (Z_f - \mu_z)\sqrt{n-3}\ ,$$

where $Z_f = 1.1513 \log \dfrac{1+r}{1-r}$ and $\mu_z = 1.1513 \log \dfrac{1+\rho}{1-\rho}$.

The variable Z is approximately standard normal.

(iv) The critical region is $Z \leq -z_{\alpha/2}$ and $Z \geq z_{\alpha/2}$ for $H_1 : \rho \neq \rho_0$;

$Z \leq -z_\alpha$ for $H_1 : \rho < \rho_0$;

$Z \geq z_\alpha$ for $H_1 : \rho > \rho_0$.

(v) Compute the value of Z from the sample data.

(vi) Decide as under:

Reject H_0, if z falls in the critical region.

Accept H_0, otherwise.

Example 21.9. A random sample of size 28 pairs from a bivariate normal population showed a correlation coefficient of 0.7. Is this value consistent with the assumption that the correlation coefficient in the population is 0.5?

(i) We set up our null and alternative hypotheses as

$$H_0 : \rho = 0.5 \text{ and } H_1 : \rho \neq 0.5.$$

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic to be used in this case is

$$Z = \frac{Z_f - \mu_z}{1/\sqrt{n-3}},$$

where $Z_f = 1.1513 \log \frac{1+r}{1-r}$ and $\mu_z = 1.1513 \log \frac{1+\rho_0}{1-\rho_0}$. The variable Z is approximately standard normal.

(iv) The critical region for $\alpha = 0.05$ is $Z < -1.96$ and $Z > 1.96$.

(v) Computations: We are given $r = 0.7$. $\rho = 0.5$ and $n = 28$.

Now $$Z_f = 1.1513 \log \frac{1+0.7}{1-0.7} = 1.1513 \log \frac{1.7}{0.3} = 0.87, \text{ and}$$

$$\mu_z = 1.1513 \log \frac{1+0.5}{1-0.5} = 1.1513 \log \frac{1.5}{0.5} = 0.55;$$

$$\therefore \quad z = \frac{0.87 - 0.55}{1/\sqrt{28-3}} = (0.32)(5) = 1.60.$$

(vi) **Conclusion.** Since the computed value $z = 1.60$ does not fall in the critical region, so we accept our null hypothesis and conclude that the correlation co-efficient in the population might be 0.5.

21.5.2. Testing Hypothesis about the Equality of Two Correlations. Let r_1 and r_2 be the correlation co-efficients of two random samples of sizes n_1 and n_2 pairs, drawn from two bivariate normal populations with correlation co-efficients ρ_1 and ρ_2. Then to test the hypothesis $H_0 : \rho_1 = \rho_2$, we calculate

$$Z_{f_1} = 1.1513 \log \frac{1+r_1}{1-r_1} \text{ and } Z_{f_2} = 1.1513 \log \frac{1+r_2}{1-r_2}.$$

Since Z_{f_1} and Z_{f_2} are approximately normally distributed, therefore the difference $Z_{f_1} - Z_{f_2}$, if $H_0 : \rho_1 = \rho_2$ is true, is approximately normally distributed with a mean zero and standard deviation $\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}$.

The test-statistic would then be

$$Z = \frac{Z_{f_1} - Z_{f_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

which is approximately standard normal.

The rest of the procedure for testing $H_0 : \rho_1 = \rho_2$ is the same.

Example 21.10. A random sample of 28 pairs from a certain bivariate normal population gave $r = 0.6$; another random sample of 23 pairs from another bivariate population gave $r = 0.4$. Test at the 0.05 significance level the hypothesis $H_0 : \rho_1 = \rho_2$ against the alternative $H_1 : \rho_1 \neq \rho_2$.

(i) Our null and alternative hypotheses are

$$H_0 : \rho_1 = \rho_2 \text{ and } H_1 : \rho_1 \neq \rho_2 .$$

(ii) The significance level is at $\alpha = 0.05$.

(iii) The test-statistic to be used under H_0, is

$$Z = \frac{Z_{f_1} - Z_{f_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

where $Z_{f_1} = 1.1513 \log \frac{1 + r_1}{1 - r_1}$ and $Z_{f_2} = 1.1513 \log \frac{1 + r_2}{1 - r_2}$

The variable Z is approximately standard normal.

(iv) The critical region is $|Z| \geq 1.96$.

(v) Computations:

$$Z_{f_1} = 1.1513 \log \frac{1 + 0.6}{1 - 0.6} = 1.1513 \log \frac{1.6}{0.4} = 0.69, \text{ and}$$

$$Z_{f_2} = 1.1513 \log \frac{1 + 0.4}{1 - 0.4} = 1.1513 \log \frac{1.4}{0.6} = 0.42, \text{ so that}$$

$$z = \frac{0.69 - 0.42}{\sqrt{\frac{1}{25} + \frac{1}{20}}} = \frac{0.27}{0.30} = 0.90$$

(vi) **Conclusion.** Since the computed value $z = 0.90$ does not fall in the critical region, so we accept our hypothesis and conclude that the data present sufficient evidence to indicate that the correlation co-efficients of the two populations are equal.

21.5.3. Testing the Hypothesis that $\rho = 0$. We are often interested in testing the null hypothesis that the population correlation co-efficient equals zero. That is, we wish to test $H_0 : \rho = 0$ (*i.e.* there is no linear correlation between the variables X and Y). We have seen that the sampling distribution of r, the sample correlation coefficient, is skew when ρ is not zero. However, when $\rho = 0$, the sampling distribution of r is symmetric. This property makes it possible to test the hypothesis $H_0 : \rho = 0$ by using the t-distribution. Thus when the random variables X and Y are normally distributed and $\rho = 0$, the statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

has a Student's t-distribution with $v = n-2$ degrees of freedom. We would reject H_0, if

$$t < -t_{\alpha,(n-2)} \text{ when } H_1 \text{ is } \rho < 0,$$

$$t > t_{\alpha,(n-2)} \text{ when } H_1 \text{ is } \rho > 0,$$

$$|t| \geq t_{\alpha/2,(n-2)} \text{ when } H_1 \text{ is } \rho \neq 0.$$

We accept H_0 otherwise and conclude that X and Y are not linearly correlated. It is interesting to note that testing the hypothesis that $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ is equivalent to testing the hypothesis that $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. Thus a test that will reject the hypothesis $H_0 : \beta = 0$, will also reject the hypothesis $H_0 : \rho = 0$. A t-test used to reject or to accept $H_0 : \rho = 0$ is satisfactory for any sample size.

An alternative way of testing the hypothesis $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ is the use of F-statistic. The statistic

$$F = \frac{r^2(n-2)}{1-r^2}$$

has, if H_0 is true, an F-distribution with $v_1 = 1$ and $v_2 = n-2$ degrees of freedom. The two tests are equal because the t-statistic is related to the F-statistic (with 1 *d.f.* in the numerator) by $t^2 = F$.

An F-test can also be used to test the hypothesis,

$H_0 : \rho^2 = 0$, the simple coefficient of determination is zero, against

$H_1 : \rho^2 \neq 0$, the coefficient of determination is not zero.

If $F > F_{\alpha,(v_1, v_2)}$, we reject H_0, otherwise we accept H_0. The rest of the procedure is the same.

To test the hypothesis about rank correlation, r_s; (when the sample size is neither too small nor too large), *i.e.* to test

H_0:The ranks of two population data sets are not correlated, against

H_1:The ranks are correlated,

the appropriate test-statistic is

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

which, if H_0 is true, has a Student's t-distribution with $v=n-2$ *df*. We would reject H_0 if t falls in the critical region.

Example 21.11. A random sample of 20 pairs of observations gives a co-efficient of correlation of 0.45. Test the hypothesis at the 0.05 level of significance that the correlation co-efficient in the population is zero.

(i) We state the null and alternative hypotheses as

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0.$$

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which, if H_0 is true, has a t-distribution with $(n-2)=8$ degrees of freedom.

(iv) The critical region is $|t| \geq t_{0.025,(18)} = 2.10$.

(v) Computations. Substituting the values, we get

$$t = \frac{0.45\sqrt{20-2}}{\sqrt{1-(0.45)^2}} = \frac{1.91}{0.89} = 2.14$$

(vi) **Conclusion.** Since the computed value of $t=2.14$ falls in the rejection region, we therefore reject our null hypothesis and may conclude that the correlation co-efficient in the population differs from zero.

Example 21.12. (a) Find the least value of r in a sample of 18 pairs from a bivariate normal population, that is significant at 5% level.

(b) How many pairs of observations must be included in a sample in order that an observed correlation coefficient of value 0.47 shall have a calculated value of t greater than 2.12. (P.U., B.A/B.Sc. 1991)

(a) The significance of r is tested by the statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Substituting the value of n, we get

$$t = \frac{r\sqrt{18-2}}{\sqrt{1-r^2}} = \frac{4r}{\sqrt{1-r^2}}$$

For a significant value of r, t should be $\geq t_{0.025,(16)} = 2.12$

That is $\quad \dfrac{4r}{\sqrt{1-r^2}} \geq 2.12$

or $\quad 16r^2 \geq (2.12)^2 (1-r^2)$

or $\quad 20.4944\, r^2 \geq 4.4944$ or $r^2 \geq \dfrac{4.4944}{20.4944} = 0.22$

$\therefore \quad |r| \geq 0.47.$

Hence the required least value of $r = 0.47$.

(b) Given $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} > 2.12$

Substituting values, we get

$$\frac{0.47\sqrt{n-2}}{\sqrt{1-(0.47)^2}} > 2.12$$

or $\quad (0.47)^2 (n-2) > (2.12)^2 (1-0.2209)$

or $\quad (n-2) > \dfrac{(4.4944)(0.7791)}{0.2209}$

or $\quad (n-2) > 15.85$ or $n > 17.85$

Hence the number of pairs must be 18.

21.5.4. Testing Hypothesis about the Equality of Several Correlations. Suppose that $r_1, r_2, \ldots, r_k$ are the simple correlation coefficients calculated from k random samples of $n_1, n_2, \ldots, n_k$ pairs of observations respectively, drawn from bivariate normal populations with corresponding correlation coefficients $\rho_1, \rho_2, \ldots, \rho_k$. We desire to test the null hypothesis that the population correlation coefficients are equal, *i.e.*

$H_0 : \rho_1 = \rho_2 = \ldots = \rho_k$ ($= \rho$, say) against the alternative hypothesis

H_1 : Not all population correlation co-efficient are equal.

To test this hypothesis, the values of r are converted to z_r-values (Fisher's z-transformation). That is, we define $Z_{f_1}, Z_{f_2}, \ldots, Z_{f_k}$ by the relation

$$Z_{f_i} = 1.1513 \log \frac{1 + r_i}{1 - r_i}.$$

We have already seen that for large samples, Z_{f_i} are all distributed normally with $E(Z_{f_i}) = 1.1513 \log \frac{1+\rho}{1-\rho}$ and $\text{Var}(Z_{f_i}) = \frac{1}{n_i - 3}$.

Furthermore, the best linear unbiased estimate of $1.1513 \log \frac{1+\rho}{1-\rho}$ is given by $\bar{Z} = \frac{\sum (n_i - 3) Z_{f_i}}{\sum (n_i - 3)}$. It has been shown that the statistic

$$u = \sum_{i=1}^{k} (n_i - 3)(Z_{f_i} - \bar{Z})^2$$

is distributed approximately as χ^2-distribution with $(k-1)$ degrees of freedom when the null hypothesis is true. The rest of the test-procedure is the same.

21.6 INFERENCE IN PARTIAL, MULTIPLE CORRELATION AND REGRESSION

Hypotheses about the partial correlation, multiple correlation and regression may be tested in the same way as those of simple correlation and regression.

21.6.1. Testing the Hypothesis that a Partial Correlation Coefficient of Order k in the Population is Zero. Assuming that Y and X_i's are jointly normally distributed and the null hypothesis that the population partial correlations of order k (*i.e.* with k secondary subscripts), estimated from a random sample of size n, are individually equal to zero, R.A. Fisher has shown that the statistic

$$t = \frac{r_{12.34\ldots p}\sqrt{n-k-2}}{\sqrt{1 - r^2_{12.34\ldots p}}}$$

follows a Student's t-distribution with $\nu = n-k-2$ *d.f.*, if the null hypothesis is true. We reject H_0, if the calculated value of t falls in the critical region, otherwise we accept it. The procedure is illustrated by Example 21.13.

An Alternative way of testing the hypothesis that a particular partial correlation of order k (k denotes the number of secondary subscripts) in the population is zero, that is

$H_0 : \rho_{12.34...p} = 0$, (*i.e.* a particular partial correlation of order k is zero), and

$H_1 : \rho_{12.34...p} \neq 0$.

is based on F-test. If $r_{12.34...p}$ is the corresponding partial correlation calculated from a sample of size n, then the statistic

$$F = \frac{r^2_{12.34...p}\,(n-k-2)}{1-r^2_{12.34...p}}$$

has, if H_0 is true, an F-distribution with $\nu_1 = 1$ and $\nu_2 = n-k-2$. We reject H_0, if $F > F_\alpha(1, n-k-2)$. Otherwise, we accept H_0.

Hypotheses about the equality of several partial correlation coefficients of order k can be tested in the same way as the equality of several simple correlation coefficients, the only difference would be to reduce the effective size of sample n_i by the number of variables held constant, *i.e.* k.

Example 21.13. From a random sample of 21 sets of values from a normal population the calculated value of a partial correlation of order three is 0.40. Is this consistent with the hypothesis that the corresponding partial correlation in the population is zero? Let $\alpha = 0.05$.

(i) We state our null and alternative hypotheses as

$H_0 : \rho_{12.345} = 0$, *i.e.*, a partial correlation of order three in the population is zero, and

$H_1 : \rho_{12.345} \neq 0$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$t = \frac{r_{12.345}\sqrt{n-k-2}}{\sqrt{1-r^2_{12.345}}}$$

which, if H_0 is true, has a t-distribution with $\nu = n-k-2$ degrees of freedom. Here $k = 3$, so $d.f. = 21 - 3 - 2 = 16$.

(iv) The critical region is $|t| \geq t_{0.025,(16)} = 2.12$

(v) Computations. Substituting the values, we get

$$t = \frac{(0.40)\sqrt{21-3-2}}{\sqrt{1-(0.40)^2}} = \frac{1.60}{0.92} = 1.74$$

(vi) **Conclusion.** Since the calculated value of $t = 1.74$ does not fall in the critical region, so we accept H_0 and conclude that the given partial correlation coefficient is consistent with the hypothesis that the corresponding partial correlation in the population is zero.

21.6.2. Testing the Hypothesis that a Multiple Correlation Coefficient is Zero. To test the hypothesis that a multiple correlation coefficient in the population is zero, we compute the corresponding multiple correlation $R_{y.12...p}$ from the sample of size n from a $(p+1)$-variate normal population and use the statistic

$$F = \frac{R^2_{y.12...p}\,(n-p-1)}{(1-R^2_{y.12...p})\,p}$$

which, if H_0 is true, has an F-distribution with $v_1 = p$ and $v_2 = n-p-1$ degrees of freedom. This test is equivalent to testing the hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ simultaneously. A test of such a hypothesis is also called a test of the *overall significance* of the estimated multiple regression. The rest of the test procedure is the same.

Example 21.14. The marks in statistics (Y) are expressed as a function of marks in mathematics (X_1) economics (X_2) and intelligence tests (X_3). For a random sample of 50 students, the multiple correlation co-efficient $R_{y.123}$ was found to be 0.582. Test the hypothesis that the multiple correlation co-efficient in the population is zero. Let $\alpha = 0.05$. (P.U., M.Sc. 1988)

(i) We state our null and alternative hypotheses as

H_0: The multiple correlation co-efficient in the population is zero,

i.e. $H_0 : \rho_{y.123} = 0$, and

$H_1 : \rho_{y.123} \neq 0$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic, if H_0 is true, is

$$F = \frac{R^2_{y.123}\,(n-p-1)}{(1-R^2_{y.123})\,p},$$

which has an F-distribution with $v_1 = p$ and $v_2 = n-p-1$ degrees of freedom. Here $p = 3$.

(iv) **Computations.** Substituting the values, we get

$$F = \frac{(0.582)^2\,(50-3-1)}{[1-(0.582)^2]\,(3)} = \frac{15.58}{1.98} = 7.87,$$

and $v_1 = 3$, $v_2 = 50-3-1 = 46$ degrees of freedom.

(v) The critical region is $F \geq F_{0.05}(3, 46) = 2.81$.

(vi) **Conclusion.** Since the computed value of $F = 7.87$ falls in the critical region, so we reject H_0 and may conclude that the multiple correlation coefficient in the population differs from zero significantly.

21.7 ANALYSIS OF VARIANCE FOR REGRESSION

The procedure of analysing or partitioning the total variation in the dependent variable Y_i, into its components: one explained by the regression line and the other residual or unexplained part, is called *analysis of variance for regression.* Two independent and unbiased estimates of the population variance are obtained by dividing the two parts by the corresponding number of degrees of freedom. Their ratio under the assumption of normality of the values of Y_i and the variables in the population having no regression, has an F-distribution with appropriate degrees of freedom. The test procedure is usually arranged in an analysis of variance table.

21.7.1. ANOVA for Simple Linear Regression and Test of $H_0 : \beta = 0$. Given a random sample of n pairs of observations (X_i, Y_i) from a bivariate normal population, in which the variables are assumed to be unrelated, *i.e.* $\beta = 0$. Let the estimated linear regression be $Y_i = a + bX_i$. If $\bar{Y}$ denotes the mean of the n values of Y_i, we can construct the following identity

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Squaring both sides and summing over all values, we get

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

The cross-product term is evaluated as below:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n} (Y_i - a - bX_i)(a + bX_i - \bar{Y})$$

$$= a\,(\Sigma Y_i - na - b\Sigma X_i) + b(\Sigma X_i Y_i - a\Sigma X_i - b\Sigma X_i^2) - \bar{Y}\,(\Sigma Y_i - na - b\Sigma X_i)$$

But $\Sigma Y_i - na - b\Sigma X_i = 0$ and $\Sigma X_iY_i - a\Sigma X_i - b\Sigma X_i^2 = 0$ as they are the normal equations when we derive a and b. Thus the cross-product vanishes and we are left with

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Hence, we find that the amount of variation in the dependent variable given by $\Sigma(Y_i - \bar{Y})^2$ and called Total Sum of Squares, has been partitioned into two parts: (i) that which has been explained by the regression line, i.e. $\Sigma(\hat{Y}_i - \bar{Y})^2$ and referred to as the Sum of Squares due to regression, and (ii) that which is left unexplained by the regression line, i.e. $\Sigma(Y_i - \hat{Y}_i)^2$ and called the Sum of Squares of residuals or errors. Thus the equation partitioning the total sum of squares may be written as

Total SS(SST) = Residual or Error SS(SSE) + Regression SS(SSR).

Now the number of degrees of freedom corresponding to $\Sigma(Y_i - \bar{Y})^2$ is n−1, as there is one restriction of computing $\bar{Y}$ from the data, the number of degrees of freedom corresponding to $\Sigma(Y_i - \hat{Y}_i)^2$ is n−2, as its computation is subject to two restrictions (i.e., computation of a and b) and corresponding to $\Sigma(\hat{Y}_i - \bar{Y})^2$, the df = (n − 1) − (n − 2) = 1.

On the hypothesis that $H_0 : \beta = 0$, i.e. the two variables are unrelated, we get two independent and unbiased estimates of the variance $\sigma^2_{Y.X}$, when the two parts, viz. Regression sum of squares and Error or Residual sum of squares are divided by the respective degrees of freedom. That is, Mean Square Error (MSE) and Mean Square Regression (MSR) are both estimates of $\sigma^2_{Y.X}$. The expected mean squares can be shown as

$$E\left(\frac{SSE}{n-2}\right) = \sigma^2_{Y.X} \text{ and } E(SSR) = \sigma^2_{Y.X} + \beta^2 \Sigma(X_i - \bar{X})^2$$

A larger value of MSR suggests that β is not zero.

Hence, if $H_0 : \beta = 0$, i.e. the population has no regression, is true, the statistic

$$F = \frac{MSR}{MSE}$$

follows the F-distribution with $\nu_1 = 1$, and $\nu_2 = n_2 - 2$ df. We would reject $H_0 : \beta = 0$ when F exceeds $F_\alpha(1, n-2)$

But $\sum Y_i - na - b\sum X_i = 0$ and $\sum X_i Y_i - a\sum X_i - b\sum X_i^2 = 0$ as they are the normal equations when we derive a and b. Thus the cross-product vanishes and we are left with

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Hence, we find that the amount of variation in the dependent variable given by $\sum(Y_i - \bar{Y})^2$ and called *Total Sum of Squares*, has been partitioned into two parts: (i) that which has been explained by the regression line, *i.e.* $\sum(\hat{Y}_i - \bar{Y})^2$ and referred to as the *Sum of Squares due to regression*, and (ii) that which is left unexplained by the regression line, *i.e.* $\sum(Y_i - \hat{Y}_i)^2$ and called the *Sum of Squares of residuals or errors*. Thus the equation partitioning the total sum of squares may be written as

$$\text{Total } SS(SST) = \text{Residual or Error } SS(SSE) + \text{Regression } SS(SSR)$$

Now the number of degrees of freedom corresponding to $\sum(Y_i - \bar{Y})^2$ is $n-1$, as there is one restriction of computing $\bar{Y}$ from the data, the number of degrees of freedom corresponding to $\sum(Y_i - \hat{Y}_i)^2$ is $n-2$, as its computation is subject to two restrictions (*i.e.*, computation of a and b) and corresponding to $\sum(\hat{Y}_i - \bar{Y})^2$, the $df = (n-1) - (n-2) = 1$.

On the hypothesis that $H_0 : \beta = 0$, *i.e.* the two variables are unrelated, we get two independent and unbiased estimates of the variance $\sigma^2_{Y.X}$, when the two parts, viz. Regression sum of squares and Error or Residual sum of squares are divided by the respective degrees of freedom, That is, Mean Square Error (MSE) and Mean Square Regression (MSR) are both estimates of $\sigma^2_{Y.X}$. The expected mean squares can be shown as

$$E\left(\frac{SSE}{n-2}\right) = \sigma^2_{Y.X} \text{ and } E(\text{SSR}) = \sigma^2_{Y.X} + \beta^2 \sum(X_i - \bar{X})^2$$

A larger value of MSR suggests that β is not zero.

Hence, if $H_0 : \beta = 0$, *i.e.* the population has no regression, is true, the statistic

$$F = \frac{\text{MSR}}{\text{MSE}}$$

follows the F-distribution with $v_1 = 1$, and $v_2 = n_2 - 2$ *df*. We would reject $H_0 : \beta = 0$ when F exceeds $F_\alpha(1, n-2)$

The computation of the various sums of squares can be simplified as below:

$$\text{Total } SS(\text{SST}) = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$\text{Residual or Error } SS(\text{SSE}) = \sum(Y_i - \hat{Y}_i)^2$$

$$= \sum(Y_i - a - bX_i)^2 = \sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i$$

Regression $SS(\text{SSR})$ can be found either by subtraction, or can be computed directly as

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2 = \sum(a + bX_i - \bar{Y})^2 = a\sum Y_i + b\sum X_i Y_i - \frac{(\sum Y_i)^2}{n}$$

$$= \frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum(X_i - \bar{X})^2} = b^2\sum(X_i - \bar{X})^2.$$

These results are usually arranged in an analysis of variance table as follows:

Source of Variation	*d.f.*	Sum of Squares	Mean Square	Computed F
Regression	1	$SSR = \sum(\hat{Y}_i - \bar{Y})^2 = b^2\sum(X_i - \bar{X})^2$	$MSR = \frac{b^2\sum(X - \bar{X})^2}{1}$	$F = \frac{MSR}{MSE}$
Error or Residual	$n-2$	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$MSE = \frac{\sum(Y_i - \hat{Y})^2}{n-2}$	--
Total	$n-1$	$SST = \sum(Y_i - \bar{Y})^2$	--	--

It is interesting to note that the above-mentioned SS can be expressed in terms of r, the correlation co-efficient. Thus we have

$$\sum(\hat{Y}_i - \bar{Y})^2 = b^2\sum(X_i - \bar{X})^2 = r^2\frac{S_y^2}{S_x^2} \cdot n \cdot S_x^2 = nr^2S_y^2, \text{ and}$$

$$\sum(Y_i - \hat{Y}_i)^2 = \sum[Y_i - \bar{Y} - b(X_i - \bar{X})]^2 \qquad [\because \hat{Y} - \bar{Y} = b(X_i - \bar{X})]$$

$$= \sum(Y_i - \bar{Y})^2 + b^2\sum(X_i - \bar{X})^2 - 2b\sum(Y_i - \bar{Y})(X_i - \bar{X})$$

$$= nS_y^2 + nb^2S_x^2 - 2br\,S_yS_x$$

$$= nS_y^2 + nr^2S_y^2 - 2nr^2S_y^2$$

$$= n(1 - r^2)\,S_y^2$$

Now $MSR = nr^2 S_y^2$ and $MSE = \dfrac{n(1-r^2)S_y^2}{n-2}$

$$\therefore \quad F = \frac{MSR}{MSE} = \frac{r^2(n-2)}{1-r^2},$$

which, when $H_0 : \beta = 0$ is true, follows the F-distribution with $v_1 = 1$ and $v_2 = n-2$ *d.f.*

We note that the statistic F is the square of t, *i.e.* $F = t^2$ when $v_1 = 1$. Hence this method is equivalent to testing the hypothesis $H_0 : \rho = 0$.

Example 21.15. Test the hypothesis that $\beta = 0$ at the 0.05 level of significance for the data in Example 21.5 by setting the results in an analysis of variance table.

The stepwise solution is given below:

(i) We state our hypotheses as

$H_0 : \beta = 0$, *i.e.*, the two variables are unrelated or there is no regression, and

$H_1 : \beta \neq 0$, *i.e.*, the two variables are related.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{\text{Regression Mean Square}}{\text{Residual Mean Square}},$$

which, if H_0 is true, has an F-distribution with $v_1 = 1$, $v_2 = n-2$ degrees of freedom, assuming that population is normally distributed.

(iv) Computations. To set up the analysis of variance table, we compute the necessary sums of squares as below:

In Example 21.5, we found $\Sigma Y_i = 1700$; $\Sigma Y_i^2 = 246{,}100$; $\Sigma X_i Y_i = 109{,}380$, $n = 12$, and the estimated regression line as $Y_i = -179.42 + 5.03\, X_i$. Now, we find

Total $$SS = \Sigma(Y_i - \bar{Y})^2 = \Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n}$$

$$= 246100 - \frac{(1700)^2}{12} = 5267;$$

Residual $SS = \sum(Y_i - \hat{Y}_i)^2 = \sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i$

$= 246100 - (-179.42)(1700) - (5.03)(109330)$

$= 246100 + 305014 - 550181 = 933$; and

Regression SS = Total SS − Residual SS

$= 5267 - 933 = 4334.$

The *ANOVA*-Table is

Source of Variation	*d.f.*	Sum of Squares	Mean Square	Computed F
Regression	1	4334	4334	$\frac{4334}{93.3} = 46.45$
Residual	10	933	93.3	--
Total	11	5267	--	--

(v) The critical region is $F \geq F_{0.05}(1, 10) = 4.96$

(vi) **Conclusion.** Since the computed value of $F = 46.45$ falls in the critical region, so we reject our hypothesis and may conclude that the variables in the population are related.

21.7.2. *ANOVA* for Multiple Regression and Testing Hypothesis about the β Parameters. Let us first consider a case of linear regression on two variables X_1 and X_2. Suppose we have a random sample of n sets of observation $[(Y_i, X_{1i}, X_{2i}), i = 1, 2, ..., n]$ from a trivariate normal population having $\beta_1 = \beta_2 = 0$, where β's are the regression coefficients.

Let the estimated regression line be $\hat{Y} = a + b_1X_1 + b_2X_2$. Then the following sum-of-squares identity is obtained:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

The **total SS** has as before $n-1$ degrees of freedom, while the **residual SS** has n minus the number of parameters estimated by the regression equation (in this case 3). These results are shown in the following *ANOVA*-table;

Source of Variation	*d.f.*	Sum of Squares	Mean Square	Computed F
Regression	2	$SSR = \sum(\hat{Y} - \bar{Y})^2$	$MSR = \frac{SSR}{2}$	$F = \frac{MSR}{MSE}$
Residual	$n-3$	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-3}$	--
Total	$n-1$	$SST = \sum(Y_i - \bar{Y})^2$	--	--

When the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ is true, MSR and MSE are found to be independent and unbiased estimates of the variance σ^2. Hence the variance ratio F has an F-distribution with $v_1=2$, $v_2=n-3$ degrees of freedom, and we reject the null hypothesis when F exceeds $F_\alpha(v_1, v_2)$. This test is frequently called a test for the *overall significance of the estimated* multiple regression, since it simultaneously tests the hypothesis that each $\beta_j = 0$ $(j > 0)$ against

$$H_1 : \text{At least one of the } \beta_j \neq 0.$$

The sums of squares can be computed by the following formulas:

$$SST = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - (\sum Y_i)^2/n;$$

$$SSE = \sum(Y_i - \hat{Y})^2 = \sum Y_i^2 - a\sum Y_i - b_1\sum X_1 Y - b_2\sum X_2 Y; \text{ and}$$

$$SSR = SST - SSE.$$

We may generalise this discussion by considering a linear regression on p variables $X_1, X_2, \ldots, X_p$. The random sample in this case will consist of n sets of values $[(Y_i, X_{1i}, X_{2i}, \ldots, X_{pi}), i = 1, 2, \ldots, n]$. The $(p+1)$-variate population is again assumed to be uncorrelated normal with variance σ^2 and we wish to test the hypothesis that the population regression co-efficients equal zero, *i.e.*, $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$. Let the estimated regression equation be

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_pX_p.$$

Then the sum-of-squares identity will be

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 + \sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2$$

Here the Residual SS will have $n-p-1$ degrees of freedom as $p+$ parameters are to be estimated and Regression SS term will have p degrees of freedom. The analysis of variance table would be as below:

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Regression	p	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Residual	$n-p-1$	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-p-1}$	--
Total	$n-1$	$SST = \sum(Y_i - \bar{Y})^2$	--	--

It has been shown that the two **Mean Squares** provide independent and unbiased estimates of σ^2, when the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, is true. Hence the statistic F has an F-distribution with $v_1 = p$, $v_2 = n - p - 1$ degrees of freedom and we reject H_0 if F exceeds $F_\alpha(p, n - p - 1)$. Rejection of H_0 implies that at least one of the independent variables $X_1, X_2, \ldots, X_p$ contributes significantly to the regression model.

The simplification of the various sum of squares-terms described above, gives interesting results. If R denotes the multiple correlation co-efficient (the order of the co-efficient corresponds to the number of independent variables used in the regression equation), then

$$\text{Regression} \quad SS = \sum(\hat{Y}_i - \bar{Y})^2 = R^2\sum(Y_i - \bar{Y})^2, \text{ and}$$

$$\text{Residual} \quad SS = \sum(Y_i - \hat{Y}_i)^2 = (1 - R^2)\sum(Y_i - \bar{Y})^2 .$$

Therefore, the statistic F, the ratio of two unbiased estimates of σ^2, will have an F-distribution with $v_1 = p$, $v_2 = n-p-1$ degrees of freedom and will be used to test the null hypothesis that the multiple correlation co-efficients in the $(p+1)$-variable population do not exist. The results can be assembled in an *ANOVA*-table as below:

Source of Variation	*d.f.*	Sum of Squares	Mean Square
Regression	p	$SSR = R^2\sum(Y_i - \bar{Y})^2$	$MSR = \frac{SSR}{p}$
Residual	$n-p-1$	$SSE = (1-R^2)\sum(Y_i - \bar{Y})^2$	$MSE = \frac{SSE}{n-p-1}$
Total	$n-1$	$SST = \sum(Y_i - \bar{Y})^2$	--

$$\text{Computed } F = \frac{MSR}{MSE} = \frac{R^2(n-p-1)}{(1-R^2)p} .$$

Hence, in this case, we observe that the statistic F depends on R, the multiple correlation co-efficient and the degrees of freedom.

Example 21.16. Consider the following set of data:

Y	4	6	5	8	12	13	19
X_1	1	1	2	3	4	5	7
X_2	2	4	3	4	8	8	11

(a) Find the estimated regression equation $\hat{Y} = a + b_1X_1 + b_2X_2$

(b) Obtain the *ANOVA* table and test the hypothesis that there is no association between either regressor and the dependent variable. Use the 0.01 level of significance. (I.U., M.Sc. 1995)

(a) The equation of the estimated multiple linear regression is $Y = a + b_1X_1 + b_2X_2$, where a, b_1 and b_2 are the least-squares estimates of α, β_1 and β_2 respectively.

The sums needed to calculate a, b_1 and b_2 are found to be

$$\Sigma X_1 = 23,\ \Sigma X_2 = 40,\ \Sigma X_1^2 = 105,\ \Sigma X_2^2 = 294,\ \Sigma X_1X_2 = 173,$$
$$\Sigma Y = 67,\ \Sigma Y^2 = 815,\ \Sigma YX_1 = 290,\ \Sigma YX_2 = 488 \text{ and } n = 7.$$

Now substituting these sums in the Normal Equations, we get

$$\begin{aligned} 7a + 23b_1 + 40b_2 &= 67 \\ 23a + 105b_1 + 173b_2 &= 290 \\ 40a + 173b_1 + 294b_2 &= 488. \end{aligned}$$

Solving them simultaneously, we obtain

$$a = 0.7379,\ b_1 = 1.0123 \text{ and } b_2 = 0.9638.$$

Hence the desired estimated multiple regression is

$$\hat{Y} = 0.7379 + 1.0123\,X_1 + 0.9638\,X_2$$

(b) To obtain the *ANOVA* table and to test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$, we proceed as below:

(i) We state the null and alternative hypotheses as

H_0: $\beta_1 = \beta_2 = 0$, *i.e.* none of the regressors is significant; and

H_1: At least one of the β_1 and β_2 is non-zero.

(ii) The significance level is set at $\alpha = 0.01$

(iii) The test-statistic to use is

$$F = \frac{MSR}{MSE}$$

which, if H_0 is true, has an F-distribution with $v_1 = 2$, $v_2 = n-3$ *d.f.*, assuming that population is normally distributed.

(iv) Computations. To set up the *ANOVA* table, we find the necessary sums of squares as below:

Total $$SS = \Sigma(Y_i - \bar{Y})^2 = \Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n}$$

$$= 815 - \frac{(67)^2}{7} = 173.7143$$

$$\text{Regression } SS = \sum(\hat{Y}_i - \bar{Y})^2$$
$$= a\sum Y + b_1\sum X_1 Y + b_2\sum X_2 Y - (\sum Y_i)^2/n$$
$$= (0.7379)(67) + (1.0123)(290) + (0.9638)(488) - (67)^2/7$$
$$= 172.0532$$

$$\text{Residual } SS = \text{Total } SS - \text{Regression } SS$$
$$= 173.7143 - 172.0532 = 1.6611$$

The *ANOVA*-table is

Source of Variation	*d.f.*	Sum of Squares	Mean Square	F
Regression	2	172.0532	86.0266	207.14
Error	4	1.6611	0.4153	
Total	6	173.7143	-	

(v) The critical region is $F \geq F_{0.01}(2, 4) = 18.00$

(vi) **Conclusion.** Since the calculated value of $F = 207.14$ falls in the critical region, so we reject our null hypothesis and conclude that there is an association between at least one of the regressors and the dependent variable.

EXERCISES

21.1 In a regression problem, find the mean and variance of

(i) b, the estimate of β,

(ii) a, the estimate of α, and

(iii) $\hat{Y}$, the estimate of the mean value $\mu_{Y.X}$ for a given X value.

21.2 (a) Show that the least squares estimates a and b are unbiased and also find their variances.

(b) What are the properties of the sampling distribution of b, the estimate of β? How is a confidence interval for β constructed?

21.3 Describe how to construct confidence interval for β, the regression co-efficient, illustrating your answer by considering the regression of Y on X from the following data:

$n = 20, \sum X = 40, \sum Y = 60, \sum X^2 = 95, \sum Y^2 = 297, \sum XY = 150.$

Mention what assumptions you have made in your conclusions.

21.4 The heights (X) and weights (Y) of 100 individuals give $\Sigma X_i = 6826$, $\Sigma Y_i = 16440$, $\Sigma X_i^2 = 466540$, $\Sigma Y_i^2 = 2766596$, $\Sigma XY = 1124828$.

(i) Determine the slope for the regression line $Y_i = \alpha + \beta X_i + \varepsilon_i$; notations have usual meanings.

(ii) Assuming normality, find a 95% confidence interval for α and β, the regression parameters. (P.U., M.Sc. 1981)

21.5 The age (X) and systolic blood pressure (Y) of 100 individuals gave the following sums:

$\Sigma X_i = 4421$, $\Sigma Y_i = 12130$, $\Sigma X_i^2 = 208349$, $\Sigma Y_i^2 = 1498976$, $\Sigma X_i Y_i = 542735$.

(a) Compute the regression line which is used to estimate $\mu_{Y.X}$.

(b) Assuming normality,

(i) construct 95% confidence intervals for α and β,

(ii) predict blood pressure (Y) for an age $X=50$ and compute a 90% confidence interval for this estimate. (P.U., M.Sc. 1984)

21.6 From 10 pairs of values (X_i, Y_i), $i = 1, 2, \ldots, 10$, the following quantities are calculated:

$\Sigma X = 311$, $\Sigma Y = 310.1$, $\Sigma XY = 10{,}074$, $\Sigma X^2 = 10100$ and $\Sigma Y^2 = 10{,}055.09$.

Assuming normality,

(i) find 90% confidence intervals for β and $\mu_{Y.X=30}$;

(ii) test the hypothesis that $\beta=1$ against the alternative that $\beta \neq 1$. (I.U., M.Sc. 1990)

21.7 Fit a regression line $Y = a + bX$ to the following data and test the hypothesis $H_0 : \beta = 0$ at 5% level, where β is the population regression coefficient.

$n = 12$, $\Sigma X = 766$, $\Sigma Y = 1700$, $\Sigma X^2 = 49068$, $\Sigma Y^2 = 246100$ and $\Sigma XY = 109380$. (P.U., B.A./B.Sc. 1987)

21.8 Given $n = 20$, $\Sigma X = 40$, $\Sigma Y = 60$, $\Sigma X^2 = 95$, $\Sigma Y^2 = 297$, $\Sigma XY = 150$. Determine the regression of Y on X and test the hypothesis that the two variables are independent. (P.U., B.A/B.Sc. 1993)

21.9 The data given below represent the heights (X) and the weights (Y) of five men. We selected the heights in advance and then observed the weights of a random group of men having the selected heights.

X:	60	62	65	70	72
Y:	130	135	158	170	185

Using 5 per cent significance level and assuming normality,

(i) test the hypothesis $H_0 : \beta = 0$ against $\beta \neq 0$.

(ii) test the hypothesis $H_0 : \beta = 6$ against $\beta \neq 6$.

(iii) predict the weight of an individual who is 66 inches in height. Give a prediction interval. (P.U., M.Sc. 1970)

21.10 Find the linea · regression equation from the following data:

X	65, 50, 55, 65, 55, 70, 65, 70, 55, 70, 50, 55
Y	85, 74, 76, 90, 85, 87, 94, 98, 31, 91, 76, 74

Assuming normality, test the hypothesis

(i) $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$;

(ii) $H_0 : \alpha = 32$ against $H_1 : \alpha \neq 32$;

at the 0.01 level of significance.

21.11 (a) Describe how to construct confidence interval for the difference between two linear regression co-efficients of two regression lines.

(b) The various sums for two sets of data, each of 4 observations are as follows:

Set	ΣX	ΣY	ΣXY	ΣX^2	ΣY^2
A	8	37	76	18	349
B	15	47	179	59	557

(i) Find the estimates of β_1 and β_2, the regression coefficients of two linear regression lines.

(ii) Construct the 95% confidence interval for $\beta_1 - \beta_2$ and test the hypothesis that $\beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$ at the 0.05 level of significance. (M.Sc., I.U., 1990, P.U., 1990)

21.12 Various sums for three sets of data, each of four observations, are as follows:

Set	ΣX	ΣY	ΣXY	ΣX^2	ΣY^2
A	8	37	76	18	349
B	15	47	179	59	557
C	7	26	47	15	174

(i) Determine the estimates of β_1, β_2 and β_3, the regression coefficients of three linear regression lines.

(ii) Test the hypotheses $\beta_1 = \beta_2$; $\beta_2 = \beta_3$ and $\beta_1 = \beta_3$ and state whether the lines can be regarded as parallel.

(P.U., M.Sc. 1987)

21.13 Use the following data to test the hypothesis that the regression is linear at the 0.05 level of significance:

X	2	2	2	3	3	4	5	5	6	6	6
Y	4	3	8	18	22	24	24	18	13	10	16

21.14 A survey of the pocket money received by children in a primary school was made by choosing at random four children of each of the ages 5, 7, 9 and 11 years. The amounts of pocket money received are given below:

Age (years)	Pocket money (Rs.)
5	2, 8, 10, 12
7	9, 13, 14, 16
9	9, 14, 16, 21
11	18, 19, 23, 36

Find the regression equation for predicting pocket money from age and test for linearity of regression. Use a 0.05 level of significance.

21.15 The amounts of a chemical compound Y, which dissolved in 100 grams of water at various temperatures, X were recorded as follows:

X°C	Y (grams)		
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

Find a regression equation and test the hypothesis that the regression is linear at the 0.05 level of significance.

21.16 (a) How will you find a confidence interval for population correlation coefficient?

(b) A correlation co-efficient of 0.2 is obtained from a random sample of 625 pairs of observations. Find the 99% confidence interval for the coefficient of correlation in the population.

21.17 (a) A random sample of size $n=23$, taken from a bivariate normal population, showed a correlation coefficient of 0.59. Compute the 90% confidence interval for ρ.

(P.U., B.A./B.Sc. 1984)

(b) The lengths (X) and breadths (Y) of 243 cuckoo eggs were measured (in mm) with the results:

$$\sum X = 5442.2, \sum X^2 = 122155.04, \sum Y = 4019.6,$$
$$\sum Y^2 = 66588.92 \text{ and } \sum XY = 90113.83.$$

Give a 95% confidence interval for the true correlation coefficient.

21.18 (a) In a class of 25 students we find a correlation co-efficient of 0.731 between the scores on two tests. Find the 95% confidence interval for the value of the correlation coefficient in the population. (I.U., M.Sc. 1995)

(b) A random sample of size 20 from a bivariate normal population showed a correlation coefficient of 0.92. Find a 98% confidence interval for the population correlation coefficient? (I.U., M.Sc. 1993)

21.19 (a) Test the hypothesis that $\rho = 0.7$, if a sample of 50 gave $r=0.6$. (P.U., B.A./B.Sc. 1983)

(b) A value of r of 0.6 is calculated from a random sample of 39 pairs of observations from a bivariate normal population. Is this value of r consistent with the hypothesis that $\rho = 0.4$?

21.20 (a) Two independent samples have 28 and 19 pairs of observations with correlation coefficients 0.55 and 0.75 respectively. Are these values of r consistent with the hypothesis that the samples have been drawn from the same population?

(b) A sample of 67 gave a correlation of 0.72 whereas another sample of 39 gave a correlation of 0.84. Test at the 0.05 level of significance the hypothesis $H_0 : \rho_1 = \rho_2$ against the alternative $H_1 : \rho_1 \neq \rho_2$.

21.21 (a) Explain the procedure for testing a hypothesis that the population correlation co-efficient equals zero.

(b) A random sample of 28 observations gave a correlation co-efficient of 0.45. Test the hypothesis at the 0.05 level of significance that the population correlation co-efficient is zero?

21.22 (a) A sample of 10 pairs of observations yields a correlation coefficient of 0.7. Is it reasonable to suppose that such a value would arise from a population where the coefficient is 0.85?

(b) Is the value of 0.7 itself significant?

(c) Another sample of 12 pairs of observations shows a coefficient of 0.9. Is this likely to be from the same population as the first?

21.23 (a) A random sample of 20 pairs of observations from a bivariate normal population gives a correlation coefficient of 0.55. Using $\alpha = 0.05$, test the hypothesis that the variables in the population are uncorrelated. (P.U., B.A/B.Sc. 1989)

(b) A random sample of 27 pairs of observations from a bivariate normal population gave a coefficient of correlation of 0.30. Is it consistent with the hypothesis that the coefficient of correlation in the population is zero? Use 5% level of significance. (P.U., B.A/B.Sc. 1996)

21.24 (a) A sample of size 12 yielded $r = 0.32$. Test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. Let $\alpha = 0.01$.

(b) A random sample of 19 pairs from a bivariate normal population showed a correlation of 0.65. Is this consistent with the hypothesis that correlation coefficient in the population is zero?

(c) A study was made to determine whether the dividend paid to the holder of a share of stock is positively correlated with the price of a share of stock. A sample of $n = 20$ stocks gave a correlation coefficient of $r = 0.78$. Use the F-test to test the hypothesis of no correlation. Let $\alpha = 0.05$.

21.25 (a) What is the least value of r in a random sample of 38 pairs that is significant (i) at the 0.05 level, (ii) at the 0.01 level? (P.U., B.A/B.Sc 1977)

(b) Find the least value of r in a sample of 27 pairs from a bivariate normal population that is significant at the 5% level. (P.U., B.A/B.Sc. 1935)

(c) How many pairs of observations must be included in a sample in order that an observed correlation co-efficient of value 0.42 shall have a calculated value of t greater than 2.72?

21.26 (a) Describe how to test a hypothesis about the equality of several correlations.

(b) Random samples of 10, 15 and 20 are drawn from a bivariate normal population, yielding $r = 0.3, 0.4, 0.49$ respectively. Form a combined estimate of ρ and test the hypothesis that the correlations are homogeneous.

21.27 Samples of 20, 30, 40 and 50 are drawn from the same parent population, yielding values of r, the sample correlation coefficient of 0.41, 0.60, 0.51, 0.48 respectively. Use these values of r to obtain a combined estimate of the population correlation coefficient. Test the hypothesis that the samples come from the same bivariate population.

21.28 (a) Show that a partial correlation coefficient $r_{12.34} = 0.5$, in a sample of 20 sets of values from a quadrivariate normal population, is significant at the 5% level. (I.U., M.Sc. 1990)

(b) From a random sample of 25 sets of values from a normal population, the calculated value of a partial correlation of order two is 0.48. Is this consistent with the hypothesis that the corresponding partial correlation in the population is zero, at 5% level of significance? (P.U., B.A/B.Sc. 1988)

21.29 (a) Given $n = 20$, $r_{12.34} = 0.51$, test by means of t-test and of F-test, the hypothesis that $\rho_{12.34} = 0$.

(b) From samples of 24, 29, 33, 38, 42, the partial correlation coefficients of order 3 are found to be 0.38, 0.54, 0.60, 0.50, 0.42. Test their homogeneity. (I.U., M.Sc. 1992)

21.30 (a) Given that

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

determine r_{12}, $r_{13.2}$ and $R_{1.23}$.

(b) Test each one of these correlation coefficients for a significance at 5% level. (M.Sc., P.U., 1988, I.U., 1995)

21.31 (a) In a sample of 20 sets of values from a trivariate normal population, $R_{1.23}$ was found to be 0.35. Show that this is not significant of correlation in the population between X_1 and the variables X_2 and X_3. (P.U., B.A/B.Sc., 1990)

(b) In a sample of 25 sets of values from a quadrivariate normal population, $R_{1.234}$ was found to be 0.4. Test the hypothesis at the 0.05 level of significance that the multiple correlation coefficient in the population between X_1 and the variables X_2, X_3 and X_4 is zero.

21.32 (a) Outline the procedure for performing a simple linear regression analysis. Set your results in an analysis of variance table and interpret them.

(b) A gauge is to be calibrated using dead weight. If X represents the standard and Y, the gauge reading, perform a linear regression analysis based on the following results from 10 observations.

$$\bar{X} = 230,\ \bar{Y} = 226,\ \Sigma(X-\bar{X})(Y-\bar{Y}) = 1532$$

$$\Sigma(X-\bar{X})^2 = 1561,\ \Sigma(Y-\bar{Y})^2 = 1539.$$

Test $H_0 : \beta = 1$, using $\alpha = 0.01$. (P.U., M.Sc. 1971)

21.33 (a) Write a note on Analysis of Variance for Regression.

(b) The data given below represents the heights (X) and the weights (Y) of five men. We selected the heights in advance and then observed the weights of a random group of men having the selected heights.

X:	60	62	65	70	72
Y:	130	135	158	170	185

Set out the analysis of variance for testing the regression.

21.34 Given that $n = 38$, $\bar{X} = 6$, $\bar{Y} = 42$, $\Sigma(X-\bar{X})^2 = 100$, $\Sigma(Y-\bar{Y})^2 = 10{,}000$, $\Sigma(X-\bar{X})(Y-\bar{Y}) = -800$; answer the following:

(i) Determine $\hat{Y} = b_0 + b_1 X$.

(ii) Partition $\Sigma(Y-\bar{Y})^2$ into two parts, one associated with the slope of the linear regression and the other associated with the deviations about the regression.

(iii) Test $H_0 : \beta_1 = 0$, using $\alpha = 0.05$.

(iv) For the observation ($X=8$, $Y=36$), compute the adjusted value of Y. (P.U., M.Sc. 1987)

(v) Interpret both b_1 and β_1.

21.35 Given the data:

Y	90	72	54	42	30	12
X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2

Test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ at $\alpha = 0.05$ by applying the analysis of variance technique.

21.36 Given the data:

Y_i	12	10	9	13	20
X_{1i}	2	2	3	4	4
X_{2i}	1	1	0	0	3

Obtain the least-squares estimates of the parameters in the multiple regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$ and test the overall significance of the regression coefficients.

(I.U., M.Sc. 1991)

22

The Analysis of Covariance

22.1 INTRODUCTION

Occasionally, when an experiment is being carried out, there is an uncontrollable variable, in addition to the main variable (variable of interest) in terms of which we wish to study the effects of classification or treatments. The uncontrollable variable that runs along with the main variable and is suspected to influence its values, is called a *concomitant variable* (or *covariate*), and is denoted by X, while the main variable is represented by Y. The data thus consist of pairs of values of the two variables (X, Y), one pair for each unit. The statistical technique applied to analyse the effect of classification or treatment on the main variable Y, after removing the effect of the concomitant variable by regression method, is called *Analysis of Covariance*. Hence, the analysis of covariance is a mixture of the regression analysis and analysis of variance. Like analysis of variance, this technique partitions the total covariation of a bivariate sample, measured by the sum of the products of the deviations of the variables from their means into component parts associated with specific sources of variation. The covariance effects are thus sorted out and a test of the existence of regression is made. Finding evidence of any regression, its effect is removed before tests are made on the significance of the treatments. *R.A.* Fisher (1890–1962) has expressed that the analysis of covariance "*combines the advantages and reconciles the requirement of the two widely applicable procedures known as regression and analysis of variance*".

To illustrate how the analysis of variance is adjusted to provide for the removal of the effects of concomitant variable X, let us consider the simple linear regression model $Y=\alpha+\beta X+\varepsilon$. Then the estimated regression line is $Y = a + bX$, where a and b are the least squares estimates of α and β. As the regression line goes through the means, we have $\bar{Y} = a + b\bar{X}$. Subtracting, we get

$$\hat{Y} - \bar{Y} = b(X - \bar{X})$$

or $\hat{y} = bx$

This implies that $e = y - \hat{y} = y - bx$

Now the sum of the squares of the errors of estimate is

$$\sum e^2 = \sum(y - bx)^2 = \sum y^2 - 2b\sum xy + b^2\sum x^2$$

$$= \sum y^2 - 2\left(\frac{\sum xy}{\sum x^2}\right)\sum xy + \left(\frac{\sum xy}{\sum x^2}\right)^2 \sum x^2 \quad \left(\because b = \frac{\sum xy}{\sum x^2}\right)$$

$$= \sum y^2 - 2\frac{(\sum xy)^2}{\sum x^2} + \frac{(\sum xy)^2}{\sum x^2} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$$

The term $\frac{(\sum xy)^2}{\sum x^2}$, which is actually the sum of squares due to the regression with 1 degree of freedom, is the amount of reduction in the sum of squares of the Y variable due to its linear regression on X. Thus, to remove the effect of the regression, the term $\frac{(\sum xy)^2}{\sum x^2}$ is to be subtracted from the sum of squares for the variable Y and the result is the *corrected* or *adjusted* sum of squares for Y. The procedure after correction for the regression is that of analysis of variance.

22.2 ONE-WAY ANALYSIS OF COVARIANCE AND PARTITIONING THE SUM OF PRODUCTS

Suppose that a random sample of n pairs of values of the two variables X and Y, is taken from a bivariate normal population and that the paired observations are classified into k groups, each group representing some factor of classification or treatment. Let (X_{ij}, Y_{ij}) be the ith pair of observations in the jth group or treatment $(j = 1, 2, \ldots, k)$ and let the number of pairs in the jth group be n_j with $\sum n_j = n$. The number of pairs in the different groups need not be equal. Then the data may be arranged as shown in the following table:

	Groups or Treatments								
	1		...	j		...	k		
	X_{11}	Y_{11}		X_{1j}	Y_{1j}		X_{1k}	Y_{1k}	
	X_{21}	Y_{21}		X_{2j}	Y_{2j}		X_{2k}	Y_{2k}	
	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	
	X_{i1}	Y_{i1}		X_{ij}	Y_{ij}		X_{ik}	Y_{ik}	
	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	
	$X_{n_1 1}$	$Y_{n_1 1}$		$X_{n_j j}$	$Y_{n_j j}$		$X_{n_k k}$	$Y_{n_k k}$	
Total	$T_{\cdot 1}$	$T'_{\cdot 1}$	...	$T_{\cdot j}$	$T'_{\cdot j}$	...	$T_{\cdot k}$	$T'_{\cdot k}$	$T_{\cdot\cdot}, T'_{\cdot\cdot}$
Mean	$\bar{X}_{\cdot 1}$	$\bar{Y}_{\cdot 1}$	...	$\bar{X}_{\cdot j}$	$\bar{Y}_{\cdot j}$	...	$\bar{X}_{\cdot k}$	$\bar{Y}_{\cdot k}$	$\bar{X}_{\cdot\cdot}, \bar{Y}_{\cdot\cdot}$

As before, let $\bar{X}_{.j}$ and $\bar{Y}_{.j}$ denote the means of the two variables in the jth group or treatment and $\bar{X}_{..}$ and $\bar{Y}_{..}$ be their grand means.

We construct the following identities:

$$X_{ij} - \bar{X}_{..} = (X_{ij} - \bar{X}_{.j}) + (\bar{X}_{.j} - \bar{X}_{..}),$$

$$Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{.j}) + (\bar{Y}_{.j} - \bar{Y}_{..}).$$

We have already shown that

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{..})^2 = \sum_j \sum_i (X_{ij} - \bar{X}_{.j})^2 + \sum_j n_j (\bar{X}_{.j} - \bar{X}_{..})^2$$

Similarly,

$$\sum_j \sum_i (Y_{ij} - \bar{Y}_{..})^2 = \sum_j \sum_i (Y_{ij} - \bar{Y}_{.j})^2 + \sum_j n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

The identity for sum of products is

$$(X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = (X_{ij} - \bar{X}_{.j})(Y_{ij} - \bar{Y}_{.j}) + (\bar{X}_{.j} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..})$$
$$+ (X_{ij} - \bar{X}_{.j})(\bar{Y}_{.j} - \bar{Y}_{..}) + (\bar{X}_{.j} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{.j})$$

Summing over all pairs of observations, we get

$$\sum_j \sum_i (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = \sum_j \sum_i (X_{ij} - \bar{X}_{.j})(Y_{ij} - \bar{Y}_{.j}) + \sum_j n_j (\bar{X}_{.j} - \bar{X}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})$$

The other two terms drop out as they are found to be zero when summed over i and j. Thus the analysis of the sum of products when the paired observations are classified into k groups, is analogous to the analysis of the sum of squares of X or Y for k groups or treatments.

Hence in problems involving k groups of paired observations of X and Y, there will be three analyses, each with $n-1$ degrees of freedom, viz.,

(i) analysis of sum of squares of X;

(ii) analysis of sum of squares of Y; and

(iii) analysis of sum of products of X and Y.

It is customary to denote the **Total SS**, the **Between or Treatment SS** and the **Within or Error SS** by the terms

$$S_{xx} = T_{xx} + E_{xx} \text{ and}$$

$$S_{yy} = T_{yy} + E_{yy} \text{ for } X \text{ and } Y \text{ respectively.}$$

Similarly, we have

$$S_{xy} = T_{xy} + E_{xy} \text{ for products.}$$

Assuming that the factor being investigated has no effect, and that the population from which the random sample of r pairs of observations

is taken, has the covariance μ_{11}, we find that the expected value of total sum of products, *i.e.*, S_{xy}, is $(n-1)\,\mu_{11}$. As the observations in the jth group may be regarded as a random sample of n_j pairs, the expected value of the sum of products in the jth group is given by

$$E\,[\sum_i (X_{ij} - \bar{X}_{.j})\,(Y_{ij} - \bar{Y}_{.j})] = (n_j - 1)\,\mu_{11}, \text{ so that}$$

$$E\,[\sum_j \sum_i (X_{ij} - \bar{X}_{.j})\,(Y_{ij} - \bar{Y}_{.j})] = \sum_j (n_j - 1)\,\mu_{11} = (n-k)\,\mu_{11}$$

We also find that

$$E\,[\sum_j n_j\,(\bar{X}_{.j} - \bar{X}_{..})\,(\bar{Y}_{.j} - \bar{Y}_{..})] = (k-1)\,\mu_{11}$$

Hence each of the terms S_{xy}, E_{xy} and T_{xy}, when divided by $n-1$, $n-k$ and $k-1$ degrees of freedom respectively, gives an unbiased estimate of the population covariance if our assumption is correct. These results can be summarised in the following table:

Source of Variation	d.f.	Sum of Squares & Products			Co-efficient of Regression
		$\sum x^2$	$\sum y^2$	$\sum xy$	
Treatment (Between groups)	$k-1$	T_{xx}	T_{yy}	T_{xy}	--
Error (Within)	$n-k$	E_{xx}	E_{yy}	E_{xy}	$b = \dfrac{E_{xy}}{E_{xx}}$
Total	$n-1$	S_{xx}	S_{yy}	S_{xy}	$b' = S_{xy}/S_{xx}$

We may calculate an estimate of the regression co-efficient, a regression line and partition the sum of squares due to **Error** and **Total**. For example, the error sum of squares is partitioned into a sum of squares due to regression of E_{xy}^2/E_{xx} with 1 degree of freedom and a sum of squares due to deviations about the regression line of $E_{yy}-E_{xy}^2/E_{xx}$ with $(n-k-1)$ degrees of freedom. The corresponding sums of squares for the **Total SS** term are S_{xy}^2/S_{xx} with 1 degree of freedom and $S_{yy}-S_{xy}^2/S_{xx}$ with $(n-2)$ degrees of freedom.

Our next problem is to find out if there is evidence of any regression between the variables X and Y. For this purpose, we compare the regression SS, *i.e.*, E_{xy}^2/E_{xx} having 1 degree of freedom with the term "$E_{yy}-E_{xy}^2/E_{xx}$" having $n-k-1$ degree of freedom by applying the F-statistic. If the regression proves to be *significant*, we construct another table containing sums of squares which have been corrected for the effect of regression, and then proceed to test the significance of the difference

between **group or treatment means.** Since the adjusted sum of squares for the **Between group or treatment means** is obtained by subtraction, we arrange the work in a tabular form as below:

Source of Variation	Regression		Adjusted Sum of Squares		
	d.f.	*SS*	*d.f.*	*SS*	Mean Square
Total	1	$\frac{S_{xy}^2}{S_{xx}}$	$n-2$	$S'_{yy}=S_{yy}-\frac{S_{xy}^2}{S_{xx}}$	---
Error	1	$\frac{E_{xy}^2}{E_{xx}}$	$n-k-1$	$E'_{yy}=E_{yy}-\frac{E_{xy}^2}{E_{xx}}$	$s_e^2=\frac{E'_{yy}}{n-k-1}$
Treatment	(By subtraction)		$k-1$	$T'_{yy}=S'_{yy}-E'_{yy}$	$s_t^2=T'_{yy}/k-1$

The quantity s_e^2 is an estimate of the variance of Y in the population after correction for regression. Similarly, s_t^2 is another estimate of the variance of Y. Hence the statistic $F = \frac{s_t^2}{s_e^2}$ has an F-distribution with $v_1=k-1$, $v_2=n-k-1$ degrees of freedom. We reject the hypothesis that there is no significant difference between the treatment or group means after adjusting for the effect of regression, if the computed value of F exceeds $F_\alpha(k-1, n-k-1)$. To interpret the results properly, a table of adjusted treatment means is constructed.

On the contrary, if the regression does not prove to be significant, the conventional analysis of variance may be carried out ignoring the values of the concomitant variable X.

The multiple-comparisons test (e.g. the *LSD* test) may be applied to the adjusted treatment means—means which are free from regression factor and which are obtained by subtracting the adjustment factor $b_{yx}(\bar{X}_{.j}-\bar{X}_{..})$ from the unadjusted means $\bar{Y}_{.j}$. That is, the corrected means are obtained from the equation

$$adj\ \bar{Y}_{.j} = \bar{Y}_{.j} - b_{yx}(\bar{X}_{.j} - \bar{X}_{..}),\ j = 1, 2, ..., k.$$

To compare two adjusted means such as $adj\ \bar{Y}_{.1}$ and $adj\ \bar{Y}_{.2}$, we find the estimated variance of a difference between two adjusted means by the formula

$$s^2 = s_e^2\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{X}_{.1} - \bar{X}_{.2})^2}{E_{xx}}\right]$$

$$= 2s_e^2\left[\frac{1}{r} + \frac{(\bar{X}_{.1} - \bar{X}_{.2})^2}{E_{xx}}\right] \text{ when } n_1 = n_2 = r$$

22.2.1. Alternative Computing Formulas for Sum of Products. We can simplify the computations of the sums of products. For example, the sum of products of the deviations of n pairs of observations (Y_i, X_i), $i = 1, 2, \ldots, n$ from their means $(\bar{X}, \bar{Y})$ is given by

$$\sum_i (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_i (X_iY_i - \bar{X}Y_i - X_i\bar{Y} + \bar{X}\bar{Y})$$

$$= \sum_i X_iY_i - n\bar{X}\bar{Y} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y}$$

$$= \sum_i X_iY_i - n\bar{X}\bar{Y}$$

$$= \sum_i X_iY_i - \frac{TT'}{n}$$

where $T = \sum_i X_i$ and $T' = \sum_i Y_i$.

Applying this result to the sum of products $\sum_i\sum_j (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..})$, we get

$$S_{xy} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = \sum_j\sum_i X_{ij}Y_{ij} - \frac{T_{..}T_{..}'}{n},$$

where $T_{..} = \sum_j\sum_i X_{ij}$, $T_{..}' = \sum_j\sum_i Y_{ij}$ and $n = \sum_j n_j$

For **Within or Error** sum of products, on summing first over the observations in the jth group, we have

$$\sum_j\sum_i (X_{ij} - \bar{X}_{.j})(Y_{ij} - \bar{Y}_{.j}) = \sum_j \left[\sum_i X_{ij}Y_{ij} - \frac{T_{.j}T_{.j}'}{n_j}\right]$$

where $T_{.j} = \sum_i X_{ij} = n_j\bar{X}_{.j}$ and $T_{.j}' = \sum_i Y_{ij} = n_j\bar{Y}_{.j}$ are the sums of the observations of X and Y in the jth group.

Next summing over the values of j, we get

$$E_{xy} = \sum_j\sum_i (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = \sum_j\sum_i X_{ij}Y_{ij} - \sum_j \frac{T_{.j}T_{.j}'}{n_j}.$$

The sum of products for Between **Treatments or Groups**, *i.e.*, T_{xy}, can be obtained by subtraction. The arithmetic can further be simplified by replacing larger numbers by smaller ones (*i.e.*, deviations) as the three sums of products are unchanged by a change of the origins of X and Y.

Example 22.1. If A, B and C are the three methods of teaching, and X and Y stand for original spelling performance and later spelling performance of each of the four students allocated to systems A, B and C as tabulated below, set up the table of analysis of covariance. What conclusions could be drawn from the table of analysis of covariance set up by you?

A		B		C	
X	Y	X	Y	X	Y
3	10	4	12	1	6
2	8	3	12	2	5
1	8	3	10	3	8
2	11	5	13	1	7

(P.U., D.St. 1964, M.A., 1969)

The necessary computations are shown below:

	A			B			C			
	X	Y	XY	X	Y	XY	X	Y	XY	
	3	10	30	4	12	48	1	6	6	
	2	8	16	3	12	36	2	5	10	
	1	8	8	3	10	30	3	8	24	
	2	11	22	5	13	65	1	7	7	$T_{..}, T'_{..}$
Total	8	37	76	15	47	179	7	26	47	30, 110
$T_{.j}^2, T'^2_{.j}$	64	1369		225	2209		49	676		

$$\sum_i \sum_j X_{ij}^2 = (3)^2 + (2)^2 + \ldots + (1)^2 = 92$$

$$\sum_i \sum_j Y_{ij}^2 = (10)^2 + (8)^2 + \ldots + (7)^2 = 1080$$

$$\sum_i \sum_j X_{ij} Y_{ij} = (3)(10) + (2)(8) + \ldots + (1)(7) = 302$$

$$\sum_j T_{.j}^2 = 64 + 225 + 49 = 338$$

$$\sum_j T'^2_{.j} = 1369 + 2209 + 676 = 4254$$

$$\sum_j (T_{.j})(T'_{.j}) = (8 \times 37) + (15 \times 47) + (7 \times 26) = 1183$$

The sums of squares and products are computed below:

Source	$\sum x^2$	$\sum y^2$	$\sum xy$
1. Correction factor	$\frac{(30)^2}{12} = 75$	$\frac{(110)^2}{12} = 1008.33$	$\frac{(30)(110)}{12} = 75$
2. Total SS	$\sum_i \sum_j X_{ij}^2 - C.F.$ $= 92 - 75$ $= 17\ (= S_{xx})$	$\sum_i \sum_j Y_{ij}^2 - C.F.$ $= 1080 - 1008.33$ $= 71.67\ (= S_{yy})$	$\sum_i \sum_j X_{ij} Y_{ij} - C.F.$ $= 302 - 275$ $= 27\ (= S_{xy})$
3. Treatment SS	$\sum_j \frac{T_{.j}^2}{r} - C.F.$ $= \frac{338}{4} - 75$ $= 9.50\ (= T_{xx})$	$\sum_j \frac{T_{.j}^2}{r} - C.F.$ $= \frac{4254}{4} - 1008.33$ $= 55.17\ (T_{yy})$	$\sum_j \frac{T_{.j}\, T'_{.j}}{r} - C.F.$ $= \frac{1183}{4} - C.F.$ $= 20.75\ (= T_{xy})$
4. Error SS	By Subtraction $(= E_{xx})$	By Subtraction $(= E_{yy})$	By Subtraction $(= E_{xy})$

These results are shown in the following table:

Source of Variation	$d.f.$	Sum of Squares and Products			Regression Co-efficient
		$\sum x^2$	$\sum y^2$	$\sum xy$	
Treatment	2	9.50	55.17	20.75	----
Error	9	7.50	16.50	6.25	$b = 0.83$
Total	11	17.00	71.67	27.00	$b' = 1.59$

Now we set up another table in which the sum of squares have been corrected for the effect of regression.

Source of Variation	Regression		Adjusted Sum of Squares		Mean Square	Computed F
	$d.f.$	SS	$d.f.$	SS		
Total	1	42.88	10	71.67−42.88 = 28.79	--	--
Error	1	5.21	8	16.50−5.21 = 11.29	$S_e^2 = 1.41$	--
Treatment	By Subtraction		2	28.79−11.29 = 17.50	$S_t^2 = 8.75$	$F = \frac{8.75}{1.41} = 6.2$

We reject our null hypothesis that there is no difference in the three methods of teaching, if $F > F_\alpha$; (2, 8). Choosing $\alpha = 0.05$, we see that the

computed value of F exceeds $F_{.05}$; (2, 8) = 4.46. Hence we reject our null hypothesis and conclude that the later spelling performance corrected for differences in original spelling performance, varies significantly from method to method.

If we do not wish to consider the original spelling performance, *i.e.*, the X-values, then significance of differences in the means of later spelling performance would be tested by the F-statistic computed as

$$F = \frac{55.17}{2} \times \frac{9}{16.50} = \frac{27.58}{1.83} = 15.07.$$

This large value of F could be ascribed to the differences in the X-values. The regression line (error) for the method A would be

$$\hat{Y}_{\cdot A} = \bar{X}_{\cdot A} + b(X_{iA} - \bar{X}_A)$$

$$= \frac{37}{4} + 0.83\,(X_{iA} - 2)$$

$$= 9.25 + 0.83\,(X_{iA} - 2)$$

In order to test the significance of the regression co-efficient, b, we compute the value of the F-statistic by the formula

$$F = \frac{\text{Regression } MSE}{\text{Deviation about Regression } MSE}$$

22.3 TWO-WAY ANALYSIS OF COVARIANCE (with no interaction).

Let a random sample of n pairs of observations (X_{ij}, Y_{ij}) from a homogeneous population, be classified according to two factors of classificiation, say, Treatments and Blocks. Assuming one pair of observation per cell, we may arrange the data in a rectangular arrary having c-columns and r-rows so that $n = rc$. Let the columns represent Treatments and Rows, Blocks. Then the pair (X_{ij}, Y_{ij}) belonging to the ith block and the jth treatment will appear in the ith row and the jth column.

For partitioning the **sums of squares** for the two variables X and Y, we have the identies

$$(X_{ij} - \bar{X}_{..}) = (\bar{X}_{i\cdot} - \bar{X}_{..}) + (\bar{X}_{\cdot j} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..}), \text{ and}$$

$$(Y_{ij} - \bar{Y}_{..}) = (\bar{Y}_{i\cdot} - \bar{Y}_{..}) + (\bar{Y}_{\cdot j} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..}).$$

Squaring both sides and summing over all values, we have

$$\sum_i\sum_j (X_{ij} - \bar{X}_{..})^2 = c\sum_i (\bar{X}_{i\cdot} - \bar{X}_{..})^2 + r\sum_j (\bar{X}_{\cdot j} - \bar{X}_{..})^2 + \sum\sum (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})^2, \text{ and}$$

$$\sum_i\sum_j (Y_{ij} - \bar{Y}_{..})^2 = c\sum_i(\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 + r\sum_j (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2 + \sum_i\sum_j (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})^2$$

where all the symbols have their usual significance.

These results may be symbolically written as

$$S_{xx} = B_{xx} + T_{xx} + E_{xx}, \text{ and}$$

$$S_{yy} = B_{yy} + T_{yy} + E_{yy},$$

where B stands for **Blocks** and T for **Treatments**, etc.

The partitioning of the sum of products is obtainined by multiplying the two-identities for the variables X and Y and by summing over all values. Thus we get

$$\sum_i\sum_j (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = c\sum_i(\bar{X}_{i\cdot} - \bar{X}_{..})(\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 + r\sum_i(\bar{X}_{\cdot j} - \bar{X}_{..})(\bar{Y}_{\cdot j} - \bar{Y}_{..}) + \sum_i\sum_j (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})$$

i.e., $S_{xy} = B_{xy} + T_{xy} + E_{xy}$

Assuming that the two factors of classification have no influence on the results, all the terms in the above identity when divided by the appropriate number of degrees of freedom, provide unbiased estimates of the population covariance μ_{11}. The sums of squares and products as partitioned above, may be summaried as follows:

Source of Variation	Degrees of freedom	Sum of Squares and Products			Regression Co-efficient
		$\sum x^2$	$\sum y^2$	$\sum xy$	
Treatments	$c-1$	T_{xx}	T_{yy}	T_{xy}	----
Blocks	$r-1$	B_{xx}	B_{yy}	B_{xy}	--
Error	$(c-1)(r-1)$	E_{xx}	E_{yy}	E_{xy}	$b = \frac{E_{xy}}{E_{xx}}$
Total	$rc-1$	S_{xx}	S_{yy}	S_{xy}	--

As before, we must first find out if there is evidence of any regression between the variables X and Y. For this purpose, the regression co-efficient, b, is obtained from error sums and its significance is found by comparing the **Regression SS**, $\frac{E_{xy}^2}{E_{xx}}$ having 1 degree of freedom with the **Deviations about Regression SS**, $\left(i.e., E_{yy} - \frac{E_{xy}^2}{E_{xx}}\right)$

having $(cr-c-r)$ degrees of freedom by using the F-statistic. If the regression proves to be significant, the sums of squares must be corrected of the effect of regression before proceeding to test the significance of the effects due to the factors of classification. For example, the corrected sum of squares for Treatments are found by considering a reduced form of the above table as shown below:

Source of Variation	Regression		Adjusted Sum of Squares	
	d.f.	SS	d.f.	SS
Total $(T+E)$	1	$\frac{(T_{xy}+E_{xy})^2}{T_{xx}+E_{xx}}$	$r(c-1)-1$	$(T+E)'_{yy} = (T_{yy}+E_{yy}) - \frac{(T_{xy}+E_{xy})^2}{T_{xx}+E_{xx}}$
Error (E)	1	$\frac{(E_{xy})^2}{E_{xx}}$	$(r-1)(c-1)-1$	$E'_{yy} = E_{yy} - \frac{E_{xy}^2}{E_{xx}}$
Treatment (T)	By Subtraction		$c-1$	$T'_{yy} = (T+E)'_{yy} - E'_{yy}$

The Error and Treatment Mean Squares are

$$s_e^2 = \frac{E'_{yy}}{(r-1)(c-1)-1} \quad \text{and} \quad s_t^2 = \frac{T'_{yy}}{c-1}$$

When our assumptions hold, the statistic $F = \frac{s_t^2}{s_e^2}$, has an F-distribution with $v_1 = c-1$, $v_2 = (r-1)(c-1) - 1$ degrees of freedom. Hence we reject the hypothesis that the treatment effects are zero, if F exceeds $F_\alpha; (v_1, v_2)$.

In a similar way, we can build up a test for the effect of blocks or the second factor of classification.

Example 22.2. Perform the analysis of covariance upon the data given below:

Blocks	Treatments					
	A		B		C	
	X	Y	X	Y	X	Y
1	5	17	6	23	4	29
2	15	16	8	16	10	25
3	12	12	15	18	15	24

Compute tests of significance and adjusted treatment means. (I.U., M.Sc. 1986)

For anlaysis of covariance, the necessary computations are given below:

Blocks	Treatments									Total	
	A			B			C			$B_{i\cdot}$	$B_{i\cdot}'$
	X	Y	XY	X	Y	XY	X	Y	XY		
1	5	17	85	6	23	138	4	29	116	15	69
2	15	16	240	8	16	128	10	25	250	33	57
3	12	12	144	15	18	270	15	24	360	42	54
$T_{\cdot j}, T_{\cdot j}'$	32	45	469	29	57	536	29	78	726	90	180
$T_{\cdot j}^2, T_{\cdot j}'^2$	1024	2025	--	841	3249	--	841	6084	--	2706	11358

(i) **Total SS and Products:**

$$S_{xx} = \sum_i \sum_j X_{ij}^2 - \frac{(T_{..})^2}{n} = (5)^2 + (15)^2 + \ldots + (15)^2 - \frac{(90)^2}{9}$$

$$= 1060 - 900 = 160.$$

$$S_{yy} = \sum_i \sum_j Y_{ij}^2 - \frac{(T_{..}')^2}{n} = (17)^2 + (16)^2 + \ldots + (24)^2 - \frac{(180)^2}{9}$$

$$= 3840 - 3600 = 240, \text{ and}$$

$$S_{xy} = \sum_i \sum_j X_{ij} Y_{ij} - \frac{(T_{..})(T_{..}')}{n} = (469 + 536 + 726) - \frac{(90)\,(180)}{9}$$

$$= 1731 - 1800 = -69$$

(ii) **Treatment SS and Products:**

$$T_{xx} = \sum_j \frac{T_{\cdot j}^2}{r} - \frac{(T_{..})^2}{n} = \frac{2706}{3} - 900 = 902 - 900 = 2,$$

$$T_{yy} = \sum_j \frac{T_{\cdot j}'^2}{r} - \frac{(T_{..}')^2}{n} = \frac{11358}{3} - 3600 = 3786 - 3600 = 186, \text{ and}$$

$$T_{xy} = \sum_j \frac{(T_{\cdot j})\,(T_{\cdot j}')}{r} - \frac{(T_{..})\,(T_{..}')}{n}$$

$$= \frac{(32 \times 45) + (29 \times 57) + (29 \times 78)}{3} - \frac{(90 \times 180)}{9}$$

$$= \frac{5355}{3} - 1800 = 1785 - 1800 = -15.$$

The analysis of covariance table is then set up as below:

Source of Variation	d.f.	Sum of Squares and Products			Regression Co-efficient
		Σx^2	Σy^2	Σxy	
Block	2	2/9	13	−1.67	
Treatment	2	2/9	54	0	
Error	4	4/9	151	−8.00	$b = \frac{-8.00}{4/9} = -18$
Total	8	8/9	218	−9.67	

Hence the estimate of the missing observation $= -b = 18$.

23.4.7. Efficiency of a RCB Design Relative to a CR Design. An experimental design is to be more *efficient* than another if it attains the same precision with less expenditure of time or money, and its relative efficiency is usually expressed in terms of a ratio of error variances. This sort of expression is considered straightforward and practical as it can be directly interpreted in terms of the amount of replication required by the design to attain a given precision. As such, the experimental units may be subjected to the same or dummy treatments on account of the fact that we are concerned only with errors. This implies that the treatments may be ignored. The analysis of variance for the *RCB* design with k treatments and r replications, is

Source of Variation	d.f.	Sum of Squares	Mean Square
Treatments	$k-1$	$(k-1)\,s_t^2$	s_t^2
Blocks	$r-1$	$(r-1)\,s_b^2$	s_b^2
Error	$(r-1)(k-1)$	$(r-1)(k-1)\,s_e^2$	s_e^2
Total	$rk-1$	--	--

For efficiency, we ignore treatments and get

(i) *With Blocks:*

Source of Variation	d.f.	SS	MS
Blocks	$(r-1)$	$(r-1)\,s_b^2$	s_b^2
Error	$r(k-1)$	$r(k-1)\,s_e^2$	s_e^2
Total	$rk-1$	--	--

Substituting these values, we get

$$LSD = (3.182)\sqrt{\frac{2(0.625)}{3}}, \qquad (\because t_{0.025,(3)} = 3.182)$$

$$= (3.182)(0.645) = 2.05.$$

To test the significance of the differences between the adjusted treatment means, we arrange them in ascending order of magnitude and draw a line under any pair of adjacent means that do not differ significantly as

A	B	C
14.62	19.19	26.19

It is observed that all the three adjusted means differ significantly as the difference between any pair exceeds the LSD.

22.4 ANALYSIS OF COVARIANCE MODELS. ONE-WAY CLASSIFICATION

Let Y_{ij} denote the ith observation of the jth treatment Y, and X_{ij} denote the corresponding concomitant variable, on which Y_{ij} has a linear regression with β as regression co-efficient. Then the *fixed effects model* for the analysis of covariance is

$$Y_{ij} = \mu + \tau_j + \beta(X_{ij} - \bar{X}_{..}) + e_{ij}, \qquad \begin{pmatrix} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots, k \end{pmatrix}$$

where μ is the true mean effect, τ_j is the effect of the jth treatment, $\bar{X}_{..}$ is the grand mean of the concomitant variable and e_{ij} is the random error assumed to be normally and independently distributed with zero mean and common variance.

The least-squares estimates are obtained by minimizing $Q = \sum_i \sum_j e_{ij}^2$ subject to the restriction that $\sum \tau_j = 0$. Then

$$Q = \sum_i \sum_j [Y_{ij} - \mu - \tau_j - \beta(X_{ij} - \bar{X}_{..})]^2$$

Differentiating Q *w.r.t.* μ and equating to zero, we have

$$\frac{\partial Q}{\partial \mu} = 0 = -2 \sum_i \sum_j [Y_{ij} - \mu - \tau_j - \beta(X_{ij} - \bar{X}_{..})]$$

This gives that $\hat{\mu} = \bar{Y}_{..}$

Again differentiating Q *w.r.t.* τ_j and equating to zero, we get

$$\frac{\partial Q}{\partial \tau_j} = 0 = -2 \sum_i [Y_{ij} - \mu - \tau_j - \beta(X_{ij} - \bar{X}_{..})]$$

Simplifying, we get

$$\hat{\tau}_j = \bar{Y}_{.j} - \bar{Y}_{..} - \beta(\bar{X}_{.j} - \bar{X}_{..})$$

which shows that each treatment mean should be adjusted for the regression on X.

Similarly, $\dfrac{\partial Q}{\partial \beta} = 0 = -2 \sum_i \sum_j (X_{ij} - \bar{X}_{..}) [Y_{ij} - \mu - \tau_j - \beta(X_{ij} - \bar{X}_{..})]$

Substituting for μ and τ_j and simplifying, we have

$$\hat{\beta} = \frac{\sum_i \sum_j (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) - \sum_i \sum_j (\bar{X}_{.j} - \bar{X}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})}{\sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 - \sum_i \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2}$$

$$= \frac{\sum_i \sum_j (X_{ij} - \bar{X}_{.j})(Y_{ij} - \bar{Y}_{.j})}{\sum_i \sum_j (X_{ij} - \bar{X}_{.j})^2}$$

$$= \frac{\text{Residual or Error cross-product of } XY}{\text{Residual or Error sum of squares of } X} = \frac{E_{xy}}{E_{xx}}$$

These results can be presented in an analysis of covariance table as before. After adjustment, the Error SS $= E_{yy} - \dfrac{E_{xy}^2}{E_{xx}}$, and

$$\textbf{Error MS} = \frac{E_{yy} - E_{xy}^2 / E_{xx}}{n - k - 1} \qquad (n = rk)$$

We wish to test the hypothesis that there are no differences among the true effects of the k treatments, *i.e.*, $H_0 : \tau_j = 0$.

In other words, our hypothesis is the comparison of

(i) $Y_{ij} = \mu + \tau_j + \beta(X_{ij} - \bar{X}_{..}) + e_{ij}$, and

(ii) $Y_{ij} = \mu + \beta(X_{ij} - \bar{X}_{..}) + e_{ij}'$

Thus we estimate the **Treatment SS** as the difference between the two **Error Sums of Squares** for the two models. From (i), we have

$$\textbf{Error SS} = E_{yy} - \frac{E_{xy}^2}{E_{xx}},$$ and from (ii), we have

$$\textbf{Error SS} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = (T_{yy} + E_{yy}) - \frac{(T_{xy} + E_{xy})^2}{T_{xx} + E_{xx}},$$ so that

Treatment SS = Error SS (*ii*) − Error SS (*i*)

Hence $F = \dfrac{\text{Treatment } MS}{\text{Error } MS}$ has an F-distribution with $v_1 = k-1$, $v_2 = n-k-1$ degrees of freedom when our hypothesis is true.

If F proves to be significant, we can apply the multiple comparisons test to look for significant comparisons. Suppose there are two treatments, then the difference between them is given by

$$\hat{\tau}_1 - \hat{\tau}_2 = \bar{Y}_{.1} - \bar{Y}_{.2} - \beta(\bar{X}_{.1} - \bar{X}_{.2}), \text{ and}$$

$$\text{Var}\,(\hat{\tau}_1 - \hat{\tau}_2) = s_e^2 \left[\frac{2}{r} + \frac{(\bar{X}_{.1} - \bar{X}_{.2})^2}{E_{xx}}\right]$$

Hence all pairs will give different variances.

In case of two-way analysis of covariance, the *fixed effects model* is

$$Y_{ij} = \mu + B_i + \tau_j + \beta(X_{ij} - \bar{X}_{..}) + e_{ij}, \qquad \left[\begin{array}{l} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots, c \end{array}\right.$$

where the letters have their usual significance. The least-squares estimates are found by minimizing $\sum_i \sum_j e_{ij}^2$ subject to the restrictions that $\sum_i B_i = 0 = \sum_j \tau_j$.

The model may be written as

$$Y_{ij} - \beta(X_{ij} - \bar{X}_{..}) = \mu + B_i + \tau_j + e_{ij}$$

or $$Z_{ij} = \mu + B_i + \tau_j + e_{ij}, \quad \text{where } Z_{ij} = Y_{ij} - \beta(X_{ij} - \bar{X}_{..})$$

The terms on the right hand side of this equation give the standard model for a two-way analysis of variance, while the terms on the left hand side are the deviations of Y_{ij} from its linear regression on X_{ij}. Thus the analysis of covariance is a mixture of the regression analysis and the analysis of variance.

22.4.1. Assumptions Made in Analysis of Covariance. The assumptions made in the analysis of covariance are similar to those made in the analysis of variance and regression analysis as the analysis of covariance is a mixture of these two techniques. However, the assumptions necessary for carrying out the covariance analysis are given as follows:

(i) The populations are normally distributed with homogeneous variances.

(ii) The samples are random and independent.

(iii) The regression is linear and the slope is not zero.

(iv) The treatment, block and regression effects are additive.

(v) The concomitant variable is fixed, implying that it is not being affected by the treatments.

(vi) The residuals e_{ij} are normally and independently distributed with zero mean and a common variance.

22.4.2. Uses of Covariance Analysis. The main uses of the technique of covariance analysis are briefly enumerated below:

(i) It increases the precision of the experiment by removing certain environmental effects that cannot be controlled by the experimental design.

(ii) The covariance adjustments remove the bias due to regression. The nature of the bias is the term $\beta(\bar{X}_{ij} - \bar{X}_{..})$.

(iii) As it partitions the total covariance or the sum of cross-products into component parts, it is therefore used for testing a regression co-efficient or the homogeneity of k linear regression co-efficients.

(iv) It assists in analysing the treatment effects properly when a concomitant variable is playing its part in producing the treatment effects.

(v) It can be profitably used to estimate missing observations by setting $Y=0$ for each missing value and introducing dummy covariate X in such a way that $X=1$ corresponds to the missing values and $X=0$ to all others.

EXERCISES

22.1 Define a covariance analysis. Discuss the appropriate model and the assumptions involved. (P.U., B.A Hons. Part III, 1969)

22.2 Describe the purpose of an analysis of covariance and outline the main stages in its calculation.

22.3 What are the uses and advantages of covariance analysis? Illustrate your answer by examples. What factors are considered in interpreting the results of the analysis? (P.U., M.Sc. 1970)

22.4 (a) Describe a sampling experiment to illustrate the analysis of covariance technique, What tests of significance are made in a covariance analysis?

(b) How treatment means are adjusted in covariance analysis?

22.5 When is the use of covariance analysis appropriate? Write down the analysis of covariance table for a two-way classification with h treatments laid out in r blocks, and indicate tests that you would carry out. (P.U., M.Sc. 1972)

22.6 Give the analysis of variance for a two-way classification, discussing the appropriate model and the assumptions involved.

If a measurement of a concomitant variable is available for each observation, explain how the analysis would be modified, discussing any further assumptions.

22.7 (a) What are the assumptions behind a covariance analysis?

(b) In the process of analyzing data by a covariance analysis, what tests of significance are made?

(c) Explain the interpretation or inferences and the course of action indicated when each of the above test is signficant; when each is nonsignificant.

22.8 Carry out an analysis of covariance with the data given below:

A		B	
X	Y	X	Y
8	17	9	19
7	15	8	20
6	15	8	17
7	18	10	20

22.9 An experiment on gain in weight of rats resulted as shown in the table, X indicates the quantity of food and Y, the gain in weight. Did the four rations A, B, C, D produce different gains among the rats? Are the gains affected materially by quantity of food?

A		B		C		D	
X	Y	X	Y	X	Y	X	Y
96	98	109	64	179	71	127	72
108	102	125	86	132	84	100	54
94	102	85	51	163	71	151	109
128	108	32	72	143	62	116	93

(P.U., B.A/B.Sc. Hons. Part II, 1964; M.A., 1960)

22.10 Using the following table, calculate an analysis of covariance of X on T. Test the significance of the regression and compare mean values (if necessary, adjusted) of X for each group.

Group							
1		2		3		4	
X	T	X	T	X	T	X	T
3.2	4.0	2.4	4.7	2.6	3.7	2.0	3.4
3.3	3.6	3.0	6.4	3.5	6.8	3.8	9.4
2.3	1.3	2.2	3.8	4.1	8.8	2.8	6.3
4.4	9.2	2.5	4.9	2.5	4.5	2.3	4.7
		3.6	8.7			2.5	5.1
		1.2	2.0				

(P.U., M.Sc., 1971)

22.11 Three groups of 5 students each were given an initial test on admission and the scores obtained are given as X. In a subsequent test, the scores obtained were given as Y. Perform the analysis of covariance to test the significance of differences between subsequent scores.

Group					
I		II		III	
X	Y	X	Y	X	Y
14	10	11	5	7	5
9	6	9	2	6	4
11	8	8	6	2	1
12	6	10	5	10	7
10	9	10	4	7	9

22.12 (a) It is desired to compare the IQ's of students in three schools. It is suspected that IQ is related to the students' grade-point average. If samples of 12 students are taken from each school, should their IQ values be adjusted for their grade-point averages by an analysis of covariance, or should an analysis of variance of their IQ's be performed directly?

(P.U., M.Sc. 1972)

(b) Following is the table of sums of squares and products:

Source of Variation	$d.f.$	Σx^2	Σy^2	Σxy
Treatment	9	4684	112	−532
Error	27	3317	516	−650

Complete the anlaysis, making appropriate tests to indicate the reason for your conclusions.

(P.U., B.Sc. Hons. Part I, 1972; M.Sc. 1970)

22.13 Each of 4 blocks was divided into 3 plots, and 3 different treatmnets A, B, C were distributed at random among the plots of each block. The rows correspond to blocks and the yields of grain and straw are denoted by X and Y respectively. Examine the covariation between the yields of grain and straw to show whether the yield of straw after correction for yield of grain varies significantly with treatment.

Blocks	Treatments					
	A		B		C	
	X	Y	X	Y	X	Y
1	65	32	75	38	72	33
2	68	26	54	20	69	30
3	71	26	71	28	69	38
4	62	33	64	29	61	30

22.14 What are the basic principles of anlaysis of covariance? Set up analysis of covariance table, indicating the nature of questions and manner in which those questions can be answered for the data on two variables X and Y presented in the table:

Blocks	Treatmnents							
	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
1	28	21	22	17	27	18	34	25
2	24	21	22	16	26	20	21	15
3	20	16	18	16	19	14	21	16
4	25	21	23	17	31	19	27	21

(P.U., M.A., 1960)

22.15 For an experiment, the results are summarised by the following sum of squares and products:

Source of Variation	d.f.	Σx^2	Σy^2	Σxy
Blocks	5	7472.6	6.31	−111.65
Treatments	6	116020.3	112.86	3598.05
Error	30	28665.1	23.23	682.20
Total	41	152158.0	142.40	4168.60

(a) Based on the error sum of squares and products, is the regression of Y on X significant at $\alpha = 0.05$?

(b) Are the differences among the treatment means for Y adjusted for variation attributed to X, significant at $\alpha = 0.05$?

(c) What conclusions do you draw from the above data about the effects of treatments? Make any additional computations that you consider necessary.

❖❖❖❖❖❖❖❖❖❖❖❖

23

Experimental Designs

23.1 INTRODUCTION

By an *experimental design*, we mean a **plan** used to collect the data relevant to the problem under study in such a way as to provide a basis for valid and objective inference about the stated problem. The plan usually consists of the selection of treatments whose effects are to be studied, the specification of the experimental layouts, the assignment of treatments to the experimental units and the collection of observations for analysis. All these steps are accomplished before any experiment is performed.

An experiment is planned to

(i) get maximum information for minimum expenditure in the minimum possible time;

(ii) avoid systematic errors;

(iii) evaluate the outcomes critically and logically;

(iv) ignore spurious effects, if any.

The following considerations go into the planning of an experiment:

(i) What is the experiment intended to do?

(ii) What is the nature of the treatments or dependent variables and how are they to be estimated?

(iii) How is the independent variable likely to affect the treatments or dependent variables?

(iv) Are the factors to be held constant or varied? If varied, whether the variation is quantitative or qualitative?

The answers to these questions enable the experimenter to state his hypothesis precisely and to plan his experimental procedure in a more effective way.

There are two types of designs; *systematic* and *random designs*; but the analysis of variance techniques are suitable to randomized designs only. The basic randomized designs are (i) Completely Randomized, (ii) Randomized Complete Blocks, and (iii) Latin Squares, which we discuss in the sections that follow.

23.2 BASIC PRINCIPLES OF EXPERIMENTAL DESIGNS

The basic principles of experimental designs are *randomization*, *replication* and *local control*. These principles make a valid test of significance possible. Each of them is described briefly in the following subsections.

23.2.1. Randomization. The first principle of an experimental design is *randomization*, which is a random process of assigning treatments to the experimental units. The random process implies that every possible allotment of treatments has the same probability. An *experimental unit* is the smallest division of the experimental material and a *treatment* means an experimental condition whose effect is to be measured and compared. The purpose of randomization is to remove bias and other sources of extraneous variation, which are not controllable. Another advantage of randomization (accompanied by replication) is that it forms the basis of any valid statistical test. Hence the treatments must be assigned *at random* to the experimental units. Randomization is usually done by drawing numbered cards from a well-shuffled pack of cards, or by drawing numbered balls from a well-shaken container or by using tables of random numbers.

23.2.2. Replication. The second principle of an experimental design is *replication*; which is a repetition of the basic experiment. In other words, it is a complete run for all the treatments to be tested in the experiment. In all experiments, some variation is introduced because of the fact that the experimental units such as individuals or plots of land in agricultural experiments, cannot be physically identical. This type of variation can be removed by using a number of experimental units. We therefore perform the experiment more than once, *i.e.*, we repeat the basic experiment. An individual repetition is called a *replicate*. The number, the shape and the size of replicates depend upon the nature of the experimental material. A replication is used

(i) to secure more accurate estimate of the *experimental error*, a term which represents the differences that would be observed if the same treatments were applied several times to the same experimental units;

(ii) to decrease the experimental error and thereby to increase *precision*, which is a measure of the variability of the experimental error; and

(iii) to obtain more precise estimate of the mean effect of a treatment, since $\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$, where n denotes the number of replications.

23.2.3. Local Control. It has been observed that all extraneous sources of variation are not removed by randomization and replication. This necessitates a refinement in the experimental technique. In other words, we need to choose a design in such a manner that all extraneous sources of variation are brought under control. For this purpose, we make use of *local control*, a term referring to the amount of balancing, blocking and grouping of the experimental units. Balancing means that the treatments should be assigned to the experimental units in such a way that the result is a balanced arrangement of the treatments. Blocking means that like experimental units should be collected together to form a relatively homogeneous group. A block is also a replicate. The main purpose of the principle of local control is to increase the efficiency of an experimental design by decreasing the experimental error. The point to remember here is that the term local control should not be confused with the word **control**. The word **control** in experimental design is used for a treatment which does not receive any *treatment* but we need to find out the effectiveness of other treatments through comparison.

23.3 THE COMPLETELY RANDOMIZED DESIGN

A *completely randomized (CR)* design, which is the simplest type of the basic designs, may be defined as a design in which the treatments are assigned to experimental units *completely at random*, that is the randomization is done without any restrictions. The design is completely flexible, *i.e.*, any number of treatments and any number of units per treatment may be used. Moreover, the number of units per treatment need not be equal. A completely randomized design is considered to be most useful in situations where (i) the experimental units are homogeneous, (ii) the experiments are small such as laboratory experiments, and (iii) some experimental units are likely to be destroyed or to fail to respond.

23.3.1. Experimental Layout. The *layout* of an experiment is the actual placement of the treatments on the experimental units, which may pertain to time, space or type of material. Suppose we have t treatments and the experimental material is divided into n experimental

units. We shall then assign the k-treatments *at random* to the n experimental units in such a way that the treatment τ_j ($j = 1, 2, \ldots, k$) is applied r_j times, with $\sum r_j = n$. When each treatment is applied the same number of times, then $r_1 = r_2 = \ldots = r_k = r$ and $\sum r_j = rk = n$. Usually, each treatment is applied (or replicated) an equal number of times.

An example of the experimental layout for a completely randomized design (*CR*) using four treatments *A*, *B*, *C* and *D*, each repeated 3 times, is given below:

C	A	B	D
C	B	C	A
A	D	D	B

The result or response of a treatment which may be a real yield, the gain in weight, the ability, etc., is generally called *yield* and is represented by the letter Y.

23.3.2. Statistical Model and Analysis. Let Y_{ij} denote the yield of the ith observation on treatment j. Then Y_{ij} may be represented by the linear model

$$Y_{ij} = \mu + \tau_j + e_{ij}, \quad \begin{cases} j = 1, 2, \ldots, k \\ i = 1, 2, \ldots, r_j \end{cases}$$

where μ represents the true mean effect, τ_j represents the effects of treatment j and e_{ij} denotes the random error, normally and independently distributed with mean zero and variance σ^2. The null hypothesis in the case of *fixed effects model*, may then be stated as

$$H_0 : \tau_j = 0 \text{ for all } j = 1, 2, \ldots, k$$

and the alternative hypothesis as H_1 : some $\tau_j \neq 0$. These hypotheses are equivalent to the following set of hypotheses

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \text{ and}$$

$$H_1 : \text{Not all } \mu \text{ are equal.}$$

For analysis of variance, the least-squares estimates are obtained by minimizing $\sum_j \sum_i e_{ij}^2$, subject to the restriction that $\sum_j \tau_j = 0$. (Remember τ_j are the deviations from μ).

Statistical Analysis. For convenience, the data of yields obtained from a *CR* design can be assembled as follows for statistical analysis:

Observation	Treatments						
	1	2	...	j	...	k	
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1k}	
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2k}	
⋮	⋮	⋮		⋮		⋮	
j	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ik}	
⋮	⋮	⋮		⋮		⋮	
r_j	$Y_{r_1 1}$	$Y_{r_2 2}$		$Y_{r_j j}$		$Y_{r_k k}$	
Total	$T_{\cdot 1}$	$T_{\cdot 2}$	...	$T_{\cdot j}$	...	$T_{\cdot k}$	$T_{\cdot\cdot}$
Means	$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$	...	$\bar{Y}_{\cdot j}$	...	$\bar{Y}_{\cdot k}$	$\bar{Y}_{\cdot\cdot}$

This is equivalent to k independent random samples and hence analogous to one-way classification. Thus the partitioning of the *total sum of squares* into components for treatments and error would be as usual, *i.e.*,

$$\sum_{j=1}^{k}\sum_{i=1}^{r_j}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2 = \sum_{j=1}^{k} r_j(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})^2 + \sum_{j=1}^{k}\sum_{i=1}^{r_j}(Y_{ij}-\bar{Y}_{\cdot j})^2$$

or $$\text{Total } SS = \text{Treatment } SS + \text{Error } SS$$

and the *ANOVA*-Table for a *CR* design would be

Source of Variation	*d.f.*	Sum of Squares	Mean Square
Treatments (τ_j)	$k-1$	$\sum_j r_j(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})^2$	$s_t^2 = MST_r$
Error (e_{ij})	$n-k$	By subtraction	$s_e^2 = MSE$
Total	$n-1$	$\sum_j\sum_i (Y_{ij}-\bar{Y}_{\cdot\cdot})^2$	---

The computations of the sums of squares are carried out as usual.

Now, if our hypothesis *that all treatment effects are zero*, is true and if the assumption of normality, independence, additivity, linearity and homogeneous variances underlying the analysis of variance hold, then the ratio

$$F = \frac{s_t^2}{s_e^2} = \frac{MS \text{ for Treatments}}{MS \text{ for Error}}$$

has an F-distribution with $v_1 = k-1$, $v_2 = n-k$ degrees of freedom.

The hypothesis is rejected at the α level of significance, when $F \geq F_{\alpha}; (v_1, v_2)$.

If the hypothesis is rejected and the difference between treatment effects proves to be significant, then the difference between two treatment means can be tested by

$$t = \frac{\bar{Y}_{\cdot j} - \bar{Y}_{\cdot j'}}{s_e \sqrt{\frac{1}{r_j} + \frac{1}{r_{j'}}}},$$

where the statistic t has Student's t-distribution with $(n-k)$ degrees of freedom.

To answer questions of the type *which treatment is the best, which treatment is the second best*, we use multiple comparisons tests.

23.3.3. Advantages and Disadvantages. The advantages and disadvantages of the completely randomized designs are given below:

(i) The design is very simple and is easily laid out.

(ii) It has the simplest statistical analysis.

(iii) It provides the maximum number of degrees of freedom for error sum of squares.

(iv) The design is flexible, *i.e.*, any number of treatments and of replications may be used.

(v) The design is applicable only to a small number of treatments.

(vi) There is possibility of entering the whole of the variation among the experimental units into the experimental error as the randomization is not restricted in any direction.

Example 23.1. An experiment was conducted to compare the yields of three varieties of potatoes. Each variety was assigned at random to equal-size plots, four times. The yields were as follows:

Variety		
A	B	C
23	18	16
26	28	25
20	17	12
17	21	14

Test the hypothesis that the three varieties of potatoes are not different in the yielding capabilities.

This experiment may be described by the following statistical mdoel:

$$Y_{ij} = \mu + \tau_j + e_{ij},$$

where Y_{ij} represents the yields of the ith observation on treatment (variety) j, and τ_j denotes the effect of the treatment j, etc. Here $j = 1, 2$ and 3 and $i = 1, 2, 3, 4$ for each treatment. To analyse these results, we proceed as below:

(i) We formulate our null hypothesis as

H_0: $\tau_j = 0$ for all $j = 1, 2$ and 3, *i.e.* there is no difference among the yielding capabilities of the three varieties of potatoes:

and the alternative hypothesis as

H_1 : Not all τ_j are equal.

(ii) We choose the significance level at $\alpha = 0.05$

(iii) The test-statistic to be used is

$$F = \frac{MS \text{ for Treatments}}{MS \text{ for Error}}$$

which, if H_0 is true, has an F-distribution with $\upsilon_1 = k - 1$ and $\upsilon_2 = n - k$ degrees of freedom.

(iv) The necessary computations are carried out as below:

	Variety			Total	$\sum\sum Y_{ij}^2$
	A	B	C		
	23 (529)	18 (324)	16 (256)	----	1109
	26 (676)	28 (784)	25 (625)	---	2085
	20 (400)	17 (289)	12 (144)	---	833
	17 (289)	21 (441)	14 (196)	---	926
$T_{.j}$	86	84	67	237	4953
$T_{.j}^2$	7396	7056	4489	18941	↑
$\sum_i\sum_j Y_{ij}^2$	1894	1838	1221	4953	check ←

$$\text{Total } SS = \sum_i\sum_j Y_{ij}^2 - \frac{T_{..}^2}{n},$$

$$= 4953 - \frac{(237)^2}{12} = 4953 - 4680.75 \doteq 272.25$$

$$\text{Treatment } SS = \sum_j \frac{T_{.j}^2}{r} - \frac{T_{..}^2}{n} = \frac{18941}{4} - 4680.75 = 54.50$$

The sum of squares for experimental error is obtained by subtraction. The analysis of variance table is

Source of Variation	*d.f.*	Sum of Squares	Mean Square	Computed *F*
Treatments (τ_j)	2	54.50	27.25	1.13
Error (e_{ij})	9	217.75	24.19	---
Total	11	272.25	---	---

(v) The critical region is $F \geq F_{0.05}(2, 9) = 4.26$

(vi) **Conclusion.** Since the computed value of $F = 1.13$ does not fall in the critical region, so we accept our null hypothesis and may conclude that there is no difference among the yielding capabilities of the three varieties of potatoes.

23.4 THE RANDOMIZED COMPLETE BLOCK DESIGN

A *randomised complete block* (*RCB*) design may be defined as one in which (i) the experimental material is divided into groups or blocks in such a manner that the experimental units within a particular block are relatively homogeneous, (ii) each block contains a complete set of treatments, *i.e.*, it constitutes a replication of treatments, and (iii) the treatments are assigned *at random* to the experimental units within each block, which means the randomization is restricted within blocks. A new randomization is made for every block. The purpose of this type of arrangement is to control a source of variation in the experimental material. The randomized complete block design is perhaps the most frequently used experimental design.

23.4.1. Experimental Layout. Suppose there are k treatments and r blocks in a randomized complete block design, then each block contains k homogeneous plots, one of each treatment. An experimental *layout* for such a design, using 6 treatments A, B, C, D, E and F in 3 blocks might be as follows:

BLOCK I	D	B	A	C	F	E
BLOCK II	C	B	E	F	D	A
BLOCK III	C	F	B	D	A	E

In actual field or laboratory, the treatments would occupy plots or the experimental units in positions corresponding to those shown in the layout, that is the treatments refer to the actual locations in the field or laboratory.

23.4.2. Statistical Model and Analysis. As each observation in a RCB design is classified by the block to which it belongs and by the treatment it receives, therefore Y_{ij} represents the observation corresponding to block i ($i = 1, 2, \ldots, r$) and treatment j ($j = 1, 2, \ldots, k$). The linear statistical model for this design would be

$$Y_{ij} = \mu + B_i + \tau_j + e_{ij}$$

where B_i represents block effect and e_{ij} are assumed to be normally and independently distributed with mean zero and variance σ^2, and where blocks and treatments are orthogonal. The least-squares estimates of the parameters μ, B_i and τ_j are obtained by minimizing the quantity $Q = \sum_j \sum_i e_{ij}^2$ subject to the restriction that $\sum_i B_i = 0$ and $\sum_j \tau_j = 0$.

The analysis of a randomized complete block experiment consists of a two-way analysis of variance test where the hypotheses (i) $H_0 : \tau_j = 0$ for all j, *i.e.*, all treatment effects are zero, and (ii) $H_0' : B_i = 0$ for all i, *i.e.*, all block effects are zero, are tested. To facilitate computations, the data may be arranged in a **Table** as shown below:

Blocks	Treatments						Block	
	1	2	...	j	...	k	Total	Means
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1k}	$B_{1\cdot}$	$\bar{Y}_{1\cdot}$
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2k}	$B_{2\cdot}$	$\bar{Y}_{2\cdot}$
⋮	⋮	⋮		⋮		⋮	⋮	⋮
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ik}	$B_{i\cdot}$	$\bar{Y}_{i\cdot}$
⋮	⋮	⋮		⋮		⋮	⋮	⋮
r	Y_{r1}	Y_{r2}		Y_{rj}		Y_{rk}	$B_{r\cdot}$	$\bar{Y}_{r\cdot}$
Total	$T_{\cdot 1}$	$T_{\cdot 2}$	...	$T_{\cdot j}$	...	$T_{\cdot k}$	$T_{\cdot\cdot}$	--
Means	$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$	...	$\bar{Y}_{\cdot j}$	...	$\bar{Y}_{\cdot k}$	--	$\bar{Y}_{\cdot\cdot}$

The unbiased estimates of the common population variance σ^2 are usually obtained by partitioning the **total sum of squares** as in the case of a two-way table. Accordingly, the partitioning of the total sum of squares into components for treatments, blocks and error would be

$$\sum_{i=1}^{r}\sum_{j=1}^{k}(Y_{ij}-\bar{Y}_{..})^2 = r\sum_{j=1}^{k}(\bar{Y}_{.j}-\bar{Y}_{..})^2 + k\sum_{i=1}^{r}(\bar{Y}_{i.}-\bar{Y}_{..})^2 + \sum_{i=1}^{r}\sum_{j=1}^{k}(Y_{ij}-\bar{Y}_{.j}-\bar{Y}_{i.}+\bar{Y}_{..})^2$$

i.e., Total SS = Treatment SS + Block SS + Error SS,

and the *ANOVA*-Table for a *RCB* design would be

Source of Variation	*d.f.*	Sum of Squares	Mean Square
Treatments (τ_j)	$k-1$	$r\sum_j(\bar{Y}_{.j}-\bar{Y}_{..})^2 = SSTr$	$s_t^2 = \frac{SSTr}{k-1}$
Blocks (B_i)	$r-1$	$k\sum_i(\bar{Y}_{i.}-\bar{Y}_{..})^2 = SSB$	$s_b^2 = \frac{SSB}{r-1}$
Error (e_{ij})	$(k-1)(r-1)$	By subtraction $= SSE$	$s_e^2 = \frac{SSB}{(k-1)(r-1)}$
Total	$rk-1$	$\sum_j\sum_i(Y_{ij}-\bar{Y}_{..})^2 = SST$	---

The computations of the sums of squares are carried out as usual. Now, if our hypothesis $H_0 : \tau_j = 0$ for all j, *i.e.*, all treatment effects are zero, is true, then the statistic

$$F_1 = \frac{s_t^2}{s_e^2} = \frac{MS \text{ for treatments}}{MS \text{ for errors}}$$

has as F-distribution with $v_1 = k-1$ and $v_2 = (k-1)(r-1)$ degrees of freedom. The hypothesis $H_0 : \tau_j = 0$ for all j, is rejected at the α level of significance, when $F_1 \geq F_\alpha; (v_1, v_2)$.

Furthermore if our hypothesis $H_0' : B_i = 0$ for all i, *i.e.*, all block effects are zero, is true, the statistics

$$F_2 = \frac{s_b^2}{s_e^2} = \frac{MS \text{ for blocks}}{MS \text{ for errors}}$$

has an F-distribution with $v_1 = r-1$ and $v_2 = (k-1)(r-1)$ degrees of freedom. The hypothesis $H_0 : B_i = 0$ for all i is rejected at the α level of significance if F_2 exceeds $F_\alpha; (v_1, v_2)$. It is important to point out that if F_2 proves to be insignificant, which means there are no differences among blocks, we have a bad design.

Usually, the blocks are chosen to be different by making the differences between them as large as possible, and hence we find no

reason to test the hypothesis that there are no differences among the blocks.

The difference between two treatment means selected at random can be tested by computing the least-significant-difference (LSD) as

$$LSD = t_{\alpha/2}\sqrt{\frac{2\,(\text{error mean square})}{r}}$$

where r is the number of blocks (replications) and the value of $t_{\alpha/2}$ is found from the table of the Student's t-distribution for error degrees of freedom.

23.4.3. Advantages and Disadvantages. The important advantages of the randomized complete block design are as follows:

(i) The source of extraneous variation is controlled by grouping the experimental material and hence the estimate of the experimental error is decreased.

(ii) The design is flexible, *i.e.*, any number (but not less than 2) of replications may be run and any number of treatments may be tested.

(iii) The experiment can be set up easily.

(iv) The statistical analysis is simple and straight forward.

(v) It is easy to adjust for the missing observations.

However, the design suffers from the following two disadvantages:

(i) It controls variability only in one direction.

(ii) It is not a suitable design when the number of treatments is very large or when the blocks are not homogeneous.

Example 23.2. Four varieties of wheat were tried in a randomized complete block design in four replications. Yield in kilogram per plot is shown in the table given below. Test the hypothesis that there is no difference in the means of four varieties.

Replicates	Varieties			
	V_1	V_2	V_3	V_4
I	2	5	4	1
II	2	3	3	1
III	4	6	6	2
IV	1	4	2	3

(P.U., B.A./B.Sc. 1971)

(i) We state our null hypothesis as

$H_0: \tau_j = 0$ for all $j = 1,2,3,4$, *i.e.*, there is no difference in the means of the four varieties of wheat; and the alternative hypothesis is

H_1 : Not all the four means are equal.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic to be used is

$$F = \frac{s_t^2}{s_e^2} = \frac{MS \text{ for treatments}}{MS \text{ for errors}}$$

which, if H_0 is true, has an F-distribution with $v_1 = k - 1$ and $v_2 = (k-1)(r-1)$ degrees of freedom.

(iv) The necessary computations are shown below:

Replicates	Varieties V_1	V_2	V_3	V_4	$B_{i\cdot}$	$B_{i\cdot}^2$	$\sum\sum Y_{ij}^2$
I	2 (4)	5 (25)	4 (16)	1 (1)	12	144	46
II	2 (4)	3 (9)	3 (9)	1 (1)	9	81	23
III	4 (16)	6 (36)	6 (36)	2 (4)	18	324	92
IV	1 (1)	4 (16)	2 (4)	3 (9)	10	100	30
$T_{\cdot j}$	9	18	15	7	49	649	191
$T_{\cdot j}^2$	81	324	225	49	679		↑
$\sum_i\sum_j Y_{ij}^2$	25	86	65	15	191	←	Check
$\bar{Y}_{\cdot j}$	2.25	4.50	3.75	1.75	--		

$$\text{Total } SS = \sum_i\sum_j Y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{n}, \text{ where } n = rk$$

$$= 191 - \frac{(49)^2}{16} = 191 - 150.06 = 40.94$$

$$\text{Treatment } SS = \sum_j \frac{T_{\cdot j}^2}{r} - \frac{T_{\cdot\cdot}^2}{n},$$

$$= \frac{679}{4} - \frac{(49)^2}{16} = 169.75 - 150.06 = 19.69$$

$$\text{Replicate } SS = \sum_i \frac{B_{i\cdot}^2}{k} - \frac{T_{\cdot\cdot}^2}{n},$$

$$= \frac{649}{4} - \frac{(49)^2}{16} = 162.25 - 150.06 = 12.19$$

The sum of squares for experimental error is obtained by subtraction.

These results are given in the following ANOVA-Table:

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Treatments	3	19.69	6.56	$F_1 = 6.50$
Replicates	3	12.19	4.06	$F_2 = 4.02$
Error	9	9.06	$s_e^2 = 1.01$	---
Total	15	40.94	---	---

(v) The regions of rejection are $F_1, F_2 \geq F_{0.05;\,(3,\,9)} = 3.86$.

(vi) **Conclusion.** Since the computed value of F_1 falls in the critical region, we therefore reject our null hypothesis and conclude that the means of the four varieties of wheat are significantly different.

Furthermore, since F-statistic indicates rejection of the null hypothesis, we therefore apply the LSD test to find out which means differ significantly

$$\text{Now} \quad LSD = t_{0.025,(9)} \times \sqrt{\frac{2s_e^2}{r}}$$

$$= 2.26\sqrt{\frac{2 \times 1.01}{4}} = 2.26 \times 0.71 = 1.60.$$

Arranging the means of the four varieties of wheat in ascending order of magnitude and drawing a line under any subset of adjacent means that do not differ significantly, we get

V_4	V_1	V_3	V_2
1.75	2.25	3.75	4.50

Example 23.3. Following is the plan of a field layout for testing four varieties A, B, C and D of wheat in each of 5 blocks. The plot yields in pounds are also indicated therein.

BLOCK I	BLOCK II	BLOCK III	BLOCK IV	BLOCK V
D 29.3	B 33.0	D 29.8	B 36.8	D 28.8
B 33.3	A 34.0	A 34.3	A 35.0	C 35.8
C 30.8	C 34.3	B 36.3	D 28.0	B 34.5
A 32.3	D 26.0	C 35.3	C 32.3	A 36.5

Perform an analysis of variance and state your conclusions.

(P.U., M.A. 1966)

For statistical analysis, we organise the data as shown in the table below:

Blocks	Treatments (Varieties of wheat)				$B_{i\cdot}$
	A	B	C	D	
I	32.3	33.3	30.8	29.3	125.7
II	34.0	33.0	34.3	26.0	127.3
III	34.3	36.3	35.3	29.8	135.7
IV	35.0	36.8	32.3	28.0	132.1
V	36.5	34.5	35.8	28.8	135.6
$T_{\cdot j}$	172.1	173.9	168.5	141.9	656.4
Means	34.42	34.78	33.70	28.38	32.82

Computations for *ANOVA*:

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{n}$$

$$= (32.3)^2 + (34.0)^2 + \ldots + (28.8)^2 - \frac{(656.4)^2}{20}$$

$$= 21725.22 - 21543.05 = 182.17$$

$$\text{Treatment } SS = \sum_j \frac{T_{\cdot j}^2}{r} - \frac{T_{\cdot\cdot}^2}{n}$$

$$= \frac{(172.1)^2 + \ldots + (141.9)^2}{5} - \frac{(656.4)^2}{20}$$

$$= 21677.50 - 21543.05 = 134.45$$

$$\text{Block } SS = \sum_i \frac{B_{i\cdot}^2}{k} - \frac{T_{\cdot\cdot}^2}{n},$$

$$= \frac{(125.7)^2 + \ldots + (135.6)^2}{4} - \frac{(656.4)^2}{20}$$

$$= 21564.51 - 21543.05 = 21.46$$

$$\text{Error } SS = \text{Total } SS - (\text{Treatment } SS + \text{Block } SS)$$

$$= 182.17 - (134.45 + 21.46) = 26.26$$

Hence the *ANOVA*-Table is

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Treatments (τ_j)	3	134.45	44.82	$F_1 = 20.47$
Blocks (B_i)	4	21.46	5.36	$F_2 = 2.45$
Error (e_{ij})	12	26.26	2.19	---
Total	19	182.17	---	---

To test $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4$ against H_1 : Not all τ_j are equal, we find that $F_1 = 20.47$, which exceeds $F_{0.05}$; (3, 12) = 3.49. We therefore reject the null hypothesis and conclude that there is a considerable difference among the means of the four varieties of wheat.

23.4.4. Randomized Complete Block Design with Replications Within Blocks. Let us assume that each treatment is replicated n times within each block, so that each block contains nk plots. The experiment becomes analogous to two-way classification with each cell containing n observations. This leads to a new statistical model in the form

$$Y_{ijl} = \mu + B_i + \tau_j + (B\tau)_{ij} + e_{ijl} \qquad \begin{array}{l} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots, k \\ l = 1, 2, \ldots, n \end{array}$$

where $(B\tau)_{ij}$ denotes the interaction between the ith block and the jth treatment with the restrictions $\sum_{i=1}^{r} (B\tau)_{ij} = 0 = \sum_{j=1}^{k} (B\tau)_{ij}$; and other symbols have their usual significance.

The partitioning of the **total sum of squares** in this case is given as follows:

$$\sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n} (Y_{ijl} - \bar{Y}_{\ldots})^2 = rn \sum_{j} (\bar{Y}_{.j.} - \bar{Y}_{\ldots})^2 \quad \rightarrow \text{Treatment SS}$$

$$+ kn \sum_{i} (\bar{Y}_{i..} - \bar{Y}_{\ldots})^2 \quad \rightarrow \text{Block SS}$$

$$+ n \sum_{i}\sum_{j} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{\ldots})^2 \quad \rightarrow \text{Interaction SS}$$

$$+ \sum_{i}\sum_{j}\sum_{l} (Y_{ijl} - \bar{Y}_{ij.})^2 \quad \rightarrow \text{Error SS}$$

The total number of degrees of freedom is partitioned as below:

Source of Variation	Degrees of freedom
Treatments (τ_j)	$k-1$
Blocks (B_i)	$r-1$
Interaction $(B\tau)_{ij}$	$(k-1)(r-1)$
Error (e_{ijl})	By subtraction
Total	$nrk-1$

The rest of the computations and analysis are carried out in the usual manner.

23.4.5. Missing Observations in RCB Design. Sometimes it so happens that some experimental units or observations are missing. For example, plants may be destroyed or uprooted, records may be lost, flasks may break, animals may die or human beings may not cooperate. These omissions, being beyond the control of the experimenter, destroy the orthogonality and the balance of the design. In such a case, it is possible to analyse the non-orthogonal data as a multiple regression, but this procedure is difficult. The simple method, as suggested by Yates, is to estimate the value of the missing observation so that a complete set of data is obtained for carrying out the usual analysis of variance. The missing observation is estimated by minimizing the error sum of squares. It has been further shown that the validity of the analysis of variance of the augmented data is not disturbed if the proportion of the missing observations is not large and if the error degrees of freedom are decreased by the number of missing observations computed.

Suppose for convenience, the observation for treatment 1 in block 1 in a *RCB* design with k treatments and r block, is missing and x is put in place of this missing observation. The data may be arranged as below:

Blocks	Treatments				Block Total
	1	2	...	k	
1	x	Y_{12}	...	Y_{1k}	$B'_{1\cdot}+x$
2	Y_{21}	Y_{22}	...	Y_{2k}	$B_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
r	Y_{r1}	Y_{r2}	...	Y_{rk}	$B_{r\cdot}$
Treatment Total	$T'_{\cdot 1}+x$,	$T_{\cdot 2}$	...	$T_{\cdot k}$	$T'_{\cdot\cdot}+x$

Performing the analysis of variance, we get the following sums of squares:

Source of Variation	Sum of Squares
Treatments	$\frac{(T'_{.1}+x)^2}{r} + \frac{T^2_{.2}}{r} + \dots + \frac{T^2_{.k}}{r} - \frac{(T'_{..}+x)^2}{kr}$
Blocks	$\frac{(B'_{1.}+x)^2}{k} + \frac{B^2_{2.}}{k} + \dots + \frac{B^2_{r.}}{k} + \frac{(T'_{..}+x)^2}{kr}$
Error	By subtraction
Total	$\sum_i \sum_j Y^2_{ij} + x^2 - \frac{(T'_{..}+x)^2}{kr}$

Thus the error sum of squares is obtained as

$$SSE = x^2 + \frac{(T'_{..}+x)^2}{kr} - \frac{(T'_{.1}+x)^2}{r} - \frac{(B'_{1.}+x)^2}{k} + \text{terms which do not}$$

involve x. To minimize SSE, we differentiate it *w.r.t.* x, equate it to zero and solve for x. We therefore get

$$\frac{\partial(SSE)}{\partial x} = 0 = 2x + \frac{2(T'_{..}+x)}{kr} - \frac{2(T'_{.1}+x)}{r} - \frac{2(B'_{1.}+x)}{k}$$

or $$\left(1 + \frac{1}{kr} - \frac{1}{r} - \frac{1}{k}\right)x + \frac{T'_{..}}{kr} - \frac{T'_{.1}}{r} - \frac{B'_{1.}}{k} = 0$$

or $$\left(\frac{kr+1-k-r}{kr}\right)x = \frac{rB'_{1.} + kT'_{.1} - T'_{..}}{kr}$$

or $$x = \frac{rB'_{1.} + kT'_{.1} - T'_{..}}{(k-1)(r-1)}$$

The formula in general may be written as

$$x = \frac{rB + kT - G}{(k-1)(r-1)}$$

where r = number of blocks or replicates,

k = number of treatments,

B = sum of the remaining values in the blocks with the missing observation,

T = sum of the remaining values of the treatment with missing observation, and

G = grand total of all the observed values.

The analysis of variance is then completed and the test of significance is carried out.

If two or more observations are missing, we use the following iterative procedure proposed by Yates:

(i) Assign guess-estimates such as the mean of the actually observed values, to all but one missing observations.

(ii) Estimate this one missing observation, using the formula for one missing observation.

(iii) Estimate then one of the other missing observations, using the estimate obtained in (ii) and the remaining previously assigned values.

(iv) Estimate the remaining observations and complete the cycle.

(v) Repeat the procedure using the estimates.

(vi) Continue the procedure until the estimates become stabilized.

(vii) Subtract one degree of freedom for each missing observation from the total and error sum of squares.

The analysis of variance is then carried out as usual, though this procedure is not strictly correct. The estimated $S.E.$ of the difference between the mean of a treatment with a missing observation and the mean of any other treatment is $\sqrt{s_e^2 \left[\frac{2}{r} + \frac{k}{r(r-1)(k-1)}\right]}$.

Example 23.4. The following data, obtained from a randomized complete block design with 3 treatments A, B and C, and 3 blocks contain one missing observation represented by x.

Blocks	Treatments		
	A	B	C
I	5	12	15
II	7	10	14
III	8	16	x

Estimate the missing observation and prepare a table for analysis of variance.

We use the formula

$$x = \frac{rB + kT - G}{(k-1)(r-1)}$$

to estimate the missing observation.

Here B = sum of the values in block III (i.e., block containing missing observation)

$= 8 + 16 = 24$;

T = sum of the values of treatment C (i.e., treatment containing missing observation).

$= 15 + 14 = 29$; and

$G = 5 + 7 + \ldots + 15 + 14 = 87$

Substituting in the formula, we get the estimate as

$$x = \frac{3(24) + 3(29) - 87}{(3-1)(3-1)} = 18.$$

Putting 18 in place of x, we carry out the necessary computations as below:

Blocks	Treatments			
	A	B	C	$B_{i\cdot}$
I	5	12	15	32
II	7	10	14	31
III	8	16	18	42
$T_{\cdot j}$	20	38	47	105

$$\text{Total } SS = \sum_i \sum_j Y_{ij}^2 - \frac{T_{..}^2}{n}$$

$$= (5)^2 + (7)^2 + \ldots + (14)^2 + (18)^2 - \frac{(105)^2}{9}$$

$$= 1383 - 1225 = 158$$

$$\text{Treatment } SS = \sum_j \frac{T_{\cdot j}^2}{r} - \frac{T_{..}^2}{n},$$

$$= \frac{(20)^2 + (38)^2 + (47)^2}{3} - \frac{(105)^2}{9} = 1351 - 1225 = 126$$

Block $SS = \sum_i \frac{B_{i.}^2}{k} - \frac{T_{..}^2}{n}$,

$$= \frac{(32)^2+(31)^2+(42)^2}{3} - \frac{(105)^2}{9}$$

$$= 1249.7 - 1225 = 24.7$$

Error SS = Total SS − (Treatment SS + Block SS)

$$= 158 - (124 + 24.7) = 7.3$$

Hence the *ANOVA*-Table is

Source of Variation	*d.f.*	Sum of Squares	Mean Square	Computed F
Treatments	2	126.0	63.0	25.9
Blocks	2	24.7	12.35	
Error	3	7.3	2.43	
Total	7	158.0	---	

It is to be noted that total and error degrees of freedom have been decreased by 1.

Now, we can test $H_0 : \tau_j = 0$ for all j in the customary manner.

23.4.6. Estimation of Missing Observation by Covariance. The analysis of covariance technique can also be used to estimate the missing observation. The procedure followed is described below:

(i) Insert zero for the missing value, *i.e.*, set $y = 0$.

(ii) Introduce a dummy covariate consisting of a one associated with the missing observation and zeros for all other observations.

(iii) Carry out the covariance analysis.

(iv) The estimate of the missing observation is given by

$$y_0 - bx_0 = -b = -\frac{E_{xy}}{E_{xx}},$$

where $y_0 = 0$, $x_0 = 1$ and E_{xy} and E_{xx} are the error sum of products and the error sum of squares obtained from the covariance analysis table.

In the case of two or more missing values, a covariate is introduced for each missing value and a multiple covariance is computed.

Example 23.5. Estimate the missing value (y_0) from the data given below by applying the covariance technique:

Blocks	Treatments		
	A	B	C
I	5	12	15
II	7	10	14
III	8	16	y_0

In order to carry out the analysis of covariance, we insert zero for the missing observation and pair the observations with the covariate composed of zeros and a one as shown in the table.

Blocks	Treatments								
	A		B		C				
	X	Y	X	Y	X	Y	$B_{i\cdot}$	$B_{i\cdot}'$	$B_{i\cdot}'^2$
I	0	5	0	12	0	15	0	32	1024
II	0	7	0	10	0	14	0	31	961
III	0	8	0	16	1	0	1	24	576
$T_{\cdot j}, T_{\cdot j}'$	0	20	0	38	1	29	1	87	2561
$T_{\cdot j}'^2$	400		1444		841		2685		

The sum of squares and products are computed below:

Source of Variation	Σx^2	Σy^2	Σxy
1. *C.F.*	$\frac{(1)^2}{9} = \frac{1}{9}$	$\frac{(87)^2}{9} = 841$	$\frac{1 \times 87}{9} = 9.67$
2. Total	$1 - \frac{1}{9} = \frac{8}{9}$	$1059 - 841 = 218$	$0 - 9.67 = -9.67$
3. Treatment	$\frac{1}{3} - \frac{1}{9} = \frac{2}{9}$	$\frac{2685}{3} - 841 = 54$	$\frac{1 \times 29}{3} - 9.67 = 0$
4. Block	$\frac{1}{3} - \frac{1}{9} = \frac{2}{9}$	$\frac{2561}{3} - 841 = 13$	$\frac{24}{3} - 9.67 = -1.67$
5. Error	(By diff.) $= \frac{4}{9}$	$= 151$	$= -8.00$

The analysis of covariance table is then set up as below:

Source of Variation	d.f.	Sum of Squares and Products			Regression Co-efficient
		Σx^2	Σy^2	Σxy	
Block	2	2/9	13	−1.67	
Treatment	2	2/9	54	0	
Error	4	4/9	151	−8.00	$b = \frac{-8.00}{4/9} = -18$
Total	8	8/9	218	−9.67	

Hence the estimate of the missing observation $= -b = 18$.

23.4.7. Efficiency of a RCB Design Relative to a CR Design. An experimental design is to be more *efficient* than another if it attains the same precision with less expenditure of time or money, and its relative efficiency is usually expressed in terms of a ratio of error variances. This sort of expression is considered straightforward and practical as it can be directly interpreted in terms of the amount of replication required by the design to attain a given precision. As such, the experimental units may be subjected to the same or dummy treatments on account of the fact that we are concerned only with errors. This implies that the treatments may be ignored. The analysis of variance for the *RCB* design with k treatments and r replications, is

Source of Variation	d.f.	Sum of Squares	Mean Square
Treatments	$k-1$	$(k-1)\,s_t^2$	s_t^2
Blocks	$r-1$	$(r-1)\,s_b^2$	s_b^2
Error	$(r-1)(k-1)$	$(r-1)(k-1)\,s_e^2$	s_e^2
Total	$rk-1$	--	--

For efficiency, we ignore treatments and get

(i) *With Blocks*:

Source of Variation	d.f.	SS	MS
Blocks	$(r-1)$	$(r-1)\,s_b^2$	s_b^2
Error	$r(k-1)$	$r(k-1)\,s_e^2$	s_e^2
Total	$rk-1$	--	--

(ii) *Without Blocks:*

Error SS without blocks $= (r-1)\, s_b^2 + r(k-1)\, s_e^2$ with $(rk-1)$ d.f.

$$\therefore \text{ Estimate of Error Variance} = \frac{(r-1)\, s_b^2 + r(k-1)\, s_e^2}{rk-1}$$

Hence the required *efficiency* = ratio of error variances

$$= \frac{(r-1)\, s_b^2 + r(k-1)\, s_e^2}{(rk-1)\, s_e^2}$$

$$= \frac{\text{estimated } MSE \text{ for a } CR \text{ design}}{MSE \text{ for the } RCB \text{ design}}$$

In this discussion, we have assumed that in either of the designs, exactly the same experimental units or plots would have been used.

As a example, the relative efficiency (RE) of the RCB design given in Example 23.3 as compared to a CR design is calculated as

$$RE = \frac{(5-1)(5.36) + (5)(4-1)(2.19)}{(5 \times 4 - 1)(2.19)}$$

$$= \frac{21.44 + 32.85}{41.61} = \frac{54.29}{41.61} = 1.30 \text{ or } 130\%$$

This indicates that the RCB design is 130% as efficient as the completely randomized design.

23.5 THE LATIN SQUARE DESIGN

The experimental error in RCB design is reduced by controlling the source of extraneous variation in one direction, *i.e.*, by grouping the experimental units in one way. When the variation is found in two directions, it becomes necessary to remove these two sources of variation simultaneously. This end is achieved by simultaneous blocking of experimental units in two mutually perpendicular directions called, *rows* and *columns*. Since each row and each column is a complete block, the grouping for a balanced arrangement is performed by imposing the restriction that each treatment must appear once and only once in each row and once and only once in each column. If there are k treatments, the experimental area will be divided into k-rows and k-columns, resulting in k^2 plots or experimental units, as the experiment is laid out in a *square pattern*. The treatments are then assigned *at random* to plots or experimental units. Such a double blocking of experimental units and a corresponding doubly restricted random assignments is called a **Latin**

Square (*LS*) design, following Euler who used Latin letters for symbols or *treatments*.

Hence a **Latin Square** (*LS*) design is an arrangement of k treatments in a $k \times k$ square, where the treatments are grouped in blocks in two directions, the directions being orthogonal to each other and to the treatments, and where the treatments appear once and only once in each direction. It should be noted that in a Latin Square design, the number of rows, the number of columns and the number of treatments must all be equal.

23.5.1. Construction and Layout. Latin Squares are always constructed by rotation, *e.g.* in the case of 4 treatments *A*, *B*, *C* and *D*, we get

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

A large number of distinct Latin squares can be derived by interchanging rows and columns of a *standard Latin square*, where a standard Latin square is a Latin Square in which the treatments in the first row and in the first column are arranged in alphabetical or numerical order. A standard Latin square is also called a *reduced* Latin square. There is only one standard square for a 2×2 Latin square, viz.,

A	B
B	A

Similarly, for a 3×3 Latin square, there is only one standard square

A	B	C
B	C	A
C	A	B

But for a 4×4 Latin square, there are four possible standard squares as shown below:

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

The possible standard forms of a 5×5 Latin square are 56 and so on. Each standard square of size $k \times k$ yields $k!\,(k-1)!$ different squares by permuting the rows except the first and all the k columns. Accordingly, we may obtain from a square of size

(i) 2×2, 2 different squares;

(ii) 3×3, 12 different squares;

(iii) 4×4, 576 different square;

(iv) 5×5, 161, 280 different square; etc.

In designing a Latin square, one square is to be selected *at random* from all the possible forms of the $k \times k$ Latin squares and thereafter, the rows and the columns are randomized for the purpose of field layout. Thus the experimental layout for the field having 5 treatments A, B, C, D and E, might be as follows:

	COLUMNS				
	1	2	3	4	5
1	$\boxed{B}$	$\boxed{E}$	$\boxed{A}$	$\boxed{C}$	$\boxed{D}$
2	$\boxed{D}$	$\boxed{A}$	$\boxed{E}$	$\boxed{B}$	$\boxed{C}$
ROWS 3	$\boxed{E}$	$\boxed{B}$	$\boxed{C}$	$\boxed{D}$	$\boxed{A}$
4	$\boxed{A}$	$\boxed{C}$	$\boxed{D}$	$\boxed{E}$	$\boxed{B}$
5	$\boxed{C}$	$\boxed{D}$	$\boxed{B}$	$\boxed{A}$	$\boxed{E}$

It is conceptually desirable that a Latin square design should occupy a square experimental area, but in certain situations, the design may be laid out otherwise, if the purpose of the *LS* design, *i.e.*, control of variation in two directions, is achieved.

23.5.2. Statistical Model and Analysis. The data obtained from a $k \times k$ Latin Square design may be arranged in tabular form as below:

	COLUMNS							
	1	2	3	4	...	k	Total	Mean
1	Y_{11A}	Y_{12B}	Y_{13C}	Y_{14D}		...	$R_{1\cdot\cdot}$	$\bar{Y}_{1\cdot\cdot}$
2	.			.			.	.
Rows 3								
4	Y_{41D}	Y_{42A}	Y_{43B}	Y_{44C}		...	$R_{4\cdot\cdot}$	$\bar{Y}_{4\cdot\cdot}$
⋮	⋮			⋮		⋮	⋮	
k	.			.		.	$R_{k\cdot\cdot}$	$\bar{Y}_{k\cdot\cdot}$
Total	$C_{\cdot 1\cdot}$	$C_{\cdot 2\cdot}$	$C_{\cdot 3\cdot}$	$C_{\cdot 4\cdot}$		...	$(G = T\ldots)$	
Mean	$\bar{Y}_{\cdot 1\cdot}$	$\bar{Y}_{\cdot 2\cdot}$	$\bar{Y}_{\cdot 3\cdot}$	$\bar{Y}_{\cdot 4\cdot}$		...	.	$\bar{Y}_{\ldots}$

The treatment means may be represented by $\bar{Y}_{..A}, \bar{Y}_{..B}, \bar{Y}_{..C}, \bar{Y}_{..D}$, etc.

The yield $Y_{ij(h)}$ of the $ij(h)$th observation in a LS design with one observation per plot or experimental unit, may be represented by the linear statistical model

$$Y_{ij(h)} = \mu + R_i + C_j + \tau_h + e_{ij(h)} \qquad i, j, h = 1, 2, \ldots, k,$$

where μ, R_i, C_j, τ_h and $e_{ij(h)}$ denote the general mean, the effect of the ith row, the effect of the jth column, the effect of the hth treatment and experimental error respectively; and where $e_{ij(h)}$ are assumed to be normally and independently distributed with mean zero and common variance σ^2. In this model, the interaction terms are assumed to be zero, and the subscript h being not independent of i and j, is placed in brackets. To test the hypotheses that

(i) all treatment-effects are zero;

(ii) all row-effects are zero; and

(iii) all column-effects are zero,

we find the least-squares estimates of μ, R_i, C_j and τ_h by minimising the quantity

$$Q = \sum\sum e^2_{ij(h)}$$

subject to the restrictions that

$$\sum_i R_i = 0, \ \sum_j C_j = 0 \text{ and } \sum_h \tau_h = 0.$$

For partitioning the **total sum of squares** into components for rows, columns, treatments and error, let us construct the following identity:

$$\begin{aligned}
[Y_{ij(h)} - \bar{Y}_{...}] &= (\bar{Y}_{i..} - \bar{Y}_{...}) && \rightarrow \text{Row term} \\
&+ (\bar{Y}_{.j.} - \bar{Y}_{...}) && \rightarrow \text{Column term} \\
&+ (\bar{Y}_{..h} - \bar{Y}_{...}) && \rightarrow \text{Treatment} \\
&+ [Y_{ij(h)} - \bar{Y}_{i..} - \bar{Y}_{.j.} - \bar{Y}_{..h} + 2\bar{Y}_{...}] && \rightarrow \text{Error}
\end{aligned}$$

Squaring both sides of the identity and summing over all values, we get

$$\begin{aligned}
\sum\sum (Y_{ij(h)} - \bar{Y}_{...})^2 &= k \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2 + k \sum_j (\bar{Y}_{.j.} - \bar{Y}_{...})^2 \\
&+ k \sum_h (\bar{Y}_{..h} - \bar{Y}_{...})^2 + \sum\sum [Y_{ij(h)} - \bar{Y}_{i..} - \bar{Y}_{.j.} - \bar{Y}_{..h} + 2\bar{Y}_{...}]^2 \\
&+ \text{cross-product terms which vanish, when summed.}
\end{aligned}$$

This **sum-of-squares** identity may give rise to some confusion with respect to the summation process over i, j and h. The confusion disappears by an intelligent understanding of the following relations:

$$\sum\sum (Y_{ij(h)} - \bar{Y}_{...})^2 = \sum_i\sum_j (Y_{ij(h)} - \bar{Y}_{...})^2 = \sum_i\sum_h (Y_{ij(h)} - \bar{Y}_{...})^2$$
$$= \sum_j\sum_h (Y_{ij(h)} - \bar{Y}_{...})^2$$

Denoting the row, column, treatment and grand totals by R_i, C_j, T_h and G respectively, the various **sums of squares** terms are simplified as

Total $$SS = \sum\sum (Y_{ij(h)} - \bar{Y}_{...})^2 = \sum_i\sum_j Y^2_{ij(h)} - \frac{G^2}{k^2};$$

Rows $$SS = k\sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2 = \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2};$$

Columns $$SS = k\sum_j (\bar{Y}_{.j.} - \bar{Y}_{...})^2 = \sum_j \frac{C_j^2}{k} - \frac{G^2}{k^2};$$

Treatments $$SS = k\sum_h (\bar{Y}_{..h} - \bar{Y}_{...})^2 = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2};$$

and the sum of squares for experimental error may be obtained by subtraction.

The resulting analysis of variance is shown in the following table:

ANOVA-Table for a $k \times k$ Latin Square Design

Source of Variation	*d.f.*	Sum of Squares	Mean Square
Rows (R_i)	$k-1$	$R_{yy} = \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2}$	$s_r^2 = \frac{R_{yy}}{k-1}$
Columns (C_j)	$k-1$	$C_{yy} = \sum_j \frac{C_j^2}{k} - \frac{G^2}{k^2}$	$s_c^2 = \frac{C_{yy}}{k-1}$
Treatments (τ_h)	$k-1$	$T_{yy} = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2}$	$s_t^2 = \frac{T_{yy}}{k-1}$
Error ($e_{ij(h)}$)	$(k-1)(k-2)$	E_{yy} = by subtraction	$s_e^2 = \frac{E_{yy}}{(k-1)(k-2)}$
Total	k^2-1	$S_{yy} = \sum_i\sum_j Y^2_{ij(h)} - \frac{G^2}{k^2}$	---

Hence to test the hypothesis that treatment means are equal, we compute the ratio

$$F = \frac{s_t^2}{s_e^2} = \frac{MS \text{ for Treatments}}{MS \text{ for Error}}$$

which conforms to an F-distribution with $v_1 = k - 1$, $v_2 = (k-1)(k-2)$ *d.f.* if the null hypothesis is true. The hypothesis would be rejected when $F \geq F_\alpha; (v_1, v_2)$.

In a similar way, we can proceed to test the hypothesis that all row-effects or column-effects are zero.

The *least significant difference* (*LSD*) to test the difference between two treatment means selected at random, is given by

$$LSD = t_{\alpha/2} \sqrt{\frac{2\, s_e^2}{k}}, \text{ where } k \text{ is the number of rows or columns.}$$

23.5.3. Advantages and Disadvantages. A Latin square design has the following advantages and disadvantages:

(i) A Latin square design reduces the error variance by controlling two sources of variation.

(ii) The analysis of a *LS* design is simple and remains relatively simple with missing observations.

(iii) A Latin square design is generally more efficient than a randomized complete block design.

(iv) A Latin square design is less flexible than a *RCB* design. It is practical only for 5 to 10 treatments. When the number of treatments exceeds 10, the design is seldom used.

(v) For a small number of treatments, a *LS* design does not provide a sufficient number of replicates to give a valid estimate of error.

(vi) Replication in a Latin square design is costly.

(vii) In agricultural experimentation, the land requirement is rigid, the actual layout in the field may be laborious and approach to the central most plot becomes difficult.

Example 23.6. Five fertilizers *A*, *B*, *C*, *D* and *E* were tested by arranging plants in a Latin square design in a field. The rows and columns in the table are rows and columns in the field. The yields in bushels per plot are as shown.

	COLUMNS									
	1		2		3		4		5	
1	B	4.9	D	6.4	E	3.3	A	9.5	C	11.8
2	C	9.3	A	4.0	B	6.2	E	5.1	D	5.4
Rows 3	D	7.6	C	15.4	A	6.5	B	6.0	E	4.6
4	E	5.3	B	7.6	C	13.2	D	8.6	A	4.9
5	A	9.3	E	6.3	D	11.8	C	15.9	B	7.6

Analyse the data for evidence at the 5 per cent level that the mean yields are not equal for the five fertilizers. Also pick up the best fertilizer.

(i) We state our hypotheses as:

H_0: The mean yields for all the five fertilizers are equal, and

H_1: Not all mean yields for the fertilizers are equal.

(ii) The level of significance is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{MS \text{ for fertilizers}}{MS \text{ for error}},$$

which, if H_0 is true, has an F-distribution with $v_1 = k - 1$ and $v_2 = (k-1)(k-2)$ degrees of freedom.

(iv) Computations.

Rows	COLUMNS					R_i	R_i^2
	1	2	3	4	5		
1	4.9	6.4	3.3	9.5	11.8	35.9	1288.81
2	9.3	4.0	6.2	5.1	5.4	30.0	900.00
3	7.6	15.4	6.5	6.0	4.6	40.1	1608.10
4	5.3	7.6	13.2	8.6	4.9	39.6	1568.16
5	9.3	6.3	11.8	15.9	7.6	50.9	2590.81
C_j	36.4	39.7	41.0	45.1	34.3	196.5	7955.79
C_j^2	1324.96	1576.09	1681.00	2034.01	1176.49	7792.55	--

Summary for Fertilizers (or Treatments). (Total for a fertilizer or treatment is obtained by adding all the observed values of the fertilizer or treatment in the square).

Fertilizer	A	B	C	D	E
Total (T_h)	34.2	32.3	65.6	39.8	24.6
Means	6.84	6.46	13.12	7.96	4.92

Now, Total $SS\ (S_{yy}) = \sum_i \sum_j Y^2_{ij(h)} - \frac{G^2}{k^2}$

$$= (4.9)^2 + (6.4)^2 + \ldots + (15.9)^2 + (7.6)^2 - \frac{(196.5)^2}{25}$$

$$= 1829.83 - 1544.49 = 285.34$$

Rows $SS\ (R_{yy}) = \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2} = \frac{7955.79}{5} - 1544.49$

$$= 1591.16 - 1544.49 = 46.67$$

Columns $SS\ (C_{yy}) = \sum_j \frac{C_j^2}{k} - \frac{G^2}{k^2} = \frac{7792.55}{5} - 1544.49$

$$= 1558.51 - 1544.49 = 14.02$$

Fertilizers $SS\ (T_{yy}) = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2}$

$$= \frac{(34.2)^2 + (32.3)^2 + \ldots + (24.6)^2}{5} - \frac{(196.5)^2}{25}$$

$$= 1741.10 - 1544.49 = 196.91$$

Error $SS\ (E_{yy}) = S_{yy} - (R_{yy} + C_{yy} + T_{yy})$

$$= 285.34 - (46.67 + 14.02 + 196.61) = 28.04$$

These results are arranged in the following analysis of variance table:

Source of Variation	*d.f.*	Sum of Squares	Mean Square	Computed *F*
Rows	4	46.67	11.67	
Columns	4	14.02	3.51	
Fertilizers	4	196.61	49.15	$F = \frac{49.15}{2.34} = 21.00$
Error	12	28.04	2.34	--
Total	24	285.34		

(v) The critical region is $F \geq F_{0.05;}(4, 12) = 3.26$

(vi) **Conclusion.** Since the computed value of $F = 21.00$ falls in the critical region, we therefore reject our null hypothesis. Hence we conclude that the data provide sufficient evidence to indicate at the 5% significance level that the mean yields for the fertilizers are not equal.

Next, we compute the *least significant difference* (*LSD*) to determine which fertilizer is to be considered the best one. Thus

$$LSD = t_{0.025,(12)} \sqrt{\frac{2\,(MSE)}{k}}$$

$$= 2.18 \sqrt{\frac{2\,(2.34)}{5}} = (2.18)\,(0.96) = 2.09 \text{ bushels}$$

Arranging the mean yields for the five fertilizers in ascending order and drawing a line under any subset of adjacent mean yields that do not differ significantly, we get

E	B	A	D	C
4.92	6.46	6.84	7.96	13.12

The fertilizer C is the best one as the mean yield for fertilizer C differs significantly from the mean yields of all other fertilizers.

23.5.4. Missing Observations in a Latin Square. The formula for a missing observation in a $k \times k$ Latin square is developed as follows:

For convenience, we assume that the observation in the first row in the first column and for treatment one, is missing. We represent it by x. Let R_1', C_1' and T_1' represent the totals of observed values in the first row, of the first column and for treatment one respectively. The total for all the observed values is denoted by G' and the totals of other rows, columns and treatments are denoted by R_i, C_j and T_h respectively, where $i, j, h = 2, 3, \ldots, k$.

To find the best estimate of the missing observation, we minimize the error sum of squares. With the above assumptions, we may write the various sums of squares as follows:

Source of Variation	Sum of Squares
Rows	$R_{yy} = \frac{1}{k}\left[(R_1' + x)^2 + \sum_{i=2}^{k} R_i^2\right] - \frac{(G'+x)^2}{k^2}$
Columns	$C_{yy} = \frac{1}{k}\left[(C_1' + x)^2 + \sum_{j=2}^{k} C_j^2\right] - \frac{(G'+x)^2}{k^2}$
Treatments	$T_{yy} = \frac{1}{k}\left[(T_1' + x)^2 + \sum_{h=2}^{k} T_h^2\right] - \frac{(G'+x)^2}{k^2}$
Error	E_{yy} = By subtraction
Total	$S_{yy} = x^2 + \sum_i \sum_j Y_{ij(h)}^2 - \frac{(G+x)^2}{k^2}$

The error sum of squares is then obtained as

$$E_{yy} = S_{yy} - (R_{yy} + C_{yy} + T_{yy})$$

$$= x^2 + \frac{2\,(G' + x)^2}{k^2} - \frac{1}{k}\left[(R_1' + x)^2 + (C_1' + x)^2 + (T_1' + x)^2\right]$$

+ terms not involving x.

Differentiating E_{yy} *w.r.t.* x, equating the derivative to zero, and solving for x, we get

$$\frac{\partial E_{yy}}{\partial x} = 0 = 2x + \frac{4(G'+x)}{k^2} - \frac{2}{k}\left[(R_1' + x) + (C_1' + x) + (T_1' + x)\right]$$

$$\text{or} \quad x\left(1 + \frac{2}{k^2} - \frac{1}{k} - \frac{1}{k} - \frac{1}{k}\right) = \frac{1}{k}(R_1' + C_1' + T_1') - \frac{2G'}{k^2}$$

$$\text{or} \quad x\left(\frac{k^2 - 3k + 2}{k^2}\right) = \frac{k\,(R_1' + C_1' + T_1') - 2G'}{k^2}$$

$$\text{Hence } x = \frac{k(R_1' + C_1' + T_1') - 2G'}{(k-1)\,(k-2)}$$

$$[\because k^2 - 3k + 2 = (k-1)\,(k-2)]$$

When the missing observation corresponds to the ith row, the jth column and the hth treatment, the formula holds and may be written as

$$x = \frac{k(R'_i + C'_j + T'_h) - 2G'}{(k-1)(k-2)}$$

We then substitute the value of x and carry out the analysis of variance in the usual way, excepting that we reduce the degrees of freedom associated with *Error* and *Total* by 1. The formula for the estimated *S.E.* of the difference between a treatment mean and the mean of a treatment with a missing observation becomes

$$\sqrt{s_e^2 \left[\frac{2}{k} + \frac{1}{(k-1)(k-2)}\right]}$$

When two or more observations are missing, the formula for a single missing observation is repeated exactly in the same way as we did in the randomised complete block design, subtracting 1 from the degrees of freedom associated with error and total for each missing observation.

23.5.5 Efficiency of Latin Squares. Let s_r^2, s_c^2 and s_e^2 denote the *mean squares* for rows, columns and error in a $k \times k$ Latin square. Then the efficiency of a *LS* design relative to a *RCB* design is estimated as

(i) taking rows as blocks, by the formula

$$\text{Efficiency} = \frac{s_c^2 + (k-1)\, s_e^2}{k\, s_e^2} = \frac{1}{k}\left[\frac{s_c^2}{s_e^2} + (k-1)\right]$$

(ii) taking columns as blocks, by the formula

$$\text{Efficiency} = \frac{s_r^2 + (k-1)\, s_e^2}{k\, s_e^2} = \frac{1}{k}\left[\frac{s_r^2}{s_e^2} + (k-1)\right]$$

To evaluate the efficiency of a *LS* design relative to a completely randomised (*CR*) design, the appropriate formula is

$$\text{Efficiency} = \frac{s_r^2 + s_e^2 + (k-1)\, s_e^2}{(k-1)\, s_e^2}.$$

23.5.6. Orthogonal Latin Squares and Graeco-Latin Squares. Two Latin squares are said to be *orthogonal* if each letter of one square occurs exactly once with every letter of the other square when they are superimposed. If we write one of the two orthogonal squares with Latin letters and the other with Greek letters, then superimposing the two squares we get another square (or design) in which each Latin or Greek letter appears exactly once in each row and in each column, and each

Latin letter appears exactly once with each Greek letter. Such a square (or design) is called a *Graeco Latin Square* (*G-LS*). For example, if we have two orthogonal Latin squares, one with Latin letters and the other with Greek letters.

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

α	β	γ	δ
γ	δ	α	β
δ	γ	β	α
β	α	δ	γ

then, by superimposing, we get the following 4×4 Graeco-Latin square

$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
$B\gamma$	$A\delta$	$D\alpha$	$C\beta$
$C\delta$	$D\gamma$	$A\beta$	$B\alpha$
$D\beta$	$C\alpha$	$B\delta$	$A\gamma$

The analysis of variance appropriate to this design is given below:

Source of Variation	*d.f.*	Sum of Squares	Mean Square
Rows	$k-1$	$R_{yy} = \sum_i \frac{R_i^2}{k} - \frac{G^2}{k^2}$	$s_r^2 = \frac{R_{yy}}{k-1}$
Columns	$k-1$	$C_{yy} = \sum_j \frac{C_j^2}{k} - \frac{G^2}{k^2}$	$s_c^2 = \frac{C_{yy}}{k-1}$
Latin Letters (Treatments)	$k-1$	$T_{yy} = \sum_h \frac{T_h^2}{k} - \frac{G^2}{k^2}$	$s_t^2 = \frac{T_{yy}}{k-1}$
Greek Letters	$k-1$	$Q_{yy} = \sum_l \frac{Q_l^2}{k} - \frac{G^2}{k^2}$	$s_q^2 = \frac{Q_{yy}}{k-1}$
Error	$(k-1) \times (k-3)$	E_{yy} = by subtraction	$s_e^2 = \frac{E_{yy}}{(k-1)(k-3)}$
Total	k^2-1	$S_{yy} = \sum_i \sum_j Y_{ij}^2 - \frac{G^2}{k^2}$	---

where R_i, C_j, T_h and Q_l are the totals of the ith row, jth column, hth Latin letter (Treatment), lth Greek letter and G stands for the grand total. The assumption necessary for the analysis of a $k \times k$ *G-LS* design is that the observations may be represented by the linear statistical model

$Y_{ij(hl)} = \mu + R_i + \ldots + \varepsilon_{ij(hl)}$, $i, j, h, l = 1, 2, \ldots, k$

where $\sum R_i = \sum C_j = \sum \tau_h = \sum Q_l = 0$, and where $\varepsilon_{ij(hl)}$ are independently and normally distributed with zero mean and common variance σ^2.

A *G-LS* design is used when we desire to control three sources of variation, one source of variation being controlled by rows, the other source by columns and the third source by Greek letters. The degrees of freedom associated with *error* would be inadequate when k is less than 4. It has been shown that a 6×6 Graeco-Latin square does not exist. Cochran and Cox have given the layouts of the Graeco-Latin Square designs for all numbers of treatments from 3 to 12 with the exception of six and ten in their book *Experimental Designs*.

It should be noted that no more than $(k-1)$ Latin squares of order k can be orthogonal. A set of $(k-1)$ orthogonal squares of order k, is called a *complete set of orthogonal Latini squares*. Such complete sets can be formed whenever k is a prime number or the power of a prime number.

23.6 SINGLE DEGREE OF FREEDOM CONTRASTS

In many experiments, we may desire to partition the Sum of Squares for Treatments into a number of components, each based on 1 degree of freedom. Let Q_j be any contrast among the treatment τ's. Then the sum of squares for this contrast is computed by

$$Q_j\,SS = \frac{(\text{contrast})^2}{r\sum(\text{contrast coefficient})^2} = \frac{Q_j^2}{r\sum c_{ij}^2}$$

where c_{ij}'s are constant co-efficients of the treatment totals in the jth contrast and r is the number of observations in each total. The quantity $\frac{Q_j^2}{r\sum c_{ij}^2}$ is a component of the sum of squares for Treatment and it represents 1 degree of freedom. The Treatment Sum of Squares with $(k-1)$ degrees of freedom may thus be partitioned into $(k-1)$ independent contrasts, each based on 1 degree of freedom. In other words, we may have

$$\text{Treatment } SS = \frac{Q_1^2}{r\sum c_{i1}^2} + \frac{Q_2^2}{r\sum c_{i2}^2} + \ldots + \frac{Q_{k-1}^2}{r\sum c_{i(k-1)}^2}$$

Let us consider an example. Suppose we wish to compare τ_1 with the mean of τ_2 and τ_3. Then

$$Q_1 = 2T_1 - (T_2 + T_3) \qquad (T_j = \text{Total of } \tau_j)$$

and the sum of square for this contrast is

$$\frac{Q_1^2}{r\sum c_{i1}^2} = \frac{[2T_1 - (T_2 + T_3)]^2}{r\,[2^2 + 2(-1)^2]}$$

Similarly, for a contrast defined by

$$Q_2 = 2\,(T_1 + T_2 + T_3) - 3\,(T_4 + T_5)$$

the sum of squares will be

$$\frac{Q_2^2}{r\sum c_{i2}^2} = \frac{[2(T_1 + T_2 + T_3) - 3(T_4 + T_5)]^2}{r\,[3(2)^2 + 2(-3)^2]}$$

The *mean squares* of these contrasts are taken as the numerator for an F-test to investigate whether the effects estimated by these contrasts are zero.

23.7 FACTORIAL EXPERIMENTS

Experiments are often planned to investigate the effects of say, different rates of fertilizers, different dates of planting, different categories of education, different intensities of a stimulus, etc. Technically speaking the independent variables such as fertilizer, planting, education, stimulus, etc. are called *factors*, while the values such as rates, dates, categories or intensities at which a factor is held fixed, are known as *levels*. It is customary to represent the *factors* by small letters a, b, c, etc., and a particular *level* by small letter with a subscript a_i, b_j, c_h, etc. A treatment is then determined by a combination of different levels of factors a_i, b_j, c_h, ..., etc. For example, if we have two factors a and b, each at 2 levels zero and one (*i.e.*, $i = 0, 1; j = 0, 1$), then the four treatment combinations are

$$a_0b_0,\ a_1b_0,\ a_0b_1 \text{ and } a_1b_1.$$

An experiment is called a **factorial experiment** if the treatments consist of all possible combinations of several levels of several factors. When each treatment combination is used the same number of times, the factorial experiment is known as a **complete factorial experiment**. The experiment is described as p^n factorial experiment, when there are n factors and each factor is considered at p levels, (the number of factors being the exponent). Thus a 2^2 or 2×2 factorial experiment means 2 factors each at 2 levels.

23.7.1. Main Effects and Interaction Effects. The comparisons among the treatment combinations are called the Effects, which are represented by capitals A, B, C, etc. These effects are composed of *main effects and interactions*. A *main effect* of a factor is defined as a measure of the *average change* in effect produced by changing the level of the factor. It is measured independently of other factors and is the effect of the factor only. Factors are said to *interact*, when they are not independent. But *interaction* in a factorial experiment, is a measure of the extent to which the effect of changing the levels of one or more factors depends on the levels of the other factors. Interactions between two factors are referred to as *first order interactions*, those concerning three factors, as *second order interactions* and so on. A main effect is sometimes regarded as an *interaction of zero order*. These concepts may also be expressed in symbols.

23.7.2. Effects in a 2^2-Factorial Experiment. Let the symbols $a_i b_j$ $(i = 0, 1; j = 0, 1)$ represent both the treatment combinations and the yields from all experimental units or plots. Then the main effect of factor a is determined as below:

Effect of factor a at level b_0 of factor $b = a_1 b_0 - a_0 b_0$

Effect of factor a at level b_1 of factor $b = a_1 b_1 - a_0 b_1$

$\therefore$ Main effect of factor "a" = Average change produced by varying factor a.

$$= \frac{1}{2}[(a_1 b_0 - a_0 b_0) + (a_1 b_1 - a_0 b_1)]$$

$$= \frac{1}{2}(a_1 - a_0)(b_1 + b_0)$$

$$= A$$

Similarly the main effect of factor "b"

$$= \frac{1}{2}[(a_0 b_1 - a_0 b_0) + (a_1 b_1 - a_1 b_0)]$$

$$= \frac{1}{2}(a_1 + a_0)(b_1 - b_0)$$

$$= B.$$

If the two factors a and b were acting independently, the effect of a at b_0 and the effect of a at b_1 or the effect of b at a_0 and the effect of b at a_1, should be equal, but, in general, they will be different. This difference is a measure of the extent to which the factors interact. Hence $A \times B$, the interaction between two factors a and b, each at 2 levels zero and 1, is given by

$$A \times B = \frac{1}{2}[(a_1 b_1 - a_0 b_1) - (a_1 b_0 - a_0 b_0)]$$

$$= \frac{1}{2}[(a_1 b_1 - a_1 b_0) - (a_0 b_1 - a_0 b_0)]$$

$$= \frac{1}{2}(a_1 - a_0)(b_1 - b_0).$$

It is clear from this relation that interaction between factors a and b, *i.e.*, AB is the same as that between b and a, *i.e.* BA.

The overall mean is represented by M and is the average of all the yields, *i.e.*,

$$M = \frac{1}{4}(a_0 b_0 + a_0 b_1 + a_1 b_0 + a_1 b_1)$$

$$= \frac{1}{4}(a_0 + a_1)(b_0 + b_1)$$

Replacing the symbols a_0 and b_0 by 1, and the symbols a_1 and b_1 by a and b, (*i.e.*, writing (1) for $a_0 b_0$, a for $a_1 b_0$, b for $a_0 b_1$ and ab for $a_1 b_1$), the preceding comparisons may be expressed as

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

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

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

$$M = \frac{1}{4}(a + 1)(b + 1)$$

These effects can be conveniently written in a table of plus and minus signs as below:

Effect	Treatment Combinations				Divisor
	1	a	b	ab	
M	+	+	+	+	4
A	−	+	−	+	2
B	−	−	+	+	2
AB	+	−	−	+	2

It should be noted that the effects A, B and AB are 3 mutually orthogonal contrasts of the yields of the 4 treatments, each based on 1

degree of freedom. Further, the treatment combinations written in the order (1), a, b, ab are referred to as *standard order* (form).

23.7.3. Effects in a 2^3-Factorial Experiment. In this case, we consider 3 factors a, b and c, each at 2 levels. Representing the treatment combinations by the symbols $a_i b_j c_h$ ($i = 0, 1; j = 0, 1; h = 0, 1$) we get 8 treatment combinations: (1), a, b, c, ab, ac, bc and abc. The main effects and interactions, as defined before, can be represented by the following relations, which on being expanded algebraically, give expressions for the yields of a treatment combination:

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

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

$$B = \frac{1}{4}(a+1)(b-1)(c+1),$$

$$C = \frac{1}{4}(a+1)(b+1)(c-1),$$

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

$$AC = \frac{1}{4}(a-1)(b+1)(c-1),$$

$$BC = \frac{1}{4}(a+1)(b-1)(c-1), \text{ and}$$

$$ABC = \frac{1}{4}(a-1)(b-1)(c-1).$$

These expressions can be presented in the following tabular form:

Effect	Treatment Combinations (in standard form)								Divisor
	(1)	(a)	(b)	(ab)	(c)	(ac)	(bc)	(abc)	
M	+	+	+	+	+	+	+	+	8
A	−	+	−	+	−	+	−	+	4
B	−	−	+	+	−	−	+	+	4
C	−	−	−	−	+	+	+	+	4
AB	+	−	−	+	+	−	−	+	4
AC	+	−	+	−	−	+	−	+	4
BC	+	+	−	−	−	−	+	+	4
ABC	−	+	+	−	+	−	−	+	4

This whole thing can be extended to include n factors, all at two levels, by considering the expression

$$\frac{1}{2^{n-1}}(a \pm 1)(b \pm 1)(c \pm 1) \ldots,$$

where a minus sign appears in any factor on right if the corresponding letter is present on the left.

23.7.4. Design and Analysis for Factorial Experiments. A factorial experiment is not considered as an experimental design because of the fact that the basic designs, namely, the *CR* design, the *RCB* design and the *LS* design, are used to carry out the factorial experiments. The point to make here is that the treatments have a factorial structure.

For the purpose of analysis of variance, the basic sums of squares are computed in the usual manner, with the addition that the *treatment sum of squares* is further partitioned into component parts of *main effects and interactions*. It is interesting to note that the best estimate of any effect is obtained from the contrasts shown in the preceding tables of plus and minus signs. The significance of all effects is tested against the error mean square. Sometimes, we may not be interested in interactions, then we pool their sums of squares, assuming that they do not exist.

Let us consider a 2^2-factorial experiment which has been carried out in a randomised complete block design with r replications. The statistical model as before, would be

$$Y_{ij} = \mu + \beta_i + \tau_j + e_{ij}.$$

But the treatment effect would be composed of two main effects A and B, and their interaction AB. The **Block SS**, the **Treatment SS** and the **Error SS** are computed in the usual manner. The **sum of squares** for **main effects** and **interaction**, which come from a breakdown of the treatment effect, are computed as below:

Defining the *effect total* by [ ], *i.e.*,

$$[A] = -[1] + [a] - [b] + [ab], \text{ etc.}$$

the pertinent formulas are

$$SS \text{ for main effect } A = \frac{(\text{Contrast})^2}{r\sum(\text{contrast coefficient})^2} = \frac{[A]^2}{4r}$$

$$SS \text{ for main effect } B = \frac{[B]^2}{4r}, \text{ and}$$

$$SS \text{ for interaction } AB = \frac{[AB]^2}{4r}.$$

With this model, the analysis of variance table would be as follows:

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Blocks	$r-1$	Usual Block SS		
Treatments	3	Usual Treatment SS		
A-Effect	1	$[A]^2/4r$		
B-Effect	1	$[B]^2/4r$		
AB-Effect	1	$[AB]^2/4r$		
Error	$3(r-1)$	By subtraction	s_e^2	---
Total	$4r-1$	Usual Total SS	--	--

To test the appropriate hypotheses regarding main effects A and B, and the interaction AB, we calculate F-statistics with error mean square in the denominator. If the treatment MS does not prove to be significant, no question of examining the effects A, B or AB arises.

The sums of squares for effects may also be obtained by constructing a 2-way table for the factors a and b.

For the analysis of variance of 2^3-factorial experiment, the treatment SS would be partitioned into 7 components, A, B, C, AB, AC, BC and ABC, each associated with 1 degree of freedom. The sum of squares for any effect is computed by $[\text{effect}]^2/8r$, where r represents the number of replicates.

Example 23.7. A 2^2-factorial experiment, *i.e.*, with 2 varieties (factors) and 2 manures (levels), was carried out in a randomized complete block design with 3 replicates. The yields given in the following table, are hypothetical.

Replicates	Treatment Combinations			
	$v_1 m_1$	$v_1 m_2$	$v_2 m_1$	$v_2 m_2$
1	5	7	8	10
2	4	4	7	5
3	6	4	9	12

Perform the analysis of variance and test the significance of varieties (factors) and manures (levels).

(i) We formulate our null hypothesis as

H_0: There is no difference between treatment combinations. If it is rejected, then the other hypotheses are

(a) The variety effects are zero.

(b) The manure effects are zero.

(c) There is no interaction effect.

These are tested against the appropriate alternative hypotheses.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{MS \text{ for the effect}}{MS \text{ for error}}$$

which, if H_0 is true, has an F-distribution with appropriate degrees of freedom.

(iv) Computations. The sums of squares are calculated as below:

Replicates	Treatment Combinations				$B_{i\cdot}$	$B_{i\cdot}^2$	$\sum_j Y_{ij}^2$
	$v_1m_1=(1)$	$v_1m_2=m$	$v_2m_1=v$	$v_2m_2=vm$			
1	5 (25)	7 (49)	8 (64)	10 (100)	30	900	238
2	4 (16)	4 (16)	7 (49)	5 (25)	20	400	106
3	6 (36)	4 (16)	9 (81)	12 (144)	31	961	277
$T_{\cdot j}$	15	15	24	27	81	2261	621
$T_{\cdot j}^2$	225	225	576	729	1755	--	↑
$\sum_i Y_{ij}^2$	77	81	194	269	621	← Check	

Now, Correction Factor, $C.F. = \frac{(81)^2}{12} = 546.75$

Total $\quad SS = \sum_i \sum_j Y_{ij}^2 - C.F. = 621.00 - 546.75 = 74.25$

Block $\quad SS = \sum_i \frac{B_{i\cdot}^2}{k} - C.F.$, where k is the number of treatments

$$= \frac{2261}{4} - C.F. = 565.25 - 546.75 = 18.50$$

Treatment $\quad SS = \sum_j \frac{T_{\cdot j}^2}{r} - C.F.$, where r is the number of replicates.

$$= \frac{1755}{3} - C.F. = 585.00 - 546.75 = 38.25$$

$\therefore$ Error $SS = 74.25 - (18.50 + 38.25) = 17.50$

To compute the sums of squares for main effects and interaction, we first find the effect-totals as

$$[V] = [vm] + [v] - [m] - [1] = 27 + 24 - 15 - 15 = 21,$$

$$[M] = [vm] - [v] + [m] - [1] = 27 - 24 + 15 - 15 = 3,$$

$$[VM] = [vm] - [v] - [m] + [1] = 27 - 24 - 15 + 15 = 3$$

Hence SS for Varieties $= \dfrac{[V]^2}{4r} = \dfrac{[21]^2}{12} = 36.75,$

SS for Manures $= \dfrac{[M]^2}{4r} = \dfrac{[3]^2}{12} = 0.75$, and

SS for Interaction, $VM = \dfrac{[VM]^2}{4r} = \dfrac{[3]^2}{12} = 0.75.$

The sums of squares for Varieties, Manures and Interaction VM can also be computed by forming a 2-way table as below:

Manures	Varieties		Total
	v_1	v_2	
m_1	15	24	39
m_2	15	27	42
Total	30	51	81

Total SS (for this table) $= \dfrac{(15)^2 + \ldots + (27)^2}{3} - C.F. = 38.25,$

which is SS for treatment combinations. (It has been divided by 3 as each sum comes from 3 plots).

Variety $SS = \dfrac{(30)^2 + (51)^2}{6} - C.F. = 36.75$

Manure $SS = \dfrac{(39)^2 + (42)^2}{6} - C.F. = 0.75$

Interaction VM = Error SS in this table

$= 38.25 - (36.75 + 0.75) = 0.75,$

which are the same.

These results are summarized in the following *ANOVA*-Table:

Source of Variation	*d.f.*	Sum of Squares	Mean Square	F	$F_{0.05}(\nu_1,\nu_2)$
Blocks	2	18.50	9.25	---	----
Treatments	3	38.25	12.75	4.37	4.76
Varieties (V)	1	36.75	36.75	12.9	5.99
Manures (M)	1	0.75	0.75	0.26	5.99
Interaction (VM)	1	0.75	0.75	0.26	5.99
Error	6	17.50	$s_e^2=2.92$	--	--
Total	11	74.25	---	---	---

(v) For critical regions, the table values of F at $\alpha = 0.05$ for appropriate degrees of freedom are given in the last column of the *ANOVA*-table.

(vi) **Conclusion.** Since the computed value of $F = 4.37$ regarding Treatments does not fall in the critical region, so we accept H_0 that there is no difference between treatment combinations. As H_0 is accepted, we therefore do no bother to test the significance of the differences between varieties or manures, etc.

23.7.5. Yates' Technique for Computing Contrasts. For large factorial experiments, there is the difficulty in writing out the table of plus and minus signs. A systematic method for computing the contrasts in a 2^n factorial experiment, developed by Yates, is given below:

(i) Write down the treatment combinations and observations in *standard form.*

(ii) Add the treatment combinations (observations) in pairs and write down in column marked (1). This fills up the top half of the column (1). For bottom half, write down the differences of pairs by subtracting the upper from the lower treatment combinations (observations). This completes column (1).

(iii) Obtain column marked (2) in exactly the same way, using the results in column (1), *i.e.*, operating on the results of the preceding column.

(iv) Continue this operation until the nth column for a 2^n experiment is reached

The column (n) corresponds to effects, the first value will be the grand total of the entire experiment. Each remaining value will be a contrast in the treatment combination in that order. The sums of squares for *Main Effects* and *Interaction Effects* are calculated by squaring each of the values given in column (n) and dividing the results by $2^n \cdot r$, where r is the number of replicates.

As an illustration, the procedure is outlined in the following table for a 2^2-factorial experiment.

Yates' Method for a 2^2-Factorial Experiment

Treatment Combination	(1)	(2)	Effect
(1)	$(1) + a$	$(1) + a + b + ab$	Total
a	$b + ab$	$a - (1) + ab - b$	A
b	$a - (1)$	$b + ab - (1) - a$	B
ab	$ab - b$	$ab - b - a + (1)$	AB

23.7.6. Advantages and Disadvantages. The advantages and disadvantages of a factorial experiment are stated as follows:

(i) A factorial experiment is usually economical.

(ii) All the experimental units are used in computing the main effects and interactions.

(iii) The use of all treatment combinations makes the experiment more efficient and comprehensive.

(iv) The interaction effects are easily estimated and tested through the usual analysis of variance.

(v) The experiment yields unbiased estimates of effects, which are of wider applicability.

(vi) A factorial experiment requires an excessive amount of experimentation when there are several factors at several levels. For example, a factorial arrangement of 8 factors, each at 2 levels requires 256 combinations and the number of combinations in case of 7 factors, each at 3 levels, would be 2187. To overcome this difficulty, we use a device, known as *fractional replication*, where certain properly chosen levels of factors are omitted.

(vii) A large number of combinations when used, cause decrease in the efficiency of the experiment. The experiment may be reduced to a manageable size by *confounding* some effects considered of little practical consequence.

(viii) The experimental set up and the resulting statistical analysis are more complex.

EXERCISES

23.1 (a) What is meant by an experimental design? Describe, in brief, the basic principles of experimental designs. (P.U., B.A./B.Sc. 1985, 91)

(b) Define and discuss the use of Randomization and Replication in designing an experiment. (P.U., B.A/B.Sc. 1988)

23.2 Discuss the importance of Randomization, Replication and Local Control in designing an experiment. What effects do they have on validity of conclusions and inherent errors of experiment? (P.U., M.A., 1961)

23.3 (a) Discuss the purpose of *replication* in the experimental design.

(b) Show mathematically how variance is analysed into two or three independent parts corresponding to recognised sources of variation of data.

23.4 (a) Describe a Completely Randomized design, its model and analysis. What are its advantages and disadvantages? (P.U., B.A/B.Sc. 1992)

(b) The following table contains the body weights of calves at 8 weeks of age. There were 3 levels of feeding given to a random sample of five calves each. The completely randomized design was used. Obtain the standard error of a feeding treatment mean for the data on body weights.

Level of Feeding		
Subnormal	Normal	Supernormal
118	142	162
122	129	173
121	134	168
126	132	183
109	135	172

23.5 Thirty five plots, of approximately equal fertility, were sown with 7 different varieties of wheat, 5 plots to each variety, the distribution of varieties among the plots being random. The following table gives the yields of grain in bushels per acre, the 7 columns corresponding to the different varieties. Do the data (fictitious) indicate a significant difference in the yields of the varieties?

13	15	14	14	17	15	16
11	11	10	10	15	9	12
10	13	12	15	14	13	15
16	18	13	17	19	14	15
12	12	11	10	12	10	11

23.6 In order to study the effect of storage condition on the moisture content of white pine lumber, five storage methods were investigated, with varying numbers of experimental units (sample boards) being stored under each condition. The data (observations in %) thus obtained are given below:

Storage Condition				
1	2	3	4	5
7.3	5.4	8.1	7.9	7.1
8.3	7.4	6.4	9.5	
7.6	7.1		10.5	
8.4				
8.3				

State your null hypothesis, both in words and symbolically, about the storage conditions. Compute the analysis of variance and state your conclusions.

23.7 The staff of a university newspaper is experimenting with different formats. A completely randomized experiment is designed to compare three different formats. Volunteer students in a journalism class are given one of three formats and, after reading the news stories, take a reading comprehension test. This gives the following statistics:

Format	n	$\bar{Y}$	s^2
1	10	90	4.37
2	10	86	3.76
3	11	83	4.21

Perform the analysis of variance, calculate the F-statistic, and determine the significance level for the hypothesis that the mean reading comprehension is the same for all formats.

23.8 (a) Describe a Randomized Complete Block Design, its model and analysis. What are its advantages and disadvantages?

(b) The analysis of variance for a *RCB* design produced the *ANOVA* table shown below:

S.V.	d.f.	SS	MS	F-ratio
Treatments	3	28.2	13.80	
Blocks	5			
Error		34.1		

(i) Complete the *ANOVA* table.

(ii) Do the data provide sufficient evidence to indicate a difference among the treatment means? Test using $\alpha = 0.01$

(iii) If the sample means for treatments A and B are $\bar{y}_A = 9.7$ and $\bar{y}_B = 12.1$ respectively, find a 90% confidence interval for $(\mu_A - \mu_B)$.

23.9 (a) Compare Randomized Complete Block experiments with Completely Randomized experiments, comparing their respective advantages and relative efficiency, with illustrations. (P.U., B.A/B.Sc. 1986)

(b) Three varieties A, B and C of a crop are tested in a randomized block design with four replications, the layout being given in the diagram appended. The plot yields in pounds are also indicated therein. Analyse the experimental yields and state your conclusions.

Replications	1	A 32.1	C 34.2	B 31.7
	2	C 30.7	A 17.0	B 32.7
	3	A 40.8	B 25.3	C 48.2
	4	B 47.9	C 59.6	A 26.8

(P.U., B.A/B.Sc. 1984)

23.10 In a randomized complete block design, in each of four blocks I, II, III and IV, four varieties of wheat A, B, C, D are grown in the layout given below and the yields are also indicated therein:

I	B 27	A 17	C 15	D 25
II	A 28	D 22	B 26	C 16
III	D 14	C 11	A 22	B 25
IV	C 18	B 18	D 19	A 17

(a) Perform the analysis of variance to test at 0.05 significance level, the differences in the yields of varieties and in blocks.

(b) What would have been the result if no blocking had been done, *i.e.*, if we consider it as a completely randomized design? (P.U., B.A/B.Sc. 1986, 96)

23.11 The following data refer to three methods of hardening steel, which are referred to as A, B and C. A long strip of steel was cut into 18 pieces and the treatments allocated at random, forming six complete blocks. The actual layout of the treatments and data are shown below:

A 813	B 647	C 713		C 814	B 759	A 795		A 705	C 652	B 598
A 774	C 617	B 559		B 580	A 687	C 539		A 581	B 480	C 437

(a) Perform a two-way analysis of variance and test the significance of treatments.

(b) What would have been the result if no blocking had been done?

23.12 (a) In a randomized block experiment b blocks are divided into p-plots each, and p treatments are applied to one of the plots in each block. If Y_{ij} be the observed value of the jth treatment in the ith block, set up an appropriate analysis of variance table.

(b) For the data given below, test for the significance between the treatment means.

Blocks	Treatments			
	1	2	3	4
I	23	19	25	23
II	20	17	24	21
III	24	20	29	27
IV	22	21	24	18

23.13 Four different kinds of fertilizers f_1, f_2, f_3 and f_4 are used to study the yield of beans. The soil is divided into three blocks each containing four homogeneous plots. The yields in pounds per acre and the corresponding treatments are as follows:

Block 1	Block 2	Block 3
$f_1 = 42.7$	$f_3 = 50.9$	$f_4 = 51.1$
$f_3 = 48.5$	$f_1 = 50.0$	$f_2 = 46.3$
$f_4 = 32.8$	$f_2 = 38.0$	$f_1 = 51.9$
$f_2 = 39.3$	$f_4 = 40.2$	$f_3 = 53.5$

(a) Conduct an analysis of variance using the randomized complete block model.

(b) Use a single-degree-of-freedom contrasts to make the following comparisons among the fertilizers: (i) (f_1, f_3) versus (f_2, f_4); (ii) f_1 versus f_3.

23.14 Four treatments are given in four blocks in a Randomized Block Design as tabled below:

Blocks	Treatments				Total
	2 CS	2 CN	1 CS	1 CN	
I					1487
II					1134
III					1892
IV					614
Total	819	904	1840	1564	5127

Correction Factor = 1642883

Total Sum of Squares = 425314

Perform the analysis of variance and find out the F-ratio for treatments. (P.U., B.A./B.Sc. 1980)

23.15 Given the following abbreviated analysis of variance for a randomized complete block design:

Source of Variation	*d.f.*	Sum of Squares	Mean Square
Blocks	9	0.4074	
Treatments	3	1.1986	
Error	27	0.6249	

(a) Complete the analysis; fill in the mean squares.

(b) Compute the standard error for a treatment mean and for the difference between 2 treatment means.

(c) The treatment means are 1.464, 1.195, 1.325 and 1.662. What mean or means do you suspect might represent different population?

(d) Estimate the efficiency of this design relative to completely randomized design.

(P.U., B.A/B.Sc. 1983, 92)

23.16 Obtain an expression for estimating a missing observation in a randomized block experiment (when a single plot is missing) and explain how you will analyse the reconstructed data.

23.17 (a) Show that $x = \dfrac{kT + rB - G}{(k-1)(r-1)}$, where T and B are the total yields for the treatment and block with the missing plot, and G is the grand total of observed values in the $(r \times k)$ randomized block design.

(b) Determine the missing observation x and compute the analysis of variance.

Blocks	Treatments				
	A	B	C	D	E
I	5	6	6	7	4
II	x	7	7	8	5
III	7	7	7	9	5

23.18 Estimate the missing observations in the following $r \times c$ randomized block experiment so as to perform the analysis of variance.

Blocks	Treatments					
	1	2	3	4	5	6
I	18.5	15.7	16.2	14.1	13.0	13.6
II	11.7	y_{22}	12.9	y_{24}	16.9	12.5
III	15.4	16.6	15.5	20.3	18.4	21.6
IV	y_{41}	18.6	12.7	15.7	16.5	18.0

23.19 (a) Explain how analysis of covariance can be used to estimate a missing observation in experimental designs.

(b) Apply the covariance technique to estimate the missing value, m from the following data. Also obtain the treatment mean square and the error mean square.

Blocks	Treatments			
	I	II	III	IV
1	4.4	6.8	6.3	6.4
2	4.0	m	4.9	7.3
3	4.5	7.0	5.9	6.7
4	3.1	6.4	7.1	6.7

23.20 (a) Describe a Latin Square Design and it analysis. What are the advantages and disadvantages of a Latin square design? (P.U., B.A/B.Sc. 1990)

(b) Prepare a layout for 6 varieties of wheat in a Latin square design.

23.21 (a) Define a Latin square design. Explain the difference between a Latin square design and a randomized complete block design. (P.U., B.A/B.Sc. 1984)

(b) Why is a Latin square design sometimes referred to as a *double blocking*?

(c) What restrictions on randomness are involved in a LS design?

23.22 (a) What is a *reduced* Latin square?

(b) Carry out the analysis of variance for the following Latin square.

V_1 (2.3)	V_2 (3.0)	V_3 (3.3)	V_4 (2.5)
V_2 (3.1)	V_3 (4.1)	V_4 (2.4)	V_1 (2.4)
V_3 (4.3)	V_4 (2.5)	V_1 (2.1)	V_2 (2.9)
V_4 (2.6)	V_1 (2.0)	V_2 (2.4)	V_3 (4.4)

(P.U., B.A./B.Sc. 1979)

23.23 The atmosphere in 4 different districts of a large town was sampled, the samples being taken at 4 different heights. Four different tests for the presence of a certain chemical were made on the samples. The arrangement is shown in the following table with the percentage by weight of the chemical as determined by the tests. Letters denote the different tests.

Heights	Districts			
	1	2	3	4
1	A 8	B 5.3	C 4.1	D 5
2	D 6.8	A 4.9	B 4.1	C 3.2
3	B 6.3	C 4.7	D 4.0	A 5
4	C 5.7	D 3.3	A 4.0	B 4.2

Is there evidence of significant variation from district to district and between heights in the percentage of the chemical present in the atmosphere?

23.24 An experiment was conducted to assess the relative resistances to abrasion of four grades of leather (*A*, *B*, *C*, *D*). A machine was used in which the samples could be tested in any one of 4 positions. Since different runs (replications) are known to yield variable results, it was decided to make four runs. A Latin Square design was used and the following results were obtained.

Runs	Positions			
	1	2	3	4
1	(*A*) 150	(*B*) 145	(*D*) 130	(*C*) 133
2	(*D*) 130	(*C*) 172	(*A*) 170	(*B*) 127
3	(*C*) 133	(*D*) 132	(*B*) 115	(*A*) 170
4	(*B*) 98	(*A*) 171	(*C*) 132	(*D*) 120

Perform the analysis of variance and test the significance of the grades of leather.

23.25 Four machines were tested in a 4×4 Latin Square design to see if they should be adjusted to produce more uniform product. If only one adjustment can be made, which machine should be adjusted and by how much? Justify your recommendations by completing the following analysis?

Source of Variation	Mean Square
Rows (Operators)	136
Columns (Time Periods)	30
Machines (Treatments)	1649
Error	81

Treatments	1	2	3	4
Means	83.00	44.75	40.00	43.00

(P.U., B.A./B.Sc., Hons. Part III, 1964)

23.26 The following is a 5×5 Latin Square for the data taken from a manurial experiment with sugarcane. The five treatments were as follows:

A: no manure; *B*: an inorganic manure; *C*,*D*,*E*: three levels of farm-yard manure.

Plan and Yield of Sugarcane (in suitable units) per plot are as follows.

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

Analyse the above data to find out if there are any differences among the treatment effects at 5% level of significance.

(P.U., B.A/B.Sc. 1985)

23.27 An agricultural experiment was conducted on the Latin square plan to test the effect on yield due to change of treatment (5 kinds) and also to variation of soil in each of two perpendicular directions. The results are set out in the Latin square below, in which letters correspond to treatments, while rows and columns correspond to the two perpendicular directions.

A 7.4	D 8.9	E 5.8	B 12.0	C 14.3
C 11.8	B 6.5	A 8.7	E 7.6	D 7.9
D 10.1	C 17.9	B 9.0	A 8.5	E 7.1
E 8.8	A 10.1	C 15.7	D 11.1	B 7.4
B 11.8	E 8.8	D 14.3	C 18.4	A 10.1

(a) Perform an analysis of variance and find out if the effects on yield are significant.

(b) If no blocking had been done in rows and columns, what would have been the conclusion?

23.28 The responsive of five monkeys to a stimulus under five different conditions during five periods consisting of successive weeks were observed according to the Latin Square design below. The numbers are the total number of responses and letters denote the conditions. Analyse the data, describing briefly any assumptions that you make.

	Periods 1	2	3	4	5
Monkeys 1	194 B	369 D	344 C	380 A	693 E
2	202 D	142 B	200 A	356 E	473 C
3	355 C	301 A	439 E	338 B	528 D
4	515 E	590 C	552 B	677 D	546 A
5	184 A	421 E	355 D	284 C	366 B

Before the experiment was carried out, it was thought that condition E would produce a greater response on the average than condition D. Do the data support this conjecture?

23.29 We wish to conduct a field experiment to test the yielding ability of 6 varieties of wheat and have available an area of land sufficient for 36 plots. Give a layout and indicate the proper partitioning of the total degrees of freedom for the following experimental designs:

(a) Completely randomized,

(b) Randomized complete block,

(c) Latin square.

Indicate, by means of arrows, the proper F-tests for testing variety differences in each design. (P.U., B.A/B.Sc. 1975, '87)

23.30 Describe Latin Square design and its analysis. How does it differ from a Randomized Complete Block design? How will you estimate missing values in a Latin Square design? Estimate the relative efficiency of the two designs mentioned above.

(P.U., M.A. Stat., 1960)

23.31 (a) Given a Latin Square design with r rows, columns and treatments. Assume that the data for one plot were lost. Show that the estimate of missing observation found by minimising error sum of squares is

$$y = \frac{r\,(R + C + T) - 2G}{(r-1)\,(r-2)}$$

where R, C and T are respectively the totals of row, column and treatment which contain the missing observation; and G is the grand total.

(b) Supposing observation for treatment D in row 1 and column 2 of question 23.27 were missing, estimate it for the purpose of analysis of variance.

23.32 (a) What is meant by a complete set of orthogonal Latin Squares, and in what circumstances might such a set prove useful? Construct a complete set of orthogonal Latin Squares of side 4.

(b) Indicate the main stages in the analysis of variance for a Graeco-Latin square. (P.U., B.Sc. Hons. Part II, 1972)

23.33 Construct a 5×5 Graeco-Latin Square experimental design, indicating the steps in the construction, and give the analysis of variance appropriate to this design.

Comment on assumptions necessary for the analysis and on the use of the design.

23.34 A composite measure of screen quality was made on screen using four lacquer concentrations, four standing times, four acryloid concentrations, (A, B, C, D) and four acetone concentrations (α, β, γ, δ). A Graeco-Latin square design was used with data recorded as follows;

Standing Time	Lacquer Concentrations			
	$\frac{1}{2}$	1	$1\frac{1}{2}$	2
30	$C\beta$ (16)	$B\gamma$ (12)	$D\delta$ (17)	$A\alpha$ (11)
20	$B\alpha$ (15)	$C\delta$ (14)	$A\gamma$ (15)	$D\beta$ (14)
10	$A\delta$ (12)	$D\alpha$ (6)	$B\beta$ (14)	$C\gamma$ (13)
5	$D\gamma$ (9)	$A\beta$ (9)	$C\alpha$ (8)	$B\delta$ (9)

Do a complete analysis of these data.

23.35 What are factorial experiments? What are their advantages over single factor experiments? Explain the terms *Main Effects* and *Interaction*, and discuss how you will separate the degrees of freedom due to main effects and interaction in a 2×2 factorial experiment.

23.36 Discuss the chief properties of a factorial experiment. Explain the meanings of main effects and interaction effects in a 2^3. factorial experiment. Give a plan for analysis of data of such an experiment. (P.U., B.A/B.Sc. Hons. Part III, 1968)

23.37 The data of the following table are from a 2×2 factorial experiment. Partition the treatment sum of squares into *main effects* and *interaction component*. Interpret the results.

Treatment combinations			
a_0b_0	a_0b_1	a_1b_0	a_1b_1
6	12	26	21
14	14	17	16
8	13	21	20
9	11	30	17
7	13	27	21

23.38 A 4×4 Latin Square was laid out to investigate the effect of 4 treatment combinations (nitrogen at levels n_0 and n_1 and phosphates et levels p_0 and p_1) on the yield of potatoes. The field plan together with plot yields is as follows:

p 19	np 31	(1) 15	n 18
np 27	(1) 10	n 21	p 21
n 23	p 17	np 30	(1) 13
(1) 12	n 20	p 15	np 26

Perform a complete analysis of the data and give your interpretation of the results.

23.39 Eight treatments have been applied to a certain crop and their effects have been reflected in the following table with two replications:

(1)	a	b	ab	c	ac	bc	abc
13	10	12	14	9	11	8	7
14	11	13	15	10	12	9	8

Give a method of analysis and interpret your result. Discuss an estimate of error in this problem.

(P.U., B.A/B.Sc. Hons Part III, 1963)

23.40 An experiment was laid to see the effect of three fertilizers n, p and k at two levels each (*i.e.*, applied and not applied). The eight combinations were randomized in each replication. Set up an analysis of variance to test the significance of Main Effects and Interactions from the data given below:

Replication	Treatments							
	(1)	n	p	np	k	nk	pk	npk
1	24	25	24	24	29	30	22	32
2	30	31	31	27	39	34	25	23
3	23	33	27	23	31	29	24	37
4	28	26	24	27	36	36	29	34

(P.U., B.A/B.Sc. Hons. Part III, 1965)

23.41 The following data were obtained from a 2^3-factorial experiment repeated 3 times

Replication	Treatment Combinations							
	(1)	a	b	ab	c	ac	bc	abc
1	12	15	24	23	17	16	24	28
2	19	20	16	17	25	19	23	25
3	10	16	17	27	21	19	29	20

Evaluate the sum of squares for all factorial effects by (i) contrast method, (ii) Yates' method.

23.42 (a) Give the structure of the analysis of covariance for a Latin square design.

(b) Set up analysis of covariance table for the following data and obtain the adjusted treatment means.

Rows	Columns			
	1	2	3	4
	Y X	Y X	Y X	Y X
1	(A) 88 46	(C) 71 44	(D) 80 45	(B) 75 44
2	(B) 74 45	(A) 100 47	(C) 84 48	(D) 74 46
3	(C) 73 47	(D) 84 46	(B) 75 47	(A) 75 47
4	(D) 86 46	(B) 73 45	(A) 86 48	(C) 85 46

❖❖❖❖❖❖❖❖❖❖❖❖❖❖

24

Nonparametric Tests

24.1 INTRODUCTION

Most of the hypothesis testing procedures presented in earlier chapters are based on certain restrictive assumptions about the distribution and parameters of the population(*s*) from which samples are drawn. The most common assumption is that of normality. As these tests are based on the assumptions that the sampled populations are approximately normal with equal variances, they are called *parametric tests*. However, there are many situations, particularly in the social sciences, where normality assumption is in fact not met and the parametric tests cannot be carried out. In such instances, the techniques that can be used, are called *nonparametric tests*.

A *nonparametric test* is one that is performed without making any restrictive assumptions about the form of the population distribution and associated parameters. A nonparametric test is also called a *distribution-free test* because it makes no specific assumptions about the shape or form of the underlying population distribution except that the population distribution is continuous. The nonparametric tests, which provide useful alternatives to the parametric tests, have several advantages, some of them are given as follows:

(i) The nonparametric tests require no assumptions about the form or parameters of the underlying populations. In other words, they do not require that samples be drawn from normal populations with specified parameters.

(ii) The nonparametric tests are relatively easy to understand.

(iii) The nonparametric tests are computationally simple and relatively quick to apply as compared with the parametric tests.

(iv) The nonparametric tests may be applied to qualitative data or data that can be ranked but not measured exactly.

However, the nonparametric tests are generally less efficient as they do not use all the information available in the sample data. Another disadvantage is that they require a larger sample size than do the corresponding parametric tests.

There are several nonparametric tests, including the chi-square tests of goodness-of-fit and of independence in contingency tables already discussed in Chapter 17. The rank correlation coefficient that measures the strength of relationship between two sets of ranked data, is also the nonparametric alternative to the simple coefficient of correlation.

24.2 THE SIGN TEST

The sign test is perhaps the simplest and the oldest nonparametric test in use. As its name implies, it is based on the signs (pluses and minuses) of the observed differences. It is used to test the null hypothesis that the probability of a + sign equals the probability of a − sign, which is equivalent to testing the hypothesis, in case of one sample, that the *population median M, has a specified value, say* M_0, because each observation is equally likely to be above the median as below it; and in case of two samples, that the *two populations are identical* (which implies that the two populations have the same distributions and share the same mean and variance).

(i) **One Sample.** To perform the test, we replace each observation by a *plus* sign or a *minus* sign depending upon whether the observation is above or below M_0, the hypothesized value of the population median. We discard any observation that equals M_0 and reduce the sample size. We denote the total number of plus and minus signs by n. The test-statistic X is defined by the number of times the less frequent sign (plus or minus) occurs. Under our null hypothesis, the sampling distribution of X is binomial with parameters $\frac{1}{2}$ and n. We determine the critical region by calculating the binomial probabilities. To reach the significance level α, we add the probabilities from both tails in case of two-tailed test, and in case of one-tailed test, the probabilities in the desired tail are added to reach α. We accept or reject the hypothesis in the usual manner. The null hypothesis in case the populations are symmetric, may be stated as $H_0 : \mu = \mu_0$.

(ii) **Two Samples.** Let X_i and Y_i denote the observations from the first sample and the second sample respectively. We replace the difference $X_i - Y_i$ by a *plus* sign if $X_i > Y_i$; by a *minus* sign if $X_i < Y_i$ and we ignore the pair if $X_i = Y_i$, *i.e.*, zero differences are dropped from the analysis. Let n represent the number of plus and minus signs and let X

stand for the number of times the less frequent sign (plus or minus) occurs. Then the sampling distribution of X is binomial with parameters $\frac{1}{2}$ and n. The rest of the procedure is the same as in one sample sign test. In case the sample sizes are not equal, some of the values of the larger sample are to be discarded (The data must be from matched pair samples). The two-sample sign test may be used to test the hypothesis $H_0 : \mu_1 = \mu_2$ when the underlying populations are assumed to be symmetric.

With large n, we use the normal approximation to the binomial distribution $b(n, \frac{1}{2})$. The statistic X is then approximately standard normal with mean $= \frac{n}{2}$ and standard deviation $= \sqrt{\frac{n}{4}}$. In other words, the test-statistic under H_0, becomes

$$Z = \frac{X - n/2}{\sqrt{n/4}}, \quad \text{without correction for continuity,}$$

or

$$Z = \frac{(X \pm \frac{1}{2}) - \frac{n}{2}}{\sqrt{n/4}}, \quad \text{with correction for continuity.}$$

We reject or accept H_0, applying the usual decision rules. In applying normal approximation to binomial distribution, n is taken *large* when both np and nq are at least 5. As $p = \frac{1}{2}$, we can therefore use the normal approximation when n exceeds 10.

Another appropriate statistic to test the hypothesis $H_0 : P$ [+ sign] $= P$ [−sign] $= \frac{1}{2}$ is the chi-square statistic, given by

$$\chi^2 = \frac{(n_1 - n_2)^2}{n_1 + n_2}, \quad \text{with 1 } df;$$

where n_1 and n_2 represent the number of plus and minus signs respectively. The sign test is simple and easy to apply. It is used in situations where the Student's t-test is not applicable. However, the t-test, when applicable, is more efficient and more powerful than a sign test. Another disadvantage of the sign test is that it considers only the signs of the differences and not their magnitude.

The *procedure* for testing the hypothesis that the population median has a specified value M_0, in case of one sample, is given below:

(i) Formulate the null and alternative hypotheses as

H_0 : Population Median $M = M_0$

H_1 : Population Median $M \neq M_0$ (or $M > M_0$ or $M < M_0$).

(ii) Decide on significance level α.

(iii) The test-statistic is X, the number of times the less frequent sign (+ or −) occurs and is binomially distributed.

If n, the number of pluses and minuses exceeds 10, the test-statistic would become

$$Z = \frac{X - n/2}{\sqrt{n/4}}$$

which, if H_0 is true, is approximately standard normal.

(iv) Computations. Subtract M_0, the hypothesized value of population median, from each observation of the sample, i.e. find the differences $X_i - M_0$. Write down a plus sign if the difference is positive and a minus sign if the difference is negative. Ignore zero differences, if any. Denote by n the number of plus and mins signs (*i.e.*, non-zero differences) and by X, the number of times the less frequent sign occurs. Compute either the extreme probabilities of the binomial variable X or the value of Z, as the case may be.

(v) The critical region depends on the test-statistic, alternative hypothesis and significance level α.

(vi) Apply the usual decision rule to reject or accept the null hypothesis.

The *procedure* in the case of two–sample sign test would be the same except the following two steps:

(i) H_0 : The two populations are identical or that they have equal medians, $M_1 = M_2$. It is tested against an appropriate alternative hypothesis.

(ii) Computations. Subtract each observation of the second sample, say Y_i, from the corresponding observation of the first sample, say X_i *i.e.* find the differences $X_i - Y_i$. Write a plus sign if the difference is positive and a minus sign if the difference is negative. Discard zero differences, if any; and so on.

Example 24.1. Use the sign test to test the hypothesis that the median of the population from which the following data is a random sample, equals 30 against the alternative that it does not: 27, 39, 30, 22, 32, 24, 25, 29, 26.

(i) We set up our hypotheses as

H_0 : Population median, $M = 30$ and $H_1 : M \neq 30$.

(ii) We specify the significance level at $\alpha = 0.05$

(iii) The test-statistic to be used is X, the number of times the less frequent sign occurs.

(iv) Computations. Subtracting 30, the hypothesized value of median from each observation and writing down the signs, we get

$$- \quad + \quad 0 \quad - \quad + \quad - \quad - \quad - \quad -$$

Now $n = 8$ as zero is ignored and $X = 2$, the number of plus signs (less frequent). Under H_0, the sampling distribution of X is binomial with $p = \frac{1}{2}$ and $n = 8$.

$$\therefore P(X \leq 2) = \left(\tfrac{1}{2}\right)^8 + 8\left(\tfrac{1}{2}\right)^8 + 28\left(\tfrac{1}{2}\right)^8 = \frac{37}{256} = 0.145$$

(v) Critical region. For a two-tailed test, to reject H_0, the computed probability should be less than 0.025.

(vi) **Conclusion.** Since the computed probability is more than $\alpha/2 = 0.025$, we therefore accept H_0 and conclude that the median of the population equals 30.

Example 24.2. An experimenter wants to determine the effectiveness of a certain reducing diet. Twelve persons were put on diet; their weights before and after they tried the diet, are shown below:

Person	1	2	3	4	5	6	7	8	9	10	11	12
Weight before	202	154	183	180	228	164	139	165	175	245	237	162
Weight after	195	154	178	199	220	157	135	180	198	206	227	155

Use the sign test, at the 5% significance level to test the hypothesis that the diet is not effective against the alternative that it is effective.

(i) The null and the alternative hypotheses are

H_0: The diet is not effective, which is equivalent to testing the hypothesis $H_0 : P\,[+\text{sign}] = P\,[-\text{sign}] = \frac{1}{2}$, and

H_1: The diet is effective, *i.e.* $p > \frac{1}{2}$, (one tailed test)

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic to be used is X, the number of times the less frequent sign occurs. Under H_0, X is binomially distributed.

(iv) **Computations.** Subtracting the weights *after* from the weights *before* they tried the diet, and writing down a plus sign for each positive difference, and a minus sign for each negative difference, we get

$$+ \quad 0 \quad + \quad - \quad + \quad + \quad + \quad - \quad - \quad + \quad + \quad +$$

Thus $n = 11$, the sum of plus and minus signs, ignoring zero difference, and $X = 3$, the number of minus signs (less frequent signs). The statistic X is the distribution of $-$signs, therefore

$$P(X \leq 3) = \left(\tfrac{1}{2}\right)^{11} + 11\left(\tfrac{1}{2}\right)^{11} + 55\left(\tfrac{1}{2}\right)^{11} + 165\left(\tfrac{1}{2}\right)^{11}$$

$$= \frac{232}{2048} = 0.113$$

(v) **Critical region.** For a one-tailed test, the computed probability is required to be less than 0.05.

(vi) **Conclusion.** Since the computed probability is more than 0.05, so we do not reject H_0.

Alternative Method (a) Since n, the number of plus and minus signs, exceeds 10, therefore the sign test can also be carried out, using normal approximation to the binomial distribution. With $\mu = np = 11 \times \frac{1}{2} = 5.5$ and $\sigma = \sqrt{npq} = \sqrt{11 \times \frac{1}{2} \times \frac{1}{2}} = 1.66$, we find

$$z = \frac{X + \frac{1}{2} - n/2}{\sqrt{n/4}}, \text{ (with continuity correction)}$$

$$= \frac{3.5 - 5.5}{1.66} = -1.20,$$

which is less than $z_{0.05} = 1.645$. Hence we cannot reject H_0, the same conclusion as obtained by applying the binomial distribution.

(b) The sign test in this case can also be carried out by the method of chi-square. Thus we have

$$\chi^2 = \frac{(n_1 - n_2)^2}{n_1 + n_2},$$

where n_1 and n_2 denote the number of plus and minus sign. Substitution gives

$$\chi^2 = \frac{(8-3)^2}{8+3} = \frac{25}{11} = 2.27$$

This value does not exceed $\chi^2_{0.05(1)} = 3.84$ and therefore the null hypothesis cannot be rejected.

24.3 THE WILCOXON SIGNED-RANK TEST FOR THE PAIRED OBSERVATIONS

The *Wilcoxon signed-rank test* was proposed by Frank Wilcoxon in 1945 and is named after him. This test is applied to paired differences when the assumption of normality is suspect. The test is more powerful than the *sign test* because it makes use of the magnitude of the differences between paired comparisons. The null hypothesis to be tested is that the two populations from which the respective members of matched pairs come, are identical. To perform the test, we rank the *absolute* values of the differences between the pairs, ignoring zero differences, if any, and in the case of ties, assigning to each of such differences the average of the ranks that would have been assigned, had they differed slightly. We then compute the sum of the ranks assigned to the positive differences and the sum of the ranks assigned to the negative differences. The test statistic, denoted by T, is based on the *smaller* sum of ranks with the sign ignored. If H_0 is true, we expect the two sums to be about equal. A large difference between the two sums provides evidence against H_0. Accordingly, we reject H_0 when the computed value of T is less than or equal to the critical value at the specified level of significance, given in a table titled, "Critical values for the Wilcoxon Signed-Rank Test". (Table 24.1). It is to be noted that in Z, χ^2, t and F tests, it is the larger values that generally provide evidence against the null hypothesis. The Wilcoxon signed-rank test is also applicable to test the hypothesis that the median of a population of differences equals some specified value, M_0, *i.e.*, $H_0 : |(X - Y) - M_0| = 0$.

With *large* n (say $n \geq 25$), it has been shown that the signed rank statistic T is approximately normally distributed with a mean

$$\mu_T = \frac{n\,(n+1)}{4},$$

and a standard deviation

$$\sigma_T = \sqrt{\frac{n\,(n+1)\,(2n+1)}{24}}$$

where n is the number of matched pairs, excluding zero differences. Hence, when $n \geq 25$, we generally use the test-statistic

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{T - \dfrac{n\,(n+1)}{4}}{\sqrt{\dfrac{n\,(n+1)\,(2n+1)}{24}}}$$

which, if H_0 is true, is approximately standard normal. We accept or reject H_0, applying the usual decision rules.

The test-*procedure* is given below:

(i) Formulate the null and alternative hypotheses as

H_0 : The two populations are identical, which is equivalent to testing the hypothesis that the *medians* of the two populations are equal or the two means are equal in the case of symmetrical populations; and

H_1 : The two populations are not identical or the two medians are not equal or one median is larger than the other, etc.

(ii) Decide on significance level α.

(iii) The test-statistic is T, the smaller sum of ranks with the sign ignored.

(iv) Computations. Find the magnitude of the differences between the paired values. Arrange the non-zero differences in order of increasing *absolute* values (i.e., discarding algebraic signs). Assign the ranks 1, 2, 3, etc. to the ranked differences. If ties occur, then assign to each such difference the average of the ranks that would have been assigned if they had differed slightly. Assign to each rank the sign of the original difference. Find the sum of the positive ranks and the negative ranks separately. Take the smaller sum of the ranks with the sign ignored as T. (In case of large n, calculate the value of Z-statistic).

(v) Critical value of T is given in the Wilcoxon signed-rank table **(Table 24.1)** for the value of n (number of matched pairs excluding zero differences) at the chosen level of significance α. This is the value *at or below* which lies the critical region.

(vi) Decide as under:

Reject H_0 if observed value of $T \leq$ the critical value, otherwise accept it. (In case of Z-statistic, reject or accept H_0, applying the usual decision rule).

The Wilcoxon Signed-Rank test is also applicable when we are given a random sample and we wish to test the null hypothesis that the population median is equal to some specified value, M_0. In this case, we match every observation in the sample with the hypothesized value M_0 and find the differences $X-M_0$. If the null hypothesis is true, the sum of the signed ranks should be close to zero. The rest of the procedure is the same as specified in the case of the paired differences.

Table 24.1. Critical Values of T for the Wilcoxon Signed-Rank Test

Pairs n	Two-tailed Test: $\alpha=0.10$; One-tailed Test: $\alpha=0.05$;	$\alpha=0.05$; $\alpha=0.025$;	$\alpha=0.02$; $\alpha=0.01$;	$\alpha=0.01$; $\alpha=0.005$;
6	2	0	--	--
7	3	2	0	--
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15
16	35	29	23	19
17	41	34	27	23
18	47	40	32	27
19	53	46	37	32
20	60	52	43	37
21	67	58	49	42
22	75	65	55	48
23	83	73	62	54
24	91	81	69	61
25	100	89	76	68

Reject H_0 if the calculated value of $T \leq$ table value at the chosen value of α. n is the number of non-zero matched pairs.

Example 24.3. Ten young recruits were put through a strenuous physical training programme by the Army. Their weights were recorded before and after the training with the following results:

Recruit	1	2	3	4	5	6	7	8	9	10
Weight *before*	125	195	160	171	140	201	170	176	195	139
Weight *after*	136	201	158	184	145	195	175	190	190	145

Use the Wilcoxon signed-rank statistic to test the hypothesis that the programme affects average weights of recruits. Let $\alpha = 0.05$

(i) We state our null and alternative hypotheses as

H_0: The programme does not affect the average weight of recruits, *i.e.* $\mu_1 - \mu_2 = 0$;

H_1: The programme affects the average weight of recruits, *i.e.* $\mu_1 - \mu_2 \neq 0$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is T, the smaller sum of ranks with the sign ignored.

(iv) Computations are shown below:

Recruit	Weight before (X)	Weight after (Y)	Difference $Y-X$	Rank of $\|Y-X\|$	Signed rank Positive	Signed rank Negative
1	125	136	+11	8	+8	
2	195	201	+ 6	6	+6	
3	160	158	− 2	1		−1
4	171	184	+13	9	+ 9	
5	140	145	+ 5	3	+ 3	
6	201	195	− 6	6		−6
7	170	175	+ 5	3	+ 3	
8	176	190	+14	10	+10	
9	195	190	− 5	3		−3
10	139	145	+ 6	6	+ 6	
Sum					45	−10=T

(v) Looking in the table for the Wilcoxon signed-rank test (Table 24.1), we find that for $n=10$ at $\alpha=0.05$ for Two-tailed test, the critical value of $T=8$ at or below which lies the critical region.

(vi) **Conclusion.** Since the observed value of $T=10$ does not lie in the critical region, so we cannot reject our null hypothesis. The data do not provide sufficient evidence to indicate that the programme affects average weight.

Example 24.4. Given that the nine observations below are a random sample from a continuous, symmetric population, use the Wilcoxon signed-rank test, at the 5% level, to test the null hypothesis that the median equals 30 against the alternative that it is less:

27, 39, 30, 22, 32, 24, 25, 29 and 26.

(i) We state our null and alternative hypotheses as

H_0 : Population Median, $M = 30$ and $H_1 : M < 30$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to be used is T, the smaller sum of ranks with the sign ignored.

(iv) Computations:

Observation X	Difference $X - M_0$	Rank of $\lvert X - M_0 \rvert$	Signed Rank (+)	Signed Rank (−)
27	−3	3		3
39	+9	8	8	
30	0	—		
22	−8	7		7
32	+2	2	2	
24	−6	6		6
25	−5	5		5
29	−1	1		1
26	−4	4		4
Sum			$T=10$	26

(v) Looking in the table for the Wilcoxon signed-rank test, we find that for $n=8$ (zero difference is ignored) at $\alpha=0.05$ for one-tailed test, the critical value of $T=5$, at or below which lies the critical region.

(vi) **Conclusion.** Since the computed value of $T=10$ is larger than the critical value, so the null hypothesis that the population median equals 30 cannot be rejected.

24.4 THE WILCOXON RANK-SUM TEST FOR INDEPENDENT SAMPLES

The Wilcoxon signed-rank test for matched pairs cannot be performed when either the two samples are independent or the two samples have different sizes. For either of these situations, another non-parametric test, known as *Wilcoxon rank-sum test* is used. The null hypothesis to be tested is that the two samples come from identical populations or that the medians of two populations are equal, against some suitable alternative. The test-statistic is denoted by R, which is the sum of the ranks assigned to the observations of the smaller sample. To carry out the test, we combine the observations of the two independent samples of size n_1 and n_2 ($n_1 \le n_2$) from two populations, assumed continuous and symmetric. We arrange the $n_1 + n_2$ observations in order of increasing magnitude and assign the ranks $1, 2, \ldots, n_1 + n_2$ to them. In the case of ties, assign the average of the ranks the tied observations occupy. We denote the sum of the ranks assigned to sample 1 (smaller sample) by R. This R is the *rank-sum statistic*. The extreme values of R provide evidence to reject the null hypothesis. The critical values of R for Wilcoxon rank-sum test for a 5% level of significance, one-or-two-tailed test are shown in Table 24.2. We reject H_0, if the

computed value of R is less than or equal to the smaller table value or greater than or equal to the larger table value for the given values of n_1 and n_2.

It has been shown that when either sample size (or both of them) exceeds 10, the statistic R is approximately normally distributed with mean

$$\mu_R = \frac{n_1 (n_1 + n_2 + 1)}{2}$$

and standard deviation $$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

The test-statistic would then become

$$Z = \frac{R - \mu_R}{\sigma_R} = \frac{R - \dfrac{n_1 (n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}},$$

which, if H_0 is true, is approximately standard normal. We accept or reject H_0, applying the usual decision rules. It has been further shown that, if the two populations are normal, the Wilcoxon Rank-Sum test is almost efficient as the two-sample t-test.

Table 24.2 Critical Values of R for the Wilcoxon Rank Sum Test

In this table there are four numbers for each value of n_1 and n_2. The top pair are the critical values for a two-tailed test for $\alpha = 0.05$, while the lower are the critical values for one-tailed test for $\alpha = 0.05$.

n_1 \ n_2	4	5	6	7	8	9	10
4	10 26 11 25	11 29 12 28	12 32 13 31	13 35 14 34	14 38 15 37	15 41 16 40	15 45 17 43
5		17 38 19 36	18 42 20 40	20 45 21 44	21 49 23 47	22 53 24 51	23 57 26 54
6			26 52 28 50	27 57 29 55	29 61 31 59	31 65 33 63	32 70 35 67
7				36 69 39 66	38 74 41 71	40 79 43 76	42 84 45 81
8					49 87 51 85	51 93 54 90	53 99 56 96
9						63 105 66 105	65 115 69 111
10							78 132 82 128

The *procedure* for testing H_0 by means of the Wilcoxon rank-sum-test is given below:

(i) State the null and alternative hypotheses as

H_0: The two samples come from identical populations, or the medians of two populations are equal;

H_1: The two samples do not come from identical populations, etc.

In the case of one-tailed test, we may state that $M_1 > M_2$, etc.

(ii) Decide on a significance level α (generally $\alpha = 0.05$).

(iii) The test-statistic to be used is R, the sum of the ranks of the smalier sample. In case of equal sample sizes, either rank sum can be used.

(iv) Computations. Combine the two samples and arrange the observations in increasing order of magnitude. For convenience, underline the observations from sample 1. Assign the ranks 1, 2, ..., $n_1 + n_2$, with ties being assigned the average of the ranks that they occupy. Find R by adding together the ranks of the underlined observations, (when n_1 and n_2 exceed 10, calculate the value of Z statistic).

(v) Critical values of R (for $\alpha = 0.05$) are looked in Table 24.2 for the given values of n_1 and n_2.

(vi) Decide as below:

Reject H_0 if the calculated value of $R \leq$ the smaller table value or $R \geq$ the larger table value, accept H_0 otherwise.

Example 24.5. The two samples below are from two populations, assumed to be identical. Test the null hypothesis, using the Wilcoxon rank-sum test, that the two medians are equal against the alternative that the median of the first population is greater than that of the second population.

Sample 1	38, 49, 45, 29, 31, 35
Sample 2	31, 42, 22, 26, 43, 37, 25, 30, 47

(i) The null and alternative hypotheses are

$H_0: M_1 = M_2$ and $H_1: M_1 > M_2$,

where M_1 and M_2 are the medians of the first and secon population respectively.

(ii) We choose the significance level at $\alpha = 0.05$.

(iii) The test-statistic to use is R, the sum of the ranks of sample 1.

(iv) **Computations.** Arranging the observations of combined samples in order of increasing magnitude and underlining the observations from sample 1 (a good way to distinguish them), we get

22, 25, 26, <u>29</u>, 30, <u>31</u>, 31, <u>35</u>, 37, <u>38</u>, 42, 43, <u>45</u>, 47, <u>49</u>.

Assigning the ranks to these observations (underlining the ranks of the observations from sample 1), we get

1, 2, 3, <u>4</u>, 5, <u>6.5</u>, 6.5, <u>8</u>, 9, <u>10</u>, 11, 12, <u>13</u>, 14, <u>15</u>

Adding the underlined ranks, we get $R = 56.5$.

(v) Looking in **Table 24.2**, we find for $n_1 = 6$, $n_2 = 9$, the critical region which consists of all values of $R \geq 63$ (one tailed test, lower pair).

(vi) **Conclusion.** Since the computed value of $R = 56.5$ does not fall in the critical region, we therefore cannot reject our null hypothesis.

Example 24.6. Given the two samples below, use the Wilcoxon rank-sum test to test the null hypothesis that the population medians are equal against the alternative that they are not equal. Let $\alpha = 0.05$

Sample 1	40, 35, 44, 42, 46, 28, 39, 50, 37, 45, 27, 35
Sample 2	32, 34, 49, 37, 48, 49, 48, 51, 45, 44, 34, 36, 50, 49, 37

(i) We state our null and alternative hypotheses as

$H_0 : M_1 = M_2$ and $H_1 : M_1 \neq M_2$,

where M_1 and M_2 are the medians of the first and second population respectively.

(ii) The significance level is set at $\alpha = 0.05$

(iii) As both n_1 and n_2 exceed 10, the test-statistic to use is

$$Z = \frac{R - \mu_R}{\sigma_R}$$

which, if H_0 is true, is approximately standard normal.

(iv) **Computations.** Arranging the observations of the combined samples in order of increasing magnitude and underlining (for identification) the observations from sample 1, we get

<u>27</u>, <u>28</u>, 32, 34, 34, <u>35</u>, <u>35</u>, 36, <u>37</u>, 37, 37, <u>39</u>, <u>40</u>, <u>42</u>, <u>44</u>, 44, <u>45</u>, 45, <u>46</u>, 48, 48, 49, 49, 49, <u>50</u>, 50, 51

Replacing these observations with their ranks (with the ranks of observations from sample 1 underlined), we get

$\underline{1}, \underline{2}, 3, 4.5, 4.5, \underline{6.5}, \underline{6.5}, 8, \underline{10}, 10, 10, \underline{12}, \underline{13}, \underline{14}, \underline{15.5}, 15.5, \underline{17.5}, 17.5, \underline{19}, 20.5, 20.5, 23, 23, 23, \underline{25.5}, 25.5, 27.$

Adding the underlined ranks, we get $R = 142.5$.

Now $n_1 = 12$ and $n_2 = 15$, therefore

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{12\,(12 + 15 + 1)}{2} = 168, \text{ and}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(12)\,(15)\,(12 + 15 + 1)}{12}} = 20.49.$$

$$\therefore z = \frac{R - \mu_R}{\sigma_R} = \frac{142.5 - 168}{20.49} = -1.24$$

(v) The critical region is $|Z| \geq 1.96$.

(vi) **Conclusion.** Since the computed value of $z = -1.24$ does not fall in the critical region, we therefore cannot reject our null hypothesis of equal medians.

24.5 THE MANN-WHITNEY U TEST

The *Mann-Whitney U test* is a nonparametric alternative to the Student's two-sample t test which requires random sampling from normal population with equal variances. The test is based on ranks and is used to determine whether or not two *independent* samples of size n_1 and n_2 come from populations having identical distributions. The null hypothesis to be tested is that the two populations are identical. To carry out the test, we arrange all the $n_1 + n_2$ observations of the combined samples in order of increasing magnitude and assign the ranks 1, 2, ..., $n_1 + n_2$ to them. In the case of ties, we assign the average of the tied ranks. We add the ranks assigned to observations in sample 1 and denote this sum by R_1. Similarly, we calculate the sum of ranks of sample 2 and denote it by R_2. We then find, for both samples, the values of U, (the statistic used in this test) as below:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1, \text{ and}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2.$$

We choose the smaller of the two values found for U_1 and U_2 as the value of the U statistic for the Mann–Whitney test *i.e.* $U = \min\,[U_1, U_2]$. We reject our null hypothesis if the calculated value of U is either $\leq$ the

smaller value or is ≥ the larger value given in a table, titled: Critical values of U for the Mann-Whitney test. (Table 24.3)

It is interesting to note that the sum of two U's will always be equal to the product of the two sample sizes, *i.e.* $U_1 + U_2 = n_1 n_2$ and the smaller value of U is always less than $\frac{n_1 n_2}{2}$. This provides a good check on the correctness of the rankings. In practice, we need to calculate either U_1 or U_2, the other can be obtained from the relation given above. It has been shown that, when H_0 is true, and n_1 and n_2 are both greater than 8, the statistic U is approximately normally distributed with mean

$$\mu_U = \frac{n_1 n_2}{2},$$

and standard deviation $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$

The test-statistic would then become

$$Z = \frac{U - \mu_U}{\sigma_U},$$

which, if H_0 is true, is approximately standard normal.

We decide whether to accept or reject the hypothesis by applying the usual decision rules.

Table 24.3. Critical Values of U for the Mann-Whitney Test

In this table there are four numbers for each value of n_1 and n_2. The top pair are the critical values for a two-tailed test for $\alpha = 0.05$, while the lower pair are the critical values for one-tailed for $\alpha = 0.05$.

n_1 \ n_2	4		5		6		7		8	
4	0 1	16 15	1 2	19 18	2 3	22 21	3 4	25 24	4 5	28 27
5			2 4	23 21	3 5	27 25	5 6	30 29	6 8	34 32
6					5 7	31 29	6 8	36 34	8 10	40 38
7							8 11	41 38	10 13	46 43
8									13 15	51 49

The Mann-Whitney U Test is also applied to an *ordered* $2 \times k$ contingency table in which two samples are classified into ranked categories or groups such as poor, fair, good, very good, excellent or so

0–20, 21–40, 41–60, 61–80, over 80, etc. In such a situation, we regard all observations (frequencies) in the same category or group as ties. We determine the average ranks for each category or group by the expression

$$r_j = \frac{T_j + 1}{2} + C, \text{ where}$$

r_j = average rank of the jth category or group,

T_j = total number of observations in the jth category or group, and

C = cumulative total of the category preceding the jth category.

We compute the sum of ranks for each sample by multiplying each average rank by the number of observations in that rank. The rest of the procedure is the same.

The following examples illustrate how the Mann-Whitney U test is used.

Example 24.7. Two plastics, each produced by a different process, were tested for ultimate strength. The measurements shown below represent breaking load in units of 1000 pounds-per-square inch.

Plastic 1	15.3,	18.7,	22.3,	17.6,	15.1,	14.8,
Plastic 2	21.2,	22.4,	18.3,	19.3,	17.1,	27.7,

Use the Mann-Whitney U test to test the hypothesis of no difference in the distributions of strengths for the two plastics at the $\alpha = 0.05$ level of significance.

(i) We state our null and alternative hypotheses as

H_0 : There is no difference in the distributions of strength, and

H_1 : The two distributions of strength are different.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is U, the smaller of the two values.

(iv) Computations. Arranging the observations of the combined samples in order of increasing magnitude and underlining the observations from sample 1, we get

<u>14.8</u>, <u>15.1</u>, <u>15.3</u>, 17.1, <u>17.6</u>, 18.3, <u>18.7</u>, 19.3, 21.2, <u>22.3</u>, 22.4, 27.7.

Replacing these values with their ranks (with the ranks of observations from sample 1 underlined), we get

<u>1</u>, <u>2</u>, <u>3</u>, 4, <u>5</u>, 6, <u>7</u>, 8, 9, <u>10</u>, 11, 12.

Totalling the ranks of the two samples separately, we have

$R_1 = 1 + 2 + 3 + 5 + 7 + 10 = 28$, and

$R_2 = 4 + 6 + 8 + 9 + 11 + 12 = 50$.

Now $U_1 = n_1 n_2 + \dfrac{n_1 (n_1 + 1)}{2} - R_1 = 6 \times 6 + \dfrac{(6)(7)}{2} - 28 = 29$, and

$U_2 = n_1 n_2 + \dfrac{n_2 (n_2 + 1)}{2} - R_2 = 6 \times 6 + \dfrac{(6)(7)}{2} - 50 = 7$.

The smaller of the two values for U_1 and U_2 is taken as U statistic, *i.e.*, $U = \min [29, 7] = 7$.

(v) Looking in **Table 24.3**, we find for $n_1 = 6$, $n_2 = 6$ and $\alpha = 0.05$ the critical region consists of all values of $U \leq 5$ and of all values of $U \geq 31$.

(vi) **Conclusion.** Since the computed value of $U=7$ does not fall in the critical region, so we are unable to reject the hypothesis of no difference in the distributions of strength.

Example 24.8. Given below are the grade point averages received by two groups of students.

Group 1	3.1, 5.3, 6.4, 6.2, 3.8, 7.5, 5.8, 4.3, 5.9, 4.9
Group 2	9.0, 5.6, 6.3, 8.5, 4.6, 7.1, 5.5, 7.9, 6.8, 5.5, 8.9

Test by the Mann-Whitney U test the hypothesis that the two groups come from identical populations. Use a 0.05 significance level.

(i) We state our null and alternative hypotheses as

H_0 : The two groups come from identical populations or equivalently, the grade point averages are equal; and

H_1 : The two populations are not identical.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is U, which becomes

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{U - \dfrac{n_1 n_2}{2}}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}},$$

as n_1 and n_2 are both greater than 8; and Z is approximately standard normal.

(iv) Computations. Arranging the observations of the combined groups in order of increasing magnitude and underlining the observations from Group 1, we get

3.1, 3.8, 4.3, 4.6, 4.9, 5.3, 5.5, 5.5, 5.6, 5.8, 5.9, 6.2,
6.3, 6.4, 6.8, 7.1, 7.5, 7.9, 8.5, 8.9, 9.0.

The corresponding ranks are

1, 2, 3, 4, 5, 6, 7.5, 7.5, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
19, 20, 21

Totalling the ranks of the two groups separately, we get

$$R_1 = 1 + 2 + 3 + 5 + 6 + 10 + 11 + 12 + 14 + 17 = 81, \text{ and}$$

$$R_2 = 4+7.5+7.5+9+13+15 + 16 + 18 + 19 + 20 + 21 = 150$$

Now the value of the test-statistic U is the smaller of

$$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1 = (10)(11) + \frac{10(10+1)}{2} - 81 = 84$$

$$U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2 = (10)(11) + \frac{11(11+1)}{2} - 150 = 26$$

or $\quad U_2 = n_1 n_2 - U_1 = 110 - 84 = 26$

Thus $U = \min [84, 26] = 26$. The mean and standard deviation of the sampling distribution of U are

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(10)(11)}{2} = 55, \text{ and}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(10)(11)(10 + 11 + 1)}{12}} = 14.2$$

Hence $z = \dfrac{U - \mu_U}{\sigma_U} = \dfrac{26 - 55}{14.2} = -2.04$

(v) The critical region is $|Z| \geq z_{0.025} = 1.96$

(vi) **Conclusion.** Since the computed value of $z = -2.04$ falls in the critical region, so we reject H_0 and conclude that the grade point averages differ significantly.

Example 24.9. The following table shows the scores received by male and female students in a certain standard test:

Score	0–4	5–9	10–14	15–19	20–25	Total
Male	3	8	17	12	6	46
Female	1	7	25	4	3	40

Apply the Mann-Whitney U test to check whether the mean scores of male and female students are the same at $\alpha = 0.05$.

(i) We state our null and alternative hypotheses as

$H_0: \mu_m = \mu_f$, *i.e.* the mean scores of male (μ_m) and female (μ_f) students are the same; and

$H_0: \mu_m \neq \mu_f$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is U, which becomes

$$Z = \frac{U - \mu_U}{\sigma_U},$$

as n_1 and n_2 are both greater than 8, and Z is approximately standard normal.

(iv) Computations. Regarding all observations in each category or group as ties, we determine their average ranks. We compute the sum of ranks for the sample of male students by multiplying each average rank by the number of observations in that rank. These computations are shown below:

Score	0–4	5–9	10–14	15–19	20–25	Total
Male	3	8	17	12	6	46
Female	1	7	25	4	3	40
Total	4	15	42	16	9	86
Cumulative Total	4	19	61	77	86	--
Average Rank	$\frac{4+1}{2}+0 = 2.5$	$\frac{15+1}{2}+4 = 12$	$\frac{42+1}{2}+19 = 40.5$	$\frac{16+1}{2}+61 = 69.5$	$\frac{9+1}{2}+77 = 82$	--
Total of ranks for male	$2.5\times3 = 7.5$	$12\times8 = 96$	$40.5\times17 = 688.5$	$69.5\times12 = 834$	$82\times6 = 492$	$R_1 = 2118$

Now $\quad U_1 = n_1 n_2 + \dfrac{n_1(n_1+1)}{2} - R_1 = (46)(40) + \dfrac{46(46+1)}{2} - 2118$

$$= 1840 + 1081 - 2118 = 803, \text{ and}$$

$$U_2 = n_1 n_2 - U_1 = 1840 - 803 = 1037$$

$$\therefore \quad U = \min[U_1, U_2] = 803$$

The mean and standard deviation of the sampling distribution of U are

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(46)(40)}{2} = 920, \text{ and}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(46)(40)(46 + 40 + 1)}{12}}$$

$$= \sqrt{13340} = 115.5$$

Hence $z = \dfrac{U - \mu_U}{\sigma_U} = \dfrac{803 - 920}{115.5} = -1.01$

(v) The critical region is $|Z| \geq z_{0.025} = 1.96$

(vi) **Conclusion.** Since the computed value of $z = -1.01$ does not fall in the critical region, so we cannot reject H_0. We conclude that the mean scores of male and female students are the same.

24.6 THE MEDIAN TEST (TWO OR MORE SAMPLES)

The *median test* is a nonparametric test used to determine whether $k(k \geq 2)$ independent random samples, which may be of unequal size, are taken from populations with the same medians. The null hypothesis, in case of two samples, may be formulated as H_0 : Median$_1$ = Median$_2$, *i.e.*, the two populations have the same medians. The test is based on the principle that about half of each sample's observations will be above and about half will be below the median, when H_0 is true. To perform the test, we combine and arrange the sample observations in order of increasing magnitude and find the median of the combined observations. For each sample, we count the number of observations that are above or below the median of the combined data, ignoring the observations that are equal to median. We set out the numbers above and below median for each sample in a 2×2 (or a 2×k, in case of k samples) contingency table like the following:

	Sample 1	Sample 2
Above median		
Below median		

The rest of the procedure is the same as that of chi-square test of independence in contingency tables. Fisher's exact probability formula should be used when the total number of observations in a 2 × 2 table is less than 20. The median test is easy and simple to apply, but it is less efficient.

Example 24.10. An identical examination was administered to two classes of students. One class contains 13 students and the other contains 17 students. The test scores of the 30 students are given below:

Class I	54, 65, 66, 71, 73, 78, 78, 80, 82, 87, 92, 93, 95.
Class II	51, 53, 54, 61, 64, 66, 67, 69, 71, 74, 76, 80, 81, 85, 89, 90, 94

At $\alpha = 0.05$, use the median test to test the hypothesis that the two classes of students are from populations with an identical median.

(i) We formulate our null and alternative hypotheses as

H_0: The median scores for class I are equal to the median scores for class II, and

H_1: The median scores for class I differ from the median scores for class II.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the chi-square statistic with 1 degree of freedom.

(iv) Computations. Arranging the combined observations in order of increasing magnitude and locating the median of the combined data, we find that median = 75. Counting the observations that are above or below 75 in each class, we find that there are 8 observations above and 5 observations below the median in class I; and there are 7 observations above and 10 observations below the median in class II. This information produces the following 2×2 contingency table:

	Class I	Class II	Total
Above median	8	7	15
Below median	5	10	15
Total	13	17	30

Now
$$\chi^2 = \frac{n\,(bc - ad)^2}{(a + c)\,(b + d)\,(a + b)\,(c + d)}$$
$$= \frac{(30)\,[(7)\,(5) - (8)\,(10)]^2}{13 \times 17 \times 15 \times 15} = 1.425$$

(v) The critical region is $\chi^2 \geq \chi^2_{0.05(1)} = 3.84$.

(vi) **Conclusion.** Since the calculated value of $\chi^2 = 1.425$ does not fall in the critical region, so we cannot reject the null hypothesis. We may conclude that the two classes of students are from populations with an identical median.

24.7 THE RUNS TEST FOR RANDOMNESS

The *runs test* is a non-parametric test used to determine whether or not the numbers in the sample form a *random sequence*. The test for randomness is based on the total number of runs. A *run* is defined as

subsequence of one or more identical symbols (or letters) that are immediately followed and preceded by different symbols (or letters) or by no symbols at all. For example, in the following sequence of plus and minus signs

$$\underbrace{+\ +}_{\text{run 1}}\ \underbrace{-\ -\ -}_{\text{run 2}}\ \underbrace{+}_{\text{run 3}}\ \underbrace{-\ -}_{\text{run 4}}\ \underbrace{+\ +\ +}_{\text{run 5}}\ \underbrace{+}\ \underbrace{-}_{\text{run 6}}$$

the number of runs is 6. The first run consists of two plus signs (*i.e.*, + +) as it is preceded by no symbol and is followed by a symbol of different kind. The second run contains 3 minus signs (*i.e.*, − − −) as it is preceded and followed by different symbols, and so on. Similarly, the following sequence of 6 Heads and 5 Tails

$$H \quad T \quad T \quad T \quad T \quad H \quad H \quad H \quad T \quad H \quad H$$

has 5 runs. Too few runs or too many runs in a sequence present evidence of non randomness.

The runs test, may be used to test the following two types of hypotheses:

(i) That the observations have been drawn at random from a single population, (*i.e.*, the sequence of observations is random).

(ii) That two random samples come from populations having identical distributions (implying that they have equal medians or means).

(i) **One Sample.** To carry out the test, we find the sample median and replace each observation by a *plus* or a *minus* sign depending on whether the observation is above or below the sample median, ignoring observations that are equal to median. We thus get a sequence of plus and minus signs. We denote the number of symbols of one kind by n_1, the number of symbols of the other kind by n_2 and the number of runs by n_r. We reject our null hypothesis if the observed value of n_r is less than or equal to the smaller number and is greater than or equal to the larger number in **Table 24.4**, for the given values of n_1 and n_2 at the 0.05 level.

It has been shown that, when both n_1 and n_2 are greater than 10, n_r is approximately normally distributed with mean

$$\mu_r = 1 + \frac{2n_1 n_2}{n_1 + n_2}$$

and standard deviation

$$\sigma_r = \sqrt{\frac{2\, n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}}$$

The test-statistic then becomes

$$Z = \frac{n_r - \mu_r}{\sigma_r}$$

which, if H_0 is true, is approximately standard normal. We apply the usual decision rules to reject or accept the null hypothesis.

Table 24.4. Critical Values of n_r in the Runs Test

In this table, there are two numbers for each value of n_1 and n_2. Any value of $n_r \leq$ the smaller number and $\geq$ the larger number is in the rejection region at the 0.05 level for one sample.

n_1 \ n_2	4	5	6	7	8	9	10
4		2 9	2 9				
5	2 9	2 10	3 10	3 11	3 11		
6	2 9	3 10	3 11	3 12	3 12	4 13	4 13
7		3 11	3 12	3 13	4 13	4 14	5 14
8		3 11	3 12	4 13	4 14	5 14	5 15
9			4 13	4 14	5 14	5 15	5 16
10			4 13	5 14	5 15	5 16	6 16

(ii) **Two Samples.** (The Wald-Wolfowitz runs test). To perform the test, we write down the $n_1 + n_2$ observations from the two samples and arrange them in one sequence according to their magnitude. We write down the letter A for each observation from sample 1 and the letter B for each observation from sample 2, thus getting a sequence of A's and B's. We count the number of runs in the whole sequence. If both n_1 and n_2 do not exceed 10, then any observed value of n_r less than or equal to the smaller number and $\geq$ the larger number in Table 24.4 is in the critical region for α=0.05. If both n_1 and n_2 are greater than 10, we use the statistic

$$Z = \frac{n_r - \mu_r}{\sigma_r}, \quad \text{(where } \mu_r = 1 + \frac{2n_1 n_2}{n_i + n_2} \text{ as before)}$$

which under H_0, is approximately standard normal.

The following examples illustrate how the runs tests for randomness are performed.

Example 24.1. Each day a sample of 10 production items was taken and the mean weight computed. Following are the first 20 daily means:

13.0 12.8 12.9 13.0 13.1 12.9 12.6 12.6 12.7 12.9

13.1 13.1 13.2 13.3 13.2 13.1 12.9 13.2 13.3 13.2

Are the number of runs below and above the median significant at the 5 percent level?

(i) We state our null and alternative hypotheses as

H_0: The plus and minus signs occur in a random manner, and

H_1: The sequence of signs is not in random order.

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic is n_r, the number of runs in one sample.

(iv) Computations. Arranging the observations in order of increasing magnitude, we find the median to be 30.05. Now, replacing each observation by a plus or a minus sign depending on whether the observation is above or below the median, we get the following sequence

$$- - - - + - - - - - + + + + + + - + + +$$

This gives $n_1 = 10, n_2 = 10$ and $n_r = 6$.

(v) Looking in table 24.4 for $n_1 = 10$, $n_2 = 10$ and $\alpha = 0.05$, the critical region consists of all values of $n_r \leq 6$ and all values of $n_r \geq 16$.

(vi) **Conclusion.** Since the observed value of $n_r = 6$ falls in the critical region, we therefore reject H_0 and may conclude that the number of runs below and above the median are significant at the 0.05 level.

Example 24.12. The following data represent the outside diameters of washers produced by two different production lines:

Line A	1.63, 1.68, 1.59, 1.64, 1.70, 1.58, 1.62, 1.71, 1.57, 1.84, 1.90, 1.96
Line B	1.65, 1.69, 1.72, 1.91, 1.74, 1.75, 1.55, 1.86, 1.87, 1.88

At the $\alpha = 0.05$ level, perform a runs test to test the hypothesis that the two random samples come from populations having the same distributions. (I.U., M.Sc., 1993)

(i) We state our null and alternative hypotheses as

H_0: The two random samples come from populations having the same distributions, and

H_1: The two samples come from populations having different distributions.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is n_r, the number of runs in two-samples. (In other words, the Wald-Wolfowitz runs test is to be used).

(iv) Computations. Arranging the observations of the two samples in one sequence according to their magnitude and replacing each observation from sample 1 by the letter A and each observation from sample 2 by the letter B, we get the following sequence of A's and B's:

$$B\ A\ A\ A\ A\ A\ A\ B\ A\ B\ A\ A\ B\ B\ B\ A\ B\ B\ B\ A\ B\ A$$

This gives $n_1 = 12$, $n_2 = 10$ and $n_r = 12$: and we compute

$$\mu_r = 1 + \frac{2n_1 n_2}{n_1 + n_2} = 1 + \frac{2\,(12)\,(10)}{12 + 10} = 11.91 \text{ and}$$

$$\sigma_r = \sqrt{\frac{2\,n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}} = \sqrt{\frac{2(12)\,(10)\,[2(12)\,(10) - 12 - 10]}{(12 + 10)^2\,(12 + 10 - 1)}}$$

$$= \sqrt{\frac{(240)\,(218)}{(484)\,(21)}} = \sqrt{5.1476} = 2.27$$

$$\therefore \quad z = \frac{n_r - \mu_r}{\sigma_r} = \frac{12 - 11.91}{2.27} = 0.04$$

(v) The critical region is $|Z| \geq z_{0.025} = 1.96$.

(vi) **Conclusion.** Since the computed value of $z = 0.04$ does not fall in the critical region, so we accept H_0 and conclude that the two random samples come from populations having identical distributions. In other words, we may state that the two lines are producing equivalent product.

24.8 THE KOLMOGOROV – SMIRNOV TESTS

The *Kolmogorov-Smirnov tests* are useful nonparametric procedures for goodness-of-fit tests. The one-sample test was proposed in 1933 by A.N. Kolmogorov while the two-sample test was developed in 1939 by N.T. Smirnov. Because of similarities between the test procedures which are based on cumulative distribution functions (or cumulative relative frequency distributions), both names are associated with the one-sample and the two-sample tests.

24.8.1. The Kolmogorov – Smirnov one-sample test is a nonparametric alternative to the *chi-square* goodness-of-fit test. The test compares a cumulative distribution functio[illegible]based on sample observations with some specified population dis[illegible] from which the

random sample has been drawn. The hypothesis to be tested is that the random sample comes from a completely specified distribution.

Let $S_n(X)$ denote the *cumulative* distribution function based on a random sample of n observations, that is, $S_n(X) = k/n$, where k is the number of sample observations less than or equal to X, and let $F_0(X)$ be the hypothesized population *cumulative* distribution function. Then, if our hypothesis that the sample observations come from the specified distribution is true, for every value of X, $S_n(X)$ is expected to be fairly close to $F_0(X)$. A large difference between them $[S_n(X) - F_0(X)]$ provides evidence for rejecting the null hypothesis. The test is therefore based on the maximum *absolute* difference D_n, defined by

$$D_n = \max |S_n(X) - F_0(X)|$$

The null hypothesis is rejected when D_n exceeds the critical value for the chosen level of significance. The critical values are obtained from special tables (see Table A-22, Steel and Torrie, second edition).

The Kolmogorov-Smirnov one-sample test is also used for testing hypothesis about discrete distributions. The test is more powerful than the chi-square test.

24.8.2. The Kolmogorov-Smirnov Two-Sample Test deals with the agreement between two sample *cumulative* distributions. The null hypothesis to be tested is that two independent samples come from identically distributed populations $[F(X_i)]$ and the alternative hypothesis is that they come from populations having different cumulative distribution functions. In symbols, the hypotheses to be tested are

$$H_0 : F_1(X_1) = F_2(X_2) \text{ for all } X, \text{ and}$$

$$H_1 : F_1(X_1) \neq F_2(X_2) \text{ for at least one } X.$$

For this purpose, we compare the sample cumulative relative frequency distributions $[S(X_i)]$ at each sample value (after ranking all values together in case of ungrouped data). The cumulative distributions of both samples are expected to be fairly close to each other if H_0 is true. A large difference between them at any point is therefore evidence for rejecting the null hypothesis. Let $S_{n_1}(X)$ and $S_{n_2}(X)$ denote the cumulative relative frequency distributions of two independent samples of size n_1 and n_2 respectively. Then the Kolmogorov-Smirnov two-sample test is based on the maximum difference D, defined by

$$D = \max [S_{n_1}(X) - S_{n_2}(X)], \text{ for a one-tailed test, and}$$

$$D = \max |S_{n_1}(X) - S_{n_2}(X)|, \text{ for a two-tailed test.}$$

In the case of small samples (n_1 and n_2 less than 40) and two-tailed test, the null hypothesis is rejected when D exceeds the critical value for the

chosen level of significance. The critical values are obtained from special tables. (see Table A-23, Steel and Torrie, 2nd edition).

When both sample sizes n_1 and n_2 are greater then 40, for a one tailed test, the test-statistic to use is

$$\chi^2 = 4D^2 \left[\frac{n_1 n_2}{n_1 + n_2} \right]$$

which has approximately a chi-square distribution with 2 d.f.

The null hypothesis is rejected if $\chi^2 \geq \chi^2_{0.05(2)} = 5.99$.

The test-statistic based on chi-square approximation may also be used with small samples.

We illustrate the procedure with the following examples:

Example 24.13. The following scores were obtained by rolling a six-sided die 10 times: 3, 4, 4, 2, 6, 6, 3, 4, 2, 5.

Use the Kolmogorov-Smirnov statistic to test at $\alpha = 0.05$, the hypothesis that it is a sample from a uniform distribution of integer values between 1 and 6 inclusive.

(i) We state our null and alternative hypotheses as

H_0 : the population distribution is uniform, and

H_1 : the population distribution is not uniform.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic is the Kolmogorov-Smirnov one-sample D statistic.

(v) Computations. Arranging the sample observations in a *cumulative* frequency distribution, we get

Scores (X)	1	2	3	4	5	6	Total
Frequency	0	2	2	3	1	2	10
Cumulative f	0	2	4	7	8	10	–

Let $S_{10}(X)$ denote the *cumulative* relative frequency (probability) distribution of 10 sample observations, and let $F_0(X)$ denote the *cumulative* probability distribution under H_0, where H_0 is the each integer would occur with probability $= \frac{1}{6}$. Then the maximum difference D is computed as follows:

Score (X)	$S_{10}(X)$	$F_0(X)$	$\lvert S_{10}(X) - F_0(X)\rvert$
1	0	1/6=0.17	0.17
2	0.2	2/6=0.33	0.13
3	0.4	3/6=0.50	0.10
4	0.7	4/6=0.67	0.03
5	0.8	5/6=0.83	0.03
6	1.0	6/6=1.00	0.00

Now $D = \max |S_{10}(X) - F_0(X)| = 0.17.$

(v) The critical value of D for $n=10$ and $\alpha=0.05$ from the table of "critical values for Kolmogorov-Smirnov one-sample test" is 0.409.

(vi) **Conclusion.** Since the observed value of $D = 0.17$ does not exceed the table value, so we cannot reject our null hypothesis, which is that the observed scores are uniformly distributed.

Example 24.14. The two independent samples given below are from two populations.

Sample 1	38, 49, 45, 29, 31, 35
Sample 2	31, 42, 22, 26, 43, 37, 25, 30, 47

Apply the Kolmogorov-Smirnov test to test the null hypothesis that the two samples come from populations having identical distributions. Use a 0.05 significance level. (I.U., M.Sc. 1995)

(i) We state our null and alternative hypotheses as

H_0: The two samples come from populations having identical distributions, or equivalently

$$F_1(X) = F_2(X), \text{ and}$$

H_1: The two samples come from different population distributions, *i.e.* $F_1(X) \neq F_2(X)$.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is the Kolmogorov-Smirnov two-sample D statistic.

(iv) Computations. We arrange all the observations together in increasing order of magnitude. We compute the sample cumulative relative frequencies at each sample value and find the differences between them at each listed point. The ordered

sample values, corresponding values of $S_6(X_1)$ and $S_9(X_2)$, and differences $S_6(X_1) - S_9(X_2)$ are given below:

Ordered Observations				
X_1	X_2	$S_6(X_1)$	$X_9(X_2)$	$\lvert S_6(X_1) - S_9(X_2)\rvert$
--	22	0	1/9	$\lvert 0-1/9\rvert = 1/9$
--	25	0	2/9	$\lvert 0-2/9\rvert = 2/9$
--	26	0	3/9	$\lvert 0-3/9\rvert = 3/9 = D$
29	--	1/6	3/9	$\lvert 1/6-3/9\rvert = 3/18$
--	30	1/6	4/9	$\lvert 1/6-4/9\rvert = 5/18$
31	31	2/6	5/9	$\lvert 2/6-5/9\rvert = 4/18$
35	--	3/6	5/9	$\lvert 3/6-5/9\rvert = 1/18$
--	37	3/6	6/9	$\lvert 3/6-6/9\rvert = 3/18$
38	--	4/6	6/9	$\lvert 4/6-6/9\rvert = 0$
--	42	4/6	7/9	$\lvert 4/6-7/9\rvert = 2/18$
--	43	4/6	8/9	$\lvert 4/6-8/9\rvert = 4/18$
45	--	5/6	8/9	$\lvert 5/6-8/9\rvert = 1/18$
--	47	5/6	9/9	$\lvert 5/6-1\rvert = 1/6$
49	--	6/6	9/9	$\lvert 1-1\rvert = 0$

Now $D = \max \lvert S_6(X_1) - S_9(X_2)\rvert = \frac{3}{9} = \frac{6}{18}$, and $n_1 = 6$ and $n_2 = 9$.

(v) The critical value of D for $n_1=6$ and $n_2=9$ from the tables of "Critical Values for Kolmogorov-Smirnov two-sample test" at $\alpha=0.05$ is $2/3 = 12/18$.

(vi) **Conclusion.** As the calculated value of D does not exceed the table value, so we cannot reject H_0. We conclude that the two samples come from populations having identical distributions.

24.9 THE KRUSKAL-WALLIS H TEST

A nonparametric alternative procedure to a one-way analysis of variance or the F-test for testing the equality of several means, is the *Kruskal-Wallis test*. This test is a generalisation of the two-sample Mann-Whitney U test. It tests the null hypothesis H_0 that all $k (k \geq 2)$ independent samples come from populations with equal means (identical

populations) or from the same population against the alternative H_1 that at least one population differs. It is an extremely useful test when the assumptions of normally distributed populations and equality of variances are not satisfied. The test was introduced in 1952 by William H. Kruskal and W. Allen Wallis.

Suppose we have k independent samples of sizes $n_1, n_2, \ldots, n_k$ ($\sum n_i = n$). We denote the observations of the ith sample by $X_{i1}, X_{i2}, \ldots, X_{in_i}$. To perform the test, we arrange *all the n observations of the k samples combined* in increasing order of magnitude and assign the ranks $1, 2, \ldots, n$ to them. In the case of ties, we assign the average of the ranks that would be assigned if there were no ties. To distinguish the sample observations, let us write the letters A, B, C, etc. for the observations of the first sample, the second sample, the third sample, etc. respectively either below the ordered observations or in a Tally Column. We then replace the observations of k samples with their corresponding ranks. we add the ranks in each sample and denote the sums by $R_1, R_2, \ldots, R_k$. We compute

$$S_k^2 = \sum_{i=1}^{k} \frac{R_i^2}{n_i}, \text{ and}$$

$$S_r^2 = \sum_{i,j} r_{ij}^2, \text{ where } r_{ij} \text{ is the rank assigned to observation } X_{ij}.$$

If there are no ties, then

$$S_r^2 = \frac{n(n+1)(2n+1)}{6}.$$

The Kruskal-Wallis statistic H is given by

$$H = \frac{(n-1)(S_k^2 - C)}{S_r^2 - C}$$

where C denotes the appropriate correction term and is given by

$$C = \frac{n(n+1)^2}{4}.$$

In the case of no ties, the statistic H simplifies to

$$H = \frac{12 S_k^2}{n(n+1)} - 3(n+1)$$

This formula may be used when there are a few ties. A generally accepted rule of thumb is that not more than 25% of observations be involved in ties.

We reject the null hypothesis H_0 : All k populations have identical distributions, for large values of H. In case there are only three samples and their sizes are fairly small (5 or fewer) we determine the significance of the statistic H by reference to the *Kruskal and Wallis' Table* which gives critical values for all combinations of the n_i's upto 5, 5, 5.

When at least one of the samples has more than 5 observations or there are more than 5 observations in each sample and H_0 is true, the sampling distribution of statistic H is approximately a chi-square distribution with $(k-1)$ degrees of freedom. We reject H_0 if the calculated value of $H \geq \chi^2_{\alpha,(k-1)}$ at the α level of significance.

The following example illustrates the procedure:

Example 24.15. The following data represent the operating times in hours for three types of scientific pocket calculators before a recharge is required:

Calculator	Operating times (in hours)
A	4.9, 6.1, 4.3, 4.6, 5.3
B	5.5, 5.4, 6.2, 5.8, 5.5, 5.2, 4.8
C	6.4, 6.8, 5.6, 6.5, 6.3, 6.6

Use the Kruskal-Wallis test, at the 0.01 level of significance, to test the hypothesis that the operating times for all three calculators are equal. (P.U., M.Sc. 1991, 95)

(i) We state our null and alternative hypotheses as

H_0: The mean operating times for all three calculators are equal, *i.e.* $\mu_1 = \mu_2 = \mu_3$; and

H_0: The mean operating times for at least two of the three calculators are not equal.

(ii) The significance level is set at $\alpha = 0.01$.

(iii) The test-statistic to use is the Kruskal-Wallis H statistic.

(iv) Computations. Arranging the observations of the combined samples in increasing order of magnitude, assigning ranks and writing the letters A, B and C to distinguish sample values, we get

Ordered Observations	4.3,	4.6,	4.8,	4.9,	5.2,	5.3,	5.4,	5.5,	5.5,	5.6
Rank	1,	2,	3,	4,	5,	6,	7,	8.5,	8.5,	10
Sample	A	A	B	A	B	A	B	B	B	C

Ordered Observations	5.8,	6.1,	6.2,	6.3,	6.4,	6.5,	6.6,	6.8
Rank	11,	12,	13,	14,	15,	16,	17,	18
Sample	B	A	B	C	C	C	C	C

To obtain the sum of ranks of each sample, we replace each observation with its corresponding rank as below:

A	4, 12, 1, 2, 6	$25 = R_1$
B	8.5, 7, 13, 11, 8.5, 5, 3	$56 = R_2$
C	15, 18, 10, 16, 14, 17	$90 = R_3$

Now $S_k^2 = \sum \frac{R_i^2}{n_i} = \frac{(25)^2}{5} + \frac{(56)^2}{7} + \frac{(90)^2}{6} = 1923,$

$$S_r^2 = \sum_{i,j} r_{ij}^2 = (4)^2 + (12)^2 + (1)^2 + \ldots + (17)^2 = 2108.5,$$

$$C = \frac{n(n+1)^2}{4} = \frac{18(18+1)^2}{4} = 1624.5$$

$$\therefore H = \frac{(n-1)[S_k^2 - C]}{S_r^2 - C} = \frac{(17)(1923 - 1624.5)}{2108.5 - 1624.5}$$

$$= \frac{(17)(298.5)}{484} = 10.48$$

OR Since there is only one tie, we can calculate the value of H by applying the formula

$$H = \frac{12 S_k^2}{n(n+1)} - 3(n+1) \qquad \left[S_k^2 = \sum_{i=1}^{k} \frac{R_i^2}{n_i}\right]$$

Thus $H = \frac{12(1923)}{18(18+1)} - 3(18+1) = 67.47 - 57 = 10.47$

(v) The critical region is $H > \chi^2_{0.01(2)} = 9.21$

(vi) **Conclusion.** Since the calculated value of $H = 10.48$ (or 10.47) falls in the critical region, we therefore reject H_0 and conclude that the mean operating times for all three calculators are not equal.

EXERCISES

24.1 Wha. do you understand by nonparametric tests? Why are such tests also called distribution-free tests? Give the advantages and disadvantages of nonparametric tests over parametric tests. (P.U., M.Sc. 1988)

24.2 (a) How do nonparametric tests differ from parametric tests? Discuss the advantages and disadvantages of non-parametric tests.

(b) Describe the Sign test. When is it most appropriately used? Explain the difference between the Wilcoxon signed-rank test and the Sign test. (I.U., M.Sc. 1985, 95)

24.3 (a) Describe the Wilcoxon signed-rank test for one sample. How does it differ from the sign test?

(b) Test the null hypothesis that the median of the population from which the data below have been obtained, equals 55 against the alternative that it is less. Use (i) the sign test, (ii) the Wilcoxon signed-rank test.

48, 51, 49, 53, 61, 59, 45, 52, 65, 47, 58, 57,

65, 56, 45, 49, 54, 63, 46, 57, 54, 53, 52, 45.

(I.U., M.Sc. 1987, 91, 94)

24.4 The following data represent the number of hours that a rechargeable hedge trimmer operates before a recharge is required: 1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2 and 1.7. Use (i) the sign test, (ii) the Wilcoxon signed-rank test to test the hypothesis at the 0.05 level of significance that this particular trimmer operates, on the average, 1.8 hours before requiring a charge.

24.5 A sample of size 8 was chosen from a population. The sample observations are given below:

2.55, 4.62, 2.93, 2.46, 1.95, 4.55, 3.11, and 0.90.

Using (i) the sign test and (ii) the Wilcoxon signed-rank test, test the hypothesis that the median of the population equals 2 against the alternative that it does not. (P.U., M.Sc. 1986)

24.6 (a) Compare the Wilcoxon signed-rank test with the sign test for matched pairs. Also explain the rationale underlying the Wilcoxon signed-rank test.

(b) Two varieties of tomato were experimented with concerni their fruit-producing abilities, measured in pounds. T following data were obtained:

Location	1	2	3	4	5	6	7	8	9	10
Variety A	3.03,	3.10,	2.35,	3.86,	3.91,	1.72,	2.65,	2.30,	2.70,	3.60
Variety B	2.28,	3.63,	2.17,	3.56,	3.73,	1.85,	1.48,	1.86,	2.76,	2.68

Apply (i) the sign test, (ii) the Wilcoxon signed-rank test at th 0.05 level of significance, to test the hypothesis that there is n difference in fruit-producing abilities of the two varieties.

24.7 It is suspected that high school graduates who plan to major i mathematics in college would score more than five points highe in a natural sciences test than in a social sciences test. Th following are the test scores of 20 such students:

Student	1	2	3	4	5	6	7	8	9	10
X	98	75	94	85	70	80	90	75	65	60
Y	87	66	80	75	62	76	78	65	50	60
Student	11	12	13	14	15	16	17	18	19	20
X	80	85	70	90	66	77	76	95	68	70
Y	70	75	60	77	55	80	65	88	60	60

Test the null hypothesis that the natural sciences test scores (X) are five points higher than the social sciences test score (Y) against the alternative that the former are higher by more than five points at $\alpha = 0.01$ by applying (i) the sign test, (ii) the Wilcoxon signed-rank test.

24.8 Explain fully the following nonparametric tests:

(a) The Wilcoxon rank-sum test.

(b) The Mann-Whitney U test.

(c) The Median test.

(d) The Runs test for randomness.

24.9 Five samples of each of two types of paint are scored as follows:

Paint I	85	87	92	80	84
Paint II	89	89	90	84	80

Analyse this with the Wilcoxon rank-sum test.

24.10 Given the two samples below, test the null hypothesis that the population medians are equal against the alternative that $M_1 < M_2$, at $\alpha = 0.05$ by applying the Wilcoxon rank-sum test.

Sample 1	26, 25, 38, 33, 42, 40, 44, 26, 25, 43, 35, 48, 37,
Sample 2	44, 30, 34, 47, 35, 46, 35, 47, 48, 34, 32, 42, 43, 49, 46, 47

24.11 (a) A taxi company tests two types of tyres on the ten cars of its fleet. The length of life of the tyres, in thousand of kilometers, is

Type A	29, 27, 23, 30.
Type B	24, 37, 35, 19, 40, 31

Use a Mann-Whitney U test to test if there is any difference in length of life of the two types of tyres.

(b) Given below are the scores obtained by two groups of trainees:

Group I	31, 28, 42, 36, 29, 51, 34, 25, 44, 33, 49
Group II	27, 45, 30, 53, 41, 39, 48, 43, 26, 37

Use a Mann-Whitney test to test if the scores for the two groups differ significantly. (I.U., M.Sc. 1993)

24.12 Given below are the ages of 29 executives of a certain company:

Men	44, 30, 34, 47, 35, 46, 35, 47, 48, 34, 32, 42, 43, 49, 46, 47
Women	26, 25, 38, 33, 42, 40, 44, 26, 25, 43, 35, 48, 37.

Test the hypothesis that the population medians are equal, *i.e.*, $H_0: M_m = M_w$ against the alternative that $M_w < M_m$, applying the Mann-Whitney U test. (I.U., M.Sc. 1992, 95)

24.13 Use the Mann-Whitney U-test to test the hypothesis that the difference between the mean scores of students in arithmetic computation in two types of school are equal, using $\alpha = 0.05$. Given the data:

Marks	0–9	10–19	20–29	30–39	40–49	50–59	Total
Residential	1	3	10	63	38	5	120
Non-Residential	4	7	25	37	13	4	90

24.14 In an experiment on the effectiveness of a teaching machine, a machine-instructed group of students was compared with a

teacher-instructed group on an achievement test. The following scores were obtained:

Scores	40–49	50–59	60–69	70–79	80–89	90–99	Total
Teacher-instructed	21	40	55	38	10	2	166
Machine-instructed	18	35	42	46	19	4	164
Total	39	75	97	84	29	6	330

Use Mann-Whitney U-test to determine whether there is a significant difference in achievements of the two groups.

(P.U., M.Sc. 1989)

24.15 (a) Discuss fully the procedure and rationale of a two-sample median test.

(b) Using the data in question 24.12, test at $\alpha = 0.05$ the hypothesis that men and women come from populations with the same median, applying the median test.

24.16 Use the median test at the $\alpha = 0.05$ level, to test the null hypothesis that the two samples are drawn from populations with the same median.

Sample 1	92, 63, 30, 78, 24, 19, 26, 79, 54, 57, 97, 46, 58, 74, 77, 80, 93, 99, 78, 50.
Sample 2	77, 87, 98, 62, 76, 47, 66, 83, 72, 80, 53, 80, 48, 75, 76, 78, 97, 53, 64, 67.

24.17 Given the information below for four samples:

Sample 1	29, 40, 43, 30, 46, 34, 32, 27
Sample 2	35, 39, 33, 31, 40, 34, 33, 29
Sample 3	34, 43, 42, 31, 35, 40, 25, 42
Sample 4	33, 34, 32, 26, 48, 30, 35, 42

Use Median test to test the hypothesis that all the four samples come from the identical population. Use $\alpha = 0.05$. (G.U., M.Sc. 1995)

24.18 (a) A true-false examination was constructed with the answers running in the following sequence:

$$T\ F\ F\ T\ F\ T\ F\ T\ T\ F\ T\ F\ F\ T\ F\ T\ F\ T\ T\ F$$

Does this sequence indicate a departure from randomness in the arrangement of T and F answers?

(b) Would the following series of boys and girls selected haphazardly at a coeducational school, be acceptable as a random sample?

G B B G G B G B B B G G G G B G B G B B G B G G G B B G B G G B G G B G B G B B G

24.19 (a) At the $\alpha = 0.05$ level, perform a runs test to determine if the following sequence of measurements is random:

3.6, 3.9, 4.1, 3.6, 3.8, 3.7, 3.4, 4.0, 3.8, 4.1, 3.9, 4.0, 3.8, 4.2, 4.1

(b) Consider the data in question 24.10 to be two random samples and use the run test to test the hypothesis that the two samples are drawn from populations having identical distributions. Let $\alpha = 0.05$.

24.20 (a) What is the function and rationale of the Kolmogorov-Smirnov one-sample test? Also explain its technique.

(b) Test the hypothesis, by the Kolmogorov-Smirnov test, that the following sample: 0.36, 0.92, −0.56, 1.86, 1.74, 0.56, −0.95, 0.24, −0.15, −0.74, 0.32, 0.82, 0.70, −0.10, −1.06, 0.15, 0.55, −0.48, −0.49, −1.26 comes from a normal distribution with mean 0.5 and variance 1. Use $\alpha = 0.05$. (I.U., M.Sc. 1987, 92)

24.21 The following data show the kidney weights in grams of 36 dogs prior to their use in experiments.

58	78	84	90	97	70	90	86	82
59	90	70	74	83	90	76	88	84
68	93	70	94	70	110	67	68	75
80	68	82	104	92	112	84	98	80

Do the data provide enough evidence that they have come from a normally distributed population with a mean of 85 grams and a standard deviation of 15 grams. Use Kolmogorov-Smirnov test.

24.22 (a) Apply the Kolmogorov-Smirnov test to the data of question 24.10. Use a two sided alternative hypothesis.

(b) Apply the Kolmogorov-Smirnov test to the data of question 24.13, to test the hypothesis that there is no difference between the two types of schools.

24.23 Four groups *A*, *B*, *C* and *D* are involved in an experiment. The cases in each group are randomly selected and the following data represent the results of the experiments:

Group A	Group B	Group C	Group D
57	43	87	60
81	61	69	70
67	42	58	75
64	45	82	72
96	39	90	79
80		91	76
68			
56			

Use the Kruskal-Wallis H test to decide whether the four groups are taken from populations with the same mean at $\alpha = 0.05$.

24.24 The following information reflects the results of the tensile strengths of the molded parts produced by different methods (A, B, C, D):

A	80, 88, 87, 86, 90, 88, 85
B	99, 91, 98, 98, 99, 96, 92, 98
C	89, 82, 81, 80, 86, 86, 86, 84
D	76, 77, 75, 78, 76, 73, 71, 80, 75, 80

Use the Kruskal-Wallis test to determine whether the mean tensile strengths are the same for all the methods.

Appendix - A: Vital Statistics

A.1 Meaning of vital statistics. There are some factors which cause changes in the size and composition of human population, e.g., births add and deaths take away some members of the population. Such factors are called *Vital Events*, and they include births, deaths, migrations (which change the size of the population), marriages, divorces, sickness, adoptions, legitimation, etc. (which affect the population composition). The collection, presentation and analysis of vital events constitute *Vital Statistics*. The term Vital Statistics may also be defined as the data systematically collected and compiled in numerical form, relating to or derived from records of vital events. Vital Statistics thus includes the whole study of man, and throws light on various social and medical problems.

A.2 Registration of Births and Deaths in Pakistan. Each birth or death that takes place in Pakistan, is an integral part of the national registration system. The Provincial Governments are required to report births and deaths they register through the local bodies to the respective registration offices. The provinces of Pakistan contain *rural* and *urban* areas, and the registration of births and deaths is carried out separately for the two types of areas.

In *rural areas*, the registration of births and deaths, and the maintenance of vital statistics is the function of rural local councils. It is obligatory on the head of the household in which a birth or death has occurred, to get the event registered within 4 days of its occurrence either at the office of the *union council* directly or to report the event to the *chowkider* (village watchman) who gets it registered at the office of the union council when he visits it. The rural local councils send a copy of entries in birth and death registers to their respective District registration offices.

In *urban areas*, the Municipal Corporations, the Municipal Committees, Town Committees and Cantonment Boards register all births and deaths occurring within their respective limits. The event is reported for its registration by the head of the household in which a birth or death has occurred. In respect of births, the *midwife/dai* attending at a birth, is also required to report the occurrence of birth to the registration office. The occurrence of a birth or death in a Government or Municipal hospital, a private hospital, an orphanage or any other public institution, is reported by the officer-in-charge of such an institution.

A.3 Uses of Vital Statistics. The important uses of vital statistics are enumerated below:

(i) The vital statistics exhibit the changing pattern of the population of a country and reveal the virility of the races. Birth and death rates computed from them, throw light upon medical facilities, hygienic conditions, standard of living, etc.

(ii) On the basis of trends in births and deaths, many manufacturers plan their production of goods such as food, clothing, medicines and other articles needed for children and aged people.

(iii) The vital statistics are used by a number of government and private agencies to draw up their social and economic plans. For instance, planning and production of housing and educational facilities, planning and operating of social security programmes, etc. depend on vital statistics.

(iv) Vital statistics in respect of incidence of diseases, number of deaths, etc. are of great use in taking remedial measures to prevent or control the spread of diseases. Arrangements for inoculation and vaccination are made on receiving information of the outbreak of any epidemic.

(v) Birth and death certificates are needed in many circumstances.

(vi) Accurately registered and systematically collected vital statistics can be used to check up the accuracy of data provided by the census.

(vii) The reproduction rates are studied to know whether the population of a country is increasing, decreasing or is stable.

(viii) The actuaries use vital statistics when estimating life insurance premium to be charged.

(ix) The business prospects of jewellers, furniture manufacturers are influenced by the trend in marriage rates.

(x) The vital statistics are the administrative and research needs of public health agencies.

A.4 Shortcomings of Vital Statistics. Some of the important defects from which our vital statistics suffer, are stated below:

(i) There is evidence that many births escape registration, especially in rural areas. Many deaths, marriages, divorces also remain unregistered. These statistics thus suffer from a large amount of under-registration, the exact extent of which is not known.

(ii) Data on the incidence of diseases, on ages at death and on the cases of death are notoriously unreliable.

(iii) Information in respect of infectious diseases is grossly under-reported. The importance of such reporting is not realized because of illiteracy of the population.

(iv) Data on vital statistics are sometimes deliberately misreported. For example, widowed or divorced may report themselves as singles or they may not tell their true ages.

(v) There arise many inaccuracies at all stages of compiling process and with respect to the presentation of vital statistics due to lack of statistical knowledge.

(vi) Delay in compilation and tabulation is also one of the main defects from which the vital statistics suffer.

A.5 Rates and Ratios. For purposes of comparison, we need *relative numbers*. In vital statistics, the commonly used relative numbers are the *rates* and *ratios*.

The *ratio* of one number, "a" to another number "c" is defined by "a divided by c". It thus indicates the relative size of two numbers. Ordinarily, a and c represent separate and distinct categories. In vital statistics, a *ratio* expresses the relation of a given kind of event to the occurrence of other events or one kind of data to another. Thus, we may write,

$$\text{Ratio} = \frac{a}{c},$$

where a denotes the number of times the given kind of event occurs, and

c denotes the number of times another event occurs.

The important types of ratios used in vital statistics are the sex ratio, child-women ratio, death ratio, birth ratio or vital index, etc.

A special type of ratio in which the denominator represents the total number of cases and the numerator represents certain fraction of this number, is called a *proportion*.

A *rate* is a type of ratio, which in vital statistics may be defined as a numerical proportion of the number of vital events to the population in which the events took place. In other words,

$$\text{Rate} = \frac{a}{a + b'}$$

where a stands for the number of times the given vital event occurs, and

b denotes the number of times, the event does not occur.

Vital rates are usually multiplied by 1,000 for ease in understanding and recording. They are generally of two types, namely, the *crude rates* and the *specific rates*. A *crude rate* is defined as a ratio of vital events occurring during a year to midyear population during the same year, without regard to any specific characteristics of the population. On the other hand, a *specific rate* is defined in terms of one or more characteristics of the population.

A.6 Sex Ratio. The ratio between males and females in a human population, is called a *sex* ratio. It is computed by dividing the number of males in a population by the number of females in the same population and the result is expressed in percentage. In other words,

$$\text{Sex Ratio} = \frac{\text{Number of Males}}{\text{Number of Females}} \times 100$$

Thus sex ratio indicates the number of males per 100 females. It is of great importance to note that a sex ratio more than 100, would imply a preponderance of males, *i.e.*, it would mean that there are more men than women. A sex ratio less than 100, signifies that there are more women then men. Sex ratios for some specific age-groups may also be computed separately.

Example A.1 Calculate Sex Ratios by Age for the 1961 census population of West Pakistan and examine your results critically.

(P.U. B.A/B.Sc. 1970)

Ages (years)	Males ('000)	Females ('000)	Sex Ratios
0 – 4	3,300	3,157	104.5
5 – 9	3,456	3,016	114.6
10 – 19	4,014	3,328	120.6
20 – 29	3,221	2,904	110.9
30 – 39	2,456	2,161	113.6
40 – 49	1,882	1,541	122.1
50 – 59	1,276	996	128.1
60 & Over	1,563	1,171	133.1
All ages	21,168	18,274	115.8

Solution: Sex ratios (number of males per 100 females) at different ages have been computed and are given in the fourth column of the above table. These ratios reveal fluctuations The sex ratio for the first age group is 104.5, which signifies that the male births are slightly more frequent than female births. The boys between the ages 10–19 are more numerous than girls in the same age group. The sex ratios for ages 40 &

A1k · ©Muhammad Imran khask

over, show unusual preponderance of men, owing to most probably a higher death rate of females at higher ages.

A.7 Child-Women Ratio. The ratio between children under 5 years of age and the women of "child bearing age" is called a *child-women ratio*. The child-bearing age is defined sometimes by age-group 15–44 and sometimes by age-group 15–49. The child-women ratio is computed by the formula

$$\text{Child-Women Ratio} = \frac{P_{0-4}}{f_{15-44}} \times 1{,}000,$$

where P_{0-4} denotes the number of children, both sexes combined, under 5 years of age, and

f_{15-44} denotes the number of females (women) between ages 15–44, (f_{15-49} is used when child-bearing age is taken as 15–49).

A.8 Birth-Death Ratio or Vital Index. The ratio between the total number of births and the total number of deaths of a population during a particular year is called *birth-death-ratio* or *vital index*. It is computed by the formula

$$\text{Vital Index} = \frac{\text{Total Number of Births}}{\text{Total Number of Deaths}} \times 100$$

A vital index more than 100, indicates that the population is increasing and is in a healthy condition. When it is less than 100, it implies that the population is decreasing, and the population is stable when the vital index equals 100.

A.9 Population Growth Rate. The *annual population growth rate* is computed by dividing the increase in population during the year by the population at the beginning of that year when the total population of a country is available each year. But for many countries including Pakistan, the total population is not available each year as the population figures are obtained by the decennial censuses of population. In such cases, the annual population growth rate is derived by the formula (compound interest formula)

$$P_n = P_0 (1 + r)^n$$

where P_0 denotes the population, at the beginning of the period/decade,

P_n denotes the population after n years,

n denotes the intercensal period, and

r denotes the unknown rate of change; the result is usually expressed in the form of percentage.

For example, the population of Pakistan according to the 1972 census was 65,309 thousand and according to the 1981 census was 84,254 thousand. For computation of growth rate the intercensal period was taken as 8.46 years as the reference dates for the 1972 census was 10th September, 1972 and for the 1981 census, it was 1st March 1981. (Source: The Pakistan Census Organization, Statistics Division, Government of Pakistan, Islamabad: 1951–81 "Population of Administrative Units; as on 4th February 1986").

Substituting the values in the formula, we get

$$84{,}254 = 65{,}309\,(1 + r)^{8.46}$$

or $$(1 + r)^{8.46} = \frac{84{,}254}{65{,}309} = 1.2901$$

Taking logs, we get

$$8.46 \log (1 + r) = \log (1.2901)$$

or $$\log (1 + r) = \frac{1}{8.46}[0.11090] = 0.01311$$

or $$1 + r = \text{Anti log } (0.01311) = 1.030$$

$\therefore$ $$r = 1.030 - 1 = 0.030$$

Thus the average growth rate of the population of Pakistan during the intercensal period 1972-1981 was 3.0 per cent per annum.

The formula given above is sometimes written as

$$r = \left(\frac{P_n}{P_0}\right)^{1/n} - 1$$

and is known as the Geometric method for estimating the annual growth rate.

A.10 Classification of Vital Rates. The commonly employed rates in vital statistics may be classified as follows:

(a) **Death Rates or Mortality Rates.** The kinds of death rates are:

(i) Crude Death Rate; (ii) Specific Death Rate,
(iii) Infant Mortality Rate, (iv) Case Fatality Rate,
(v) Standardized Death Rate.

(b) **Birth Rates or Natality Rates.** The commonly used birth rates are:

(i) Crude Birth Rate, (ii) Specific Birth Rate,
(iii) Standardized Birth Rate.

OPPO A1k · ©Muhammad Imran Khan

(c) **Reproduction Rates.** There are two important types of such rates, namely,

(i) Gross Reproduction Rate, (ii) Net Reproduction Rate.

(d) **Morbidity rates or Sickness Rates**

(e) **Marriage Rates**

(f) **Divorce Rates, etc.**

A.11 Crude Death Rate. For a given area, the crude death rate may be defined as a ratio of total registered deaths of some specified year to the total midyear population in the same year, multiplied by 1,000. It is computed as follows:

$$\text{C.D.R.} = \frac{D}{P} \times 1{,}000,$$

where C.D.R. stands for crude death rate,

D denotes the total number of deaths from all causes during a calendar year, and

P denotes the midyear total population (which is taken as an estimate of the average population during the whole calendar year) during the same year.

The crude death rate is perhaps the most widely used vital rate as it is easily understood and quickly computed. It merely tells that so many persons have died during a year, that is it represents the probability of dying for persons in the population. Mortality as it is a well known fact, varies with age, sex, race, occupation, but the crude death rate ignores all these factors, therefore it is likely to be misleading and hence should not be used for comparison between areas. The crude death rate measures the decrease in the population due to deaths and thus has a value as an index of mortality.

A.12 Specific Death Rates. When death rates are computed for some specific class of people or specific age-group of a population, they are called *specific death rates*. The kind of specificity must be stated. The most important and widely applicable specific death rates are the age-specific death rates and age-sex-specific death rates.

Age-specific death rates are computed by the formula

$$\text{A.S.D.R.} = \frac{d_i}{P_i} \times 1{,}000,$$

where *A.S.D.R.* denotes the age-specific death rate,

d_i denotes the number of deaths occurring in the ith age-group (*i.e.*, specified) during a given year, and

P_i denotes the midyear population figure of the same age-group.

The age-specific death rate measures the risk of dying in each of the age-groups selected for the computation. When age-specific death rates are computed separately for males and females, they are called *age-sex-specific death rates*. They are used in the computation of net reproduction rates and life tables.

Example A.2 Calculate age-sex specific death rates for the following data for selected age-groups:

Age-group (years)	Midyear population		No. of Deaths	
	Males	Females	Males	Females
10–14	90,229	69,379	1,043	959
15–19	83,442	56,371	931	917
20–29	164,377	103,014	1,111	914
30–39	111,878	68,724	1,039	885
40–49	69,008	48,777	1,017	812
50–59	39,732	29,420	1,008	774

Solution. Calculation of age-sex specific death rates for selected age-groups:

Males:

The age-specific death rate for the age-group 10–14 is

$$\text{A.S.D.R.} = \frac{\text{Number of deaths occurring in age-group 10–14}}{\text{Midyear population in age-group 10–14}} \times 1000$$

$$= \frac{1043}{90229} \times 1000 = 11.56$$

The age-specific death rate for the age-group 15–19 is

$$\text{A.S.D.R.} = \frac{\text{Number of deaths occurring in age-group 15–19}}{\text{Midyear population in age-group 15–19}} \times 1000$$

$$= \frac{931}{83442} \times 1000 = 11.16$$

Similarly, we compute the age-specific death rates for the remaining age-groups of male population and for all age-groups of female population. The age-specific death rates calculated for the data are shown in the table that follows:

Age-group (years)	Age-specific death rates per 1000 persons per year	
	Males	Females
10–14	11.56	13.82
15–19	11.16	16.27
20–29	6.76	8.87
30–39	9.29	12.88
40–49	14.73	16.65
50–59	25.37	26.31

A.13 Infant Mortality Rate. It is defined as a ratio of registered deaths of infants during a specified year to the total live births registered in the same year. The formula thus becomes

$$\text{I.M.R.} = \frac{d_0}{B} \times 1{,}000,$$

where I.M.R. stands for infant mortality rate,

d_0 denotes the number of deaths (excluding foetal deaths) under one year of age registered during a given year in a locality, and

B denotes the number of live births (A child showing any evidence of life is registered as "live birth" registered during the same year in the same locality.

The infant mortality rate does not provide an accurate measure of the risk of death during the first year of life because

(i) infants are usually under-enumerated,

(ii) the babies who die immediately after birth, are often not registered as live births,

(iii) some of the deaths under one year of age during a calendar year must have been of infants who had been born in the preceding calendar year, and

(iv) sometimes infant deaths are not separated from stillbirths and abortions.

The infant mortality rate is however frequently used for many purposes. A low infant mortality rate signifies that the maternity cases are well attended, medical care facilities are adequate, hygienic conditions are good, etc. Thus it serves as an indicator of the level of healthiness" of a society.

A.14 Standardized Death Rates. The crude death rates of two localities or in two occupations cannot be compared because mortality rates differ with age, sex, climate, occupation, etc. Aged people even though well housed and well fed, die comparatively at a higher rate than the young people. Good climate is conducive to long life whereas bad climate is injurious to health. Moreover, mortality is highest at the extremes of age. To eliminate such spurious effects, we compute what are called the *corrected* or *standardized death rates*. Standardized death rates can be computed either by

(i) **direct method,** *i.e.*, by applying different age-specific death rates of the population being studied (or given population), to a *standard population* (which may be any population selected to be the basis of the comparison), or

(ii) **Indirect method,** *i.e.* by applying different age-specific death rates of a standard population to the population being studied.

Direct Method. The direct method for computing the standardized death rate consists in calculating the number of deaths that would be *expected* to occur in a standard population if the age-specific death rates of the given population were to apply; and dividing the sum of the expected deaths by the standard population. That is the *age-adjusted or standardized death rate is computed by the formula*

$$\text{S.D.R.} = \frac{\text{Expected deaths in standard population}}{\text{Total standard population}} \times 1{,}000$$

$$= \frac{\sum \frac{d_i}{p_i} \times P_i}{\sum P_i} \times 1{,}000,$$

where S.D.R. stands for standardized death rate,

d_i denotes the number of deaths occurring in the given population in the ith age-group during a calendar year,

p_i denotes the midyear population figures in the given population in ith age-group during the same year, and

P_i denotes the midyear population figure in the standard population in ith age-group in the same year.

If the death rates are adjusted both for age and sex, the expected deaths are calculated as follows:

$$\text{Expected deaths for males } (m) = \frac{d_{im}}{p_{im}} \times P_{im},$$

Expected deaths for females (f) $= \frac{d_{if}}{p_{if}} \times P_{if}$.

The standardized death rate or the age-sex adjusted death rate is then given by the formula

$$\text{S.D.R.} = \frac{\sum \frac{d_{im}}{p_{im}} \times P_{im} + \sum \frac{d_{if}}{p_{if}} \times P_{if}}{\sum P_{im} + \sum P_{if}} \times 1{,}000.$$

This sort of standardized death rate is easy to compute and to explain. The choice of standard population being subjective, may influence the comparison of standardized death rates.

Indirect Method. When the age-specific death rates for the given population are not available or there are fluctuations in age-specific death rates, we determine the number of deaths that would be *expected* in the given population, assuming that the age specific death rates of the *standard* population had prevailed for one year in the given population. The ratio of total deaths observed in the given population to expected deaths, multiplied by the crude death rate of the standard population yields the age-adjusted or standardized death rate by the indirect method. In other words,

$$\text{S.D.R} = \frac{\text{Observed deaths in the given population}}{\text{Expected deaths in the given population}} \times \text{C.D.R of Standard population.}$$

The number of expected deaths for the *i*th age-group is computed by the formula

$$\text{Expected death} = \frac{D_i}{P_i} \times p_i,$$

where D_i denotes the number of deaths in the *i*th age-group in *standard* population during a year,

P_i denotes the midyear population figure in the *i*th age-group of standard population during the same year, and

p_i denotes the midyear population figure in the *i*th age-group of the given population in the same year.

If adjustment is also to be made with respect to sex, the expected deaths both for males and females are calculated separately in a similar way and the formula applied accordingly. In other words, we use the formula

$$\text{S.D.R} = \frac{\sum d_{im} + \sum d_{if}}{\sum \left(\frac{D_{im}}{P_{im}} \times p_{im}\right) + \sum \left(\frac{D_{if}}{P_{if}} \times p_{if}\right)} \times \text{C.D.R of standard population}$$

$$\text{where C.D.R. (Standard Population)} = \frac{\sum D_{im} + \sum D_{if}}{\sum P_{im} + \sum P_{if}} \times 1{,}000$$

These methods can also be applied for computing the standardized accidents rates in factories when the number of the employees and accidents according to suitable age-groups are available; true unemployment rates, given the number of persons and those seeking employment in a given area according to suitable age-groups.

Example A.3. Given the following data for two districts:

Age-group (years)	District *A*		District *B*	
	Population	Deaths	Population	Deaths
0–14	500	2	400	1
15–29	2,000	8	300	1
30–44	2,000	12	1,000	5
45–59	1,000	10	2,000	18
60–74	500	20	2,000	70
75 & over	100	15	400	50

Compute Crude Death Rates for Districts *A* and *B* and point out the fallacies. Considering District *A* as the standard population, calculate the standardized death rate for District *B*, using direct method.

(P.U., B.A/B.Sc. 1967)

Solution: The computations for crude death rates and age-specific death rates in both the districts to point out the fallacies, if any, are given below:

Age-group (years)	District *A*			District *B*		
	Population	Deaths	A.S.D.R	Population	Deaths	A.S.D.R
0–14	500	2	4.0	400	1	2.5
15–29	2,000	8	4.0	300	1	3.3
30–44	2,000	12	6.0	1,000	5	5.0
45–59	1,000	10	10.0	2,000	18	9.0
60–74	500	20	40.0	2,000	70	35.0
75 & over	100	15	150.0	400	50	125.0
Total	6,100	67	11.0	6,100	145	23.8

OPPO A1k · ©Muhammad Imran khask

$$\text{Crude Death rate for District } A = \frac{67}{6{,}100} \times 1{,}000 = 11.0$$

$$\text{Crude Death rate for District } B = \frac{145}{6{,}100} \times 1{,}000 = 23.8$$

To compute the standardized death rate for District B by the direct method, we first calculate the expected deaths in District A (*Standard Population*) as experienced in District B by the formula

$$\text{Expected death} = \frac{d_i}{p_i} \times P_i$$

These deaths are given below:

Age-group (years)	Expected deaths in Standard Population as experienced in District A	
0–14	$\frac{1}{400} \times 500$	= 1
15–29	$(\frac{1}{300}) \times 2{,}000$	= 7
30–44		10
45–59		9
60–74		18
75 & over	$(\frac{50}{400}) \times 100$	= 12
Total		57

$$\text{Hence standardized death rate} = \frac{57}{6{,}100} \times 1{,}000 = 9.3$$

The crude death rate for District B, which contains a very large proportion of older people, *i.e.* 72% of District B's population is over the age 45 years, comes to 23.8 unusually high, whereas the crude death rate for District A, which contains a smaller proportion of older people *i.e.*, only 26% population is over the age 45 year, is 11.0, unusually low. The death rates at all age-groups in District B are *lower* than those in District A but the crude death rate of B is more than double the crude death rate of A. The fallacy of this sort of unusual behaviour lies in the fact that the like is not being compared with the like. For such a situation, standardized death rates are calculated. Standardizing has lowered the death rate for District B, which turns out to be 9.3.

Example A.4 Calculate the crude and standardized death rates from the following data.

Age-group (years)	District A		Standard	
	Population (p)	Deaths (d)	Population (P)	Deaths (D)
0–9	20,210	728	231,428	6,960
10–24	29,875	213	302,170	1,972
25–44	37,294	462	293,149	4,759
45–64	16,739	741	150,122	7,030
65 & over	4,925	823	46,500	8,314
Total	109,043	2,967	1023,369	29,035

Solution.

$$\text{Crude death rate for District } A = \frac{2,967}{109,043} \times 1,000 = 27.21$$

$$\text{Crude death rate for Standard} = \frac{29,035}{1023,369} \times 1,000 = 28.37$$

The computations for the standardized death-rates both by (i) *Direct* method, *i.e.* by the formula

$$\text{expected death} = \frac{d_i}{p_i} \times P_i, \text{ and}$$

(ii) *Indirect* method, *i.e.* by the formula

$$\text{expected death} = \frac{D_i}{P_i} \times p_i, \text{ are shown below:}$$

Age (years)	Expected deaths in Standard Population (i)	Expected deaths in District A (ii)
0–9	8,336	608
10–24	2,154	195
25–44	3,632	605
45–64	6,646	784
65 & over	7,770	881
Total	28,538	3,073

$$\text{Standardized death rate (Direct)} = \frac{28,538}{1023,369} \times 1,000 = 27.89$$

$$\text{Standardized death rate (Indirect)} = \frac{2,967}{3,073} \times 28.37 = 27.39$$

A.15 Crude Birth Rate. It is a ratio of total registered live births ring a calendar year to the total midyear population during the same

year and multiplied by 1,000. This is analogous to the crude death rate. It is computed by the formula

$$\text{C.B.R.} = \frac{B}{P} \times 1{,}000,$$

where C.B.R. stands for crude birth rate,

B denotes the total number of live births registered during a given year, and

P denotes the midyear total population during the same year.

It is a rough measure of the production capacity of the population. The fertility level of one area cannot be compared with that of another area by crude birth rates.

The difference between the crude birth rate and the crude death rate presents the *crude rate of natural increase* of the population per year per thousand people. It is computed as

$$\text{Crude rate of natural increase} = \frac{\text{Births} - \text{Deaths}}{\text{Midyear Population}} \times 1{,}000$$

If births are more than deaths, additions are being made to the population, but if deaths are more, persons are being removed from the population.

The crude birth rate is sometimes called the "crude live-birth rate" to distinguish it from the "still-birth rate" which is computed as follows:

$$\text{Still-birth Rate} = \frac{Bs}{B} \times 1{,}000,$$

where Bs denotes the number of still-births during any year, and

B denotes the total number of births (both live and still) in the same year.

It provides a measure of reproductive wastage.

Note. A still birth may be defined as a birth of foetus showing no evidence of life after complete expulsion or extraction if 28 weeks of gestation has been reached. A still birth is also called a foetal death.

A.16. Age-Specific Birth Rate. While calculating crude birth rate, we have taken into consideration the total midyear population, but the number of births depends upon the number of married women of child-bearing age. Moreover, births vary with the age of the parents, etc. Age-specific birth rates (A.S.B.R.) are therefore generally computed for women by the formula

$$\text{A.S.B.R} = \frac{b_i}{P_{if}} \times 1{,}000$$

where A.S.B.R. stands for age-specific birth rate,

b_i denotes the number of births registered during the year to women of ith age-group, and

P_{if} denotes the mid-year population figure of women in the same age-group.

Age-specific birth rates are used to compare the natality of different areas as its computation takes account of the age-sex composition of population.

A.17. Standardized Birth Rate. The crude birth rates cannot be used for inter-area comparison because the number of births in any area depends upon the number of married women between the child-bearing ages. The crude birth rate for an area containing a very large proportion of population outside the child-bearing ages, is likely to be smaller in spite of the fact that the number of children per woman in the area is greater. Such sort of diversity is reduced by standardizing the birth rates. The standardized birth rate is computed in a manner similar to the standardized death rate by taking a standard population of women between the child-bearing ages given according to different age-groups.

A.18. General Fertility Rate. *Fertility* means *actual* production of children. It should be distinguished from *fecundity*, which means the *physiological ability* to produce children, irrespective of whether or not children have been produced. It is to be noted that the opposite of *sterility* is *fecundity* and not *fertility*. Fertility is measured from birth statistics, but there is no direct measurement for fecundity.

The *general fertility rate* is a ratio of all live births registered during a year to the number of women of child-bearing age. It is computed by the formula

$$\text{G.F.R.} = \frac{B}{P_{if}} \times 1{,}000,$$

where G.F.R. stands for general fertility rate,

B denotes the total number of live-births registered during the year, and

P_{if} denotes the midyear population of women of child-bearing age.

The fertility rate is general in the sense that it attributes all births to all women in child-bearing age-groups. But the number of births depends upon the number of married women of child-bearing age

OPPO A1k · ©Muhammad Imran khask

Moreover, fertility varies with a number of factors such as age, duration of marriage, occupation, social class, religion, area of residence, etc. The fertility rate specific for age is defined and computed by the following formula:

$$\text{Age-Specific Fertility Rate} = \frac{B_i}{P_{if}} \times 1{,}000,$$

where B_i denotes the number of live births occurring to mothers of the ith age-group during a year,

P_{if} denotes the midyear female population of the same age-group during the same year.

The terms *age-specific birth rates and the age-specific fertility rates* are used interchangeably.

A.19. Total Fertility Rate. The *total fertility rate* is obtained by aggregating the age-specific fertility rates for women of each reproductive age. When the age-specific fertility rates are given or computed for 5-year age-groups, the aggregate is to be multiplied by 5, the number of years in each age-group, as it is the sum of the rates at every individual age which is required. In other words, the total fertility rate is computed by the formula

$$\text{T.F.R.} = 5 \sum (\text{age-specific fertility rate})$$

$$= 5 \sum_i \frac{B_i}{P_{if}} \times 1000,$$

where T.F.R. stands for total fertility rate,

B_i denotes the total live births in ith age-group; and

P_{if} denotes the midyear female population in the same age-group during the same year.

The *total fertility rate* thus provides the total number of babies who would be born (ignoring mortality) to a hypothetical group of fertility rates while passing through the reproductive period.

A.20. Gross Reproduction Rates. While deriving the general fertility rates, the sex of the child born and the mortality are not taken into consideration. Taking sex into account, we get Gross Reproduction Rate (G.R.R.) We obtain Net Reproduction Rate (N.R.R.), when adjustment in respect of both sex and mortality is made:

The *gross reproduction rate* (G.R.R.) may thus be defined as the sum of age-specific birth rates of child-bearing age, restricted to female births only. It is assumed that

(i) none of the female babies would die or would migrate before reaching the end of child-bearing age, and

(ii) the fertility rate would remain unchanged through-out the child-bearing age.

When age-specific fertility rates are given or computed for age-groups, each such rate should be multiplied by the number of years in each age-group and then added, *or* the resulting figure may be multiplied by the span of age-groups instead of multiplying each age-group. The classification of female population of child-bearing age, taken as 15–49 years, into 5 years age-groups such as 15–19, 20–24, ..., 45–49, yields *seven* age-groups. The G.R.R. is therefore computed by the formula

$$\text{G.R.R.} = 5 \sum_{i=1}^{7} \frac{b_{if}}{P_{if}}$$

where b_{if} denotes the number of live female births registered during a year, to mothers of age i, (i is an age-group of five years); and

P_{if} denotes the midyear female population of the same age-group.

Sometimes the births of both sexes may be given and the computations are made by using these births. The G.R.R. in such a case is obtained by multiplying the resulting sum by the ratio of females to total births.

The gross reproduction rate estimates the average number of female babies produced by one married woman throughout her reproductive life.

A.21 Net Reproduction Rate. It is a well known fact that all the female babies born do not survive till they reach the child-bearing age. To adjust the rate for mortality, we therefore take into account the survival rates (*i.e.* the probability of daughters surviving from birth to age-group of mother) for girls born to women in different age-groups. The N.R.R. is thus defined by the average number of female babies who would become mothers when they attain their child-bearing age. It is computed by the formula

$$\text{N.R.R.} = 5 \sum \frac{\text{Female births}}{\text{Female population}} \times \text{probability of survival.}$$

Hence the N.R.R. estimates the number of female babies that would be produced by women throughout their life time if they were to experience at each age-group certain *fertility* and *mortality rates*. Thus it estimates the average number of potential mothers who survive.

A net reproduction rate of 1 signifies that the present female population is exactly maintaining itself and the population is considered stable. If it is less than 1, it implies that the number of potential mothers is decreasing and hence the population is declining. The population in this case is considered as heading towards extinction. If it is greater than 1, the population will increase.

Example A.5. Calculate age-specific fertility rates, total fertility rate, gross reproduction rate and net reproduction rate for the following data. Assume sex ratio at birth to be 105.2 per cent.

Age-group (years)	Female Population	Registered Births	Probability of Survival
15 – 19	13,472	1,223	0.9694
20 – 24	8,812	1,744	0.9668
25 – 29	5,419	1,709	0.9632
30 – 34	5,070	708	0.9584
35 – 39	5,018	369	0.9519
40 – 44	4,860	25	0.9424

Solution. A sex ratio = 105.2 per cent implies that there are 1052 males for 1000 females. We first calculate the number of female births by multiplying the registered births of each age-group by the factor 1000/2052. The female births are shown in column 4 of table given on next page.

Next, we calculate the age-specific birth (fertility) rates for 1000 women by the formula

Age-specific fertility rate

$$= \frac{\text{Number of live births in } i\text{th age-group}}{\text{Mid-year female population in } i\text{th age-group}} \times 1000$$

For example, age-specific fertility (birth) rate for age-group 15–19, is

$$\text{A.S.B.R.} = \frac{B_i}{P_{if}} \times 1000 = \frac{1223}{13472} \times 1000 = 90.78,$$

and for age-group 20–24, is

$$\text{A.S.B.R.} = \frac{1744}{8812} \times 1000 = 197.91, \text{ and so on.}$$

These results are set out in column 5 of the table on next page. The sum multiplied by 5 yields the total fertility rate per 1000 women.

Next, we compute age-specific birth rates for daughters only by dividing the female births in the ith group by the corresponding midyear female population per woman. They appear in column 6 of the table on next page. These calculations relate to G.R.R. Then, we multiply each age-specific birth rate for daughters only by the corresponding probability of survival to find the expected survivors of female births per woman. These figures are given in column 8. The sum of these figures multiplied by 5 gives N.R.R.

OPPO A1k · ©Muhammad Imran khask

Age group (years) (1)	Midyear Female Population (2)	Registered Births (3)	Female Babies (4)	Age-specific fertility rates per 1000 women (5)	Age-specific fertility rates for daughters only (6)	Probability of survival (7)	Expected Survivors of female births per woman (8)
15–19	13,472	1,223	596	90.78	0.0442	0.9694	0.0428
20–24	8,812	1,744	850	197.91	0.0965	0.9668	0.0933
25–29	5,419	1,709	833	315.37	0.1537	0.9632	0.1480
30–34	5,070	708	345	139.64	0.0680	0.9584	0.0652
35–39	5,018	369	180	73.54	0.0359	0.9519	0.0342
40–44	4,860	25	12	5.14	0.0025	0.9424	0.0024
Total	42,651	5,778	2,816	822.38	0.4008	--	0.3859
Multiplying by 5 for single age figures				4,112	2.004	--	1.9295

Hence T.F.R. $= 5\sum$ (age-specific birth rates) = 4,112 per 1000 women,

G.R.R. $= 5\sum$ (age-specific fertility rate for daughters only)
= 2.004, per woman; and

N.R.R. $= 5\sum$ (expected survivors of female births)
= 1.93 per woman.

Example A.6. Compute the gross and net reproduction rates for the following data:

Age-group (years)	Female Population (000)	Female births	Probability of Survival
15 – 19	1558	18,900	0.914
20 – 24	1112	71,100	0.899
25 – 29	1595	96,900	0.884
30 – 34	1629	64,200	0.868
35 – 39	1627	34,900	0.852
40 – 44	1522	10,800	0.834
45 – 49	1401	800	0.813

Solution. The necessary computations are given below:

Age-group (years)	Mean fertility	Female offspring of survivors
15 – 19	$\frac{18,900}{1558,000} = 0.012$	$0.012 \times 0.914 = 0.011$
20 – 24	0.064	0.057
25 – 29	0.061	0.054
30 – 34	0.039	0.034
35 – 39	0.021	0.018
40 – 44	0.007	0.006
45 – 49	0.001	0.001
Total	0.205	0.181

OPPO A1k · @Muhammad Imran khask

A.8 Calculate the age-sex specific death rates from the following data for selected age-groups:

Age-group (years)	Midyear Population		Number of Deaths	
	Males	Females	Males	Females
10 – 14	46,600	37,283	615	600
15 – 19	37,249	28,688	704	601
20 – 29	75,883	57,626	704	641
30 – 39	54,857	39,612	777	668
40 – 49	36,799	28,708	772	666
50 – 59	24,452	18,395	806	706
60 & over	21,702	15,244	830	693

A.9 (a) Define Rates and Ratios and explain the difference between them with examples.

(b) Compute specific mortality rates from cancer per 100,000 population from the following data:

Age in years	Population as of census data	Deaths from Cancer
20 – 29	725,369	79
30 – 39	700,213	238
40 – 49	609,616	660
50 – 59	548,400	1,582
60 – 69	402,054	2,342
70 & over	287,280	3,402

(P.U., B.A./B.Sc. 1968)

A.10 Point out the difference between crude and standardized death rates, and explain the direct and indirect method of standardizing. Bring out clearly the uses of standardized rates.

A.11 (a) Why are death rates standardized?

(b) Calculate the Crude and Standardized Death Rates of 1964 population by (i) Direct and (ii) Indirect methods, using 1941 population as standard.

Age-group (years)	1941		1964	
	Population	Deaths	Population	Specific death rates
0 – 9	10,000	220	15,000	20
10 – 19	11,000	132	12,000	11
20 – 49	7,000	105	8,000	14
50 & Over	2,000	90	5,000	42

(P.U., B.A./B.Sc. 1983)

OPPO A1k · ©Muhammad Imran khask

$$G.R.R. = 5 \times \sum \frac{\text{Female births}}{\text{Female Population}} = 5 \times 0.205 = 1.025$$

$$N.R.R. = 5 \times 0.181 = 0.905.$$

The net reproduction rate turns out to be less than 1. It signifies that each mother produces less than one mother per generation. The population is thus considered as heading towards extinction.

EXERCISES

A.1 Describe the meaning of vital statistics and discuss its uses and shortcomings in detail. (P.U. B.A./B.Sc., 1972)

A.2 Define vital statistics and discuss its scope and limitations. (P.U. B.A/B.Sc. 1969, 82)

A.3 What are vital statistics? Describe the system for collection of vital statistics in Pakistan. Discuss its strong and weak points and suggest remedies. (P.U., M.A., 1963)

A.4 Define and explain the following:
(i) Crude Death Rate, (ii) Crude Birth Rate,
(iii) Sex Ratio, (iv) Age-specific Death Rate
(v) Infant Mortality Rate, (vi) Vital Index. (P.U., B.A./B.Sc. 1971)

A.5 Define Death Rate, Birth Rate and Morbidity Rate. Justify your definitions by taking an example in each case. (P.U., B.A/B.Sc. 1969)

A.6 Differentiate between Rates and Ratios and explain what do you mean by the Crude Death Rates and Age-Specific Death Rates. (P.U., B.A/B.Sc. 1970)

A.7 (a) Explain (i) vital events, (ii) sources of vital data, and (iii) vital index.
(b) Calculate Age-specific Death Rates per 1,000 persons of Pakistan from the following data:

Age (in completed years)	1964 estimated Population, Both sexes in 1000	1964 estimated Deaths in 100
All ages	42,390	8,274
Under 1	1,363	3,175
1 – 4	5,578	1,841
5 – 9	6,595	401
10 – 19	8,121	255
20 – 29	6,818	329
30 – 39	5,150	309
40 – 49	3,686	316
50 – 59	2,403	277
60 & over	2,679	1,370

(P.U., B.A./B.Sc. 1970)

OPPO A1k · ©Muhammad Imran khask

A.12 Point out the difference between Crude and Standardized Death Rates. Calculate the crude and standardized death rates of the local population from the following data, using direct method.

Age-group (years)	Standard Population	No. of deaths in Standard Population	Local Population	No. of deaths in Local Population
0 – 9	600	18	400	16
10 – 19	1000	5	1500	6
20 – 59	3000	24	2400	24
60 & over	400	20	700	21

(P.U., B.A/B.Sc. 1974)

A.13 (a) Explain with suitable illustrations the object of standardizing various vital statistics relating to births, deaths and marriages.

(b) Calculate the crude death rate and the standardized death rate for the data:

Age (years)	District A				Standard Population ('000)	
	Population		No. of deaths			
	males	females	males	females	males	females
0 – 4	2,110	2,010	30	27	59	55
5 – 14	3,340	3,230	6	8	109	102
15 – 34	7,320	7,310	16	20	177	180
35 – 59	7,960	8,750	70	57	121	122
60 & over	3,240	4,280	196	230	34	41

A.14 Following table contains the numbers and age-distribution of male employees at two factories, and the number of accidents:

Age (years)	Factory I		Factory II	
	No. of employees	No. of accidents	No. of employees	No. of accidents
under 21	330	28	400	38
21 – 29	570	40	720	67
30 – 39	710	45	810	60
40 – 49	780	55	390	34
50 – 59	690	54	250	25
60 & over	250	25	80	11

(i) Calculate the crude accident rate per cent for each factory.

(ii) Take the age distribution of Factory I as a standard and calculate the standardized accident rate per cent for Factory II.

15 Define gross and net reproduction rates. Explain how you would compute the net reproduction rate and what interpretations can be made if it is 1, less than 1 or greater than 1.

16 With the help of hypothetical example, explain the difference between Gross and Net Reproduction Rates. Using some arbitrary figures, give the interpretation of these rates in population growth. (P.U., B.A/B.Sc. 1967)

17 Calculate age-specific fertility rates, total fertility rate, gross-reproduction rate and net reproduction rate from the following data, assuming sex-ratio at birth to be 106.18 per cent.

Age-group (years)	Female Population (000)	Registered Births	Probability of Survival
15 – 19	1,424	27,639	0.9645
20 – 24	1,531	226,817	0.9607
25 – 29	1,653	280,506	0.9554
30 – 34	1,658	194,526	0.9489
35 – 39	1,741	113,966	0.9416
40 – 44	1,669	32,363	0.9324
45 – 49	1,561	2,215	0.9201

(P.U., B.A/B.Sc. 1996)

(a) What is the Net Reproduction Rate, how is it calculated, and what purpose does it serve?

(b) Calculate Net Reproduction Rate from the following data:

Age-group (years)	Female Population	Female Births	Probability of Survival
15 – 19	87	4	0.850
20 – 24	63	11	0.800
25 – 29	55	8	0.700
30 – 34	41	6	0.650
35 – 39	33	4	0.600
40 – 44	36	2	0.500

(P.U., B.A/B.Sc. 1967)

A.19 Compute the gross and net reproduction rates for the following data:

Age-group (years)	Female Population (000)	Female live births	Survival rate
15 – 19	1399	15,133	0.9694
20 – 24	1422	94,155	0.9668
25 – 29	1521	102,676	0.9632
30 – 34	1756	72,490	0.9584
35 – 39	1451	31,402	0.9519
40 – 44	1689	10,640	0.9424
45 – 49	1667	700	0.9279

(P.U., B.A/B.Sc. 1985)

A.20 (a) Describe methods of calculating gross and net reproduction rates. What are the relative merits of the net reproduction rate?

(b) From the following data, calculate the gross and net reproduction rates, assuming sex ratio at birth to be 105.2 per cent.

Age-group (years)	Female Population	Registered Births	Survivors among females out of 100
15 – 19	8981	1835	634
20 – 24	5875	2616	602
25 – 29	3613	2563	568
30 – 34	3380	1062	530
35 – 39	3345	558	488
40 – 44	3248	37	444

(P.U., B.A/B.Sc. 1984)

A1k · ©Muhammad Imran khask

APPENDIX - B: Statistical Tables

Table 1. *Significant Ranges for Duncan's Multiple Range Test*

$q_{.05}(p, v)$

v	p = 2	3	4	5	6	7	8	9	10	20	50	100
1	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0
2	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09
3	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50
4	3.93	4.01	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02
5	3.64	3.74	3.79	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83
6	3.46	3.58	3.64	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68
7	3.35	3.47	3.54	3.58	3.60	3.61	3.61	3.61	3.61	3.61	3.61	3.61
8	3.26	3.39	3.47	3.52	3.55	3.56	3.56	3.56	3.56	3.56	3.56	3.56
9	3.20	3.34	3.41	3.47	3.50	3.52	3.52	3.52	3.52	3.52	3.52	3.52
10	3.15	3.30	3.37	3.43	3.46	3.47	3.47	3.47	3.47	3.48	3.48	3.48
11	3.11	3.27	3.35	3.39	3.43	3.44	3.45	3.46	3.46	3.48	3.48	3.48
12	3.08	3.23	3.33	3.36	3.40	3.42	3.44	3.44	3.46	3.48	3.48	3.48
13	3.06	3.21	3.30	3.35	3.38	3.41	3.42	3.44	3.45	3.47	3.47	3.47
14	3.03	3.18	3.27	3.33	3.37	3.39	3.41	3.42	3.44	3.47	3.47	3.47
15	3.01	3.16	3.25	3.31	3.36	3.38	3.40	3.42	3.43	3.47	3.47	3.47
16	3.00	3.15	3.23	3.30	3.34	3.37	3.39	3.41	3.43	3.47	3.47	3.47
17	2.98	3.13	3.22	3.28	3.33	3.36	3.38	3.40	3.42	3.47	3.47	3.47
18	2.97	3.12	3.21	3.27	3.32	3.35	3.37	3.39	3.41	3.47	3.47	3.47
19	2.96	3.11	3.19	3.26	3.31	3.35	3.37	3.39	3.41	3.47	3.47	3.47
20	2.95	3.10	3.18	3.25	3.30	3.34	3.36	3.38	3.40	3.47	3.47	3.47
30	2.89	3.04	3.12	3.20	3.25	3.29	3.32	3.35	3.37	3.47	3.47	3.47
40	2.86	3.01	3.10	3.17	3.22	3.27	3.30	3.33	3.35	3.47	3.47	3.47
60	2.83	2.98	3.08	3.14	3.20	3.24	3.28	3.31	3.33	3.47	3.48	3.48
100	2.80	2.95	3.05	3.12	3.18	3.22	3.26	3.29	3.32	3.47	3.53	3.53
∞	2.77	2.92	3.02	3.09	3.15	3.19	3.23	3.26	3.29	3.47	3.61	3.67

v = degrees of freedom.

Table 2. *Significant Ranges for Duncan's Multiple Range Test*

$q_{.01}(p, \upsilon)$

υ	p = 2	3	4	5	6	7	8	9	10	20	50	100
1	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
2	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0
3	8.26	8.5	8.6	8.7	8.8	8.9	8.9	9.0	9.0	9.3	9.3	9.3
4	6.51	6.8	6.9	7.0	7.1	7.1	7.2	7.2	7.3	7.5	7.5	7.5
5	5.70	5.96	6.11	6.18	6.26	6.33	6.40	6.44	6.5	6.8	6.8	6.8
6	5.24	5.51	5.65	5.73	5.81	5.88	5.95	6.00	6.00	6.3	6.3	6.3
7	4.95	5.22	5.37	5.47	5.53	5.61	5.69	5.73	5.8	6.0	6.0	6.0
8	4.74	5.00	5.14	5.23	5.22	5.40	5.47	5.51	5.5	5.8	5.8	5.8
9	4.60	4.86	4.99	5.08	5.17	5.25	5.32	5.36	5.4	5.7	5.7	5.7
10	4.48	4.73	4.88	4.96	5.06	5.13	5.20	5.24	5.28	5.55	5.55	5.55
11	4.39	4.63	4.77	4.86	4.94	5.01	5.06	5.12	5.15	5.39	5.39	5.39
12	4.32	4.55	4.68	4.76	4.84	4.92	4.96	5.02	5.07	5.26	5.26	5.26
13	4.26	4.48	4.62	4.69	4.74	4.84	4.88	4.94	4.98	5.15	5.15	5.15
14	4.21	4.42	4.55	4.63	4.70	4.78	4.83	4.87	4.91	5.07	5.07	5.07
15	4.17	4.37	4.50	4.58	4.64	4.72	4.77	4.81	4.84	5.00	5.00	5.00
16	4.13	4.34	4.45	4.54	4.60	4.67	4.72	4.76	4.79	4.94	4.94	4.94
17	4.10	4.30	4.41	4.50	4.56	4.63	4.68	4.73	4.75	4.89	4.89	4.89
18	4.07	4.27	4.38	4.46	4.53	4.59	4.64	4.68	4.71	4.85	4.85	4.85
19	4.05	4.24	4.35	4.43	4.50	4.56	4.61	4.64	4.67	4.82	4.82	4.82
20	4.02	4.22	4.33	4.40	4.47	4.53	4.58	4.61	4.65	4.79	4.79	4.79
30	3.89	4.06	4.16	4.22	4.32	4.36	4.41	4.45	4.48	4.65	4.71	4.71
40	3.82	3.99	4.10	4.17	4.24	4.30	4.34	4.37	4.41	4.59	4.69	4.69
60	3.76	3.92	4.03	4.12	4.17	4.23	4.27	4.31	4.34	4.53	4.66	4.66
100	3.71	3.86	3.98	4.06	4.11	4.17	4.21	4.25	4.29	4.48	4.64	4.65
∞	3.64	3.80	3.90	3.98	4.04	4.09	4.14	4.17	4.20	4.41	4.60	4.68

υ = degrees of freedom.

Table 3. The Normal Distribution

P	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	∞	2.5758	2.3263	2.1701	2.0537	1.9600	1.8808	1.8119	1.7507	1.6954
.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
.3	1.0364	1.0152	.9945	.9741	.9542	.9346	.9154	.8965	.8779	.8596
.4	.8416	.8239	.8064	.7892	.7722	.7554	.7388	.7225	.7063	.6903
.5	.6745	.6588	.6433	.6280	.6128	.5978	.5828	.5681	.5534	.5388
.6	.5244	.5101	.4958	.4817	.4677	.4538	.4399	.4261	.4125	.3989
.7	.3853	.3719	.3585	.3457	.3319	.3186	.3055	.2924	.2793	.2663
.8	.2533	.2404	.2275	.2147	.2019	.1891	.1764	.1637	.1510	.1383
.9	.1257	.1130	.1004	.0878	.0753	.0627	.0502	.0376	.0251	.0125
	P	.002	.001	.000,1	.000,01	.000,001		.000,000,1		
	Z	3.0902	3.2905	3.8906	4.4172	4.8916		5.3267		

The value of P for each entry is found by adding the column heading to the value in the left hand margin. The corresponding value of Z is the deviation such that the probability of an observation falling outside the range from $-Z$ to $+Z$ is P. For example, $P=.03$ for $Z=2.1701$; so that 3 per cent of normally distributed values will have positive or negative deviation exceeding the standard deviation in the ratio 2.1701 at least.

"Table 3 is taken from Table 1 of Fisher and Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, published by Oliver and Boyd Ltd., Edinburgh, and printed by permission of the authors and publishers."

OPPO A1k · ©Muhammad Imran khask

Table 4. Areas for a Standard Normal Distribution

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0159	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2083	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2380	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3880
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3990	.3997	4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4430	.4441
1.6	.4452	.4463	.4474	.4485	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4690	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4758	.4762	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4865	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4980	.4980	.4981
2.9	.4981	.4982	.4983	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	49865	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	49903	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993

Table 5. Fisher-z Values (z_f)

r	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.01000	0.02000	0.03001	0.04002	0.05004	0.06007	0.07012	0.08017	0.09024
0.1	0.10034	0.11045	0.12058	0.13074	0.14093	0.15114	0.16139	0.17167	0.18198	0.19234
0.2	0.20273	0.21317	0.22366	0.23419	0.24477	0.25541	0.26611	0.27686	0.28768	0.29857
0.3	0.30952	0.32055	0.33165	0.34283	0.35409	0.36544	0.37689	0.38842	0.40006	0.41180
0.4	0.42365	0.43561	0.44769	0.45990	0.47223	0.48470	0.49731	0.51007	0.52298	0.53606
0.5	0.54931	0.56273	0.57634	0.59014	0.60415	0.61838	0.63283	0.64752	0.66246	0.67767
0.6	0.69315	0.70892	0.72500	0.74142	0.75817	0.77530	0.79281	0.81074	0.82911	0.84795
0.7	0.86730	0.88718	0.90764	0.92873	0.95048	0.97295	0.99621	1.02033	1.04537	1.07143
0.8	1.09861	1.12703	1.15682	1.18813	1.22117	1.25615	1.29334	1.33308	1.37577	1.42192
0.9	1.47222	1.52752	1.58902	1.65839	1.73805	1.83178	1.94591	2.09229	2.29756	2.64665

Table 6. Square and Square Roots

n^2	$\sqrt{n}$	$\sqrt{10n}$
1.00	1.000	3.162
1.21	1.049	3.317
1.44	1.095	3.464
1.69	1.140	3.606
1.96	1.183	3.742
2.25	1.225	3.873
2.56	1.265	4.000
2.89	1.304	4.123
3.24	1.342	4.243
3.61	1.378	4.359
4.00	1.414	4.472
4.41	1.449	4.583
4.84	1.483	4.690
5.29	1.517	4.796
5.76	1.549	48.99
6.25	1.581	5.000
6.76	1.612	5.099
7.29	1.643	5.196
7.84	1.673	5.292
8.41	1.703	5.385
9.00	1.732	5.477
9.61	1.761	5.568
10.24	1.789	5.657
10.89	1.817	5.745
11.56	1.844	5.831
12.25	1.871	5.916
12.96	1.897	6.000
13.69	1.924	6.083
14.44	1.949	6.164
15.21	1.975	6.245
16.00	2.000	6.325
16.81	2.025	6.403
17.64	2.049	6.481
18.49	2.074	6.557
19.36	2.098	6.633
20.25	2.121	6.708
21.16	2.145	6.782
22.09	2.168	6.856
23.04	2.191	6.928
24.01	2.214	7.000
25.00	2.236	7.071
26.01	2.258	7.141
27.04	2.280	7.211
28.09	2.302	7.280
29.16	2.324	7.348

n	n^2	$\sqrt{n}$	$\sqrt{10n}$
5.5	30.25	2.345	7.416
5.6	31.36	2.366	7.483
5.7	32.49	2.387	7.550
5.8	33.64	2.408	7.616
5.9	34.81	2.429	7.681
6.0	36.00	2.449	7.746
6.1	37.21	2.470	7.810
6.2	38.44	2.490	7.874
6.3	39.69	2.510	7.937
6.4	40.96	2.530	8.000
6.5	42.25	2.550	8.062
6.6	43.56	2.569	8.124
6.7	44.89	2.558	8.185
6.8	46.24	2.608	8.246
6.9	47.61	2.627	8.307
7.0	49.00	2.646	8.367
7.1	50.41	2.665	8.426
7.2	51.84	2.683	8.485
7.3	53.29	2.702	8.544
7.4	54.76	2.720	8.602
7.5	56.25	2.739	8.660
7.6	57.76	2.757	8.718
7.7	59.29	2.775	8.775
7.8	60.84	2.793	8.832
7.9	62.41	2.811	8.888
8.0	64.00	2.828	8.944
8.1	65.61	2.846	9.000
8.2	67.24	2.864	9.055
8.3	68.89	2.881	9.110
8.4	70.56	2.898	9.165
8.5	72.25	2.915	9.220
8.6	73.96	2.933	9.274
8.7	75.69	2.950	9.327
8.8	77.44	2.966	9.381
8.9	79.21	2.983	9.434
9.0	81.00	3.000	9.487
9.1	82.81	3.017	9.539
9.2	84.64	3.033	9.592
9.3	86.49	3.050	9.644
9.4	88.36	3.065	9.695
9.5	90.25	3.082	9.747
9.6	92.16	3.098	9.798
9.7	94.09	3.114	9.849
9.8	96.04	3.130	9.899
9.9	98.01	3.146	9.950

REFERENCES

1. Armitage, P and Berry, G — *Statistical Methods in Medical Research*, 2nd ed. Blackwell Scientific Publications London—1988.
2. Bhattacharyya, G.K. and Johnson R.A. — *Statistical Concepts and Methods*. New York, John Wiley & Sons. — 1977.
3. Bowen, E.K. and M.K. Starr — *Basic Statistics for Business and Economics*. McGraw-Hill Book Company—1982.
4. Chao, L.L. *Introduction to Statistics*—Monterey, Calif. Books/Cole Publishing Co., 1980.
5. Cochran, W.G. *Sampling Techniques*—3rd ed. New York, John Wiley & Sons—1977.
6. Dixon, W.J. and F.J. Massey, Jr. — *Introduction to Statistical Analysis*, 3rd ed. New York. MeGraw-Hill Book Company, 1969.
7. Francis, A — *Advanced Level Statistics*. Stanley Thornes (Publishers) Ltd. Cheltenhem, Glos. — 1986.
8. Freund, J.E.—*Modern Elementary Statistics*, 5th ed. Englewood Cliffs, N.J. Prentice Hall, Inc. 1979.
9. Gilchrist, Warren—*Statistical Modelling*. New York, John Wiley & Sons, Ltd. 1984.
10. Groebner, D.F. and P.W. Shannon. — *Business Statistics*. Ohio, Charles E. Merrill Publishing Co. 1981.
11. Guenther, W.C. — *Concepts of Statistical Inference*. New York, McGaw-Hill Company — 1965.
12. Hicks, C.R. — *Fundamental Concepts in the Design of Experiments*. New York, Holt, Rinehart and Winston, 1964.
13. Hines, William W, and Montgomery, Dougles C — *Probability and Statistics in Engineering and Management Science*. 2nd ed. John Wiley & Sons — 1980.
14. Hoel, P.G. — *Introduction to Mathematical Statistics*: 4th ed. New York, John Wiley and Sons.
15. Hughes, A. and D. Grawoig — *Statistics: A Foundation for* [illegible] Inc. 1971.

18. Lapin, L.L.— *Statistics: Meaning and Method*, 2nd ed. New York. Harcourt Brace Jovanovich Inc. 1980.

19. Larson, H.J. — *Introduction to the Theory of Statistics* — New York, John Wiley & Sons, Inc. 1973.

20. Larsen, Richard J. & Marx, Morris L — *An Introduction to Mathematical Statistics and its Applications*. 2nd ed. 1986 Prentic-Hall, Englewood Cliffs, N.J.

21. Mann, Prem S — *Introductory Statistics*. 2nd ed. 1995, John Wiley & Sons, Inc.

22. McClave, James T. and Bensen, P. George — *Statistical for Business and Economics*. 5th ed. 1991, Maxwell Macmillan International, N.Y.

23. Mendenhall, W.—*Introduction to Probability and Statistics*, 5th ed. North Scituate Mass: Duxbury Press, 1979.

24. Miller, J.C.—*Statistics for Advanced Level*. Cambridge University Press—1985.

25. Mosteller, F., R.E.K. Rourke, and G.B. Thomen, Jr.—*Probability with Statistical applications*—Reading, Mass: Addison-Wesley Publishing Co. Inc. 1970.

26. Mood, A.M., F.A. Graybill and D.C. Boes—*Introduction to the theory of Statistics*, 3rd ed. McGraw-Hill Inc. 1974.

27. Ostle, B. and R.W. Mensing—*Statistics in Research*, 3rd ed. The Iowa State University-Press, Ames, Iowa, 1975.

28. Siegel, S.—*Nonparametric Statistics for the Behavioral Sciences*. New York, McGraw-Hill Book Company, 1956.

29. Sprent, Peter - *Applied Nonparametric Statistical Methods*. Chapman and Hall, Ltd., London, 1989.

30. Steel, Robert G.D. and Torrie, James H.—*Principles and Procedures of Statistics*, 2nd ed. McGraw-Hill Book Company, 1980.

31. Walpole, R.E.—*Introduction to Statistics*. 3rd ed. New York: Macmillan Publishing Co. Inc. 1982.

32. Walpole, R.E. and R.H. Myers—*Probability and Statistics for Engineers and Scientists*. 2nd ed. New York. Macmillan Publishing Co. Inc. 1978.

33. Wilburn, A.J.—*Practical Statistical Sampling for Auditors*. New York, Marcel Dekker, Inc. 1984.

34. Wonnacott and Wonnacott—*Introductory Statistics for Business and Economics*, 2nd ed. New York. John Wiley & Sons.

ANSWERS TO EXERCISES

Chapter 14, Pp. 56-65

14.16 The possible distinct samples of size 2 are (2, 4), (2, 6), (2, 8), (2, 10), (4, 6), (4, 8), (4, 10), (6, 8), (6, 10), (8, 10).
The sample means are 3, 4 5, 6, 5, 6, 7, 7, 8, 9.
The sample variances are 1, 4, 9, 16, 1, 4, 9, 1, 4, 1.

$$\mu = 6, \sigma^2 = 8, \mu_{\bar{x}} = 6, E(S^2) = 5.$$

14.22 (a) $n_1 = 8, n_2 = 9, n_3 = 11, n_4 = 12$
(b) $n_1 = 900, n_2 = 1800, n_3 = 1500, n_4 = 1200, n_5 = 600.$

14.26 (b) (i)

$\bar{x}_i$	4,	4.5	5,	5.5,	6,	6.5,	7,	7.5,	8
f_i	1,	2,	3,	4,	5,	4,	3,	2,	1

(ii) $\mu_{\bar{x}} = 6, \sigma_{\bar{x}} = 1.$

14.27 (i) 15

(ii)

$\bar{x}_i$	3	5	6	7	8	9	10
$f(\bar{x})$	1/15,	2/15,	4/15,	2/15,	1/15,	4/15,	1/15

14.28 (a)

$\bar{x}_i$	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
$f(\bar{x})$	1/81,	4/81,	10/81,	16/81,	19/81,	16/81,	10/81,	4/81,	1/81

14.29 (iii) $5.14 \pm 3(1.81)$

14.30 20, $\mu_{\bar{x}} = 9$

14.31

$\bar{x}$	3	4	5	6	7	8	9	10	11	12
$f(\bar{x})$	1/64,	3/64	,6/64	10/64,	12/64,	12/64,	10/64,	6/64,	3/64,	1/64

$\mu = 7.5, \sigma^2 = 11.25;\ \mu_{\bar{x}} = 7.5, \sigma^2_{\bar{x}} = 3.75.$

14.32 (a) $\mu = 7, \sigma^2 = 5.2$; (b) (i) $\sigma^2_{\bar{x}} = 0.87$, (ii) $\sigma^2_{\bar{x}} = 1.73$;
(c) 0.87; (d) 1.73.

14.33 $\sigma^2_{\bar{x}} = 5/3, \sigma^2_{(med)} = 7/2, \mu_{\bar{x}} = 5 = \mu = \mu_{(med)}$

14.35 (a) Approximately normal; (b) 0.0668.

14.36 0.5762.

14.37 (a) $\mu_{\bar{x}} = 68.5, \sigma_{\bar{x}} = 0.54$; (b) 154

14.38 (i) 11, (ii) 30.

14.39 (a) 0.80; (b) (i) $\mu_{\bar{x}} = 5.3, \sigma^2_{\bar{x}} = 0.0225$, (ii) 0.9082

14.40 (b) 0.0764.

14.41 (a) (i) 0.057 (ii) 0.013, (iii) 0.930 (b) 0;

14.42 (b) (i) 0.0127, (ii) 0.8561

14.43

$\bar{x}_1-\bar{x}_2$	0	0.5	1	1.5	2	2.5	3	3.5	4
$f(\bar{x}_1-\bar{x}_2)$	1/81,	2/81,	7/81,	10/81,	17/81,	16/81,	16/81,	8/81,	4/81

$\mu_{\bar{x}_1-\bar{x}_2} = 2.33,\ \sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{7}{9}$

14.44 0.0043, 14.45. 0.7070

14.47 The sampling distribution of proportions is:

$\hat{p}$	0/3	1/3	2/3	3/3
$f(\hat{p})$	1/20	9/20	9/20	1/20

$\mu_{\hat{p}} = 0.5,\ \mathrm{Var}(\hat{p}) = 0.05$

14.48 The sampling distribution of proportions is:

$\hat{p}$	0	1/3	2/3	1
$f(\hat{p})$	1/35	12/35	18/35	4/35

$\mu_{\hat{p}} = 4/7,\ \sigma^2_{\hat{p}} = \frac{8}{147}$

14.49 (a) (i) 0.0877 (ii) 0.0065 (b) 0.0228

14.50. 0.9774

14.51 (b) 0.6826.

14.52

S_i^2	0	0.25	1	2.25	4
$f(S_i^2)$	5/25	8/25	6/25	4/25	2/25

1.

Chapter 15, Pp. 113-122

15.4 (a) $\bar{X} = 5.9$, variance = 5.29, $m_3 = 6.288$
(b) (i) 12.17, 0.21; (ii) 12.41, 0.14.

15.5 (b) $\mu = 15,\ \sigma^2 = 11.67;\ E(\bar{X}) = 15,\ E(s^2) = 11.67.$

15.6 (b) 89%

15.9 (a) (i) $\bar{X}_1$ and $\bar{X}_2$ (ii) 140.6%, 71.1% (b) (i) T_1, T_2 and T_4, (ii) T_1.

15.15 (i) $\hat{p} = \frac{X}{n}$, (ii) $\hat{p} = \frac{\bar{X}}{n}$.

15.16 (b) 33.43, 5.10.

15.17 (a) $\hat{\lambda} = \bar{x}$ (b) $\hat{\sigma}^2 = \frac{\sum x^2}{n}$.

15.20 (b) $77.28 < \mu < 85.12$ 15.21 (b) $63.61 < \mu < 71.45$

15.22 (b) $95.06 < \mu < 104.49$ 15.23 (b) $-2.27 < \mu < 2.67$

15.24 (a) 90% (b) $1010.9 < \mu < 1014.1$

15.25 (b) $182.16 < \mu < 197.84$

15.26 (a) $30 < \mu < 40$; (b) $3.485 < \mu < 3.515$

15.27 $166.42 < \mu < 177.58$

15.28 $\bar{x}_c = 85.08$, $s_p^2 = 7.2372$, $84.62 < \mu < 85.54$.

15.29 (b) $-1.6 \le \mu_1 - \mu_2 \le 0$.

15.31 (b) $4.45 < \mu_1 - \mu_2 \le 15.55$

15.32 (a) $33.26 < \mu_1 < 34.60$; (b) $26.84 < \mu_2 < 28.04$;
(c) $5.59 < \mu_1 - \mu_2 \le 7.39$

15.33 (a) 1.55; (b) $0.26 < \mu_1 - \mu_2 \le 2.84$.

15.34 $53.63 < \mu_1 - \mu_2 \le 70.37$

15.35 (b) $0.053 < p < 0.107$, (c) 0.1206 to 0.3060

15.36 (a) $0.504 < p < 0.696$ (b) $0.592 < p < 0.664$, (c) $0.701 < p < 0.819$

15.37 (b) $-0.088 < p_1 - p_2 < 0.188$.

15.38 $-0.1414 < p_1 - p_2 < -0.0986$.

15.39 (a) 0.556, (b) 1005.

15.40 (a) 103; (b) 97; (c) 225

Chapter 16, Pp. 161-168

16.5 (b) $\alpha = 0.0548$; $\beta = 0.3504$, 0.6177, 0.8281

16.6 (a) (i) 0.323, (ii) 0.583 (b) (i) 195; (ii) 0.0548; (iii) 0.9452.

16.7 (b) $H_0 : \mu \le 800$ and $H_1 : \mu > 800$; $\bar{X}$ is the test-statistic.

16.8 β are: 0.1977, 0.4920, 0.7939, 0.9995
Powers are: 0.8023, 0.5080, 0.2061, 0.0005

16.10 (a) $z_1 = -11.96$, $z_2 = -8.04$ so $\beta < 0.2$; (b) $n = 19$

16.11 $n = 78$.

16.12 (b) $z = 4.67$; reject $H_0 : \mu = 75$.

16.13 (b) $z = -2.88$; reject H_0 in favour of $H_1 : \mu < 45$.

16.14 (a) $z = 1.23$; accept H_0. (b) $z = -1.93$; reject H_0.

16.15 (a) $z = -0.45$; accept H_0, (b) $z = -2$; reject H_0

16.16 $z = 8.97$; reject H_0.

16.17 (a) $z = -4.69$. The sample would not be so regarded.
(b) $z = -1.78$. The sample can be so regarded.

16.18 (a) $z = 5$. The process is out of control
(b) $z = -2.5$; reject the claim.

16.19 (a) $z = 0.534$; accept H_0 (b) $z = 4.22$; reject H_0

16.20 $z = -2.33$; reject H_0

16.21 $z = -2.48$, significant at 5%; insignificant at 1%.

16.22 $z = -2.00$; accept H_0

16.23 (a) $z = 9.92$; reject H_0. Australians are on the average taller than Englishmen.

(b) $z = 4.24$; reject H_0. The two brands differ in quality.

16.24 $z = 2.16$, accept H_0

16.25 (b) $z = 1.94$; reject H_0. It could be concluded that his shooting has improved.

16.26 (a) $z = 2.63$; reject H_0. The coin in biased.

(b) $z = 0.51$; accept H_0. The data are consistent with an equal sex division in the population.

16.27 $z = 3.59$; reject H_0 in favour of H_1.

16.28 (a) $z = 0.487$; accept H_0. The given information is consistent with the hypothesis.

(b) $z = -1.80$; reject H_0.

16.29 (a) $z = -4.73$; reject H_0. The claim is not legitimate.

(b) $z = 3.21$; reject the claim.

16.30 $z = -5.7$; reject H_0. The company's claim cannot be accepted.

16.31 (a) $z = 1.27$; accept H_0. The difference is not significant.

(b) $z = -3.12$; reject H_0.

16.32 (a) $z = 0.10$; accept H_0. The machine has not been improved.

(b) $z = 2.10$; reject H_0. The blue envelope will help sales.

16.34 $z = -1.025$; accept H_0

16.35 Accept $H_0 : p = 0.05$. The coin is fair.

Chapter 17, Pp. 222-238

17.4 (b) (i) 55.47, 78.80, 129.64. (ii) 55.57, 79.10, 129.93.

17.5 (b) $13.82 < \sigma^2 < 65.53$.

17.6 (a) $0.135 < \sigma^2 < 0.953$ (b) $0.0286 < \sigma^2 < 0.213$

17.7 (a) $15.00 < \sigma^2 < 64.05$ (b) $9.34 < \sigma^2 < 49.55$

17.8 $s_p^2 = 30.83$; $24.20 < \sigma^2 < 40.83$.

17.9 (b) $\chi^2 = 10$; accept H_0.

17.10 (a) $\chi^2 = 15.75$; accept $H_0 : \sigma^2 = 20$; $8.65 < \sigma^2 < 22.74$

(b) $\chi^2 = 33.33$; accept H_0.

17.11 (a) $\chi^2 = 24.108$; reject H_0 at 5% and accept H_0 at 1% level of significance.

(b) $\chi^2 = 8.71$; accept H_0.

17.12 (a) $\chi^2 = 46.2$; accept H_0 (b) $\chi^2 = 16$; accept H_0 that $\sigma = 0.9$ years.

17.13 (b) $u = 1.46$. Population variances may be regarded homogeneous.

17.14 (b) $u = 0.47$; accept the hypothesis of equal variances.

17.18 (a) $\chi^2 = 5.00$; accept H_0. (b) $\chi^2 = 34.73$; reject H_0

17.19 (a) $\chi^2 = 4.23$; accept H_0. (b) $\chi^2 = 10.144$; reject H_0

17.20 $\chi^2 = 3.78$; accept H_0. The data are consistent with theory.

17.21 (a) $\chi^2 = 4.50$; reject H_0. The coin is not fair.
(b) $\chi^2 = 4.07$; reject H_0. The dice are not fair.

17.22 (a) $\chi^2 = 0.37$ or $z = 0.61$; accept H_0 of equal sex division.
(b) $\chi^2 = 2.60$ or $z = -1.94$. Unable to say that treatment is effective.

17.23 (a) $\chi^2 = 8.47$; reject H_0. The data contradict the stated hypothesis.
(b) $\chi^2 = 10.0$; reject H_0. The distribution of grades is not uniform.

17.24 (a) $\chi^2 = 23.28$; reject H_0. (b) $\chi^2 = 57.90$; reject H_0.

17.25 (a) $\chi^2 = 18.73$; reject H_0 that die is balanced.
(b) Assuming that the total number of births is *evenly* spread over the period (*i.e.*, on the basis of number of days in the various months), the expected number of births is 50,959 for a month of 31 days, 49,315 for a month of 30 days and 46,027 for February.
$\chi^2 = 322.62$; reject H_0. The data suggest the presence of seasonality.

17.26 (b) $\chi^2 = 19.63$. The data are not consistent with the hypothesis.

17.27 (b) $\chi^2 = 2.79$. The data conform to the binomial distribution.

17.28 $\chi^2 = 5.80$. The dice were unbiased.

17.29 (a) $\chi^4 = 4.74$; accept H_0. The coins are well-balanced.
(b) $\chi^2 = 1.12$; accept H_0. The fit is good.

17.30 $\chi^2 = 1.39$. The fit is good.

17.31 $\chi^2 = 3.52$. The data fit a Poisson distribution.

17.32 $\chi^2 = 21.04$; reject H_0. The data do not fit a Normal distribution.

17.33 $\chi^2 = 2.101$; accept H_0. The data follow a Normal distribution.

17.34 (b) $(\alpha BC) = 12$, $(A\beta C) = 16$, $(AB\gamma) = 32$, $(A\beta\gamma) = 2$, $(\alpha\beta C) = 107$, $(\alpha B\gamma) = 66$, $(\alpha\beta\gamma) = 25$ and $(ABC) = 240$.

17.35 125.

17.36 (b) $(\alpha\beta\gamma) = -57$. The numbers reported in the various groups are not consistent.

17.37 Attributes are independent.

17.38 (i) Positive association. (ii) Negative association. (iii) A and B are independent.

7.41 (b) $\chi^2 = 5.489$; reject H_0. The results are associated.

17.42 (a) $\chi^2 = 57.25$; reject H_0. The two classifications are associated.
(b) $\chi^2 = 4.87$; reject H_0.

17.43 $\chi^2 = 30.50$. Attributes are highly associated.

17.44 (a) $\chi^2 = 1.85$; accept the hypothesis of independence.
(b) $\chi^2 = 26.66$, reject hypothesis of independence.

17.45 $\chi^2 = 162$; accept H_0.

17.46 $\chi^2 = 26.6$. There is association.

17.47 $\chi^2 = 69.15$; reject H_0

17.48 $\chi^2 = 134.55$; reject H_0.

17.49 $\chi^2 = 51.28$; reject H_0. The claim status is not independent of the policyholder's age.

17.50 $\chi^2 = 7.47$; accept H_0 that two variables are independent.

17.52 $\chi^2 = 6.11$, $C = 0.12$. Accept H_0.

17.53 χ^2 (without Yates' correction) $= 4.62$; χ^2 (with Yates' correction) $= 4.01$; reject H_0.

17.54 (b) $\chi^2 = 0.57$; accept H_0.

17.55 (b) Reject the null hypothesis.
(c) $p = 0.24$; reject H_0

17.56 (a) $p = 0.1098$;
(b) $p = 0.3114$; not independent.

17.57 (a) $\chi^2 = 6.29$; accept H_0 that the proportion of defectives is about the same for all three shifts.
(b) $\chi^2 = 110$; reject H_0.

17.58 (b) $\chi^2 = 9.65$;

17.59 $\chi^2 = 27.95$; reject H_0.

17.60 $\chi^2 = 6.52$; accept H_0 that there is no significant difference in two groups.

Chapter 18, Pp. 265-272

18.4 (i) -5, (ii) 4.5, (iii) 9, (iv) $= 25$

18.5 (b) $569 < \mu < 581$.

18.6 (a) $9.84 < \mu < 12.16$ when σ is known; $9.48 < \mu < 12.52$ when σ is unknown.
(b) $(-2.27, 2.67)$ when σ is known; $(-1.68, 1.88)$ when σ is not known.

18.7 (a) $87.0 < \mu < 107.4$.
(b) $12.3 < \mu < 16.7$.

18.8 $33.52 < \mu < 35.28$

18.9 (a) $0.3 < \mu_1 - \mu_2 < 15.80$

(b) $-4.6 < \mu_1 - \mu_2 < 16.0$

18.10 $8.80 < \mu_1 - \mu_2 < 15.80$

18.11 $0.50 < \mu_1 - \mu_2 < 1.74$

18.13 (a) $t = 1.0$; accept H_0. (b) $t = 1.5$; accept H_0.

(c) $t = 0.44$; accept H_0 (d) $t = -1.25$; accept H_0.

(e) $t = -0.74$; accept H_0.

18.14 $t = -4.45$; reject H_0. The given values are not consistent.

18.15 (a) $t = 1.84$; accept H_0. (b) $t = -1.71$; accept H_0.

18.17 $t = 3.16$; reject H_0 at 0.05 level but not at 0.01 level. It is advisable to check the machines.

18.18 $t = -0.36$; accept H_0. The claim is vindicated.

18.19 (b) $t = 0.42$; accept H_0.

18.20 $t = 1.007$; accept H_0.

18.21 (a) $t = -0.099$; accept H_0. The soldiers are on the average not taller than sailors.

(b) $t = 3.05$; reject H_0. Electrification does exert some effect on tillering.

18.22 $t = -0.68$; accept H_0.

18.23 (b) $t = 1.386$; accept H_0. (c) $t = 2.608$; reject H_0.

18.24 (a) $t = 1.27$; accept H_0.

(b) $t = -0.39$; accept H_0. No significant difference in the strength of the two tyeps of ropes.

18.25 $t = 1.32$; accept H_0.

18.26 $t' = 2.82$ with $v = 14$; reject the hypothesis of equal means.

18.27 (b) $t = 1.17$; accept H_0.

18.28 (a) $t = 3.44$; reject H_0.

(b) $t = 0.86$; accept H_0. The same conclusion has not been reached.

18.29 (a) $t = 2.16$; reject H_0. Food B is better than food A.

(b) $t = 4.32$; reject H_0. Food B is better than food A.

18.30 $t = 1.71$; accept H_0.

18.31 $t = 2.48$; reject H_0. Cars equipped with radial tires give better fuel economy.

18.32 $t = -1.95$; accept H_0.

Chapter 19, Pp. 290 -293

19.4 (b) $1.30 < \sigma_1^2/\sigma_2^2 < 7.50$.

(c) $0.68 < \sigma_1^2/\sigma_2^2 < 6.59$.

19.5 (b) $0.36 < \sigma_1^2/\sigma_2^2 < 7.95$; $0.60 < \sigma_1/\sigma_2 < 2.82$.

(c) $0.152 < \sigma_2^2/\sigma_1^2 < 1.84$.

19.6 (b) (i) $F = 3.125$; reject H_0.

(ii) $F = 2.48$; accept H_0 (the role of two samples interchanged).

(iii) $F = 0.47$; reject H_0 (two-tailed test).

19.7 (a) $F = 5.33$; reject H_0 (two-tailed test)

$F = 0.18$; accept H_0 (one-tailed test).

(b) $F = 1.57$. Two methods of teaching are equally variable.

19.8 $F = 1.28$; accept H_0.

19.9 $F = 4.287$; do not reject H_0.

19.10 (a) $F(6, 5) = 1.23$; difference is not significant at $\alpha = 0.05$ (two tailed).

(b) $t = 0.54$; difference is not significant at $\alpha = 0.05$.

19.11 $F = 5.88$; variances are not equal.

19.12 $F = 1.69$; accept H_0.

Chapter 20, Pp. 336-346

20.7 (a) The analysis of variance table is

Source of Variation	*d.f.*	*SS*	*MS*	*F*
Between samples	3	186	62.00	2.88
Within samples	8	172	21.50	
Total	11	358	--	

20.8 (a) The analysis of variance table is

Source of Variation	*d.f.*	*SS*	*MS*	*F*
Between machines	3	1166.95	388.98	6.00
Within machines	16	1037.60	64.85	
Total	19	2204.55	--	

The machines are significantly different with respect to items produced.

20.9 (a) The *ANOVA* Table is:

Source of Variation	*d.f.*	*SS*	*MS*	*F*
Between salesmen	2	3426	1713	5.56
Within salesmen	6	1726	287.67	
Total	8	5152	--	

Reject H_0. The differences between salesmen are significant.

(b) The *ANOVA* Table is:

S.V.	d.f.	SS	MS	F
Treatments	3	359.19	119.73	5.95
Residuals	12	258.25	21.52	
Total	15	617.44	--	

The treatments differ signficiantly.

20.10 The *ANOVA* Table is:

S.V.	d.f.	SS	MS	F
Between Tube types	2	190.08	95.04	2.86
Error	21	697.75	33.23	
Total	23	887.83	--	

Accept H_0.

20.11 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Between Samples	4	1225.6	306.4	3.59
Within Samples	35	2986.0	85.31	
Total	39	4211.6	--	

Reject H_0.

20.12 The *ANOVA* table is:

S.V.	d.f.	SS	MS	F
Between Groups	2	91.40	45.70	2.10
Within Groups	15	327.04	21.80	
Total	17	418.44	--	

Accept H_0.

20.13 The analysis of variance table is

S.V.	d.f.	SS	MS	F
Between Methods	2	22.38	11.19	6.6
Within Methods	10	16.99	1.70	
Total	12	39.37	--	

Reject H_0. The methods differ significantly.

20.14 The *ANOVA* table is:

S.V.	d.f.	SS	MS	F
Between Teachers	2	124.55	62.28	0.21
Within Teachers	29	8720.33	300.70	
Total	31	8844.88	--	

Accept H_0. There is no significant difference.

20.15 (a) $u = 2.18$. The variances are homogeneous.

(b) The analysis of variance table is

S.V.	d.f.	SS	MS	F
Between samples	3	349.62	116.54	1.18
Within samples	20	1967.71	98.39	
Total	23	2317.33	--	

Accept H_0: all means are equal.

20.16 (b) The *ANOVA* table is:

S.V.	d.f.	SS	MS	F
Between samples	3	1084.95	361.65	24.72
Within samples	16	234.02	14.63	
Total	19	1318.97	--	

Reject H_0.

20.17 (a) The *ANOVA* table is:

S.V.	d.f.	SS	MS	F
Types of plants	2	14.9452	7.4726	13.70
Error	51	27.8135	0.5454	
Total	53	42.7587	--	

The rubber content of the three types of plants is different.

(b) (i) $t = 0.26$; accept H_0. (ii) $|t| = 5.05$ with 51 d.f.

20.18 The analysis of variance table is

S.V.	d.f.	SS	MS	F
Between schools	2	2577	1288	13.49
Within schools	272	25977	95.5	
Total	274	28554	--	

Reject H_0. The difference between the means of students in the different types of school is significant.

20.19 The *ANOVA* table is:

S.V.	d.f.	SS	MS	F
Between Groups	3	87.73	$s_1^2 = 29.24$	6.22
Between Rations	2	54.12	$s_2^2 = 27.06$	5.76
Error	6	28.21	$s_3^2 = 4.70$	
Total	11	170.06	--	

There is a significant difference between rations at the 5% level.

20.20 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Factor A	4	392	98	0.97
Factor B	3	330	110	1.09
Error	12	1212	101	--

Accept both the hypotheses.

20.21 The ANOVA table is:

S.V.	d.f.	SS	MS	F
Between Breeds	3	781.75	260.75	$F_1=6.81$
Between Rations	2	438.04	219.02	$F_2=5.72$
Error	6	229.63	38.37	
Total	11	1449.42	--	

There is significant difference both in breeds and between rations.

20.22 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Between fertilizers	4	46.20	11.55	2.03
Between varieties	3	28.55	9.52	1.68
Error	12	68.20	5.68	
Total	19	142.95	--	

These data could have arisen from a population in which there was no difference between the yields of varieties and the fertilizers did not differ in their effect.

20.23 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Between subjects	4	45.28	11.32	4.86
Between treatments	3	23.24	7.75	$F=3.33$
Error	12	27.94	2.33	--
Total	19	96.46	--	--

There is no difference between the treatment means.

20.24 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Between Salesmen	3	600	200	0.5
Between Districts	2	3200	1600	4.0
Error	6	2400	400	--
Total	11	6200	--	

The salesmen were equally capable and that all districts were equally profitable to work.

20.25 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Between Subjects	2	595.60	297.80	2.93
Between Students	4	1329.73	332.43	3.27
Error	8	813.07	101.63	--
Total	14	2738.40	--	

Accept both the hypotheses.

20.26 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Columns	2	20.33	10.165	7.04
Rows	1	56.89	56.89	39.4
Interaction	2	1.45	0.725	0.5
Error	12	17.33	1.444	--
Total	17	96.00	---	---

20.27 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Diets	3	81.64	27.213	173.33
Drugs	2	12.65	6.325	40.29
Interaction	6	0.68	0.113	0.72
Error	12	1.88	0.157	---
Total	23	96.85	---	---

Differences between drugs are significant but the differences due to the interaction of drugs and diets are not significant.

20.30 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Varieties	4	270.27	67.57	6.39
Error	10	105.68	10.57	---
Total	14	375.95	---	---

Reject H_0. ($F_{0.05}(4,10) = 3.48$)

20.31 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Between Coater types	3	1.53	0.51	5.67
Between Days	2	0.21	0.10	1.11
Error	6	0.54	0.09	--
Total	11	2.28	--	

Difference between coater types is significant at the 5% level.

	M	L	A	K
$\bar{Y}_{.j}$:	4.27	4.73	4.87	5.27

20.35 (i) The *ANOVA*-table is

S.V.	d.f.	SS	MS
Between subgrade	3	120	40
Error	16	160	10
Total	19	280	--

(ii) A fixed-effects model: $Y_{ij} = \mu + \tau_j + e_{ij}$.

(iii) $F = 4$; reject H_0 at the 5% level.

(iv) One set of orthogonal contrasts is:

$$C_1 = T_{\cdot 1} - T_{\cdot 2}$$
$$C_2 = T_{\cdot 1} + T_{\cdot 2} - 2\,T_{\cdot 3}$$
$$C_3 = T_{\cdot 1} + T_{\cdot 2} + T_{\cdot 3} - 3\,T_{\cdot 4}$$

Chapter 21. Pp. 388-396

21.3 $1.03 < \beta < 2.09$.

21.4 (i) $b = 4.4096$.

(ii) $-263.08 < \alpha < -10.11$ or $159.87 < A < 168.93$ if $\mu_{Y \cdot X} = A + B(X - \bar{X})$; $2.56 < \beta < 6.26$

21.5 (a) $\hat{Y} = 99.13 + 0.502X$, or $\hat{Y} = 121.3 + 0.502\,(X - 44.21)$

(b) (i) $86.7 < \alpha < 111.55$ or $118.16 < A < 124.44$ if $\mu_{Y \cdot X} = A + B(X - \bar{X})$; $0.2258 < \beta < 0.7782$

(ii) Prediction interval: $97.89 < Y < 150.57$

21.6 (i) $0.92 < \beta < 1.09$; $29.34 < \mu_{Y \cdot X = 30} < 30.46$

(ii) $t = 0.119$; accept H_0

21.7 $\hat{Y} = -179.41 + 5.03X$; $t = 6.89$, reject H_0.

21.8 $\hat{Y} = -1 + 2X$, $t = 4.35$, reject H_0.

21.9 (i) $t = 9.23$; reject $H_0 : \beta = 0$.

(ii) $t = -3.188$; reject $H_0 : \beta = 6$.

(iii) 156.49; $139.26 < Y_i < 173.72$ (prediction interval)

21.10 $\hat{Y} = 30.056 + 0.897X$; (i) $t = 5.37$, reject H_0.

(ii) $t = -0.167$, accept $H_0 : \alpha = 32$.

21.11 (b) (i) $b_1 = 1$, $b_2 = 1$,

(ii) $-4.14 < \beta_1 - \beta_2 < 4.14$; t = 0 so accept $H_0 : \beta_1 = \beta_2$.

21.12 (i) $b_1 = 1, b_2 = 1, b_3 = 0.5454$.

(ii) For $H_0 : \beta_1 = \beta_2$, $t = 0.43$, accept H_0. For $H_0 : \beta_1 = \beta_3$, $t = 0.33$; accept H_0. For $H_0 : \beta_2 = \beta_3$, $t = 0.43$; accept H_0

21.13 $F = 14.60$; regression is non-linear.

21.14 $F = 0.60$; regression is linear.

21.15 $F = 1.36$; regression is linear.

21.16 (b) $0.100 < \rho < 0.297$

21.17 (a) $0.30 < \rho < 0.78$ (b) $0.503 < \rho < 0.640$

21.18 (a) $0.47 < \rho < 0.87$ (b) $0.771 < \rho < 0.973$

21.19 (a) $z = -1.19$; accept H_0. (b) $z = 1.62$; accept H_0

21.20 (a) $z = -1.09$; accept H_0; values are consistent with hypothesis

(b) $z = -1.49$; accept H_0.

21.21 (b) $t = 2.57$; reject H_0 in favour of H_1.

21.22 (a) $z = -1.03$; accept H_0 (b) $t = 2.77$; value is significant
(c) $z = -1.19$; accept H_0.

21.23 (a) $t = 2.79$; reject H_0. (b) $t = 1.58$; accept H_0.

21.24 (a) $t = 1.07$; accept H_0 (b) $t = 3.53$; reject H_0.
(c) $F = 27.97$; reject $H_0 : \rho = 0$.

21.25 (a) (i) 0.32; (ii) 0.41 (b) 0.38; (c) $n = 37$.

21.26 (b) 0.397; $u = 0.44$, correlations are homogeneous.

21.27 0.508; $u = 0.81$, accept H_0.

21.28 (a) $t = 2.31$; reject H_0 (b) $t = 2.51$; reject H_0

21.29 (a) $t = 2.37$; reject H_0. $F = 5.62$; reject H_0.
(b) $u = 1.64$; accept H_0. Correlations are homogeneous.

21.30 (a) $r_{12} = -0.89$, $r_{13.2} = -0.905$, $R_{1.23} = 0.98$
(b) Correlations are significant.

21.31 (a) $F = 1.19$; accept H_0 (b) $F = 1.33$; accept H_0.

21.32 (b) The analysis of variance table for regression is:

S.V.	d.f.	SS	MS	F
Regression	1	1503.5	1503.5	338.6
Residual	8	35.5	4.44	---
Total	9	1539.0	--	--

$b = 0.98$, $s_b = 0.053$ and $t = -0.38$; accept $H_0 : \beta = 1$ at $\alpha = 0.01$

21.33 (b) The analysis of variance table for regression is:

S.V.	d.f.	SS	MS	F
Regression	1	2083.94	2083.24	85.34
Residual	3	73.26	24.42	---
Total	4	2157.20	--	--

Reject H_0: The variables are related.

21.34 (i) $\hat{Y} = 90 - 8X$
(ii) Regression $SS = \sum(\hat{Y} - \bar{Y})^2 = 6400$; Residual $SS = 3600$.
(iii) $t = -8$; reject H_0. (iv) Adjusted $Y = 52$

21.35 The analysis of variance for regression is:

S.V.	d.f.	SS	MS	F
Regression	2	3949.7	1974.85	101.64
Residual	3	58.3	19.43	---
Total	5	4008.0	--	--

Reject H_0: $\beta_1 = \beta_2 = 0$ at $\alpha = 0.05$.

21.36 $\hat{\beta}_0 = 3.88$, $\hat{\beta}_1 = 2.09$, $\hat{\beta}_2 = 2.65$

ANOVA for overall significance test

S.V.	d.f.	SS	MS	F
Regression	2	70.69	35.345	17.20
Error	2	4.11	2.055	---
Total	4	74.80	--	--

Accept H_0: $\beta_1 = \beta_2 = 0$ at $\alpha = 0.05$.

Chapter 22, Pp. 414-418

22.8 The analysis of covariance table is:

S.V.	d.f.	Sum of Squares and Products			Adjusted Results			
		Σx^2	Σy^2	Σxy	$SS\left(=\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}\right)$	df	MS	F
Between Treatments	1	6.125	15.125	9.625	---	--	--	--
Within (Error)	6	4.750	12.750	4.000	9.382	5	$s_e^2 =$ 1.876	
Total $(T+E)$	7	10.875	27.875	13.625	10.805	6	--	--
Treatments adjusted					1.423	1	$s_t^2 =$ 1.424	0.76

$b = 0.842$. To test $H_0 : \beta = 0$, $F = 1.79$; reject H_0.

22.9 The analysis is:

S.V.	d.f	Σx^2	Σy^2	Σxy	Regression
Between means	3	7009.75	2826.69	−1376.12	
Within means	12	4674.00	2675.75	1862.75	0.3985
Total	15	11,683.75	5502.44	486.62	0.0416

S.V.	Regression		Adjusted SS		MS
	d.f.	SS	d.f.	SS	
Total	1	20.27	14	5482.17	--
Error	1	742.37	11	1933.38	175.8
Means	---	---	3	3548.79	1182.9

Using unadjusted y values, $F = 4.23$; reject the null hypothesis

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D.$$

By analysis of covariance, $F = 6.73$; reject $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ after adjustment for quantity of food.

22.10 The analysis of covariance table is:

S.V.	d.f.	Sum of Squares and Products			Adjusted Results			
		$\sum x^2$	$\sum t^2$	$\sum xt$	$SS\left[\sum x^2 - \frac{(\sum xt)^2}{\sum x^2}\right]$	df	MS	F
Treatments	3	2.18	5.50	−0.49				--
Error	15	9.13	96.22	28.95	9.13−8.71=0.42	14	0.03	
Total (T+E)	18	11.31	101.72	28.46	11.31−7.96=3.35	17	--	--
Treatments adjusted					3.35−0.42=2.93	3	0.98	32.67

To test $H_0 : \beta = 0$; $F = 290.33$; reject H_0. $s = 0.11$

22.11 The Analysis of Covariance table is:

S.V.	d.f.	Sum of Squares and Products			Adjusted Results			
		$\sum x^2$	$\sum y^2$	$\sum xy$	$\sum y^2$	df	MS	F
Between Groups	2	59.73	31.6	25.6	--	---	--	--
Within Groups	12	53.20	58.8	34.6	36.30	11	3.30	--
Total	14	112.93	90.4	60.2	58.31	13	--	--
Treatments adjusted					22.01	2	11.005	3.33

Reject H_0 at $\alpha = 0.05$.

22.12 The Analysis of Covariance table is:

S.V.	d.f.	Sum of Squares and Products			Adjusted		
		$\sum x^2$	$\sum y^2$	$\sum xy$	$\sum y^2$	df	MS
Treatment	9	4684	112	-532	--	---	--
Error	27	3317	516	-650	388.63	26	14.95
Total (T+E)	36	8001	628	-1182	453.38	35	--
Treatments adjusted					64.75	9	7.19

To test $H_0 : \beta = 0$, $F = 8.52$: reject H_0.

To test H_0 : there is no differences between means, $F = 2.28$; accept H_0.

22.14 The Analysis of Covariance is:

S.V.	d.f.	Sum of Squares and Products			Adjusted		
		$\sum x^2$	$\sum y^2$	$\sum xy$	$\sum y^2$	df	MS
Total	15	291	131.44	165.75	--	--	--
Blocks	3	163.5	52.69	92.50	--	--	--
Treatments	3	54.0	26.19	24.00	--	--	--
Error	9	73.5	52.56	29.24	19.56	8	2.445
(T+E)	12	127.5	78.75	73.25	36.37	11	--
Treatments adjusted					17.11	3	5.703

To test $H_0 : \beta = 0$, $F = 13.50$: reject H_0.

To test the hypothesis of no difference in adjusted Y-values, $F = 2.33$

22.15 (a) $b = 0.0238$. $F = 67.36$; reject $H_0 : \beta = 0$.

(b) The analysis of covariance table then becomes

S.V.	d.f.	Sum of Squares and Products			Adjusted		
		Σx^2	Σy^2	Σxy	Σy^2	df	MS
$T + E$	36	144685.4	136.09	4280.25	9.47	35	--
E	30	28665.1	23.23	682.20	6.99	29	0.241
Treatments adjusted					2.48	6	0.413

$F = 1.71$; the differences among the treatment means for Y after adjusting for variation attriuted to X are not significant.

Chapter 23, Pp. 464-476

23.4 (b) The *ANOVA*-Table is:

S.V.	d.f.	SS	MS
Between Levels	2	7268	3634
Within Levels	12	493	$s_e^2 = 41$
Total	14	7761	--

$S.E. = 2.87$

23.5 The analysis is:

S.V.	d.f.	SS	MS	F
Between Varieties	6	41.6	6.93	1.1
Within Varieties	28	174.0	6.21	---
Total	34	215.6	--	--

The data do not indicate significant difference.

23.6 Hypothesis H_0 : There is no difference in the effects of storage conditions, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

The *ANOVA*-Table is:

S.V.	d.f.	SS	MS	F
Between conditions	4	12.08	3.02	3.32
Within conditions	9	8.21	0.91	---
Total	13	20.29	--	--

Accept H_0.

23.7 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Between Formats	2	257.42	128.71	31.24
Within Formats	28	115.27	4.12	---
Total	30	372.69	--	--

Significance level $\cong 0.001$.

23.9 (b) The *ANOVA*-Table is:

S.V.	d.f.	SS	MS	F
Between Varieties	2	400.40	200.20	2.09
Between Replication	3	528.97	176.32	1.87
Error	6	575.05	95.84	--
Total	11	1504.42	--	--

Accept the hypotheses.

23.10 (b) The *ANOVA*-Table is:

Source	d.f.	SS	MS	F
Treatments	3	168	56	3.32
Blocks	3	72	24	1.42
Error	9	152	16.89	--
Total	15	392	--	--

Accept the hypotheses.

(b) If no blocking had been done, then the *ANOVA*-Table would become

S.V.	d.f.	SS	MS	F
Between Treatments	3	168	56	3.00
Within Treatments	12	224	18.67	---
Total	15	392	--	--

Accept H_0.

23.11 (a) The *ANOVA*-Table is:

Source	d.f.	SS	MS
Treatments	2	49884.1	24942.1
Blocks	5	149700.4	29940.1
Error	10	18724.6	1872.5
Total	17	218309.1	--

$F = 13.3$: reject H_0.

(b) The *ANOVA*-Table is:

S.V.	d.f.	SS	MS
Treatments	2	49884.1	24942.1
Error	15	168425.0	11228.0
Total	17	218309.1	--

$F = 2.22$: accept H_0.

23.12 (b) $F = 7.67$; reject H_0. The treatment means are significantly different.

23.13 (a) F (blocks) = 8.30; significant at $\alpha = 0.05$;
F (fertilizers) = 6.11; significant at $\alpha = 0.05$;

(b) (i) $F = 17.37$; significant at $\alpha = 0.05$.

(ii) $F = 0.96$; not significant.

23.14 The analysis of variance is:

S.V.	d.f.	SS	MS	F
Treatments	3	187035	62345	31.68
Blocks	3	220563	73521	--
Error	9	17716	1968	---

23.15 (a)

S.V.	MS	F
Blocks	0.0453	--
Treatments	0.3995	17.3
Error	0.0231	---

(b) *S.E.* for one treatment mean = 0.048; and

S.E. for difference between two treatment means = 0.068

(c) *LSD* = 0.140. The treatment means are arranged in ascending order and a line is drawn under the set for means that are not significantly different.

$\bar{Y}_{.j}$:	T_2	T_3	T_1	T_4
	1.195	1.325	1.464	1.662

(d) 122%

23.17 (b) $x = 6$. The analysis is:

S.V.	d.f.	SS	MS	F
Between Tests	4	17.6	4.4	38.6
Between Ages	2	5.2	---	---
Error	7	0.8	0.114	---

23.18 The missing values are $y_{41} = 15.8$, $y_{22} = 14.2$, $y_{24} = 13.9$. The analysis would be

S.V.	d.f.	SS	MS	F
Blocks	3	58.34	19.45	--
Treatments	5	12.75	2.55	0.38
Error	12	79.73	6.64	---
Total	20	150.82	--	--

ANSWERS

23.19 (b) $m = 6.5$. Treatment $MS = 6.66$ and Error $MS = 0.51$.

23.22 (b) The *ANOVA*-Table is:

S.V.	d.f.	SS	MS	F
Rows	3	0.12	0.04	--
Columns	3	0.70	0.23	--
Treatments	3	7.67	2.557	44.86
Error	6	0.34	0.057	---
Total	15	8.83	--	--

The varieties are significantly different at the 5-percent level.

23.23 F (districts) = 23.47; the variation among districts is significant.
F (heights) = 4.72; no significant variation among the heights.
F (tests) = 3.09; no difference between sensitivity of the tests.

23.24 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Positions	3	1558.5	519.5	---
Runs	3	777.5	259.2	---
Treatments	3	4548.5	1516.2	22.70
Error	6	400.5	66.8	---

The grades of leather are significantly different.

23.25 $F = 20.36$. $LSD = 15.57$. The treatment No. 1 is to be adjusted by 27 to get a uniform product.

23.26 The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Columns	4	183.7584	45.9396	---
Rows	4	141.0784	35.2696	---
Treatments	4	348.2384	87.0596	3.44
Error	12	304.0952	25.3413	---

Reject H_0. Treatment means differ significantly.

23.27 (a) The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Columns	4	13.02	3.26	---
Rows	4	46.95	11.74	---
Treatments	4	190.89	47.72	18.9
Error	12	30.32	2.53	---

The variation due to treatments is highly significant.

(b) The analysis of variance table is:

S.V.	d.f.	SS	MS	F
Treatments	4	190.89	47.72	10.58
Error	20	90.29	4.51	---

Reject H_0.

23.28 F (periods) = 9.19; significant difference in periods.
F (conditions) = 6.62; conditions are significantly different.
F (monkeys) = 17.09; significant difference in monkeys.
F (comparison) = 2.23; data fail to support the conjecture.

23.37 The analysis is:

S.V.	d.f.	SS	MS	F
Treatments	3	697.75	232.58	21.54
A	1	594.05	594.05	55.00
B	1	2.45	2.45	0.23
AB	1	101.25	101.25	9.38
Error	16	172.80	10.80	---
Total	15	870.55	--	--

23.38 The *ANOVA*-Table is:

S.V.	d.f.	SS	MS	F
Columns	3	2.25		
Rows	3	16.75		
Treatments	3	530.75	176.92	24.0
N	1	342.25	342.25	46.7
P	1	182.25	182.25	24.7
NP	1	6.25	6.25	0.8
Error	6	44.00	7.33	---
Total	15	593.75	--	--

23.39 The interactions (being regarded as trivial) mean squares may be used to provide an estimate of error mean square.

23.40 The *ANOVA*-Table is:

S.V.	d.f.	SS	MS	F
Replication	3	75.84		
Treatments	7	329.46	47.07	3.54
N	1	19.53	19.53	1.48
P	1	81.28	81.28	6.15
K	1	124.03	124.03	9.37
NP	1	9.03	9.03	0.68
NK	1	7.03	7.03	0.53
PK	1	19.53	19.53	1.48
NPK	1	69.03	69.03	5.22
Error	21	277.92	13.23	---
Total	31	683.22	--	--

23.41 $SSA = 2.6667$, $SSB = 170.6667$, $SSC = 104.1667$
$SS(AB) = 1.5$, $SS(AC) = 42.6667$, $SS(BC) = 0$, $SS(ABC) = 15$

23.42 (b) The covariance table is:

S.V.	Regression d.f.	Regression SS	Adjusted SS d.f.	Adjusted SS SS	MS	F
T+E	1	364.19	6	390.37	--	--
E	1	86.01	5	310.36	62.07	--
T	(By diff)		3	80.01	26.67	0.43

Chapter 24. Pp. 510-515

24.3 (b) (i) $z = -1.02$; accept H_0. (ii) $z = -1.17$; accept H_0.

24.4 (b) (i) $z = -0.95$; accept H_0. (ii) $T = 13$; accept H_0.

24.5 (i) $P(X \le 2) = 0.1445$; accept H_0. (ii) $T = 6$; accept H_0.

24.6 (b) $\chi^2 = 1.6$; accept H_0.

24.9 $R = 26$; accept H_0. The difference between the two types of paint is not significant.

24.10 $z = -1.84$; reject H_0.

24.11 (a) $U = 7$; accept H_0 (b) $z = -0.56$; accept H_0.

24.12 $z = -1.84$; reject H_0.

24.15 (b) $\chi^2 = 2.33$; accept H_0.

24.16 No significant difference.

24.17 $\chi^2 = 3.877$, medians are equal.

24.18 (a) There exists nonrandomness.
(b) $z = 1.78$; accept the hypothesis of a random sample.

24.19 (a) $n_r = 8$; sequence is random. (b) $z = -0.51$; accept H_0.

24.20 (b) $D = 0.24$; accept H_0.

24.21 $D = 0.1485$; accept H_0.

24.22 (a) $D = 0.36$; accept H_0 (b) $\chi^2 = 16.51$; reject H_0.

24.23 $H = 11.32$; reject H_0.

24.24 $H = 27.49$; reject H_0.

ANSWER TO APPENDIX A

A.7 19.5, 232.9, 33.0, 6.1, 3.1, 4.8, 6.0, 8.6, 11.5, 51.11

A.8 Males: 13.20, 18.90, 9.28, 14.16, 20.28, 32.96, 38.25
Females: 16.09, 20.95, 11.12, 16.86, 23.20, 38.38, 45.46.

A.9 (b) 11, 34, 108, 288, 582, 1184.

A.11 (b) (i) 18.85, 16.77 (ii) 16.78

A.12 C.D.R. = 13.4, S.D.R. = 14.0.

A.13 (b) C.D.R. = 13.32, S.D.R. = 9.02

A.14 (i) 7.42, 8.87 (ii) 9.26

A.17 A.S.F.R. = 19.41, 148.15, 169.70, 117.33, 65.46, 19.39, [illegible].42.
T.F.R = 2704 per 1,000 women G.R.R. = 1.31, N.R.R. = 1.25

A.18 (b) 2.385

A.19 (b) G.R.R. = 1.071, N.R.R. = 1.030

A.20 (b) G.R.R. = 4.512, N.R.R. 2.567.

INDEX

A

α, 126
Acceptance Region, 125
Allocation of sample sizes, 17
  equal, 17
  Neyman, 18
  proportional, 18
  optimum, 18
Alternative Hypothesis, 123
Analysis of Covariance, 397
  assumptions in, 413
  introduction, 397
  for estimation of missing observation, 438
  LSD-test, 401, 410
  models of, 411
  one-way analaysis and partitioning the sum of products, 398
  two-way analysis, 405
  uses of, 414
Analysis of Variance, 295
  assumptions in, 310
  for linear regression, 380
  for multiple regression, 384
  introduction, 295
  LSD-test, 324
  LS estimate of effects, 332, 334
  models, 331
  partitioning SS, 298
  partitioning *df*, 300
  one-way analysis, 296, 305
  two-way ANOVA, 311, 312
  two-way ANOVA with interaction, 318
  several observations per cell, 318
  table of, 301, 315
Area sampling, 22
Association, of attributes, 205
  co-efficient of, 205
  dissociation, 205
  measure of, 205
Assumptions
  in analysis of covariance, 413
  in analysis of variance, 310
  in *F*-distribution, 228
  in *t*-distribution, 245
Attributes, 200
  association of, 205
  consistence, 203
  independence, 204
  ultimate class-frequencies, 201

B

β, 126
Bartlett's test, 185
Basic Experimental Designs (See experimental designs)
Behrens-Fisher test, 258
Bias, 16
Binonial Distribution, 50, 159
BLUE, 78
Brandt-Snedecor formula, 220

C

Central Limit Theorem, 30, 41
Chi-Square Distribution, 169
  derivation of, 169
  properties of, 171
  table of, 175
  tests based on, 181
Chi-square tests in contingency tables, 207
  of equality of variances, 184
  of goodness of fit, 194, 195

of homogeneity, 219
of equality of several correlation co-efficients, 376
of equality of several proportions, 216
of independence, 200
in two-by-two tables, 211, 215
about variance, 181
of p's in multinomial distribution, 190
Cluster, 22
Cluster Sampling, 22
Co-efficient, of association, 205
of colligation, 206
of confidence, 94
of contingency, 213
Completely Randomised Designs, 421
advantages and disadvantages, 424
experimental layout, 421
statistical model and analysis, 422
Composite Hypothesis, 124
Concomitant variable, 397
Confidence Intervals (limits)
definition of, 94
for α, 352
for correlation coefficients, 368
for difference of means, 103, 247
for mean, 95, 246
for mean value, 352
for poportion, 107
for difference in proportions 108
for regression co-efficient, 350
for small samples, 246
for variance of a normal population, 178
one sided, 110, 247
for several sample variances, 180
for variance ratio, 282
interpretation of, 101
Consistence, 203
Consistent estimator, 76
Contingency, tables, 207
co-efficinet of, 213
exact test for, 215
Continuity, correction for, Yates' correction, 214
Contrasts, 330
single degree of freedom, 453
Correlation, significance of, 370
Covariance analysis (see anlaysis of covariance)
Cramer-Rao inquality, 78
Criteria for Estimators, 69
Critical region, 125

D

Degrees of Freedom, 169, 208
Designs, Experimental (see experimental design)
Dichotomy, 201
Difference between means
confidence interval for, 103, 247
sampling distribution of, 44
testing hypothesis, 143
Difference between Proportions
confidence interval for, 108
sampling distribution, 54
testing hypothesis, 152
Dissociation (see association)
Distribution, binomial, 50, 159
chi-square, 169
F-, 273
Fisher's z-, 273
normal, tests based on, 137
t-, 239
Double Sampling, 23

Duncan's Multiple Range Test, 328

E

Efficiency of
latin square design, 451
randomized block design, 440
Efficient Estimator, 77
Errors of
first kind, 125
mean square, 78
sampling and non-sampling, 5
second kind, 125
Error Sum of Squares, 299
Estimate, 67
Estimation, 67
interval, 68
linear, 68
method of, 85
point, 68
Estimation of Missing
observations by covariance, 438
in complete block experiment, 434
in latin square experiment, 449
Estimator, 20, 67
criteria for, 69
Exact Test for Independence, 215
Expected Mean Square, 334
Experimental Designs, 419
basic principles, 420
completely randomized design, 421
factorial design, 454
Graeco-latin Square, 451
introduction, 419
latin square design, 441
local control, 421
randomization, 420
randomized complete block design, 426
replication, 420

F

F-distribution, 273
assumptions in, 278
derivation of, 274
introduction, 273
properties of, 276
tables of, 278, 279, 280, 281
test based on, 284
Factorial Experiments, 454
advantages and disadvantages, 463
design and analysis, 458
main effects and interaction, 455
2^2–factorial experiment, 455
2^3–factorial experiment, 457
Yates' technique. 462
Finite correction factor, 28
Finite population, 1
correction for, 28
Fisher's exact test, 215
Fisher-Neyman Factorization critierion, 80
Fisher's z-distribution, 273
Fisher's z-transformation, 368
Fixed-effects model, 331, 411

G

Goodness of Fit Test, 194, 195
Graeco-Latin Squares, 451

H

Homogeneity, Test for, 219
of k-variances, 185
of several estimated correlation coefficients, 376

Homoscedasticity, 311
Hypothesis, alternative, 123
composite, 124
exact, 124
formulation of, 135
general procedure for testing, 137
inexact, 124
null, 123
simple, 124

I

Independene, of attribute, 204
Interaction, 311

J

Jacobian, 241, 275
Judgement Sample, 23

K

Karl Pearson's approximation, 187

L

Lagrange's multiplier, 333
Large samples, 29
Latin Square Designs, 441
advantages and disadvantages, 446
construction and layout, 442
efficiency in, 451
estimation of missing data in, 449
graeco-latin squares, 451
LSD, 446
orthogonal latin squares, 451
standard squares, 442
statistical model and analysis 443
Least Significance Difference test, 324
Least Squares Estimates of effects 332
Level, of confidence, 94
of significance, 129
Likelihood function, 87
Linear additive model, 331
Linear Regression, test for, 357, 365
on one variable, 380
on several variables, 384
Local Control, 421

M

Method of
maximum likelihood, 85
moment, 92
least squares, 93
Mean Square Error, 78
Minimum Variance Estimator, 78
Missing data, 434
Models, linear additive, 331
analysis of covariance, 411
analysis of variance, 332
Moment Generating function, 30
Multiple Comparisons Tests, 324
Multiple correlation, significance test of, 379
Multiphase Sample, 23
Multistage Sample, 22

N

Neyman Allocation, 18
Non-parametric tests, 477
Kolmogorov-Smirnov tests, 502
Krushal -- Wallis H test, 506
Mann-Whitney U test, 491
median test, 497
runs test for randomness, 498

sign test, 478
Wilcoxon rank-sum test, 487
Wilcoxon signed-rank test, 483
Non-probability Sampling, 5, 23
Non-response, 4
Non-sampling error, 5
Normal Distribution,
area under, 545
table of, 544
tests based on, 137
Null Hypothesis, 123

O

One-tailed Test, 130
One-way Classification, 296
Operting Characteristic Curve, 129
Optimum Allocation, 18
Orthogonal Latin Squares, 451

P

Paired observations, 259
Parameter, 1
Partial correlation, significance test of, 377
Partitioning of Sum of Squares and D.F.
in analysis of covariance, 398
in analysis of variance, 298
Point Estimation, 68
criteria for, 69
methods of, 85
Pooled Estimate,
from two or more samples, 82
of variance, 225
Population, finite or infinite, 1
size of , 1
hypothetical or existence, 2
discrete, 2
Sampled or target, 2
Prediction Interval, 355
Probability Sampling, 5, 9
Proportional Allocation, 18
Power, curve, 129
of a test, 128
Purposive Sampling, 5, 23

Q

Quota Sampling, 5, 24

R

Randomized Complete Blocks Design, 426
advantages and disadvantages, 429
efficiency in, 440
estimatioin of missing data in, 434
experimental layout, 426
statistical model and analysis, 427
with replications within blocks, 433
Random Effects Model, 332
Random Number Table, 7, 8
Random Sampling, simple, 5, 9
Region of Rejection (see critical region)
Regression, linear, one variable, 380
several variables, 384
Relative Efficiency, 440, 451
Replication, 420
Residual, 381

S

Sample, 1
Sample Design, 4
area, 22

cluster, 22
multiphase, 23
multi-stage, 22
non-probability, 5, 23
probability, 5, 9
purposive, 5, 23
quota, 5, 24
sequential, 23
simple random, 5, 9
stratified random, 5, 17
systematic random, 5, 21
Sample size, 111, 113, 133
Sample Survey, 4
Sampling, advantages of, 3
bias, 6
error, 5
introduction, 1
units, 1
with and without replacement, 5
Sampling Distribution, 24
of a, 349
of b, 348
of the means, 25
of differences between means, 44
of a proportion, 49
of differences between proportions, 54
of variance, 55
of $\hat{Y}$, 349
Sampling Frame, 4
Scheffe's method, 330
Several Sample Variances, 180
lignficance, level, 129
tests of, 129
imple Hypothesis, 124
ingle Degree of Freedom Contrasts, 453
xory's, method, 211
Snedecor's F-distribution (see F-distribution)
Standard Error, 25
Statistic, 1
Statistical Models (see models)
Stratified Random Sampling, 5, 17
Student-Newman-Keuls Multiple Range Tst, 327
Student's t-distribution (see t-distribution)
Sub-sampling, 22
Sufficient Estimator, 80
Systematic Sampling, 5, 21

T

t-distribution, 239
assumptions in, 245
definition of, 239
derivation of, 240
introduction, 239
for difference between two means, 245
paired test, 259
properties of, 242
tables of, 243, 244
tests based on, 249
Target Population, 1
Tests of signficance, 129
Testing of Hypothesis, 67, 123
introduction, 123
about α, 361
about β, 357
general procedure of, 137
about correlation co-efficients, 370, 371, 374
about equality of regression efficients, 362, 372

about equality of several correlations, 376
about differences between means, 143, 249, 257, 296
about proportion, 149
about differneces between two proportions, 152
about equality of two standard deviations, 156
about equality of two variances, 284
about equality of k variances, 185
about homogeneity, 219
about independence, 207
about linear regression, 357, 365
about means, 139, 141, 142, 249
about mean value $\mu_{Y.X}$, 362
about multiple correlation, 379
about partial correlations, 377
about regression co-efficient, 357
about variance, 181, 362
of standard deviation, 156
based on binomial dist., 159
based on normal distribution, 137
based on small samples, 249
Test-statistic, 124
Two-phase sampling, 23
Two-tailed test, 130
Two-way Classification, 296
Type I and Type II Errors, 125

U

Unbiased Estimator,
definition, 70
of population variance, 71

V

Variance, Analysis of (see analysis of variance)
Variance Ratio, 273

Y

Yule's co-efficient of association, 205
Yates' correction for continuity, 214
Yates' technique for contrasts, 462

Z

z-distribution, 273
Z-tests, 137–158
z-transformation of r, 368

OTHER BOOKS OF INTEREST

For B.A. & B.Sc. Students

1. **Mathematical Methods - Text Book**
   Dr. S. M. Yusuf, Abdul Majid, Prof. Ch. Muhammad Amin
2. **Key to above**
3. **Calculus with Analytic Geometry - Text Book**
   Dr. S. M. Yusuf, Prof. Ch. Muhammad Amin
4. **Key to above**
5. **Introduction to Mechanics - Text Book**
   Dr. S. M. Yusuf
6. **Key to above**
7. **Vector Analysis - Text Book**
   Dr. S. M. Yusuf
8. **Key to above**
9. **Vector Analysis and Mechanics - Text Book**
   Muhammad Ashraf
10. **Key to above**
11. **An Introduction to Numerical Analysis**
    Dr. Muhammad Iqbal
12. **Key to above**
13. **Numerical Analysis**
    Dr. S. M. Yusuf
14. **Key to above**
15. **Metric Spaces - Text Book**
    Prof. Ch. Muhammad Amin
16. **Refresher Course in Modern Physics, Electricity and Megnetism, Waves & Oscillation, Thermodynamics and Mechanics**

ILMI KITAB KHANA
www.ilmikitabkhana.com

Find us on: facebook