PAGES MISSING WITHIN THE BOOK ONLY

TIGHT BINDING BOOK

UNIVERSAL
LIBRARY
OU_160152
UNIVERSAL
LIBRARY

Contents

CHAPTER 1

INTRODUCTION

			Page
1·1.	Pure and Applied Statistics	...	1
1.2.	Characteristic features of the Science of Statistics	...	1
1·3.	Main divisions of statistical theory	...	3
1·3·1.	Design of experiments	...	3
1·3·2.	Statistical inference	...	3
1·4.	Reference system	...	4
	Exercises 1	...	4

CHAPTER 2

FREQUENCY DISTRIBUTION AND MEASURES OF LOCATION

2·1.	Frequency distributions	...	6
2·2.	Graphical representation of a frequency distribution	...	8
2·2·1.	Histogram	...	9
2·2·2.	Frequency polygon and frequency curve	...	10
2·2·3.	Cumulative frequency curve or the Ogive	...	10
2·3.	Forms of frequency curves	...	11
2·4.	Comparison of frequency distributions	...	12
2·5.	Measures of location	...	13
2·5·1.	Arithmetic mean	...	13
	Change of origin and scale, Properties of mean	...	15
2·5·2.	Geometric mean and Harmonic mean	...	16
2·5·3.	Median, Quartiles	...	17
2·5·4.	Graphical location of partition values	...	19
2·5·5.	Mode	...	21
2·6.	Requisites for an ideal measure of central tendency	...	21
	Exercises 2	...	21

CHAPTER 3

MEASURES OF DISPERSION, SKEWNESS & KURTOSIS, MOMENTS OF FREQUENCY DISTRIBUTIONS

3·1.	Dispersion	...	31
3·2.	Measures of dispersion	...	31
3·2·1.	Range, Interquartile differences	...	32
3·2·2.	Average deviation or mean deviation	...	32
3·2·3.	Minimal property of average deviation	...	33
3·2·4.	Variance, Standard deviation	...	35
3·2·5.	Relation between standard and root-mean square deviation	...	36
3·2·6.	Mean difference	...	39
3·3.	Moments	...	40

		Page
3·3·1.	Moments about the mean in terms of moments about any point and conversely ...	40
3·3·2.	Effect of change of origin and scale on moments ...	41
3·3·3.	Sheppard's corrections to moments of grouped frequency distributions ...	41
3·4.	Symmetrical distributions. Skewness ...	42
3·5.	Pearson's β- and γ- coefficients ...	44
3·6.	Factorial moments. Absolute moments ...	45
3·7.	Moments for a bivariate distribution ...	45
	Exercises 3 ...	46

CHAPTER 4

THEORY OF PROBABILITY

4·1.	Introduction ...	52
4·2.	Definition of probability ...	53
4·2·1.	Explanation of terms ...	53
4·2·2.	Mathematical definition of probability ...	54
4·2·3.	Statistical definition of probability ...	56
4·3.	Notations ...	57
4·4.	Theorem of total probability ...	58
4·4 1.	Generalization of theorem of total probability ...	59
4·4·2.	Geometrical interpretation of the above theorem ...	61
4·4·3.	Bool's Inequality ...	62
4·5.	Theorem of compound probability ...	67
4·6.	Bayes Theorem ...	70
4·7.	Set theoretic formulation of probability theory ...	71
4·7·2.	Pairwise independent events ...	76
4·8.	Axiomatic development of probability theory ...	79
	Exercises 4 ...	81

CHAPTER 5

RANDOM VARIABLES, MATHEMATICAL EXPECTATION, MOMENT GENERATING AND CUMULATIVE FUNCTIONS

5·1.	Random Variable ...	92
5·1·1.	Random Variance ...	92
5·1·2.	Distribution Function ...	93
5·1·3.	Discrete Random Variable ...	94
5·2·1.	Mathematical Expectation ...	97
5·2·2.	Expectation of a sum of Random Variables ...	98
5·2·3.	Expectation of a Product of Independent Random Variables ...	100
5·3.	Covariance ...	101
5·3·1.	Correlation Coefficient ...	101
5·3·2.	Expectations and Variance of a Linear Combination of Random Variables ...	102
5·4.	Conditional Expectation—Conditional Variance ...	106

		Page
5·5.	Moment Generating Function ...	107
5·5·1.	Effect of change and scale on M.G.F. ...	108
5·5·2.	A Property of Moment Generating Function ...	108
5·5·3.	Factorial Moment Generating Function ...	108
5·6.	Cumulants ...	109
5·6·1.	Effect of change of Origin and Scale on Cumulants ...	109
5·6·2.	Additive Property of Cumulants ...	109
5·7.	Relations between Moments and Cumulants ...	110
5·7·1.	Cumulants in terms of Moments ...	110
5·7·2.	Moments in terms of Cumulants ...	110
5·8.	Probability Generating Function ...	112
5·9.	Moments of Bivariate Probability distribution ...	114
	Exercises 5 ...	116

CHAPTER 6

SPECIAL DISCRETE PROBABILITY DISTRIBUTIONS

6·1.	Introduction ...	127
6·2.	Binomial distribution ...	127
6·2·1.	Binomial frequency distribution ...	129
6·2·2.	First four moments of the binomial distribution ...	130
6·2·3.	Recurrence relation ...	131
6·2·4.	Moment generating and cumulative functions of the binomial distribution ...	133
6·3.	Poisson distribution – Its derivation from binomial distribution ...	134
6·3·1.	Another approach to Poisson distribution ...	136
6·3·2.	Mean and variance of the Poisson distribution ...	138
6·3·3.	Moment generating and cumulative functions of Poisson distribution ...	138
6·3·4.	Reproductive property of Poisson distribution ...	139
6·3·5.	Mode of the Poisson distribution ...	139
6·4.	Other discrete distributions ...	140
6·4·1.	Causal distribution ...	140
6·4·2.	Uniform or Rectangular distribution ...	141
6·4·3.	Geometric distribution ...	141
6·4·4.	Negative binomial distribution ...	141
6·4·5.	Hypergeometric distribution ...	144
6·4·6.	Multinomial distribution ...	146
	Exercises 6 ...	148

CHAPTER 7

CONTINUOUS R. V. ARD PROBABILITY DISTRIBUTIONS

7·1.	Univariate distribution ...	159
7·2.	Definitions of moments, etc. for univariate continuous probability distributions ...	166
7·3.	Characteristic function ...	175

		Page
7·3·1.	Fourier's Inversion Theorem ...	176
7·4.	Chebychevs Lemma ...	184
7·4·1.	Bernoulli's theorem ...	190
7·5.	Continuous Bivariate probability distributions ...	192
7·6.	Bivariate transformations ...	201
7·6·1.	Univariate distributions ...	201
7·6·2.	Bivariate transformation ...	203
	Exercises 7 ...	207

CHAPTER 8

SPECIAL CONTINUOUS PROBABILITY DISTRIBUTIONS

8·1.	Normal distribution ...	226
8·1·1.	Chief features of the normal curve ...	229
8·1·2.	Some properties of the normal distribution ...	230
8·2.	Additive property of normal variates. Theorem ...	233
8·3.	Table of ordinates ...	235
8·4.	The normal probability integral or area under the normal curve ...	237
8·5.	Alternative proof of the approach of standardised binomial probability distribution to normal distribution ...	242
8·6.	Gamma distribution ...	244
8·6·1.	Additive property of gamma variates ...	245
8·7.	Beta distribution ...	245
8·7·1.	Beta distribution of the first kind ...	245
8·7·2.	Beta distribution of the second kind ...	246
8·8.	Cauchy distribution ...	247
8·9·1.	Exponential distribution ...	249
8·9·2.	Erlang Family of distributions ...	249
8·9·3.	Weibull distribution ...	250
8·9·4.	Double exponential distribution ...	250
8·9·5.	Logarithmic normal distribution ...	250
8·10.	Pearson's distributions ...	251
8·11.	Hermite polynomials ...	255
8·12.	Orthogonal property of polynomials $H_r(x)$ ...	257
8·13.	Gram-Charlier series of type A ...	258
	Exercises 8 ...	259

CHAPTER 9

METHOD OF LEAST SQUARES AND CURVE FITTING

9·1.	The method of least squares ...	269
9·2.	Curve fitting ...	272
9·2·1.	Fitting of parabolic curves ...	272
9·3.	Fitting of the curves of type $y=ab^x$ and $y=ax^b$ ...	276
9·4.	Fitting of Logistic Curve ...	277
	Exercises 9 ...	277

Page

CHAPTER 10

CORRELATION AND REGRESSION

10·1. Correlation coefficient ... 280
10·1·1. Change of origin and scale ... 281
10·1·2. Theorem ... 281
10·1·3. Limits for the coefficient of correlation ... 282
10·2. Bivariate frequency distribution ... 283
10·3·1. Linear regression ... 285
10·3·2. Equations of the lines of regression, Regression coefficients ... 286
10·4. Standard error of estimate ... 292
10·5. Correlation of ranks ... 293
10·6. Correlation ratio ... 295
10·6·1. Theorem ... 297
10·7. Bivariate normal distribution ... 300
10·7·1. Properties of the Normal Correlation Distribution ... 302
10·8. Intra-class correlation coefficient ... 305
Exercises 10 ... 307

CHAPTER 11

MULTIPLE AND PARTIAL CORRELATION

11·1. Introduction ... 317
11·2. Distribution of three (or more) variates ... 317
11·3. Determination of regression coefficients. Equation to regression plane ... 319
11·4. Properties of residuals ... 320
11·5. Variance of a residual ... 321
11·6. Multiple correlation coefficient ... 322
11·7. Partial correlation coefficient ... 323
Exercises 11 ... 326

CHAPTER 12

THEORY OF SAMPLING

12·1. Sampling ... 331
12·2. Sampling of attributes ... 332
12·2·1. Mean and standard deviation in simple sampling of attributes ... 333
12·3. Large samples. Tests of significance for a proportion ... 334
12·4. Comparison of large samples. Difference of proportions ... 337
12·5. Sampling of values of a variable ... 340
12·5·1. Sampling distribution ... 340
12·6. Sampling distribution of the mean ... 341
12·7. Fiducial limits for unknown mean ... 343
12·8. Testing the significance of the difference between the means of two large samples ... 345
12·8·2. Testing the significance of the difference between standard deviations of two large samples ... 347

		Page
12·9·1.	Standard errors of statistics. Introduction ...	348
12·9·2.	Standard errors of sample moments about the arbitrary value ...	349
12·10.	Theorem ...	351
12·11.	Sampling from a finite population (without replacement) Exercises 12 ...	360

CHAPTER 13

EXACT SAMPLING DISTRIBUTIONS

13·1.	Introductions ...	367
13·2.	Distributions of the sum of the squares of independent standard normal variates ...	368
13·2·1.	Chief features of the χ^2 probability curve ...	370
13·2·2.	Some properties of the χ^2 distribution ...	372
13·3.	Simultaneous sampling distributions of the mean and variance in random samples drawn from a normal population ...	374
13·4.	Distribution of the sum and ratio of two independent chi-square variates ...	378
13·5.	Student's t-distribution ...	382
13·5·1.	Chief features of the t-probability curve ...	385
13·5·2.	Some properties of t-distribution ...	385
13·6.	Distribution of t for comparison of two sample means ...	387
13·7.	Snedecor's F distribution ...	388
13·7·1.	Some properties of the F distribution and F curve ...	389
13·8.	Fisher's z-distribution ...	393
13·9.	Inter-relation between t, χ^2 and F distributions ...	395
13·9·1.	Theorem ...	395
13·9·2.	Theorem ...	395
	Exercises 13 ...	397

CHAPTER 14

EXACT SAMPING DISTRIBUTIONS (*Continued*)

14·1.	Introduction ...	404
14·2.	Analytical method ...	404
14·3.	Method of characteristic functions ...	406
14·3·1.	Distribution of the mean for the Poisson distribution ...	406
14·3·2.	Distribution of the mean normal distribution ...	406
14·4.	Method of induction ...	407
14·5.	Distribution of the range ...	408
14·6.	Distribution of quantiles ..	412
14·7.	Distribution of the correlation coefficient ...	415
14·7·1.	Alternative method of derivation of the distribution of the correlation coefficient when $\rho=0$ ...	419
14·8.	Fisher's z transformation ...	420
14·8·1.	Fisher's z transformation when $\rho=0$ ...	421
14·9.	Distribution of regression coefficient in normal samples ...	421
	Exercises 14 ...	422

Page

CHAPTER 15

TESTS OF SIGNIFICANCE BASED ON t, F AND Z DISTRIBUTIONS

15·1. Introduction ... 425
15·2. Tests of significance based on t distribution ... 427
15·2·1. Testing the significance of the mean of a random sample from a normal population ... 427
15·2·2. Testing the significance of the difference between two sample means ... 428
15·2·3. Testing the significance of an observed correlation coefficient ... 433
15·2·4. Testing the significance of an observed partial correlation coefficient ... 434
15·2·5. Testing the significance of an observed regression coefficient ... 435
15·2·6. Testing the significance of a rank correlation coefficient ... 436
15·3. Tests based on the F distribution ... 436
15·3·1. Testing the significance of the ratio of two independent estimates of the population variance ... 436
15·3·2. Testing the significance of a correlation ratio ... 437
15·3·3. Testing the significance of a regression function ... 438
15·3·4. Testing for non-linearity of regression ... 438
15·3·5. Testing the significance of an observed multiple correlation ... 438
15·4. Further applications of the F distribution ... 439
15·4·1. Testing homogeneity of means. Analysis of Variance ... 439
15·4·2. Analysis of variance for two factors of classification ... 443
15·5. Tests of significance based on Fisher's z transformation ... 445
15·5·1. To test whether r differs from ρ ... 446
15·5·2. Testing the significance of difference between two independent correlation coefficients ... 446
15·5·3. Combination of estimates of a correlation coefficient ... 447
15·5·4. Tests of significance for partial correlation coefficients ... 448
Exercises 15 ... 448

CHAPTER 16

TESTS OF SIGNIFICANCE BASED ON THE CHI-SQUARE DISTRIBUTION

16·1. Introduction ... 458
16·2. χ^2 as a test of goodness of fit ... 458
16·2·1. Proof of the distribution of χ^2 ... 459
16·2·2. Remarks on the distribution of χ^2 ... 461
16·3. χ^2 test as a test of independence ... 464
16·3·1. Contingency table ... 464
16·3·2. Calculation of χ^2 to test the hypothesis of independence ... 465
16·3·3. Number of degrees of freedom associated with a $s \times t$ contingency table ... 465
16·4. Further consideration of 2×2 contingency tables ... 468
16·1. Yates correlation for continuity ... 468

		Page
16·4·2.	Exact treatment of 2×2 contingency table ...	469
16·5.	Other tests of significance based on χ^2 distribution ...	470
16·5·1.	Test for variance of a normal population ..	470
16·5·2.	Combination of probabilities from independent test of significance ...	471
16·5·3.	Testing homogeneity of several independent estimates of population variance. Bartlett Test ...	472
16·5·4.	Testing homogeneity of several independent estimates of an ordinary or partial correlation coefficient ...	472
16·6.	Association of attributes and contingency ...	472
16·6·1.	Measures of Association for contingency tables ...	474
	Exercises 16 ...	475

CHAPTER 17

STATISTICAL THEORY OF POINT ESTIMATION

17·1.	Introduction ...	482
17·2.	Fisher's criteria for a best estimator ...	482
17·2·1.	Consistent estimator ...	482
17·2·2.	Efficient estimator ...	485
17·2·3.	Efficient estimator ...	488
17·3.	Methods of maximum likelihood ...	492
17·3·1.	Independence tests in contingency tables ...	496
17·3·2.	Properties of maximum likelihood estimators ...	498
17·4.	Other methods of estimation ...	500
17·4·1.	Method of moments ...	500
17·4·2.	Method of minimum variance ...	501
17·4·3.	Method of minimum χ^2 ...	501
17·4·4.	Method of inverse probability ...	502
17·4·5.	Method of least squares ...	503
17·5.	Interval estimation ...	507
17·5·1.	95 per cent confidence for intervals for special cases ...	505
17·5·2.	Confidence intervals for large samples ...	508
	Exercises 17 ...	510

CHAPTER 18

TESTING OF HYPOTHESES, SEQUENTIAL ANALYSIS, DISTRIBUTION-FREE METHODS, DECISION THEORY

18·1.	Introduction ...	51
18·2.	Testing of hypothesis ...	51
18·2·1.	Simple and composite hypothesis ...	517
18·2·2.	Two types of errors, critical region, power function, Operating characteristic function ...	
18·2·3.	Testing a simple hypothesis H_0 against an alternative simple hypothesis H_1 ...	
18·2·4.	Testing a simple hypothesis against a composite hypothesis ...	

		Page
18·2·5.	Likelihood ratio test ...	521
18·2·6.	Bias in Statistical Tests ...	521
18·3.	Sequential analysis ...	522
18·3·1.	Sequential likelihood ratio test ...	523
18·3·2.	Power function and average sample size ...	523
18·4.	Distribution-free methods ...	524
18·4·1.	Comparison with parametric case ...	525
18·4·2.	Point estimation and confidence intervals ...	525
18·4·3.	Sign test ...	525
18·4·4.	Run test ...	526
18·4·5.	Median test ...	527
18·4·6.	Wilcoxon-Mann-Whitney test ...	531
18·4·7.	Kolmogrov-Smirnov test ...	532
18·5.	Statistical Decision Theory ...	532
18·5·1.	Basic Concepts ...	532
18·5·2.	Minimax Decision Rule ...	534
18·5·3.	Bayes Decision Rule ...	535
18·5·4.	An Example ...	536
18·5·5.	Games against Nature ...	537
18·5·6.	Admissible decision Functions and Complete clauses of decision Functions ...	537
	Exercises 18 ...	537

CHAPTER 19

ELEMENTS OF STOCHASTIC PROCESSES

19·1.	Basic concepts and illustrations ...	541
19·2.	Gambler's Ruin problems ...	542
19·2·1.	Probability of Ruin ...	542
19·2·2.	Expected duration of the game ...	543
19·2·3.	Gambler's ruin with correlation ...	544
19·3.	Random walk models ...	545
19·4.	Markov chains ...	546
19·5.	Poisson process ...	549
19·6.	The simple birth process ...	550
19·7.	The simple birth and death process ...	553
19·8.	Birth, death and immigration processes ...	557
19·9.	Generalized Birth and Death processes ...	558
19·10.	The Simplest Trunking Problem. Infinitely many channels ...	559
19·11.	Waiting lines for a finite Number of channels ...	560
19·12.	Servicing of machines with only one serviceman ...	562
19·13.	Servicing of machines : Several repairmen ...	564
19·14.	A power supply problem ...	565
19·15.	Waiting line for a single channel ...	566
19·16.	Queues with several servicing stations ...	568

		Page
19·17.	Arrival and Service Time Distributions ...	569
19·18.	Queuing processes ...	571
	Exercises 19 ...	573
	TABLES	
1.	Ordinates of the normal curve ...	236
2.	Area under the normal curve ...	238
3.	Values of Chi-square ...	374
4.	Values of mod. t ...	386
5.	Values of F, 5% and 1% point ...	394
	APPENDICES	
I.	Some Continuous Distributions ...	577
II.	Some Discrete Distributions ...	579
III.	Bibliography ...	580
	Index ...	584

1

Introduction

1·1. Pure and Applied Statistics. Like Mathematics, the Science of Statistics may be classified into :

(*i*) *Pure Statistics* or *Mathematical Statistics* or the *Theory of Statistics* ; and

(*ii*) *Applied Statistics.*

In Pure Statistics, we develop the research tools which are then applied to specific problems in different fields of Applied Statistics such as Physics and Chemistry, Anthropology and Biology, Agriculture and Animal Husbandry, Economics and Commerce, Education and Psychology, Medicine and Public Health, Industry and Insurance, etc. As in Mathematics, here also a theory of Statistics has been possible, since the same type of rules are required in all the diverse fields of applications. Again, as in Mathematics, Pure and Applied Statistics act and react and thus enrich each other. Pure Statistics continues to develop newer and stronger tools for Applied Statistics, while Applied Statistics continues to present new and challenging problems for Pure Statistics. Pure Statistics not only creates new tools but also tries to perfect existing ones by placing them on more and more rigorous foundations, and in this process it depends much on recent developments in Pure Mathematics. At the same time, Applied Statistics dealing, as it does, with the application of statistical methods to specific problems continues to find new uses for the existing tools and the new tools that are being created.

In this book, we shall be concerned mainly with the theory of Statistics though we shall very often draw upon Applied Statistics for illustrating the theory. The Theory of Statistics is based on the *Theory of Probability* though it supplements that theory by certain additional axioms. We shall accordingly also study and in fact lay a great deal of emphasis on the Theory of Probability.

1·2. Characteristic Features of the Science of Statistics. Before embarking on the study of the Science of Statistics, it is

desirable to define Statistics. In 1935, Dr. W.F. Willcox listed over a hundred definitions of Statistics and the list was even then far from being exhaustive. In fact, no definition can give a complete picture of a fast growing subject like Statistics which is bound to outgrow any definition within which it may be sought to be encompassed and as such, we shall not attempt to define Statistics. However, we give below some of the characteristic features of the Science of Statistics in order to help the reader in understanding the nature of the subject.

(*i*) *Statistics is concerned with reduction of data or with obtaining correct facts from figures.* The data which are collected or compiled in the course of an enquiry are usually very voluminous and as such it is not possible to draw any conclusions from the data in its original form. It is, therefore, desirable to express all the relevant information contained in the unweildy data by means of a comparatively few numerical values. The layman's concept of Statistics as the science of collecting and tabulating numerical data, of representing data by means of charts and diagrams, of calculating the mean, median, mode, standard deviation, etc. of the data, though fifty years out of date, represents one important aspect of the reduction of data.

(*ii*) *Statistics deals with populations or aggregate of individuals rather than with individuals.* Thus 'Milton was blind' is not a statistical statement whereas 'The average height of an Indian is five feet' is a statistical statement. It should be noted that the kinetic theory of gases is a statistical theory as its results apply to the population of gas molecules and not to any individual molecule.

(*iii*) *Statistics deals with variation,* for a completely homogeneous population will be specified in terms of one individual only.

(*iv*) *Statistics deals with only numerically specified populations.* Thus the population of words in a dictionary is not of statistical interest unless we want to know how many of the words begin with the various letters of the alphabet or with some other property of words capable of being expressed in numerical form. Qualitative characteristics like intelligence, beauty, goodness, etc. do not form part of statistical studies unless numerical equivalents are assigned to these in terms of, for example, intelligence quotients, beauty measurements, etc.

(*v*) *Statistics deals with populations which occur in nature and are subject to a large number of random causes.* Thus, tables of logarithms and of other mathematical functions, though they constitute

numerical populations, are not of statistical interest, for they are determined by exact mathematical laws ; while populations of births and deaths, of imports and exports, of yield of crops are of statistical interest as they are affected by a large number of random causes.

(*vi*) *The logic used in statistical inference is inductive.* In statistical methods, we draw inferences about the population from the information contained in a *sample* which is only a part of the population and thus we pass from the particular to the general.

(*vii*) *Statistical inferences are uncertain.* This is naturally so, since the logic used is inductive. The chief merit of statistical theory is that, under certain conditions, it enables us to get a more or less precise idea of the measure of this uncertainty in terms of probability statements. Thus though it may not be possible to say that a certain result will hold in all cases, we can with a certain measurable degree of confidence say that the result is likely to hold in a known percentage of cases with the experiment or observation is repeated a large number of times.

(*viii*) *Statistical results might lead to fallacious conclusions if they are quoted shorn of their context or manipulated.* Further, statistical conclusions should not be based on incomplete information. Thus the argument that drinking is bad for longevity since 99% of the persons who drink beer die before reaching the age of 100 years is statistically defective, since we are not told what percentage of persons who do not drink beer die before reaching that age.

1·3. Main Divisions of Statistical Theory. The statistical theory can be classified into two main divisions.

(1) Design of Experiments, and

(2) Statistical Inference.

We discuss these briefly in the following paragraphs.

1·3·1. Design of Experiments. The subject of design of experiments whether in Science, Agriculture, Industry or Sample Surveys tries to lay down the basic principles and designs for collecting data in most economic and useful forms. We shall not, however, be concerned with this aspect of the theory in this book.

1·3·2. Statistical Inference. Statistical inference is the science of drawing valid conclusions from the given data which are assumed to have been so collected as to satisfy the basic requirements of the statistical theory. The problem of statistical inference may be formulated as follows :

The given data are regarded as a sample from some population and then valid inferences about the population are drawn from the

information supplied by the sample. Thus the inference drawn from the sample is inductive in nature and has some degree of certainty. The first stage is of *specification* that is, we try to specify the population from which the sample might have been drawn. In some cases, the theory of a particular problem may help us in specifying the population but in most cases experience aided by the study of a large number of standard populations is the only guide. The specification of the mathematical form of the population will, in general, involve a number of constants called the *parameters* of the populations. These parameters are to be *estimated* from the sample. In general, it would be possible to build a number of estimates from the sample called *statistics*, for the same population parameters. The *Theory of Estimation* consists in trying to find the *best estimators* for all the parameters. The possibility of discrimination arises, as due to the inductive nature of the inference, all estimators would be subject to error and the *theory of sampling distributions* helps us in evaluating these errors. The correct specification of the population and satisfactory estimation of the parameters together with *measures of reliability* of these estimators constitute the solution of the problems of statistical inference.

1·4. Reference System. The chapters have been divided into numbered sections ; the numbering begins anew in each chapter. The chapter number is prefixed to the section number and separated from it by a dot. Thus section 4·2 refers to section 2 of chapter 4 and 4·2·1 refers to sub-section 1 of section 2 of chapter 4. The equations are numbered anew in each chapter and equation numbers are always enclosed in parentheses. Thus equation (6·5·2) refers to the second equation of fifth section of the sixth chapter. Similarly, solved examples in each chapter are numbered anew. Thus example 15·5 means the fifth solved example of chapter 15.

EXERCISE 1

1. Discuss the following :

(*a*) "The Science of Statistics is essentially a branch of Applied Mathematics and may be regarded as Mathematics applied to observational data." —Fisher.

(*b*) "Statistics in the branch of scientific method which deals with the data obtained by counting or measuring the properties of population of natural phenomena."—Kendall.

(*c*) "Statistics may be regarded as (*i*) the study of population, (*ii*) the study of variation, (*iii*) the study of methods of reduction of data."—Fisher.

2. Criticise the following arguments :

(*a*) The average depth of a river is 4 feet and the average height of the members of a family is $4\frac{1}{2}$ feet, therefore the family can safely cross the river.

(*b*) 10% of the cases treated by a surgeon proved fatal. Out of the first ninety cases this year only five died, therefore out of the next ten cases, five must die.

(*c*) The average income of a king and his household servants is Rs. 2,000 p.m. Therefore all the servants must be fabulously paid.

(*d*) All the persons who have written to the firm about the medicine have praised it. Therefore the medicine must be effective.

(*e*) On measuring ten Englishmen and ten Indians it was observed that the average height of the Englishmen exceeds that of the Indians. Therefore the Englishmen are taller than the Indians.

(*f*) 99% of the numbers who voted for the Army Estimates were military officers. Therefore it was unfair to suppose that the voting was unbiased.

(*g*) The death rate in the American Navy during the Spanish American War was nine per thousand while in the city of New York it was sixteen per thousand. It was safer to be a sailor in American Navy than to live in New York city.

(*h*) From personal discussion with my friends, I can definitely say that most of the people in our country subscribe to the point of view which I hold.

(*i*) People's capitalism in America is a statistical myth in spite of most Americans holding shares in industry.

(*j*) In a country 20,000 vaccinated persons died of small-pox. Therefore vaccination is useless.

(*k*) The examination result of school *A* was 75% in a particular year. In the same year and at the same examination only 400 out of a total of 600 students were successful in school *B*. Therefore the teaching standard of *A* was decidedly better.

3. Comment on the following :

(*i*) A : According to this chart of standard heights and weights, you weigh too much.

B : No. According to this chart, I am not sufficiently tall.

(*ii*) A person asks the same set of 25 questions from all the members of a large joint family and also measures the size of their shoes. From this he deduces that there is a high degree of relationship between knowledge and size of the shoes.

(*iii*) In every country where the sale of contraceptives is allowed by the government, the birth-rate is declining. Therefore contraceptives must exert an influence on the birth-rate.

[Two unrelated factors may appear to be related, if they are both related to some third factor.]

4. (*a*) Are the following statistical statements or not ? Give reasons.

(*i*) Indians are hard working.

(*ii*) The attraction between two particles varies inversely as the square of the distance between them.

(*iii*) The production of wheat per acre in 1955 in a certain district was 78 maunds.

(*b*) Give some examples of statistical theories from Economics, Genetics and Physics.

(*c*) Out of 100 packets in a box one is drawn at random and its weight is found to be 10 lbs. What inference, if any, can you draw about the populations of packets ?

5. Write short notes on :

(*a*) Relation between Pure and Applied Statistics.

(*b*) Statistical inference.

2

Frequency Distributions and Measures of Location

2·1. Frequency Distributions. A frequency distribution is defined whenever (*i*) the values which the variate takes and (*ii*) the frequencies or the number of times each variate value is taken are specified.

As an example, consider the marks obtained by a class of 60 students in a Statistics paper, the maximum marks allotted to the paper being 50 :

22, 47, 9, 42, 31, 17, 13, 15, 18, 13, 2, 21, 27, 38, 15, 0, 33, 10, 34, 29, 26, 16, 25, 33, 36, 10, 24, 22, 26, 19, 14, 36, 18, 25, 21, 33, 35, 25, 18, 28, 25, 17, 38, 10, 5, 31, 24, 3, 12, 16, 33, 18, 26, 29, 27, 29, 28, 35, 26, 27.

Here the variable is the 'number of marks'. The data in the original form as above are called *ungrouped data*. When arranged in ascending or descending order of magnitude, the data are said to be arranged in an *array*. Counting the number of times each variate value occurs, we get the following **Frequency Table**.

TABLE I

Marks in Statistics of 60 Students

Marks in Statistics	*Number of Students*	*Cumulative Frequency*	*Marks in Statistics*	*Number of Students*	*Cumulative Frequency*
0	1	1	24	2	29
2	1	2	25	4	33
3	2	4	26	4	37
9	1	5	27	3	40
10	3	8	28	2	42
12	1	9	29	3	45
13	2	11	31	2	47
14	1	12	33	4	51
15	2	14	34	1	52
16	2	16	35	2	54
17	2	18	36	2	56
18	4	22	38	2	58
19	1	23	42	1	59
21	2	25	47	1	60
22	2	27		60	

The frequency table given above is an improvement over the arrangement of figures in an array, as it presents a clearer idea of the data, but to get a still better idea of the structure of the frequency distribution, we re-classify the data into groups as follows :

TABLE II

Frequency distribution of marks in Statistics of 60 students

Marks in Statistics (Class)	*Number of Students (Frequency)*	*Cumulative Frequency*
0— 5	4	4
6—10	1	5
11—15	7	12
16—20	11	23
21—25	6	29
26—30	16	45
31—35	7	52
36—40	6	58
41—45	1	59
46—50	1	60

This type of representation of frequencies is called a *grouped frequency distribution* or simply a grouped distribution. In the above example, 60 students have been divided into 10 *groups viz.*, 0—5, 6—10,....., 46—50. The groups are also known as *classes* and the boundary figures under the column heading 'marks' are called *class limits*. The figures on the left side of the classes are called *lower limits* while figures on the right are known as *upper limits*. Thus, for the class 21—25, 21 is the lower limit and 25 is the upper limit. The difference between the upper and lower limits of a class is called the *width* of the class or simply the *class interval*. The interval of the class 21—25 is 25—21, *i.e.*, 4. The variate value which lies mid-way between the lower and upper limits of a class is called its *mid-value* or *central value*, *e.g.*, the central value of the class 21—25 is 23. The number of individuals or observations in each class is called the *frequency* of that class or simply *class frequency*. The frequency of the class 21—25 is 6. The *cumulative frequency* corresponding to any variate value is the number of observations less than or equal to that value. The cumulative frequency corresponding to a class is the total number of observations less than or equal to the upper limit of the class, *i.e.*, it is the total of all the frequencies up to and including that class. The frequency corresponding to the upper limit of the class itself may, however, have to be left out of the total if, the class $a-b$ is defined as 'a and less than b'. In both the above tables, the cumulative frequencies are shown in the last column.

The following points should be borne in mind while tabulating or classifying an observed frequency distribution :

(*i*) For the sake of ease in computation and comparison, the class-intervals should, wherever possible, be of equal width. However, in some cases, *e.g.*, in the frequency distribution of the number of persons paying income-tax in different ranges of income or that of the number of children dying in different age-groups, it may be desirable to use the unequal class-intervals due to greater concentration of observations in certain class-intervals.

(*ii*) The number of class-intervals should not ordinarily exceed 20 and should not, in general, be less than 10. With more than 20 classes, the computations become necessarily tedious whereas with less than 10 classes, a great deal of accuracy may be lost.

(*iii*) The class-intervals should be defined with precision *i.e.*, it should be made clear whether the class $a-b$ means all values of the variate x for which $a \leqslant x < b$ or $a < x \leqslant b$. For continuous variates, it is desirable to take into account the accuracy of measurement. Thus if statures are measured correct to the nearest centimetre it is better to make class-intervals as 145·5—146·5 rather than 145—46.

(*iv*) The class interval should be an integer as far as possible.

Note. While dealing with a continuous variable, it is not possible to classify the data into groups of the type 0—5, 6—10, etc. Consider, for an example, the distribution of ages in years of students in a college. Clearly we cannot arrange the data in the form *e.g.*, 16—20, 21—25, etc. since there can be students having their ages between 20 and 21. Thus in case of a continuous variable, the data should be classified as 16—20, 20—24, 24—28, etc., it being understood that the students with ages less than 20 are included in the class 16—20. Thus a *continuous frequency distribution* has the form :

Class (Age in years)	*Frequency*
16—20	f_1
20—24	f_2
24—28	f_3
...	...

2·2. Graphical Representation of a Frequency Distribution. The graphical representation makes the unwieldy data readily intelligible and brings to light the salient features of the data at a glance. It makes visual comparison of data easier. It is also useful in locating certain *Partition Values*. (*See* 2.5·4).

The following types of graphs are used in representing frequency distributions :

1. Histogram. 2. Frequency Polygon and Frequency Curve. 3. Cumulative Frequency Curve or the '*Ogive*'.

2·2·1. Histogram. In drawing the histogram of a given grouped frequency distribution, we first mark off along the axis of x all the class-intervals on a suitable scale. With the class-intervals as bases, we draw rectangles with *areas* proportional to the frequencies of the class-intervals. For equal class-intervals, the heights of the rectangles will be proportional to the frequencies, while for unequal class-intervals, the heights will be proportional to the ratios of the frequencies to the widths of classes. Fig. 1 is the histogram of the following frequency distribution showing the number of boys on the registers of primary schools in a certain state :

Age (*in years*)	*No. of boys* (*in thousands*)	*Age* (*in years*)	*No. of boys* (*in thousands*)
2—5	150	13—14	496
5—11	3066	14—15	143
11—12	497	15—17	162
12—13	477		

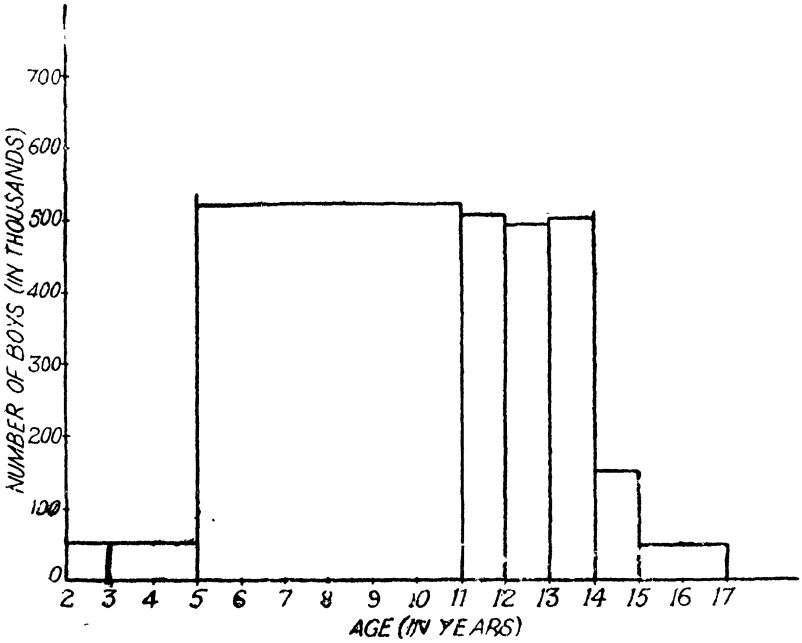


Fig. 1. Histogram of frequency distribution given above.

Here we first construct a rectangle with the class 2—5 as base and height equal to $\frac{150}{3}$ *i.e.* 50. Adjacent to this reactangle, we

construct another rectangle with the class 5—11 as base and $\frac{3066}{6}$ *i.e.*, 511 as its height. Proceeding in this way, we get a set of rectangles constituting the histogram.

Notes : 1. To draw the histogram for an ungrouped frequency distribution of a variate, we shall have to assume that the frequency corresponding to the variation value x is spread over the interval $x-\frac{1}{2}h$ to $x+\frac{1}{2}h$, where h is the jump from one value to the next.

2. The histogram makes it appear as if the frequency in each class is uniformly distributed over that class which, in general, need not be true. The histogram is, however, particularly appropriate to cases in which the frequency changes rapidly or the class-intervals are not of the same width.

2·2·2. Frequency Polygon and Frequency Curve. For an ungrouped frequency distribution, the frequency polygon is obtained by plotting points with abscissae as the variate values and the ordinates as the corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the abscissae of the points are the mid-values of the class-intervals.

For equal class-intervals, the frequency polygon can be obtained by joining the middle points of the upper sides of the adjacent rectangles of the histogram by means of straight lines.

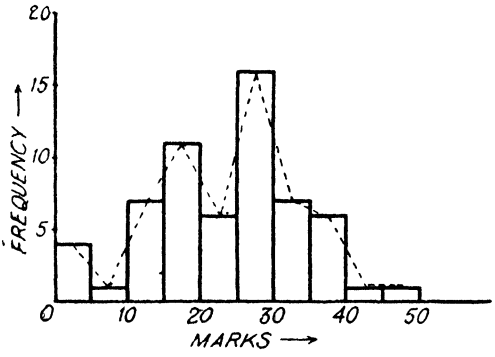


Fig. 2. Histogram and Frequency Polygon for the distribution given on page 7.

If the class-intervals are of small widths, the frequency polygon can be approximated by a smooth *frequency curve*. The frequency curve can be obtained by drawing a smooth free-hand curve through the vertices of the frequency polygon. In Fig. 2, the frequency polygon for the continuous frequency distribution given on page 7 has been shown in dotted lines.

2·2·3. Cumulative Frequency Curve or the Ogive. We first form the cumulative frequency table and then plot the points with the upper limits of the classes as abscissae and the corresponding cumulative frequency as ordinates. The points are joined by a free-hand smooth curve to give *the cumulative frequency curve or the*

Ogive. In Fig. 3, the ogive of the frequency distribution given on page 7 has been shown.

If instead of the ordinary cumulative frequency we take the ordinate as the percentage cumulative frequency *i.e.*, as the ordinary cumulative frequency expressed as a percentage of the total frequency, we get the *percentage cumulative frequency curve* which is useful for comparing different frequency distributions as the frequency distributions are thus adjusted to a uniform standard.

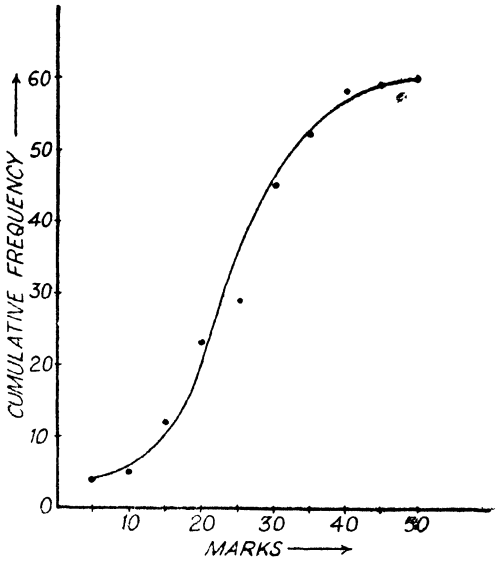


Fig. 3. Ogive of the distribution given on page 7.

2·3. Forms of Frequency Curves. The following are types of curves which we generally get in the graphical representation of frequency distributions.

(*a*) *Symmetrical Curve.* A frequency curve is said to be symmetrical if it can be folded along a vertical line (ordinate) so that the two halves of the figure coincide. An important type of a symmetrical curve is a bell-shaped curve which has got a single smooth hump in the middle and 'tails off' gradually at either end.

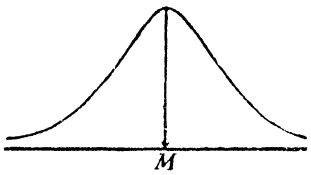


Fig. 4. Symmetrical curve.

(*b*) *Moderately Asymmetrical or Skew Curve.* A curve is said to be skew if it lacks in symmetry. In a skew curve, observations tend to pile up at one or the other end of the curve. Thus a skew curve may



Fig. 5 (*a*). Negative skew curve.

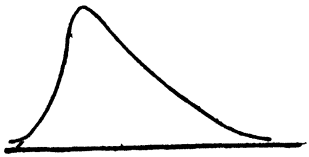


Fig. 5. (*b*). Positive skew curve.

have a long tail to the positive (right) side in which case it is said to be positive skew curve or to the negative (left) side in which case it is known as negative skew curve.

(*c*) *Extremely Asymmetrical (or J-shaped) and U-shaped Curves,* Such curves are so called as they have the shape of a '*J*' or a '*U*'.

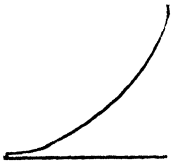


Fig. 6(*a*). Negative *J*-shaped curve.



Fig. 6(*b*). Positive *J*-shaped curve.

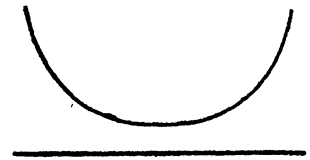


Fig. 7. *U*-shaped curve.

U-shaped curve exhibits a maximum frequency at the ends of the range and a minimum towards the centre as in Fig. 7.

Ex. The number of persons travelling by public buses at different times in the day in a capital city is given below :

Time	7 A.M.	8 A.M.	9 A.M.	10 A.M.	11 A.M.	12 NOON
No. in thousand	1	3	5	9	3	1
Time	1 P.M.	2 P.M.	3 P.M.	4 P.M.	5 P.M.	6 P.M.
No. in thousand	2	3	5	9	5	1
Time	7 P.M.	8 P.M.	9 P.M.			
No. in thousand	1	5	2			

What is the form of the frequency curve between

(*i*) 7 A.M. and 10 A.M. (*ii*) 7 A.M. and 1 P.M. (*iii*) 10 A.M. and 5 P.M. (*iv*) 5 P.M. and 6 P.M. (*v*) 1 P.M. and 7 P.M. (*vi*) 7 A.M. and 9 P.M. ?

2·4. Comparison of Frequency Distributions. In statistical work, it is frequently necessary to compare two frequency distributions. If they are of different types a precise comparison is difficult and is usually not required. If, however, they are of the same type, a comparison can be made in terms of the values of the following four types of measure :

(*a*) *Measures of location or central tendency* given the whereabouts of some central value, *e.g.*, the maximum and thus, to say,

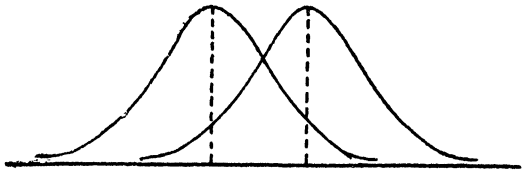


Fig. 8(*a*). Difference in central value.

locating the distribution.

(*b*) ***Measures of dispersion*** giving the degee of scatter about the central value, and thus giving the measures of variability or lack

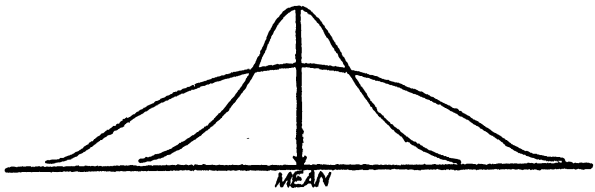


Fig. 8 (*b*). Difference in dispersion.

of homogeneity of the data.

(*c*) ***Measures of skewness*** measuring the degree of departure from symmetry.

(*d*) ***Measure of kurtosis***, measuring the degree of 'flatness' of the 'top' as compared with the 'normal' curve (defined in chap. 8).



Fig. 8(*c*). Difference in symmetry.

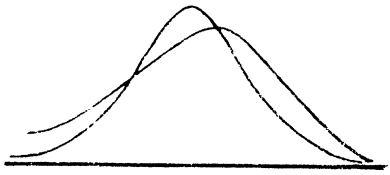


Fig. 8(*d*). Difference in kurtosis

The measures of dispersion, skewness and kurtosis will be discussed in the next chapter.

2·5. Measures of Location or Measures of Central Tendency. The various types of measures of location, in common use, are :

(1) Arithmetic Mean or simply the Mean.
(2) Geometric Mean.
(3) Harmonic Mean.
(4) Median and Quartiles.
(5) Mode.

2·5·1. Arithmetic Mean. Let the variate x take the values $x_1, x_2, \ldots\ldots, x_n$, then arithmetic mean, denoted by $\bar{x}$, is defined by

$$\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \ldots(2\cdot1)$$

If, the variate-values are not of equal 'importance', we may attach to them 'weights' $w_1, w_2, \ldots\ldots, w_n$ as measures of their importance, so that the weighted mean, $\bar{x}_w$, is defined by

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n}{w_1 + w_2 + \ldots\ldots + w_n} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i} \qquad \ldots(2\cdot2)$$

If the variate-value x_1 occurs f_1 times, x_2 occurs f_2 times and so on, then

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots + f_nx_n}{f_1 + f_2 + \ldots\ldots + f_n} = \frac{1}{N}\sum_{i=1}^{n} f_ix_i \qquad \ldots(2\cdot3)$$

where $N = f_1 + f_2 + \ldots\ldots + f_n = \sum_{i=1}^{n} f_i$ = total frequency.

This formula expresses that the arithmetic mean $\bar{x}$ is the weighted mean of the values $x_1, x_2, \ldots\ldots, x_n$ whose weights are the respective frequencies $f_1, f_2, \ldots\ldots, f_n$.

Notes : 1. The arithmetic mean may be regarded as the x-co-ordinate of the centre of gravity of a plane of the same shape as the histogram of the distribution.

2. The arithmetic mean of a grouped frequency distribution is computed by the above formula after the classes have been replaced by their mid-values.

Example 2·1. *Given the following frequency distribution, calculate the mean.*

Monthly Wages (in Rs.)	*No. of Workers*	*Monthly Wages (in Rs.)*	*No. of Workers*
12·5—17·5	2	37·5—42·5	4
17·5—22·5	22	42·5—47·5	6
22·5—27·5	19	47·5—52·5	1
27·5—32·5	14	52·5—57·5	1
32·5—37·5	3		

(*M.A. Punjab 1943*)

The calculation of Σfx is shown in the following table :

Mid-value x	*No. of workers* f	*Product* fx	*Mid-value* x	*No. of workers* f	*Product* fx
15	2	30	45	6	270
20	22	440	50	1	50
25	19	475	55	1	55
30	14	420			
35	3	105	Total	72	2005
40	4	160			

Hence $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_ix_i$ = Rs. $\frac{2005}{72}$ = Rs. $27\frac{41}{72}$.

Change of Origin and Scale. Shifting the origin to a point whose abscissa is a, the equation (2·3) takes the form

$$\bar{x}-a=\frac{1}{N}\sum_{i=1}^{n} f_i(x_i-a)$$

or
$$\bar{x}=a+\frac{1}{N}\sum_{i=1}^{n} f_i d_i, \text{ where } d_i=x_i-a \qquad \ldots(2\cdot3')$$

This formula gives a Short Method for the computation of the arithmetic mean. The arbitrary number 'a' is called the provisional or assumed or working arithmetic mean.

In case of frequency tables having equal class-intervals say of width h, it is convenient to use a different scale. We take a new unit $(-1/h)$ times the original unit. Thus, if u be the new variate, we have

$$u=(x-a)/h \quad \text{or} \quad x=a+hu \qquad \ldots(2\cdot3'')$$

$$\therefore \quad \Sigma fx=\Sigma af+\Sigma hfu,$$

Dividing both sides by Σf, we get

$$\bar{x}=a+h\bar{u} \qquad \ldots(2\cdot3'')$$

Example 2·2. *Solve Example 2·1 by applying the formula (2·31'').*

Let us take 30 as the provisional mean, *i.e.* $a=30$ and $h=5$, so that $u=(x-30)/5$.

x	f	$u\left(=\frac{x-a}{h}\right)$	$f\times u$
15	2	−3	− 6
20	22	−2	−44
25	19	−1	−19
30	14	0	0
35	3	1	3
40	4	2	8
45	6	3	18
50	1	4	4
55	1	5	5
	$\Sigma f=72$		$\Sigma fu=-31$

Hence $\bar{x}=a+h\bar{u}=30+\frac{5\times-31}{72}=$ Rs. $27\frac{61}{72}$.

Properties of the Arithmetic Mean.

1. *The sum of the deviations of the values x_i from their mean $\bar{x}$ is zero.*

We have $\sum_{i=1}^{n} f_i(x_i-\bar{x})=\sum_{i=1}^{n} f_i x_i-\bar{x}\sum_{i=1}^{n} f_i=N\bar{x}-N\bar{x}=0.$

2. *If* $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, *be the arithmetic means of* k *distributions with respective frequencies* $n_1, n_2, \ldots, n_k$, *then the mean* $\bar{x}$ *of the whole distribution with frequency* $N(=\sum_{i=1}^{k} n_i)$ *is given by the formula*

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{x}_i.$$

Let x_{ij} be an observation of the i-th distribution and let its frequency be f_{ij}, then

$$\bar{x}_i = \frac{\sum_j f_{ij}\, x_{ij}}{\sum_j f_{ij}} = \frac{\sum_j f_{ij} x_{ij}}{n_i}.$$

If $\bar{x}$ denotes the mean of all the observations, then

$$\bar{x} = \frac{\sum\sum_{i\ j} f_{ij} x_{ij}}{\sum\sum_{i\ j} f_{ij}} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i} = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{x}_i.$$

Weighted Mean. So far we noticed that all the observations in a distribution had equal importance. But it may not be so in actual practice. Thus if the n observations $x_1, x_2, \ldots\ldots, x_n$ have weights $w_1, w_2, \ldots\ldots w_n$, then the weighted mean, $\bar{x}_w$, is given by

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \ldots\ldots + x_n w_n}{w_1 + w_2 + \ldots\ldots + w_n} = \frac{\sum_{i=1}^{n} x_i\, w_i}{\sum_{i=1}^{n} w_i} \qquad \ldots(1)$$

Thus the formula for the weighted mean is the same as the formula for the simple arithmetic mean provided that w_i is the same as f_i, $i=1, 2, \ldots\ldots, n$. Further, if the weights assigned to each of the observations are equal, then the weighted mean is the same as the simple arithmetic mean. The formula (1) gives rise to higher value than the simple mean if smaller weights are assigned to smaller observations and larger weights to larger observations. Further, (1) leads to smaller value than the simple mean if the small weights are assigned to larger observations and large weights to smaller observations.

2·5·2. Geometric Mean and Harmonic Mean. If the n non-zero and positive variate-values $x_1, x_2, \ldots\ldots, x_n$ occur $f_1, f_2, \ldots\ldots, f_n$ times respectively, then geometric mean G, is given by

$$G = \left[x_1^{f_1} \; x_2^{f_2} \; \ldots\ldots x_n^{f_n} \right]^{\frac{1}{N}} \text{ where } N = \sum_{i=1}^{n} f_i$$

i.e. $$\log G = \frac{1}{N} \sum_{i=1}^{n} (f_i \log x_i). \qquad \ldots(2{\cdot}4)$$

Thus *the logarithm of the geometric mean is the weighted mean of the different values log x_i whose weights are the frequencies f_i.*

The geometric mean gives more weight to logarithms and is not much affected by fluctuations of sampling.

The harmonic mean, H, of n non-zero different variate-values x_i with frequencies f_i is given by

$$\frac{1}{H} = \frac{\sum_{i=1}^{n} \left(f_i \times \frac{1}{x_i} \right)}{\sum_{i} f_i} = \frac{1}{N} \sum_{i=1}^{n} \left(f_i \times \frac{1}{x_i} \right) \qquad \ldots(2{\cdot}5)$$

Thus *the harmonic mean of the variate-values is the reciprocal of the arithmetic mean of their reciprocals.*

The harmonic mean gives much importance to small items.

2·5·3. Median. Quartiles. The *median* is defined as that value of the variate for which there are equal frequencies of greater and smaller values, that is, the total frequency above (or below) this value is half the overall total frequency. Thus if the n values of the variate in an ungrouped frequency distribution are arranged in ascending order of magnitude, the median is the middle value if n is odd ; it is taken to be the arithmetic mean of the two middle values if n is even, though in this case any value between the two middle values will satisfy the definition.

Quantiles or *Partition values* are the values of the variate which divide the total frequency into a number of equal parts. Some important partition values are *quartiles, deciles* and *percentiles.*

Quartiles are those values which divide the total frequency into four equal parts ; deciles and percentiles divide into ten and one hundred equal parts respectively. Suppose the values of the variate have been arranged in ascending order of magnitude, then the value of the quartile having the position mid-way between the lower extreme and the median is the *lower* or *first quartile* and is denoted by Q_1 and that between the median and the upper extreme is the *upper* or *third quartile* and is denoted by Q_3. The median is thus one of the quartiles and is sometimes denoted by Q_2. Also clearly

median is the fifth decile D_5 and fiftieth percentile P_{50}. The difference between the upper and lower quartiles, *i.e.*, Q_3-Q_1 is called the *inter-quartile range*. For a grouped frequency distribution, the quartiles, the deciles and the percentiles are given by

$$\left.\begin{aligned} Q_i &= l+\frac{\left(\frac{iN}{4}-C\right)}{f}\times h,\ i=1,2,3,\\ D_j &= l+\frac{\left(\frac{jN}{10}-C\right)}{f}\times h,\ j=1,2,\ldots,9\\ P_k &= l+\frac{\left(\frac{kN}{100}-C\right)}{f}\times h,\ k=1,2,\ldots,99 \end{aligned}\right\}^{*} \quad \ldots(2\cdot6)$$

where l is the lower limit of the class in which the particular quantile lies, f is the frequency of this class. h is the width of this class.

C is the cumulative frequency up to and including the class preceding the class in which the particular quantile lies and N is the total frequency.

Example 2·3. *Calculate the three quartiles for the following frequency distribution of the number of marks obtained by 49 students in a class.*

Marks group	*No. of students*	*Marks group*	*No. of students*
5—10	5	25—30	5
10—15	6	30—35	4
15—20	15	35—40	2
20—25	10	40—45	2

(*B.A. Punjab, 1942*)

*The cumulative frequency upto the ith quartile Q_i is $\frac{iN}{4}$ and, therefore $\frac{iN}{4}-C$ is the frequency between the lower limit of the quartile class and Q_i Assuming that the frequency f in the quartile is distributed uniformly over the length h, which will give us this frequency is, $\frac{\frac{iN}{4}-C}{f}\times h$ and hence Q_i is given by (2·6). Similarly, the formulas for D_j and P_k can be obtained.

The following is the cumulative frequency table :

Class	*Frequency*	*Cumulative Frequency*
5—10	5	5
10—15	6	11
15—20	15	26
20—25	10	36
25—30	5	41
30—35	4	45
35—40	2	47
40—45	2	49
	Total 49	

To find Q_1, we have $\frac{N}{4}=\frac{49}{4}$ and this clearly lies in class 15—20.

$\therefore$ $l=15,\quad h=5,\quad C=11,\quad f=15.$

Hence $Q_1=15+\frac{\frac{49}{4}-11}{15}\times 5=15{\cdot}41$ marks.

For median, $\frac{N}{2}=24\frac{1}{2}$ and, therefore, median class is 15—20.

Hence median $=15+\frac{\frac{49}{2}-11}{15}\times 5=19{\cdot}50$ marks.

To find Q_3, we have $\frac{3N}{4}=\frac{147}{4}$ and this lies in the class 25—30.

$\therefore$ $l=25,\quad C=36,\quad f=5,\quad h=5.$

Hence $Q_3=25+\frac{(\frac{147}{4}-36)}{5}\times 5=25{\cdot}75$ marks.

25·4. Graphical Location of Partition Values. For finding graphically the values of the median, quartiles and other partition values we must first draw cumulative frequency curve or the *Ogive*. Now to determine the value of the median we mark a point along

the Y-axis corresponding to $N/2$. From this point, draw a line parallel to the X-axis and find the point where this line cuts the curve. Draw a line perpendicular to X-axis from this point. The distance between the origin and the foot of the perpendicular is the median.

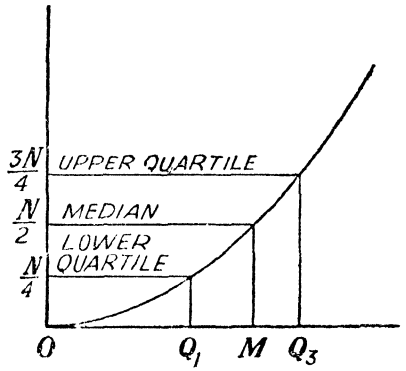


Fig. 9. Graphical location of the partition values.

For finding the value of Q_1 take a point on the Y-axis corresponding to $N/4$ and then proceed as above.

The method is applicable for determining the other partition values also from the Ogive.

Example 2·4. *Marks obtained by 80 students are given below : Read the value of the median of the distribution.*

Marks	*Frequency*	*Marks*	*Frequency*
0—10	3	30—40	30
10—20	9	40—50	18
20—30	15	50—60	5

(B.A. Hons., Delhi, 1957 ; I.A.S. 1954)

The cumulative frequency table is as under :

Marks	*Frequency*	*Cumulative Frequency*	*Marks*	*Frequency*	*Cumulative Frequency*
0—10	3	3	30—40	30	57
10—20	9	12	40—50	18	75
20—30	15	27	50—60	5	80

Plot the points (10, 3) ; (20, 12) ; (30, 27) ; (40, 57) ; (50, 75) ; (60, 80). Join them by a free hand smooth curve to obtain the Ogive.

To determine the median, we have $N/2=40$.

From the point 40 on the Y-axis, draw a straight line parallel to the axis of x cutting the Ogive at a point. Read the abscissa of this point. It is found out to be 34·3 marks (see Fig. 10).

Verification.

$$\text{Median}=30+\frac{40-27}{30}\times 10 \text{ marks}$$

$$=30+\tfrac{13}{3}=34\cdot3 \text{ marks.}$$

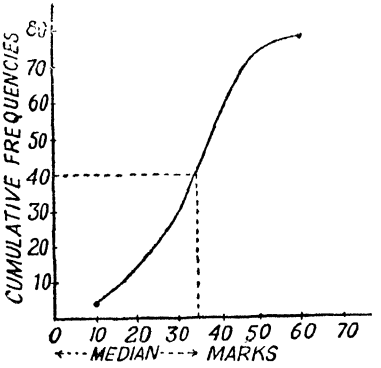


Fig. 10. Graphical location of tl median of the distribution of mark obtained by 80 students.

Ex. 1. Determine graphically the quartiles of the distribution given in Example 2·4.

Ex. 2. Draw a graph showing the number of students getting more than 0, 10, 20, 30, 40, 50 marks. Show that its intersection with the Ogive determines the median. Explain also how this curve can be used to determine the quartiles and the deciles.

2·5·5. Mode. The mode or modal value of the distribution is that value of the variate for which the frequency is maximum.

For a frequency distribution, the mode is computed by the following formula :

$$*\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h \qquad \ldots(2\cdot7)$$

where l is the lower limit of the modal class (*i.e.*, class having maximum frequency), f_m is the maximum frequency, f_1 and f_2 are the frequencies of the classes preceding and following the modal class.

Example 2·5. *Find the mode of the distribution given in Example 2·3, page* [illegible].

Clearly 15—20 is the modal class

$\therefore \quad l = 15,\ h = 5,\ f_1 = 6,\ f_2 = 10,\ f_m = 15,$

Hence $\text{mode} = 15 + \frac{15-6}{30-6-10} \times 5 = 18{\cdot}21$ marks.

2·6. Requisites for an ideal measure of central tendency. G. Udny Yule has suggested the following chief characteristics of an ideal measure of central tendency :

1. It should be rigidly defined.
2. It should be easily understandable and easy to calculate.
3. It should be based on all the observations.
4. It should be amenable to algebraic treatment.
5. It should be least affected by fluctuations of sampling.
6. It should not be affected much by the extreme values.

EXERCISES 2

1. Make a frequency table having grades of wages with class-interval of two annas each from the following data of daily wages received by 30 labourers in a certain factory. Daily wages in annas are :

14, 16, 16, 14, 22, 13, 15, 24, 12, 23, 14, 20, 17, 21, 22,
18, 18, 19, 20, 17, 16, 15, 11, 12, 21, 20, 17, 18, 19, 23.

(B.A. Punjab 1945)

---

*For computing the mode, we divide the modal class in the ratio

$$f_m - f_1 : f_m - f_2$$

Another formula for calculating the mode is :

$$\text{Mode} = l + \frac{f_2}{f_1 + f_2} \times h$$

where f_1 and f_2 are the frequencies of the classes preceding and following the modal class.

Ans.

Wages in annas	*No. of labourers*	*Wages in annas*	*No. of labourers*
11—13	3	19—21	5
13—15	4	21—23	4
15—17	5	23—25	3
17—19	6		

2. The weights in grams of 50 apples picked out at random from a consignment are as follows :

106, 107, 76, 82, 109, 107, 115, 93, 187, 95, 123, 125, 111,
92, 86, 70, 126, 68, 130, 129, 139, 119, 115, 128, 100, 186,
84, 99, 113, 204, 111, 141, 136, 123, 90, 115, 98, 110, 78,
90, 107, 81, 131, 75, 84, 104, 110, 80, 118, 82.

Form the grouped frequency table by dividing the variate range into intervals of equal width, each corresponding to 20 grams, in such a way that the mid-value of the first class corresponds to 70 grams.

Ans.

Weight in grams	*No. of apples*	*Weight in grams*	*No. of apples*
60—80	5	140—160	1
80—100	14	160—180	0
100—120	17	180—200	2
120—140	10	200—220	1

3. (*a*) Show how the limit of classes are defined in grouping a given data. (*B.A. Hons., Delhi, 1949*)

(*b*) Mention the considerations which help you in deciding the class-limits and class-intervals ? (*M. Sc. Agra, 1947*)

(*c*) What are the principles governing the choice of (1) the number of class-intervals (2) the length of the class-intervals and (3) the mid-point of the class-intervals. (*I.A.S. 1957*)

4. Form an ordinary frequency table from each of the following tables :

(*a*) *Marks*	*No. of students*	(*b*) *Marks*	*No. of students*
Below 10	3	Above 0	30
,, 20	8	,, 10	26
,, 30	17	,, 20	21
,, 40	20	,, 30	14
,, 50	22	,, 40	10
		,, 50	0

Ans.

(*a*) *Marks*	*No. of students*	(*b*) *Marks exceeding*	*Not exceeding*	*No. of students*
0—10	3	0 —	10	4
10—20	5	10 —	20	5
20—30	9	20 —	30	7
30—40	3	30 —	40	4
40—50	2	40 —	50	10
		50 and above		0
Total	22		Total	30

5. (*a*) Write a short note on the use of graphical methods in Statistics.

(*b*) Write short notes on the following :

(*i*) Graphical representation of distributions. (*B.A. Hons. Delhi, 1951*)

(*ii*) Histogram. (*B.A. Hons. Delhi, 1956*)

(*iii*) Ogive. (*M. Sc., Agra, 1953 ; B A. and B.Sc., Lucknow, 1948*)

(*iv*) Frequency polygon and frequency curve. (*B.A. Hons., Delhi, 1951*)

(*c*) Is there any relation between frequency polygon and frequency curve ? (*B.A., Punjab, 1953*)

6. Draw the histogram and frequency polygon for the data in Exs. 1. and 2, pages 21 and 22.

7. Represent the following distribution by a frequency polygon :

Source	*Frequency*	*Source*	*Frequency*
90—99	2	50—59	14
80—89	12	40—49	3
70—79	22	30—39	1
60—69	20	20—29	1

8. The following table shows the frequency distribution for the number of students per teacher in 750 colleges and professional schools :

Students x	1	4	7	10	13	16	19	22	25	28
Frequency f	7	46	165	195	189	89	28	19	9	3

Draw a histogram and a frequency polygon. (*B.A. Hons., Delhi, 1958*)

[**Hint.** 1, 4, 7, etc., are the mid-points of the classes 0—2, 3—5, 6—8, etc.]

9. Draw a histogram representing the following frequency distribution :

Wages (monthly) Rs.	*Frequency*	*Wages (monthly)* Rs.	*Frequency*
15	2	35	9
20	20	40	4
25	26	45	3
30	16		

[**Hint.** 15, 20....., 45 are the mid-points of the classes 12·5—17·5, 17·5—22·5,....., 42·5—47·5.]

10. Draw a histogram and a frequency polygon for the following distribution :

Degree of cloudiness x	*Frequency* f	*Degree of cloudiness* x	*Frequency* f
10	580	4	45
9	150	3	68
8	196	2	75
7	75	1	130
6	55	0	220
5	40		

[**Hint.** Take x as the central value and then plot.]

11. (*a*) Draw a histogram for the following fictitious data showing the distribution of incomes liable to surcharge in the financial year ending 31st March, 1957.

Income (in thousands)			Income (in thousands)		
Exceeding	*Not exceeding*	*No. of Persons*	*Exceeding*	*Not exceeding*	*No. of Persons*
2·0	2·5	32,143	20·0	25·0	950
2·5	3·0	28,516	25·0	30·0	650
3·0	4·0	24,812	30·0	40·0	212
4·0	5·0	15,910	40·0	50·0	170
5·0	7·5	12,615	50·0	75·0	50
7·5	10·0	9,720	75·0	100·0	25
10·0	15·0	4,330	100·0 and above		10
15·0	20·0	3,240			

(*b*) Group the frequencies in the intervals 2—5, 5—10, etc. and draw the frequency polygon and histogram.

12. Draw the histogram and the frequency curve for the following distribution of 727 deaths from scarlet fever (for various ages) classified according to age :

Age in years	*Number of deaths*	*Age in years*	*Number of deaths*
0—5	322	45—	6
5—	213	50—	7
10—	70	55—	5
15—	27	60—	0
20—	26	65—70	1
25—	17		
30—	12		——
35—	11	Total	77
40—	10		——

13. Draw a cumulative frequency curve for the following data :

Marks	*No. of students*	*Marks*	*No. of students*
0—4	4	12—16	8
4—8	6	16—20	4
8—12	10		

14. (*a*) Construct an Ogive from the following data relating to corn stalks (Height in feet). Locate graphically the median, quartiles and deciles :

Exceeding	*Not exceeding*	*No. of corn stalks*
3	4	92
4	5	211
5	6	612
6	7	1802
7	8	2029
8	9	913
9	10	243

(*b*) Draw a cumulative graph of the following distribution showing the monthly wages of a group of workmen.

Wages (Rs.)	*No. of workmen*	*Wages (Rs.)*	*No. of workmen*
20—	8	25—	25
21—	10	26—	15
22—	11	27—	9
23—	16	28—	6
24—	25		

(*I.A.S. 1950*)

15. The following marks were given to a batch of candidates :

66, 62, 45, 59, 32, 51, 56, 60, 51, 49,
25, 42, 54, 54, 58, 70, 43, 58, 50, 52,
38, 67, 50, 59, 48, 65, 71, 30, 46, 55,
82, 51, 63, 45, 53, 40, 36, 56, 70, 52,
67, 55, 57, 30, 63, 42, 74, 58, 44, 55.

Draw a cumulative frequency curve and from this curve, find the proportion of candidates receiving more than 50 marks. *(I.A.S. 1953)*

16. Draw an Ogive curve showing the number of candidates obtaining marks higher than assigned value from the following data :

Marks obtained	0—10	11—20	21—30	31—40	41—50	51—60	61—70	71—80
No. of candidates	33	108	221	439	529	495	322	153

17. The following table gives the frequency distribution of marks in two examination papers :

Class	*Frequency*		*Class*	*Frequency*	
	A	B		A	B
11—15	8	5	56—60	53	65
16—20	15	16	61—65	48	48
21—25	39	24	66—70	45	32
26—30	47	39	71—75	39	25
31—35	52	50	76—80	32	18
36—40	53	61	81—85	23	13
41—45	54	68	86—90	17	9
46—50	56	63	91—95	13	9
51—55	59	72	96—100	8	5

Draw the cumulative frequency curve of mark distribution and answer the following questions :

(*a*) For paper A, if 30% of the candidates are to be failed, what is the pass mark ?

(*b*) For paper B, if pass mark is 33, what per cent of candidates fail ?

[**Hint.** (*a*) Divide the ordinate of the highest point of the Ogive into 100 equal divisions. Through the 30th division, draw a horizontal line and find its intersection with the curve. The abscissa of this point is the required pass mark.

[**Ans.** (*a*) 39, (*b*) 18]

18. Give one illustration each of the type of data for which you would expect the frequency curve to be :

(*i*) fairly symmetrical, (*ii*) positively skewed,
(*iii*) negatively skewed, (*iv*) J-shaped.

(B.A. Punjab, Sept. 1953)

19. (*a*) What are the various measures of location of a distribution and to what purposes are they used ? *(B.A. Hons., Delhi, 1955, 57)*

(*b*) In what circumstances would you consider the Arithmetic Mean, the Geometric Mean, the Harmonic Mean respectively, the most suitable statistic to describe the central tendency of a distribution ? (*I.A.S., 1964*)

20. The frequency distribution below gives the cost of production of sugarcane in different holding, obtain the arithmetic mean.

Class	*Frequency*	*Class*	*Frequency*
2—6	1	18—	52
6—	9	22—	36
10—	21	26—	19
14—	47	30—34	3

[**Ans.** 19·21]

21. The yields of grain (x lbs.) from 500 small plots are grouped into classes with a common class-interval (0·2 lb.) in the table below, the values of x given being the mid-values of the class. Show that the mean of the distribution is 3·95 lbs., the median is 3·95 lbs., and the quartiles 3·63 lbs. and 4·28 lbs.

x	f	x	f	x	f	x	f	x	f
2·8	4	3·4	47	4·0	88	4·6	35	5·2	4
3·0	15	3·6	63	4·2	60	4·8	10	—	—
3·2	20	3·8	78	4·4	59	5·0	8	—	—

(*B.A. Hons., Delhi, 1952*)

22. Obtain the median and the mode for the following table :

No. of days absent	*No. of students*	*No. of days absent*	*No. of students*
Less than 5	29	Less than 30	644
,, ,, 10	224	,, ,, 35	650
,, ,, 15	465	,, ,, 40	653
,, ,, 20	582	,, ,, 45	655
,, ,, 25	634		

[**Ans.** median=12·2 days, mode=11·4 days]

23. Find the median and the two quartiles for the following table ;

Marks	*No. of students*	*Marks*	*No. of students*
Below 10	15	Below 50	96
,, 20	35	,, 60	127
,, 30	60	,, 70	198
,, 40	84	,, 80	250

[**Ans.** median=59·3 marks, Q_1=31·04 marks, Q_3=68·52 marks]

24. Find the median, lower and upper quartiles, 8th decile and 5th percentile for the following distribution of 45 workers :

Monthly wages (in Rs.)	*No. of workers*	*Monthly wages (in Rs.)*	*No. of workers*
1—2·99	6	9—10·99	21
3—4·99	53	11—12·99	16
5—6·99	85	13—14·99	4
7—8·99	56	15—16·99	4

[**Ans.** median=Rs. 6·49, Q_1=Rs. 5·05, Q_3=Rs. 8·40, D_8=Rs. 8·86, P_{56} Rs.=6·80]

25. Find the median, lower and upper quartiles, 4th decile and 40th percentile for the following distribution :

Marks	*No. of students*	*Marks*	*No. of students*
0—4	10	14—18	5
4—8	12	18—20	8
8—12	18	20—25	4
12—14	7	25 and over	6

[**Ans.** median=10·9 marks. Q_1=6·5 marks, Q_3=18·12 marks, D_4=10 marks, P_{40}=10 marks]

26. (*a*) Calculate the modal and the median wages of the following distribution of wages per thousand employees in a certain factory :

Daily wages (in Rs.)	*No. of employees*	*Daily wages (in Rs.)*	*No. of employees*
2—4	3	14—	204
4—	13	16—	139
6—	43	18—	69
8—	102	20—	25
10—	175	22—	6
12—	220	24—26	1

[**Ans.** Modal wages=Rs. 13·50, median wage=Rs. 13·50]

(*b*) The expenditure of 100 families is given as under :

Expenditure	0—10	10—20	20—30	30—40	40—50
No. of families	14	?	27	?	15

The median and mode for the distribution are Rs. 25 and Rs. 29 respectively. Calculate the missing frequencies. [**Ans.** 25, 26]

27. Calculate (*i*) the unweighted mean of the prices in column III and (*ii*) the mean obtained by weighing each price by the quantity consumed.

I	II	III
Articles of food	*Quantity consumed in maunds*	*Price in Rs. per maund*
Flour	11·50	5·8
Ghee	5·60	58·4
Sugar	0·28	8·2
Potatoes	0·16	2·5
Oil	0·35	20·0
Total	17·89	94·9

[**Ans.** (*i*) Rs. 18·98, (*ii*) 22·55]

28. Determine graphically the values of the median and the two quartiles of the data given in exercises 14, 15 and 16 on pages 24 and 25.

29. A distribution consists of 3 components with frequencies 45, 40 and 65 having their means 2, 2·5 and 2 respectively. Prove that the mean of the combined distribution is 2·13 approximately.

30. (*a*) Show that the arithmetic mean of the first n natural numbers is $\frac{1}{2}(n+1)$.

(*b*) Show that the weighted arithmetic mean of first n natural numbers whose weights are equal to the corresponding numbers is equal to $\frac{1}{3}(2n+1)$.

(*c*) The arithmetic mean of n numbers of a series is $\bar{x}$. The sum of the first $(n-1)$ numbers is k. Show that the nth number is $n\bar{x}-k$.

(*d*) The frequencies of values 0, 1, 2....., n of a variable are given by

$$q^n,\ {}^nc_1\, q^{n-1}p,\ {}^nc_2\, q^{n-2}\, p^2,\ldots\ldots,\ p^n$$

where $p+q=1$. Show that the mean is np.

31. If $x_1, x_2,\ldots\ldots, x_n$ are non-zero positive numbers, prove that $A \geqslant G$.

[**Hint.** Use the result $\dfrac{x_1+x_2+\ldots+x_n}{n} \geqslant (x_1.\ x_2.\ldots.\ x_n)^{\frac{1}{n}}$, the equality sign holding when the x's are equal.]

32. Show that $\left[\sum_{i=1}^{n} w_i\right]\left[\sum_{i=1}^{n} w_i(x_i-\bar{x}_w)^2\right] = \sum_{i=1}^{n}\ \sum_{j>i} w_iw_j(x_i-x_j)^2$

where $\sum_{i=1}^{n} w_i \neq 0$ and $\bar{x}_w$ is the weighted mean of x_i's with weights w_i.

[**Hint.** R.H.S. $= \frac{1}{2}\sum_{1}^{n}\sum_{1}^{n} w_i.w_j.(x_i-x_j)^2 = \frac{1}{2}\sum_{1}^{n}\sum_{1}^{n} w_iw_j\{(x_i-\bar{x}_w)-(x_j-\bar{x}_w)\}^2$
$=$ etc.

(*B.Sc. Hons. Sub. 1968*)

33. Show that for J-shaped distribution with the maximum frequency towards the lower values of the variate, the median is nearer to Q_1 (lower quartile) than to Q_3 (upper quartile). (*M.Sc. Agra 1954*)

34. A distribution $x_1, x_2,\ldots\ldots, x_n$ with frequencies $f_1, f_2,\ldots\ldots, f_n$ is transformed into the distribution $X_1, X_2,\ldots\ldots, X_n$ with the same corresponding frequencies of the relation $X_i=ax_i+b$, $i=1, 2,\ldots\ldots, n$ where a and b are constants. Show that the mean, mode and median of the new distribution are given in terms of those of the first distribution by the same transformation.

(*B.A., Aligarh, 1960*; *B.Sc. Agra, 1961*)

35. A variate takes values $a, ar, ar^2,\ldots\ldots, ar^{n-1}$ each with frequency unity. Show that the *A.M.* A, is $\dfrac{a(1-r^n)}{n(1-r)}$, the *G.M.*, G, is $a.r^{\frac{n-1}{2}}$ and the *H. M.*, H, is $\dfrac{an(1-r)r^{n-1}}{1-r^n}$. Prove that $AH=G^2$. Prove also that $A>G>H$ unless $n=1$ when all the three means coincide.

36. A person travels n equal distances with velocities $v_1, v_2,\ldots. v_n$. Show that his average velocity (total distance/total time) cannot exceed $(v_1+v_2+\ldots+v_n)/n$. When will it be equal to this value ?

37. An aeroplane flies around a square the sides of which measure 100 miles each. The aeroplane covers at a speed of 100 miles per hour the first-side, at 200 m.p.h. the second side, at 300 m.p.h. the third side, and at 400 m.p.h. the fourth side. Use the correct mean to find the average speed round the square.

[**Ans.** H.M. 192·3 m.p.h.]

38. (A) Which type of average would be suitable for the following cases ? Give reasons :

(*a*) Size of agricultural holdings.

(*b*) Comparison of intelligence.

(*c*) Marks obtained in any examination.

(*d*) The average speed of a man travelling 1 mile at 5 m.p.h., 2 miles at 7 m.p.h. and $3\frac{1}{2}$ miles at 8 m.p.h.

(*e*) *Per capita* income in several countries.

(B) What measure do you suggest for each of the following distributions:

(*a*) Incomes of workers in a factory,

(*b*) Heights of students, and

(*c*) Number of petals of flowers.

(*B.A. Hons. Pb., 1968 ; M.A., Punjab, 1960*)

[**Ans.** (*a*) mode (*b*) median (*c*) mode (*d*) harmonic mean (*e*) weighted mean]

39. A population of values is symmetrically distributed about the constant k. Examine whether the following statements are correct :

(*a*) the mean is k,

(*b*) the mean coincides with the mode,

(*c*) the distribution cannot be bimodal,

(*d*) $\frac{1}{2}$ (upper quartile+lower quartile) is equal to k,

(*e*) the median coincides with mean.

(*M.A., Eco., Statistics, Delhi, 1959*)

[**Ans.** (*a*) Correct (*b*) Incorrect (*c*) Incorrect (*d*) Correct (*e*) Incorrect.]

40. In a frequency table the upper boundary of each interval has a constant ratio to the lower boundary. Show that the geometric mean, G, may be expressed by the formula

$$\log G = x_0 + \frac{C}{N} \sum_i f_i (i-1)$$

where x_0 is the logarithm of the mid-values of the first interval and C is the logarithm of the ratio between the upper and lower boundaries.

(*B.Sc. Agra, 1962 ; B.Sc. Meerut, 1970*)

41. If $\bar{x}_1 = \frac{1}{n} \sum_1^n x_i$, $\bar{x}_2 = \frac{1}{n} \sum_2^{n+1} x_i$ and $\bar{x}_3 = \frac{1}{n} \sum_3^{n+2} x_i$, then show that

(*a*) $\bar{x}_2 = \bar{x}_1 + \frac{1}{n}(x_{n+1} - x_1)$ and

(*b*) $\widetilde{x_3} = \bar{x}_2 + \frac{1}{n}(x_{n+2} - x_2)$ (*B.Sc., Calcutta, 1966*)

42. If G_x be the geometric mean of Nx's and G_y be the geometric mean of Ny's, then prove that the geometric mean G of the $2N$ values is given by

$$G^2 = G_x . G_y.$$

43. The following table shows the distribution of 100 families according to their expenditure per week. Number of families according to expenditure groups Rs. 10—20 and Rs. 30—40 are missing from the table. The median and mode are given to be Rs. 25 and 24. Calculate the missing frequencies and then arithmetic mean of the data.

Expenditure :	0—10	10—20	20—30	30—40	40—50
No. of families :	14	?	27	?	15

(*B.Sc., Jiwaji, 1971*)

Ans. 25, 25]

3

Measures of Dispersion, Skewness and Kurtosis. Moments of Frequency Distributions

3·1. Dispersion. It should be clear that a measure of central tendency by itself can exhibit only one of the important characteristics of distribution and, therefore, while studying a distribution it is equally important to know how the variates are clustered around or scattered away from the point of central tendency. Thus if the distribution of marks obtained by two groups of students have the same average marks, we will certainly like to know whether the two groups consist of students of average and near-average intelligence or are made up of a large number of very bright and very dull students. Fig. 8 (*b*), given on page 13, clearly shows that the two distributions centre round the same point, *viz.*, the mean, yet they differ in variation from the mean. Such variation is variously called as *dispersion*, *spread*, *scatter*, or *variability*, and we consider in the next section some measures that have been proposed to get a quantitative measure of this dispersion.

3·1·1. Characteristics of an ideal measure of Dispersion. The following are the requisites for an ideal measure of dispersion :

1. It should be rigidly defined.
2. It should be easily understandable and easy to calculate.
3. It should be based on all the observations.
4. It should be amenable to further mathematical treatment.
5. It should be best affected by fluctuations of sampling.

3·2. Measures of Dispersion. The measures of dispersion can be classified into three groups :

(*i*) Measures of the difference between representative variat values, such as the *range*, the *interquartile* or the *interdecile range*

(*ii*) Measures obtained from the deviations of every variate value from some central value, such as the *mean deviation from the mean* or the *mean deviation from the median* or the *standard deviation.*

(*iii*) Measures obtained from the variations of all the variates among themselves such as the *mean difference.*

3·2·1. Range. Interquartile Differences. Range is the difference between the extreme values of the variate, that is, range is equal to $x_n - x_1$, where x_n and x_1 are the greatest and the least values of the variate. It is easily understood and computed, but depends exclusively on the two extreme values, while it is desirable that a measure of dispersion or any other measure should depend on all the variate values. Nevertheless, the range is a very useful measure in industrial engineering work, specially in *statistical quality control* work.

The interquartile range is the difference between the upper and lower quartiles. The *semi-interquartile* range or the *quartile deviation* is half of it *i.e.*

$$Q = \frac{Q_3 - Q_1}{2} \qquad \ldots(3\cdot1)$$

The *interdecile range* is the difference between the ninth and the first decile and contains eighty per cent of the total frequency, while the interquartile range contains fifty per cent. Both these measures are easily calculated and give a fairly good idea of the spread of the distribution and are commonly used in elementary descriptive statistics, but in advance theory their use is rather rare because of the mathematical difficulty of handling them.

3·2·2. Average Deviation or Mean Deviation. It would be useless to take the sum of the deviations from the mean as a measure of dispersion since some deviations are positive and some are negative and their algebraic sum is zero. (Compare § 2·5·1, page 13).

However, if we take the mean of the absolute values of the deviations, we get a useful measure of dispersion called the mean deviation from the mean. Thus

$$\text{Mean deviation from the mean} = \frac{1}{N} \sum_{i=1}^{n} f_i \mid x_i - \bar{x} \mid \qquad \ldots(3\cdot2)$$

where $N = \sum_{i=1}^{n} f_i$, $\bar{x} = \frac{1}{N} \sum_{1}^{n} f_i x_i$.

this formula, $\bar{x}$ has to be replaced by median or mode according
e mean deviation from median or mode is required.

The mean deviation is based on all the values of the variate and hence it is a better measure of dispersion. However, there is an element of artificiality in mean deviation since the deviations are not taken with their proper signs. The mean deviation is also difficult to handle mathematically.

Example 3·1. *Calculate the mean deviation (from the mean) of the following distribution :*

Marks	0–10	10–20	20–30	30–40	40–50
No of students	5	8	15	16	6

We first calculate mean by applying formula (2·3″), page 15 and then apply formula (3·2), page 32 to find mean deviation.

Mid-value	$u=\frac{x-25}{10}$	f	fu	$x-\bar{x}$	$f\mid x-\bar{x}\mid$
5	−2	5	−10	−22	110
15	−1	8	−8	−12	96
25	0	15	0	−2	30
35	1	16	16	8	128
45	2	6	12	18	108
	Total	50	10		472

$$\text{Mean}=a+\frac{h}{N}\Sigma fu=25+\frac{10}{50}\times 10=27 \text{ marks.}$$

$$\text{Mean deviation}=\frac{1}{N}\Sigma f_i\mid x_i-\bar{x}\mid=\frac{472}{50}=9{\cdot}44 \text{ marks.}$$

Ex. Calculate mean deviation from mode of the above distribution.

3·2·3. Minimal Property of Average Deviation. *The sum of the absolute deviations from the median is a minimum.* In other words, *the mean deviation from median is less than that measured from any other value.*

Proof. Let O be the median, A be a point distant c from O on the right of O. Let D_O, D_A denote the sum of absolute deviations from O and A respectively and let D_{OL}, D_{OR} denote the contributions to D_O by the observations on the left and the right of O respectively. Similarly, let D_{AL}, D_{AR} denote the contributions to D_A by the same set of observations.

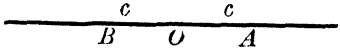


Fig. 11 (*a*)

Then

$$\begin{aligned} D_A-D &=(D_{AL}+D_{AR})-(D_{OL}+D_{OR}) \\ &=(D_{AL}-D_{OL})+(D_{AR}-D_{OR}) \end{aligned}$$

The first factor on the right is positive and has the values $(N/2)c$. The second factor may be positive or negative but in either case cannot exceed $(N/2)c$ in magnitude since the contributions from $N/2$ observations on the right of O can be both positive and negative.

$$\therefore \qquad D_A \leqslant D_O.$$

Similarly, $$D_B \geqslant D_O.$$

In fact, the equality sign will arise only if the median is indeterminate and can be any value in a certain range *e.g.*, in the figure any point in the range CD can be regarded as the median and the sum of the absolute deviations from any point of CD will be same and less than that measured from any value outside the range. Thus we have proved that the mean deviation is least when measured from the median.

$\frac{N}{2}$ OBSERVATION C-D $\frac{N}{2}$ OBSERVATION

Fig. 11(*b*).

Aliter. Mean deviation from any point a,

$$= \frac{1}{N} \sum_{i=1}^{n} f_i \mid x_i - a \mid = \frac{1}{N} S$$

where $N = \sum_{i=1}^{n} f_i$ and $S = \sum_{i=1}^{n} f_i \mid x_i - a \mid$

We shall now show that mean deviation or S will be minimum when a is the median.

Let $$x_m < a < x_{m+1}$$

$\Rightarrow$ $x_i < a$, $i = 1, 2, \ldots, m$, and $x_i > a$, $i = m+1, m+2, \ldots, 2n$

$$\text{Now } S = \sum_{i=1}^{m} f_i \mid x_i - a \mid + \sum_{i=m+1}^{n} f_i \mid x_i - a \mid$$

$$= \sum_{i=1}^{m} f_i (a - x_i) + \sum_{i=m+1}^{n} f_i (x_i - a)$$

$$= a\left[\sum_{i=1}^{m} f_i - \sum_{i=m+1}^{n} f_i\right] + \left[-\sum_{i=1}^{m} f_i x_i + \sum_{i=m+1}^{n} f_i x_i\right]$$

For S to be minimum or maximum, we have

$$0 = \frac{\partial S}{\partial a} = \left[\sum_{i=1}^{m} f_i - \sum_{i=m+1}^{n} f_i\right]$$

$$\Rightarrow \qquad \sum_{i=0}^{m} f_i = \sum_{i=m+1}^{n} f_i$$

Thus a is the median. For values of a to the left of the median S is a decreasing function of a and for values of a to the right of the

median, S is an increasing function, and also $0=\frac{\partial S}{\partial a}$, it follows that the mean deviation from median is less than that measured from any other value.

3·2·4. Variance. Standard Deviation. The *mean square deviation* of the variate x from any value a is defined as the mean value of the square of the deviations of x from a and is, thus, given by

$$\mu_2'=s^2=\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i-a)^2 \qquad \ldots(3\cdot3)$$

The positive square root of this quantity is called the *root-mean square deviation* taken from a. In the particular case when deviation is taken from the mean $\bar{x}$, the mean-square deviation is called the *variance* of x and is denoted by μ_2 or σ_x^2 so that we have

$$\mu_2=\sigma_x^2=\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i-\bar{x})^2. \qquad \ldots(3\cdot4)$$

The standard deviation (*s.d.*) of x is the positive square root of the variance of x and is denoted by σ_x. The *coefficient of variation* is defined as $100\times(s.d.\div\text{mean})$. It should be noted that this ratio is dimensionless since *s.d.* and mean have the dimensions of the variate.

Note. The formula for variance *viz.* equation (3·4) can be put as :

1. $$\sigma_x^2=\frac{1}{N}\sum_{1}^{n} f_i\,(x_i-\bar{x})^2=\frac{1}{N}\sum_{1}^{n} f_i\,(x_i^2-2x_i\bar{x}+\bar{x}^2)$$

$$=\frac{1}{N}\sum_{1}^{n} f_i\,x_i^2-2\bar{x}.\frac{1}{N}\sum_{1}^{n} f_i x_i+\bar{x}^2.\ \frac{1}{N}\sum_{1}^{n} f_i$$

$$=\frac{1}{N}\sum_{1}^{n} f_i x_i^2-2\bar{x}.\ \bar{x}+\bar{x}^2.1\ =\frac{1}{N}\sum_{1}^{n} f_i x_i^2-\bar{x}^2 \qquad \ldots\ldots(3\cdot4')$$

$$=\frac{1}{N}\sum_{1}^{n} f_i x_i^2-\left(\frac{1}{N}\Sigma f_i x_i\right)^2 \qquad \ldots\ldots(3\cdot4'')$$

2. $$\sigma_x^2=\frac{1}{N}\sum_{1}^{n} f_i\ (x_i-\bar{x})^2=\frac{1}{N}\Sigma f_i\ \overline{(x_i-a}-\overline{\bar{x}-a)^2}$$

$$=\frac{1}{N}\sum_{1}^{n} f_i\,(x_i-a)^2-2.\ \frac{1}{N}\sum_{1}^{n} f_i\,(x_i-a).(\bar{x}-a)$$

$$+(\bar{x}-a)^2.\ \frac{1}{N}\sum_{1}^{n} f_i$$

$$=\frac{1}{N}\sum_{1}^{n} f_i\,(x_i-a)^2-\left[\frac{1}{N}\sum_{1}^{n} f_i\,(x_i-a)\right]^2$$

$$\sigma_x^2=\frac{1}{N}\sum_1^n f_i d_i^2-\left(\frac{1}{N}\sum_1^n f_i d_i\right)^2 \text{ where } d_i=x_i-a \quad ...(3\cdot4''')$$

3·2·5. Relation between Standard and Root-mean Square Deviations.

We have
$$s^2=\frac{1}{N}\sum_{i=1}^n f_i\,(x_i-a)^2$$
$$=\frac{1}{N}\sum_i f_i\,(x_i'+d_i)^2,$$
where $x_i'=x_i-\bar{x}$ and $d=\bar{x}-a$

$$\therefore \quad s^2=\frac{1}{N}\sum_{i=1}^n f_i x'^2+\frac{2d}{N}\sum_{i=1}^n f_i x_i+\frac{1}{N}d^2\sum_{i=1}^n f_i$$

The second term on the right vanishes since the sum of the deviations from the mean is zero, and accordingly

$$s^2=\sigma_x^2+d^2 \quad \text{or } \mu'_2=\mu_2+d^2 \qquad ...(3\cdot5)$$

Clearly s^2 will be least when $d=0$, that is, when $a=\bar{x}$.

Hence *the root-mean square deviation is least when deviations are measured from the mean i.e., the standard deviation is least possible root-mean square deviation.*

Note. (3·5) given above can be put in the following form :

$$\sigma_x^2=\frac{1}{N}\sum_i f_i\,(x_i-a)^2-(\bar{x}-a)^2$$
$$=\frac{1}{N}\sum_i f_i X_i^2-\left[\frac{1}{N}\sum_i f_i\,(x_i-a)\right]^2 \text{ where } X_i=x_i-a$$

Properties of the standard deviation

We shall now show that the standard deviation is unaltered by the *change of origin* but its value gets changed by the *change of scale.* Let us therefore take 'a' as the new origin and scale be 'h' times the original scale so that

$$u=\frac{x-a}{h}\Rightarrow x=a+hu \text{ and } \bar{x}=a+h\,\bar{u} \qquad ...(3\cdot6)$$

Thus from (3·4), we have

$$\sigma_x^2=\sum_{i=1}^n fh^2\,(u_i-u)^2=h^2\sigma_u^2 \qquad ...(3\cdot7)$$

$$\Rightarrow \quad \sigma_x=h\sigma_u$$

where σ_x^2 and σ_u^2 denote variances of variates x and u respectively.

Example 3·2. *Find the mean and the variance of the distribution in which the values of x are 1, 2,.....,N, the frequency of each being unity.*

Here $\bar{x} = \dfrac{1+2+\ldots+N}{N} = \frac{1}{2}(N+1)$.

The mean square deviation from $x=0$ is

$$s^2 = \frac{1^2+2^2+\ldots+N^2}{N} = \frac{1}{6}(N+1)(2N+1)$$

Hence $\sigma^2 = \frac{1}{6}(N+1)(2N+1) - \frac{1}{4}(N+1)^2$

$= \frac{1}{12}(N^2-1)$.

Example 3·3. *Find the s.d. and the coefficient of variation from the following table giving wages of 230 persons :*

Wages (in Rs.)	*No. of persons*	*Wages (in Rs.)*	*No. of persons*
70–80	12	110–120	50
80–90	18	120–130	45
90–100	35	130–140	20
100–110	42	140–150	8

We first calculate the mean and standard deviation.

Wages (in Rs.)	*Mid-value x*	*No. of persons f*	$u = \dfrac{x-105}{10}$	u^2	fu	fu^2
70–80	75	12	−3	9	−36	108
80–90	85	18	−2	4	−36	72
90–100	95	35	−1	1	−35	35
100–110	105	42	0	0	0	0
110–120	115	50	+1	1	+50	50
120–130	125	45	+2	4	+90	180
130–140	135	20	+3	9	+60	180
140–150	145	8	+4	16	+32	128
Total		230			+125	753

Mean $= a + h\bar{u} = 105 + 10 \times \frac{125}{230} =$ Rs. 110·4

Using formula (3·7), we have

$S.D. = 10 \times \sqrt{\frac{753}{230} - (\frac{125}{230})^2} =$ Rs. 17·3.

$\therefore$ Coefficient of variation (C.V.) $= 230 \times \dfrac{17{\cdot}3}{110{\cdot}4} = 15{\cdot}67$.

Example 3·4. *(Variance of the combined group) The standard deviation of two sets containing n_1 and n_2 members are σ_1 and σ_2 respectively, being measured from their respective means m_1 and m_2. If the two sets are grouped together as one set of (n_1+n_2) members,*

show that the standard deviation, σ, of this set, measured from its mean is given by

$$\sigma^2=\frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2}+\frac{n_1n_2}{(n_1+n_2)^2}.\ (m_1-m_2)^2.$$

(B.A., Hons., Delhi, 1949 ; B A., Madras, 1948, 50, 53)

Let s_1^2 and s_2^2 be the mean square deviations of the two sets respectively, then the mean square deviation, s^2 of the combined set is given by

$$s^2=\frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} f_i\,(x_i-a)^2$$

$$=\frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1} f_i\,(x_i-a)^2+\sum_{i=n_1+1}^{n_1+n_2} f_i(x_i-a)^2\right]$$

$$=\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2}$$

$$=[n_1(\sigma_1^2+d_1^2)+n_2(\sigma_2^2+d_2^2)]/(n_1+n_2)$$

or
$$s^2=\frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2}+\frac{n_1d_1^2+n_2d_2^2}{n_1+n_2} \qquad \ldots(3\cdot8)$$

where $d_1=m_1-a$, $d_2=m_2-a$, a being an arbitrary number.

Let a be taken as the mean of the whole set, then

$$a=m=(n_1m_1+n_2m_2)/(n_1+n_2),\ s^2=\sigma^2.$$
$$d_1=m_1-a=n_2(m_1-m_2)/(n_1+n_2),$$
$$d_2=n_1\,(m_2-m_1)/(n_1+n_2).$$

Substituting the values of a, d_1, d_2 and s^2 in (3·8) and simplifying we get the required result.

Example 3·5. *Show that, for any discrete distribution, the standard deviation is not less than the mean deviation from the mean.*
(B. Sc. Gujarat, 1971 ; M.Sc. Allahabad 1967 ; M.A. Eco. Statistics ; Delhi 1958 ; B.A. Hons., Delhi 1966 ; I.A.S. 1968, 1961 ; M.Sc. Agra 1943)

We have to show that S.D. $\geq$ Mean deviation from mean

i.e.,
$$\frac{1}{N}\sum_i f_i(x_i-\bar{x})^2 > \left[\frac{1}{N}\sum_i f_i\ |\ x_i-\bar{x}\ |\ \right]^2$$

or $(\Sigma f_iy_i^2)\,(\Sigma f_i) \geqslant (\Sigma f_iy_i)^2$, where $y_i=|\ x_i-\bar{x}\ |$, $N=\Sigma f_i$

or
$$\Sigma f_i^2y_i^2+\sum_{i\neq j}\Sigma f_if_j(y_i^2+y_j^2) \geqslant \Sigma f_i^2y_i^2+2\sum_{i\neq j}\Sigma f_if_jy_iy_j$$

or
$$\sum_{i\neq j}\Sigma f_if_j(y_i-y_j)^2 \geqslant 0$$

i.e.,
$$\sigma_y^2 \geqslant 0$$

which is always true. Hence the result follows.

Note. The equality sign in (3·9) will hold when all the variate values are equal.

Ex. Show that the variances σ^2 of a distribution with k component distributions is given by

$$N\sigma^2=\sum_{i=1}^{k} n_i(\sigma_i^2+d_i^2)$$

where m_i is the mean of the i-th component with frequency n_i, m is the mean of the whole distribution, $d_i=m_i-m$ and N is the total frequency $\sum_{i=1}^{k} n_i$.

3·2·6. Mean Difference. For a frequency distribution the coefficient of mean difference is defined by

$$\Delta_1=\frac{1}{N(N-1)}\sum_{j=1}^{n}\sum_{\substack{i=1\\ i\neq j}}^{n} |\, x_i-x_j \,|\, f_i f_j$$

or by

$$\Delta_1=\frac{1}{N^2}\sum_{j=1}^{n}\sum_{\substack{i=1\\ i\neq j}}^{n} |\, x_i-x_j \,|\, f_i f_j$$

depending on whether we exclude or include zero differences. The difference lies only in the denominator and is unimportant for large N.

This coefficient was first defined by Gini (1912) and has a certain theoretical attraction, since it depends on all the variate values and the differences of the values among themselves and not on deviations from some central value. It is, however, difficult to compute and the appearance of the absolute values makes it difficult to be handled mathematically.

The dimensionless measure corresponding to mean difference is Gini's coefficient of concentration which is defined by

$$G_1=\frac{\Delta_1}{2\mu_1'}\text{, where } \mu_1' \text{ is given by (3·3), page 40.}$$

Ex. 1. Prove that $\frac{1}{N^2}\sum_{j=1}^{n}\sum_{\substack{i=1\\ i\neq j}}^{n}(x_i-x_j)\,f_i f_j=0.$

Ex. 2. Prove that $\frac{1}{N^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(x_i-x_j)^2 f_i f_j=2 \text{ var } x.$

Ex 3. Find G_1 for the distribution in which the variate takes values a, $a+d$,.....$a+n-1d$, each with the same frequency.

3·3. Moments. We define the rth moment about any point a as μ_r' where

$$\mu_r' = \frac{1}{N} \sum_{i=1}^{n} f_i(x_i - a)^r, \quad \sum_{i=1}^{n} f_i = N. \qquad \ldots(3\cdot10)$$

Thus rth moment can be regarded as the arithmetic mean of the variate $(x-a)^r$.

Thus rth moment about the mean $\bar{x}$ is written without dash so that

$$\mu_r = \frac{1}{N} \sum_{i=1}^{n} f_i(x_i - \bar{x})^2 \qquad \ldots(3\cdot11)$$

The moments about the mean are of much importance because of their values being independent of the origin of reference.

Putting $r=0, 1$ respectively in (3·10) and (3·11), we have

$$\mu_0' = \mu_0 = \frac{1}{N} \sum_{i=1}^{n} f_i = 1 \qquad \ldots(3\cdot12)$$

$$\mu_1' = \frac{1}{N} \sum_{i=1}^{N} f_i(x_i - a) = \bar{x} - a = d, \text{ say} \qquad \ldots(3\cdot13)$$

$$\mu_1 = \frac{1}{N} \sum_{i=1}^{N} f_i(x_i - \bar{x}) = 0 \qquad \ldots(3\cdot14)$$

3·3·1. Moments about the Mean in terms of Moments about any point and conversely.

We have

$$\mu_r = \frac{1}{N} \sum_{i=1}^{n} f_i(x_i - \bar{x})^r = \frac{1}{N} \sum_{i=1}^{n} f_i(\overline{x_i - a} - \overline{\bar{x} - a})^r$$

$$= \frac{1}{N} \sum_{i=1}^{n} f_i(X_i - d)^r \quad \text{where } X_i = x_i - a,\ d = \bar{x} - a$$

$$= \frac{1}{N} [\Sigma f_i X_i^r - {}^rC_1 d \Sigma f_i X_i^{r-1} + {}^rC_2 d^2 \Sigma f_i X_i^{r-2} - \ldots$$

$$+ {}^rC_{r-1}(-d)^{r-1} \Sigma f_i X_i + (-d)^r \Sigma f_i]$$

$$= \mu_r' - {}^rC_1 d\mu'_{r-1} + {}^rC_2 d^2 \mu'_{r-2} - \ldots + (-1)^{r-1} r d^{r-1} \mu_1' + (-1)^r d^r$$

or $$\mu_r = \mu_r' - {}^rC_1 \mu_1' \mu'_{r-1} + {}^rC_2 \mu_1'^2 \mu'_{r-2} - \ldots + (-1)^{r-1}(r-1)\mu_1'^r \qquad \ldots(3\cdot15)$$

In particular,

$$\mu_2 = \mu_2' - \mu_1'^2 \qquad \ldots(3\cdot16)$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \qquad \ldots(3\cdot17)$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 \qquad \ldots(3\cdot18)$$

These three formulae should be committed to memory. It should be noticed that in each of these relations, the sums of the coefficients of various terms of the right member is equal to zero and that each term of the right member is of the same dimension as the term on the left.

Conversely,

$$\mu_r' = \frac{1}{N}\sum_i f_i(x_i-a)^r = \frac{1}{N}\sum_i f_i(\overline{x_i-\bar{x}}+\overline{\bar{x}-a})^r$$

$$= \frac{1}{N}\sum_i f_i(x_i'+d)^r \quad \text{where } x_i' = x_i-\bar{x},\ d=\bar{x}-a$$

$$= \frac{1}{N}\left[\Sigma f_i x_i'^r + {}^rC_1 d\Sigma f_i x_i'^{r-1} + {}^rC_2 d^2\Sigma f_i x_i'^{r-2} + \ldots\ldots \right.$$

$$\left. + {}^rC_{r-1}d^{r-1}\Sigma f_i x_i' + d^r\Sigma f_i\right]$$

$$= \mu_r + {}^rC_1\mu_{r-1}d + {}^rC_2\mu_{r-2}d^2 + \ldots\ldots + {}^rC_2\mu_2 d^{r-2} + d^r.$$

In particular,

$$\mu_2' = \mu_2 + d^2,\ \mu_3' = \mu_3 + 3d\mu_2 + d^3,\ \mu_4' = \mu_4 + 4d\mu_3 + 6d^2\mu_2 + d^4.$$

Ex. Show that the sum of the coefficients of the various terms of the right member of (3·15) is equal to zero.

3·3 2. Effect of Change of Origin and Scale on Moments. Let us introduce a new variate u related to x by

$$u = \frac{x-a}{h} \text{ so that } x-a = hu,\ \bar{x}-a = h\bar{u} \text{ and } x-\bar{x} = h(u-\bar{u}).$$

$$\therefore \quad \mu_r' = \frac{1}{N}\sum_i f_i(x_i-a)^r = h^r.\ \frac{1}{N}\sum_i f_i u_i^r$$

and

$$\mu_r = \frac{1}{N}\sum_i f_i(x_i-\bar{x})^r = h^r.\ \frac{1}{N}\sum_i f_i(u_i-\bar{u})^r.$$

Thus the *rth moment of the variate x is h^r times the corresponding moment of the variate u.*

In particular, taking $a=\bar{x}$ and $h=\sigma_x$ we find that the distribution of u has zero mean and unit variance and $(x-\bar{x})'\sigma_x$ is said to be the form of the variate in *standard measure.*

3·3·3. Sheppard's corrections to moments of grouped frequency distributions. In grouped frequency distributions, we had assumed that the frequencies are concentrated at the mid-points of the class-intervals. This assumption is, however, only an approximation. W.F. Sheppard proved that if

(*a*) the frequency distribution is continuous and,

(*b*) the frequency tapers off to zero in both directions, then the corrections to the various moments are :

$$\left.\begin{aligned}
&\mu_1' \text{ (corrected)} = \mu_1', \\
&\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12}, \\
&\mu_3 \text{ (corrected)} = \mu_3, \\
&\mu_4 \text{ (corrected)} = \mu_4 - \tfrac{1}{2}h^2\mu_2 + \frac{7}{240}h^4
\end{aligned}\right\} \quad \ldots(3{\cdot}19)$$

where h is the width of the class-intervals and the moments on the right side refer to the moments of the grouped frequency distribution.

Charlier Check.

The following identities

$$\Sigma f(x+1) = \Sigma fx + N$$

$$\Sigma f(x+1)^2 = \Sigma fx^2 + 2\Sigma fx + N$$

$$\Sigma f(x+1)^3 = \Sigma fx^3 + 3\Sigma fx^2 + 3\Sigma fx + N$$

$$\Sigma f(x+1)^4 = \Sigma fx^4 + 4\Sigma fx^3 + 6\Sigma fx^2 + 4\Sigma fx + N$$

are found useful in checking the values of Σfx, Σfx^2, etc. These identities are named after their discoverer.

3·4. Symmetrical and Skew Distributions. A distribution is said to be symmetrical when the frequencies are symmetrically distributed about the mean, that is, when the values equidistant from the mean have equal frequencies. For example, the following distribution is symmetrical about its mean 5.

x :	1	2	3	4	5	6	7	8	9
f :	3	4	6	9	10	9	6	4	3

Also for a symmetrical distribution, the mean, mode and median coincide. Further in the case of such a distribution, median lies half-way between the two quartiles, *i.e.*, $Q_3 - \text{median} = \text{median} - Q_1$.

Skewness means lack in symmetry. In other words, for an unsymmetric distribution, it is the departure from symmetry. Skewness indicates whether the curve is turned more to one side than to the other, that is whether the curve has a longer tail on one side. Skewness is said to be *positive* if the curve is more elongated to the right side, that is to say, if the mean of the distribution is to the right of or greater than the mode. In the reverse case, skewness is said to be *negative*.

For distributions of moderate skewness, there is an empirical relationship between the mean, the mode and the median expressed by the equation :

$$\text{mean} - \text{mode} = 3\ (\text{mean} - \text{median}) \quad \ldots(3{\cdot}20)$$

The object of measuring skewness is to estimate the direction in which and also the extent to which the distribution is distorted from the symmetrical distribution. Karl Pearson's coefficient of skewness is given by the formula :

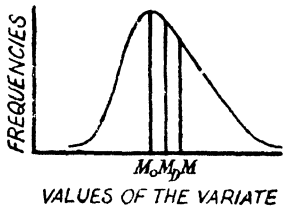


Fig. 12. Asymmetrical distribution.

$$\text{Coeff. of skewness} = \frac{\text{mean} - \text{mode}}{\text{standard deviation}} = \frac{3\ (\text{mean} - \text{median})}{\text{standard deviation}} \quad \ldots(3{\cdot}21)$$

This is clearly a pure number since both the numerator and the denominator have the same dimensions. Also it is zero for a symmetrical distribution.

However, if the standard deviation is not known, a rough measure of coefficient of skewness is obtained by

$$\text{Coeff. of skewness} = \frac{(Q_3 - \text{median}) - (\text{median} - Q_1)}{(Q_3 - \text{median}) + (\text{median} - Q_1)}$$

$$= \frac{Q_3 + Q_1 - 2\ \text{median}}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \quad \ldots(3{\cdot}22)$$

This is evidently a pure number and is zero for a symmetrical distribution.

$$\text{In terms of } \beta_1 \text{ and } \beta_2, \text{ coeff. of skewness} = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}.$$

For a symmetrical distribution, $\beta_1 = 0$.

Example 3·6. *Find out a coefficient of dispersion for quartile deviation and a measure of skewness from the following table giving the wages of 230 persons* :

Wages (in Rs.)	*No. of persons*	*Wages (in Rs.)*	*No. of persons*
70–80	12	110–120	50
80–90	18	120–130	45
90–100	35	130–140	20
100–110	42	140–150	8

Here $N/2=115$ and this lies in the class 110—120

Hence median, $Q_3=110+\frac{115-107}{50}\times 10=$Rs. 11·16

Similarly, $Q_1=90+\frac{57\,5-30}{35}\times 10=$Rs. 97·88

and $Q_3=120+\frac{172\cdot5-157}{45}\times 10=$Rs. 123·44

By example 3·3, page 37,

Mean=Rs. 110·4,

Standard deviation =Rs. 17·3

Coefficient of dispersion for quartile deviation

$$=\frac{Q_3-Q_1}{Q_3+Q_1}=\frac{25\cdot56}{221\cdot32}=0\cdot1$$

Measure of skewness $=\frac{\text{mean}-\text{mode}}{\text{standard deviation}}=\frac{110\cdot4-116\,5}{17\cdot3}$

$=-0\cdot33$

Measure of skewness $=\frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}=\frac{-1\cdot88}{25\cdot56}=-0\cdot098.$

3·5. Pearson's β and γ-Coefficients. Karl Pearson defined the four important coefficients as follows :

$$\left.\begin{aligned}\beta_1&=\frac{\mu_3^2}{\mu_2^3}, & \gamma_1&=\sqrt{\beta_1}\\ \beta_2&=\frac{\mu_4}{\mu_2^2} & \gamma_2&=\beta_2-3\end{aligned}\right\} \quad \ldots(3\cdot23)$$

The values of the four coefficients depend upon the first four moments.

It is to be noted that these four coefficients are all pure numbers, that is to say they are independent of origin and scale of the measurement of the variate [since μ_n has the dimension (variate)^n hence the above quotients have dimension zero].

In the case of a symmetrical distribution, all the moments of odd order about the mean are equal to zero, since the terms of the sum in $\mu_{2r+1}=\frac{1}{N}\sum_i f_i(x_i-\bar{x})^{2r+1}$ cancel in pairs and hence β_1 is also equal to zero. Thus β_1 gives a *measure of departure from symmetry*, that is to say, *of skewness.*

The term *Kurtosis*, due to Karl Pearson, is described as :

'Given two frequency distributions which have the same variability as measured by the standard deviation, they may be relatively more or less flat topped than the 'normal curve' (discussed in Chapter 8). A frequency curve may, in other words, be symmetrical but it may fail to be equally flat topped with the normal curve. The relative flatness of top is called the *kurtosis* and can be measured by β_2. It will be proved in § 8·1·2 that for a normal distribution, $\beta_2 = 3$. Hence for any distribution, the quantity $\beta_2 - 3$ is called *excess of kurtosis* or simply its *excess*. Curves with values of β_2 less than 3 are called *Platykurtic* while those with values of β_2 greater than 3 are called *Leptokurtic*.

3·6. Factorial Moments. Absolute Moments. The *factorial moment* of order r about the origin ($x=0$) of a distribution is defined as

$$\mu'_{(r)} = \frac{1}{N} \sum_i f_i x_i^{(r)}, \qquad N = \sum_i f_i \qquad \ldots(3\cdot24)$$

where $x^{(r)} = x(x-1)(x-2)\ldots(x-r+1)$.

The *absolute moment* of order r about the origin ($x=0$) of a distribution is defined as

$$\nu_r = \frac{1}{N} \sum_i f_i \mid x_i \mid^r, \qquad N = \sum_i f_i \qquad \ldots(3\cdot25)$$

so that the mean deviation about any point is the first absolute moment about that point.

Ex. (*i*) Prove the following inequality for the absolute moments defined by (3·25) :

$$\nu_r^{2r} \leqslant \nu^r{}_{r-1}\, \nu^r{}_{r+1}$$

(*ii*) Replacing r successively by 1, 2,....., r and multiplying all the inequalities thus formed, show that

$$\nu_r^{\frac{1}{r}} \leqslant \nu_{r+1}^{\frac{1}{r+1}}, \; r=1, 2,\ldots\ldots, \quad (M.A., \textit{Poona, 1959})$$

[**Hint.** (*i*) Use the fact that $\frac{1}{N} \sum_i f_i \left[a \mid x_i \mid^{\frac{r+1}{2}} + b \mid x_i \mid^{\frac{r-1}{2}} \right]^2$ is non-negative. In other words, the quadratic form in a and b *viz.*, $\nu_{r-1}a^2 + 2\nu_r ab + \nu_{r+1}b^2$ is non-negative, the condition for which is $\nu_{r-1}\nu_{r+1} - \nu^2{}_r \geqslant 0$.]

3·7. Moments for a Bivariate Distribution. Let X and Y be two variates, *e.g.*, ages of bride and bridegroom ; heights and weights of students, quantity of manure applied and the corresponding yield obtained. Further, let f_i be the frequency of the pair (x_i, y_i), $i=1, 2,\ldots, n$. Then we define the moment μ_{rs}' about the

origin (0, 0) of order r in and s in y by

$$\mu_{rs}' = \frac{1}{N} \sum_{i=1}^{n} f_i x_i^r y_i^s. \qquad \ldots(3{\cdot}26)$$

where $$N = \sum_{i=1}^{n} f_i$$

In particular,

$$\mu_{10}' = \frac{1}{N} \sum_{i=1}^{n} f_i x_i = \bar{x},$$

$$\mu_{01}' = \frac{1}{N} \sum_{i=1}^{n} f_i y_i = \bar{y},$$

$$\mu_{11}' = \frac{1}{N} \sum_{i=1}^{n} f_i x_i y_i$$

Similarly, we define moment μ_{rs} about the mean $(\bar{x}, \bar{y})$ by

$$\mu_{rs} = \frac{1}{N} \sum_{i=1}^{n} f_i (x_i - \bar{x})^r (y_i - \bar{y})^s \qquad \ldots(3{\cdot}27)$$

In particular,

$$\mu_{10} = 0, \ \mu_{01} = 0$$

$$\mu_{11} = \frac{1}{N} \sum_{i=1}^{n} f_i (x_i - \bar{x}) (y_i - \bar{y}) \qquad \ldots(3{\cdot}28)$$

$$= \frac{1}{N} \sum_{i=1}^{n} f_i x_i y_i - \bar{x} \ \frac{1}{N} \Sigma f_i y_i - \bar{y} \frac{1}{N} \Sigma f_i x_i + \bar{x}\bar{y}$$

$$= \mu_{11}' - \bar{x}\bar{y} \qquad \ldots(3{\cdot}29)$$

This formula is very important in the study of *correlation* between two variates. μ_{11} is often called the *covariance* between the variates x and y and is denoted by cov (x, y).

Also $$\mu_{20} = \sigma_x^2 = \frac{1}{N} \sum_{i=1}^{n} f_i (x_i - \bar{x})^2 = \mu_{20}' - \bar{x}^2 \qquad \ldots(3{\cdot}30)$$

$$\mu_{02} = \sigma_y^2 = \frac{1}{N} \sum_{i=1}^{n} f_i \ (y_i - \bar{y})^2 = \mu_{02}' - \bar{y}^2 \qquad \ldots(3{\cdot}31)$$

EXERCISES 3

1. (*a*) What do you understand by dispersion ? How is it measured ? What are the different measures of dispersion ? Explain their uses.

(*b*) How does skewness differ from dispersion ? What is the object of measuring skewness ?

(*c*) Write short notes on :

(*a*) Variance, (*b*) Quartile deviation. (*c*) Coefficient of variation.

(*d*) Skewness and kurtosis. (*e*) Positive and negative skewness.

(*d*) Show that for any frequency distribution :

(*i*) kurtosis is greater than unity.

(*ii*) coefficient of skewness is less than 1 numerically

(*B.A. Hons., Delhi 1956 ; I.A.S. 1953, 54, 56*)

2. Calculate the arithmetic mean and the standard deviation of the following values of the world's annual gold output (in millions of pounds) for 20 different years :

94	95	96	93	87	79	73	69	68	67
78	82	83	89	95	103	108	117	130	97

(*M. Sc. Agra 1959*)

[**Ans.** Mean=90·15, standard deviation=15·99 approx.]

3. The following table gives the monthly wages of workers in a factory. Compute the standard deviation, Quartile deviation, coefficient of variation and skewness.

Monthly wages (in Rs.)	*No. of workers*	*Monthly wages (in Rs.)*	*No. of workers*
12·5—17·5	2	37·5—	4
17·5—	22	42 5—	6
22·5—	19	47·5—	1
27·5—	14	52 5—57·5	1
32·5—	3		

[**Ans.** S.D.=Rs. 8·85, Q.D.=Rs. 5·145, C.V.=31·8. Skewness=0·7]

4. Find the mean and s.d. of the following series :

Expenditure	*No. of students*
Below Rs. 5	6
,, 10	16
,, 15	28
,, 20	38
,, 25	46

[**Ans.** Mean=Rs. 12·94, S.D.=Rs. 2]

5. The value of the arithmetic mean and the standard deviation of the following frequency distribution of a continuous variable derived by using both arbitrary working origin and scale are 135·3 lb. and 9·6 lb. respectively. Determine the actual class-intervals.

u :	−4	−3	−2	−1	0	1	2	3	Total
f:	2	5	8	18	22	13	8	4	80

(*I.A.S., 1952*)

[**Hint.** Use $\bar{x}=a+h\bar{u}$, $\sigma_x=h\sigma_u$]

6. For a frequency distribution of marks in History of 200 candidates (grouped in intervals 0—5, 5—10, ...) the mean and standard deviation (s.d.) were found to be 40 and 15. Later it was discovered that the score 43 was misread as 53 in obtaining the frequency distribution. Find the corrected mean and s.d. corresponding to the corrected frequency distribution. (*I.A.S., 1957*)

[**Ans.** Mean 39·95. Standard deviation=14·974 approx.]

7. A computor while calculating the mean and the standard deviation of 25 observations, obtained the following values :

mean=56 inches, standard deviation=2 inches.

It was later discovered at the time of checking that he had wrongly copied down an observation as 64. What is the mean and s.d. if correct value is omitted ? [Ans. Mean=55·67″. S.D.=1·18″ approx.]

8. An analysis of the monthly wages paid to workers in two firms *A* and *B*, belonging to the same industry, gives the following results :

	Firm A	*Firm B*
Number of wage earners	586	648
Average monthly wages	Rs. 52·5	Rs. 47·5
Variance of the distribution of wage	100	121

(*a*) Which firm, *A* or *B*, pays out the larger amount as monthly wages ?

(*b*) In which firm, *A* or *B* is there greater variability in individual wages ?

(*c*) What are the measures of (*i*) average monthly wages and (*ii*) the variability in individual wages of all the workers in the two firms, *A* and *B* taken together ? (*B. A. Punjab 1961* ; *I. A. S., 1951*)

[Ans. (*a*) Firm *B* pays a larger amount as monthly wages.

(*b*) There is greater variability in individual wages in firm *B*

(*d*) (*i*) Combined arithmetic mean=Rs. 49·9,

(*ii*) Combined standard deviation of the two series=Rs. 10·5.

9. The scores of two golfers for 10 rounds each are :

A : 58, 59, 60, 54, 65, 66, 52, 75, 69, 52

B : 84, 56, 92, 65, 86, 78, 44, 54, 78, 68

Which may be regarded as the more consistent player ?

(*B.A. Punjab, 1954*)

[Hint. Compare the coefficients of variation.]

[Ans. *A* is more consistent player]

10. Goals scored by two teams *A* and *B* in a football season were as follows :

No. of goals scored in a match	*Number of matches*	
	A	*B*
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

By calculating the coefficient of variation in each case, find which team be considered more consistent. (*I.A.S., 1954*)

[Ans. Team *B*]

11. (*a*) The means of two samples of sizes 50 and 100 respectively are 54·4 and 50·3 and the standard deviations are 8 and 7. Obtain the mean and standard deviation of the sample of size 150 obtained by combining the two samples.

(*B. Sc. Delhi, 1966* ; *B.A. Punjab 1953* ; *B.A. & B. Sc. Lucknow, 1947*)

[Ans. Mean=51·57. S.D.=7·5 approx.]

(*b*) The first of two samples has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15·6 and standard deviation $\sqrt{13{\cdot}44}$, find the standard deviation of the second group.

(*B. Sc. Delhi 1960, M A., Econ., Delhi 1959*)

[Ans. 4]

12. (*a*) Show that, if the variable takes the values 0, 1, 2,....., n with frequencies proportional to the binomial coefficients 1, ${}^nC_1$, ${}^nC_2$..., ${}^nC_n$ respectively then the mean of the distribution is $\frac{1}{2}n$, the mean square deviation about origin is $\frac{1}{4}n(n+1)$ and the variance is $\frac{1}{4}n$.

(*b*) Show that, if the variable takes the values 0, 1, 2, ..., n with frequencies given by the terms of the binomial series q^n, ${}^nC_1q^{n-1}p$, ${}^nC_2q^{n-2}p^2$, ..., p^n where $p+q=1$, then the mean square deviation is n^2p^2+npq and the variance is npq.

13. Show that, if deviations are small compared with the mean M of a variable X, so that $\left(\frac{x}{M}\right)^3$ and higher powers of $\frac{x}{M}$ may be neglected, and σ is the s.d. of X, (*B.A. Delhi, 1971*)

(*a*) (*i*) $G=M\left(1-\frac{1}{2}\frac{\sigma^2}{M^2}\right)$ (*ii*) $M^2-G^2=\sigma^2$ (*B. Sc., Agra 1955*)

(*ii*) $H=M\left(1-\frac{\sigma^2}{M^2}\right)$

(*M. Sc., Agra 1953, 63 ; B.A. Hons., Delhi 1966*)

(*b*) mean $(\sqrt{x})=\sqrt{M}\left(1-\frac{\sigma^2}{8M^2}\right)$ *approx.*

(*c*) $V=\sqrt{\frac{2(M-G)}{M}}$ where V is the coefficient of variation.

14. From a sample of n observations, the arithmetic mean and variance are calculated. It is then found that one of the values, x_1, is in error, and should be replaced by x_1'. Show that the adjustment to the variance to correct this error is

$$\frac{1}{n}(x_1'-x_1)\left[(x_1'+x_1)-\frac{x_1'-x_1+2T}{n}\right]$$

where T is the total of the original results.

(*M.A., Agra 1967 ; B.Sc., Lucknow 1962*)

Hint. Corrected variance $=\frac{1}{n}\{x_1'^2+x_2^2+\ldots+x_n^2\}-\left\{\frac{T-x_1+x_1'}{n}\right\}^2$

15. Find the mean deviation from the mean and the standard deviation of the A.P. a, $a+d$, $a+2d$, ..., $a+nd$ and prove that the latter is greater than the former. (*B.A. Punjab 1956*)

[**Ans.** Mean deviation $=d\times\frac{n(n+1)}{2n+1}$, S.D. $=d\times\sqrt{\frac{n(n+1)}{3}}$]

16. If in a series of measurements we obtain m_1 values of magnitude x_1, m_2 of magnitude x_2, and so on, and if $\bar{x}$ is the mean value of all the measurements, prove that the standard deviation is

$$\sqrt{\frac{\Sigma m_r(k-x_r)^2}{\Sigma m_r}-\delta^2}$$

where $\bar{x}=k+\delta$, and k is any constant. *(M.A., Agra 1967 ; B.Sc. Aligarh)*

Hint. $\sigma^2=\Sigma m_r(x_r-\bar{x})^2/\Sigma m_r$, $\bar{x}=\Sigma m_r x_r/\Sigma m_r$

$=\Sigma m_r(x_r-k+k-\bar{x})^2/\Sigma m_r=$ etc.

17. *(a)* The first three moments of a distribution about the value 2 of the variable are 1, 16 and -40. Show that the mean is 3, the variance 15 and $\mu_3=-86$. Also show that the first three moments about $x=0$ are 3, 24 and 76.

(b) The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Show that the mean is 7, variance 16, $\mu_3=-64$ $\mu_4=162$, $\beta_1=1$ and $\beta_2=0{\cdot}63$. *(B. Sc., Luck., 1966, B. Sc. Kanpur, 1969)*

(c) For a distribution, the mean is 10, variance is 16, γ_1 is $+1$ and β_2 is 4. Find the first four moments about the origin. *(B. Sc. Delhi 1969)*

[**Ans.** $\mu_1'=10$, $\mu_2'=116$, $\mu_3'=1544$ and $\mu_4'=23184$].

18. *(a)* Calculate the first four moments about the mean for the following data. Also calculate β_1 and β_2.

x:	1	2	3	4	5	6	7	8	9
f:	1	6	13	25	30	22	9	5	2

(b) In a certain distribution, the first four moments about the point 4 are $-1{\cdot}5$, 17, -30 and 108. Calculate β_1 and β_2 and state whether the distribution is leptokurtic or platykurtic. *(B.A., Delhi, 1971 ; B.A. & B. Sc., Lucknow 1949)*

[**Ans.** *(a)* $\mu_1'=-0{\cdot}09$, $\mu_2=2{\cdot}49$, $\mu_3=0{\cdot}7$, $\mu_3=18{\cdot}33$ approx., $\beta_1=0{\cdot}03$ approx., $\beta_2=3$ approx.

(b) $\mu_2=14{\cdot}75$, $\mu_3=39{\cdot}75$, $\mu_4=142{\cdot}31$, curve is platykurtic.]

19. Show that, if the class interval of a grouped distribution is less than one-third of the calculated standard deviation (*S.D.*), Sheppard's adjustment makes a difference of less than $\frac{1}{2}\%$ in the estimate of *S.D.*

(B.A. & B.Sc., Lucknow, 1948)

20. Show that for distributions,

(i) $\beta_2 \geqslant 1$. *(M. Sc., Agra, 1953)*

(ii) $\beta_2 > \beta_1$ *(B.A. Delhi 1971 ; B.A. Lucknow, 1964, 66)*

Hint. *(i)* We are to prove $\frac{1}{N}\ \Sigma\ f_i y_i > \left[\frac{1}{N}\ \Sigma\ f_i\ y_i\right]^2$

where $y_i=(x_i-\bar{x})^2$

i.e. $\sigma_y^2 \geqslant 0$.

21. Show the first factorial moments are related to ordinary moments by the equations

$$\mu'_{(1)}=\mu_1'=\bar{x},\ \mu'_{(2)}=\mu'_2-\bar{x}$$

$$\mu'_{(3)}=\mu_3'-3\mu_2'+2\bar{x},\ \mu'_{(4)}=\mu_4'-6\mu_3'+11\mu_2'-6\bar{x}$$

22. Show that, for bivariate distributions, the moments about the mean are invariant for the change of origin but not for the change of scale.

23. For a distribution of 250 heights, calculations standard deviation, β_1 and β_2 were 54 inches, 3 inches, 0 a. tively. It was however discovered on checking that the two it the original data were wrongly written in place of correct values 6 respectively. Calculate the correct frequency constants. (*B.A. Hons.*

[**Ans.** Mean remains unaltered, *s.d.*=2·968, β_1=·004,

24. In calculating the moments of a frequency distribution based observations, the following results were obtained :

mean=9, variance=19, β_1=0·7, β_2 is $+ve$, β_2=4

But later on it was noticed that one observation 12 was read as 21. Obtain the correct value of the first four central moments. (*I.A.S. 1963*; *I.S.I. 1965*)

[**Ans.** Corrected mean=8·91, μ_2=17·64, μ_3=57·05, μ_4=1257·15, β_1=0·54, β_2=4·04]

25. If r be the range and $S=\left[\frac{1}{n-1}\Sigma(x_i-\bar{x})^2\right]^{\frac{1}{2}}$ be the standard deviation of a set of observations $x_1, x_2, \ldots\ldots, x_n$, then

$$S \leqslant \frac{r}{2}\cdot\left(\frac{n}{n-1}\right)^{\frac{1}{2}}$$ (*B.A. Hons. Delhi, 1967*)

26. Prove that

(*a*) $\mu_r' = \sum_{j=0}^{r} {}^{r}C_{j}\, \mu'_{r-j}\, \mu_1'^{j}$

(*b*) $\mu_r = \sum_{j=0}^{r} {}^{r}C_{j}\, \mu'(-\mu_1')^{j}$ (*B. Sc. Sardar Patel 1967*)

Hint. (*a*) $\mu_r'(a) = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i-a)^r = \frac{1}{N}\sum_{1}^{n} f_i\left[(x_i-b)-(a-b)\right]^r$

$$= \frac{1}{N}\Sigma f_i\,[(x_i-b)-c]^r$$

where $c=a-b$.

27. (*a*) If δ_r is the rth absolute moment about zero, use the mean value of

$$\left[u\,|x|^{\frac{r-1}{2}} + v\,|x|^{\frac{r+1}{2}}\right]^2$$

to show that $\delta_r^{2r} < \delta_{r-1}^{r}\,\delta_{r+1}^{r}$

From this, derive the following inequalities :

(*i*) $\delta_r^{r+1} \leqslant \delta_{r+1}^{r}$ (*ii*) $\delta_r^{1/r} \leqslant \delta_{r+1}^{1/r+1}$ (*I.A.S. 1966*)

(*b*) If λ and μ are positive integers, $\lambda<\mu$ and if β_k stands for the absolute moment of a distribution, show that

$$\beta_\lambda^{1/\lambda} \leqslant \beta_u^{1/\mu}$$ (*I.A.S. 1962*)

(*c*) Prove the following relation for Pearson's β-coefficients of skewness and kurtosis : $\beta_2 > \beta_1 + 1$ (*B.Sc. Delhi 1971, B. Sc. Hons. Gauhati, 1968*)

4

Theory of Probability

4·1. Introduction

In the theory of probability we deal with *events* which are the outcomes of an experiment repeated under essentially the homogeneous and similar conditions. An experiment is defined as an action which we conceive, do or intend to do. The experiments are of two types :

(*a*) **Deterministic experiment.** Here chance does not play any part or role in the outcome of an experiment. In other words, conditions completely determine the outcome. We can completely predict the outcome.

Thus for example (*i*) The pressure p and volume v of a gas are known to be related by an equation of the form

$$pv^k = c$$

from which we can determine p as soon as we know v ; k and c being constants.

Hence the pressure p is *known* as soon as the volume v of gas is known, c being a constant.

(*ii*) In the Logistic curve

$$y = \frac{a}{1+be^{x}},$$

y is completely *determined* as soon as x is known, a and b being constants.

(*iii*) The distance of a particle moving in a st. line is *known* as soon the time t, the accelerating 'a' and the initial velocity u is known since we have the relation

$$s = ut + \tfrac{1}{2}at^2.$$

(*b*) **Non-deterministic experiment.** Here the chance plays an importont role in the outcome of an experiment. In other

words, the conditions do not determine completely the outcomes and the phenomena is non-deterministic or probabilistic.

Thus for example :

(*i*) In rolling an honest or unbiased die, one could predict with certainty which face may turn up.

(*ii*) In drawing a card from a well-shuffled pack of cards, one can not say with certainty which card will be drawn.

(*iii*) One cannot guess whether it will be a rainy day or a sunny day tomorrow.

4·2. Definition of Probability

4·2·1. Explanation of various terms. Before definition of probability is given we explain certain terms which are used in its definition.

Trial and Event. Let the real or conceptual experiment be repeated under essentially the same condition and let it result in any of the possible outcomes. Then the experiment is called a trial and the possible outcomes are known as the events or cases. Thus tossing of a coin is a trial and the turning up of a head or a tail is known as the event.

Equally likely cases. The cases (or possibilities) are said to be *equally likely* when we have no reason to expect any one rather than the other. For example, as the result of drawing a card from a well-shuffled pack, any card may appear in the draw so that the 52 different cases are equally likely.

Mutually exclusive cases. Cases are called mutually exclusive (m. e.) or incompatible if the occurrence of one of them precludes the occurrence of all the others. On the contrary, cases are compatible if it is possible for them to happen simultaneously. For instance, in the rolling of two dice, the cases of the face marked 5 appearing on one dice and face 5 appearing on the other are compatible since in rolling two dice it is possible to get face 5 on each.

Exhaustive Cases. Cases are said to be exhaustive when they include all possible cases. In throwing a coin, the turning up of a head and of a tail are exhaustive cases.

Favourable cases. The cases which entail the happening of an event are said to be *favourable* to the event. For example, in the tossing of a dice, the number of cases favourable to the appearance of a multiple of 3 are two, *viz.*, 3 and 6.

We shall now consider two definitions of probability :

(*i*) Mathematical or *a priori* or classical

(*ii*) Statistical or empirical.

4·2·2. Mathematical (or a priori) definition of probability. *If there are* n **exhaustive, mutually exclusive** *and* **equally likely** *cases and* m *of them are favourable to an event* A, *the probability of the happening of* A *is defined as the ratio* m/n. *Laplace's definition*].

Thus denoting the probability (or chance) of the happening of an event A by p, we have

$$p = \frac{m}{n} \qquad \ldots(4\cdot1)$$

This gives the numerical measure of probability. Clearly, p is a positive number not greater than unity so that $0 \leqslant p \leqslant 1$.

Since the number of cases in which the event A will not happen is $n-m$, the probability q that the event A will not happen is given by

$$q = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p, \qquad \ldots(4\cdot2)$$

so that $p+q=1$, $0 \leqslant q \leqslant 1$.

An event A is *certain* to happen if all the cases are favourable to it and then the probability of its happening is unity, while if it is *certain not* to happen, its probability is zero.

We solve two examples to illustrate this definition of probability.

Example 4·1. *An urn contains 9 balls two of which are red, three blue and four black. Three balls are drawn from the urn at random, that is every ball has an equal chance of being included in the three. What is the chance that*

(*i*) *the three balls are of different colours* ?

(*ii*) *two balls are of the same colour and third of different* ?

(*iii*) *the balls are of the same colour* ?

(*i*) The number of ways in which three balls can be drawn from 9 balls is ${}^9C_3$ or 84, and these are equally likely, exhaustive and mutually exclusive cases.

A red ball can be drawn in 2 ways, a blue in 3 and a black in 4 ways, so that the total number of ways of drawing three differently coloured balls is $2 \times 3 \times 4 = 24$.

Therefore the required chance is $\frac{2}{8}\frac{4}{4}$ or $\frac{2}{7}$.

(*ii*) Two blue balls can be drawn in ${}^3C_2$ ways and then a red or black ball in 6 ways so that the two blue balls and another coloured ball can be drawn in $6 \times {}^3C_2$ or 18 ways. Two black balls can be

drawn in ${}^4C_2$ ways and another coloured ball in 5 ways so that two black and another coloured ball can be drawn in $5 \times {}^4C_2$ or 30 ways. Similarly, the number of ways in which two red balls and another coloured ball can be drawn in $7 \times \overset{2}{C}_2$ or 7 ways. The total number of ways in which two balls of the same colour and a third of a different colour may be drawn is $18+30+7=55$.

Therefore the required chance is $\frac{55}{84}$.

(*iii*) Three blue balls can be drawn in 1 way and 3 black balls in $\overset{4}{C}_3$ or 4 ways so that the total number of ways of drawing three balls of the same colour is 5 and the chance is $\frac{5}{84}$.

Example 4·2. *In a hand at whist, what is the chance that the 4 kings are held by a specified player ?*

In a hand at whist, the specified player must have 13 cards. including 4 kings. Thus he should have 9 cards of other type or types. 4 kings can be drawn in ${}^4C_4$ ways and 9 cards can be drawn out of the remaining 48 cards in ${}^{48}C_9$ ways. The total number of ways in which 4 kings and 9 cards of other type or types can be drawn is ${}^4C_4 \times {}^{48}C_9$ from a full pack of 52 cards. Any 13 cards can be drawn in ${}^{52}C_{13}$ ways.

Therefore the requird chance $= \frac{{}^4C_4 \times {}^{48}C_9}{{}^{52}C_{13}} = \frac{11}{4165}$.

Objections to a priori or classical or mathematical definition of probability are :

1. It involves the word equally likely or equally probable. What happens when cases are not equally likely ? For example, if we have a coin known to be biased in favour of heads, that is a head is more likely to appear than a tail than what is the probability of a head ? The classical definition leaves us completely helpless here.

2. The classical definition fails when the total number of possible outcomes is infinite. It also restricts value of probability to rational numbers only.

3. It may not be practically possible to enumerate all the equally likely cases.

Example 4·3. *Each coefficient in the equation* $ax^2+bx+c=0$ *is determined by throwing an ordinary die. Find the probability that the equation will have real roots.* (*M.Sc., Bombay, 1965*)

The equation $ax^2+bx+c=0$ will have its root real if,

$$b^2 \geqslant 4ac.$$

Now each coefficient of the given equation is determined by throwing an ordinary die so that each of the coefficients can take the value 1 to 6.

Hence total number of possible outcomes or equally likely cases $=6\times6\times6=216$.

The number of favourable cases (satisfying $b^2 \geqslant 4ac$) are obtainable as follows :

ac	a	c	$4ac$	$b(b^2 \geqslant 4ac)$	*No. of cases*
1	1	1	4	2,3,4,5,6	5
2	1 2	2 1	8	3,4,5,6	$2\times4=8$
3	1 3	3 1	12	4,5,6	$2\times3=6$
4	1 4 2	4 1 2	16	4,5,6	$3\times3=9$
5	1 5	5 1	20	5,6	$2\times2=4$
6	1 6 2 3	6 1 3 2	24	5,6	$4\times2=8$
8	2 4	4 2	32	6	$2\times1=2$
9	3	3	36	6	1
				Total	43

Clearly ac cannot take the values 7, 10, 11, 12 so as to satisfy $b^2>4ac$ and since none of the coefficients can be greater than 6.

Hence the required probability $=\frac{43}{216}$.

4·2·3. Von Mises's Statistical (or Empirical) definition of probability. *If trials be repeated a great number of times under essentially the same conditions then the limit of the ratio of the number of times that an event happens to the total number of trials as the number of trials increases indefinitely is called the probability of the happening of that event. It is assumed that the ratio approaches a finite and a unique limit.*

Symbolically, if m be the number of times in which the event A happens in a series of n trials, then the probability, p, of the happening of this event is given by

$$P(A)=p=\lim_{n\to\infty}\frac{m}{n}, \qquad \ldots(4\cdot3)$$

provided that the limit is finite and unique.

Note. The two definitions of probability are apparently different. In the 'a priori' definition, it is the relative frequency of favourable cases to the total number of cases while in empirical definition, it is the limit of the relative frequency of the happening of the event under consideration. However, in Sec. 7·4 we shall see how the two definitions can be reconciled.

4·3. Notations. Let $A_1, A_2, \ldots\ldots, A_n$, be n events, then in dealing with the theory of probability, we shall find it convenient to use the following notations :

$P(A_i)$ denotes the probability for the event A_i to materialize.

$P(\bar{A}_i)$ denotes the probability for the event A_i ***not*** to materialize.

$\bar{A}_i$ will be called the event opposite to A_i. Clearly A_i and $\bar{A}_i$ are exhaustive and mutually exclusive and (4·2) gives

$$P(A_i)+P(\bar{A}_i)=1.$$

$P(A_i+A_j)$ denotes the probability of occurrence of *at least* one of the events A_i, A_j.

Similarly, $P(A_1+A_2+\ldots+A_n)$ denotes the probability of occurrence of at least one of the n events $A_1, A_2, \ldots, A_n$.

$P(A_i\,A_j)$ denotes the probability of the occurrence of both the events A_i and A_j.

A_iA_j will be called a *compound* event. $P(A_i \mid A_j)$ denotes the conditional probability for the event A_i to happen when *it is known* that the event A_j has already happened.

$P(A_i \mid A_jA_k)$ denotes the conditional probability for the event A_i to happen when *it is known* that A_j and A_k have already happened.

Two events are said to be *independent* (in probability sense) if the probability of happening of one does not depend on the happening or not happening of the other. For example, if from a bag containing a white and b black balls, we draw a ball and replace it before the next draw is made, the change of the ball being white or black in the second draw is independent of the result of the first draw, that is of the colour of the first ball.

4·4. Additive Law of Probability or Theorem of Total Probability. *The probability that one of several* **mutually exclusive** *events $A_1, A_2, \ldots, A_n$, will happen is the sum of the probabilities of the separate events.* In symbols,

$$P(A_1+A_2+\ldots+A_n)=P(A_1)+P(A_2)+\ldots+P(A_n)$$

Proof. Let N be the total number of exhautive, mutually exclusive and likely cases and out of these let $m_1, m_2, \ldots, m_n$ be respectively favourable to the events $A_1, A_2, \ldots, A_n$. Since the n events are mutually exclusive these $m_1, m_2, \ldots, m_n$ cases are quite distinct and non-overlapping. Therefore the total number of cases which are favourable to either A_1 or A_2 or ..... or A_n is $m_1+m_2+\ldots+m_n$.

Hence
$$P(A_1+A_2+\ldots+A_n)=\frac{m_1+m_2+\ldots+m_n}{N}$$
$$=\frac{m_1}{N}+\frac{m_2}{N}+\ldots+\frac{m_n}{N}$$
$$=P(A_1)+P(A_2)+\ldots+P(A_n)$$
$$=\sum_{i=1}^{n} P(A_i)$$

which proves the theorem.

Cor. 1. If an event A consists of n mutually exclusive forms $A_1, A_2, \ldots, A_n$ so that A happens wherever any one of these events happens and conversely, then

$$A=A_1+A_2+\ldots+A_n$$

and
$$P(A)=P(A_1)+P(A_2)+\ldots+P(A_n)$$

Cor. 2. If n mutually exclusive events $A_1, A_2, \ldots, A_n$ are exhaustive also, so that the probability of *at least* one of the n events to materialize is a certainty, then

$$P(A_1+A_2+\ldots+A_n)=1$$

or
$$P(A_1)+P(A_2)+\ldots+P(A_n)=1.$$

Def. Two events A and B are said to be *equivalent* $(A=B)$, if the happening or not happening of either implies the happening or not happening of the other.

Ex. 1. Prove that

(*i*) $A+B\equiv B+A$ (*ii*) $AB\equiv BA$ (*iii*) $AB\neq\bar{A}\bar{B}$

(*ii*) $\overline{A+B}\neq\bar{A}+\bar{B}$ (*iv*) $\overline{A_1+A_2+\ldots+A_n}\equiv\bar{A}_1\bar{A}_2\cdots\bar{A}_n$

(*iii*) $A+A\equiv AA\equiv A$ (*iii*) $\bar{A}+\bar{B}\equiv\bar{A}\bar{B}$

Ex. 2. Prove that

(*i*) $P(A\bar{A})=0$

(*ii*) $P(A \mid B)=P(A)$ and $P(B \mid A)=P(B)$, if A and B are independent events.

(*iii*) $P(AB)=0$ if A and B are mutually exclusive. ³),

Ex. 3. Let A, B, C be three arbitrary events. Find expressions for the following events :

(*i*) Only A occurs ; (*ii*) Both A and B, but not C, occur ; (*iii*) All three events occur ; (*iv*) At least one event occurs ; (*v*) At least two events occur ; (*vi*) One and no more occurs ; (*vii*) Two and no more occurs ; (*viii*) None occurs ; (*ix*) Not more than two occurs :

[**Ans.** (*i*) $A\bar{B}\bar{C}$; (*ii*) $AB\bar{C}$; (*iii*) ABC ; (*iv*) $A+B+C$; (*v*) $AB+BC+CA$; (*vi*) $ABC+\bar{A}BC+A\bar{B}C+AB\bar{C}$; (*vii*) $AB\bar{C}+BC\bar{A}+CA\bar{B}$; (*viii*) $\bar{A}\,\bar{B}\,\bar{C}$; (*ix*) $\bar{A}\,\bar{B}\,C$)]

Ex. 4. Use the theorem of total probability to show that

(*i*) $P(A)=P(AB)+P(A\bar{B})$

(*ii*) $P(A_1+A_2+\cdots+A_n)+P(\bar{A}_1\bar{A}_2\ldots\ldots\bar{A}_n)=1$

(*iii*) $P(AB)+P(\bar{A}B)+P(A\bar{B})+P(\bar{A}\bar{B})=1$

(*iv*) $P(A+B)=P(A\bar{B})+P(\bar{A}B)+P(BA)$.

4·4.1. Generalization of the Theorem of Total Probability.

If the events A_1, A_2,.....,A_n, are not mutually exclusive, then the probability of the happening of at least one of the n events is given by $P(A_1+A_2+\cdots+A_n)$

$$=\sum_{i=1}^{n}P(A_i)-\sum_{\substack{i,j=1\\ i\neq j}}^{n}P(A_iA_j)+\ldots\ldots+(-1)^{n-1}P(A_1A_2\ldots A_n)$$

Consider first two events A and B only.

Since AB and $A\bar{B}$ are two exhaustive and mutually exclusive forms in which A can occur, we have

$$P(A)=P(AB)+P(A\bar{B})$$

$$P(B)=P(BA)+P(B\bar{A})$$

$$=P(AB)+P(\bar{A}B),$$

since the order of events is immaterial to the probability of the compound event.

Adding, we get

$$P(A)+P(B)=P(AB)+[P(\bar{A}B)+P(\bar{A}\bar{B})+P(AB)]$$

By the theorem of total probability the expression within the square brackets represents the probability $P(A+B)$ of the occurrence of at least one of the events A and B.

Hence $$P(A+B)=P(A)+P(B)-P(AB) \qquad \ldots(4\cdot4)$$

Let B mean the occurrence of at least one of the events A_1 and A_2, then (4·4) gives

$$P(A+A_1+A_2)=P(A)+P(A_1)+P(A_2)-P(A_1A_2)$$
$$-P(AA_1)-P(AA_2)+P(AA_1A_2).$$

Rearranging and modifying the notations slightly, we have

$$P(A_1+A_2+A_3)=P(A_1)+P(A_2)+P(A_3)-P(A_1A_2)-P(A_2A_3)$$
$$-P(A_3A_1)+P(A_1A_2A_3)$$
$$=\sum_{i=1}^{3} P(A_i)-\sum_{\substack{i,j=1\\ i\neq j}}^{3} P(A_iA_j)+P(A_1A_2A_3). \qquad \ldots(4\cdot5)$$

The general law for n events, which may be proved by induction, is

$$P(A_1+A_2+\ldots+A_n)=\sum_{i=1}^{n} P(A_i)-\sum_{\substack{i,j=1\\ i\neq 1}}^{n} P(A_iA_j)+\sum_{\substack{i,j,k=1\\ i\neq j\neq k}}^{n} P(A_iA_jA_k)$$
$$-\ldots\ldots+(-1)^{n-1}P(A_1A_2\ldots\ldots A_n) \qquad \ldots(4\cdot6)$$

where the second sum is over all combinations of the numbers 1, 2, ...,n taken two at a time, the third is over all combinations of the numbers taken three at a time, and so forth.

Cor. 1. If $A_1, A_2, \ldots\ldots, A_n$ are n mutually exclusive events, then

$$P(A_iA_j)=0,\ P(A_iA_jA_k)=0,\ P(A_1A_2\ldots\ldots A_n)=0.$$

Consequently (4·6) reduces to

$$P(A_1+A_2+\ldots+A_n)=\sum_{i=1}^{n} P(A_i)=P(A_1)+P(A_2)+\ldots+P(A_n)$$

which is the theorem of total probability.

Cor. 2. $$P(A_1+A_2+\ldots+A_n)\geqslant \sum_{i=1}^{n} P(A_i)-\sum_{\substack{i,j=1\\ i\neq j}}^{n} P(A_iA_j)$$

From (4·5), $$P(A_1+A_2+A_3)\geqslant\sum_{i=1}^{3}P(A_i)-\sum_{\substack{i,j=1\\ i\neq j}}^{3} P(A_iA_j)$$

This result can be obtained by mathematical induction.

Note 1. If $P(A_i)$ be written as p_i, $P(A_iA_j)$ as p_{ij}, $P(A_iA_jA_k)$ as p_{ijk} and so on, then (4·6) can be written as

$$P(A_1+A_2+\ldots+A_n)=\Sigma p_i-\Sigma p_{ij}+\Sigma p_{ijk}-\ldots+(-1)^{n-1}p_{12\cdots n}$$
$$=S_1-S_2+S_3-\ldots+(-1)^{n-1}S_n. \qquad \ldots(4\cdot6')$$

where $S_1, S_2, \ldots\ldots$ denote the sum of the probabilities of events taken one, two, $\ldots\ldots$ at a time.

Note 2. Consider any two events A and B. The events $A-AB$, AB, $B-AB$ are always pairwise mutually exclusive and we have

$$A=(A-AB)+AB,\ B=AB+(B-AB),$$
$$A+B=(A-AB)+AB+(B-AB)$$

$$\therefore \quad P(A)=P(A-AB)+P(AB),\ P(B)=P(AB)+P(B-AB),$$
$$P(A+B)=P(A-AB)+P(AB)+P(B-AB).$$

Eliminating $P(A-AB)$ and $P(B-AB)$, we get

$$P(A+B)=P(A)+P(B)-P(AB).$$

Ex. 1. (*a*) For any two events A_1 and A_2, prove that

$$P(A_1+A_2)\leqslant P(A_1)+P(A_2)$$

(*b*) Arrange for any two events A_1 and A_2 the following probabilities in descending order of magnitude : $P(A_1)$, $P(A_1A_2)$, $P(A_1+A_2)$, and $P(A_1)+P(A_2)$

(*B.A., Hons, Delhi, 1962* ; *M.A., Statistics, Delhi 1960*)

[**Ans.** $P(A_1A_2)\leqslant P(A_1)\leqslant P(A_1+A_2)\leqslant P(A_1)+P(A_2)$]

Ex. 2. Use the theorem of Total probability to show that

(*i*) $P(A_1+A_2+A_3)=P(A_1\bar{A}_2\bar{A}_3)+P(A_2\bar{A}_3\bar{A}_1)+P(A_3A_2\bar{A}_1)+P(A_1A_2\bar{A}_3)$
$+(P\bar{A}_2A_3\bar{A}_1)+P(A_3\bar{A}_1\bar{A}_2)+P(A_1A_2A_3)$

(*ii*) $P(A_1)=P(A_1\bar{A}_2A_3)+P(A_1A_2\bar{A}_3)+P(A_1A_2\bar{A}_3)+P(A_1A_2A_3)$

(*iii*) $P(A_1A_2)=P(A_1A_2A_3)+P(A_1A_2\bar{A}_3)$

Hence deduce that

$$P(A_1+A_2+A_3)=P(A_1)+P(A_2)+P(A_3)-P(A_1A_2)-P(A_2A_3)-P(A_3A_1)+P(A_1A_2A_3)$$

Ex. 3. Assuming the above theorem to be true for n events and less, prove that it also holds for $n+1$ events.

Ex. 4. Prove the above theorem by the *method of inclusion and exclusion*.

[To find the *L.H.S.* of (4·6′), we have to find the number of *distinct* cases which are favourable to $A_1, A_2, \ldots, A_n$. The formula will be proved if we show that each case occurs once and once only as the *R.H.S.* of (4·6′). Consider a particular case. We can easily number the cases so that this case is favourable to $A_1, A_2, \ldots, A_m$ only. Then this case occurs ${}^mC_1$ times in S_1, ${}^mC_2$ times in S_2, ..., ${}^mC_m$ times in S_m and none at all in the remaining terms *i.e.*, it occurs ${}^mC_1-{}^mC_2+\ldots+(-1)^{m-1}{}^mC_m=1-(-1)^m$ times on the *R.H.S.*]

4·3·2. Geometrical interpretation of the above theorem. Let us suppose that an event A_1 is said to occur when the *sample point* lies within a certain *simple space* bounded by a curve C_1. The space outside the curve will, of course, correspond to $\bar{A}_1$. Similarly, the event A_2 will be said to occur when the sample point lies inside another curve C_2. Thus in Fig. 13, the whole area shaded by parallel lines corresponds to A_1, the area shaded by parallel lines only to A_1A_2, and the dotted area to A_2A_1. The shaded area shown by dots and crossed lines *i.e.*, the area common to the two curves corresponds to A_1A_2. The entire area bounded by curves C_1 and C_2 corresponds to A_1+A_2.

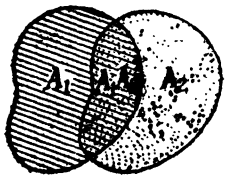


Fig. 13. A general sample space.

Area for A_1+A_2=Area for A_1+Area for A_2−Area for A_1A_2.

Dividing both sides by the total number of sample points we get

$$P(A_1+A_2)=P(A_1)+P(A_2)-P(A_1A_2)$$

Similarly, Fig. 14 explains equation (4·5).

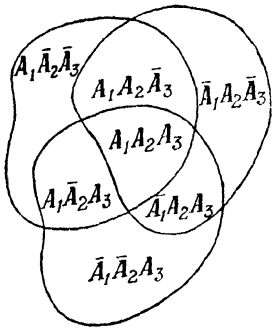


Fig. 14.

4·4·3. Bool's Inequality :

For n events $A_1, A_2, \ldots\ldots, A_n$

$$P(A_1+A_2+\ldots+A_n) \leqslant P(A_1)+P(A_2)+\ldots+P(A_n).$$

The equality sign, holding, when events are mutually exclusive.

Since $P(A+B)\leqslant 1$,

we have $P(A)+P(B)-P(AB)\leqslant 1$

$\Rightarrow$ $P(AB)\geqslant P(A)+P(B)-1$

$\therefore$ $P(ABC)\geqslant P(A)+P(BC)-1$

$\geqslant P(A)+P(B)+P(C)-2$

Generalizing,

$$P(A_1A_1\ldots A_n)\geqslant P(A_1)+P(A_2)+\ldots\ldots+P(A_n)-(n-1)$$

Similarly, $P(\bar{A}_1\bar{A}_2\ldots\ldots\bar{A}_n)\geqslant P(\bar{A}_1)+P(\bar{A}_2)+\ldots\ldots$

$$+P(\bar{A}_n)-(n-1)$$

$\Rightarrow$ $P(\overline{A_1A_2\ldots\ldots A_n})\geqslant[1-P(A_1)]+[1-P(A_2)]+\ldots\ldots$

$$+[1-P(A_n)]-(n-1)$$

Hence

$$P(A_1)+P(A_2)+\ldots\ldots+P(A_n)\geqslant 1-P(\overline{A_1A_2\ldots\ldots A_n})$$

$$=P(A_1+A_2+\ldots\ldots+A_n)]$$

Example 4·4. *A and B take turns in throwing two dice, the first to throw 9 being awarded the prize. Show that if A has the first throw, their chances of winning are in the ratio 9 : 8.*

(B.Sc. Delhi 1972)

9 can be made up in 4 ways *viz.*, 3, 6 ; 4, 5 ; 5, 4 ; 6, 3 with two disc and thus the chance of throwing 9 is 4/36 or $\frac{1}{9}$ and the chance of not throwing is $1-\frac{1}{9}$ or 8/9.

If A is to win, he should throw 9 in either the first, third or fifth.....throw.

The probability of A's throwing 9 on the first throw is $\frac{1}{9}$, in the third throw, it is $\frac{8}{9}\times\frac{8}{9}\times\frac{1}{9}$, or $(\frac{8}{9})^2\times\frac{1}{9}$, in the fifth throw it is $\frac{8}{9}\times\frac{8}{9}\times\frac{8}{9}\times\frac{8}{9}\times\frac{1}{9}$ or $(\frac{8}{9})^4\times\frac{1}{9}$ and so on.

Hence p_A = A's total chance of winning

$$=\tfrac{1}{9}+(\tfrac{8}{9})^2\times\tfrac{1}{9}+(\tfrac{8}{9})^4\times\tfrac{1}{9}+\ldots\ldots$$

$$=\tfrac{1}{9}/[1-(\tfrac{8}{9})^2]=\tfrac{9}{17}$$

In other words, *odds* are 9 to 8 in *favour* of A or 8 to 9 *against* A. B can win in second, fourth, or sixth, ...throw.

$$\therefore \quad p_B = \tfrac{8}{9}\times\tfrac{1}{9}+\tfrac{8}{9}\times\tfrac{8}{9}\times\tfrac{8}{9}\times\tfrac{1}{9}+\ldots\ldots$$
$$= \tfrac{8}{9}[\tfrac{1}{9}+(\tfrac{8}{9})^2\times\tfrac{1}{9}+(\tfrac{8}{9})^4\times\tfrac{1}{9}+\ldots\ldots]$$
$$= \tfrac{8}{9}\times\tfrac{9}{17}=\tfrac{8}{17}.$$

$\therefore$ A's chance to B's is in the ratio $\frac{9}{17} : \frac{8}{17}$ *i.e.*, 9 : 8.

Note. It will be observed that p_B could have been written down immediately from the relation $p_B = 1 - p_A$

Example 4·5. *Six cards are drawn with replacement from an ordinary deck. What is the probability that each of the four suits will be represented at least once among the six cards ?*

We shall first find the probability that all suits do not appear. Let A symbolize the appearance of all the suits and B the non-appearance of at least one of the suits. Since either A or B is certain to happen, $P(A+B)=1$ and since A and B are mutually exclusive, $P(A)=1-P(B)$.

To get $P(B)$, let us classify the possible outcomes favourable to B into four sets. B_1 is the set of all outcomes in which spades are absent, B_2 is the set for which hearts are absent, B_3 and B_4 are the sets for which diamonds and clubs are absent respectively.

These sets are overlapping : an outcome which consists of only spades and hearts falls in B_3 and in B_4. Clearly

$$P(B)=P(B_1+B_2+B_3+B_4)=\sum_{i=1}^{4} P(B_i)-\sum_{i\neq j} P(B_iB_j)+\sum_{i\neq j\neq k} P(B_iB_jB_k) - P(B_1B_2B_3B_4)$$

The probability $P(B_1)$ that a spade will not appear in the six draws is $(\frac{3}{4})^6$, and the value is the same for all B_i ; hence

$$\Sigma P(B_j)=4(\tfrac{3}{4})^6$$

The probability $P(B_1B_2)$ that neither spades, nor hearts will appear in the six draws is $(\frac{1}{2})^6$ and is the same for all six pairs of the four suits taken two at a time ; hence

$$\Sigma P(B_iB_j)=6(\tfrac{1}{2})^6$$

Similarly, $$\Sigma P(B_iB_jB_k)=4(\tfrac{1}{4})^6$$

and clearly $$P(B_1B_2B_3B_4)=0$$

$\therefore$ Required probability $= P(A)$
$$= 1-4(\tfrac{3}{4})^6+6(\tfrac{1}{2})^6-4(\tfrac{1}{4})^6=\cdot 381.$$

Example 4·6. *n letters to each of which corresponds an envelope, are placed in the envelopes at random. (i) What is the probability that no letter is placed in the right envelope ? (ii) What is the probability that exactly r letters are placed in the right envelopes ?*

(M.A., Bombay, 1955 ; M.A., Delhi 1957, 1966, I.C.A.R., Delhi 1948 ; M.A., Calcutta, 1956 ; M.A., Nagpur, 1953)

Let A_i, $i=1, 2, \ldots, n$ denote the event that the nth letter is placed in the right envelope. Then the

$$\text{Required probability} = P(\overline{A_1 A_2 \ldots\ldots A_n})$$
$$= 1 - P(A_1 + A_2 + \ldots + A_n)$$
$$= 1 - \sum_{i=1}^{n} P(A_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} P(A_i A_j) - \sum_{i \neq j \neq k} P(A_i A_j A_k) + \ldots + (-1)^n P(A_1 A_2 \ldots A_n)$$

Now $$P(A_i) = \frac{1}{n}, \quad i=1, 2, \ldots, n$$

$$P(A_i A_j) = P(A_i) P(A_j \mid A_i)$$
$$= \frac{1}{n} \cdot \frac{1}{n-1}, \text{ for all } i, j, \; i<j$$

$$P(A_i A_j A_k) = P(A_i) P(A_j \mid A_i) P(A_k \mid A_i A_j)$$
$$= \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} \text{ for all } i, j, k, \; i<j<k$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$P(A_1 A_2 \ldots A_n) = \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} \ldots\ldots \frac{1}{2} \cdot 1$$

Hence the required probability

$$= 1 - {}^nC_1 \cdot \frac{1}{n} + {}^nC_2 \cdot \frac{1}{n(n-1)} + {}^nC_3 \cdot \frac{1}{n(n-1)(n-2)}$$
$$\ldots\ldots + (-1)^n \cdot \frac{1}{n(n-1)\ldots 3.2.1}$$
$$= \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots\ldots + (-1)^n \cdot \frac{1}{n!}$$

Aliter. (*i*) Suppose that u_n is the number of ways in which all letters go wrong. Consider any two letters. If these occupy each other's envelope the number of ways in which the remaining $n-2$ letters go wrong is u_{n-2} and there are $n-1$ ways in which two letters can be interchanged. But if one letter occupies another's place and not *vice versa* which can happen in $(n-1)$ ways there are u_{n-1} ways in which the others can go wrong. Hence we have the *difference equation*

$$u_n = (n-1)(u_{n-1} + u_{n-2})$$

Re-writing we have

$$u_n - n u_{n-1} = -u_{n-1} - (\overline{n-1}\ u_{n-2})$$
$$= \ldots \quad \ldots \quad \ldots$$
$$= (-1)^{n-2}(u_2 - 2u_1).$$

But $u_1=0$ and $u_2=1$

and thus $u_n-nu_{n-1}=(-1)^n$

or $\frac{u_n}{n!}-\frac{u_{n-1}}{(n-1)!}=\frac{(-1)^n}{n!}$

Putting $n=2, 3, \ldots\ldots, n$ and adding, we get

$$u_n=n!\left\{\frac{1}{2!}-\frac{1}{3!}+\ldots+\frac{(-1)^n}{n!}\right\}$$

$\therefore$ Required probability $=\frac{u_n}{n!}=\frac{1}{2!}-\frac{1}{3!}+\ldots\ldots+\frac{(-1)^n}{n!}$

i.e., the first $(n-1)$ terms of $1-e^{-1}$.

(*ii*) r letters can be chosen in ${}^nC_r$ ways and the rest arranged in u_{n-r} ways.

$$\therefore \quad \text{Required probability}=\frac{{}^nC_r u_{n-r}}{n!}=\frac{1}{r!}\cdot\frac{u_{n-r}}{(n-r)!}$$

$$=\frac{1}{r!}\left\{\frac{1}{2!}-\frac{1}{3!}+\ldots+\frac{(-1)^{n-r}}{(n-r)!}\right\}.$$

Example 4·7. *Two similar decks of n different cards are put into random order and matched against each other. If a card occupies the same position in both the decks we speak of a match (coincidence). Find the probability of at least one match.*

Since we are interested in the relative order in two decks we can take without loss of generality one deck to be in its natural order. Let A_i' be the event that a match occurs at the i-th place. This means that card number 'i' is in the i-th place while the other $(n-1)$ cards may be in any order.

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

Similarly, probability that cards numbered i and j are in their natural places while the remaining $(n-2)$ cards may be in any arbitrary order is given by

$$p_{ij}=\frac{(n-2)!}{n!}=\frac{1}{n(n-1)}.$$

Thus sum S_r of formula (4·6′) contains ${}^nC_r$ terms each of which is $\frac{(n-r)!}{n!}$ and therefore

$$S_r={}^nC_r.\frac{(n-r)!}{n!}=\frac{1}{r!}$$

Hence $$P(A_1+A_2+\ldots+A_n)=S_1-S_2+S_3-\ldots\ldots\pm S_n$$

$$=1-\frac{1}{2!}+\frac{1}{3!}-\ldots\ldots\text{to } n \text{ terms.}$$

4·4·4. Theorem. *For any integer m with $1 \leqslant m \leqslant N$, the probability $P_{(m)}$ that exactly m among the N events $A_1, A_2, \ldots\ldots, A_N$ occur simultaneously is given by*

$$P_{(m)} = S_m - \binom{m+1}{m} S_{m+1} + \binom{m+2}{m} S_{m+2} \cdots\cdots \pm \binom{N}{m} S_N$$

$$= \sum_{k=m}^{N} (-1)^{k-m} \binom{k}{m} S_k \qquad \ldots(i)$$

where $$S_m = \sum_{k=1}^{m} p_{i_{k_1}} \; p_{i_{k_2}} \; \ldots\ldots p_{i_{k_n}}$$

Proof. Let E be an arbitrary sample point, and suppose that it is contained in exactly n among the N events A_j.

Then $P\{E\}$ appears as a contribution to $P_{(m)}$ only if $n = m$.

Now $P\{E\}$ appears as a contribution to the section

$$S_1 = \sum_{i=1}^{n} p_i, \; S_2 = \sum_{\substack{i,j=1 \\ i<j}}^{n} p_{ij}, \ldots\ldots, \quad S_n = \sum_{k=1}^{n} p_{i_{k_1} i_{k_2} \cdots\cdots i_{k_n}}$$

and not to $S_{n+1}, S_{n+2}, \ldots\ldots, S_N$.

Thus if $n < m$, $P\{E\}$ does not contribute to the right side of (i).

If $n = m$, then $P\{E\}$ appears in one and only 1 term of S_m.

If $n > m$, the contributions of $P\{E\}$ to the terms $S_m, S_{m+1}, \ldots$ $\ldots, S_n$, on the right of (i) cancel. Hence the total contribution of $P\{E\}$ to the right side of (i) is, therefore,

$$\binom{n}{m} - \binom{m+1}{m}\binom{n}{m+1} + \binom{m+2}{m}\binom{n}{m+2} - \ldots\ldots$$

$$\ldots + (-1)^{n-m} \binom{n}{m}\binom{n}{n}$$

$$= \sum_{k=0}^{n-m} (-1)^k \binom{m+k}{k}\binom{n}{m+k}$$

$$= \sum_{k=0}^{n-m} (-1)^k \frac{(m+k)!}{k!\,m!} \cdot \frac{n!\,(n-m)!}{(m+k)!\,(n-m-k)!\,(n-m)!}$$

$$= \binom{n}{m} \sum_{k=0}^{n-m} (-1)^k \binom{n-m}{k}$$

$$= \binom{n}{m}\left[1 - \binom{n-m}{1} + \binom{n-m}{2} \ldots\right.$$

$$\left. \ldots\ldots + (-1)^{n-m} \binom{n-m}{n-m}\right]$$

$$=\binom{n}{m}(1-1)^{n-m}=0.$$

Cor. The probability P_m that at least m among the N events occur simultaneously is given by

$$P_m=P_{(m)}+P_{(m+1)}+P_{(m+2)}+\ldots\ldots P_{(N)}$$
$$=S_m-\binom{m}{m-1}S_{m+1}+\binom{m+1}{m-1}S_{m+2}\cdots$$
$$\ldots\ldots\pm\binom{N-1}{m-1}S_N$$

Ex. Prove that

(*i*) $S_m-(m+1)\,S_{m+1}\leqslant P_{(m)}\leqslant S_m$

(*ii*) $S_m-m\,S_{m+1}\leqslant P_m<S_m$.

(M.Sc. Bombay 1971 ; M.Sc. Gauhati 1967 ; M.Sc. Delhi 1968)

4·5. Multiplicative Law of Probability or Theorem of Compound Probability. *The probability of the simultaneous occurrence of two events A and B is equal to the probability of A multiplied by the conditional probability of B, given that A has occurred (or it is equal to the probability of B multiplied by the conditional probability of A given that B has occurred). In symbols,*

$$P(AB)=P(A)P(B\mid A)$$
$$=P(B)P(A\mid B).$$

Proof. Let N denote the total number of mutually exclusive and equally likely cases among which m cases are favourable to the event A. The cases favourable to both the events A and B are included in the m cases favourable to A. Let their number be m_1. Then the probability $P(AB)$ that both the events A and B will happen is given by

$$P(AB)=\frac{m_1}{N}=\frac{m}{N}\cdot\frac{m_1}{m}.$$

The ratio m/N is the probability of A, that is $m/N=P(A)$.

Assuming the occurrence of A there are only m equally likely cases left out of which m_1 are also favourable to B. Hence the ratio m_1/m represents the conditional probability $P(B\mid A)$ of B, supposing that A has occurred.

Hence $$P(AB)=P(A)P(B\mid A)$$

Since the compound event AB involves A and B symmetrically, also have

$$P(AB)=P(B)P(A\mid B).$$

Cor. 1. If A and B are independent events then $P(B\mid A)$ is

the same as $P(B)$. Thus for two independent events the theorem of compound probability becomes

$$P(AB)=P(A)P(B).$$

In general, *if the constituent events of a compound event are mutually independent, the probability of the compound event is equal to the product of the probabilities of the constituent events.*

In symbols,

$$P(A_1.A_2...A_n)=P(A_1)P(A_2).....P(A_n)$$

provided that $P(A_i)=P(A_i \mid A_j...A_p)$ for all $i, j, ..., p$, $i\neq j\neq ...\neq p$.

Cor. 2. $P(AB \mid C)=P(A \mid C)P(B \mid AC)$.

The above proof holds in this case provided N represents the total number of equally likely and mutually exclusive events favourable to C.

Cor. 3. $P(ABC)=P(A)P(B \mid A)P(C \mid AB)$

For $P(ABC)=P[A(BC)]=P(A)P(BC \mid A)$

$=P(A)P(B \mid A)P(C \mid AB)$ [using Cor. 2]

Cor. 4. $P(A_1A_2...A_n)=P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2).....$

$P(A_n \mid A_1A_2 ..A_{n-1})$

$$=P\left(A_{i_1}\right)P\left(A_{i_2}\middle|A_{i_1}\right)P\left(A_{i_3}\middle|A_{i_1}\ A_{i_2}\right)...$$

$$P\left(A_{i_n}\middle|A_{i_1}\ A_{i_2}\ .....\ A_{i_{n-1}}\right)$$

where $i_1, i_2, ..., i_n$ denote any permutation of the integers $1, 2, ..., n$.

Example 4·8. *A bag contains 4 red and 3 blue balls. Two drawings of 2 balls are made. Find the chance that the first drawing gives 2 red balls and the second drawing 2 blue balls*

(*a*) *if the balls are returned to the bag after the first draw,*

(*b*) *if the balls are not returned.*

(*a*) The number of ways in which any two balls may be drawn is ${}^7C_2$. The number of ways in which two red balls may be drawn is ${}^4C_2$. Therefore, the probability of drawing two red balls is ${}^4C_2/{}^7C_2=\frac{2}{7}$.

Similarly, the probability of drawing two blue balls is ${}^3C_2/{}^7C_2=\frac{1}{7}$.

Therefore, the chance of the compound event is $\frac{2}{7}\times\frac{1}{7}=\frac{2}{49}$.

(*b*) As before the chance of drawing 2 red balls is $\frac{2}{7}$. The balls are not returned to the bag and the number of ways in which the two balls may be drawn is ${}^5C_2$ so that the chance of drawing blue balls is $\frac{{}^3C_2}{{}^5C_2}=\frac{3}{10}$.

Therefore the chance of compound event is $\frac{2}{7}\times\frac{3}{10}=\frac{3}{35}$.

Note. In (*a*) of this example the two events are independent but in (*b*) they are clearly not independent since the chance of drawing two blue balls is affected by the first draw.

Example 4·9. *The probability of n independent events are $p_1, p_2, \ldots, p_n$. Find an expression for the probability that at least one of the events will happen.* (*I.A.S., 1951*)

Use the result to find the chance of obtaining at least one 6 in a throw of four dice.

For the rth event, the chance that it will not happen is given by $1-p_r$, $r=1, 2,\ldots, n$. The chance, therefore, that none of the events will happen is given by $(1-p_1)(1-p_2)\ldots\ldots(1-p_n)$, and since it is certain that either no event will happen or that at least one event will happen the probability that at least one event will happen is given by $1-(1-p_1)(1-p_2)\ldots\ldots(1-p_n)$.

The probability of obtaining a 6 in a throw of a dice is $\frac{1}{6}$ so that using the above result the chance of at least one 6 in throw of four dice is given by $1-\left(\frac{5}{6}\right)^4=\frac{671}{1296}$.

Example 4·10. *For any three events A, B and C, prove that*

(*i*) $P(A+B \mid C)=P(A \mid C)+P(B \mid C)-P(AB \mid C)$

(*ii*) $P(A\bar{B} \mid C)+P(AB \mid C)=P(A \mid C)$.

(*B.A. Hons. Delhi 1966*)

We have

(*i*) $P(A+B)=P(A)+P(B)-P(AB)$

$\therefore$ $P(AC+BC)=P(AC)+P(BC)-P(ABC)$

Hence

$$\frac{P(AC+BC)}{P(C)}=\frac{P(AC)}{P(C)}+\frac{P(BC)}{P(C)}-\frac{P(ABC)}{P(C)}$$

$\therefore$ $\frac{P[(A+B)C]}{P(C)}=P(A \mid C)+P(B \mid C)-P(AB \mid C)$

$\Rightarrow$ $P(A+B \mid C)=P(A \mid C)+P(B \mid C)-P(AB \mid C)$

(*ii*) $P(A\bar{B} \mid C)+P(AB \mid C)$

$$=\frac{P(A\bar{B}C)}{P(C)}+\frac{P(ABC)}{P(C)}$$

$$=\frac{P(A\bar{B}C)+P(ABC)}{P(C)}$$

$$=\frac{P(AC)}{P(C)}$$

$$=P(A \mid C).$$

Ex. 1. The probability that a teacher will give an unannounced test during any class meeting is $\frac{1}{5}$. If a student is absent twice, what is the probability that he will miss at least one test ? [Ans. $\frac{9}{25}$]

Ex. 2. A can solve 75% of the problems in this book and B can solve 70%. What is the probability that either A or B can solve a problem chosen at random ? *(B.A. Punjab 1961 S.)* [Ans. $\frac{37}{40}$]

4·6. Bayes Theorem. *An event A can be explained by a set of exhaustive and mutually exclusive hypotheses $B_1, B_2, \ldots\ldots, B_n$. Given*

(i) 'a priori' probabilities $P(B_1), P(B_2), \ldots\ldots, P(B_n)$ corresponding to a total absence of knowledge regarding the occurrence of A and

(ii) conditional probabilities $P(A \mid B_1), P(A \mid B_2), \ldots P(A \mid B_n)$.

(a) It is required to form the 'a posteriori' probabilities $P(B_1 \mid A), P(B_2 \mid A), \ldots\ldots, P(B_n \mid A)$.

(b) Further, find the probabilities of materialisation of another event C, given the probabilities

$$P(C \mid AB_1),\ P(C \mid AB_2),\ \ldots\ldots,\ P(C \mid AB_n).$$

Proof. *(i)* By the theorem of compound probability,

$$P(AB_i) = P(B_i)P(A \mid B_i) = P(A)P(B_i \mid A)$$

$$\therefore \quad P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{P(A)} \qquad \ldots(4\cdot7)$$

Since the event A can materialise in the mutually exclusive forms $AB_1, AB_2, \ldots\ldots, AB_n$, we have by the theorem of total probability

$$P(A) = P(AB_1) + P(AB_2) + \ldots + P(AB_n)$$

$$= \sum_{i=1}^{n} P(B_i)P(A \mid B_i) \qquad \ldots(4\cdot8)$$

$$\therefore \quad P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)} \qquad \ldots(4\cdot9)$$

(ii) $$P(C \mid A) = \sum_{i=1}^{n} P(CB_i \mid A) = \sum_{i=1}^{n} P(B_i \mid A)P(C \mid AB_i)$$

$$P(C \mid A) = \frac{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)P(C \mid AB_i)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)}$$

using (4·9)... (4·10)

Note. The formula (4·9) is also known as the formula for 'a posteriori' probability. This name is explained by the fact that this formula gives the probability of relation of the events B_i with respect to the occurrence of A, $P(B_i)$ is known as a *Priori* probability.

Ex. 1. The contents of three urns 1, 2, 3 are as follows :

1 white, 2 red, 3 black balls

2 white, 3 red, 1 black balls

3 white, 1 red, 2 black balls

An urn is chosen at random and from it two balls are drawn at random. The two balls are one red and one white. What is the probability that they come from the second urn ? *(B.Sc. Hons. Cal. 1968)*

[Hint. $P[B_1]=P[B_2]=P[B_3]=\frac{1}{3}$; $P[A \mid B_1]=\frac{2}{15}$, $P[A \mid B_2]=\frac{2}{5}$

$P[A \mid B_3]=\frac{1}{5}$, $P[B_1 \mid A]=\frac{2}{11}$, $P[B_2 \mid A]=\frac{6}{11}$, $P[B_3 \mid A]=\frac{3}{11}$.]

Ex. (2) From an urn containing 3 white and 5 black balls, 4 balls are transferred into an empty urn. From this urn 2 balls are taken and both happen to be white. What is the probability that the three balls taken from the same urn will be white ?

[Hint. $P(B_1)=\frac{3}{7}$, $P(B_2)=\frac{1}{14}$; $P(A \mid B_1)=\frac{1}{6}$, $P(A \mid B_2)=\frac{3}{6}$, $P(C \mid AB_1)=\frac{1}{2}$, $P(C \mid AB_2)=\frac{3}{4}$, $P(C \mid A)=\frac{7}{12}$.]

4·7·1. Set-theoretic Formulation of Probability Theory. The above discussion of the theory of probability is restrictive in the sense that it applies only when all the cases under consideration are equally likely. We want to generalise it to the case when all the cases are not necessarily equally likely. However the generalisation should include the earlier discussion as a special case.

The above discussion is also quite intuitive. We want to make it more rigorous and axiomatic. The definitions and axioms will of course, be motivated by the results obtained above. In our discussion, we shall identify events with sets and as such some concepts from *set theory* will also be useful.

A set is a well-defined collection of objects, so that given any object, we can say whether it belongs to the set or not. Each object of the set is called its *element.* The set of all elements under consideration is called the *universal set* and the set which contains no element is called the *null set* or the empty set and is denoted by ϕ. The *union* $A \cup B$ of the sets A and B is set of all those elements which belong to A or to B or to both. The *intersection* $A \cap B$ of the two sets A and B is the set of all the elements common to A and B. The *complement* $\bar{A}$ of a set A is the set of all those elements of the universal set which do not belong to A. Two sets A and B are said to be disjoint if they have no common elements *i.e.* if $A \cap B=\phi$. A set A is said to be a *subset* of a set B if whenever an

element x belongs to A, it also belongs to B and we write $A \subset B$. If $A \subset B$ and $B \subset A$, the two sets are the same and we write $A=B$.

Definition of probability. When we perform an experiment, actual or conceptual, there may be a number of possible outcomes $e_1, e_2, \ldots\ldots, e_n$, say. We assume that these outcomes are *mutually exclusive i.e.* only one of the outcomes is possible at a time and these are *exhaustive i.e.* one of these outcomes must occur. Such outcomes are called *elementary events* and the set of all such outcomes *i.e.* the set of all elementary events is called the *sample space.*

Thus if the experiment consists of throwing of a six-faced die, the possible outcomes are : turning up of 1, 2, 3, 4, 5 or 6. These outcomes are mutually exclusive since only one of the numbers may turn up at a time, and they are exhaustive, since one of these must occur (we, of course, assume that the die will not stand on an edge !) Thus there are six elementary events and the sample space consists of 6 points viz. 1, 2, 3, 4, 5, 6.

When we throw up a coin, the sample space consists of two points viz. Head and Tail or H and T.

When we throw up two dice simultaneously, the sample space consists of 36 elementary events viz.

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

where the elementary event (3, 2) means that the first die shows 3 and the second die shows 2.

With three dice, the sample space will consist of 216 elementary events or points viz. all ordered triplets of the form (i, j, k) where i, j, k can take the values 1 to 6.

Our experiments may also be conceptual *e.g.*, we may think of a rational number, so that an elementary event is any rational number. In this case our sample space is countably infinite or *denumerable i.e.* in this case we can set up a one-one correspondence between elements of the set of elementary events and the members of the set of natural numbers. If our experiment is thinking of a point on a line or in a plane, the sample space is non-countably infinite. We shall confine ourselves, however, mainly to finite sample spaces.

We now assign real numbers $p_1, p_2, \ldots\ldots, p_n$ to each of the elementary events $e_1, e_2, \ldots\ldots, e_n$ in such a way that

(*i*) $p_i \geqslant 0$ $(i=1, 2, \ldots\ldots, n)$

(*ii*) $p_1+p_2+ \ldots\ldots +p_n=1$.

These numbers are called *probabilities of elementary events* and the assignment of these numbers satisfying these two conditions is called an *allowable* or *permissible assignment of probabilities.*

The assignment in which $p_1=p_2=\ldots\ldots=p_n=1/n$ *i.e.*, in which all the elementary events are assigned the same probability is called the *natural assignment of probabilities.*. We make this assignment when we have reasons to believe that all outcomes are "*equally likely*" *i.e.* when there is no reason to believe that any one outcome is more likely than ony other *e.g.* if we have an 'unbiased die', we may assign the probability 1/6 to each of the 6 elementary events and we may assign probability 1/36 to each of the 36 elementary events.

In the case of a biased die which is leaded on one side, we cannot use the natural assignment of probabilities. How do we then assign the probabilities to the elementary events ? One method is to throw the die a large number of times and note the number of times it shows a fixed number say 2. If a die is thrown repeatedly a large number of times under more or less identical conditions, it will be found that the ratio of number of times the die shows 2 to total number of times it is thrown becomes almost stationary at some number. This ratio may be assigned as the probability for the elementary event corresponding to throwing up of number 2. In the theory of probability, we shall however assume that such a permissible assignment of probabilities of elcmentary events is already given to us.

We now define an *event* as any subset on the sample space, and probability of an event E is the sum of the assigned probabilities to the elements of the events.

Thus if,

$$E=e_{i_1}+e_{i_2}+\ldots\ldots+e_{i_m}$$

then,

$$P(E)=p_{i_1}+p_{i_2}+\ldots\ldots+p_{i_m}$$

when $e_{i_1}, e_{i_2}, \ldots\ldots, e_{i_m}$ are the m elementary events which are elements of E.

There are 2^n subsets of the sample space S and each of these will have a probability associated with it. It is obvious that

$$P(\phi)=0, \qquad P(S)=1$$

and that probability of any other event will lie between 0 and 1.

In the case of natural assignment of probabilities, the probability of each elementary element is $\frac{1}{n}$ and

$$P(E)=\frac{m}{n}=\frac{N(E)}{N(S)}$$

i.e. probability of an event is the number of elementary elements in E to the number of elementary events in S. This is sometimes expressed by defining probability or an event E as the ratio of the number of cases favourable to E to the total number of cases when the cases are exclusive, exhaustive and equally likely. This definition is useful only for the natural assignment of probabilities. This was the definition we used earlier.

For the natural assignment of probabilities, it is easily seen that

P(Head in the case of a coin throwing)$=\frac{1}{2}$

P(Odd number turning up in the case of die)$=\frac{3}{6}=\frac{1}{2}$.

P(Sum of points being 7 in the case of two dice)$=\frac{6}{36}=\frac{1}{6}$

Now suppose the assignment of probabilities for the die is,

$$p_1=\frac{1}{12},\ p_2=\frac{1}{12},\ p_3=\frac{1}{12},$$

$$p_4=\frac{1}{3},\ p_5=\frac{1}{12},\ p_6=\frac{1}{3}$$

This is allowable since the sum of these numbers is unity and each one is a non-negative real number, then

$$P(\text{odd number})=p_1+p_3+p_5=\frac{1}{12}+\frac{1}{12}+\frac{1}{12}=\frac{1}{4}$$

$$P(\text{even number})=p_2+p_4+p_6=\frac{1}{12}+\frac{1}{3}+\frac{1}{3}=\frac{3}{4}.$$

Theorem of Total Probability.

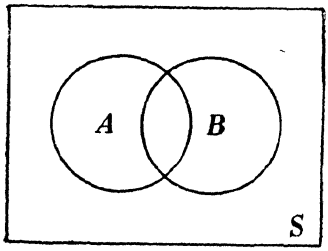


Fig. 15.

$P[A \cup B]$=sum of probabilities assigned to all elementary events in $A \cup B$.

Now the elementary events in $A \cup B$ can be obtained by taking all the elementary events in A and all the elementary events in B together, but in this process, we consider the elementary events in $A \cap B$ twice and

therefore we remove all the elementary events in $A \cap B$ once, so that we get

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

In the same way, we get

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C),$$

since to get all the elementary events in $A \cup B \cup C$, we first combine together all the elementary events in A, B and C, but in this pairwise common events have been counted twice, so we remove them once, but in this removal, we have removed the elementary events of $A \cap B \cap C$, once too often, and so we include these events once more.

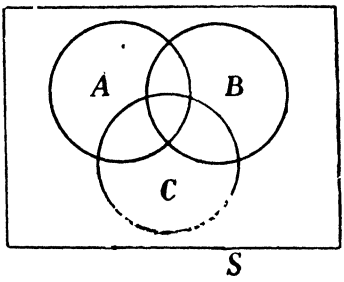


Fig. 16.

Two events are said to be mutually exclusive if they have no elementary event in common *i.e.* if the happening of one precludes the happening of the other *i.e.* if the corresponding subsets of S are disjoint.

Thus A and B are disjoint if

$$A \cap B = \phi$$

For two mutually exclusive events

$$P(A \cap B) = P(\phi) = 0$$

Similarly a number of events are said to be mutually exclusive if the corresponding sets are all disjoint.

For *mutually exclusive events*

$$P(A \cup B) = P(A) + P(B)$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

and in general,

$$P(A_1 \cup A_2 \cup \ldots\ldots \cup A_m) = P(A_1) + P(A_2) + \ldots\ldots + P(A_m)$$

Theorem of Compound Probability. Suppose an event B is known to have happened then $B \neq \phi$ or $P(B) \neq 0$.

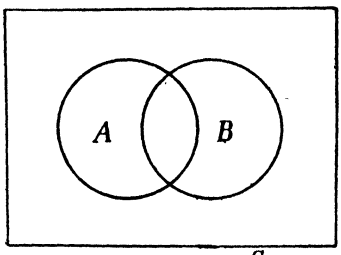


Fig. 17.

We now define the conditional assignment of probability to an elementary event $\{e_i\}$ by

$$P(\{e_i\} \mid B) = \frac{P(\{e_i\})}{P(B)} \quad \text{if } e_i \in B$$

$$= 0 \quad \text{if } e_i \notin B$$

This defines an allowable assignment of probabilities in S, since

$$(i)\ P(\{e_i\}) \geqslant 0,\ P(B) > 0, \frac{P(\{e_i\})}{P(B)} \geqslant 0$$

$$(ii)\ P(\{e_i\} \mid B) + P(\{e_2\} \mid B) + \ldots\ldots, + P\{e_n\} \mid B)$$

$$= \frac{\Sigma' P(\{e_i\})}{P(B)} = 1,$$

where Σ' denotes summation over these elementary events which are elements of B (for other elementary events, the contribution is 0).

We now define the conditional probability $P(A \mid B)$ of the happening of A when B is known to have occurred as sum of the probabilities $P(\{e_i\} \mid B)$ over all the elements of A. However the contribution to the sum is 0, from all those elements of A which are not elements of B *i.e.* the contribution is only from the elements of $A \cap B$ so that we get

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

or

$$P(A \cap B) = P(B) P(A \mid B)$$

Similarly

$$P(B \cap A) = P(A) P(B \mid A)$$

But

$$P(A \cap B) = P(B \cap A)$$

so that we get,

$$P(A \cap B) = P(B) P(A \mid B) = P(A) P(B \mid A)$$

This is known as the *theorem* of *compound probability*.

Now an event A is said to be *independent* of an event B if

$$P(A \mid B) = P(A)$$

i.e. if the probability of happening of A is independent of the happening of B. For such an event, the theorem gives

$$P(A \cap B) = P(B) P(A \mid B) = P(B) P(A)$$

showing that,

$$P(A) = P(A \mid B)$$

i.e. that B is independent of A.

Here we have assumed that $P(A) \neq 0$, $P(B) \neq 0$, that A and B are non-null sets or A and B are not impossible events. Thus if A is independent of B, then B is independent of A. We can speak of A and B as independent events.

4·7·2. Pairwise Independent Events. For two *independent events* A and B

$$P(A \cap B) = P(A) P(B)$$

The *converse* of this is also true *i.e.* if $P(A\cap B)=P(A)P(B)$, then A and B are independent events.

Proof. $P(A\cap B)=P(A)P(B)$ and $P(A\cap B)=P(A)P(B\mid A)$ $=P(B)P(A\mid B)$ showing that,

$$P(B)=P(B\mid A) \text{ and } P(A)=P(A\mid B)$$

and the result is established.

For three events A, B, C, repeated use of the theorem compound probability, gives

$$P(A\cap B\cap C)=P[(A\cap B)\cap C]=P(A\cap B)P(C\mid A\cap B)$$
$$=P(A)P(B\mid A)P(C\mid A\cap B)$$

Since $A\cap B\cap C$ is unchanged by the change of order of the events, we get

$$P(A\cap B\cap C)=P(A)P(B\mid A)P(C\mid A\cap B)=P(A)P(C\mid A)P(B\mid A\cap C)$$
$$=P(B)P(C\mid B)P(A\mid B\cap C)=P(B)P(A\mid B)P(C\mid A\cap B)$$
$$=P(C)P(B\mid C)P(A\mid B\cap C)=P(C)P(A\mid C)P(B\mid A\cap C)$$

If A, B, C are *pairwise independent*, then these give

$$P(A\cap B\cap C)=P(A)P(B)P(C\mid A\cap B)=P(A)P(C)P(B\mid A\cap C)$$
$$=P(B)P(C)P(A\mid B\cap C)$$

Example 4·11. *Show by an example that if C is independent of A and B separately, it will not be necessarily independent of $A\cap B$.*

Consider the simultaneous throw of two die.

Let A be the event that the number on the first die is odd.

Let B be the event that the number on the second die is odd.

Let C be the event that the sum of numbers on the two die is odd.

It is easily seen by considering the 36 points in the sample space that,

$$P(A)=\frac{1}{2}, \qquad P(B)=\frac{1}{2}, \qquad P(C)=\frac{1}{2}$$

$$P(A\cap B)=\frac{1}{4}, \quad P(A\cap C)=\frac{1}{4}, \quad P(B\cap C)=\frac{1}{4}$$

$$P(A\cap B\cap C)=0$$

Since $P(A\cap C)=P(A)P(C)$, $P(B\cap C)=P(B)P(C)$, C is independent of both A and B and since

$$P(A\cap B\cap C)=P(A)P(B)P(C\mid A\cap B)=0$$
$$P(C\mid A\cap B)=0=P(C)$$

so that C is not independent of $A\cap B$.

Thus the result $P(A\cap B\cap C)=P(A)P(B)P(C)$ is not necessarily true for pairwise independent events. We require an additiona condition.

Condition for Mutual Independence of n Events. We accordingly define events $E_1, E_2, \ldots\ldots, E_n$ to be mutually independent only if,

$$P(E_i \cap E_j) = P(E_i)P(E_j)$$
$$P(E_i \cap E_j \cap E_k) = P(E_i)P(E_j)P(E_k)$$
$$P(E_i \cap E_j \cap E_k \cap E_l) = P(E_i)P(E_j)P(E_k)P(E_l)$$
$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$
$$P(E_1 \cap E_2 \ldots\ldots \cap E_m) = P(E_1)P(E_2)\ldots\ldots P(E_m)$$

for all unequal $i, j, k, l, \ldots\ldots$ The total number of conditions imposed is ${}^{m}C_2 + {}^{m}C_3 + \ldots\ldots + {}^{m}C_m = 2^m - m - 1$.

Example 4·12. *Prove that if A, B and C are random events in a sample space and if A, B, C are pairwise independent and A is independent of $(B \cup C)$, then A, B and C are mutually independent.*

(M.A. Delhi, 1965)

Since A, B, C are pairwise independent, we have

$$P(A \cap B) = P(A)P(B)$$
$$P(B \cap C) = P(B)P(C)$$
$$P(A \cap C) = P(A)P(C)$$

and $$P(A \cap B \cup C) = P(A)P(B \cup C) \quad \ldots (1)$$

Now
$$P[A \cap B \cup C] = P[(A \cap B) \cup (A \cap C)]$$
$$= P(A \cap B) + P(A \cap C) - P[(A \cap B) \cap (A \cap C)]$$
$$= P(A)P(B) + P(A)P(C) - P(A \cap B \cap C)$$

Also
$$P(A)P(B \cup C) = P(A)[P(B) + P(C) - P(B \cap C)]$$
$$= P(A)P(B) + P(A)P(C) - P(A)P(B \cap C)$$

Hence using (1), we have

$$P(A)P(B) + P(A)P(C) - P(A \cap B \cap C)$$
$$= P(A)P(B) + P(A)P(C) - P(A)P(B \cap C)$$
$$\Rightarrow \quad P(A \cap B \cap C) = P(A)P(B \cap C) = P(A)P(B)P(C)$$

Hence the result is established.

Example 4·13. *An event A is known to be independent of the events B, $B \cup C$ and $B \cap C$. Show that it is also independent of C.*

Since A and B are independent,

$$P(A \cap B) = P(A)P(B)$$

Similarly $P[A \cap (B \cup C)] = P(A)P(B \cup C)$

$$P[A \cap (B \cap C)] = P(A)P(B \cap C)$$

$$\therefore \quad P(A \cap B) + P(A \cap C) - P(A \cap B \cap C)$$
$$= P(A)\,[P(B) + P(C) - P(B \cap C)]$$

$$\Rightarrow \quad P(AB) + P(AC) - P(A \cap B \cap C)$$
$$= P(A)P(B) + P(A)P(C) - P(A)P(B \cap C)$$

$$\Rightarrow \quad P(A \cap C) = P(A)P(C)$$

Hence A and C are independent.

4·8. It is obvious that the term universal set, null set, subset in set theory correspond respectively to sample space, impossible event, event in probability theorem. Also the notation $E+F$ and $E.F$ we used correspond to $E \cup F$ and $E \cap F$ of set theory.

Axiomatic Development of Probability Theory. In the axiomatic development of probability theory, probability is defined as a function defined on events (subsets of S) *i.e.*, it is a rule which associates to each event A a certain real number $P(A)$ which satisfies the following three axioms.

Axiom I. $P(A) \geqslant 0$, *i.e.*, the probability of every event is non-negative.

Axiom II. $P(S)=1$, *i.e.*, the probability of a certain event is unity.

Axiom III. If $A_1, A_2, \ldots\ldots$ are a countable number of sub-events of S such that

$$A_1 \cup A_2 \cup A_3 \ldots\ldots = S \text{ and } A_i \cap A_j = \phi (i \neq j),$$

then $$P(A_i \cup A_j \cup A_k \cup \ldots\ldots) = P(A_i) + P(A_j) + P(A_k) + \ldots\ldots$$

i.e., the probability of a union of disjoint events is the sum of probabilities of the events themselves.

From these axioms, we easily deduce the following theorems :

Theorem 1. $P(\phi)=0$

Proof : $S \cup \phi = S,\ S \cap \phi = \phi$

and therefore $P(S \cup \phi) = P(S) = P(S) + P(\phi)$, so that $P(\phi)=0$

Theorem 2. $P(\bar{A}) = 1 - P(A)$.

Proof : $A \cup \bar{A} = S,\ A \cap \bar{A} = \phi,$

$$P(A \cup \bar{A}) = P(S) = 1 = P(A) + P(\bar{A})$$

so that $$P(\bar{A}) = 1 - P(A).$$

Theorem 3. $A \subset B \Rightarrow P(A) \leqslant P(B)$

Proof : $A \subset B \Rightarrow A \cup (B \cap \bar{A}) = B \cup A = B$

Also $A \cap (B \cap \bar{A}) = \phi$, so that

$$P[A \cup (B \cap \bar{A})] = P(A) + P(B \cap A) = P(B)$$

But $P(B \cap \bar{A}) \geqslant 0$, so that theorem 3 follows.

Theorem 4. $0 \leqslant P(A) \leqslant 1$ for any A in S

Proof : $A \subset S \Rightarrow P(A) \leqslant P(S) \Rightarrow P(A) \leqslant 1$

since from axiom $P(A) \geqslant 0$, it follows that $0 \leqslant P(A) \leqslant 1$.

Definition. The conditional probability of event A, given B, is defined by

$$P(A \mid B)=\frac{P(A\cap B)}{P(B)}$$

wherever $P(B)\neq 0$

We can built up a rigorous theory of probability based on the above axioms and definition. Here we define probability as a function satisfying certain axioms. The theorem of total probability is taken as an axiom and the theorem of compound probability is taken as a definition.

This approach is similar to that of Euclidean geometry where points, lines etc., are undefined terms satisfying certain axioms and where we prove certain theorems based on these axioms. Geometry becomes practical when we give intuitive meaning to points and lines. In the same way the theorem of this deductive theory of probability will be useful when probability is given an intuitive meaning as in the earlier sections of this chapter.

Example 4·13. *If A and B are events such that $A\subset B$ then $P(A)\leqslant P(B)$.*

Since $A\subset B$ we have from figure or otherewise

$$B=A\cup(B\cap\bar{A})$$

$$\therefore \quad P(B)=P(A)+P(B\cap\bar{A})$$

since A and $B\cap\bar{A}$ are mutually exclusive events.

Hence $P(B)\geqslant P(A)$ since $P(B\cap A)\geqslant 0$

Example 4·14. *If for two mutually exclusive events A and B $P(B)=2P(A)$ and $A\cup B=S$, find $P(A)$.*

We have $P(A)+P(B)=P(A\cup B)=P(S)=1$

or $3P(A)=1$

$$\therefore \quad P(A)=\tfrac{1}{3}.$$

Example 4·15. *Prove the following laws, assuming that in each case the conditional probabilities are defined.*

(*a*) $P(B)=1$, then $P(A \mid B)=P(A)$.

(*b*) If A and B are disjoint and $P(B)>0$, then $P(A \mid B)=0$.

(*c*) If $A_1\subset A_2$, then $P(A_1 \mid B)\leqslant P(A_2 \mid B)$

(*d*) $P(A_1\cup A_2 \mid B)=P(A_1 \mid B)+P(A_2 \mid B)-P(A_1\cap A_2 \mid B)$.

(*a*) Since $P(B)=1$

$\therefore$ $B=S=$ entire sample space and $P(S)=1$

Now $P(A \mid B)=P(A\cap B)/P(B)=P(A\cap S)/P(S)=P(A)$.

(*b*) Since $A\cap B=\phi$, we have $P(A\cap B)=P(\phi)=0$.

$\therefore \quad P(A \mid B)=P(A \cap B)/P(B)=0.$

(*c*) Since $A_1 \subset A_2$, we have $A_1 \cap B \subseteq A_2 \cap B$

$\therefore \quad P(A_1 \cap B) \geqslant P(A_2 \cap B)$

or $\quad P(B)P(A_1 \mid B) \leqslant P(B)B(A_2 \mid B).$

Cancelling $P(B)$ which is not zero we obtain the required result.

(*d*) $$\begin{aligned} P(A_1 \cup A_2 \mid B) &= P(B \cap (A_1 \cup A_2)] \\ &= P[BA_1 \cup BA_2]/P(B) \\ &= [P(BA_1)+P(BA_2)-P(A_1A_2 \mid B)]/P(B) \\ &= P(A_1 \mid B)+P(A_2 \mid B)-P(A_1A_2 \mid B) \end{aligned}$$

Ex. 1. For any two events A and B, prove that

$$P[(A \cap B) \cup (B \cap \bar{A}]=P(A)+P(B)-2P(A \cap B).$$

Ex. 2. Let A, B and C be three events associated with an experiment $\{A, B, C\}$ and let $P(B)=\frac{3}{2}P(A)$, $P(C)=\frac{1}{3}P(B)$, find $P(A)$. [**Ans.** $\frac{1}{3}$]

Ex 3. Suppose that A and B are events for which

$$P(A)=x, P(B)=y \text{ and } P(A \cap B)=z.$$

Express each of the following probabilities in terms of x, y and z.

(*a*) $P(\bar{A} \cup \bar{B})$ (*b*) $P(\bar{A} \cap B)$

(*c*) $P(\bar{A} \cup B)$ (*d*) $P(\overline{A \cap B})$

(*e*) $P(A \cap \bar{B})$ (*f*) $P[(\overline{A \cap B})]$

[**Ans.** (*a*) $1-z$, (*b*) $y-z$, (*c*) $1-x+z$, (*d*) z, (*e*) $x-z$, (*f*) $1-z$]

Ex. 4. If A, B, C are mutually independent events, then $A \cup B$ and C are also independent.

EXERCISES 4

1. (*a*) Define independent and mutually exclusive events. Can two events be mutually exclusive and independent simultaneously ?

(*B.A., Punjab, 1953*)

[**Ans.** No]

(*b*) Define mutually exclusive events and state whether they can be regarded as a special case of dependent or independent events.

(*B.A., Hons., Delhi, 1957*)

(*c*) Give the 'a priori' and statistical definition of probability. Which of these is more general and why ? (*B.A., Punjab, 1957*)

2. (*a*) Discuss and criticise the following :

$P(A)=\frac{2}{3}P(B)=\frac{1}{4}P(C)=\frac{1}{6}$ for the probabilities of three mutually exclusive events A, B and C. (*Certificate Exam., Calcutta, 1952*)

(*b*) A, B and C are three independent events such that

$P(AB\bar{C})=\frac{1}{4}$, $P(A\,B\bar{C})=\frac{1}{8}$ and $P(\bar{A}\,B\bar{C})=\frac{1}{12}$. Prove that $P(A)=\frac{1}{2}$ and $P(\bar{A}\,\bar{B}\,\bar{C})=\frac{1}{4}$. (*B.Sc. Bombay, 1970*)

3. (*a*) Given $p_1=P(A)$, $p_2=P(B)$, $p_3=P(A \cap B)$, express the following in terms of p_1, p_2, p_3 :

(*i*) $P(\bar{A} \cup \bar{B})$ (*ii*) $P(\bar{A} \cap B)$ (*iii*) $P(\bar{A} \cup B)$ (*iv*) $P(\bar{A} \cup \bar{B})$ (*v*) $P(A \cap \bar{B})$

(*vi*) $P(\bar{A} \mid \bar{B})$ (*vii*) $P(\overline{A \cup B})$ (*viii*) $\overline{P(A \cap B)}$. (*B.A., Hons., 1971*)

[**Ans.** (*i*) $1-p_3$, (*ii*) p_2-p_3, (*iii*) $1-p_1+p_3$, (*iv*) $1-p_1-p_2+p_3$, (*v*) p_1-p_3

(*vi*) $\frac{1-p_1-p_2+p_3}{1-p_2}$, (*vii*) $1-p_1-p_2+p_3$, (*viii*) $1-p_3$]

(*b*) In terms of probabilities $p_1=P(A)$, $p_2=P(B)$, $p_3=P(AB)$, express

(*i*) $P(A+B)$, (*ii*) $P(A \mid B)$, (*iii*) $P(\bar{A}B)$, (*ir*) $P[\bar{A}\cap A\cup B]$,

(*v*) $P[A\cup(\bar{A}\cap B)]$, (*vi*) Condition that A and B are mutually exclusive,

(*vii*) Condition that A and B are mutually independent.

(*B.Sc. Meerut, 1970 ; B.A. Hons. Delhi, 1969*)

[**Ans.** (*i*) $p_1+p_2-p_3$, (*ii*) $\frac{p_3}{p_2}$, (*iii*) p_2-p_3

(*iv*) p_2-p_3 (*v*) $p_1+p_2-p_3$]

4. (*a*) If A and B are two events and the probability $P(B) \neq 1$, prove that

$$P(A \mid \bar{B}) = \frac{P(A)-P(A\cap B)}{1-P(B)}$$

and hence deduce that

$$P(A\cap B) \geqslant P(A)+P(B)-1.$$

Also show that $P(A) >$ or $< P(A \mid B)$ according as $P(A \mid \bar{B}) >$ or $< P(A)$.

(*M.Sc. Bombay, 1969 ; M.Sc. Gauhati, 1967 ; B.Sc., Sardar Patel, 1968*)

(*b*) If A and B are mutually exclusive and $P(A\cup B)\neq 0$, then prove that

$$P(A \mid A\cup B) = \frac{P(A)}{P(A)+P(B)}$$

(*c*) If $A\cap B\cap C=\varphi$, then

$$P(B\cup C \mid A)=P(B \mid A)+P(C \mid A)$$

(*d*) Prove : (*i*) $P(A\cap\bar{B} \mid C)=P(A \mid C)-P(A\cap B \mid C)$

(*ii*) $P(\bar{A}_1\cup A_2 \mid \bar{A}_3)=1-P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2)$.

(*B.Sc. Bombay, 1968*)

[**Hint.** (*a*) We have $P(A \mid \bar{B})=[P(A)-P(AB)]/[1-P(B)]$ ...(*i*)

Since $P(A \mid \bar{B})\leqslant 1$, we have

$$P(A)-P(AB)\leqslant 1-P(B)$$

$\Rightarrow$ $P(AB)\geqslant P(A)+P(B)-1$ ...(*ii*)

Clearly we have

$$\frac{P(A \mid \bar{B})}{P(A)} = \frac{P(\bar{B} \mid A)}{P(\bar{B})} = \frac{1-P(B \mid A)}{1-P(B)}$$

Hence if *i.e.*, if *i.e.*, if

$$\left.\begin{aligned} P(A \mid \bar{B}) &> P(A) \\ 1-P(B \mid A) &> 1-P(B) \\ P(B \mid A) &< P(B) \\ P(A \mid B) &< P(A) \end{aligned}\right\} \quad \text{...(iii)}$$

since $\frac{P(A \mid B)}{P(A)} = \frac{P(B \mid A)}{P(B)}$. Hence the result.]

5. If, A, B and C are three arbitrary events and we denote by

$$S_1=P(A)+P(B)+P(C)$$
$$S_2=P(A\cap B)+P(B\cap C)+P(C\cap A)$$
$$S_3=P(A\cap B\cap C)$$

prove that exactly one of the three events occurs is given by $S_1-2S_2+3S_3$.

(*M.A. Statistics, Delhi, 1965*)

6. It is 8 : 5 against a person who is now 40 years old living till he is 70 and 4 : 3 against a person now 50 living till he is 70. Find the probability that

one at least of these persons will be alive 30 years hence.

[Ans. Required probability $=1-\frac{8}{13}\times\frac{4}{7}=\frac{52}{91}$]

7. A and B are two independent witnesses (*i.e.*, there is no collusion between them) in a case. The probability that A will speak the truth is x, and the probability that B will speak the truth is y, A and B agree in a certain statement. Show that the probability that this statement is true is $\frac{xy}{1-x-y+2xy}$.

(*B.Sc., Lucknow, 1949*)

8. (*a*) Let p be the probability that a man aged x years will die in a year. Find the probability that out of n men, $A_1, A_2, \ldots\ldots, A_n$, each aged x, A_1 will die and be the first to die. [*M.A. Punjab, 1968 ; I A .S. 1954*)

$$\left[\text{Ans. } \frac{1}{n}\ [1-(1-p)^n]\ \right]$$

(*b*) Two dice are thrown. Let A and B be two events described as follows :

A : the event that sum of faces is odd,

B : the event that at least one ace occurs.

Find the probabilities of the events (*i*) BA (*ii*) $A+B$ (*iii*) $A\bar{B}$.

(*B.Sc., Bombay, 1967*)

$$\left[\text{Ans. } (i)\ \frac{6}{36}\ (ii)\ \frac{23}{36}\ (iii)\ \frac{5}{36}\right]$$

9. A and B throw alternately with a pair of dice. A wins if he throws 6 before B throws 7, and B if he throws 7 before A throws 6. If A begins, show that his chance of winning is $\frac{30}{61}$. (Huyghen's Problem)

(*M.Sc., Agra, 1945 ; B.A., Hons., Delhi, 1956, 1960*)

10. Consider families with exactly two children. Let b and g stand for a boy and a girl respectively and first letter for the older child, we have four possibilities bb, bg, gb and gg. These are the four sample points, and we associate probability 1/4 with each,

(*a*) Given that the family has a boy, what is the probability that the other one is also a boy,

(*b*) Given that the family has the older child as a boy, what is the probability that the other one is also a boy.

[Ans. (*i*) $\frac{1}{3}$, (*ii*) $\frac{1}{2}$]

11. Three groups of children contain respectively 3 girls and 1 boy : 2 girls and 2 boys, 1 girl and 3 boys. One child is selected at random from each group. Show that the chance that the three selected consists of 1 girl and 2 boys is $\frac{13}{32}$.

(*Certificate Exam., Calcutta, 1951 ; M.Sc., Agra, 1955, 1959 ; B.Sc., Agra, 1958*)

12. Out of $(2n+1)$ tickets consecutively numbered, three are drawn at random. Find the chance that the numbers on them are in (*i*) Arithmetical Progression. (*ii*) Geometrical Progression if number of tickets is 100.

(*B.Sc., Delhi, 1973, 75 ; M.Sc., Agra, 1950 ; B.Sc., Agra, 1959*)

$$\left[\text{Ans. } (i)\ \frac{3n}{4n^2-1},\ (ii)\ \frac{53}{100\times49\times33}\right]$$

13. (*a*) A player tosses a coin and is to score one point for every head turned up and two for every tail. He is to play on until his score reaches or

passes n. If p_n is the chance for attaining exactly n, show that $p_n=\frac{1}{2}(p_{n-1}+p_{n-2})$ and hence find the value of p_n. *(B.A., Punjab, 1959)*

Hint. $p_n-p_{n-1}=-\frac{1}{2}(p_{n-1}-p_{n-2})$.

$$\left[\text{Ans. } \frac{1}{3}\left\{2+(-1)^n\frac{1}{2^n}\right\}\right]$$

(b) The chance of success in a single trial is $\frac{1}{3}$. If p_n is the probability that there are no two consecutive successes is n trials, prove that

$$p_n=\frac{2}{3}p_{n-1}+\frac{2}{3}p_{n-2}.$$

(B.Sc., Bombay, 1969)

(c) There are n similar biased dice such that the probability of obtaining a 6 with each one of them is the same and equal to p. If all the dice are rolled once, show that p_n, the probability that an odd number of 6's is obtained satisfies the difference equation

$$p_n+(2p-1)p_{n-1}=p$$

and hence derive an explicit expression for p_n.

Hint. $p_n-\frac{1}{2}=(1-2p)[p_{n-1}-\frac{1}{2}]$. [Ans. $p_n=\frac{1}{2}[1+(1-2p)^n]$

14. Each of the n urns contains a white and b black balls. One ball is transferred from the first urn into the second, then one ball from the latter into the third and so on. Finally one ball is taken from the last urn, what is the probability of its being white ? *(M.Sc., Punjab, 1970)*

$$\left[\text{Ans. } p_n=\frac{a}{a+b}\right]$$

[**Hint.** If p_k be the probability of drawing a white ball from the kth urn then

$$p_k=\frac{a+1}{a+b+1}p_{k-1}+\frac{a}{a+b+1}(1-p_{k-1})=\frac{1}{a+b+1}p_{k-1}+\frac{a}{a+b+1}$$

Putting $k=2$ and using $p_1=\frac{a}{a+b}$, we get $p_2=\frac{a}{a+b}$. Proceeding in this manner and using the method of induction, we conclude that $p_n=\frac{a}{a+b}$].

15. *(a)* Four different objects 1, 2, 3, 4 are distributed at random on four places marked 1, 2, 3, 4. What is the probability that none of objects occupies the place corresponding to its number ?

(b) A person writes four letters and four envelopes. If the letters are placed in the envelopes at random, what is the chance that not more than one letter is placed in the correct envelope ?

[**Hint.** *(a)* See solved example 4·5.] [Ans. *(a)* $\frac{3}{8}$ *(b)* $\frac{17}{24}$]

16. *(a)* Prove that the following theorem in probability for the events E_1, E_2, and E_3 :

(i) $P(\bar{E}_1\bar{E}_2)=1+P(E_1E_2)-P(E_1)-P(E_2)$. *(B.A., Punjab, April, 1954)*

(ii) $P(\bar{E}_1+\bar{E}_2+\bar{E}_3)=1-P(E_1)P(E_2 \mid E_1)P(E_3 \mid E_1E_2)$,

where $P(\bar{E})$ is the probability that E does not occur. *(B.A., Punjab Sept., 1954)*

(b) If $A+B$ means the events A and $\bar{A}B$ and if $A-B$ means the event AB prove that

(*i*) $P[(A-AB)+B]=P(A+B)$ and

(*ii*) $\overline{P(A+B+C)}=P(\bar{A}\bar{B}\bar{C})$ (*M.A., Statistics, Delhi, 1960*)

(*c*) For any two events A and B, find whether the following statements are true or not:

(*i*) $P(A \mid B)+P(\bar{A} \mid B)=1$

(*ii*) $P(A \mid B)+P(A \mid \bar{B})=1$. (*B.A., Hons., Delhi 1966*]

[**Ans.** (*i*) Correct (*ii*) Incorrect]

(*d*) Examine whether the following relations are correct.

(*i*) $(A \cup B)-C=A \cup (B-C)$

(*ii*) $\overline{(A \cup B \cup C)}=A'B'C'$ (*M.A., Statistics, Delhi 1969*)

[**Ans.** (*i*) incorrect (*ii*) correct)

17. What is the probability that at least one of the players in a bridge game will *get* a complete suit of cards ?

(*M.A., Statistics, Delhi 1965* ; *I.C.A.R., Delhi 1974* ; *M.A. Delhi, 1959*)

[**Hint.** Required probability $=P(A_1+A_2+A_3+A_4)=\Sigma p_i-\Sigma\Sigma p_{ij}+\Sigma\Sigma\Sigma p_{ijk}-p_{1234}$

$$=\frac{4c_1\times 4}{52c_{13}}-\frac{4c_2\times 4\times 3}{52c_{13}\times 39c_{13}}+\frac{4c_3\times 4\times 3\times 2}{52c_{13}\times 39c_{13}\times 26c_{13}}-\frac{4c_4\times 4\times 3\times 2\times 1}{52c_{13}\times 39c_{13}\times 26c_{13}\times 13c_{13}}$$

$$=\frac{16\,.\,13\,!\,39\,!-72(13\,!)^2\,26\,!+72(13\,!)^4}{52\,!}\Bigg]$$

18. In a lottery m tickets are drawn at a time out of the total number of n tickets and returned before the next drawing is made. Show that the probability that in k drawings each of the numbers 1, 2, 3, ..., n will appear at least once is given by

$$p_k=1-{}^nC_1\left(1-\frac{m}{n}\right)^k+{}^nC_2\left(1-\frac{m}{n}\right)^k\left(1-\frac{m}{n-1}\right)^k-\ldots$$

(*M.Sc., Bombay, 1960* ; *B.A., Hons., Delhi, 1962* ; *B.A., Hons., Calcutta, 1965* ; *M.A., Stat. Delhi, 1955* ; *M.A., Patna, 1956*)

[**Hint.** $p_k=1-P$(ticket numbered i does not appear in any one of the drawings.]

19. We have k varieties of objects each variety consisting of the same number of objects. These objects are drawn one at a time and replaced before the next drawing. Show that the probability, p_n, that n and no less drawings will be required to produce objects of all varieties is given by

$$p_n=\sum_{0}^{k-1}(-)^r\binom{k-r}{r}\left(\frac{k-1-r}{k}\right)^{n-1}$$

i.e., $$k^{n-1}p_n=(k-1)^{n-1}+{}^{k-1}c_1(k-2)^{n-1}+{}^{k-1}c_2(k-3)^{n-1}-\ldots\ldots$$

(*M.A. Stat. Delhi, 1962*)

20. Cards are dealt one by one from a well shuffled pack until an ace appears. Show that the probability that exactly n cards are dealt *before* the first ace, is $\dfrac{(51-n)(50-n)(49-n)}{13.51.50.49}$.

(*B.A., Hons., Delhi, 1955, 1971* ; *M.A. Delhi, 1954* ; *I.C.A.R., Delhi, 1950*)

If cards continue to be dealt until a second ace appears, prove that the probability that exactly r cards are dealt in all *before* the second ace, is

$$\frac{r(51-r)(50-r)}{13\cdot 17\cdot 50\cdot 49}$$ (*I.C.A.R., Delhi, 1950*)

[**Hint.** (*i*) Reqd. prob.$=\dfrac{4}{52-n}\times\dfrac{{}^{48}c_n}{{}^{52}c_n}$. (*ii*) Reqd. prob.$=\dfrac{{}^4c_1\times{}^{48}c_{r-1}}{{}^{52}c_r}\times\dfrac{3}{52-r}$]

(*b*) Cards are dealt one by one from an ordinary pack (without re placement) until two aces have appeared. Find the most probable number of cards to be turned up. [**Ans.** 18]

[**Hint.** The most probable number to be turned up will be the value of x which gives the chance of success at the xth trial *viz.*,

$$\frac{(x-1)(52-x)(51-x).4.3}{52.51.50.49}$$

its greatest value. Hence etc.]

21. Consider n independent tosses of a coin which has probability p of falling heads. Let S be the number of tosses in which the coin falls head. Find the probability that S is even. [**Ans.** $\frac{1}{2}\{1+(q-p)^n\}$]

22. (*a*) Show that the chance of throwing 15 with 3 dice is $\frac{5}{108}$.

(*b*) A and B throw with 3 dice ; if A throws 8, what is B's chance of throwing a higher number.

(*c*) If $6n$ tickets numbered 0, 1, 2,..... , $6n-1$ are placed in a bag and three are drawn out, find the chance that the sum of the numbers is equal to $6n$.

(*d*) The probability of occurrence of an event A is 0·7, the probability of non-occurrence of B is 0·5 and that of at least one of A and B not occurring is 0·6. Find the probability that at least one of A and B occurs.

[**Hint.** (*a*) The required probability is the coefficient of x^{15} in the expansion of

$$\frac{1}{16}(x^1+x^2+x^3+x^4+x^5+x^6)^3,$$

that is, in $\frac{1}{6^3}\cdot\frac{x^3(1-x^6)^3}{(1-x)^3}$. Hence etc.]

[**Ans.** (*b*) $\frac{7}{27}$. (*c*) $3n/(6n-1)(6n-2)$. (*d*) 0·8]

(*d*) $P(A)=0{\cdot}7$, $P(\bar{B})=0{\cdot}5$, P (at least one of A and B not occurs) $=P(\overline{A+B})=1-P(A+B)$ so that $P(AB)=0{\cdot}4$. Hence P (at least one of A and B occurs) $=P(A+B)=P(A)+P(B)-P(AB)=$ etc.

23. An urn contains a white and b black balls. Balls are drawn one by one until only those of the same colour are left. What is the probability that they are white ? (*M.A. Banaras, 1949 ; M.A., Punjab, 1960*)

[**Sol.** Reqd. prob. $=\left(\sum_{x=0}^{a-1} {}^{b+x-1}C_{b-1}\right)\div {}^{a+b}C_b=\sum_{0}^{a-1}\left({}^{b+x-1}C_x\right)\div {}^{a+b}C_b$

$$={}^{a+b-1}C_{a-1}\div {}^{a+b}C_b=\frac{a}{a+b}.$$

For ${}^{b-1}c_0+{}^{b}c_1+{}^{b+1}c_2+.....+{}^{b+a-2}C_{a-1}$

$=$ term independent of x in $\left[(1+x)^{b-1}+\frac{(1+x)^b}{x}+\frac{(1+x)^{b+1}}{x^2}+...\right.$

$$\left......+\frac{(1+x)^{b+a-2}}{x^{a-1}}\right]$$

$=$,, ,, ,, in $(1+x)^{b-1}\left[\left(\frac{1+x}{x}\right)^a-1\right]\div\left[\left(\frac{1+x}{x}\right)-1\right]$

$=$,, ,, ,, in $(1+x)^{a+b-1}\div x^{a-1}$

$={}^{a+b-1}C_{a-1}={}^{a+b-1}C_b$.]

24. Of three independent events, the probability that the first only should happen is $\frac{1}{4}$; the probability that the second should happen is $\frac{1}{8}$, and the probability that the third only should happen is $\frac{1}{12}$. Obtain the unconditional probability of the three events. *(B.A. Hons. Delhi, 1969)*

[Ans. $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$]

25. A coin is tossed $(m+n)$ times $(m>n)$. Show that the probability of getting m consecutive heads is

$$\frac{n+2}{2^{m+1}}.$$

(M.A., Madras, 1969 ; M.A., Banaras, 1957 ; Indian Statistical Institute, Calcutta, 1948)

[**Hint.** The chance that the event has already occurred at the end of $(m+n-1)$ tosses is p_{n-1} and is only increased with an additional toss by the fact that an incomplete run of $(m-1)$ heads preceded by a tail may be followed by a head. The chance for this is $1/2^{m+1}$. We have $p_n=p_{n-1}+1/2^{m+1}$. Write $n-1$, $n-2$,....., 1 for n and add using $p_0=1/2^m$]

26. An urn contains N balls, pN of which are white, the rest black. From it are drawn n_1 balls r_1 of which are found to be white. The balls are then replaced, and n_2 balls are now extracted among which r_2 are white balls. Show that the most probable value of p for such asample is $p=(r_1+r_2)\div(n_1+n_2)$. *(I.C.A.R., Delhi, 1948)*

[**Hint.** Maximize the expression

$$f(p)={}^{n_1}C_{r_1}\, p^{r_1}(1-p)^{n_1-r_1} \times {}^{n_2}C_{r_2}(1-p)^{n_2-r_2}\ ]$$

27. An urn contains N identical balls, black and white in unknown proportion, there being at least one white and one black ball. A ball is drawn out and replaced x times, the balls being mixed after each drawing with the result that first r white balls are seen. If all compositions of the population are equally likely, prove that the most probable value of R, the number of white balls in the urn is given by $R/N=r/n$. *(M.A., Bombay, 1954)*

28. (*a*) What is the probability that two integers chosen at random will be prime to each other ?

(*b*) Four possible integers are chosen at random. Find the probability of their having a common factor.

(*c*) A sportsman's chance of shooting an animal at a distance $r(>a)$ is a^2/r^2. He fires when $r=2a$, and if he misses he reloads and fires when $r=3a$, $4a$,..... If he misses at distance na, the animal escapes. What are the odds against the sportsman ? *(M.A., Patna 1957)*

[**Hint.** (*a*) Let a and b be any two integers and r a prime number. The possible remainders are 0, 1, 2,...,$r-1$, then the chance of a being divisible by r is $1/r$ and same is that for b. The integers being chosen at random, the probability of both a and b having r as a factor is $1/r^2$ and therefore not being divisible by r is $1-\frac{1}{r^2}$. Reqd. prob. $=\prod\left(1-\frac{1}{r^2}\right)=6/\pi^2$ where r takes every prime value 2, 3, 5,.....

(*b*) Use the formula $\prod\left(1-\frac{1}{p^4}\right)=\frac{90}{\pi^4}$,

where p denotes every prime number 2, 3, 5,.....].

(*c*) The chance of escape of the animal in the end $= \prod_{2}^{n}\left(1-\frac{a^2}{i^2a^2}\right)$

$$= \prod_{2}^{n}\left(1-\frac{1}{i^2}\right) - \prod_{i=r}\frac{i-1}{i}\cdot\prod_{i=2}^{n}\frac{i+1}{i} = \frac{1}{2}\cdot\frac{n+1}{n} = \frac{n+1}{2n}.$$ [Ans. $n+1 : n-1$]

29. From a pack of 52 cards an even number of cards is drawn. Show that the probability of half of these cards being red is $\left[\frac{52!}{(26!)^2} - 1\right] \Big/ (2^{51}-1)$

(*M. Sc. Meerut, 1968*)

30. (*a*) If each of n independent trials can result in one of the k incompatible events $E_1, E_2, \ldots\ldots, E_k$ with respective probabilities $p_1, p_2, \ldots\ldots, p_k$; $(p_1+p_2+\ldots\ldots+p_k=1)$ show the probabilities to have n_1 events E_1, n_2 events $E_2, \ldots\ldots$, n_k events E_k where $n_1+n_2+\ldots\ldots+n_k=n$, is given by

$$P_{n_1, n_2, \ldots\ldots, n_k} = \frac{n!}{n_1!\, n_2!\, \ldots\ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots\ldots p_k^{n_k}.$$

(*b*) A bag contains 5 white, 7 green, 12 red and 14 black balls. The balls are indistinguishable from each other except by colour. A ball is drawn and replaced after its colour has been noted, on ten occasions. If any ball is likely to be drawn as any other, show that the probability that of the ten balls seen 3 will be white, 3 green, 2 red and 2 black is

$$\frac{10!}{3!\,3!\,2!\,2!}\left(\frac{5}{38}\right)^3\left(\frac{7}{38}\right)^3\left(\frac{12}{38}\right)^2\left(\frac{14}{38}\right)^2$$

31. A bag contains n balls, k drawings are made in succession, and the ball on each occasion is found to be white, find the chance that the next drawing will give a white ball when the balls are not replaced after each drawing.

(*M.A., Patna, 1954*)

$$\left[\text{Ans. Reqd prob.} = 1 - {}^kC_1\frac{(n-1)!}{n!} + {}^kC_2\frac{(n-2)!}{n!} + \ldots\ldots + \frac{(-1)^k(n-k)!}{n!}\right]$$

32. In a bolt factory machines A, B, C manufacture respectively 25, 35 and 40 per cent of the total. Of their output 5, 4 and 2 per cent are defective bolts. A bolt is drawn at random from the product and is found defective. What are the probabilities that it was manufactured by machines A, B, or C ?

(*M.A., Madras, 1960* ; *M.A., Statistics, Delhi, 1959, 66*)

$$\left[\text{Ans. } \frac{125}{345}, \frac{140}{345}, \frac{180}{345}\right]$$

33. The probability of the occurrence of two independent events is known to be each greater than $\frac{1}{2}$. It is given that probability of the first event occurring simultaneously with the second not occurring is 3/25. Also the probability of the second occurring simultaneously with the first not occurring is 8/25. Find the probabilities of the respective events.

[**Hint.** $P(A_1 \cap \overline{A_2}) = 3/25$, $P(\overline{A_1} \cap A_2) = 8/25$ $\therefore$ $P(A_1) - P(A_1)P(A_2) = 3/25$, $P(A_2) - P(A_1)P(A_2) = 8/25$, solve for $P(A_1)$ and $P(A_2)$].

34. (*a*) Three newspapers, A, B, C, are published in a certain city. It is estimated from a survey that of the adult population :

20% read A, 16% read B, 14% read C, 8% read both A and B, 5% read both A and C, 4% read both B and C, 2% read all three. What percentage

reads at least one of the papers ? Of those that read at least one, what percentage reads both A and B ? *(M.A., Delhi, 1955, 68)*

[Ans. (*i*) 35% (*ii*) 28%]

(*b*) In a bolt factory machines A, B and C manufacture respectively 25%, 35% and 40%. Of the total of their output 5, 4, 2 percents are defective bolts, A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by machine A, B or C. (Use Baye's theorem). *(M.A., Delhi, 1969, 68 ; B.Sc. Hons. Cal. 1968)*

$$\left[\text{Ans. } \frac{23}{69}, \frac{28}{69}, \frac{16}{69}\right]$$

35. There are two urns. In the first urn there are 3 white and 2 black balls and in the second 1 white and 4 black. From any arbitrarily chosen urn, one ball is taken. What is the probability for its being white if probability of choosing any ball from both the urns is 0·5 each ? [Ans. 0·4]

36. The probability that a family chosen at random has exactly n children is αp^n $(0 < p < 1)$. Suppose also that the chance of any child having blue eyes is a $(0 < a < 1)$, independently of the others. Prove that the probability of a family chosen at random having exactly k children with blue eyes is

$$\alpha p^k a^k/[1-p(1-a)]^{k+1},\ k \geqslant 1.$$

(M.A., Punjab, 1953)

37. Let the probability p_n that a family has exactly n children be αp^n, $n \geqslant 1$ and $p_0 = 1-\alpha p(1+p+p^2+\ldots\ldots)$. Suppose that all sex distributions of n children have the same probability. Show that probability of family containing k boys is

$$2\alpha p^k/(2-p)^{k+1},\ k \geqslant 1.$$

(M.A., Statistics, Delhi, 1964 ; B.A., Hons., Calcutta, 1954)

Given that a family includes at least one boy, show that the probability that there are two or more boys is $p/(2-p)$.

38. A man is given n keys of which one fits the lock. He tries them successively without replacement. Find the probability that at the rth trial the lock will open. Also determine the expected number of trials and their variance (*i*) if successful keys are not eliminated from further selection and (*ii*) if they are.

(M.A., Statistics, Delhi, 1964, 68)

[Ans. (*a*) $\frac{1}{n}$ or $\left(1-\frac{1}{n}\right)^{r-1}\left(\frac{1}{n}\right)$

(*b*) (*i*) $\mu = n$, $\sigma^2 = n(n-1)$, (*ii*) $\mu = \frac{1}{2}(n+1)$, $\sigma^2 = \frac{1}{12}(n^2-1)$]

39. (*a*) An event A is known to be stochastically independent of the events B, $B \cup C$, and $B \cap C$ show that it is also stochastically independent of C. *(B.Sc., Delhi, 1971)*

(*b*) A and B are independent events both of which are neither certain nor possible. Prove that the opposite events $\bar{A}$ and $\bar{B}$ are independent. *(B.Sc., Delhi, 1961)*

40. (*a*) If A and B are independent events, prove that (*i*) A and $\bar{B}$ are independent, (*ii*) $\bar{A}$ and B are independent and (*iii*) $\bar{A}$ and $\bar{B}$ are independent.

[Hint. (*iii*) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1-P(A \cup B) = 1-[P(A)]+P(B)$
$-P(A \cap B)] = [1-P(A)][1-P(B)] = P(\bar{A})P(\bar{B})]$

(B.Sc., Poona, 1969 ; B.Sc., Bombay, 1967)

(*b*) A, B and C are three events such that A and B are independent and $P(C)=0$. Show that A, B and C are independent.

41. The chance of success in each trial is p. If p_k is the probability that there are even number of success in k trials, prove that

$$p_k=p+p_{k-1}(1-2p)$$

Deduce that $p_k=\frac{1}{2}[1+(1-2p)^n]$. (*B.Sc., Bombay, 1970*)

[**Hint.** The event can happen in the following mutually exclusive ways :

(*i*) in $(n-1)$ trials, even number of successes occur and then a failure with probability $(1-p)$ at the kth trial.

(*ii*) in $(n-1)$ trials, odd number of successes and then a success at the nth trial.

Hence $$p_k=(1-p)p_{k-1}+(1-p_{k-1})p$$
$$=p+(1-2p)p_{k-1}.$$

Also $$p_k-\tfrac{1}{2}=(1-2p)[p_{k-1}-\tfrac{1}{2}].$$

42. A group of $2N$ boys and $2N$ girls is divided into two equal groups. Find the probability 'p' that each group will be equally divided into boys and girls.

Estimate 'p' using Stirling's formula. (*M.A. Stat. Delhi, 1963*)

$\left[\text{Use } n! \simeq \sqrt{2\pi}\, e^{-n}\, n^{n+\frac{1}{2}}\right]$ $\left[\text{Ans. } \frac{(2N!)^4}{(N!)^4(4N!)} \cong \sqrt{\frac{2}{\pi N}}\right]$

43. (*a*) What is the probability that at least two out of n people have the same birth date ? (*B.Sc., Delhi, 1971*)

(*b*) What is the probability that

(*i*) the birthdays of twelve people will fall in twelve different calendar months (assume equal probabilities for the 12 months) ?

(*ii*) the birthdays of six people will fall in exactly two calendar months ?

(*B.Sc., Poona, 1965*)

$$\left[\text{Ans. } (a)\ 1-\left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\cdots\cdots\left(1-\frac{n-1}{365}\right)\right.$$

$$\left.(b)\ (i)\ \frac{12}{12^{12}} \quad (ii)\ \frac{\binom{12}{2}\left(2^6-2\right)}{12^6}\right]$$

44. A and B have equal chances of winning a single game. A wants n games and B $n+1$ games to win a match. Show that the odds in favour of A are $1+P$ to $1-P$ where

$$P=\frac{2n!}{(n!)^2\,2^{2n}}.$$

(*M.Sc., Kerala, 1969*)

[**Hint.** Probability of A winning at least n games out of $2n$ games

$=^{2n}c_n q^n p^n+^{2n}c_{n+1}[q^{n-1}p^{n+1}+\ldots+^{2n}c_{2n}p^{2n}$ where $p=q=\frac{1}{2}$

$=\frac{1}{2^{2n}}[\frac{1}{2}\{(1+1)^{2n}-^{2n}c_n)\}]$=etc.]

45. (*a*) Each coefficient of an equation $Ax^2+Bx+C=0$ is determined by the score thrown up in a throw of an unbiased six-faced die having numbers

from 1 to 6. Obtain the probability that the equation has (*i*) equal and (*ii*) real roots. (*M.A., Poona, 1965*)

[Ans. (*i*) $\frac{5}{216}$ (*ii*) $\frac{43}{216}$]

(*b*) The variates a and b are uniformly distributed in the intervals $0 \leqslant a \leqslant 6$ and $0 \leqslant b \leqslant 9$. Find the probability that the equation $x^2-ax+b=0$ has two real roots. [Ans. $\frac{29}{70}$]

(*c*) The variates a and b are uniformly distributed over the interval $(-1, 1)$. Find the probability that the equation $x^2+2ax+b=0$ has equal roots.

[Ans. $\frac{3}{9}=\frac{1}{3}$]

(*d*) If the random variable Y is uniformly distributed over (0, 5), what is the probability that the roots of the equation $4x^2+4xY+Y+2=0$ are real ?

(*M.A. Old Course Delhi 1974*)

47. In a bridge game North and South have ten spades between them. Find the probability that either East and West has no spade (there are 13 spades in pack of 52 cards and a player has 13 cards. The four players are designated by the positions they occupy, *viz.* North, South, East and West).

(*B.Sc., Hons., Calcutta, 1968*)

[Ans. $\frac{11}{50}$, $\frac{13}{50}$]

48. Find the probability for a poker hand *i.e.* a hand of five cards taken at random from a full pack of 52 cards to be a (*i*) royal flush, *i.e.* ten, jack, queen, king and ace in a single suit, (*ii*) full house *i.e.* one pair and one triple of cards with equal face values, (*iii*) straight, *i.e.* five cards in a sequence regardless of suit (*iv*) four of a kind. (*M.Sc., Gauhati, 1963*)

[Ans. Let $\overset{52}{C}_5=q$, then (*i*) $\frac{4}{q}$ (*ii*) $13 \cdot 12 . \frac{4}{q}$

(*iii*) $\frac{9 \cdot 4^5}{q}$ (*iv*) $13 \cdot 12 . \frac{4}{q}$]

49. A and B engage themselves in a game each step of which consists in one of them winning a counter from the other. At beginning, A has a counters and B has b counters and in each successive step the probability of A winning a counter from B is p. The game is terminated when either of the two has all $a+b$ counters what is the probability of A's winning when he has n counters.

(*a*) Consider the two cases (*i*) $p \neq \frac{1}{2}$ (*ii*) $p=\frac{1}{2}$ (*M.Sc., Bombay, 1958*)

(*b*) Consider the two cases (*i*) $p \neq q$ (*ii*) $p=q$

(*M.Sc., Gauhati, 1953* ; *M.Sc., Mysore, 1965*)

50. Let A B and C be any three events. If $P(A \cap \bar{B} \cap \bar{C})=\frac{1}{4}$, $P(A \cap B \cap C)=\frac{1}{4}$, $P(A \cap B \cap C)=\frac{1}{4}$ and $P(A \cap B \cap C)=\frac{1}{4}$. Verify whether A, B and C are (*i*) pairwise independent (*ii*) completely independent.

[**Hint.** Use Venn diagram.] (*B.Sc., Hons., Calcutta, 1967*)

5

Random Variables, Mathematical Expectation, Moment Generating Function and Cumulative Function

5·1·1. Random Variables. Let a real number X be associated with the outcome of a random experiment E. Thus if E be the random experiment consisting of two tosses of a coin, then X, the number of heads, will take the values 0, 1 or 2. Presenting the outcomes and the values in a tabular form, we have

Outcome (w) :	TT	TH	HT	HH
Values of X (w) :	0	1	1	2

Thus it is clear that to each outcome w of E, there corresponds a real number $X(w)$. Further since the sample points correspond to outcomes, we conclude that $X(w)$ is defined for each $w \in S$, the sample space and thus a random variable is defined to a real function on S. Explicitly, we define a one-dimensional random variable as a real-valued function defined on sample space S which is associated with a given random experiment. Further, if w be the outcome and $X(w)$ be the real-valued function and range $(-\infty, +\infty)$ with domain S, then for every real number a, the event is $[w : X(w) \leqslant a]$. We say that $P(X \leqslant a)$ is the probability of the set of outcomes w for which $X(w) \leqslant a$, *i.e.*

$$P(X \leqslant a) = P[w : X(w) \leqslant a]$$

Notations. Let x represent a real number, then the set of all w in S such that $X(w) = x$ is denoted by writing $X = x$, so that

$$P(X = x) = P[w : X(w) = x]$$

Similarly
$$P[X \leqslant a] = P\{w : X(w) \in [-\infty, a]\}$$
$$P[a < X \leqslant b] = P\{(w : X(w) \in [a, b]\}$$
$$P(X = a \text{ or } X = b) = P\{X = a \cup X = b\}$$
$$P(X = a, X = b) = P[X = a \cap X = b\}$$

Example. *Consider the tossing of a coin so that our sample space consists of two outcomes either head or tail. Thus*

$$S=\{w_1, w_2\} \quad \text{where } w_1=\text{Head}, w_2=\text{Tail}$$

and

$$X(w)=\begin{cases}1 & \text{if } w=\text{Head appears}\\ 0 & \text{if } w=\text{Tail appears}\end{cases}$$

Note. Here the random variable $X(w)$ takes only two values as w can be either head or tail. Such a random variable is known as Bernoulli random variable (r. v.).

Remark. If X_1 and X_2 are random variables and (*r. v.*) c is a constant, then

(*i*) cX_1 is a r.v.

(*ii*) X_1+X_2 is a r.v.

(*iii*) X_1-X_2 is a r.v.

(*iv*) max. $[X_1, X_2]$ is a r.v.

(*v*) min. $[X_1, X_2]$ is a r.v.

Ex. 1. If X is a one-dimensional r.v. and $a<b$, show that

$$P(a<X\leqslant b)=P(X\leqslant b)-P(X\leqslant a).$$

Ex. 2. Two honest die are rolled, each outcome being an ordered pair (a, b) where each of a and b is an integer from 1 to 6. Let X be the value $a+b$ assigned to the outcome (a, b). Determine the probabilities of events $X=5$, $X=7$.

Ex. 3. Let the sample space S consists of three elements w_1, w_2 and w_3. Let $P(w_1)=\frac{1}{4}$, $P(w_2)=\frac{1}{2}$ and $P(w_3)=\frac{1}{4}$. If X is a r.v. defined on S by $X(w_1)=10$, $X(w_2)=-3$, $X(w_3)=15$, find $P(-2\leqslant X\leqslant 2)$. (*M.Sc., Banaras, 1969*)

Hint. $P(a\leqslant X\leqslant b)=P(X=a)+P(a<X\leqslant b)=P(X=a)+P(X\leqslant b)-P(X\leqslant a)$.

5·1·2. Distribution Function. Let X be a r.v. Then the function $F_X(x)$ defined as

$$F_X(x)=P(X\leqslant x)=P\{w : X(w)\leqslant x\}, \quad -\infty<x<\infty$$

is called the distribution function of X. It will be convenient to denote the distribution function (d.f.) by $F(x)$ or by F.

Important Properties of distribution function

I. If F is the d.f. of the r.v. X and if $a<b$, then

$$P(a<X\leqslant b)=F(b)-F(a)$$

Proof. Clearly, the events '$a<X\leqslant b$' and '$X\leqslant a$' are disjoint.

Also $\quad 'a<X\leqslant b' \cup 'X\leqslant a'=X\leqslant b$

$\therefore \quad P(a<X\leqslant b)+P(X\leqslant a)=P(X\leqslant b)$

$\therefore \quad P(a<X\leqslant b)=P(X\leqslant b)-P(X\leqslant a)=F(b)-F(a)$

Cor. $\quad P(a<X<b)=P(a<X\leqslant b)-P(X=b)$

$$=F(b)-F(a)-P(X=b)$$

Similarly, $P(a\leqslant x\leqslant b)=F(b)-F(a)+P(X=a)$

$$P(a\leqslant X<b)=F(b)-F(a)+P(X=a)-P(X=b)$$

II. If F is the d.f. of the r.v. X, then

(*i*) $0 \leqslant F(x) \leqslant 1$

(*ii*) $F(x) \leqslant F(y)$ if $x < y$

i.e. F is a monotonically non-decreasing function.

Proof. (*i*) Since $0 \leqslant P(X \leqslant x) \leqslant 1$, the result (*i*) follows

(*ii*) We have

$$F(y) - F(x) = P(x < X \leqslant y) \geqslant 0$$

Hence $F(y) \geqslant F(x)$ when $x < y$

$\therefore$ It follows $F(x) \leqslant F(y)$ when $x < y$

III. If F is d.f. of a r.v. X, then

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

and

$$F(+\infty) = \lim_{x \to +\infty} F(x) = 1$$

Proof. Considering the sample space S as a countable union of disjoint events, we have

$$S = [\bigcup_{n=1}^{\infty} -n < x \leqslant -n+1] \cup [\bigcup_{n=0}^{\infty} n < X \leqslant n+1]$$

$$\therefore \quad P(S) = \sum_{n=1}^{\infty} P(-n < X \leqslant n+1) + \sum_{n=0}^{\infty} P(n < X \leqslant n+1)$$

$$\therefore \quad 1 = \lim_{a \to \infty} \sum_{n=1}^{a} [F(-n+1) - F(-n)] + \lim_{b \to \infty} \sum_{n=0}^{b} [F(n+1) - F(n)]$$

$$= \lim_{a \to \infty} [F(0) - F(-a)] + \lim_{b \to \infty} [F(b+1) - F(0)]$$

$$= F(0) - F(-\infty) + F(\infty) - F(0)$$

$$1 = F(\infty) - F(-\infty).$$

Also from property (2), it follows that

$$F(\infty) \leqslant 1 \quad \text{and} \quad F(-\infty) \geqslant 0$$

Hence $F(\infty) = 1$ and $F(-\infty) = 0$.

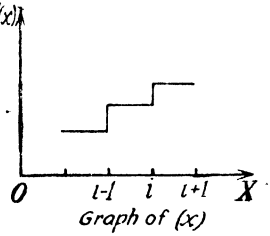
5·1·3. Discrete random variable. Probability distribution and discrete distribution function. Let S be the sample space associated with a given random experiment. If S be discrete, then a real-valued function defined on S is called a discrete random variable and is usually denoted by X.

Let X take at most a countably infinite number of values x_1, x_2, ...With each possible outcome, we associate a number $p_i = P(X = x_i) = p(x-1) \geqslant 0$ called the probability of x_i. The numbers p_i clearly satisfy the conditions

(*i*) $p(x_i) \geqslant 0 \ \forall i$ (*ii*) $\sum_{i=1}^{\infty} p(x_i) = 1$

The function p is called the probability mass function of the random variable X and the set $\{p(x_i)\}$ is called the *probability distribution of X*.

The function $F(x) = \sum_{i:\, x_i \leqslant x} p_i$, $\sum_{i=1}^{\infty} p_i = 1$, $p_i \geqslant 0$ graph of $f(x)$ is known as the discrete distribution function. However, if $x_i = i$ then $F(x)$ (integer), is called the 'step function with jump p_i at i'. Further, $F(x)$ will be constant between any two consecutive integers, *e.g.* between $i+k-1$, $i+k$; k and i being integers.



Graph of f(x)

Fig. 18

Example 5·1. *A random variable X has the following probability function :*

x :	-2	-1	0	1	2	3
$p(x)$:	$0{\cdot}1$	k	$0{\cdot}2$	$2k$	$0{\cdot}3$	$3k$

(*i*) *Find k* (*ii*) *Evaluate* $P(X<2)$, $P(X \geqslant 2)$, $P(-2<X<2)$ (*iii*) *Find the minimum value of k such that* $P(X \leqslant 1) > {\cdot}36$ (*iv*) *Determine the distribution function of X* (*v*) *Evaluate the mean of X.*

(*i*) Now $\sum_{x=-2}^{3} p(x) = 1$ gives

$$0{\cdot}1 + k + 0{\cdot}2 + 2k + 0{\cdot}3 + 3k = 1$$

Thus $$k = \frac{1}{15}.$$

(*ii*) $P(X<2) = P(X=-2) + P(X=-1) + P(X=0) + P(X=1)$

$$= 0{\cdot}1 + k + 0{\cdot}2 + 2k$$

$$= 3k + 0{\cdot}3 = 0{\cdot}5$$

$$P(X \geqslant 2) = P(X=2) + P(X=3)$$

$$= 0{\cdot}3 + 3k = 0{\cdot}3 + 0{\cdot}2 = 0{\cdot}5$$

Otherwise. $P(X \geqslant 2) = 1 - P(X<2) = 1 - 0{\cdot}5 = 0{\cdot}5$

$$P(-2<X<2) = P(X=-1) + P(X=0) + P(X=1)$$

$$= k + 0{\cdot}2 + 2k = 3k + 0{\cdot}2$$

$$= 0{\cdot}2 + 0{\cdot}2 = 0{\cdot}4.$$

(*iii*) $$P(X \leqslant 1) = P(X=-2) + P(X=-1) + P(X=0) + P(X=1)$$

$$= 0{\cdot}1 + k + 0{\cdot}2 + 2k = 3k + 0{\cdot}3$$

By hyp., $P(X\leqslant 1)>\cdot 36$

$\Rightarrow \quad 3k+0\cdot 3>\cdot 36 \quad \Rightarrow \quad 3k>\cdot 06$

$\Rightarrow \quad k=\cdot 02.$

(*iv*) $F(x)=0,\ x<-2$

$=\frac{1}{15},\ -2\leqslant x<-1$

$=\frac{14}{15},\ -1\leqslant x<3.$

(*v*) $E(x)=\Sigma xp(x)$

$=(-2)(0\cdot 1)+(-1)k+0(0\cdot 2)+1(2k)+2(0\cdot 3)+3(3k)$

$=10k+0\ 4=\frac{16}{15}.$

Example 5·2. *From a lot of 25 items containing 5 defectives, a sample of 4 items was drawn at random (i) without replacement (ii) with replacement. Find the probability distribution of X, the number of defectives.*

(*i*) We draw a sample of 4 items from a lot of 25 items and this can be done in $\binom{25}{4}$ ways which are equally likely. The number of ways in which out of 5, x defectives can be chosen and out of remaining 20 items, $4-x$ non-defectives can be chosen $4\binom{5}{x}\times\binom{20}{4-x}$. Consequently

$$P(X=x)=\binom{5}{x}\times\binom{20}{4-x}\div\binom{25}{4}.$$

(*ii*) When there is drawn a sample of 4 items with replacement, the number of ways $=\binom{4}{x}$.

Further the probability of choosing one defective out of 5 defectives is $\frac{1}{5}$ so that the probability of choosing x defectives out of 5 is $\left(\frac{1}{5}\right)^x$. The probability of not choosing a defective is $\frac{4}{5}$ so that the probability of not choosing $4-x$ non-defective items is $\left(\frac{4}{5}\right)^{4-x}$. Hence the probability of choosing x defective and $4-x$ non-defective items is

$$P(X=x)=\binom{4}{x}\times\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{4-x}.$$

5·21. Mathematical Expectation. Let $\psi(x)$ be a function of the discrete random variable X so that it takes the value $\psi(x_i)$ when x takes the value x_i ; and let p_i be the probability of this value of the random function

i.e. $$p_i = P = P(X = x_i).$$

The expected value of the function $\psi(x)$ is given by

$$E[\psi(x)] = \sum_{i=1}^{n} p_i \psi(x_i), \qquad \text{where} \qquad \sum_{i=1}^{n} p_i = 1.$$

When $$\psi(x) = x^r, \qquad E(X^r) = \sum_{i=1}^{n} p_i x_i^r \qquad \ldots(5{\cdot}1)$$

and is defined as the rth moment about $x=0$ of the discrete probability distribution. It is denoted by μ_r'.

Also $$\mu_r = E(X - \bar{X})^r = \sum_{i=1}^{n} p_i (x - \bar{x})^r \qquad \ldots(5{\cdot}2)$$

In particular $$\mu_1' = E(X) = \sum_{i=1}^{n} p_i x_i \qquad \ldots(5{\cdot}3)$$

and is commonly called the *expected value* or the *expectation* of the discrete random variable X, provided the series $\Sigma p_i r_i$ is convergent.

Also $$\mu_2 = E[X - E(X)]^2 \qquad \ldots(5{\cdot}4)$$

and is called the *variance* of the distribution x and is denoted by Var (X).

Example 5·3. *What is the expected value of the number of points that will be obtained in a single throw with an ordinary dice ?*

Hence the variate X is the number of points on a dice. Its values are 1, 2, 3, 4, 5, 6 each having the probability $\frac{1}{6}$.

Hence $E(X) = p_1x_1 + p_2x_2 + \ldots\ldots + p_6x_6 = \frac{1}{6}.1 + \frac{1}{6}.2 + \ldots + \frac{1}{6}.6 = \frac{7}{2}$.

Example 5·4. *What is the expectation of the number of failures preceding the first success in an infinite series of independent trials with constant probability p of success in a trial ?*

(*B.A. Hons. Delhi, 1961* ; *M.A. Madras, 1959* ; *M.A. Agra, 1964* ; *B.Sc. Hons. Gauhati, 1971* ; *M.A. Bombay, 1953* ; *B.Sc. Hons. Sub. Statistics, Delhi, 1966*)

If X denotes the number of failures preceding the first success we find that X takes the values 0, 1, 2, 3,.....with probabilities

$$p,\ qp,\ q^2p,\ q^3p, \ldots\ldots \text{where } q = 1 - p,$$

Hence $$E(X) = 0\,p + 1.qp + 2.q^2p + 3.q^3p + \ldots$$
$$= pq(1 + 2q + 3q^2 + \ldots\ldots)$$
$$= pq(1-q)^{-2} = \frac{q}{p} = \frac{1}{p} - 1.$$

Example 5·5. *A person draws cards one by one from a pack until he draws all the aces. How many cards he may be expected to draw?* (*Indian Statistical Institute, Calcutta, 1941*)

Suppose in x drawings he draws all the aces. This means that in $(x-1)$ drawings he should draw 3 aces and the xth draw should be an ace. The probability of such an event is

$$\frac{{}^{4}C_{3}\times{}^{48}C_{x-4}}{{}^{52}C_{x-1}}\times\frac{1}{52-x+1}=\frac{4(x-1)(x-2)(x-3)}{52\times51\times50\times49}$$

Hence required expectation

$$=\sum_{x=4}^{52}x.\frac{4(x-1)(x-2)(x-3)}{52.51.50.49}$$

$$=\frac{4}{52.51.50.49}\times\left[\sum_{x=4}^{52}x^4-6\sum_{x=4}^{52}x^3+11\sum_{x=4}^{52}x^2-6\sum_{x=4}^{52}x\right]$$

Ex. 1. Show by an example that the mathematical expectation need not be finite.

[**Hint.** Suppose the elementary probability law of the discontinuous random variable k is $p(k)=\frac{e^{-1}}{k!}$, $k=0, 1, 2,\ldots\ldots\infty$.

$\therefore\ E(k!)=\sum_{k=0}^{\infty}\left[k!.\frac{e^{-1}}{k!}\right]=e^{-1}\sum_{0}^{\infty}1$ Hence, which is not finite.]

Ex. 2. Let a random variable X take the values $X_k=(-1)^k 2^k k^{-1}$, $k=1, 2, 3,\ldots$, with probabilities $p_k=2^{-k}$. Find $E(X)$. [**Ans.** $-\log 2$]

Ex. 3. Show that if X is a discrete random variable and a is a constant,

(*i*) $E(a)=a$; (*iii*) $E(aX)=aE(X)$;

(*ii*) $E[a\psi(X)]=aE[\psi(X)]$; $\psi(X)$ is any function of X. (*iv*) $E(X-\bar{X})=0$.

Ex. 4. Show that (*i*) $Var(aX)=a^2\ Var(X)$, a being a constant.

(*ii*) $\mu_r(aX)=a^r\mu_r(X)$, a being a constant.

Ex. 5. Two unbiased dice are thrown. Find the expected value of the sum of numbers of points drawn. [**Ans.** 7]

Ex. 6. Prove that $Var(X)=E(X^2)-[E(X)]^2$. Deduce that $E(X^2)\geqslant[E(X)]^2$.

Ex. 7. Prove that moments about the mean are invariant for change of origin, but not for change of scale.

Ex. 8. If $P(X\leqslant 0)=1$ and if $E(X)$ exists, then $E(X)\leqslant 0$.

Ex. 9. Prove that $|E(X)|\leqslant E|X|$.

[**Hint.** $|E(X)|=|\Sigma x_i p_i|<\Sigma|x_i p_i|=\Sigma|x_i|\,|p_i=E|X|$.]

5·2·2. Expectation of a Sum of Random Variables. Theorem. *The expectation of the sum of two random variables is equal to the sum of their expectations.*

In symbols, if X and Y are random variables, then

$$E(X+Y)=E(X)+E(Y).$$

Proof. Let X and Y be two stochastic variates. Let X assume the m values $x_1, x_2, \ldots\ldots, x_m$ with probabilities $p_1, p_2, \ldots\ldots, p_m$ and Y assume the n values $y_1, y_2, \ldots\ldots, y_n$ with probabilities $p_1', p_2', \ldots\ldots, p_n'$. Then the sum $X+Y$ is a stochastic variate which can take the mn values x_i+y_j, $(i=1, 2, \ldots\ldots, m\ ;\ j=1, 2, \ldots\ldots, n)$. Let p_{ij} denote the probability of X assuming the value x_i, and at the same time, Y assuming the value y_j, *i.e.*, $p_j=P(X=x_i,\ Y=y_j)$. As X assumes a definite value x_i, Y assumes one of the values $y_1, y_2, \ldots$ $\ldots, y_n$ so that the sum $\sum_{j=1}^{n} p_{ij}$ represents the probability p_i of X assuming the value x_i that is, $\sum_{j=1}^{n} p_{ij}=p_i$.

Similarly, $\sum_{i=1}^{m} p_{ij}$ represents the probability of Y assuming the value y_j that is, $=p_j'$.

By definition,
$$E(X+Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}(x_i+y_j)$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}x_i + \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}y_j$$
$$= \sum_{i=1}^{m} x_i \left(\sum_{j=1}^{n} p_{ij}\right) + \sum_{j=1}^{n} y_j \left(\sum_{i=1}^{m} p_{ij}\right)$$
$$= \sum_{i=1}^{m} x_i p_i + \sum_{j=1}^{n} y_j p_j'$$

Thus $E(X+Y)=E(X)+E(Y)$.

Cor. 1. The expectation of the sum of a number of discrete random variables is the sum of their expectations. In symbols,

$$E(X+Y+Z+\ldots)=E(X)+E(Y)+E(Z)+\ldots$$

For $E(X+Y+Z+\ldots)=E(X+\overline{Y+Z+\ldots}),$
$$=E(X)+E(\overline{Y+Z+\ldots}),$$
$$=E(X)+E(Y)+E(Z+\ldots),$$
$$=E(X)+E(Y)+E(Z)+\ldots$$

Cor. 2. $E(aX+b)=aE(X)+b$.

Cor 3. $E[\psi_1(x)+\psi_2(x)]=E[\psi_1(x)]+E[\psi_2(x)]$.

Example 5·6. *What is the mathematical expectation of the sum of points on a dice ?*

Denoting by x_i the number of points on the ith dice, the sum of points on n dice will be

$$s=x_1+x_2+\ldots+x_n$$

$\therefore \quad E(s)=E(x_1+x_2+\ldots+x_n)=E(x_1)+E(x_2)+\ldots+E(x_n)$.

But for every single dice, $E(x_i)=\frac{7}{2}$, $i=1, 2,\ldots, n$. (*See* example 5·3).

Hence $E(s)=\dfrac{7n}{2}$.

Example 5·7. *Suppose that we have a series of n independent trials, the probability of success being p_i for ith trial. Show that the mean number of successes is* $\sum_{i=1}^{n} p_i$. *(B.A., Hons., Delhi, 1955)*

Let us associate with ith trial a random variable x_i which has the value 1 in case of success and the value 0 in case of a failure. Then the number of successes in n trials is given by

$m=x_1+x_2+\ldots+x_n$, so that $E(m)=E(x_1)+E(x_2)+\ldots+E(x_n)$

Since x_i can take the values 1 and 0 with probabilities p_i and $1-p_i$, we have

$$E(x_i)=1.p_i+0(1-p_i)=p_i$$

Hence $E(m)=p_1+p_2+\ldots+p_n=\sum_{i=1}^{n} p_i$.

Ex. 1. If x is a continuous random variable, show that *(i)* $Var\,(x)=E(x)^2-[E(x)]^2$; *(ii)* $Var\,(ax)=a^2\,Var\,(x)$, *(iii)* $E(ax+by)=aE\,(x)+bE(y)$, a and b are constants. *(B.Sc. Delhi, 1971)*

Ex. 2. Thirteen cards are drawn simultaneously from a deck of 52. If aces count 1, face cards 10 and others according to denomination, find the expectation of the total score on the 13 cards. [**Ans.** 85]

5·2·3. Expectation of a Product of Independent Random Variables.

Definition. *Two random variables are said to be independent if probability of either taking a particular value does not depend on what value the other variate takes.*

Theorem. *The expectation of the product of two independent random variables is equal to the product of their expectations ; or in symbols, if X and Y are two independent random variables, then*

$$E(XY)=E(X)E(Y).$$

Proof. Let the random variable X assume the m values x_1, $x_2,\ldots\ldots, x_m$ and the variable Y the n values $y_1, y_2,\ldots\ldots, y_n$. Let the probability of X assuming the value x_i be p_i' and the probability of Y assuming the value y_j be p_j'. Since the variables X and Y are independent, by the theorem of compound probability, the probability that X assumes the value x_i and at the same time Y assumes the value y_j is p_ip_j'.

$$\text{Hence } E(XY)=\sum_{i=1}^{m}\sum_{j=1}^{n} p_i p_j' x_i y_j=\sum_{i=1}^{m} p_i x_i \sum_{j=1}^{n} p_j' y_j$$

$$=\sum_{i=1}^{m} p_i x_i E(Y)=E(Y)\sum_{i=1}^{m} p_i x_i$$

$$=E(X)E(Y).$$

Note. This theorem can be extended to the case of several independent variables, so that if $X, Y, Z,\ldots$ are independent stochastic variables,

$$E(XYZ.....)=E(X)E(Y)E(Z).....$$

Ex. 1. Find the expected value of the product of points on n dice ?

Ans. $\left(\frac{7}{2}\right)^n$

Ex. 2. Prove that if X, Y are two random variables whose expectations exists and if $P(x \leqslant y)=1$, then $E(X) \leqslant E(Y)$.

Ex. 3. If $x_1, x_2,\ldots\ldots, x_n$ are independent in pairs, are they necessarily independent variables ?

5·3. Covariance. Definition. If X and Y are two random variables with the respective expected values (or means) $\bar{X}$ and $\bar{Y}$, then the covariance between X and Y is defined by the relation

$$Cov\ (X, Y)=E_1(X-\bar{X})(Y-\bar{Y})] \qquad \ldots(5\cdot5)$$

Cor. If X and Y are independent variables, then (5·5) reduces to

$$Cov\ (X, Y)=E(X-\bar{X})\ E(Y-\bar{Y})=0,$$

since $$E(X-\bar{X})=E(X)-E(\bar{X})=\bar{X}-\bar{X}=0.$$

Thus *the covariance of two independent variables is equal to zero.*

The *converse* of this is not necessarily true, *i.e.*, if the covariance between two variables is zero, it is not necessary that they are independent. Thus let u and v be two independent variables with same variance and let

$$X=u+v, \qquad Y=u-v,$$

so that $$\bar{X}=\bar{u}+\bar{v} \qquad \bar{Y}=\bar{u}-\bar{v}.$$

Now $$Cov(X, Y)=E(X-\bar{X})(Y-\bar{Y})=E(u-\bar{u}+v-\bar{v})(u-\bar{u}-v+\bar{v})$$

$$=E(u-\bar{u})^2-E(v-\bar{v})^2=\sigma_u^2-\sigma_v^2=0,$$

but X and Y are not necessarily independent *e.g.*, if u and v denote the number of points on two dice, then their sum and difference X and Y are either both even or both odd and thus are dependent.

5·3·1. Correlation Coefficient. Definition. The ratio

$$r_{XY}=\frac{Cov\ (X, Y)}{\sqrt{[(Var\ (X)\ Var\ (Y)]}}=\frac{E(X-\bar{X})(Y-\bar{Y})}{\sqrt{E(X-\bar{X})^2}\sqrt{E(Y-\bar{Y})^2}}=\frac{\mu_{11}}{\sigma_X\sigma_Y} \qquad \ldots(5\cdot5')$$

is defined as the correlation coefficient between X and Y.

Obviously $r_{XY}=r_{YX}$

and both have the same sign as $Cov\ (X, Y)$.

Note. (5·5) can be put in the form :

$$\begin{aligned} Cov\ (X, Y) &= E(XY-\bar{X}Y-\bar{Y}X+\bar{X}\bar{Y}) \\ &= E(XY)-E(\bar{X}Y)-E(\bar{Y}X)+E(\bar{X}\bar{Y}) \\ &= E(XY)-\bar{X}E(Y)-\bar{Y}E(X)+\bar{X}\bar{Y} \\ &= E(XY)-\bar{X}\bar{Y}-\bar{Y}\bar{X}+\bar{X}\bar{Y} \\ &= E(XY)-E(X)E(Y) \qquad \ldots(5\cdot5'') \end{aligned}$$

Ex. Show that, if a and b are constants,

(*i*) $Cov\ (aX, bY)=ab\ Cov\ (X, Y)$,

(*ii*) $Cov\ (X+a, Y+b)=Cov\ (X, Y)$,

(*iii*) $Cov\ (X, X)=Var\ (X)$.

(*iv*) $Cov\left(\frac{X-\bar{X}}{\sigma_x}, \frac{Y-\bar{Y}}{\sigma_y}\right)=\frac{1}{\sigma_x\sigma_y} Cov\ (X, Y)$

(*v*) $r_{ax+b,\ cy+d}=r_{ax,\ cy}=\pm r_{x,\ y}$

according as a and c have the same or opposite signs.

5·5. Expectation and Variance of a Linear Combination of Random Variables.

Let $x_1, x_2 \ldots\ldots x_n$, be n random variables with finite variances $\sigma_1^2, \sigma_2^2, \ldots\ldots, \sigma_n^2$, then we have to find the variance of u where

$$u=a_1x_1+a_2x_2+\ldots\ldots+a_nx_n.$$

a's being constants.

We have $$\begin{aligned} E(u) &= E(a_1x_1+a_2x_2+\ldots\ldots+a_nx_n) \\ &= E(a_1x_1)+E(a_2x_2)+\ldots+E(a_nx_n) \\ &= a_1E(x_1)+a_2E(x_2)+\ldots+a_nE(x_n) \end{aligned}$$

$$\therefore \quad u-E(u)=a_1[x_1-E(x_1)]+a_2[x_2-E(x_2)]+\ldots+a_n[x_n-E(x_n)]$$

Squaring both sides and taking expected values, we get

$$\begin{aligned} Var\ (u) &= a_1^2\ Var\ (x_1)+a_2^2\ Var\ (x_2)+\ldots+a_n^2\ Var\ (x_n) \\ &\quad +2a_1a_2\ Cov\ (x_1, x_2)+\ldots\ldots+2a_{n-1}a_n\ Cov\ (x_{n-1}, x_n) \\ &= \sum_{i=1}^{n} a_i^2\ Var\ (x_i)+2\sum_{\substack{i,\ j=1 \\ j\neq 1}}^{n} a_ia_j\ Cov\ (x_i, x_j) \qquad \ldots(5\cdot6) \\ &= \sum_{i=1}^{n} a_i^2\ \sigma_i^2+2\sum_{\substack{i,\ j=1 \\ j\neq 1}}^{n} a_ia_j\sigma_i\sigma_j r_{ij} \quad \text{using } (5\cdot5'), (5\cdot6) \end{aligned}$$

where r_{ij} is correlation coefficient between x_i and x_j.

Cor. 1. Let $a_1=1, a_2=1, a_3=a_4=\ldots=a_n=0$, so that

$$Var\ (x_1+x_2)=Var\ (x_1)+Var(x_2)+2\ Cov\ (x_1, x_2) \qquad \ldots(5\cdot7)$$

Cor. 2. Let $a_1=1, a_2=-1, a_3=a_4=\ldots=a_n=0$ so that

$$Var(x_1-x_2)=Var(x_1)+Var(x_2)-2Cov\ (x_1, x_2) \qquad \ldots(5\cdot7')$$

Cor. 3. Let x_1 and x_2 be independent variables, then

$Cov\ (x_1, x_2)=0,$

$Var\ (x_1 \pm x_2)=Var\ (x_1)+Var\ (x_2)$ ...(5·7″)

Cor. 4. Let $a_1=a_2=a_3=\ldots=a_n=1/n$, so that $u=\bar{x}$

Also let x_i's be independent variables and let $Var\ (x_i)=\sigma^2$, then

$$Var\ (\bar{x})=\frac{\sigma^2}{n^2}+\frac{\sigma^2}{n^2}+\ldots\ldots\text{up to } n \text{ times}=\frac{\sigma^2}{n}.$$

Cor 5. $Var\ (aX+b)=a^2\ Var\ X.$

(B.A. Hons., Delhi 1949, 55)

Example 5·8. *Find the variance of the number of successes in a series of n independent trials in which the probability of success in the ith trial is p_i.*

Using the notation of Example 5·6, page 99, we have

$$Var\ (m)=Var(x_1+x_2+\ldots+x_n)=Var(x_1)+Var(x_2)+\ldots+Var(x_n)$$

Now x_i^2 takes the value 1 if there is a success and zero in case of failure so that

$$E(x_i^2)=1.\ p_i+0\ .(1-p_i)=p_i.$$

Hence the variance of the number of successes in the ith trial is

$$Var(x_i)=E[x_i-E(x_i)]^2=E[x_i]^2-2x_iE(x_i)+\{E(x_i)\}^2$$
$$=E(x_i^2)-\{E(x_i)\}^2=p_i-p_i^2=p_iq_i\ ;\ q_i=1-p_i$$

$$\therefore\quad Var\ (m)=\sum_{i=1}^{n} p_iq_i \leqslant \frac{n}{4} \text{ as } p_i(1-p_i) \leqslant \tfrac{1}{4}$$

Cor. If $p_i=p$ then $Var\ (m)=pq+pq+\ldots+pq=npq.$

Example 5·9. *A box contains 2^n tickets among which ${}^nC_i$ tickets bear the number $i(i=0, 1, 2, \ldots, n)$. A group of m tickets is drawn. Let S denote the sum of their numbers. Find E(S) and Var S.*

(B.A., Hons. Calcutta, 1956 ; M.A , Delhi 1958 ; M.Sc., Calcutta, 1945, 48)

Let $x_1, x_2, \ldots\ldots, x_m$ be the variable representing the numbers on the first, second ..... m-th tickets.

Now x_i is a variable which takes the values 0, 1, 2, ..., n with respective probabilities

$$\frac{{}^nc_0}{2^n}, \frac{{}^nc_1}{2^n}, \ldots\ldots, \frac{{}^nc_n}{2^n}$$

$$\therefore\quad E(x_i)=\frac{1}{2^n}\ [1.{}^nc_1+2.{}^nc_2+\ldots\ldots+n.{}^nc_n]$$

$$=\frac{n}{2^n}[1+{}^{n-1}c_1+{}^{n-1}c_2+\ldots\ldots+{}^{n-1}c_{n-1}]$$

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

Hence $E(S)=E(x_1+x_2+\ldots+x_m)=\sum_{i=1}^{m} E(x_i)=\frac{mn}{2}$

$$E(x_i^2)=1^2.\frac{{}^nc_1}{2^n}+2^2.\frac{{}^nc_2}{2^n}+\ldots\ldots+n^2.\frac{{}^nc_n}{2^n}$$

$$=\frac{n}{2^n}[(1+1)^{n-1}+(n-1)(1+1)^{n-2}]$$

$$=\frac{n}{2^n}[2^{n-1}+(n-1)2^{n-2}]=\frac{n(n+1)}{4}.$$

$$E(x_ix_j)=\sum_{\substack{i\,j\\ j\neq i}}\left[ij\frac{{}^nc_i}{2^n}\cdot\frac{{}^nc_j}{2^n-1}\right]+\left[\sum_{i=1}^{n} i^2\frac{{}^nc_i}{2^n}.\frac{{}^nc_i-1}{2^n-1}\right]$$

$$=\sum_j\left[j.\frac{{}^nc_j}{2^n-1}\left(\sum_{i=1}^{n} i\frac{{}^nc_i}{2^n}-j.\frac{{}^nc_j}{2^n}\right)+\sum_{i=1}^{n}\frac{i^2({}^nc_i)^2}{2^n(2^n-1)}\right.$$

$$\left.-\sum_{i=1}^{n} i^2\frac{{}^nc_i}{2^n(2^n-1)}\right]$$

$$E(x_ix_j)=\sum_i\left[j\cdot\frac{{}^nc_j}{2^n-1}\left(\frac{n}{2}-j.\frac{{}^nc_j}{2^n}\right)\right]+\sum_{j=1} j^2\frac{({}^nc_j)^2}{2^n(2^n-1)}-\sum_{j=1}\frac{j\,{}^nc_j}{2^n(2^n-1)}$$

$$=\frac{1}{2^n-1}\left[\frac{n}{2}\cdot n.2^{n-1}-\frac{1}{2^n}\sum_{j=1}^{n} j^2{}^nc_j\right]$$

$$=\frac{1}{2^n-1}\left[n^2.2^{n-2}-\frac{n(n+1)}{4}\right]$$

$$\therefore\quad E(S_2)=\sum_{i=1}^{m} E(x_i^2)+\sum_{\substack{j\\ i\neq j}}\sum_{i} E(x_ix_j)$$

$$=\frac{mn(n+1)}{4}+\frac{m(m-1)n}{2^n-1}\left[n.2^{n-2}-\frac{n+1}{4}\right]$$

Hence $Var(S)=E(S^2)-[E(S)]^2=\frac{mn}{4}-\frac{mn(m-1)}{4(2^n-1)}$, on simplification.

Example 5·10. *Use the relation that* $E(Ax^a+Bx^b+Cx^c)^2$ *is always* $\geqslant 0$, *x being a random variable and E denotes the mathematical expectation, show that*

$$\begin{vmatrix} \mu_{2a} & \mu_{a+b} & \mu_{a+c} \\ \mu_{a+b} & \mu_{2b} & \mu_{b+c} \\ \mu_{a+c} & \mu_{b+c} & \mu_{2c} \end{vmatrix} \geqslant 0,$$ *μ_n being the nth moment about mean.*

Hence or otherwise show that $\beta_2-\beta_1-1\geqslant 0$. *Show also that* $\beta_2\geqslant 1$.

(M.Sc. Banaras, 1967 ; I.C.A.R., Delhi, 1950)

We are given that $E(Ax^a+Bx^b+Cx^c)^2$ *is always* $\geqslant 0$ *i.e.*, $E[A^2x^{2a}+B^2x^{2b}+C^2x^{2c}+2ABx^{a+b}+2BCx^{b+c}+2CAx^{c+a}]$ is always $\geqslant 0$.

If x denotes deviations from the mean, this gives

$A^2\mu_{2a}+B^2\mu_{2b}+C^2\mu_{2c}+2AB\mu_{a+b}+2BC\mu_{b+c}+2CA\mu_{c+a}$ is always $\geqslant 0$.

We know that the conditions that

$$a'x^2+b'y^2+c'z^2+2f'yz+2g'zx+2h'xy$$

is non-negative for all values of x, y and z are

$$(i)\ a'\geqslant 0, \qquad (ii)\ \begin{vmatrix} a' & h' \\ h' & b' \end{vmatrix} \geqslant 0, \qquad (iii)\ \begin{vmatrix} a' & h' & g' \\ h' & b' & f' \\ g' & f' & c' \end{vmatrix} \geqslant 0.$$

In the present case

$$a'=\mu_{2a},\ b'=\mu_{2b},\ c'=\mu_{2c},\ f'=\mu_{b+c},\ g'=\mu_{c+a},\ h'=\mu_{a+b}.$$

$\therefore$ We get the required result on substituting the values of a', b', c', f', g', h' in (iii).

Putting in this result $a=0$, $b=1$, and $c=2$ and remembering that $\mu_1=0$, we get

$$\begin{vmatrix} 1 & 0 & \mu_2 \\ 0 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix} \geqslant 0$$

i.e.,
$$\mu_2\mu_4-\mu_3^2-\mu_2^3 \geqslant 0.$$

For β_2 to be finite, μ_2 should be $\geqslant 0$.

Hence on dividing by μ_2^3, we have

$$\frac{\mu_4}{\mu_2^2}-\frac{\mu_3^2}{\mu_2^3}-1\geqslant 0$$

i.e.,
$$\beta_2-\beta_1-1\geqslant 0$$

Further since $\beta_1\geqslant 0$, we have $\beta_2\geqslant 1$, *otherwise* from (ii) putting $a=1$, $b=1$ we get

$$\begin{vmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 \end{vmatrix} \geqslant 0 \quad \text{or} \quad \mu_4\mu_2\geqslant\mu_3^2$$

or
$$\beta_2 \geqslant 1.$$

Ex. 1. Show that

$$Cov\left(\sum_{i=1}^{n} a_ix_i,\ \sum_{j=1}^{n} b_jx_j\right) = \sum_{j=1}^{n}\sum_{i=1}^{n} a_ib_j\ Cov\ (x_i, x_j).$$

Ex. 2. Prove that

$$\mu_{2a}'\mu_{2b}'\geqslant(\mu_{a+b}')^2$$

and

$$\mu_{2a}\mu_{2b}\geqslant(\mu_{a+b})^2$$

Deduce (i)
$$\mu_2'\geqslant\mu_1'^2$$

(ii)
$$\beta_2\geqslant 1.$$

Ex. 3. Prove that

$$\begin{vmatrix} \mu_{2a} & \mu_{a+1} & \mu_{a+c} & \mu_{a+d} \\ \mu_{b+a} & \mu_{2b} & \mu_{b+c} & \mu_{b+d} \\ \mu_{c+a} & \mu_{c+b} & \mu_{2c} & \mu_{c+d} \\ \mu_{d+a} & \mu_{d+b} & \mu_{d+c} & \mu_{2d} \end{vmatrix} \geqslant 0.$$

5·4. Conditional Expectation. Conditional Variance.

Def. Let X be a random variable which takes the values x_i with probability $p_i(Y)$ subject to the occurrence of the event Y, that is

$$p_i(Y=y_j)=P(X=x_i \mid Y=y_j),\ i=1, 2, \ldots\ldots, n$$
$$=P\{X=x_i) \cap Y=y_j\}/P(Y=y_j)$$

Then the *conditional expectation* of X_j given Y, is defined as

$$E(X \mid Y=y_j)=\sum_{i=1}^{n} p_j(Y=y_j)x_j$$
$$=\frac{\Sigma\, x_i\, P(X=x_i\ ;\ Y=y_j)}{P(Y=y_j)}$$

and the *conditional variance* of X, given Y, is defined as

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

The *Conditional Covariance* may be defined similarly.

Theorem I. *The expected value of X is equal to the expectation of the conditional expectation of X, given Y.* Symbolically,

$$E(X)=E[E(X \mid Y)]$$

Proof. $E[E(X \mid Y)]=\sum_j E(X \mid Y_j)=\sum_i \sum_j x_i p_i(Y_j)$

$$=\sum_i \sum_j x_i \frac{P(X=x_i,\ Y=y_j)}{p_j}$$
$$=\sum_i x_i\, P(X=x_i)$$
$$=E(X)$$

Theorem II. *The variance of X can be regarded as consisting of two parts, the expectation of the conditional variance and the variance of the conditional expectation.* Symbolically,

$$Var(X)=E[Var\ (X \mid Y)]+Var[E(X \mid Y)]$$

Proof. $Var(X)=E[X-E(X)]^2$

$$=E[\overline{X-E(X \mid Y)}+\overline{E(X \mid Y)-E(X)}]^2$$
$$=E[X-E(X \mid Y)]^2+E[E(X \mid Y)-E(X)]^2 \ldots(5{\cdot}8)$$

the middle term being zero since it is

$$E[XE(X \mid Y)-XE(X)+E(X).E(X \mid Y)-[E(X \mid Y)]^2 \qquad \ldots(5{\cdot}9)$$

and $E[XE(X \mid Y)]=E[E(X \mid Y).\ E(X \mid Y)]$

and similarly, $E[X^{n/Y)}]=E[E(X \mid Y).E(X)]$

so that the terms in (5·9) cancel. Hence the result.

Theorem III. $Cov\ (X, Y)=E\ Cov\ (X, Y \mid Z)$
$+Cov\ [E(X \mid Z)]\ E(Y \mid Z).$

Proof. $Cov\ (X, Y)=E(X, Y)-E(X)E(Y)$
$=EE(XY \mid Z)-EE(X \mid Z)EE(Y \mid Z),$
(by theorem I)
$=E\{Cov(X, Y \mid Z)+E(X \mid Z)E(Y \mid Z)\}$
$-EE(X \mid Z)EE(Y \mid Z)$
$=E\ Cov(X, Y \mid Z)+EE(X \mid Z)E(Y \mid Z)$
$-EE(X \mid Z)$
$=E\ Cov\ (X, Y \mid Z)+Cov\ [E(X \mid Z)]E(Y \mid Z).$

5·5. Moment Generating Function. The moment generating function (*m.g. f.*) of the discrete probability distribution of a random variable X about the point a is defined by

$$M_a(t)=E[e^{t(X-a)}]=\sum_i e^{t(x_i-a)}p_i \qquad \ldots(5\cdot10)$$

where t is an auxiliary variable.

The *m.g. f.* is a function of the real parameter t only. The subscript 'a' shows the point about which the *m.g. f.* is found out.

Expanding the exponential in (5·10), we get

$$M_a(t)=\left[\sum_i p_i+t\sum_i p_i(x_i-a)+\frac{t^2}{\lfloor 2}\sum_i p_i(x_i-a)^2+\ldots\ldots\right.$$
$$\left.\ldots+\frac{t^r}{\lfloor r}\sum_i p_i(x_i-a)^r+\ldots\right]$$
$$=1+t\mu_1'+\frac{t^2}{2!}\mu_2'+\ldots.+\frac{t^r}{r!}\mu_r'+\ldots \qquad \ldots(5\cdot11)$$

where μ_r' is the moment of order r about the point a.

Thus we see that the function $M_a(t)$ defined by (5·10) generates moments and that is why it is called the moment generating function.

From (5·11), we find μ_r'=coefficient of $t^r/r!$ in the expansion of $M_a(t)$ in ascending powers of t.

This gives a method of finding the moments. However, it may be found more conveniently to evaluate the moment by the following formula obtained by differentiating the relation (5·11) r times with respect to t and then putting $t=0$. Thus

$$\mu_r'=\left[\frac{d^r}{dt^r}\ M_a(t)\right]_{t=0} \qquad \ldots(5\cdot12)$$

The equation (5·10) can be written as

$$M_a(t)=e^{-at}\ \Sigma e^{tx_i}\ p_i$$

or
$$M_a(t)=e^{-at}\ M_0(t) \qquad \ldots(5\cdot13)$$

Thus *the m.g. f. about the point a is equal to* e^{-at} *times m.g. f. about the origin.*

Note. For many probability density functions, the integral $\int_{-\infty}^{\infty} e^{tx} f(x)\,dx$ or the sum $\sum_i e^{tx_i} p_i$ may not exist for real values of t. Thus if $f(x)=6/\pi^2x^2$, $x=1, 2, 3\ldots\ldots$, then the *m.g. f.* does not exist.

5·5·1. Effect of Change of Origin and Scale on M.G.F. Let us introduce a new variate U related to X by

$$U=\frac{X-a}{h} \text{ so that } X=a+hU$$

By definition,

$$M_o(t)=\sum_i e^{tX_i}\ p_i=\sum_i e^{t(a+hU_i)}p_i=e^{at}\ \sum_i e^{thU_i}p_i$$

Writing t/h for t on both sides, we get

$$M.G.F. \text{ for } \frac{X-a}{h}\ =e^{-\frac{at}{h}}\ M_o\left(\frac{t}{h}\right). \qquad \ldots(5\cdot14)$$

5·5·2. A Property of Moment Generating Function. *The moment generating function of the sum of two independent variables is the product of their moment generating functions.*

Proof. Let X and Y be two independent variables, then the *m.g. f.* of the sum $X+Y$ with respect to origin is

$$E[e^{t(X+Y)}]=E(e^{tX}\ .\ e^{tY})=E(e^{tX}).E(e^{tY}), \qquad \ldots(5\cdot15)$$

since X and Y are independent variables and so are e^{tX} and e^{tY}. Hence the theorem.

Exercise. Generalize the theorem and give its proof.

5·5·3. Factorial Moment Generating Function. The factorial moment generating function $w(t)$ of the discrete probability distribution of X about the origin is defined as

$$w(t)=E[(1+t)^X] \qquad \ldots(5\cdot16)$$

Since $(1+t)^n=\sum_{r=0}^{\infty}\frac{X^{(r)}\ t^r}{r\,!}$,

we get $\mu_{(r)}'=$coefficient of $\frac{t^r}{r\,!}$

in the expansion of $w(t)$ in ascending power of t.

5·6. Cumulants. If the logarithm of the *m.g. f.* of a distribution can be expanded as a convergent series in powers of t *viz.*,

$$K(t)=\log_e M_o(t)$$
$$\equiv k_1 t+k_2\frac{t^2}{2!}+\ldots+\frac{k_r t^r}{r!}+\ldots \qquad \ldots(5\cdot17)$$

then the coefficients k_1, k_2,..... are called the first cumulant, second cumulant.....of the distribution and $K(t)$ is called the *cumulative function.*

Differentiating r times the members of the relation (5·17) with respect to t and putting $t=0$, we have

$$k_r=\left[\frac{d^r}{dt^r}\log_e M_o(t)\right]_{t=0}=\left[\frac{d^r}{dt^r}K(t)\right]_{t=0} \qquad \ldots(5\cdot18)$$

5·6·1. Effect of Change of Origin and Scale on Cumulants.

Let $u=\dfrac{X-a}{h}$, so that $X=a+hu$.

The cumulative function about $X=0$ is given by

$$K(t)=\log_e M_o(t)=\log_e[\Sigma e^{txi}p_i]=\log_e[\Sigma e^{t(a+hu_i)}p_i]$$
$$=\log_e[e^{at}\Sigma e^{thu_i}p_i]=at+\log_e M_o(th)$$

or
$$K(t)=at+k_1(th)+k_2\frac{(th)^2}{2!}+\ldots+k_r\frac{(th)^r}{r!}+\ldots$$
$$=[a+(k_1h)]t+(k_2h^2)\frac{t^2}{2!}+\ldots+(k_rh^r)\frac{t^r}{r!}+\ldots$$

Thus *except the first cumulant, the r-th cumulant of the distribution of X is h^r times the r-th cumulant of the distribution of u.* In other words, *cumulants are not invariant for the change of scale.* Also we notice that *all the cumulants except the first are invariant for the change of origin.*

Note. The cumulants (or semi-invariants) owe their name to two very remarkable properties. In the first place, all cumulants except the first are independent of origin of reference. Secondly, if the scale of measurement of the variable is altered by multiplying all values by a constant h, the rth cumulant is multiplied by h^r.

5·6·2. Additive Property of Cumulants. *The r-th cumulant of the sum of independent stochastic variables is the sum of the r-th cumulants of the independent variables.*

Taking logarithms of both sides of (5·15), we get

$\log_e(m.g.f.\text{ for } X+Y)=\log_e(m.g.f.\text{ for } X)+\log_e(m.g.f.\text{ for } Y)$.

Equating coefficients of like powers of t, we get the required result.

5·7. Relations between Moments and Cumulants.

5·7·1. Cumulants in terms of moments. From (5·17) we have the identity

$$k_1 t + k_2 \frac{t^2}{2!} + k_3 \frac{t^3}{3!} + \ldots = \log_e\left(1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots\right) \quad \ldots(5\cdot19)$$

On comparing the coefficients of like powers of t, we find

$$\left.\begin{aligned} k_1 &= \mu_1', \\ k_2 &= \mu_2' - \mu_1'^2 = \mu_2, \\ k_3 &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \mu_3, \\ k_4 &= \mu_4' - 4\mu_3'\mu_1' - 3\mu_2'^2 + 12\mu_2'\mu_1'^2 - 6\mu_1'^4 \\ &= \mu_4 - 3\mu_2^2 \end{aligned}\right\} \quad \ldots(5\cdot20)$$

These results can be *directly* obtained as follows :

From (5·13), $M_{\mu_1'}(t) = e^{-\mu_1' t} M_o(t)$

$$\therefore \qquad \log_e M_{\mu_1'}(t) = -\mu_1' t + \log_e M_o(t)$$

$$\text{or } \log_e\left(1 + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \ldots\ldots\right) = -\mu_1' t + k_1 t + k_2 \frac{t^2}{2!} + \ldots\ldots \quad \ldots(5\cdot21)$$

Equating the coefficients of like powers of t after expanding the left hand member, we obtain the required results.

Example 5·11. *Show that* $\mu_r' = \sum_{j=1}^{r} {}^{r-1}c_{j-1}\mu'_{r-j}k_j$.

(B.Sc. Shivaji, 1971 ; B.Sc. Hons. Gauhati 1969)

Differentiating the equation (5·19) with respect to t and cross-multiplying, we get

$$\mu_1' + t\mu_2' + \ldots + \frac{t^{r-1}}{(r-1)!}\mu_r' + \ldots\ldots$$

$$= \left[1 + t\mu_1' + \ldots + \frac{t^{r-j}}{(r-j)!}\mu'_{r-j} + \ldots\ldots\right] \times \left[k_1 + k_2 t + \ldots + k_j \frac{t^{j-1}}{(j-1)!} + \ldots\right]$$

Comparing the coefficient of t^{r-1} on both sides we get the required result.

Ex. Show that $\beta_1 = k_3^2/k_2^3$ and $\beta_2 = (k_4/k_2^2) + 3$.

5·7·2. Moments in terms of Cumulants. From (5·19), we have

$$1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \ldots + \mu_r' \frac{t^r}{r!} + \ldots\ldots$$

$$= \exp\left[k_1 t + \frac{k_2 t^2}{2!} + \ldots \frac{k_r t^r}{r!} + \ldots\right]$$

$$= \exp[k_1 t] \exp\left[\frac{k_2 t^2}{2!}\right] \ldots\ldots \exp\left[\frac{k_r t^r}{r!}\right] \ldots\ldots$$

$$= \left[1 + k_1 t + k_1^2 \frac{t^2}{2!} + \ldots\right] \cdot \left[1 + k_2 \frac{t^2}{2!} + \ldots\right] \ldots\ldots$$

$$\left[1 + k_r \frac{t^r}{r!} + \ldots\right] \ldots$$

Equating the coefficient of t^r on both sides, we have

$$\mu_r' = \sum_{m=0}^{r} \Sigma \frac{k_{p_1}^{\pi_1}}{(p_1!)^{\pi_1} \pi_1!} \cdot \frac{k_{p_2}^{\pi_2}}{(p_2!)^{\pi_2} \pi_1!} \ldots\ldots \frac{k_{p_m}^{\pi_m}}{(p_m!)^{\pi_m} \pi_m!} \; r!$$

where the second summation extends over all non-negative values of π's such that

$$p_1\pi_1 + p_2\pi_2 + \ldots\ldots + p_m\pi_m = r.$$

Note. This can be expressed in a determinental form by using the generalised expansion of exp x (published by Dr. G.S. Mahajani and Dr. Ram Bihari). The generalised expansion is

$$\text{Exp}\left[a_1 x + a_2 x^2 + \ldots\right] = 1 + D_1 x + D_2 \frac{x^2}{2!} + \ldots + \frac{D_r x^r}{r!} + \ldots$$

where

$$D_r = \begin{vmatrix} a_1 & -1 & 0 & 0 & \ldots 0 \\ 2a_2 & a_1 & -2 & 0 & \ldots 0 \\ 3a_3 & 2a_2 & a_1 & -3 & \ldots 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ ra_r & (r-1)a_{r-1} & \ldots & .. & \ldots a_1 \end{vmatrix}$$

Hence $\exp\left\{k_1 t + k_2 \frac{t^2}{2!} + \ldots + k_r \frac{t^r}{r!} + \ldots\ldots\right\}$

$$= 1 + D_1 t + D_2 \frac{t^2}{2!} + \ldots + D_r \frac{t^r}{r!} + \ldots\ldots$$

where

$$D_r = \begin{vmatrix} k_1 & -1 & 0 & 0 \ldots\ldots 0 \\ \frac{2k_2}{2!} & k_1 & -2 & 0 \ldots\ldots 0 \\ \frac{3k_3}{3!} & \frac{2k_2}{2!} & k_1 & -3 \ldots\ldots 0 \\ \ldots & \ldots & \ldots & \ldots \quad \ldots \\ \ldots & \ldots & \ldots & \ldots \quad \ldots \\ \frac{rkr}{r!} & \frac{(r-1)k_{r-1}}{(r-1)!} & \ldots & \ldots\ldots\ldots k_1 \end{vmatrix}$$

Equating the coefficient of t^r on both sides, we have

$$\mu_r' = D_r \qquad \ldots(5{\cdot}22)$$

In particular,

$$\mu_1' = D_1 = k_1,$$

$$\left.\begin{aligned}
\mu_2' = D_2 &= \begin{vmatrix} k_1 & -1 \\ k_2 & k_1 \end{vmatrix} = k_1^2 + k_2 \\
\mu_3' = D_3 &= \begin{vmatrix} k_1 & -1 & 0 \\ k_2 & k_1 & -2 \\ \frac{k_3}{2} & k_2 & k_1 \end{vmatrix} = k_1^3 + 3k_1k_2 + k_3 \\
\mu_4' = D_4 &= \begin{vmatrix} k_1 & -1 & 0 & 0 \\ k_2 & k_1 & -2 & 0 \\ \frac{k_3}{2} & k_3 & k_1 & -3 \\ \frac{k_4}{6} & \frac{k_3}{2} & k_2 & k_1 \end{vmatrix} = k_1^4 + 6k_1^2k_2 + 4k_1k_3 + 3k_2^2 + k_4
\end{aligned}\right\} \qquad \ldots(5{\cdot}23)$$

5·8. Probability Generating Function. If $a_0, a_1, a_2, \ldots,$ is a sequence of real numbers and if

$$A(s) = a_0 + a_1 s + a_2 s^2 + \ldots = \sum_{i=1}^{\infty} a_i s^i$$

converges in some interval $-s_0 < s < s_0$, then the function $A(s)$ is known as the generating function of the sequence $\{a_i\}$. If the sequence $\{a_i\}$ is bounded, then $A(s)$ clearly converges for at least $|s| < 1$.

If $a_i = c$ for all i where c is a constant, then

$$A(s) = c(1 + s + s^2 + s^3 + \ldots\ldots) = c.\frac{1}{1-s}$$

For $\qquad a_i = \dfrac{1}{i\,!},$

$$A(s) = \sum_{i=0}^{\infty} \frac{1}{i\,!}\, s^i = e^s$$

If $a_i = p_i = P(X = i)$, then the expression

$$P(s) = p_0 + p_1 s + p_2 s^2 + \ldots\ldots = \sum_{i=0}^{\infty} p_i s^i = E(sX)$$

is known as the probability generating function (*p.g. f.*) of the random variable X. Also the probability of $X=i$ is the coeff. of s^i and hence the reason for the name *p.g. f.* Clearly if $s=e^t$ then $P(e^t)$ is the moment generating function. This provides another method of finding M.G.F. from $P(s)$.

Theorem. *The probability generating function of the sum of a number of independent random variables is equal to the product of their p.g. f.'s.*

Proof. We shall prove the result for two independent random variables though the proof is general.

Let X_1 and X_2 be two independent random variables. Then

$$P_{X_1+X_2}(s)=E(s^{X_1+X_2})$$
$$=E(s^{X_1})E(s^{X_2}),$$
since X_1 and X_2 are independent
$$=P_{X_1}(s)\ P_{X_2}(s)$$

which proves the result.

In general,

$$P_{X_1+X_2+\ldots+X_n}(s)=\prod_{i=1}^{n} P_{X_1}(s)$$

Cor. If the independent *r. v.* $X_1, X_2, \ldots\ldots X_n$ are all identically distributed and hence have the same *p.g. f.*, $P_X(s)$. Then

$$P_{X_1+X_2+\ldots+X_n}(s)=[P_X(s)]^n.$$

Example 5·12. *If x is the number scored with the throw of an unbiased die, find the generating function of $\{x\}$.*

Here $p_i=\frac{1}{6}$ for $i=1, 2, \ldots\ldots, 6$

Hence the generating function $=\frac{1}{6}s+\frac{1}{6}s^2+\ldots\ldots+\frac{1}{6}s^6$
$$=\tfrac{1}{6}s(1-s^6)/(1-s).$$

Example 5·13. *Let X be a random variable with generating function $p(s)$. Find generating function of (a) $X+1$ (b) $2X$. (c) $p(X\leqslant n)$.* (*M.A. Delhi 1969, 74*)

(*a*) We have $P_X(s)=\sum_{k=0}^{\infty} p_k s^k$

$\therefore$ Probability generating function of $(X+1)$

$$=\sum_{k=0}^{\infty} p_k s^{k+1}=s\sum_{k=0}^{\infty} p_k s^k=sP(s).$$

(*b*) Probability generating function of $2X$

$$= \sum_{k=0}^{\infty} p_k s^{2k} = \sum_{k=0}^{\infty} p_k (s^2)^k = P(s^2).$$

(*c*) Let $P(X \leqslant n) = q_n$ and $P(X=n) = p_n$

so that $$q_n = P(X \leqslant n) = P(X=0) + P(X=1) + \ldots + P(X=n)$$
$$= p_0 + p_1 + \ldots\ldots + p_n$$

$$\therefore \quad Q(s) = \sum_{n=0}^{\infty} q_n s^n = \sum_{n=0}^{\infty} (p_0 + p_1 + \ldots\ldots + p_n) s^n$$
$$= p_0 s^0 + (p_0 + p_1)s + (p_0 + p_1 + p_2)s^2 + \ldots\ldots$$
$$= p_0(1 + s + s^2 + \ldots\ldots) + p_1 s(1 + s + s^2 + \ldots\ldots)$$
$$+ p_2 s^2 (1 + s + s^2 + \ldots\ldots)$$
$$= p_0 \frac{1}{1-s} + p_1 \frac{s}{1-s} + p_2 \frac{s^2}{1-s} + \ldots$$
$$= \frac{1}{1-s} \sum_{n=0}^{\infty} p_n s^n = \frac{1}{1-s} \cdot P(s).$$

Ex. Obtain the *p.g.f.* for the distribution

(*a*) $P(X=r) = {}^nC_r\, p^r (1-p)^{n-r}$, $r = 0, 1, 2, \ldots\ldots$

(*b*) $P(X=r) = (1-p)p^{r-1}$, $r = 1, 2, \ldots\ldots$

(*c*) $P(X=r) = \dfrac{e^{-m} m^r}{r!}$, $r = 0, 1, 2, \ldots\ldots$

(*d*) $P(X=r) = p(1-p)^r$, $r = 1, 2, \ldots$ Ans. (*iv*) $P(y) = \dfrac{p_s}{1-(1-p)^s}$

5·9. Moments of Bivariate Probability distribution. The expectation of a function of two dimensional random variables (X, Y) is $E[g(X, Y)] = \sum_i \sum_j x_i y_j \; ; \; P(X=x_i, Y=y_j)$

In particular the (r, s)th moments of order r in X and s in Y is defined as

$$\mu_{rs}' = E(X^r Y^s) = \sum_i \sum_j X_i^r Y_j^s \cdot P(X=x_i, Y=y_j)$$

Example 5·14. *Two ideal dice are thrown. Let X_1 be the score on the first die and X_2 the score on the second die. Let Y denote the maximum of X_1 and X_2 i.e. max (X_1, X_2)*

(*a*) *Write down the joint distribution of Y and X_1.*

(*b*) *Find $E(Y)$, Var (Y) and Cov (Y, X_1).*

(*M.Sc. Kerala 1967 ; M.Sc. Dibrugarh, 1969*)

Solution. Clearly X_1 and X_2 can each take the values 1 to 6 with probability 1/6.

$$P(Y=1, X_1=1) = P(X_1=1, X_2=1)$$

$$= P(X_1=1)P(X_2=1)=\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{36}$$

$$P(Y=2, X_1=1)=P(X_1=1, X_2=2)$$

$$= P(X_1=1)P(X_2=2)=\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{36}$$

$$P(X_1=2, Y=1)=0$$

$$P(X_1=2, Y=2)=P(X_1=2, X_2=1)$$

$$+P(X_1=2, X_2=2)=\frac{1}{36}+\frac{1}{36}=\frac{2}{36}$$

and so on. We find the following joint probability table of X and Y.

Y_1 \ X	1	2	3	4	5	6	marginal total
1	1/36	1/36	1/36	1/36	1/36	1/36	6/36
2	0	2/36	1/36	1/36	1/36	1/36	6/36
3	0	0	3/36	1/36	1/36	1/36	6/36
4	0	0	0	4/36	1/36	1/36	6/36
5	0	0	0	0	5/36	1/36	6/36
6	0	0	0	0	0	6/36	6/36
marginal total	1/36	3/36	5/36	7/36	9/36	11/36	1

$$E(Y)=1.\frac{1}{36}+2.\frac{3}{36}+3.\frac{5}{36}+4.\frac{7}{36}+5.\frac{9}{36}+6.\frac{11}{36}$$

$$=\frac{1}{36}(1+6+15+28+45+66)=\frac{161}{36}$$

$$E(Y^2)=1^2.\frac{1}{36}+2^2.\frac{1}{36}+3^2.\frac{5}{36}+4^2.\frac{7}{36}+5^2.\frac{9}{36}+6^2.\frac{11}{36}$$

$$=\frac{791}{36}$$

$$\text{Var } Y=E(Y^2)-[E(Y)]^2=\frac{791}{36}-\left(\frac{161}{36}\right)^2=\frac{2555}{1296}$$

$$E(X_1)=\frac{6}{36}(1+2+3+4+5+6)=\frac{126}{36}=\frac{21}{6}.$$

Cov $(X_1, Y)=E(X_1 Y)-E(X_1)\,E(Y)$

Where $E(X_1, Y)=1.\frac{1}{36}+2.\frac{1}{36}+3.\frac{1}{36}+4.\frac{1}{36}$

$+5.\frac{1}{36}+6.\frac{1}{36}+4.\frac{2}{36}+6.\frac{1}{36}+8.\frac{1}{36}$

$+10.\frac{1}{36}+12.\frac{1}{36}+9.\frac{3}{36}+12.\frac{1}{36}$

$+15.\frac{1}{36}+18.\frac{1}{36}+16.\frac{4}{36}+20.\frac{1}{36}$

$+24.\frac{1}{36}+25.\frac{5}{36}+3.\frac{1}{36}+36.\frac{6}{36}$

or $$E(X_1, Y)=\frac{616}{36}$$

Cov. $(x_1, y)=\frac{616}{36}-\frac{21}{6}.\frac{161}{6}=-\frac{2765}{36}$.

EXERCISE 5

1. (*a*) Define (*i*) a discrete random variable (*ii*), probability distribution (*iii*) distribution function.

(*b*) State and prove the important properties of distribution function.

(*c*) When the distribution function is called the 'step' function ?

(*d*) Let $x_1, x_2, \ldots\ldots$ be a countable number of points and $p(x_i)$ be the probability of x_i. Then show that, if F is the distribution of X,

$$F(x_i)-F(x_{i-1})=p(x_i).$$

2. A random variable X has the probability function

x :	0	1	2	3	4	5	6	7	8
$p(x)$:	a	$3a$	$5a$	$7a$	$9a$	$11a$	$13a$	$15a$	$17a$

(*i*) Determine the value of a.

(*ii*) Find $P(X<2)$, $P(X\geqslant 6)$, $P(3<X<5)$

(*iii*) Find the minimum value of a such that $P(X\leqslant 3)>\cdot 5$.

(*iv*) What is the distribution function of X ?

[**Ans.** (*i*) $\frac{1}{81}$ (*ii*) $\frac{4}{81}, \frac{45}{81}, \frac{1}{9}$ (*iii*) $\frac{1}{32}$]

3. From a lot containing 12 items, 4 of which are defective, 5 are chosen at random. If X be the number of defectives found in the sample, write down

(*i*) the probability distribution of X.

(*ii*) $P(X\leqslant 1)$, (*iii*) $P(1<X\leqslant 3)$.

(*iv*) Verify $\sum_{x=0}^{4} p(x)=1$.

[Assume that the sample is drawn without replacement.]

[**Ans.** (*i*) $p(x)=\binom{4}{x}\binom{8}{5-x}\div\binom{12}{5}$, $x=0, 1, 2, \ldots\ldots$

(*ii*) $P(X\leqslant 1)=P(X=0)+P(X=1)=\left[\binom{8}{5}+\binom{4}{1}\binom{8}{4}\right]\div\binom{12}{5}=$ etc

(*iii*) $P(1<X<3)=P(X=2)=\binom{4}{2}\binom{8}{3}\div\binom{12}{5}=$ etc.]

4. Define a random variable and its distribution. Obtain the distribution function of the total number of heads occurring in three tosses of an unbiased coin. *(B.Sc. Kerala 1970)*

5. Define the distribution function of a random variable. Let a coin with probability p for the occurrence of a head be tossed until a head appears. Let X denote the number of tails observed. Find $P(X=r)$ and hence obtain its distribution function. *(B.Sc. Kerala, 1971)*

[Ans. $P(X=r)=(1-p)^r p$]

6. A random variable X has the following probability function :

x :	0	1	2	3
$p(x)$:	0·1	0·2	0·3	0·4

Find the probability function of Y and the mean of Y where $Y=x^2+1$.

[Ans.

y :	1	2	5	10
$p(y)$:	0·1	0·2	0·3	0·4

mean of $Y=[0^2.(0\cdot1)+1^2(0\cdot2)+2^2(0\cdot3)+3^2(0\cdot4)]+1=6.$]

7. A can have any integral value from 0 to $a-1$, the chance of any value x varying as $a-x$, and B can have any integral value from 0 to $b-1$, the chance of any value y varying as $b-y$ $(a>b)$. If A and B are independent, find $P(A=B)$.

Hint. $P(A=B)=\sum_{x=0}^{b} P[(A=x)\cap(B=x)]=\sum_{x=0}^{b} P(A=x)P(B=x)$

where $P(A=x)=y_1(a-x)$

with $1=\sum_{x=0}^{a-1} P(A=x)=\sum_{x=0}^{a-1} y_1(a-x)$

$$=y_1\left[a\,.\,a-\frac{a(a-1)}{2}\right]=y_1\frac{a(a+1)}{2}$$

$\Rightarrow$ $y_1=2/a(a+1)$

and, $\therefore$ $P(A=X)=\dfrac{2(a-x)}{a(a+1)}$

Similarly, $P(B=y)=y_2(b-y)$ with $1=\sum_{y=0}^{b-1} y_2(b-y)$

giving $P(B=y)=\dfrac{2(b-y)}{b(b+1)}$.

8. An experiment consists of three independent tosses of a fair coin. Let X=the number of heads, Y=number of head runs, Z=length of head run, a head run being defined as consecutive occurrence of atleast two heads, its length then being the number of heads occurring together in three tosses of the coin.

(*a*) Find the probability function of the random variables (*i*) X (*ii*) Y (*iii*) Z (*iv*) $X+Y$ (*v*) XY.

(*b*) Write down the probability table of (*i*) $X+Y,=U$, (*ii*) $XY=V$.
(*B.Sc. Kerala 1969*)

[**Ans.**

Elementary Event	X	Y	Z	$X+Y$	XY
HHH	3	1	3	4	3
HHT	2	1	2	3	2
HTH	2	0	0	2	0
HTT	1	0	0	1	0
THH	2	1	2	3	2
THT	1	0	0	1	0
TTH	1	0	0	1	0
TTT	0	0	0	0	0

(*b*) (*i*)

u :	0	1	2	3	4
$p(u)$:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

(*ii*)

v :	0	1	2	3
$p(v)$:	$\frac{5}{8}$	0	$\frac{2}{8}$	$\frac{1}{8}$

9. Define for a discrete random variable, the following terms :
(*i*) Mathematical Expectation. (*B.A. Hons., Delhi, 1958*)
(*ii*) Cumulative Function. (*B.A. Hons., Delhi, 1956*)
(*iii*) Moment Generating Function. (*B.Sc., Agra*)
(*iv*) Probability Generating Function. (*B.Sc. Shivaji 1970*)
Show how they are related to each other.

10. (*a*) From an urn containing 3 red and 2 white balls, a man is to draw 2 balls at random without replacement, being promised 20 P. for each red ball he draws, and 10 P. for each white one. Find his expectation. [**Ans.** 32 P]

(*b*) If the roots of the equation $x^2-ax+b=0$ are real and b is positive but otherwise unknown ; what are the expectations and variance.

$$\left[\text{Ans. } \frac{a^2}{6}, \frac{a^2}{72}\right]$$

11. A man draws 2 balls from a bag containing 3 white and 5 black balls. If he is to receive Rs. 70 for every white ball which he draws and Rs. 7 for every black ball, find his expectation. [**Ans.** Rs. 61·25]

12. The face cards are discarded from two decks of cards ; then one card is drawn from each deck. What is the expectation of the sum of the numbers drawn ? [**Ans.** 11]

13. (*a*) On the average how many times must a die be thrown until one gets a 6 ? [**Ans.** 6]

(*b*) Four coins are tossed. What is the expectation of the number of heads ? [**Ans.** 2

(*c*) A coin is tossed until a head appears. Find the expectation of the number of tosses. [**Ans.** 2]

(*d*) Two unbiased dice are thrown. Find the expected value of the sum of numbers of points on them. [**Ans.** 7]

14. In a lottery m tickets are drawn out of n tickets numbered from 1 to n. What is the expectation and dispersion of the some of the numbers drawn ?
(*M.A. Bombay, 1952, 1953*)

$$\left[\text{Ans. } \frac{m(n+1)}{2}, \frac{m(n^2-1)}{12}\left(1-\frac{m-1}{n-1}\right)\right]$$

15. A bag contains a coin of value M and a number of other coins whose aggregate value is m. A person draws one at a time till he draws the coin of value M. Find the value of his expectation.

(M.A., Statistics Delhi, 1958 ; M.Sc., Agra 1963)

$\left[\text{Ans. } M+\frac{m}{2}\right]$

16. Balls are taken one by one out of an urn containing a white and b black balls until the first white ball is drawn. Show the expectation of the number of black balls preceding the first white ball is $b/(a+1)$.

(Indian Statistical Institute, Calcutta 1947)

Sol. Let x_1 denote the number of black balls preceding the first white ball. Let x_i denote the number of black balls following the $(i-1)th$ white balls and preceding the i-th white ball. Let x_{a+1} denote the number of black balls left after all the white balls have been drawn.

Then evidently, $x_1+x_2+\ldots\ldots+x_{a+1}=b$

and therefore, $E(x_1)+E(x_2)+\ldots\ldots+E(x_{a+1})=b.$

Now the probability of every sequence of balls *i.e.*, of every system of numbers $x_1, x_2, \ldots\ldots, x_{a+1}$ is the same *viz.*, $\frac{a!\,b!}{(a+b)!}$.

Thus $E(x_1)=E(x_2)=\ldots\ldots=E(x_{a+1})=M$, say

Hence $(a+1)M=b \quad \therefore \quad M=\frac{b}{a+1}$ **Ans.**

17. A box contains a white and b black balls ; c balls are drawn. Find the expectation of the number of white balls drawn. *(M.A. Punjab Sept., 1956)*

Sol. Let a variable x_i be defined as follows :

$x_i=1$ if i-th ball drawn is white, $i=1, 2, \ldots\ldots, c$

$=0$ if ,, ,, ,, ,, black.

Then the number of white balls s among the c balls drawn is given by

$$s=x_1+x_2+\ldots\ldots+x_c$$

Now $P(x_i=1)=\frac{a}{a+b}, \quad P(x_i=0)=\frac{b}{a+b}$

$\therefore \quad E(x_i)=1.\frac{a}{a+b}+0.\frac{b}{a+b}=\frac{a}{a+b}$

$\therefore \quad E(s)=\frac{a}{a+b}+\frac{a}{a+b}+\ldots+\frac{a}{a+b}=\frac{ca}{a+b}$.

18. An urn contains pN white and qN black balls, the total number of balls being N, $p+q=1$. Balls are drawn one by one (without being returned to the urn) until a certain number n of balls is reached. What is the dispersion of the number of white ball drawn ?

(M.A., Statistics, Delhi, 1961 ; M.A., Delhi, 1959)

Sol. Let $x_i=1$ if the i-th ball drawn is white

$=0$,, ,, ,, black

then we have

$E(x_i)=p$, $E(m)=np$, $E(x_i^2)=p$, $E(x_i-p)^2=E(x_i^2)-p^2=p-p^2=pq$, $q=1-p$

and $E(x_i-p)(x_j-p)=E(x_ix_j)-p^2=\frac{pN}{N}.\frac{pN-1}{N-1}-p^2.$

$$=p.\left[\frac{pN-1}{N-1}-p\right]=-\frac{pq}{N-1}, \quad i\neq j$$

$$\therefore \text{Dispersion} = E(m-np)^2 = E\left[\sum_{i=1}^{n}(x_i-p)^2 + 2\sum_{i\neq j}(x_j-p)(x_i-p)\right]$$

$$= \sum_{t=1}^{n} E(x_i-p)^2 + 2\sum_{i\neq j} E(x_i-p)(x_j-p)$$

$$= npq + 2.\ {}^nC_2\left(-\frac{pq}{N-1}\right) = npq\,\frac{N-n}{N-1} \quad \textbf{Ans.}$$

19. In a lottery containing n numbers (1, 2, ..., n), m numbers are drawn at a time. Let x_i represent the frequency of a specified number in N drawings. Prove that

$$E(x_i) = Np,\ E(x_i-Np)^2 = Npq.\ E(x_i-Np)(x_j-Np) = E(x_ix_j) - N^2p^2$$
$$= N^2p(p'-p),\ i \neq j$$

where $$p = \frac{m}{n}\ ,\ q = 1-p,\ p' = \frac{m-1}{n-1}\ .$$

(M. Sc., Madras, 1967)

Hint. Let Y_{ij} denote the random variable associated with the *j-th* draw such that

$$y_{ji} = \begin{cases} 1 \text{ if number } i \text{ comes on } j\text{-th draw} \\ 0 \text{ otherwise} \end{cases}$$

Then $X_i = \sum_{i=j}^{N_i} Y_{ji}$ is the frequency of a specified number i in N drawings.

Then $E(X_i) = E[\sum_{j=1}^{N} Y_{ji}] = \sum_{j=1}^{N} E(Y_{ji}) = \sum_{j=1}^{N}\left[1.\ \frac{m}{n} + 0.\left(1-\frac{m}{n}\right)\right] = \frac{m}{n}\ N = Np.$

Also $E(X_iX_j) = E\left[\left(\sum_{j=1}^{N} Y_{ji}\right)\left(\sum_{k=1}^{N} Y_{kj}\right)\right] = E\left[\sum_{k=1}^{N} Y_{ki}Y_{kj} + \sum_{\substack{k=1 \\ k\neq l}}^{N} Y_{ki}\,Y_{lj}\right]$

$$= \sum_{k=1}^{N} E\,Y_{kj}Y_{ki} = \sum_{k\neq l} E(Y_{kj}Y_{ki}) = \sum_{k=1}^{N} \frac{\binom{n-2}{m-2}}{\binom{n}{m}} + E(Y_{ki})E(Y_{kj}) = \text{etc.}]$$

20. Given a lot of size m containing s items of a specified kind. If items are to be drawn *without* replacement until i of the s items have been drawn, show that on the average $\frac{i(m+1)}{s-1}$ drawings will be necessary.

(I.S I. Calcutta 1954; M.A., Calcutta 1963)

21. A box containing k varieties of objects, the number of objects of each variety being the same. These objects are drawn one at a time and put back before the next drawing. Denote by n the smallest number of drawings which produce objects of all varieties find $E(n)$ and $E(n^2)$. Also for large k, find $E(n)$ and Var n.

***Sol.** Let x_{r+1} denote the number of drawings required to produce any of the $(k-r)$ varieties where r varieties have already been drawn. x_{r+1} is a stochastic variate taking the values 1, 2, 3, ..., with respective probabilities.

$$\frac{k-r}{k},\ \frac{r}{k}\cdot\frac{k-r}{k}\cdot\left(\frac{r}{k}\right)^2\cdot\frac{k-r}{k}\cdot\left(\frac{r}{k}\right)^3\cdot\frac{k-r}{k},\ \ldots\ldots$$

*This solution is due to Dr. V.N. Murti.

Hence by definition

$$E(x_{r+1})=\frac{k-r}{r}\sum_{n=1}^{\infty} n\left(\frac{r}{k}\right)^{n-1}=\frac{k-r}{k}\left(1-\frac{r}{k}\right)^{-2}=\frac{k-r}{k}\cdot\frac{k^2}{(k-r)^2}$$

$$=\frac{k}{k-r}$$

$$E(x_{r+1}^2)=\frac{k-r}{k}\sum n^2\left(\frac{r}{k}\right)^{n-1}=\frac{k(k+r)}{(k-r)^2}.$$

Since balls are replaced each time before fresh drawing is made the stochastic variates $x_1, x_2, \ldots, x_k$ are independent.

The last number of drawings required to produce all the k varieties is evidently equal to

$$n=x_1+x_2+\ldots+x_k.$$

$$\therefore\quad E(n)=E(x_1+x_2+\ldots+x_k)=\sum_{r=0}^{k-1}E(x_{r+1})=\sum_{r=0}^{k-1}\frac{k}{k-r}=k\left(1+\tfrac{1}{2}+\ldots+\frac{1}{k}\right)$$

Similarly $$E(n^2)=k^2\left(\frac{1}{1^2}+\frac{1}{2^2}+\ldots+\frac{1}{k^2}\right)-k\left(1+\frac{1}{2}+\ldots+\frac{1}{k}\right)$$

$$+k^2\left(1+\frac{1}{2}+\ldots\ldots+\frac{1}{k}\right)^2.$$

For large n, $$E(n)=\frac{(r-1)k}{k-r+1}-k\log_{\frac{k}{k-r+1}}.$$

22. A bag contains $2n$ contents, of which half are marked with odd numbers and half with even numbers, the sum of all the numbers being S. A man is to draw counters. If the sum of the numbers drawn is odd, he is to receive that number of rupees, if even, he is to pay that number of rupees. Show that the expectation is $S/n(2n-1)$ rupees.

23. A vessel contains tickets numbered 1 to N. Find the expectation and variance of the largest numbers X drawn in n drawings when random sampling with replacement is used. Also find for large N and fixed k, $E(X)$ and $Var\,(X)$.

[**Hint.** The event $X\leqslant k$ means that each of n number drawn is less than or equal to k so that $P(X\leqslant k)=(k/N)^n$.

$$p_k=P(X=k)=P(X\leqslant k)-P(X\leqslant k-1)=\left(\frac{k}{N}\right)^n-\left(\frac{-1}{N}\right)^n.$$

$$\therefore\quad E(X)=\sum_{k=1}^{N}kp_k=\sum_{k=1}^{N}\frac{k}{N^n}\,[k^n-(k-1)^n]=\sum_{k=1}^{N}\frac{1}{N^n}\left[k^{n+1}-k(k-1)^n\right]$$

$$=\frac{1}{N^n}\sum_{i=1}^{N}[k^{n+1}-(k-1)^{n+1}-(k-1)^n]$$

$$=\frac{1}{N^n}\left[N^{n+1}-\sum_{k=1}^{N}(k-1)^n\right]\sim\frac{1}{N^n}\left[N^{n+1}-\int_0^N t^n\,dt\right]=\frac{nN}{n+1}$$

24. (*a*) With usual notation,* prove that $\dfrac{\partial\mu_r'}{\partial k_j}={}^rC_j\,\mu_{r-j}'$

(*b*) Deduce that $\mu_2=k_2$, $\mu=k_3$, $\mu_4=k_4+3k_2^2$. (*M.A., Madras, 1959*)

[**Hint.** (*a*) Differentiate the following identity with respect to k_j and identify powers of t.

$$\exp\left\{k_1t+\frac{k_2t^2}{2!}+\ldots\ldots+\frac{k_rt^r}{r!}+\ldots\right\}=1+\mu_1't+\frac{\mu_2't^2}{2!}+\ldots\ldots+\frac{\mu_r't^r}{r!}+\ldots$$

(*b*) Put $j=1$ in the result of (*a*)].

25. (*a*) Assuming the additive property of cumulants prove that the variance of the sum of a number of independent variates is the sum of their variances.

(*b*) Let x_1 and x_2 be two stochastic random variables having variance k and 2 respectively. If the variance of $y=3x_2-x_1$ is 25, find k. [**Ans.** 7]

(*c*) Suppose that X is a random variable for which $E(X)=10$ and Var $X=25$. Find the positive values of a and b such that $Y=aX-b$ have expectation 0 and variance 1. (*B.Sc., Sardar Patel, 1970*)

[**Ans.** $a=\frac{1}{5}$, $b=2$]

(*d*) Show that if $\bar{x}$ is the mean of n independent stochastic variables, then

$$M_{\bar{x}}(t)=\left[M_X\left(\frac{t}{n}\right)\right]^n$$

and if X's are identically distributed independent stochastic variables then

$$\sum_{i=1}^{n} M_{X_i}(t)=[M_{X_i}(t)]^n$$

26. Show that the variance of the weighted mean

$$\sum_{i=1}^{n} w_ix_i \Big/ \sum_{i=1}^{n} w$$

of n independent stochastic variates x_i is a minimum when the weights are inversely proportional to the corresponding variances. Also show that this minimum variance is H/n when H is the harmonic mean of the variance σ_i^2, $i=1, 2, \ldots, n$.

(*M.A., Statistics, Delhi, 1960* ; *M.A., Patna, 1955*)

[**Hint.** $\sigma^2=$variance of the weighted mean$=\frac{\Sigma w_i^2\sigma_i^2}{(\Sigma w_i)^2}$.

For minimum of σ^2, $\frac{d\sigma^2}{dw_i}=0$, which gives $w_i\sigma_i^2=\frac{\Sigma w_i^2\sigma_i^2}{\Sigma w_i}$

showing that $w_i\sigma_i^2$ is the same for all values of i, *i.e.*, $w_i \propto (1/\sigma_i^2)$. Hence the result.]

27. In a sequence of Bernoullian trials in each of which the probability of success is p, show that

(*a*) (*i*) the expected value of X_1, the number of trials up to and including the first success in $1/p$; and that of x_k, the number of trials up to and including k-th success is also $1/p$;

(*ii*) the expected value of $X_1+X_2+\ldots+X_n$ of the number of trials up to and including the nth success is $\frac{n}{p}$ and variance is $\frac{n(1-p)}{p^2}$.

(*b*) if X be the length of run of either success or failures starting with the first trial, show that

$$E(x)=\frac{q}{p}+\frac{p}{q} \text{ and var } (x)=\left(\frac{p}{q}-\frac{q}{p}\right)^2+\left(\frac{p}{q}+\frac{q}{p}\right).$$

[**Hint.** (*b*) $P(x=r)=P(SS\ldots SF)+P(FF\ldots FS)=p^rq+q^rp$.
$\quad\quad\quad

The difference equation gives us $M_x = a+b-b\left(1-\frac{1}{a+b}\right)^x$.

32. An urn contains n tickets numbered from 1 to n. These tickets are drawn one by one so that a certain number appears in the first place, another number in the second place and so on. Find the mathematical expectation of the number of coincidence (we shall say that there is a coincidence when the number on a ticket corresponds to the place it occupies). **[Ans. 1]**

33. (*a*) Show that if X is non-negative

$$E\left(\frac{1}{X}\right) \geq \frac{1}{E(X)}. \qquad \textit{(M.A., Statistics, Delhi, 1970)}$$

(*b*) If x, y are independent continuous random variables with mean $\mu(x)$ and $\mu(y)$ and variances σ_x^2 and σ_y^2 respectively, show that

(*i*) $\sigma^2(xy) = \sigma^2(x)\,\sigma^2(y) + \mu^2(x)\,\sigma^2(y) + \mu^2(y)\,\sigma^2(x)$.

(M.A., Statistic s, Delhi, 1969)

[**Hint.** $\sigma_{xy}^2 = E\{(x-\mu(x))(y-\mu(y)) + y\mu(x) + x\mu(y) - 2\mu(x)\mu(y)\}^2 = \text{etc.}$]

(M.A., Statistics, Delhi, 1970)

(*ii*) $\dfrac{Var\,(xy)}{[E(x)]^2[E(y)]^2} = C_x^2 C_y^2 + C_x^2 + C_y^2$

where $C_x = \sqrt{Var\,x}/E(x) = \sigma_x/\mu(x)$, $C_y = \sqrt{Var\,y}/E(y) = \sigma_y/\mu(y)$

(B.Sc., Poona, 1967)

34. Let x be the number of failure preceding the kth success in an infinite set of Bernoullian trials with constant probability p of success. Find the probability function $g(x)$ of x and also its mean and variance. Further show that

$$\sum_{x=0}^{n-k} g(x) = \sum_{t=k}^{n} \binom{n}{t} p^t q^{n-t}.$$

(Bombay, B.Sc., 1969)

35. A random variable x assumes the values $\lambda_1, \lambda_2, \ldots\ldots$, with probabilities $\mu_1, \mu_2, \ldots\ldots$, show that

$$p_k = \frac{1}{k!} \sum_{j=0}^{\infty} \mu_j e^{-\lambda_j} (\lambda_j)^k,\ \lambda_j > 0,\ \sum_{j=0}^{\infty} \mu_j = 1,\ k = 0, 1, \ldots$$

is a probability distribution. Find its generating function and prove that its mean equals $E(x)$ and its variance equals $Var\,(x) + E(x)$.

(M.Sc., Lucknow, 1967)

$$\left[\text{Ans. } P(s) = \sum_{0}^{\infty} p_k s^k = \sum_{i=0}^{\infty} \lambda_i e^{\lambda_j (s-1)}\right]$$

36. Suppose that shipment involving a varying number of parts arrive each day. If N is the number of items in the shipment the probability distribution of the random variable N is given as :

n :	10	11	12	13	14	15
$P(N=n)$:	·05	·10	·10	·20	·35	·20

The probability that any particular part is defective is the same for all parts and equal to 0·10. If X is the number of defective parts arriving each day what is the expected value of X? [Ans. 1·33]

[**Hint.** $E(X)=E[E(X(N)]=E(0{\cdot}10N)=$ etc.]

37. In a fairy tale, the fairy grandmother assumes the queen that her baby prince will not die until the following condition is fulfilled:

A scroll is prepared and on the day of the birth and every subsequent birthday a letter of the Greek alphabet, chosen at random, is entered on the scroll. On the day that all the twentyfour letters of the alphabet appear, he will die. Show that the expectation of his life is 89·6 years nearly.

[It is given that $\log_e 24=3{\cdot}17805$ and

$$\lim\left[1+\frac{1}{2}+\ldots\ldots+\frac{1}{n}-\log x\right]=\gamma \text{ (Euler's constant)}=0{\cdot}57721)$$

(M.A., Statistics, Delhi, 1968)

38. Let X and Y be two random variables each taking three values -1, 0, 1 and having the joint probability distribution :

Y \ X	−1	0	1	Total
−1	0	·1	·1	·2
0	·2	·2	·2	·6
1	0	·1	·1	·2
Total	·2	·4	·4	1·0

(*a*) Find $E(X)$, $E(Y)$, Var X and Var Y. Are X and Y independent?

(*b*) Given that $Y=0$, what is the conditional probability distribution of X?

(*c*) Find Var $(Y \mid X=-1)$. *(B.Sc. Poona, 1967 ; J. & K. U. M.Sc., 1968)*

Ans. $E(X)=0$, $E(Y)=0$, $E(XY)=0 \Rightarrow E(XY)=EXEY \Rightarrow X$ and Y are independent. Var $X=·4$, Var $Y=·56$,

$P(X=-1 \mid Y=0 \mid =P(X=-1 \cap Y=0) \mid P(Y=0)=·2{\cdot}6=\frac{1}{3}$

$P(X=0 \mid Y=0 \mid =\frac{1}{3}$, $P(X=1 \mid Y=0)=\frac{1}{3}$

Var $(Y \mid X=-11)=E(Y \mid X=-11)]^2-[E(Y \mid X=-11)]^2=0-0^2=0.$

39. Two stochastic variables X and Y take the values 1, 2, 3 and their probabilities are as follows :

Y \ X	*1*	*2*	*3*	*Total*
1	0·1	0·1	0·1	0·3
2	0·1	0·2	0·1	0·4
3	0·1	0·1	0·1	0·3
Total	0·3	0·4	0·3	1

Find $E(X)$, $E(Y)$, Var X, Var Y, EXY, $E(X+Y)$ and Cov (X, Y). Determine whether X and Y are independent.

(*B.Sc. Poona 1966* ; *M.Sc. Madras 1969*)

[**Ans.** $EX=2$, $EY=2$, Var $X=·6$, Var $Y=·6$, $E(X+Y)=4$,]

6

Special Discrete Probability Distributions

6·1. Introduction. In chapter 5, we explained what is meant by discrete probability distribution. In this chapter, we shall study some important discrete probability distributions.

6·2. *Binomial Distribution. Suppose that a trial is repeated so that we have a series of n trials. Let us call the occurrence of an event a 'success' and its non-occurrence a 'failure'. Let p be the probability of a success and q be the probability of a failure in a single trial, so that $p+q=1$. We shall assume that the trials are *independent* and the probability p of successes is the same in every trial. The number of successes in n trials may be 0, 1, 2,....., r, ..., n and is obviously a chance variate. Let us find the probability that the variate takes a particular value, say r. By the theorem of compound probability, the probability that the first r trials are successes and the remaining $n-r$ are failures that is, the sequence of successes and failures is $\underbrace{S\,S.....S}_{r}\ \underbrace{F\,F.....F}_{n-r}$, where S denotes a success and F a failure is $p^r\,q^{n-r}$. Clearly this is also the probability for the r successes and $n-r$ failures to occur in *any particular-definite specified order*. But we are interested in any r trials being successes and since r trials can be chosen out of n in $\binom{n}{r}$ mutually exclusive

---

*The binomial distribution was discovered early in the eighteenth century by the Swiss mathematician, Jakob Bernoulli (1654—1705). It was published posthumously in 1713.

ways, by the theorem of total probability, the chance $P(r)$ of r successes in a series of n independent trials is given by

$$P(r)=\binom{n}{r} p^r q^{n-r}=b(r\,;\,n,\,p)=f(r) \qquad \ldots(6{\cdot}1)$$

Thus the chance variate, *i e.*, the number of successes, can take the values 0, 1, 2,....., r,... . ..., n with corresponding probabilities

$$q^n,\ \binom{n}{1} q^{n-1}p,\ \binom{n}{2} q^{n-2}p^2,\ldots\ldots,\ \binom{n}{r} q^{n-r}p^r,\ldots\ldots,\ p^n.$$

The probability distribution of the number of successes so obtained is called the binomial probability distribution for the obvious reason that the probabilities are the various terms in the binomial expansion of $(q+p)^n$.

The sum of the probabilities

$$=\sum_{r=0}^{n} b(r\,;\,n,\,p)=\sum_{r=0}^{n}\binom{n}{r} p^r q^{n-r}=(q+p)^n=1.$$

The binomial distribution contains two independent constants, *viz*, n and p (or q). They are called *parameters* of the binomial distribution. The distribution function is given by

$$F(r)=\sum_{0<t<x}\binom{n}{t} p^t q^{n-t}$$

If $p=q=\frac{1}{2}$, the binomial distribution is *symmetrical* distribution and when $p\neq q$, it is a skew distribution.

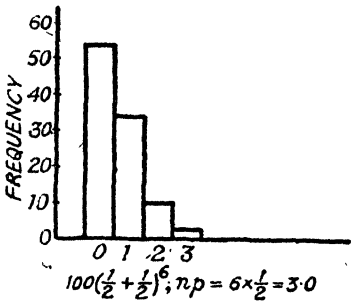


Fig. 6·1 (*a*)

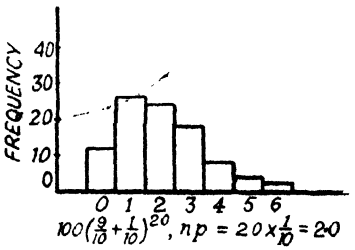


Fig. 6·1 (*b*)

In figures given above are plotted two binomials for

(*a*) $p={\cdot}5$, $\quad n=6$

(*b*) $p={\cdot}1$, $\quad n=20$

Ex. 1. Show that the probability of at least r successes in n Bernoullia trials is $\sum_{x=r}^{n}\binom{n}{x}p^x q^{n-x}$ and that of at most r successes is $\sum_{x=0}^{r}\binom{n}{x}p^x q^{n-x}$

Plot these probabilities against r for $n=6$, $p=\cdot5$ and $n=8$, $p=\frac{1}{3}$.

Ex. 2. A and B play a game in which A's chance of winning is $\frac{2}{3}$. In a series of 8 games what is the chance that A will win at least 6 games ? **Ans.** $\frac{1024}{2187}$

6·2·1. Binomial Frequency Distribution. Let the n independent trials constitute one experiment and let this experiment be repeated N times. Then we expect r successes to occur $N\binom{n}{r}p^r q^{n-r}$ times. This will be called the *expected frequency* of r successes in N experiments and the possible number of successes together with the expected frequencies will be said to constitute the binomial (expected) frequency distribution. In practice, the observed frequencies will differ from the expected frequencies due to chance causes. For N sets, each of n trials the expected frequencies of 0, 1, 2,..., n successes are given by the successive terms in the binomial expansion of $N(q+p)^n$, $q+p=1$.

Example 6·1. *Find the most probable number of successes in a series of n independent trials, the probability of success in each trial being p.* *(B.A., Hons. Delhi, 1949, 1956)*

We are to find the number of successes which has a greater probability than any other.

The chance of r successes will be greater than or equal to that of $r-1$ or $r+1$ successes if

$$b(r-1\ ;\ n,\ p)\leqslant b(r\ ;\ n,\ p)\geqslant b(r+1\ ;\ n,\ p)$$

or $${}^nC_{r-1}p^{r-1}q^{n-r+1}\leqslant {}^nC_r p^r q^{n-r}\geqslant {}^nC_{r+1}p^{r+1}q^{n-r+1},\ q=1-p$$

Simplifying, we get $\frac{r}{n-r+1}\cdot\frac{q}{p}\leqslant 1\geqslant \frac{n-r}{r+1}\cdot\frac{p}{q}$.

i.e., if $rq\leqslant np-rp+p$ and $rq+q\geqslant nq-rp$

i.e., if $np-q\leqslant r\leqslant np+p$ or $(n+1)p-1\leqslant r\leqslant np+p$. ...(6·2)

Case I. $(n+1)p=k$ (an integer) then $P(r)$ increases till $r=k$ and then $P(r)=P(r+1)$ and after that it begins to decrease.

Thus if $p=\frac{2}{3}$ and $n=8$, then $(n+1)p-1=9\times\frac{2}{3}-1=5$ so that 5 and 6 are the most probable numbers.

Case II. $np=k^1+f$ where k^1 is an integer and f is a proper fraction, $P(r)$ is maximum when $r=[np+p]$ where $[x]$ means the integral part of x.

Thus if $p=\frac{2}{3}$, the most probable number of successes in 10 trials is $[\frac{20}{3}+\frac{2}{3}]$, that is 7.

The most probable number of successes obviously gives the *mode* of the Binomial distribution and its value is $np+p$. If

$(n+1)p-1$ is an integer, the distribution is bimodal and the second mode is at $[(n+1)p-1]$.

Ex. Find the most probable number of heads in 99 tossings of a biased coin given that the probability of a head in a single tossing is $\frac{2}{5}$.

(M.A., Delhi, 1961 ; M.A., Punjab, 1959)

[Ans. 60]

6·2·2. First Four Moments of the Binomial Distribution. Take an arbitrary origin at 0 success. Hence

$$\mu_1' = \sum_{r=0}^{n} r.{}^nC_r p^r q^{n-r}$$

$$= 0.q^n + 1.{}^nC_1 q^{n-1}p + 2.{}^nC_2 q^{n-2}p^2 + \ldots\ldots + np^n$$

$$= np[q^{n-1} + {}^{n-1}C_1 q^{n-2}p + {}^{n-1}C_2 q^{n-3}p^2 + \ldots\ldots + p^{n-1}]$$

$$= np(q+p)^{n-1} = np. \qquad \ldots(6\cdot3)$$

$$\mu_2' = \sum_{r=0}^{n} r^2.{}^nC_r p^r q^{n-r}$$

$$= \sum_{r=0}^{n} [\{r + r(r-1)\}]\ {}^nC_r p^r q^{n-r}]$$

$$= \sum_{r=0}^{n} r\,{}^nC_r p^r q^{n-r} + \sum_{r=0}^{n} r(r-1)\ {}^nC_r p^r q^{n-r}$$

$$= np + n(n-1)p^2 \sum_{r=2}^{n} {}^{n-r}C_{r-2}\ p^{r-2}q^{n-r}$$

$$= np + n(n-1)p^2(q+p)^{n-2}$$

$$= np[1 + (n-1)p] = np(q+np)$$

$$= npq + n^2p^2$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = npq. \qquad \ldots(6\cdot4)$$

$$\mu_3' = \sum_{r=0}^{n} r^3 . {}^nC_r\ p^r q^{n-r}$$

Writing r^3 as $r(r-1)(r-2) + 3r(r-1) + r$ and proceeding in a similar manner, we have

$$\mu_3' = n(n-1)(n-2)p^3(q+p)^{n-3} + 3n(n-1)p^2(q+p)^{n-2} + np(q+p)^{n-1}$$

$$= n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= npq(q-p) \text{ on simplification.} \qquad \ldots(6\cdot5)$$

Writing r^4 as $r(r-1)(r-2)(r-3) + 6r(r-1)(r-2) + 7r(r-1) + r$, we get

$$\mu_4' = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)$$

$$\left(\begin{matrix} n \\ x \end{matrix} \right) p^x q^{n-x} = np.$$

Easily we find

$$\mu_4=3n^2p^2q^2+npq(1-6pq)=npq[1+3(n-2)pq)]. \quad ...(6\cdot6)$$

Hence $\beta_1=\dfrac{\mu_3^2}{\mu_2^3}=\dfrac{(q-p)^2}{npq}=\dfrac{(1-2p)^2}{npq}$, $\gamma_1=\dfrac{1-2p}{\sqrt{npq}}$

$$\beta_2=\frac{\mu_4}{\mu_2^2}=3+\frac{1-6pq}{npq}, \quad \gamma_2=\frac{1-6pq}{npq}.$$

Note. (1) The skewness is positive for $p<\frac{1}{2}$, negative for $p>\frac{1}{2}$ and zero for $p=\frac{1}{2}$.

(2) As n the number of trials increases indefinitely, $\beta_1\to0$ and $\beta_2\to3$; both γ_1 and γ_2 tend to zero.

6·2·3. Recurrence Relation. We have

$$b(x\ ;\ n,\ p)=f(x)=\frac{n\ !}{x\ !\ (n-x)\ !}p^xq^{n-x}$$

and $$b(x+1\ ;\ n,\ p)=f(x+1)=\frac{n\ !}{(x+1)\ !\ (n-x-1)\ !}p^{x+1}\ q^{n-x-1}$$

so that $$f(x+1)=\frac{n-x}{x+1}\cdot\frac{p}{q}f(x) \quad ...(6\cdot6)$$

Thus $f(x+1)$ can be obtained from $f(x)$ if we know n and p.

Example 6·2. Fittig Binomial Distribution. *The following data are the number of seeds germinating out of 10 on damp filter paper for 80 sets of seeds. Fit a binomial distribution to these data* :

x :	0	1	2	3	4	5	6	7	8	9	10	Total
f :	6	20	28	12	8	6	0	0	0	0	0	80

(*B.Sc., Agra, 1955*)

Here $n=10$, $N=80$ and $\Sigma f_i=80$

Also the arithmetic mean $=\dfrac{\Sigma f_ix_i}{\Sigma f_i}=\dfrac{174}{80}$

i.e., by (6·3) $$np=\frac{174}{80}$$

$$\therefore\quad p=\frac{174}{800}=0\cdot2175, \qquad q=1-p=\cdot7825.$$

Hence the binomial distribution to be fitted $80(\cdot7825+\cdot2175)^{10}$. From (6·6′) or the successive terms in the expansion of this give the expected or theoretical frequencies of 0, 1, 2, ..., 10 successes. These are approximately

6·9, 19·1, 24·0, 17·8, 8·6, 2·9, 0·7, 0·1, 0, 0, 0.

Example 6·3. *Prove for the binomial distribution the following recurrence relation :* $\mu_{r+1}=pq\left(nr\mu_{r-1}+\frac{d\mu_r}{dp}\right)$.

(M.A., Calcutta, 1950 ; M.A., Punjab, 1953 S ; M.A., Statistics, Delhi 1959 ; B.A., Hons & M.A., Madras, 1960 ; B.A., Hons., Delhi 1952)

By means of this relation, find μ_2, μ_3 *and* μ_4.

(B.A., Hons., Delhi 1952, 65)

We have $\mu_r=\sum_{j=0}^{n} f_j(j-\mu_1')^r=\sum_{j=0}^{n} {}^nc_j p^j q^{n-j}(j-np)^r$

Hence $\frac{d\mu_r}{dp}=\Sigma {}^nc_j jp^{j-1}q^{n-j}(j-np)^r$

$$-\Sigma(n-j){}^nc_j p^j q^{n-j-1}(j-np)^r$$
$$-nr\Sigma {}^nc_j p^j q^{n-j}(j-np)^{r-1}$$

$$=\sum_{j=0}^{n}[{}^nc_j p^{j-1}q^{n-j-1}(j-npj^r(j-np+jp)]-nr\mu_{r-1}$$

$\therefore$ $pq\frac{d\mu_r}{dp}=\sum_{j=0}^{n} {}^rc_j p^j q^{n-j}(j-np)^{r+1}-nprq\mu_{r-1}$

or $\mu_{r+1}=pq\left(\frac{d\mu r}{dp}+nr\mu_{r-1}\right)$.

Putting $r=1, 2, 3$ successively, we get

$$\mu_2=pq\left[\frac{d\mu_1}{dp}+n\cdot 1\ \mu_0\right]=pq(0+n)=npq, \text{ since } \mu_0=1$$

$$\mu_3=pq\left[\frac{d\mu_2}{dp}+n\cdot 2\mu_1\right]=pq[nq-np+2n.0]=npq(q-p)$$

$$\mu_4=pq[n(6p^2-6p+1)+3n.np(1-p)]$$
$$=npq[1-6p(1-p)+3npq]=3n^2p^2q^2+npq(1-6pq).$$

Example 6·4. *A random sample of size n is drawn from an infinite population in which the individuals fall into one or the other of the two categories with probability p and 1−p. The numbers in the sample falling into the two categories are* n_1 *and* n_2 $(n_1+n_2=n)$. *Show that the covariance between* n_1 *and* n_2 *is* $-np(1-p)$. *Hence obtain the variance of the difference in the proportion* $\frac{n_1}{n}-\frac{n_2}{n}$.

(*a*) Clearly n_1 is a Binomial variate.

$\therefore$ $E(n_1)=np$, $E(n_2)=E(n-n_1)=n-E(n_1)=n-np=nq$.

$\therefore$ $Cov(n_1, n_2)=E(n_1-np)(n_2-nq)=E(n_1-np)(n-n_1-n+np)$
$=-E(n_1-np)^2=-Var(n_1)=-npq=-np(1-p)$.

$$(b)\ Var\left(\frac{n_1}{n}-\frac{n_2}{n}\right)=Var\ \frac{n_1}{n}+Var\left(\frac{n_2}{n}\right)-2\ Cov\left(\frac{n_1}{n},\ \frac{n_2}{n}\right)$$

$$=\frac{1}{n^2}\ Var\ (n_1)+\frac{1}{n^2}\ Var\ (n_2)-\frac{2}{n^2}\ Cov\ (n_1,\ n_2)$$

$$=\frac{1}{n^2}\cdot\ [npq+npq+2npq]=\frac{4pq}{n}\cdot$$

Ex. If x is binomial variate with the probability law

$$P(x)={}^nC_x p^x q^{n-x},\ q=1-p,\ x=0,\ 1,\ 2,\ldots,\ n$$

prove that $Cov\left(\frac{x}{n},\ \frac{n-x}{n}\right)=\frac{-pq}{n}$ *(M.A., Statistics, Delhi, 1959)*

6·2·3. Moment Generating Function and Cumulative Function of the Binomial Distribution.

The probability generating function $=\sum_{x=0}^{n} t^x\,{}^nC_x p^x q^{n-x}=(q+pt)^n$.

M.G.F. about origin is given by

$$M_0(t)=E(e^{tx})=\sum_{x=0}^{n} e^{tx}\,{}^nC_x\,p^x q^{n-x}=\sum_{x=0}^{n}\,{}^nC_x\,q^{n-x}(pe^t)^x$$

$$=(q+pe^t)^n \qquad \ldots(6\cdot57)$$

Differentiating $M_0(t)$ once, twice, etc., with respect to t and putting $t=0$ and using (5·12), we get

$$\mu_1'=np \qquad \mu_2'=npq+n^2p^2,\ \mu_2=npq$$

$$\mu_3=npq(q-p),\qquad \mu_4=npq[1+3(n-2)pq].$$

Since $M_a(t)=e^{-at}M_0(t)$ we get the *m.g.f.* of the binomial distribution with respect to its mean np.

$$M\mu_1'(t)=e^{-npt}\ (q+pe^t)^n=(qe^{-pt}+pe^{-qt})^n.$$

Writing the values of the exponentials and arranging, we get

$$M\mu'_1(t)=\left[1+pq\frac{t^2}{2\,!}+pq(q^2-p^2)\frac{t^3}{3\,!}+pq(q^3+p^3)\frac{t^4}{4\,!}+\ldots\right]^n$$

Expanding the right side and equating the coefficients of like powers of t on either side, we deduce that

$$\mu_2=npq,\ \mu_3=npq(q-p),\ \mu_4=npq[1+3(n-2)pq]$$

The cumulative function of the binomial distribution about its mean np is given by

$$K(t)=\log\ (qe^{-pt}+pe^{-qt})^n$$

$$=n\ \log\left[1+pq\frac{t^2}{2\,!}+pq(q^2-p^2)\frac{t^3}{3\,!}+pq(q^3+p^3)\frac{t^4}{4\,!}+\ldots\right]$$

$$=n\left[pq\frac{t^2}{2\,!}+pq(q^2-p^2)\frac{t^3}{3\,!}+pq(q^3+p^3)\frac{t^4}{4\,!}+\ldots\right]$$

$$-\frac{n}{2}\left[pq\frac{t^2}{2\,!}+\ldots\ldots\right]^2+\ldots\ldots$$

Equating the coefficients of $\frac{t^2}{2!}$, $\frac{t^3}{3!}$, $\frac{t^4}{4!}$ on either side we get

$$k_2=npq\ ;\ k_3=npq\,(q^2-p^2)=npq(q-p),$$
$$k_4=npq(q^3+p^3)-3np^2q^2=npq[(q+p)^3-3pq(q+p)-3pq]$$
$$=npq(1-6pq).$$

Example 6·5. *Show that the factorial moment generating function $w(t)$ of the binomial distribution $(q+p)^n$ is $(1+pt)^n$. Hence or otherwise, show that $\mu'_{(r)}=n^{(r)}p^r$.* (*M.A. Delhi, 1959*)

First Method. $w(t)=E[1+t]^x=\sum_{x=0}^{n} {}^nC_x p^x q^{n-x}.(1+t)^x$

$$=[q+p(1+t)]^n=(1+pt)^n$$

$\therefore$ $\mu'_{(r)}=$ *coeff.* of $t^r/r!$ in the expansion of $w(t)$

$$=n^{(r)}p^r.$$

Second Method. Differentiating $(q+p)^n$ r times partially with respect to p and multiplying by p^r, we have

$$n^{(r)}(q+p)^{n-r}p^r=\sum_{j=r}^{n} j^{(r)}{}^nC_j p^j q^{n-j}=\mu'_{(r)}$$

or $$\mu'_{(r)}=n^{(r)}p^r.$$

Thus $\mu'_{(1)}=np$, $\mu'_{(2)}=n(n-1)p^2$, $\mu_{(2)}=\mu'_{(2)}-\mu'^2_{(1)}=npq$.

6·3. *Poisson Distribution. The binomial distribution assumes a very convenient form in the limiting case when p (or q) is very small and n is large and the average number of successes np (or nq)$=m$ is a finite constant.

In the binomial distribution, the probability of r successes is given by

$$P(r)={}^nC_r p^r q^{n-r}$$

We write this as

$$P(r)=\frac{n!}{r!\,(n-r)!}\left(\frac{m}{n}\right)^r\left(1-\frac{m}{n}\right)^{n-r} \qquad \left[\because\ p=\frac{m}{n}\right]$$

$$=\frac{m^r}{r!}\left(1-\frac{m}{n}\right)^n\cdot\frac{\left(1-\frac{m}{n}\right)^{-r}}{n^r}\cdot\frac{n!}{(n-r)!}$$

By making n tend to infinity, we have

*It was first discovered by a French Mathematician S.D. Poisson in 1837. It is also sometimes known as "The Law of Small Numbers".

$$\lim_{n\to\infty} P(r)=\frac{m^r}{r\,!}\,e^{-m}\lim_{n\to\infty}\left[\frac{1}{n^r}\cdot\frac{n\,!}{(n-r)\,!}\right]$$

Using *Stirling's formula for n !, *viz*,

$$n\,!\sim\sqrt{2\pi}e^{-n}n^{n+1/2}$$

where e is the Naperian base, 2·71828..., we have

$$\lim_{n\to\infty} P(r)=\frac{m^re^{-m}}{r\,!}\lim_{n\to\infty}\frac{\sqrt{2\pi}e^{-n}n^{n+1/2}}{n^r\sqrt{2\pi}e^{-n+r}(n-r)^{n-r+1/2}}$$

$$=\frac{m^re^{-m}}{r\,!\,e^r}\lim_{n\to\infty}\left\{\frac{1}{\left(1-\frac{r}{n}\right)^n\left(1-\frac{r}{n}\right)^{-r+1/2}}\right\}$$

$$=\frac{m^re^{-m}}{r\,!}\text{, since } e^r\,.\,e^{-r}=1,\ r \text{ being finite} \quad \ldots(6{\cdot}6)$$

The chances of 0, 1, 2,.....successes in the finite sequence are e^{-m}, me^{-m}, $\frac{m^2e^{-m}}{2\,!}$, ...respectively. In other words the chance variate, the number of successes, assumes the values 0, 1, 2,....., r..., with the corresponding probabilities e^{-m}, me^{-m}, $\frac{m^2e^{-m}}{2\,!}$,..., $\frac{m^re^{-m}}{r\,!}$,...

The probability distribution of the number of successes (with the assumptions mentioned above) is called the Poisson distribution.

The sum of the probabilities $=\sum_{r=0}^{\infty}\frac{m^re^{-m}}{r\,!}=e^{-m}e^m=1$.

The Poisson distribution involves only one parameter *viz.*, m, Fig. 6·2 shows the Poisson distributions from $m=3$ and $m=10$.

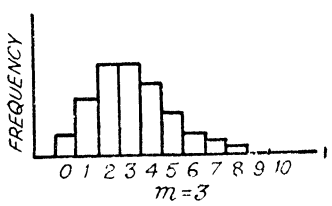


Fig. 6·2 (*a*)

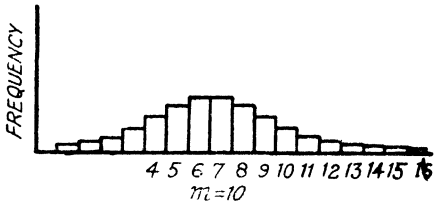


Fig. 6·2 (*b*)

The following are some examples of Poisson variates :

(1) The number of typographical errors per page in type material.

(2) The number of defective screws per box of 100 screws

(3) The number of bacterial colonies in a given culture per unit area of microscope slab.

*Stirling formula which states that the limiting value of the ratio of the two exprssions, as n tends to infinity is unity, was published in 1730.

(4) The number of deaths in a district in one year by a rare disease.

Ex. Show that $\lim_{n\to\infty} \frac{n!}{(n+r)!} \cdot \frac{1}{n^r} = \lim_{n\to\infty} \frac{n(n-1)\ldots(n-r+1)}{n \cdot n \cdot n}$

$$= \lim_{n\to\infty} \left(1-\frac{1}{n}\right) \ldots \left(1-\frac{r-1}{n}\right) = 1 .$$

Hence show how the result of section 6·3 can be deriyed without using Stirling formula.

Alternative Derivation of Poisson Distribution.

$$\lim_{n\to\infty} P(0) = \lim_{n\to\infty} q^n = \lim_{n\to\infty} (1-p)^n = \lim_{n\to\infty} \left(1-\frac{m}{n}\right)^n = e^{-m}$$

$$\text{Also } \lim_{n\to\infty} \frac{P(r+1)}{P(r)} = \lim_{n\to\infty} \frac{n-r}{r+1} \cdot \frac{p}{q} = \lim_{n\to\infty} \frac{n-r}{r+1} \cdot \frac{m/n}{1-m/n}$$

$$= \lim_{n\to\infty} \frac{1-r/n}{1-m/n} \cdot \frac{m}{r+1} = \frac{m}{r+1} .$$

$$\therefore \text{ Lim } P(r) = \text{Lim } \frac{P(r)}{P(r-1)} \cdot \text{Lim } \frac{P(r-1)}{P(r-2)} \ldots\ldots \text{Lim } \frac{P(1)}{P(0)} \text{Lim } P(0)$$

$$= \frac{m}{r} \cdot \frac{m}{r-1} \ldots\ldots\ldots \frac{m}{1} \cdot e^{-m} = \frac{e^{-m}m^r}{r!} .$$

In this derivation, we have not employed Stirling's formula.

Another *Derivation of Poisson* Distribution

The $p\ g.\ f.$ of the Binomial distribution is $P_B(t) = (q+pt)^n$

$$= [1+p(t-1)]^n$$

Putting $np=m$, we have. $P_B(t) = \left[1 + \frac{m}{n}(t-1)\right]^n$

$$\therefore \quad P(t) = \lim_{n\to\infty} P_B(t) = e^{m(t-1)} = e^{-m} \sum_{i=0}^{\infty} \left(\frac{mt}{r}\right)^r$$

$$\therefore \quad P(x=r) = \frac{e^{-m}m^r}{r!} .$$

6·3·1. Another Approach to Poisson Distribution. Let the random variate X be the number of calls during time t at a telephone switch board. We assume (a) that the probability of a call in time dt is $\lambda\, dt$, (b) that the calls are independent.

Let $P_x(t)$ denote the probability of x calls in time t so that $P_x(t+dt)$ is the chance of x calls in time $t+dt$. There are two mutually exclusive possibilities.

(*i*) x calls in time t and no call in time dt.

(*ii*) $x-1$ calls in time t and 1 call in time dt.

[The possibility of more than one call in time dt is neglected as it would be of order $(dt)^2$].

$$\therefore \qquad P_x(t+dt)=P_x(t)(1-\lambda\, dt)+P_{x-1}(t)\lambda\, dt$$

or $$\frac{P_x(t+dt)-P_x(t)}{dt}=\lambda\,[P_{x-1}(t)-P_x(t)]$$

Taking limit as $dt\to 0$, we have

$$\frac{dP_x(t)}{dt}=\lambda\,[P_{x-1}(t)-P_x(t)]$$

Putting $$P_x(t)=\frac{(\lambda t)^x}{x\,!}f(t) \qquad \ldots(6{\cdot}6')$$

we get $$\frac{(\lambda t)^x}{x\,!}f'(t)+f(t)\,\frac{\lambda^x}{x\,!}xt^{x-1}=\lambda\left[\frac{(\lambda t)^{x-1}}{(x-1)\,!}\,f(t)-\frac{(\lambda t)^x}{x\,!}\,f(t)\right]$$

or $$\frac{tf'(t)}{x}+f(t)=f(t)-\frac{\lambda t}{x}f(t)$$

or $$\frac{f'(t)}{f(t)}=-\lambda$$

Integrating $$f(t)=e^{-\lambda t}$$

since from (6·6′) $P_0(t)=f(t)$ so that $f(0)=P_0(0)=1$.

Hence $$P_x(t)=\frac{e^{-\lambda t}(\lambda t)^x}{x\,!},\ x=0,\ 1,\ 2,\ldots\ldots$$

Also $$\sum_{x=0}^{\infty}P_x(t)=e^{-\lambda t}\sum_{x=0}^{\infty}\frac{(\lambda t)^x}{x\,!}=e^{-\lambda t}\cdot e^{\lambda t}=1.$$

Hence the number of calls in a fixed time t is a Poisson variate with parameter λt.

The following are further examples of Poisson variates :

(1) The number of cars passing through a certain street in time t.

(2) The number of suicides or deaths by heart attack in time t.

(3) The number of radioactive atoms decaying in time t.

(4) The number of cosmic-ray particles tripping a Geiger-Muller counter in time interval t.

(5) The number of pieces of a certain merchandise sold by a store in time t.

(6) The number of fragments received by a surface area t from a fragmentation bomb.

Ex. 1. Explain how in each of the above examples, the condition assumed in the derivation of the Poisson distribution are satisfied.

Ex. 2. Give some examples of the occurrence of Poisson distribution in different fields. *(M.A., Delhi, 1960 ; M.A., Patna, 1956)*

Ex. 3. If X is a random Poisson variate such that $P(X=1)=P(X=2)$, find $P(X=4)$. $\left\{\text{Ans. } \frac{2e^{-2}}{3}\right\}$

Ex. 4. If X is a random Poisson variate, show that $P(X\geqslant n)-P(X\geqslant n+1)=P(X=n)$. *(B.A., Delhi, 1971)*

6·3·2. Mean and Variance of the Poisson Distribution.

$$\mu_1' = \sum_{r=0}^{\infty} \frac{r.m^r e^{-m}}{r!} = 0.e^{-m} + 1 \cdot \frac{me^{-m}}{1!} + \frac{2.m^2e^{-m}}{2!} + \ldots\ldots$$

$$= me^{-m}\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \ldots\ldots\right) = me^{-m}e^{m} = m.$$

$$\mu_2' = \sum_{r=0}^{\infty} \frac{r^2 m^r e^{-m}}{r!} = \sum_{0}^{\infty}\left[\{r+r(r-1)\}\frac{m^r e^{-m}}{r!}\right]$$

$$= \sum_{0}^{\infty} \frac{rm^r e^{-m}}{r!} + m^2 \sum_{0}^{\infty} \frac{m^{r-2}e^{-m}}{(r-2)!}$$

$$= m + m^2 \qquad \left(\because \sum_{r=2}^{\infty} \frac{m^{r-2}}{(r-2)!} = e^m\right)$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = m$$

Note. These results can be deduced from the mean and variance of the binomial distribution by putting $p=m/n$ and then letting n tends to infinity. It should be noted that mean and variance of Poisson probability distribution have the same value m. However if in a Poisson frequency distribution the variate is measured in certain units, the mean of the distribution will be m units and variance would be m square units.

Ex. 1. Prove that $\mu_2'=m[(m+1)^2+m]$, $\mu_4'=m[m^3+6m^2+7m+1]$. Deduce that $\mu_3=m$, $\mu_4=3m^2+m$.

Ex. 2. Deduce the first four moments about the mean of the Poisson distribution from those of the binomial distribution.

Ex. 3. Prove that, as m tends to infinity, both γ_1 and γ_2 tend to zero.

6·3·3. Moment Generating Function and Cumulative Function of the Poisson Distribution.

M.g.f. and cumulative function of the Poisson distribution with respect to origin are given by

$$M_0(t) = \sum_{x=0}^{\infty} e^{tx} \frac{m^x e^{-m}}{x!} = e^{-m} \sum_{x=0}^{\infty} \frac{(me^t)^x}{x!} = e^{-m}e^{me^t} = exp\{m(e^t-1)\}$$

$$K_0(t)=m(e^t-1)=m\left(t+\frac{t^2}{2!}+\frac{t^3}{3!}+\ldots\ldots\right)$$

$$k_r=\text{coeff of } \frac{r^r}{r!} \text{ in } K_0(t)=m, \text{ for all } r.$$

Thus all the cumulants of the Poisson distribution are equal to m.

Ex. 1. Find the *m.g.f.* about the mean for the Poisson variate and deduce the values of the first four moments about the mean.

Ex. 2. If x is a Poisson variate with mean m, show that $(x-m)/\sqrt{m}$ is a variate with zero mean and unity variance. Find the *m.g.f.* for this variate and show that it approaches $exp(-\frac{1}{2}t^2)$ as m tends to infinity.

6·3·4. Reproductive Property of the Poisson Distribution. Theorem. *If two independent variables x_1 and x_2 have Poisson distributions with means m_1 and m_2 respectively, their sum $y=x_1+x_2$ is a Poisson variable with mean m_1+m_2.*

Proof. Let $M_1(t)$, $M_2(t)$ be the moment generating functions of x_1 and x_2 and let $M(t)$ be the *m.g.f.* of their sum, then

$$M_1(t)=exp[m_1(e^t-1)],\ M_2(t)=exp[m_2(e^t-1)]$$

By § 5·6·2, the *m.g.f.* of the sum x_1+x_2 is given by

$$M(t)=M_1(t)\times M_2(t)=exp\ [m_1(e^t-1)]\ exp\ [m_2(e^t-1)]$$
$$=exp\ [(m_1+m_2)(e^t-1)]$$

but this is the *m.g.f.* of a Poisson distribution with mean m_1+m_2. Hence the theorem.

Ex. 1. Show that if x_i $(i=1, 2,\ldots\ldots, n)$ are independent variates having Poisson distribution with means m_i their sum $\sum_{i=1}^{n} a_i x_i$ is a Poisson variate with mean $\sum_{i=1}^{n} a_i m_i$, where a's are constants.

Ex. 2. Show that $f(r+1)=\frac{m}{r}\ f(r)$ and hence $f(1)$, $f(2)$,..... , can be obtained in terms of $f(0)$.

6·3·5. Mode of the Poisson Distribution. The mode is clearly that value of r for which $e^{-m}m^r/r!$ is greater than the term that precedes it and the term that follows it.

$$\therefore \qquad \frac{m^{r-1}\,e^{-m}}{(r-1)!} \leqslant \frac{m^r\,e^{-m}}{r!} \geqslant \frac{m^{r+1}\,e^{-m}}{(r+1)!}$$

These give $r \leqslant m$ and $r+1 \geqslant m$

i.e., $$m-1 \leqslant r \leqslant m.$$

Thus if m is not an integer, the mode is the integral value between $m-1$ and m. If m is an integer, then the probability of the

Poisson variate taking the value $m-1$ or m is the same and hence either can be taken as mode of the Poisson distribution with mean m. The distribution in this case can be regarded as a bimodal one.

Example 6·7. *Find the probability that at most 5 defective fuses will be found in a box of 200 fuses if experience shows that 2 per cent of such fuses are defective.*

Hence $m=np=200(\cdot 02)=4$.

Hence using (6·6), the required probability is given by

$$P(x\leqslant 5)=\sum_{x=0}^{5}\frac{e^{-4}\,4^x}{x\,!}=e^{-4}\left(1+4+\frac{4^2}{2}+\frac{4^3}{6}+\frac{4^4}{24}+\frac{4^5}{120}\right)$$

$$=\cdot 785 \qquad [e^{-4}=0\cdot 0183]$$

Example 6·8. *Fit a Poisson distribution to the following data which gives the number of yeast cells per square for 400 squares.*

No. of cells per square (x)	0	1	2	3	4	5	6	7	8	9	10	Total
No. of squares (f)	103	143	98	42	8	4	2	0	0	0	0	400

The expected or theoretical frequency for r successes is $Ne^{-m}m^r/r\,!$, but m is not given in this example. The mean of the Poisson distribution is m $\therefore$ mean m of the distribution of x is estimated by

$$m=\frac{\Sigma fx}{\Sigma f}=\frac{529}{400}=1\cdot 32$$

Assuming that the distribution of x is Poisson in the expected frequencies of 0, 1, 2,....., 10 are the successive values of

$$400\frac{e^{-1\cdot 32}\,(1\cdot 32)^x}{x\,!}$$

Since $e^{-1\cdot 32}=0\cdot 2674$, the expected frequency of 0, 1, ..., 10 cells are 107, 141, 93, 41, 14, 4, 1, 0, 0, 0, 0.

6·4. Other Discrete Distributions.

6·4·1. Causal Distribution. Here the variate takes with certainty one value only, say $x=u$, so that

$$P(x=u)=1$$

$$P(x=t)=0 \qquad t\neq u.$$

Also $$P(x\leqslant t)=0, \qquad t<u.$$

$$=1. \qquad t\geqslant u.$$

The distribution has one parameter u. The distribution obviously corresponds to a causal description met in the exact sciences and this shows that the statistical description of nature includes the causal distribution as a particular case.

6·4·2. Uniform or Rectangular Distribution. Here the variate takes the values $x_1, x_2, \ldots\ldots, x_n$ (and in particular the values $1, 2, \ldots\ldots, n$) each with the same probability $1/n$ so that

$$P(x=x_i)=\frac{1}{n},\ i=1, 2, \ldots\ldots, n$$

Clearly $$\sum_{i=1}^{n} P(x=x_i)=1$$

Such a distribution arises in the tossing of a coin or casting of a dice.

6·4·3. Geometric Distribution. In a Bernoullian sequence of trials, probability of success being p, let x be the number of failures preceding the first success then x is a random variate with probability distribution given by

$$P(x)=q^x p, \qquad x=0, 1, 2, \ldots\ldots, q=1-p.$$

since the first x trials are to be failures and the last trial is to be a success.

Clearly $$\sum_{x=0}^{\infty} P(x)=\sum_{x=0}^{\infty} q^x p=\frac{p}{1-q}=1$$

$$M.G.F.=\sum_{x=0}^{\infty} e^{tx}q^x p=p\sum_{0}^{\infty}(qe^t)^x=\frac{p}{1-qe^t}=p(1-qe^t)^{-1}$$

Ex. 1. Show that the mean and variance of the geometric distribution $P(x)=q^x p$, $x=0, 1, 2, \ldots$, are respectively qp^{-1}, qp^{-2}.

(B.A. Hons. Delhi, 1968, 1971)

Ex. 2. Show that the mode of the distribution $P(x)=(\frac{1}{2})^x$, $x=1, 2, 3, \ldots$, is 1.

6·4·4. Negative Binomial Distribution. Let $P(x)$ be the probability that exactly $x+k$ trials will be required to produce k successes in an indefinite series of Bernoullian trials. Clearly the last trial must be a success and its probability is p. Among the other $(x+k-1)$ trials, there must be $k-1$ successes and the probability of this is ${}^{x+k-1}c_{k-1}p^{k-1}$, q^x, $q=1-p$.

Hence multiplying the two probabilities, we get

$$P(x=k)={}^{x+k-1}c_{k-1}p^k q^x,\ x=0, 1, 2, \ldots, \text{and } k>0 \qquad \ldots(6\cdot7)$$

$$\therefore \quad \sum_{0}^{\infty} P(x=k) = p^k \sum_{x=0}^{\infty} {}^{x+k-1}c_{k-1}\, q^x = p^k(1-q)^{-k} = 1,$$

since $${}^{x+k-1}c_{k-1} = \frac{(k+x-1)(k+x-2)\ldots k}{x\,!} = (-1)^x\, {}^{-k}C_x$$

where $${}^{-k}c_x = \frac{(-k)(-k-1)\ldots(-k-x+1)}{x\,!} = (-1)^x\, {}^{x+k-1}C_{k-1}$$

Thus $P(x)$ can also be written as

$$P(x) = {}^{-k}c_x p^k (-q)^x, \; x=0, 1, 2, \ldots \qquad \ldots(6{\cdot}8)$$

This distribution is called *Pascal's distribution*. It contains two parameters p and k. If $k=1$ it reduces to the geometric distribution and in this sense it is a generalisation of this distribution.

In the above distribution, k is an integer but the distribution remains meaningful even if k is not an integer but $k \geqslant 0$. In this case the distribution is called the *Negative Binomial Distribution.*

Alternative forms of this distribution are obtained by putting

(*i*) $k = \frac{1}{\beta}$, $p = \frac{1}{1+\beta\mu}$, $q = \frac{\beta\mu}{1+\beta\mu}$, so that

$$P(x) = \frac{(1+\beta)(1+2\beta)\ldots[1+(x-1)\beta]}{x\,!}\,(1+\beta\mu)^{-\frac{1}{\beta}}\left(\frac{\mu}{1+\beta\mu}\right)^x \qquad \ldots(6{\cdot}8')$$

In this form the distribution with two parameters β and μ is known as *Polya's distribution.*

If $\beta = 1$, then

$$P(x) = \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^x \qquad \ldots(6{\cdot}8'')$$

which is another form of geometric distribution.

(*ii*) $p = \frac{1}{Q}$, $q = \frac{P}{Q}$, so that $Q-P=1$, since $p+q=1$. Thus

$$P(x)Q^{-k} \cdot {}^{x+k-1}C_{k-1}\left(\frac{P}{Q}\right)^x$$

$$= Q^{-k} \cdot {}^{-k}C_x \left(\frac{-P}{Q}\right)^x, \; x=0, 1, 2, \ldots \qquad \ldots(6{\cdot}9)$$

which is the general term of the binomial expansion of $(Q-P)^{-k}$.

The m.g.f. of the form (6·8) of the negative binomial distribution given by

$$M(t)=\sum_{x=0}^{\infty} Q^{-k}e^{tx}\,{}^{x+k-1}C_{k-1}\left(\frac{P}{Q}\right)^x=\sum_{0}^{\infty} Q^{-k}\,{}^{x+k-1}C_{k-1}\left(\frac{Pe^t}{Q}\right)^x$$

$$=(Q-Pe^t)^{-k} \qquad \ldots(6{\cdot}10)$$

$$E(x)=\left[\frac{dM}{dt}\right]_{t=0}=[kPe^t(Q-Pe^t)^{-(k+1)}]_{t=0}=kP$$

$$E(x)=\left[\frac{d^2M}{dt^2}\right]_{t=0}=[kPe^t(Q-Pe^t)^{-(k+1)}]$$

$$+k(k+1)P^2e^{2t}(Q-Pe^t)^{-(k+2)}]_{t=0}$$

$$=kP+k(k+1)P^2$$

$$\text{Var}\,(x)=E(x^2)-[E(x)]^2=kP+kP^2=kPQ.$$

The similarity of *m.g.f.* to that of the binomial probability distribution should be noted. It is to be noted that for the negative binomial $\mu_2>\mu_1'$ whereas for the ordinary binomial $\mu_2<\mu_1'$.

The negative binomial distribution has been found to occur in many biological situations and can come about as a result of clustering (or contagian) among the successes of an otherwise binomial population *e.g.*, death of insects, number of insect bites per apple. The distribution also arises in inverse sampling from a binomial population and as a weighted average of Poisson distributions.

Poisson distribution as the limiting case of negative binomial distribution.

If we let $P\to 0$ and $k\to\infty$ in such a way that limit $kP=m$ (a finite constant), then (6·8) takes the form

$$P(x)=\lim_{\substack{k\to\infty\\P\to 0}}(1+P)^{-k}\,\frac{(k+x-1)\ldots k}{x\,!}\left(\frac{P}{1+P}\right)^x$$

$$=\lim_{\substack{k\to\infty\\P\to 0}}\left(1+\frac{m}{n}\right)^{-k}\frac{1}{x\,!}\left(1+\frac{x-1}{k}\right)\ldots 1.m^x\left(1+\frac{m}{k}\right)^{-x}$$

$$=\frac{e^{-m}m^x}{x\,!}.$$

Ex. 1. How is the Poisson distribution related to the negative binomial distribution ? *(M.A., Punjab, 1958)*

Ex. 2. Show that for the negative binomial $(Q-P)^{-k}$, where $Q-P=1$, cumulative function $K(t)=-k\log[(1-P(e^t-1)]$. Hence deduce that $k_1=kP$, $k_2=kPQ$. Find k_3 and k_4 also.

$$\left[\text{Ans.}\quad \begin{matrix} k_3=kq\,(1+q)/p^3 \\ k_4=kq\,(1+4q+q^2)/p^4 \end{matrix}\right]$$

Ex. 3. Show that the probability generating function of (6–7) is $[p/(1-qs)]^r$.

Ex. 4. Prove the following recurrence formula for negative binomial distribution $\binom{x+k-1}{k-1}p^k q^k$.

$$f(x+1, k, p)=\frac{x+k}{x+1}\, qf(x, k, p).$$

6·4·5. *Hypergeometric Distribution. Suppose r balls are drawn one at a time, *without replacement*, from a bag containing m white and n black balls. The probability that x will be white out of the r balls so drawn is

$$P(x)=\frac{{}^mC_x\ {}^nC_{r-x}}{{}^{m+n}C_r},\quad x=0,\ 1,\ 2,\ \ldots\ldots,\ r,\ r\leqslant m,\ r\leqslant n \qquad \ldots(6\cdot 11)$$

where $\sum_{x=0}^{r} P(x)=1$ since $\sum_{x=0}^{r} {}^mC_x\ {}^nC_{r-x}={}^{m+n}C_r$

which can be proved by equating the coefficient of y^r in

$$(1+y)^m(y+1)^n=(1+y)^{m+n}.$$

The distribution with (6·11) as the probability law is known as Hypergeometric distribution.

$$\text{Mean}=E(x)=\sum_{x=0}^{r} xP(x)=\sum_{x=0}^{r}\frac{x.{}^mC_x\ {}^nC_{r-x}}{{}^{m+n}C_r}$$

$$=\frac{m\sum_{x=1}^{r}{}^{m-1}C_{x-1}\ {}^nC_{r-x}}{{}^{m+n}C_r}=\frac{m}{{}^{m+n}C_r}\sum_{y=0}^{r-1}{}^{m-1}C_y\ {}^nC_{r-y-1},\ x-1=y$$

$$\mu_1'=\frac{m}{{}^{m+n}C_r}\times{}^{m-1+n}C_{r-1}=\frac{mr}{m+n}.$$

$$\mu_2'=E(x^2)=\sum_{x=0}^{r}x^2P(x)=\sum_{x=0}^{r}x(x-1)\ P(x)+\sum_{x=0}^{r}xP(x)$$

Now $\sum_{x=0}^{r}x(x-1)P(x)$

$$=\frac{\Sigma x(x-1)\ {}^mC_x\ {}^nC_{r-x}}{{}^{m+n}C_r}=\frac{m(m-1)}{{}^{m+n}C_r}\sum_{x=2}^{r}{}^{m-2}C_{x-2}\ {}^nC_{r-x}$$

or $$\mu_2'=\frac{m(m-1)}{{}^{m+n}C_r}\sum_{y=0}^{r}{}^{m-2}C_y\ {}^nC_{r-2-y}=\frac{m(m-1)}{{}^{m+n}C_r}\ {}^{m-2+n}C_{r-2}$$

---

*The distribution is so termed since the *m.g.f.* can be expressed in terms of Hypergeometric Function.

$$\mu_2' = \frac{m(m-1)\,.\,r(r-1)}{(m+n)(m+n-1)}\,.$$

$$\therefore \quad \mu_2 = \frac{m(m-1)\,.\,r(r-1)}{(m+n)(m+n-1)} + \frac{mr}{m+n} - \left(\frac{mr}{m+n}\right)^2$$

$$= \frac{mnr(m+n-r)}{(m+n)^2(m+n-1)}\,.$$

Alternative Form. The hypergeometric distribution can also be obtained as follows :—

Suppose a bag contains Np white and Nq black balls $(p+q=1)$. The probability that a sample of n can be drawn containing r white and $(n-r)$ black balls is

$$^{Np}C_r \; ^{Nq}C_{n-r} / ^{N}C_n$$

or $$^{n}C_r (NP)^{(r)} (NQ)^{(n-r)} / N^{(n)}$$

where $$x^{(r)} = (x-1)(x-2)\ldots(x-r+1)$$

Consider the discontinuous distribution

$$f(x) = \frac{1}{N^{(n)}} \sum_{r=0}^{n} \{^{n}C_r (NP)^{(r)} \; (Nq)^{(n-r)} x^r\} \qquad \ldots(6{\cdot}12)$$

$$= \frac{(Nq)^{(n)}}{N^{(n)}} \sum_{r=0}^{n} \left(\frac{(Np)^{(r)} n^{(r)}}{(Nq-n+r)^{(r)}} \cdot \frac{x^r}{r} \right)$$

$$= \frac{(Nq)^{(n)}}{N^{(n)}} F(\alpha, \beta : \gamma, x)$$

where $\alpha = -n$, $\beta = -Np$, $\gamma = Nq - n + 1$ and the hypergeometric function.

$$F(\alpha, \beta ; \gamma, x) = 1 + \frac{\alpha\beta}{\gamma} \cdot \frac{x}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)} \frac{x^2}{2!} + \ldots$$

It can be proved that the function $F(\alpha, \beta ; \gamma, x)$ satisfies the following differential equation

$$x(1-x) \frac{d^2F}{dx^2} + \{\gamma - (\alpha+\beta+1)x\} \frac{dF}{dx} - \alpha\beta F = 0$$

Putting $x = e^t$, $\alpha = -n$, $\beta = -Np$, $\gamma = Nq - n + 1$, this reduces to

$$(1-e^t) \left[\frac{d^2\varphi}{dt^2} - (n+Np) \frac{d\varphi}{dt} - nNp\varphi \right] - Nn\, p\varphi + N \frac{d\varphi}{dt} = 0$$

But $\varphi = \Sigma \mu_r' \, t^r / r !$ Therefore, equating the constant term,

$$\mu_1' = np$$

The distribution when transferred about its mean np gives ϕe^{npt} for φ and then we have

$$(1+e^{\theta})\left[\frac{d^2\varphi}{dt^2}+\frac{d\varphi}{dt}\{n(p-q)-Np\}+(N-n)pq^n\varphi\right]+\frac{Nd\varphi}{dt}=0$$

Equating coefficient of t, t^2, ..., we get

$$\mu_2=npq(N-n)/(N-1)$$

$$\mu_2=npq(q-p)(N-n)(N-2n)/[(N-1)(N-2)].$$

Ex. 1. Deduce the moments of the binomial distribution.

Ex. 2. Show that at $N\to\infty$, (6·12) approaches the binomial distribution

6·4·6. Multinomial Distribution. The Binomial distribution can easily be generalized to n repeated independent trials when each trial may result in one of the several outcomes say E_1, E_2, ..., E_k with respective probabilities p_1, p_2, ..., p_k in each trial where $p_1+p_2+...+p_k=1$.

The probability that n trials will result in E_1 occuring n_1 times, E_2 occurring n_2 times, ..., E_k occurring n_k times in a *fixed definite order* is

$$p_1^{n_1}\ p_2^{n_2}\ \ldots\ldots\ p_k^{n_k}\ ,\ \sum_{i=1}^{k} n_i=n.$$

But we are interested in events occurring in any order. The number of mutually exclusive ways in which this can happen is

$$\frac{n\ !}{n_1\ !\ n_2\ !\ \ldots\ldots\ n_k\ !}$$

Hence the required probability is

$$P(n_1, n_2, \ldots, n_k)=\frac{n\ !}{n_1\ !\ n_2\ !\ldots n_k}\ p_1^{n_1}\ p_2^{n_2}\ \ldots p_k^{n_k},\ \sum_{i=1}^{k} n_i=n \quad \ldots(6\cdot13)$$

This distribution is called multinomial probability distribution as the expression is the general term of the multinomial expansion of

$$(p_1+p_2+...+p_k)^n$$

We notice that this is not a univariate distribution but is a multivariate distribution involving k variates n_1, n_2, ..., n_{k-1}, n_k but only $(k-1)$ variates are independent since $\sum_{1}^{k} n_i=n$. It has application in sampling *with replacement* of which individuals are classified into more than two categories.

The means, variances, covariances can be obtained as follows :

Since $\Sigma \frac{n!}{n_1!\, n_2!\ldots\ldots n_k!} p_1^{n_1} p_2^{n_2}\cdots p_k^{n_k} = (p_1+p_2+\ldots+p_k)^n,$

$$p_i \frac{\partial}{\partial p_i} \Sigma \frac{n!}{n_1!\, n_2!\ldots n_k!} p_1^{n_1} p_2^{n_2}\cdots p_k^{n_k} = \Sigma n_i P(n_1, n_2, \ldots, n_k)$$

$$= p_i \frac{\partial}{\partial p_i} (p_1+p_2+\ldots+p_k)^n,$$

$$= np_i\,(p_1+p_2+\ldots+p_k)^{n-1}.$$

$$p_i \frac{\partial}{\partial p_j} \Sigma n_i P(n_1, n_2, \ldots, n_k) = \Sigma n_i n_j P(n_1, n_2, \ldots, n_k)$$

$$= p_j \frac{\partial}{\partial p_j} np_i(p_1+\ldots+p_k)^{n-1}$$

$$= n(n-1)p_i p_j\,(p_1+\ldots+p_k)^{n-2}$$

Similarly $p_i \frac{\partial}{\partial p_i} \Sigma n_i P(n_1, \ldots, n_k) = \Sigma n_i^2 P(n_1, n_2, \ldots, n_k)$

$$= np_i(p_1+\ldots+p_k)^{n-1} + n(n-1)p_i^2(p_1+\ldots+p_k)^{n-2}$$

These results give

$$E(n_i) = np_i,$$

$$E(n_i n_j) = n(n-1)p_i p_j,\ i \neq j$$

$$(En_i^2) = np_i + n(n-1)p_i^2$$

$$\mathrm{Var}(n_i) = E(n_i^2) - [E(n_i)]^2 = np_i(1-p_i)$$

$$\mathrm{Cov}\,(n_i, n_j) = E(n_i n_j) - E(n_i)E(n_j) = -np_i p_j.$$

Ex. 1. Show that

$M_0(t) = e^u$. for Causal distribution.

$M_0(t) = \frac{1}{n} e^t \frac{(e^{nt}-1)}{e^t-1}$ for Uniform distribution.

$M_0(t) = \left(\frac{p}{1-qe^t}\right)^k$ for Pascal distribution.

$M_0(t) = [1+\beta\mu\,(1-e^t)]^{-1/\beta}$ for Polya's distribution.

Deduce *m.g.f.* about the mean in each case. Hence or otherwise obtain the first four moments about the mean for each distribution.

Ex. 2. Show that the sum of k independent geometric variates with the same parameter a is a Pascal variate. Hence or otherwise prove that the mean and variance of the Pascal distribution are kq/p and kq/p^2 respectively. Show also that the sum of a number of independent Pascal variates (with same p) is a Pascal variate.

Ex. 3. The *cumulative distribution function* (*c.d.f.*) $F(x)$ of a random *variable* X is defined by

$$F(x) = P[X \leqslant x]$$

Prove that $F(x+0)-F(x-0)=P[X=x]$

For the uniform distribution prove that

$$F(x)=\frac{[x]}{n}.$$

and draw its graph. [Here $[x]$ denotes the integral part of x].

Ex. 4. In question 27 of Exercise 5, show that the size of the sample S_r can be written as $1+X_1+X_2+\ldots\ldots+X_{r-1}$, where X_1 is the number of drawings following the selection of the i-th element and up to and including the drawing of the next new element. Prove that X_i is a Pascal variate with $p_i=(N-i)/N$. Hence find the mean and variance of S_r.

Ex. 5. In examples 5·7, 5·8 and questions 26, 27 and 28 of Ex. 5 we have found the means and variances of some distribution *without first finding the distributions.* What are these distributions ? Show also that the *m.g.f.* for the distribution of the examples is

$$\prod_{i=1}^{n}(q_i+p_ie^t)$$

Hence find the mean, variance and μ'_3 for this distribution.

Ex. 6. Show that Poisson's distribution is the limiting form of Polya's distribution as $\beta\to 0$.

Ex. 7. Show that the *m.g.f.* of the multinomial distribution is

$$[p_1e^{t_1}+p_2e^{t_2}+p_3e^{t_3}+\ldots+p_ke^{t_k}]^n,\ \sum_1^k pi=1$$

and from this deduce that the means, variances and covariance are given be

$$E(n_i)=np_i,$$

$$Var\ (n_i)=np_i(1-p_i),$$

$$Cov\ (n_i, n_j)=-np_ip_j\ i\neq j$$

(M.A., Bombay, 1955 ; M.A. Poona, 1959)

EXERCISE 6

1. (*a*) Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or six ? *(B.A., Hons., Delhi, 1949, 56)*

(*b*) Take 100 sets of 10 tosses of an unbiased coin. In how many cases do you expect to get 7 heads at least ? *(B.A., Madras, 1947)*

[**Ans.** (*a*) 233 (*b*) 17 times]

2. A perfect cubic die is thrown a larger number of times in sets of 8. The occurrence of a 5 or 6 is called a success. In what proportion of the sets would you expect to get 3 successes ? [**Ans.** 27·31 per cent]

3. (*a*) Assuming that half the proportion are consumers of rice, so that the chance of an individual being a consumer is $\frac{1}{2}$, and assuming that 100 investigators each took ten individuals to see whether they are consumers, how many investigators would you expect to report that three people or less were consumers ? *[M.A. Eco., Statistics, Delhi, 1958 ; M.A., Punjab, 1952)*

(*b*) In the above example, if 1,000 investigators each choose 100 individuals, how many would you expect to report that more than 60 persons were consumers ? (*B.A. Luck., 1950*)

[Ans. (*a*) 17 ; (*b*) 18]

4. An irregular six-faced die is thrown, and the expectation that in 10 throws it will give five even numbers is twice the expectation that it will give four even numbers. How many times in 10,000 sets of 10 throws would you expect it to give no even number ? (*B.A. Hons., Delhi, 1954*)

5. Eight mice are selected at random and they are divided into two groups of 4 each. Each mouse in group *A* is given a dose of certain poison *a* which is expected to kill one in four ; each mouse in group *B* is given a dose of certain poison *b* which is expected to kill one or two. Show that nevertheless, there may be fewer deaths in group *B* than in group *A* and find the probability of this happening. (*I.A.S. 1962*)

[Ans. $\frac{525}{4096}$]

6. The following data due to Weldon shows the results of throwing 12 dice 4096 times ; a throw of 4, 5, or 6 being called success.

Success	*Frequency*	*Success*	*Frequency*
0	—	7	847
1	7	8	536
2	60	9	257
3	198	10	71
4	430	11	11
5	731	12	
6	948		
		Total	4096

Find the expected frequencies and compare the actual mean with those of the expected distribution. Calculate the standard deviation.

(*B.Sc., Agra 1966* ; *B.A., Hons., Delhi, 1953*)

[Ans. Expected frequencies are :
1, 12, 66, 220, 495, 792, 924, 792, 495, 220, 66, 12, 1.
Expected mean=6, Actual mean=6·139]

7. Show that if two symmetrical binomial distributions ($p=q=\frac{1}{2}$) of degree n (and of the same number of observations) are so superposed that the rth term of the one coincides with the $(r+1)$th term of the other, the distribution formed by adding superposed terms is a symmetrical binomial of degree $(n+1)$. (*M.Sc., Agra, 1949, 1953*)

8. The incidence of occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the probability that out of six workers 4 or more will catch the disease ? [Ans. $\frac{53}{3125}$]

9. In a precision bombing attack there is a 50% chance that any one bomb will strike the target. Two direct hits are required to destroy the target completely. How many bombs must be dropped to give a 99% chance or better of completely destroying the target ? [Ans. 11]

10. (*a*) Five dice are tossed 96 times. Find the expected frequencies when throwing of a 4, 5 or 6 is regarded as a success.

(*b*) Two dice are thrown until a seven is obtained. Find the most probable number of throws and also the expected number of throws.

(*B.A. & B.Sc., Lucknow ; M.A., Punjab*)

[**Ans.** (*a*) 3, 15, 30, 30, 15, 3 (*b*) 1, 6]

11. (*a*) If probability of success is ·01, how many trials are necessary in order that probability of at least one success is $>\frac{1}{2}$? [**Ans.** 69]

(*b*) Bring out the fallacy, if any, in the following statement :

The mean of a binomial distribution is 5 and its s.d. is 3.

(*B.A. Hons., Delhi, 1952*)

[**Ans.** It makes $q=1{\cdot}8$ which is wrong].

(*c*) If X and Y are independent binomial variates with parameters p_1 and p_2 and indices n_1 and n_2 respectively, obtain the probability that $X+Y=r$.

(*d*) If X and Y are independent binomial variates $b(5, \frac{1}{2})$ and $b(7, \frac{1}{2})$, find $P(X+Y=3)$.

12. Show that, for the binomial distribution $(q+p)^n$, the factorial moments about the mean are

$$\mu_{(2)}=npq, \quad \mu_{(3)}=-2npq(p+1). \qquad (I.C.A.R., Delhi, 1950)$$

13. Obtain the characteristic function (C.F.) of the standard binomial variate $\frac{r-np}{\sqrt{npq}}$ where r is the number of successes obtained in n independent trials, each with constant probability p of success $q=1-p$. Obtain the limit of this function as n tends to infinity. (*B.A., Hons., Delhi, 1953*)

[The C.F. about $x=0$ is given by $\varphi_x(t)=E(e^{it\overline{x}})=(q+pe^{it})^n$] [**Ans.** $\varphi(t)=e^{-\frac{1}{2}t^2}$]

14. Show that for the binomial distribution $(q+p)^n$, $q=1-p$

$$k_{r+1}=pq\frac{dk_r}{dP}, \quad r>1.$$

Hence deduce that $k_4=npq(1-6pq)$.

(*M.A., Delhi, 1959, 1966 ; M.Sc. Calcutta, 1943*)

[**Hint.** $k_r=\left[\frac{d^r}{dt^r}\{n\log(q+pe^t)\}\right]_{t=0}$ $\therefore \frac{dk_r}{dp}=n\left[\frac{d^r}{dt^r}\left(\frac{-1+e^t}{q+pe^t}\right)\right]_{t=0}$

Also $k_{r+1}=\left[\frac{d^{r+1}}{dt^{r+1}}n\log(q+pe^t)\right]_{t=0}=np\left[\frac{d^r}{dt^r}\left(\frac{e^t}{q+pe^t}\right)\right]_{t=0}$

Now show that $k_{r+1}-pq\frac{dk_r}{dp}$ simplifies to zero.]

15 (*a*) If $K(t)$ be the cumulative function about the origin of the binomial distribution of size n, show that

$$\frac{d}{dt}K(t)=n\left[1+e^{-(z+t)}\right]^{-1} \quad \text{where } z=\log_e\left(\frac{p}{q}\right).$$

(*b*) By expanding the right hand side in powers of t by Taylor's theorem, show that $k_r=n\frac{d^{r-1}p}{dz^{r-1}}$ where k_r is the rth cumulant.

(*c*) Hence or otherwise, obtain the recurrence relation :

$$k_{r+1}=pq\frac{dkr}{dp}, \quad r>1. \qquad (B.A. Hons., Calcutta, 1949)$$

[**Hint.** (*a*) $M_0(t)=(q+pe^t)^n$ $\therefore$ $K(t)=n\log(q+pe^t)$

$$\therefore \quad \frac{d}{dt}K(t)=\frac{npe^t}{q+pe^t}=np(qe^{-t}+p)^{-1}$$

$$=n\left(1+\frac{q}{p}e^{-t}\right)^{-1}=\text{etc.}\ ]$$

16. For Bernoulli's distribution $(q+p)^n$, prove the following results:

(*a*) (*i*) $\sum_{r=0}^{n} {}^nC_r p^r q^{n-r}(r-np)=0$, (*ii*) $\sum_{r=0}^{n} {}^nC_r p^r q^{n-r}(r-np)^2=npq$.

(*b*) $\mu_r=\left(p\frac{\partial}{\partial p}\right)^r (q+p)^n$ when μ_r is the rth moment about the origin.

Hence obtain the coefficient β_1 of the Binomial distribution.

(B.Sc., Kerala, 1965)

[**Hint.** (*a*) Differentiate the following identity twice with respect to u and set $u=0$

$$e^{-npu}(pe^u+q)^n=\sum_{r=0}^{n} {}^nC_r p^r q^{n-r} e^{(r-np)u}.]$$

17. (*a*) Show that in the Binomial distribution $(p+q)^n$, the ratio of the mean deviation about the mean to standard deviation is

$$\frac{2}{\sqrt{npq}}\,{}^{n-1}C_{\mu-1}\,p^{\mu-1}\,q^{n-\mu}$$

where μ is the greatest integer contained in $np+1$. Find the limit of the above ratio as n tends to infinity.

(*b*) Find the mean deviation about the mean of Poisson distribution with parameter m. *(B.A. Hons., Delhi, 1971)*

[**Sol.** (*a*) By definition, mean deviation, M_d, about the mean is given by

$$M_d=\sum_{r=0}^{n} {}^nC_r p^r q^{n-r}\,|\,r-np\,|\,.$$

But $$0=\sum {}^nC_r p^r q^{n-r}(r-np).$$

$\therefore$ $$M=2\sum_{r>np} {}^nC_r p^r q^{n-r}(r-np),$$

the summation being for all integers $r>np$.

Let $$F(x, y)=\sum_{r>np} {}^nC_r x^r y^{n-r}$$

$$\therefore \quad \sum_{r>np} {}^nC_r p^r q^{n-r}(r-np)=p\frac{\partial F(p, q)}{\partial p}-npF(p, q)$$

But, by Euler's theorem, $nF(p, q)=P\frac{\partial F(p, q)}{\partial p}+q\frac{\partial F(p, q)}{\partial q}$

$$\therefore \quad M_d=2\sum_{r>np} {}^nC_r p^r q^{n-r}(r-np)=2pq\left(\frac{\partial F}{\partial p}-\frac{\partial F}{\partial q}\right)$$

$$=2npq\,{}^{n-1}C_{\mu-1}p^{\mu-1}\,q^{n-\mu},\quad \mu\leqslant np+1<\mu+1.$$

Alitei. (*a*) and (*b*) For a neater proof see problem 161, page 114, second edition *Advanced Problems in Statistics* by H.C. Saxena.

18. Let λ and μ_r denote the mean and central rth moment of a Poisson distribution respectively. Obtain the following recurrence formula :

$$\mu_{r+1}=r\lambda\mu_{r-1}+\lambda\frac{d\mu_r}{d\lambda}, \text{ where } \mu_r=\sum_{j=0}^{\infty}\frac{e^{-\lambda}\lambda^j}{j!}(j-\lambda)^r$$

and hence obtain β_1 and β_2 of the Poisson distribution.

(M.A. Patna, 1956 ; M.A. Delhi, 1956, 1960 ; M.A. Bombay, 1953 ; M.A. Punjab, 1956 ; B A. Hons., Delhi, 1961 ; M.A. Eco. Statistics, Delhi, 1958)

[**Ans.** $\beta_1=1/\lambda$, $\beta_2=3+1/\lambda$]

19. From records of 10 Russian army corps kept over 20 years, the following data was obtained showing the number of deaths caused by the kicks of a horse. Determine, m, the average number of deaths per army corps per annum, and calculate the theoretical Poisson frequencies (Data of L. von Bortkiewicz) and compare them with the observed frequencies.

Number of deaths per army corps per annum	0	1	2	3	4	Total
Frequency of occurrence	109	65	22	3	1	200

$[e^{-0\cdot61}=0\cdot5436]$ *(M.Sc. Agra)*

[**Ans.** $m=0\cdot61$; 109, 66, 20, 4, 1]

20. In certain factory turning out razor blades, there is a small chance $\frac{1}{500}$ for any blade to be defective. The blades are supplied in packets of 10. Use the Poisson distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades respectively in a consignment of 10,000 packets, given that $e^{-\cdot02}=0\cdot9802$.

(B.Sc., Agra, 1956 ; B. Sc., Lucknow, 1947, 49, ; M.Sc., Agra, 1955, 50)

[**Ans.** $m=10\times\frac{1}{500}=\cdot02$; 9802, 196 and 2 packets respectively]

21. *(a)* If x is a Poissionian variate with mean m, what would be the expectation of e^{-kx}. kx where k is a constant ? *(I.A.S. 1948)*

[**Ans.** $m.k.\exp\{m(e^{-k}-1)-k\}$]

(b) Show that expectation of $(e^{-kx})=e^{-m(1-e^{-k})}$.

(M.A., Delhi, 1957 ; B.A. Hons., Delhi, 1960)

Hence show that, if $\bar{x}$ be the arithmetic mean of n independent random variables $x_1, x_2, \ldots\ldots, x_n$, each having Poisson distribution with mean m, then $e^{-\bar{x}}$ as an estimate of e^{-m} is biased although $\bar{x}$ is an unbiased estimate of m.

(B.A., Hons., Delhi, 1960)

22. Six coins are tossed 6,400 times. Using the Poisson distribution, what is the approximate probability of getting six heads x times ?

(M.A., Bombay, 1956)

$\left[\text{Ans. } \frac{e^{-100}(100)^x}{x!}\right]$

23. For Poisson distribution with parameter m, prove that

(i) $\beta_1=\frac{1}{m}$, $\beta_2=3+\frac{1}{m}$ *(B.Sc., Agra, 1951)*

(ii) $m\sigma\gamma_1\gamma_2=1$. *(B.Sc., Agra, 1956)*

(iii) $\mu'(r)=m^r$, $\mu(2)=m$, $\mu(3)=-2m$, $\mu(4)=3m(m+2)$,

(M.Sc., Agra, 1957 ; B.Sc. Bombay, 1968)

24. A distributor of bean seeds determines from extensive tests that 5 per cent of a large batch of seeds will not germinate. He sells the seeds in packages of 100 and guarantees 20 per cent germination. What is the probability that a given package will violate the guarantee ?

$$\left[\text{Ans. Required probability} = \sum_{r=21}^{100} \left(\frac{e^{-10}10^r}{r!}\right) = 1 - \sum_{r=0}^{20} \left(\frac{e^{-10}10^r}{r!}\right)\right]$$

25. A manufacturer of cotter pins knows that 5 per cent of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality ? $[e^{-5} = \cdot 006738]$

$$\left[\text{Ans. } P(x>10) = \sum_{x=11}^{100} \frac{e^{-5}.5^x}{x!} = 1 - \sum_{x=0}^{10} \left(\frac{e^{-5}.5^x}{x!}\right)\right]$$

26. A source of liquid is known to contain bacteria with the mean number of bacteria per cubic centimetre equal to 3. Ten 1-cubic centimetre test tubes are filled with the liquid. Assuming the Poisson distribution is applicable calculate the probability (*a*) that all 10 test tubes will show growth, that is, contain at least 1 bacterium each and (*b*) that exactly 7 test tubes will show growth. $[e^{-3} = \cdot 04978]$

[Ans. (*a*) $(1-e^{-3})^{10}$ (*b*) ${}^{10}C_7(1-e^{-3})^7(e^{-3})^3$]

27. Suppose the number of telephone calls on an operator received from 9 00 to 9·05 a m. follow a Poisson distribution with mean 3. Find the probability that (*a*) the operator will receive no calls in that time interval tomorrow ; (*b*) in the next 3 days the operator will receive a total of 1 call in that time interval.

[Ans. (*a*) e^{-3} (*b*) $3 \times (e^{-3})^2(1-e^{-3})$]

28. A car hire firm has two cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1·5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused. $[e^{-1\cdot 5} = 0\cdot 2231]$

(*B.A., Hons., Calcutta, 1955*)

[Ans. 0·2231, 0·1913]

29. A large number of observations on a given solution, which contained bacteria, were made taking samples of 1 c.c. each noting down the number of bacteria present in each sample. Assuming the Poisson distribution and given that 10 per cent samples contained no bacteria, find the average number of bacteria per c.c. (*M.A., Statistics, Delhi, 1959*)

[Ans. $\log_e 10$ or 2·3026]

30. Prove that the sum of two independent Poissonian variates with means m_1 and m_2 is also a Poissonian variate with mean m_1+m_2.

(*B.A., Hons., Delhi, 1951*)

31. (*a*) If x and y are independent Poisson variates with mean m_1 and m_2, prove that the probability that $(x-y)$ has the value r is the coefficient of t^r in $\exp\{m_1t+m_2t^{-1}-m_1-m_2\}$

(*B.A., Hons , Delhi, 1948, 55, 60*)

(*b*) Does the difference of two independent Poisson variates follow the Poisson distribution ? Give reasons.

(*M.A., Delhi, 1969 ; B A., Hons , Delhi, 1957*)

Hint. No, since the *m.g.f.* of the difference is $\exp[m_1(e^t-1)-m^2(e^{-t}-1)]$ which cannot be put in the form in which Poisson variate has the *m.g.f.* Further the difference cannot be a Poisson variate is evident from the fact that it may have negative or positive value.]

32. (*a*) If K_1 and K_2 are two independent Poisson variates with means 1 and 2 respectively, show that $P(x_1+x_2<4)=13e^{-3}$.

(*b*) Let x and y be the independent Poisson variates. The variance of $X+Y$ is 9 and $P(X=3 \mid X+Y=6)=\frac{5}{54}$. Find the mean of X.

(B.Sc. Kerala, 1970.)

33. Obtain the characteristic function for the standard Poisson variate $\frac{r-m}{\sqrt{m}}$ where r is the number of successes in n independent trials. Obtain the limit of this function as $m\to\infty$. [**Ans.** $e^{-\frac{1}{2}t^2}$]

34. If the probability of exactly r successes is

$$P_r=\frac{1}{r!}\sum_{k=0}^{n-r}\frac{(-1)^k}{k!};\ r=0,\ 1,\ldots\ldots,\ n.$$

Show that, as $n\to\infty$, this tends to a Poisson probability function.

35. A special dice with $(n+1)$ faces is marked in its faces the numbers

$$0,\ \frac{1}{n},\ \frac{2}{n},\ \ldots\ldots,\ \frac{n-1}{n}\text{ and }\frac{n}{n}.$$

The dice is unbiased. For distribution of the random variable corresponding to the number on the uppermost face, find (*a*) the expected value (*b*) the standard deviation (*c*) β_1 and (*d*) β_2. *(I.A.S. 1963)*

[**Ans.** Expected value$=\frac{1}{2}$, standard deviation$=\sqrt{\frac{(n+3)}{21n}}$, $\beta_1=0$]

36. If t is positive and real, show that function defined by

$$f(x)=e^{-t}(1-e^{-t})^{x-1},\ x=1,\ 2,\ 3,$$

can represent a probability density function of the variate x. Find the mean and variance of the distribution. *(M.A. Patna, 1967)*

[**Ans.** mean$=e^t$, $\mu_2=e^{2t}-(e^t)^2$]

37. Compound Poisson distribution. For given λ, X is a Poisson variate with mean λ while λ is itself a continuous random variate having the distribution given by

$$\frac{a^\nu}{\lfloor\nu}\ e^{-a\lambda}\ \lambda^{\nu-1}\ d\lambda,\ (\lambda>0,\ a>0,\ \nu>0).$$

Find the overall distribution of X when the random fluctuations in λ are allowed for. *(M.A. Bombay 1956 ; I.C.A.R. Delhi, 1949 ; M.A., 1958)*

Calculate the mean and variance of the resulting distribution and show that the variance is always greater than the mean. *(M.A., Delhi, 1954, 60)*

[**Ans.** Negative binomial distribution, $\mu_1'=\frac{\nu p}{q}$, variance$=\frac{\nu(p+q)q}{p^2}$]

38. Compound Binomial distribution. Suppose that the probability of an insect laying n eggs is $e^{-\lambda}\ \lambda^n/n!$ and that the probability of an egg deve-

loping is p. Assuming natural independence of the eggs, show that the probability of a total of k survivors in given by the Poisson distribution with parameter λp. Hence find the mean and variance of the distribution obtained.

(*B.Sc. Poona 1967, 71*)

Hint. $P(X=k)=\sum_{n=0}^{\infty}\frac{e^{-\lambda}\lambda^n}{n!}\,{}^nC_k p^k(1-p)^{n-k}=\frac{e^{-\lambda}p^k}{k!}\sum_{n=k}^{\infty}\frac{\lambda^n(1-p)^{n-k}}{(n-k)!}$

$$=\frac{e^{-\lambda}p^k\lambda^k}{k!}\sum_{n=k}^{\infty}\frac{[\lambda(1-p)]^{n-k}}{(n-k)!}=\frac{e^{-\lambda}p^k\lambda^k e^{\lambda(1-p)}}{k!}=\frac{e^{-\lambda p}(\lambda P)^k}{k!}\Big]$$

39. The number of female insects in a given region follows a Poisson distribution with mean m. The number of eggs laid by each insect follows a Poisson distribution with mean λ. Find the probability distribution of the number of eggs in the region and its mean and variance. (*B.A., Hons., Calcutta, 1955*)

[Ans. The probability distribution is $\frac{e^{-m}\lambda^x}{x!}\sum_{s=0}^{\infty}\frac{(me^{-\lambda})^s s^x}{s!}$

$m.g.f.=\exp[m\{\exp\{(\lambda e^{\theta}-1)\}-1)\}]$

From this we find mean$=m\lambda$, variance$=m\lambda(\lambda+1)$].

40. Prove the identity :

$$p^n+{}^nC_1p^{n-1}q+{}^nC_2p^{n-2}q^2+\ldots+{}^nC_kp^{n-k}q^k=\left(\sum_{x=0}^{k}{}^nC_x p^{n-x}q^x\right)$$

$$=\int_0^p x^{n-k-1}(1-x)^k dx\div\int_0^1 x^{n-k-1}(1-x)^k dx$$

(*M.A., Statistics, Delhi, 1959*)

[**Hint.** *First Method.* Take derivatives with respect to p.

Second Method. Integrate the numerator of right side by parts repeatedly with $(1-x)^k$ as the first function.

Third Method. Use Taylor's theorem in the form

$$f(a+h)=f(a)+hf'(a)+\ldots\ldots+\frac{h^{r-1}}{(r-1)!}f^{r-1}(a)+\int_0^1\frac{h^r(1-t)^{r-1}}{(r-1)!}f^r(a+th)dt$$

and put $a=q$, $h=p$, $f(a+h)=(q+p)^n$,

$$\therefore\ (q+p)^n=\sum_{j=0}^{r-1}{}^nC_j p^j q^{r-j}+R_r \text{ where } R_r=\frac{n!}{(n-r)!\,r!}\int_0^p x^{r-1}(1-x)^{n-r}dx$$

or $R_r=\frac{B_p(r,n-r+1)}{B(r,n-r+1)}$ and $t=1-\frac{x}{p}$.]

41. Show that the hypergeometric distribution (§ 6·4·5) tends to Binomil distribution if $m, n\to\infty$ and $\frac{m}{m+n}\to p$.

42. The multinomial variables $n_1, n_2, \ldots, n_k$ with $n_1+n_2+\ldots+n_k=n$ (constant) are distributed in the form

$$\frac{n!}{n_1!\,n_2!\ldots n_k!}p_1^{n_1}p_2^{n_2}\ldots\ldots p_k^{n_k}, \quad \sum_{i=1}^{k} p_i=1.$$

Show that the mean and variance of $\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}$ are given by

$$(k-1) \text{ and } 2(k-1)\left(1-\frac{1}{n}\right)+\frac{1}{n}\left(\sum_{i=1}^{k}\frac{1}{p_i}-k^2\right)$$

(B.A., Hons., & M.A., Madras, 1957)

[**Hint.** Let $x_i=\frac{(n_i-np_i)}{\sqrt{np_i}}$ so that $\sum_{i=1}^{k} x_i\sqrt{p_i}=0$ and therefore $\chi^2=\sum_{1}^{k}x^2{}_i$

The joint m *g.f.* of the variables $n_1, n_2, \ldots, n_k$ is $(p_1e^{t}{}_1+p_2e^{t}{}_2+\ldots+p_ke^{t}{}_k)^n$

The joint *m.g.f.* of the variables $x_1, x_2, \ldots, x_k$ is

$$\phi(t_1, t_2, \ldots\ldots, t_k)=e^{-\sqrt{\Sigma p_i}\,t}\left(p_1e^{\frac{t_1}{\sqrt{np_1}}}+\ldots\ldots+\left(p_ke^{\frac{t_k}{\sqrt{np_k}}}\right)^n\right.$$

Now write down Maclaurin expansion of this function and obtain $E\chi^2$, etc.]

43. If $r_1, r_2, \ldots\ldots, r_k$ be k independent Poisson variates with parameters $\lambda_1, \lambda_2, \ldots\ldots, \lambda_k$, prove that (*i*) $r=\sum_{i=1}^{k} r_i$ is a Poisson variate, and (*ii*) the conditional distribution $P(r_1, r_2\ldots\ldots, r_k/r)$, when the sum $r=(r_1+r_2+\ldots\ldots+r_k)$ is fixed is multinomial.

[**Hint.** $P(r_1, \ldots\ldots, r_k)=e^{-(\lambda_1+\ldots+\lambda_k)}\frac{\lambda_1{}^{r_1}}{r_1!}\ldots\ldots\frac{\lambda_k{}^{r_k}}{r_k!}$

$$P(r=r_1+\ldots+r_k)=e^{-\lambda}\sum\frac{\lambda_1{}^{r_1}}{r_1!}\ldots\ldots\frac{\lambda_k{}^{r_k}}{r_k!}=\frac{e^{-\lambda}\lambda^r}{r!}, \quad \lambda=\lambda_1+\ldots+\lambda_k$$

$$\therefore \quad P(r_1, \ldots\ldots, r_k/k_1)=\frac{P(r_1\ldots\ldots r_k)}{P(r)}=\frac{r!}{r_1!\ldots r_k!}\left(\frac{\lambda_1}{\lambda}\right)^{r_1}\ldots\left(\frac{\lambda_k}{\lambda}\right)^{r_k}$$

which is multinomial with probabilities $\lambda_1/\lambda, \ldots\ldots, \lambda_k/\lambda$ in the classes.]

44. When n is large and $np_i=\lambda_i$ $(i=1, 2, \ldots\ldots k)$ is moderate, show that the multinomial distribution is approximated by

$$\exp\left[-\sum_{i=1}^{k}\lambda_i\right]\prod_{j=1}^{r}\lambda_j{}^{k_j}/\prod_{j=1}^{r}k_j$$

Prove also that the sum of all the terms is unity.

45. Derive the distribution of the random number N, where N is the smallest number of independent trials required to obtain a given number k of occurrences of an event of which the probability is p in each trial.

Show that as p becomes small the distribution of $2Np$ tends to that of a χ^2 distribution with $2k$ degrees of freedom.

[**Ans.** The distribution is ${}^{N-1}C_{k-1}\,p^k(1-p)^{N-k}$, $N=k, k+1, \ldots$]

$$mgf=\left(1+\frac{e^{-\theta}-1}{p}\right)^{-k} \qquad \therefore \quad \mu_1{}'=\frac{k}{p}, \quad \mu_2=\frac{k}{p^2}+\frac{k}{p}$$

In standard measure *m.g.f.* tends to $(1-\theta)^{-k}$ and hence the result.

46. Show that the probability $P(r+1)$ of at least $(r+1)$ success in $(x+1)$ independent repeated Bernoulli trials with probability p of success at each trial, is equal to

$$(r+1)\binom{n+1}{r+1}\int_0^p x^r(1-x)^{n-r}dx$$

[**Hint.** $P(r+1)$ may be regarded as a function of p for r and n fixed.

Verify $\frac{d}{dp}\ P(r+1)=\frac{(n+1)\,!}{r\,!\,(n-r)\,!}p^rq^{n-r}$.]

47. Show that for the hypergeometric distribution

$$\sum_{x=0}^{n} {}^nC_x\frac{(Np)^{(x)}\,(Nq)^{(n-x)}}{N^{(n)}}.\quad x=0,\ 1,\ldots,\ n$$

the rth factorial moment about origin is

$$\frac{n^{(r)}(Np)^{(r)}}{N^{(r)}}. \qquad (I.C.A.R.\ Delhi)$$

[**Hint.** $\mu'_{(r)}=\sum_{x=0}^{n}\frac{1}{N^{(r)}}\cdot\binom{n}{x}M^{(x)}\ (N-M)^{(n-x)}\cdot(x)^{(r)}$

$$=\frac{1}{N^{(n)}}\cdot\sum_{y=0}^{n-r}M^{(n-y)}(N-M)^{(y)}\ (n-y)^{(r)}]$$

where $y=n-x$, $Nq=N-M$.

$$=\frac{M^{(r)}\ n^{(r)}}{N^{(r)}}\sum_{y=0}^{n-r}(M-r)^{n-r-y}(N-M)^{(y)}$$

$$=\frac{M^{(r)}\ n^{(r)}}{N^{(r)}}(N-r)^{(n-r)}=\frac{M^{(n)}\ n^{(r)}\ N^{(n)}}{N^{(n)}\ .\ N^{(r)}}$$

$$=\frac{n^{(r)}\ (Np)^{(r)}}{N^{(r)}}\ ]$$

48. Show that if X_1 and X_2 are independent Poisson variates with parameters λ_1 and λ_2, the conditional distribution of X_1 given X_1+X_2 namely $P(X_1=k,\ X_1+X_2=n)$ is binomial with probability $\lambda_1/(\lambda_1+\lambda_2)$.

(M.Sc. Bombay, 1970 ; M.Sc. Gauhati, 1967)

[**Hint.** $P[X_1=k\mid(X_1+X_2=n)=\frac{P[X_1=k\cap X_1+X_2=n]}{P(X_1+X_2=n)}$

$$=P(X_1=k)\ P(X_2=n-r)$$

$$\div P(X_1+X_2=n)$$

$$=\text{etc.}$$

49. Determine the mean deviation about mean for the negative binomial distribution.

50. If X has a Poisson distribution with parameter λ, show that

(*i*) $\sum_{x=0}^{k} e^{-\lambda} \frac{\lambda^x}{x!} = \frac{1}{\Gamma(k+1)} \int_{\lambda}^{\infty} e^{-z} z^k \, dz$ (*M.Sc. Poona, 1970*)

(*ii*) $\sum_{x=\lambda+\alpha\sqrt{\lambda}+\frac{1}{2}}^{\lambda+\beta\sqrt{\lambda}-\frac{1}{2}} \frac{e^{-\lambda} \lambda^x}{x!} \to \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{1}{2}t^2} \, dt$ as $\lambda \to \infty$.

(*M.A. Delhi, 1974*)

(*iii*) $P(X>k) = \frac{1}{\Gamma k+1} \int_{0}^{\lambda} e^{-t} t^k \, dt$ (*B.Sc. Hons. Calcutta 1970*

7

Continuous Random Variables and Probability Distributions

7·1. **Univariate Distributions.** We have so far dealt with discrete probability distributions, *i.e.*, distributions in which the stochastic variate takes either a finite set of values (as in Binomial and Hypergeometric Distributions) or at most a denumerably infinite set of values (as in Poisson and Negative Binomial Distributions). But when we deal with variates like heights and weights, we find that such variates can take a non-enumerably infinite set of values or more precisely they can take any value in the given interval $a \leqslant X \leqslant b$ of the *arithmetic continuum.* Such variates are called continuous variates and their probability distributions are accordingly known as *continuous probability distributions.* The concept of a continuous distribution is extremely useful for the following reasons :

(*i*) It is a logical necessity.

(*ii*) It is a theoretical necessity since in Statistics we deal not only with distributions occurring in nature but also with theoretical distributions and such distributions are very often continuous distributions.

(*iii*) It is easier mathematically to deal with a continuous distribution through the use of techniques of differentiation and integration than with a discrete distribution with a large number of values for the variate.

(*iv*) Continuous distributions are very good approximations to discrete distributions.

In continuous probability distributions, since we cannot associate a finite probability with each variate value as the sum of the probabilities will not converge (except possibly when the probability associated with all but a denumerable set of points is zero), we find

it more useful to associate probabilities with intervals. Thus $f(x)$, a continuous function of x, defines the probability distribution of the stochastic variate x by the relation that the probability of the value of the variate falling in the infinitesimal interval $x-\frac{1}{2}dx$ to $x+\frac{1}{2}dx$ is expressible in the form $f(x)dx$. Symbolically,

$$P(x-\tfrac{1}{2}dx \leqslant X \leqslant x+\tfrac{1}{2}dx)=f(x)dx.$$

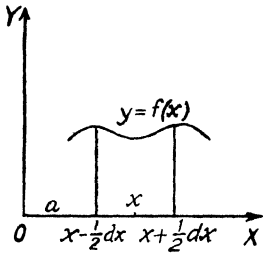


Fig. 15.

$f(x)$ is called the *probability density function* or simply the *density function* and $f(x)dx$ is called *probability differential*. The continuous curve $y=f(x)$ is called the *probability density curve* or briefly the *probability curve* and when this is symmetrical, the distribution is said to be *symmetrical*.

The range of the variate may be finite or infinite but even when the range is finite it is convenient to consider the range as infinite by simply postulating that the density function is zero outside the specified range. Thus if X takes values in the interval (a, b) only and its density function there is $\phi(x)$, its distribution is equivalent to that of the variate with density function defined as follows :

$$\begin{aligned} f(x)&=0, && x<a \\ &=\phi(x) && a\leqslant x\leqslant b \\ &=0, && x>b. \end{aligned}$$

The density function $f(x)$ possesses the following *properties* :

$$\left.\begin{aligned} &(i) && f(x)\geqslant 0 \text{ for very } x \\ &(ii) && \int_{-\infty}^{+\infty} f(x)dx=1 \end{aligned}\right\} \qquad \ldots(7{\cdot}1)$$

Property (i) is obviously necessary since negative probability has no meaning. Property (ii) corresponds to the requirement that the probability of an event that is certain to occur should be equal

to unity. The probability that the variate X falls in any interval (c, d) is given by

$$P(c \leqslant X \leqslant d) = \int_c^d f(x)dx.$$

The geometrical interpretation of the integral in the above expression is that it denotes the area bounded by the curve $y=f(x)$, the ordinates $x=c$, $x=d$ and the axis of x while

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

means that the total area under the curve is unity.

Let $$F(x) = \int_{-\infty}^{x} f(x)\ dx = P(X \leqslant x)$$

then the function $F(x)$ is the probability that the value of the variate X will be $\leqslant x$.

Thus $$F(b) = P(X \leqslant b)$$

and $$F(b) - F(a) = \int_a^b f(x)dx = P(a \leqslant X \leqslant b)$$

$F(x)$ is called the *cumulative distribution function* (*c.d.f.*) of x or simply *distribution function.* The *c.d.f.* $F(x)$ has obviously the following *properties* :

(*i*) $$F'(x) = f(x) \geqslant 0$$

so that $F(x)$ is a non-decreasing function,

(*ii*) $$F(-\infty) = \lim_{x \to -\infty} \int_{-\infty}^{x} f(x)\ dx = \int_{-\infty}^{\infty} f(x)\ dx = 0.$$

(*iii*) $$F(\infty) = \lim_{x \to \infty} \int_{-\infty}^{x} f(x)\ dx = \int_{-\infty}^{+\infty} f(x)\ dx = 1.$$

(*iv*) $F(x)$ is continuous on the right side.

(*v*) Discontinuities of $F(x)$ are at the most countable.

Since $F'(x) = f(x)$, the density function can be obtained by differentiating $F(x)$.

Example 7·1. *(a) Is the function defined as follows a density function ?*

$$f(x)=0 \qquad x<2$$
$$=\tfrac{1}{18}(3+2x), \qquad 2\leqslant x\leqslant 4$$
$$=0, \qquad x>4.$$

(b) Find the probability that a variate having this density will fall in interval $2\leqslant x\leqslant 3$.

(a) $f(x)$ is clearly $\geqslant 0$ for every x in the given interval and

$$\int_{-\infty}^{\infty} f(x)dx=\int_{-\infty}^{2} 0.dx+\int_{2}^{4}\frac{1}{18}(3+2x)+\int_{4}^{\infty}0.dx=1.$$

Hence the function $f(x)$ satisfies the requirements for density function.

(b) Required probability $=P(2\leqslant x\leqslant 3)$

$$=\int_{2}^{3}\frac{1}{18}(3+2x)dx=\tfrac{4}{9}.$$

Alternately,

$$F(x)=0 \qquad (x<2)$$
$$=\int_{2}^{x}\frac{1}{18}(3+2x)dx=\tfrac{1}{18}(x^2+3x-10), \qquad (2\leqslant x\leqslant 4)$$

$$\therefore \quad P(2\leqslant x\leqslant 3)=P(x\leqslant 3)-P(x<2)=F(3)-F(2)$$
$$=\tfrac{1}{18}(9+9-10)-0=\tfrac{4}{9}.$$

Example 7·2. *Find the constant k so that the function $f(x)$ defined as follows may be a density function :*

$$f(x)=\frac{1}{k}, \qquad a\leqslant x\leqslant b$$
$$=0, \qquad \textit{elsewhere.}$$

Find also the cumulative distribution of the random variable X when k satisfies the requirements for $f(x)$ to be a density function.

(i) If $k>0$, $f(x)\geqslant 0$ for every x. Hence for $f(x)$ to be a density function, we have

$$1=\int_{a}^{b} f(x)dx=\int_{a}^{b}\frac{1}{k}dx$$
$$=\frac{b-a}{k}, \text{ so that } k=b-a.$$

(*ii*) We have $F(x)=0 \qquad x<a$

$$=\int_a^x \frac{1}{k}\,dx=\frac{x-a}{b-a}, \qquad a\leqslant x\leqslant b$$

$$=1 \qquad x>b$$

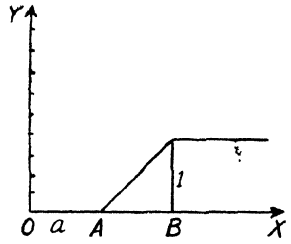


Fig. 17. Rectangular distribution.

Its graph is shown in Fig. 18.

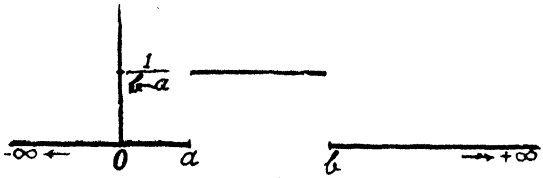


Fig. 18.

Note. Whenever $f(x)$ is constant throughout its range the variate is said to have a *uniform* or *rectangular distribution* of probability.

Example 7·3. *Verify that the following is a distribution function* :

$$F(x)=\begin{cases} 0 & x<-a \\ \frac{1}{2}\left(\frac{x}{a}+1\right), & -a\leqslant x\leqslant a \\ 1 & x>a \end{cases}$$

In order that $F(x)$ may be a distribution function, it must satisfy the five conditions given on page 161. It can be easily verified that the conditions hold good in this case.

Further,

$$f(x)=\frac{d}{dx}F(x)=\begin{cases} \frac{1}{2a}, & -a\leqslant x\leqslant a \\ 0, & \text{otherwise} \end{cases}$$

Now, $$\int_{-\infty}^{+\infty} f(x)dx=\int_{-a}^{a}\frac{1}{2a}\,dx+\int_{-\infty}^{-a}0.dx+\int_{a}^{\infty}0.dx=1$$

Hence $f(x)$ is the probability density function and $F(x)$ is the distribution function.

Example 7·4. *The probability distribution function of a random variable X is*

$$f(x)=\begin{cases} x & \text{for} \quad 0<x\leqslant 1 \\ 2-x & \text{for} \quad 1\leqslant x<2 \\ 0 & \text{for} \quad x\geqslant 2 \end{cases}$$

Compute the cumulative distribution function of X.

(B.Sc. Poona 1968 ; I.S.I. Calcutta 1969)

For any x in the range $x\leqslant 0$, we have

$$F(x)=\int_{-\infty}^{x} f(x)dx=\int_{-\infty}^{x} 0.dx=0$$

For any x in $0<x<1$,

$$F(x)=\int_{-\infty}^{0} f(x)dx+\int_{0}^{x} f(x)dx$$

$$=\int_{-\infty}^{0} 0\,dx+\int_{0}^{x} x\,dx=\tfrac{1}{2}x^2$$

For any x in $1<x\leqslant 2$,

$$F(x)=\int_{-\infty}^{0} f(x)dx+\int_{0}^{1} f(x)dx+\int_{1}^{x} f(x)dx$$

$$=\int_{-\infty}^{0} 0\,dx+\int_{0}^{1} x\,dx+\int_{1}^{x} (2-x)dx$$

$$=0+\frac{x^2}{2}\Big|_{0}^{1}+2x-\frac{x^2}{2}\Big|_{1}^{x}=\frac{1}{2}+2x-\frac{x^2}{2}-\frac{3}{2}$$

$$=2x-\tfrac{1}{2}x^2-1.$$

For every x in $x\geqslant 2$,

$$F(x)=\int_{-\infty}^{0} f(x)dx+\int_{0}^{1} f(x)dx+\int_{1}^{2} f(x)dx+\int_{2}^{\infty} f(x)dx$$

$$=\int_{-\infty}^{0} 0.dx+\int_{0}^{1} x\,dx+\int_{1}^{2} (2-x)dx+\int_{2}^{\infty} 0.dx$$

$$=0+\left[\frac{1}{2}x^2\right]_0^1+\left[2x-\frac{x^2}{2}\right]_1^2+0$$

$$=\tfrac{1}{2}+(2-\tfrac{3}{2})=1.$$

Example 7·5. *An angle θ is chosen at random between $-\pi/2$ and $\pi/2$ and a line is drawn through the point $P(o, a)$ at any angle θ with the x-axis. Find the density function for the point x at which this line crosses the x-axis.*

The position of X is determined by its distance $OX=x$ from O, this distance being positive or negative according as X is to the right or to the left of the point O. $\angle OPX=\theta$. The relation between θ and x is $x=a\tan\theta$ or $\theta=\tan^{-1}(x/a)$.

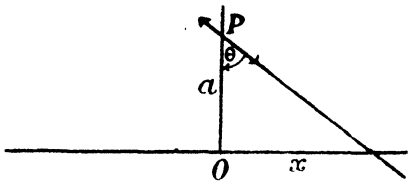


Fig. 19.

Consequently the relation between $d\theta$ and dx, obtained by differentiation, is

$$d\theta=\frac{a\,dx}{a^2+x^2}.$$

Assuming that the variable θ has uniform distribution of probability in the interval $-\pi/2$ to $+\pi/2$, the probability that θ will lie

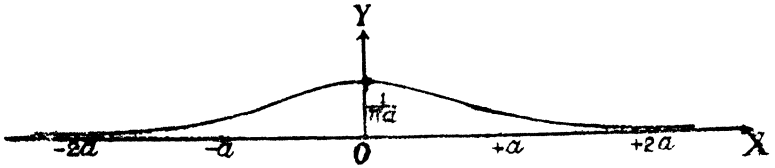


Fig. 20. Cauchy's curve $f(x)=\frac{1}{\pi}\cdot\frac{a}{x^2+a^2}$.

in the interval θ to $\theta+d\theta$ is $d\theta/\pi$ and the probability that the distance of x from O will be contained between x and $x+dx$ is the same.

Hence the probability density function for the distribution of the variable x is

$$f(x)=\frac{1}{\pi}\cdot\frac{a}{a^2+x^2}$$

Clearly
$$\frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{a\,dx}{a^2+x^2}=1.$$

The distribution is called *Cauchy's distribution*. It is named after its discoverer Cauchy (1789—1857).

Ex. Ratio active particles are being omitted from a point source P and scintillate against a plane sensitive screen, placed at unit distance from the source. If the particles travel in straight lines all directions are equally likely,

show that the frequency (probability) distributions of scintillations along any line on the screen passing through the foot of the perpendicular Q from P is given by

$$\frac{dx}{\pi\{1+(x-a)^2\}}$$

where a is the distance of O and x that of any point from the origin taken anywhere on the line.

7·2. Definitions of Moments etc. for Univariate Continuous Probability Distributions. In extending the definition and theorems of discrete probability distributions to continuous probability distributions we shall observe that formally it consists in replacing p_i by $f(x)dx$, x_i by x and summation over i by integration over the permissible range for x.

Let X be a continuous random variable with probability density function $f(x)$ and *c.d.f.* $F(x)$. The mathematical expectation of any function $\phi(x)$ is given by

$$E[\phi(x)]=\int_{-\infty}^{\infty}\phi(x)\,f(x)dx=\int_{-\infty}^{\infty}\phi(x)dF(x) \qquad \ldots(7\cdot2)$$

provided this integral *exists* and is called the mean value or expected value of $\phi(x)$.

In particular,

$$\mu_1'=E(x)=\int_{-\infty}^{\infty}xf(x)dx=\int_{-\infty}^{\infty}xdF(x)$$

provided the integral exists, *i.e.*, if $\int_{-\infty}^{+\infty}|x|\,f(x)dx<\infty$.

$$\mu_r'\text{ (about origin)}=E(x^r)=\int_{-\infty}^{\infty}x^rf(x)dx \qquad \ldots(7\cdot3)$$

$$\mu_r=E[(x-\bar{x})^r]=\int_{-\infty}^{\infty}(x-\bar{x})^r\,f(x)dx \qquad \ldots(7\cdot4)$$

$$M_0(t)=E(e^{tx})=\int_{-\infty}^{\infty}e^{tx}f(x)dx \qquad \ldots(7\cdot5)$$

$$M_u(t)=E\left[e^{t\left(\frac{x-\bar{x}}{\sigma}\right)}\right]=\int_{-\infty}^{\infty}e^{t\left(\frac{x-\bar{x}}{\sigma}\right)}f(x)dx$$

where $u=(x-\bar{x})/\sigma$.

or $$M_u(t)=e^{-\frac{t\bar{x}}{\sigma}}\int_{-\infty}^{\infty}e^{\frac{tx}{\sigma}}f(x)dx=e^{-\frac{t\bar{x}}{\sigma}}M_o\left(\frac{t}{\sigma}\right) \quad ...(7\cdot6)$$

Geometric mean G, is given by

$$\log G=\int_{-\infty}^{\infty}\log x.f(x)dx=E(\log x) \quad ...(7\cdot7)$$

provided the integral exists. Of course, when x is negative, $f(x)$ should be zero for G to be real.

Harmonic mean, H, is defined by

$$\frac{1}{H}=\int_{-\infty}^{\infty}\frac{1}{x}f(x)dx=E\left(\frac{1}{x}\right) \quad ...(7\cdot8)$$

Median 'a' is the value of the variate such that x is equally likely to be greater than or less than a, *i.e.*,

$$P(x\leqslant a)=P(x\geqslant a)$$

or $$\int_{-\infty}^{a}f(x)dx=\tfrac{1}{2}=\int_{a}^{\infty}f(x)dx \quad ...(7\cdot9)$$

The lower and the upper quartiles Q_1 and Q_3 are given by

$$\int_{-\infty}^{Q_1}f(x)dx=\tfrac{1}{4}=\int_{Q_3}^{\infty}f(x)dx \quad ...(7\cdot10)$$

Mode or modal value is that value of the variate for which the probability density is maximum so that if probability density function $f(x)$ is continuous and differentiable, the mode is defined by the root satisfying

$$\left.\begin{aligned}f'(x)&=0\\f''(x)&<0\end{aligned}\right\} \quad ...(7\cdot11)$$

provided that the solution of $f'(x)=0$ lies within the permissible range of the variate.

Mean deviation from the mean $\bar{x}$ is given by

$$\text{M D.}=\int_{-\infty}^{\infty}|x-\bar{x}|f(x)dx \quad ...(7\cdot12)$$

In terms of the cumulative distribution function,

$$\left.\begin{aligned} F(Q_1)&=\tfrac{1}{4} \\ F(Q_2 \text{ or Median})&=\tfrac{1}{2} \\ F(Q_3)&=\tfrac{3}{4} \\ F(D_r)&=\frac{r}{10} \end{aligned}\right\} \qquad \ldots(7{\cdot}13)$$

where D_r is the rth decile.

Mean difference (due to Gini) is defined by

$$\Delta_1=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\, x-y \,|\, f(x)f(y)\, dx\, dy \qquad \ldots(7{\cdot}14)$$

Theorems on Expectation

(*a*) If X and Y are continuous random variables, then

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

(*b*) If X and Y are independent continuous random variables, then

$$E(XY)=E(X)\ E(Y).$$

Proof. (*a*) By definition,

$$E(X+Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)\, f_{XY}(x, y)\, dx\, dy$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{XY}(x, y)dxdy+\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y\, f_{XY}(x, y)\, dx\, dy$$

$$=\int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f_{XY}(x, y)\, dy\right] dx+\int_{-\infty}^{+\infty} y\left[\int_{-\infty}^{+\infty} f_{XY}(x, y)dx\right]dy$$

$$=\int_{-\infty}^{\infty} x\, f_X(x)\, dx+\int_{-\infty}^{\infty} y\, f_Y(y)\, dy$$

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

(*b*) $E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\, f_{XY}(x, y)\, dx\, dy$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\, f_X(x)\, f_Y(y)\, dx\, dy,$$

since X and Y are independent

$$=\left[\int_{-\infty}^{+\infty} x\, f_X(x)dx\right]\left[\int_{-\infty}^{+\infty} y\, f_Y(y)\, dy\right]$$

$$=E(X)\, E(Y).$$

Example 7·6. *If μ_r' exists, then μ_s' exists for all $1 \leqslant s \leqslant r$, i.e. if $E(x^r)$ exists, then $E(x^s)$ exists for all $1 \leqslant s \leqslant r$.*

(M. Sc., Nagpur, 1968)

We have

$$E(|x|^s)=\int_{-\infty}^{\infty} |x|^s\, dF(x)=\int_{-1}^{1} |x|^s\, dF(x)+\int_{|x|>1} |x|^s\, dF(x)$$

$$\leqslant \int_{-1}^{1} |x|^r\, dF(x)+\int_{|x|>1} |x|^r\, dF(x) \quad [\because |x|^s \leqslant |x|^r]$$

since $|x|^s < |x|^r$ for $s<r$ and $|y| > 1$.

Hence $E(|x|^s) \leqslant \int_{-1}^{1} dF(x)dx + \int_{|x|>1} |x|^r\, dF(x)$

Thus $E(|x|^s) \leqslant 1+E(|x|^r)$

Hence the result follows.

Example 7·7. *Show that the mean deviation from median is less than that measured from any other value.*

(B.A. Hons., Delhi, 1950; I.A.S. 1947)

The mean deviation about an arbitrary point a is given by

$$F(a)=\int_{-\infty}^{\infty} |x-a|\, f(x)dx$$

where $$\int_{-\infty}^{\infty} f(x)dx=1 \qquad \ldots(1)$$

$$\Rightarrow \qquad \int_{-\infty}^{a} (a-x)\, f(x)dx+\int_{a}^{\infty} (x-a)\, f(x)dx$$

Differentiating with respect to a under the sign of integration*, we have

$$\frac{d}{da}F(a)=\int_{-\infty}^{a} f(x)\ dx-\int_{a}^{\infty} f(x)\ dx,$$

and $$-\frac{d^2}{da^2}F(a)=f(a)+f(a)=2f(a)>0,$$

assuming that $f(x)$ is not zero at the median value.

Thus for $F(a)$ to be minimum, we should have

$$\int_{-\infty}^{a} f(x)\ dx=\int_{a}^{\infty} f(x)\ dx$$

Also from (1), $$\int_{-\infty}^{a} f(x)\ dx+\int_{a}^{\infty} f(x)\ dx=1$$

Thus $$\int_{-\infty}^{a} f(x)dx=\tfrac{1}{2}=\int_{a}^{\infty} f(x)dx$$

Example. 7·8. *A continuous distribution of a variable* X *in the range* $(-3, 3)$ *is defined by*

$$f(x)=\frac{1}{16}(3+x)^2,\quad -3\leqslant x\leqslant -1$$

$$=\frac{1}{16}(2-6x^2),\quad -1\leqslant x\leqslant 1$$

$$=\frac{1}{16}(3-x)^2,\quad 1\geqslant x\geqslant 3.$$

verify that the area under the curve is unity. Show that the mean is zero. (*B.Sc. Jiwaji, 1971, B.Sc. Allahabad, 1959*)

If $f(x, \alpha)$ is a continuous function of both variables x and α possessing continuous partial derivatives $\frac{\partial^2 f}{\partial x \partial \alpha}$, $\frac{\partial^2 f}{\partial \alpha \partial x}$ and a and b are differentiable functions of α, then

$$\frac{d}{dx}\left[\int_{a}^{b} f(x, \alpha)dx\right]=\int_{a}^{b}\frac{\partial f}{\partial \alpha}\ dx+f(b, \alpha)\ \frac{db}{d\alpha}-f(a, \alpha)\frac{da}{d\alpha}.$$

We have

$$\int_{-3}^{+3} f(x)dx = \int_{-3}^{-1} f(x)dx + \int_{-1}^{+1} f(x)dx + \int_{1}^{3} f(x)dx$$

$$= \int_{-3}^{-1} \frac{(3+x)^2}{16}\,dx + \int_{-1}^{+1} \frac{6-2x^2}{16}\,dx + \int_{1}^{3} \frac{(3-x)^2}{16}\,dx$$

$$= \frac{(3x+x)^3}{48}\Bigg|_{-3}^{-1} + \frac{6x-\frac{2x^3}{3}}{16}\Bigg|_{-1}^{+1} + \frac{(3-x)^3}{-48}\Bigg|_{1}^{3}$$

$$= \frac{1}{48}\left[\ 8+32+8\ \right] = 1.$$

Hence the area under the curve is unity.

$$\text{Mean} = E(x) = \int_{-3}^{+3} x\,f(x)\,dx$$

$$= \int_{-3}^{-1} \frac{x(3+x)^2}{16}\,dx + \int_{-1}^{1} \frac{x(6-2x)^2}{16} + \int_{1}^{3} \frac{x(3-x)^2}{16}\,dx$$

$$= \frac{1}{16}\left[\int_{-3}^{-1} (9x+6x^2+x^3)dx + \int_{-1}^{1} (6x-2x^3)dx + \int_{1}^{3} (9x-6x^2+x^3)dx\right.$$

$= 0$, on simplification.

so that a is the median value.

Example 7·9. *Show that for the exponential distribution*

$$dP = y_0 e^{-\frac{x}{\sigma}}\,dx,\quad 0 \leqslant x < \infty,\quad \sigma > 0,\quad y_0 \text{ being a constant}$$

the mean and the s d. are each equal to σ *and that the interquartile range is* $\sigma \log_e 3$. *Also find* μ_r' *and show that* $\beta_1 = 4$ *and* $\beta_2 = 9$.

(*M.A. Punjab, 1958* ; *I.A.S., 1954* ; *B.A. Hons., Delhi, 1953*)

The constant y_0 is given by

$$y_0 \int_0^{\infty} e^{-\frac{x}{\sigma}}\,dx = 1, \text{ so that } y_0 = \frac{1}{\sigma}$$

$$\mu_c' \text{ (about } x=0) = \frac{1}{\sigma}\int_0^{\infty} x^r e^{-\frac{x}{\sigma}}\,dx = r\,!\,\sigma^r \text{ by Gamma integral.}$$

$\therefore \quad \mu_1' = \sigma,\ \mu_2' = 2\sigma^2;\ \mu_3' = 6\sigma^3,\ \mu' = 24\sigma^4$

$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = \sigma^2,\ \mu_3 = 2\sigma^3,\ \mu_4 = 9\sigma^4.$

Hence $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = 4$ and $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 9.$

Also $\displaystyle\int_0^{Q_1} \frac{1}{\sigma} e^{-\frac{x}{\sigma}} dx = \frac{1}{4}$ so that $e^{-\frac{Q_1}{\sigma}} = \frac{3}{4}$ ...(A)

and $\displaystyle\int_0^{Q_3} \frac{1}{\sigma} e^{-\frac{x}{\sigma}} dx = \frac{3}{4}$ or $e^{-\frac{Q_3}{\sigma}} = \frac{1}{4}.$ ...(B)

Dividing (A) by (B), we get

$$e^{\frac{Q_3 - Q_1}{\sigma}} = 3 \quad \text{or} \quad Q_3 - Q_1 = \sigma \log_e 3.$$

Example 7·10. *Show that for the rectangular distribution*

$$dP = dx,\ 0 \leqslant x \leqslant 1$$

$\mu_1' = \frac{1}{2},\ \mu_2 = 1\frac{1}{12}$, *mean deviation* $= \frac{1}{4}$.

(B.Sc. Mysore, 1971 ; B.Sc. Delhi, 1973 ; I.A.S. 1950)

Here $f(x) = 1, \quad 0 \leqslant x \leqslant 1$

$\therefore$ We have

$$\mu_1' = \int_0^1 x\, dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}, \qquad \mu_2' = \int_0^1 x^2\, dx = \frac{1}{3}$$

$$\therefore \quad \mu_2 = \frac{1}{3} - \frac{1}{4} = 1\frac{1}{12}$$

Mean deviation (about the mean)

$$= \int_0^1 |\, x - \tfrac{1}{2}\,|\, dx = \int_0^{\frac{1}{2}} (\tfrac{1}{2} - x) dx + \int_{\frac{1}{2}}^1 (x - \tfrac{1}{2}) dx$$

$$= \left[\frac{x}{2} - \frac{x^2}{2}\right]_0^{\frac{1}{2}} + \left[\frac{x^2}{2} - \frac{x}{2}\right]_{\frac{1}{2}}^1 = \frac{1}{4}.$$

Ex. 1. Show that the mean deviation from the mean is not greater than the standard deviation.

Ex. 2. If (*i*) the origin is taken at the median, (*ii*) $a > 0$ and (*iii*) $F(a)$ denotes the mean deviation from the point a, show that

$$F(a) - F(0) = \int_{-\infty}^{a} (a - x) f(x) dx + \int_a^{\infty} (x - a) f(x) dx - \int_{-\infty}^{0} (-x) f(x) dx - \int_0^{\infty} x f(x) dx$$

$$=2\int_0^a (a-x)\,f(x)dx>0.$$

Similarly, if $a<0$, show that $F(a)-F(0)>0$,

Now deduce the result of Example 7·7.

Ex. 3. For rectangular distribution

$$dP=dx,\ 1\leqslant x\leqslant 2$$

show that $$A>G>H.$$

Ex. 4. (*a*) For the triangular distribution

$$dP=\frac{1}{a}\left[1-\frac{|x-b|}{a}\right]dx,\ |x-b|<a.$$

show that the mean is b and variance is $a^2/6$.

(*b*) For the triangular distribution with the density function

$$f(x)=1-|1-x|,\ 0<x<2,$$

show that the mean is 1, variance $\frac{1}{6}$ and $\mu_3=0$ (*B.Sc. Shivaji, 1970*)

Example 7 11. *Show that for a continuous distribution*

$$\triangle_1=2\int_{-\infty}^{\infty} F(x)\{1-F(x)\}dx.$$

In particular for the rectangular distribution

$$dF=dx,\ 0\leqslant x\leqslant 1,$$

show that $$\triangle_1=\tfrac{1}{3}.$$

$$\triangle_1=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|x-y|\,dF(x)dF(y) \text{ where } dF(x)=f(x)dx$$

$$=2\int_{-\infty}^{\infty}\int_{y=-\infty}^{x}(x-y)dF(x)dF(y)$$

$$=2\int_{-\infty}^{\infty}dF(x)\int_{-\infty}^{x}(x-y)dF(y). \qquad ...(1)$$

$$\text{Now } \int_{-\infty}^{x}(x-y)dF(y)=\Big[(x-y)F(y)\Big]_{-\infty}^{x}+\int_{-\infty}^{x}F(y)dy=\int_{-\infty}^{x}F(y)dy$$

$$\therefore \text{ From (1), } \Delta_1 = 2\int_{-\infty}^{\infty} dF(x)\phi(x) \text{ where } \phi(x) = \int_{-\infty}^{x} F(y)dy$$

$$= \int_{-\infty}^{x} F(x)dx$$

$$\therefore \qquad \Delta_1 = 2\Big[F(x)\phi(x)\Big]_{-\infty}^{\infty} - 2\int_{-\infty}^{\infty} F(x)F(x)dx$$

$$= 2\int_{-\infty}^{\infty} F(x)\{1 - F(x)\}dx,$$

since

$$\Big[F(x)\phi(x)\Big]_{-\infty}^{\infty} = 1\int_{-\infty}^{\infty} F(x)dx - (0)(0) = \int_{-\infty}^{\infty} F(x)dx.$$

In the particular case,

$$F(x) = \int_0^x dF = \int_0^x dx = x$$

$$\Delta_1 = 2\int_0^1 x(1-x)dx = \tfrac{1}{6}.$$

Example 7·12. *Show that the variance may be defined as half the mean of all possible variate difference, that is to say without reference to deviation from a central value, the mean.*

(I.A.S., 1949 ; I.C.A.R., Delhi, 1948)

$$E^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-y)^2 f(x)\, f(y)dxdy$$

$$= \int_{-\infty}^{\infty} x^2 f(x)dx \int_{-\infty}^{\infty} f(y)dy - 2\int_{-\infty}^{\infty} xf(x)dx \int_{-\infty}^{\infty} yf(y)dy + \int_{-\infty}^{\infty} f(x)dx \int_{-\infty}^{\infty} y^2 f(y)dy$$

$$= \mu_2'.\ 1 - 2\mu_1'.\ \mu_1' + 1\ .\ \mu_2'$$

$$= 2(\mu_2' - \mu_1'^2) = 2\mu_2 = 2 \text{ (variance).}$$

Hence the result follows.

Ex. 1. Show that, if y and z are independent random values of a variate x, the expected value of $(y-z)^2$ is twice the variance of the distribution of x.

Ex. 2. Show that if mean exists, $\Delta_1 = 4\int_{-\infty}^{+\infty} x[F(x)-\tfrac{1}{2}]dF$ where $dF=f(x)dx$

In particular, for the rectangular distribution $dF=dx$, $0 \leqslant x \leqslant 1$, show that $\Delta_1=\frac{1}{3}$.

[**Hint.** $\Delta_1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |x-y|\, dF(x)dF(y) = 2\int_{-\infty}^{\infty} dF(x)\int_{-\infty}^{x}(x-y)F(y)$

$$= 2\int_{-\infty}^{\infty} dF(x)\left\{\Big(E(y)(x-y)\Big)_{-\infty}^{x} + \int_{-\infty}^{\infty} F(y)dy\right\} = 2\int_{-\infty}^{\infty} dF(x)\int_{-\infty}^{x} F(y)dx$$

$$= 2\int_{-\infty}^{\infty} dF(x)\Big[yF(y)\Big]_{-\infty}^{x} - \int_{-\infty}^{x} ydF(y)$$

$$= 2\int_{-\infty}^{\infty} dF(x)\Big[xF(x)\Big]_{-\infty}^{x} - 2\int_{-\infty}^{\infty} dF(x)\int_{-\infty}^{x} ydF(y)$$

$$= 2\int_{-\infty}^{\infty} dF(x)\Big[xF(x)\Big]_{-\infty}^{x} - 2\int_{-\infty}^{\infty} ydF(y)\int_{y}^{\infty} dF(x)$$

$$= 2\int_{-\infty}^{\infty} dF(x)\Big[xF(x)\Big]_{-\infty}^{x} - 2\int_{-\infty}^{\infty} ydF(y)\int_{y}^{\infty} dF(x)$$

$$= 2\int_{-\infty}^{\infty} dF(x)\Big[xF(x)\Big]_{-\infty}^{x} - 2\int_{-\infty}^{\infty} ydF(y)\Big[1-F(y)\Big]$$

$$= 2\int_{-\infty}^{\infty} xF(x)dF(x) - 2\int_{-\infty}^{\infty} x(1-F(x)dF(x)) = \text{etc.}$$

7·3. Characteristic Function. The characteristic function of a random variable X for the continuous probability distribution with density function $f(x)$ is defined as

$$\phi_X(t) = E\,(e^{itX}) = \int_{-\infty}^{\infty} e^{itX} f(x)dx = \int_{-\infty}^{\infty} e^{itX} dF(x)$$

$\phi(t)$ has the following *properties* :

(*i*) $\phi(t)$ is continuous in t.

(*ii*) $\phi(t)$ is defined in every finite t interval.

(*iii*) $\phi(0)=1$.

(*iv*) $\phi(t)$ and $\phi(-t)$ are conjugate quantities.

$$(v) \quad |\phi(t)| \leqslant \int_{-\infty}^{\infty} |e^{itx}| f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1 = \phi(0).$$

Since $|\phi(t)| \leqslant 1$, characteristic function $\phi(t)$ always exists.

These are necessary but not sufficient conditions for $\phi(t)$ to be a characteristic function of some continuous distribution. Thus $\exp(t^4)$ satisfies all the above five conditions, yet cannot be a characteristic function since it would mean that the distribution has zero mean and zero variance. In fact no very simple set of sufficient conditions for $\phi(t)$ to be a characteristic function are known. One of the simplest set of necessary and sufficient conditions, due to Cramer, is given by the following theorem :

In order that a given bounded and continuous function $\phi(t)$ should be the characteristic function of a distribution it is necessary and sufficient that $\phi(0)=1$ and that the function

$$\phi(z)=\int_0^{\infty}\int_0^{c} \phi(t-z)e^{ix(t-b)}\,dtdx$$

is real and non-negative for all real x and all $c>0$.

The *advantage* of the characteristic function over *M.G.F.* is that while the integral for the *M.G.F.* may be divergent, integral for the *C.F.* cannot be so in view of property (*v*) above and, therefore, in proving theorems the *C.F.* is invariably used. The use of the *C.F.* also enables us to solve the converse problem, *i.e.*, given the characteristic function, to find the probability density function.

7·3·1. Fourier's Inversion Theorem.

Let $F(x)$ and $\phi_X(t)$ be the distribution function and characteristic function respectively of a random variable X, then

$$F(x)-F(0)=\operatorname*{Lim}_{t\to\infty} \frac{1}{2\pi}\int_{-t}^{t}\frac{1-e^{itx}}{it}\phi(t)dt = P.\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1-e^{itx}}{it}\phi(t)dt$$

where P denotes the principal value of the integral.

If $F(x)$ is continuous everywhere and $dF(x)=f(x)\ dx$, we can write

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt \qquad \ldots(7\cdot15)$$

In the latter case the inversion theorem is equivalent to the Fourier Integral theorem *viz.*

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} e^{it(x-y)}F(y)dy$$

and can be easily deduced from the latter by the substitution

$$\phi(t)=\int_{-\infty}^{\infty} e^{-ity}f(y)dy \qquad \ldots(7\cdot16)$$

The reciprocal relation between the density function $f(x)$ and the characteristic function $\phi(t)$ is obvious on comparing (7·15) and (7·16).

Example 7·13. *Show that the distribution for which the characteristic function is* $e^{-|t|}$ *has the density function*

$$f(x)=\frac{1}{\pi}\cdot\frac{1}{1+x^2},\quad -\infty<x<\infty.$$

(M.A., Delhi 1966 ; M.A., Poona, 1959 ; M.A. Bombay, 1956 ; M.A., Nagpur, 1953)

We have $$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-|t|}e^{-itx}dt \qquad \ldots(7\cdot15)$$

$$=\frac{1}{2\pi}\left[\int_{-\infty}^{0} e^{t}e^{-itx}dt+\int_{0}^{\infty} e^{-t}e^{-itx}dt\right]$$

or $$f(x)=\frac{1}{2\pi}\int_{0}^{\infty} e^{-t}(e^{itx}+e^{-itx})dt=\frac{1}{\pi}\int_{0}^{\infty} e^{-t}\cos tx\ dt.$$

Integrating by parts gives

$$f(x)=\frac{1}{\pi}\left[-e^{-t}\cos tx\right]_{0}^{\infty}-\frac{x}{\pi}\int_{0}^{\infty} e^{-t}\sin tx\ dt$$

$$= \frac{1}{\pi} + \frac{x}{\pi}\left[e^{-t}\sin tx\right]_0^{\infty} - \frac{x^2}{\pi}\int_0^{\infty} e^{-t}\cos tx\, dt$$

$$= \frac{1}{\pi} - x^2 f(x).$$

Thus $$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, \quad -\infty < x < \infty.$$

Example 7·14. *Show that if θ is a non-negative function of the random variate x and*

$$\alpha(t) = \int_{-\infty}^{+\infty} \theta^t d\, F(x) = \int_{-\infty}^{+\infty} \theta^t f(x)dx$$

then the density function of θ, if any, is given by

$$f(\theta) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \theta^{-t-1}\alpha(t)dt$$

(M.A., Nagpur, 1953 ; I.C.A.R., Delhi, 1952)

We have $$\alpha(t) = \int_{-\infty}^{\infty} \theta^t dF(x) = E(\theta^t) = E(e^{t \log \theta}) = E(e^{tv})$$

where $$v = \log_e \theta.$$

Thus $\alpha(t)$ is the moment generating function of v, if the distribution of v exists. Moreover, if it exists it can be found with the help of the inversion theorem thus :

Characteristic function of $v = \phi(T) = E(e^{iTv}) = \alpha(iT)$

$$\therefore \quad g(v) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ivT}\phi(T)dT = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{-vt}\alpha(t)dt$$

where $$iT = t.$$

Distribution of v is $dF = g(v)dv$ so that to find the distribution of θ, we write $dF = g(v)\frac{dv}{d\theta}$. $d\theta = \frac{1}{b}g(v)d\theta = f(\theta)d\theta$, where $f(\theta) = \frac{1}{\theta}g(v)$.

Hence the frequency function of θ, if any, is given by

$$f(\theta) = \frac{1}{\theta}g(v) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty}\frac{1}{\theta}\, e^{-t \log \theta}\, \alpha(t)dt = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty}\frac{1}{\theta}\,\theta^{-t}\alpha(t)dt$$

$$= \frac{1}{2\pi i}\int_{-i\infty}^{+i\infty} \theta^{-t-1}\alpha(t)dt.$$

Example 7·15. *If x_1, x_2, x_3, x_4 are chance variables whose distribution has density function given by*

$$\frac{1}{(2\pi)^2} \quad e^{-\frac{1}{2}(x_1^2+x_2^2+x_3^2+x_4^2)}$$

show that $z=x_1x_2-x_3x_4$ has the density function given by $\frac{1}{2}e^{-|z|}$.
(M.A. Bombay, 1956 ; M.A. Delhi 1951)

$$\phi_z(t) = \frac{1}{(2\pi)^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{it(x_1x_2-x_3x_4)} \quad e^{-\frac{1}{2}\Sigma x_i^2}\,dx_1dx_2dx_3dx_4$$

$$= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{\frac{1}{2}(x_1^2-2it\,x_1x_2+x_2^2)}dx_1dx_2$$

$$\times \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x_3^2+2it\,x_3x_4+x_4^2)}dx_3dx_4$$

$$= \frac{1}{\sqrt{1+t^2}} \cdot \frac{1}{\sqrt{1+t^2}} = \frac{1}{1+t^2}$$

using $$\frac{1}{\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(ax^2+2hxy+by^2)}dxdy = \frac{1}{\sqrt{ab-h^2}}$$

$$\therefore \quad f(z)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-itz}\,dt}{1+t^2} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\cos tz\,dt}{1+t^2} = \tfrac{1}{2}e^{-|z|}$$

a, b, h are all positive.

Ex. 1. Are e^{-t^4}, $e^{-\frac{1}{2}t^2}$, $|\sin t|$, $t^{\log t}$, e^{it} and $\frac{1}{1+t^2}$ characteristic functions ? If not, why not ? If so, find the corresponding density functions.
(M.A. Statistics, Delhi, 1955 ; M.A. Bombay, 1956 ; M.A. Delhi, 1951)

Ex. 2. A variate has moments $\mu_r'=r!$. Find the characteristic function and deduce its distribution.

[**Hint.** $\phi(t)=\sum_{r=0}^{\infty}\frac{(it)^r}{r!}=\sum_{0}^{\infty}(it)^r=\frac{1}{1-it}$ [**Ans.** $f(x)=e^{-x}$, $0\leqslant x\leqslant\infty$]

Ex. 3. Find the frequency density of the distribution for which $k_r=(r-1)!$
(M.A. Statistics, Delhi, 1961)

[**Hint.** $\log\phi(t)=\sum_{r=1}^{\infty}\frac{k_r(it)^r}{r!}=$etc.] [**Ans.** $f(x)=e^{-x}$, $0\leqslant x<\infty$]

Ex. 4. A variate x has moments $\mu'_r=\frac{r!}{(r/2)!}$ when r is even and $\mu_r'=0$ when r is odd. Deduce the distribution of the variate from its characteristic function. *(M.A. Statistics, Delhi, 1959)*

Ex. 5. Starting with the theorem of Fourier transform, prove the results of Ex. 7·14, page 178. *(I.C.A.R. Delhi, 1950)*

Theorems on Characteristic Function (c.f.)

1. *If $\phi_X(t)$ is the c.f. of the random variable X, then $e^{ita}\phi_X(at)$ is the c.f. of the random variable $aX+b$, a and b being real constants.*

Proof. We have, by def.,

$$\phi_{aX+b}\phi(t)=E[e^{it(aX+b)}]=E[e^{itb}\cdot e^{itaX}]$$

Cor. $\phi_{ax}(t)=\phi_x(at)=e^{itb}E[e^{itax}]=e^{itb}\phi_x(at)$

2. *If ϕ is the c.f. of X, then $\bar{\phi}$ is the c.f. of $-X$.*

Proof. $\phi_X(t)=E[e^{it(-X)}]=E[e^{-itX}]$

$$=\phi_X(-t)=\bar{\phi}(t)$$

3. *If the distribution function of a random variable X is symmetrical about zero i.e., $f(-x)=f(x)$, then $\phi(t)$ is real valued and even function of t.*

Proof. We have

$$\phi(t)=E(e^{itx})=\int_{-\infty}^{+\infty} e^{itx}f(x)\,dx$$

$$=\int_{+\infty}^{-\infty} e^{-ity}f(-y)(-dy)\ ;\ x=-y$$

$$=\int_{-\infty}^{+\infty} e^{-ity}f(y)dy$$

$$=\int_{-\infty}^{+\infty} e^{-itx}f(x)dx$$

$$=\phi(-t)$$

Thus $\phi(t)$ is an even function of t.

Also $\bar{\phi}(t)=E[e^{-itx}]$

$$=E[e^{i(-t)x}]$$

$$=\phi(-t)=\phi(t)$$

Hence $\phi(t)$ is a real-valued function of t.

4. *The c.f. of the sum of independent random variables is equal to the product of their individual characteristic functions.*

In symbols, if $X_1, X_2, \ldots, X_n$ are n independent random variables, then

$$\phi_{X_1+X_2+\ldots+X_n}(t)=\phi_{X_1}(t)\phi_{X_2}(t)\ldots\phi_{X_n}(t)=\prod_{i=1}^{n}\phi_{X_i}(t)$$

Proof. $\phi_{X_1+X_2\cdots+X_n}(t) = E[e^{it(X_1+X_2+\cdots+X_n)}]^2$

$$= E(e^{itX_1}\ e^{itX_2}\ \ldots e^{itX_n}]$$

$$= E[\prod_{j=1}^{n} e^{itX_j}]$$

$$= \prod_{j=1}^{n} E[e^{itX_j}] = \prod_{j=1}^{n} \phi_{X_j}(t)$$

which establishes the result to be proved.

Cor. If X_j's are identically distributed, then

$$\phi_{\sum_{j=1}^{n} X_j}(t) = [\phi_X(t)]^n$$

Effect of change of origin and of scale on Characteristic Function. Let us introduce a new variable u where

$u = \dfrac{X-a}{h}$, a and h being real constants then by theorem I, it follows that

$$\phi_u(t) = e^{-\frac{iat}{h}}\ \phi_X\left(\frac{t}{h}\right)$$

In particular, if a=mean=μ, h=s.d.=σ, then the c.f. of the standard variable u is given by

$$\phi_u(t) = e^{-\frac{it\mu}{\sigma}}\ \phi_X\left(\frac{t}{\sigma}\right).$$

It should be noted that a necessary and sufficient condition for the convergence of the sequence $\{F_n(x)\}$ to be a distribution function $F(x)$ is that for every t, the sequence $\{\phi_n(t)\}$ converges to a limit $\phi(t)$ which is continuous for the value $t=0$. When the condition is satisfied, then limit $\phi(t)$ is identical with the c.f. of $F(x)$.

Example. 7·16. *Let $F_n(x)$ be the distribution function defined by*

$$F_n(x) = 0 \qquad \text{for} \qquad x \leqslant -n$$

$$= \frac{x+n}{2n} \qquad \text{for} \quad -n < x < n$$

$$= 1 \qquad \text{for} \qquad x \geqslant n$$

Is the limit of $F_n(x)$ a distribution function ? If not, why ?

Now
$$\phi_n(t) = \int_{-n}^{n} e^{itx} \frac{1}{2n} dx$$
$$= \frac{1}{2n}\left[\frac{e^{itn} - e^{-itn}}{it}\right] = \frac{\sin nt}{nt}$$

$\therefore$
$$\phi(t) = \lim \phi_n(t)$$
$$= \lim_{n\to\infty} \frac{\sin nt}{nt} = \begin{cases} 1 \text{ if } t=0 \\ 0 \text{ if } t\neq 0 \end{cases}$$

Thus $\phi(t)$ is discontinuous at $t=0$.

7·4. Geometrical Probability. When points are chosen 'at random' in a straight or curved line or in a plane area or in a volume or when lines are drawn 'at random' on a plane area, same lengths or angles or areas equally likely and then distribution is uniform. Problems of this type can usually be easily solved by geometrical interpretation of the conditions given.

Example 7·17. *Three points are taken at random on the circumference of a circle. Find the probability that they lie on the same semicircle.*

Since we are concerned with relative positions only, we may regard one point as fixed ; call this as A. Selecting one of the other two points as B, we define the position of B by the arc ACB measured anti-clockwise from CA to CB, C being the centre. Let the third point be D. There are two cases according as $\angle ACB$ $(=\theta)$ is (i) $<\pi$, or (ii) $>\pi$.

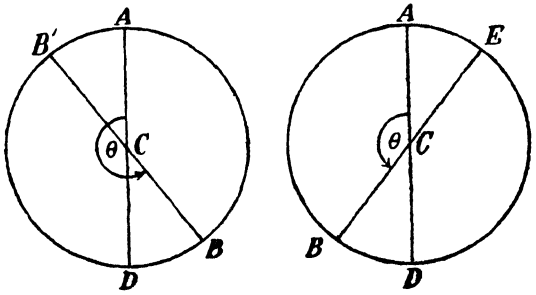


Fig. 21.

Let AD and BE be the diameters. Then the points A, B and D will be on the semi-circle provided that B is not the arc DE. The probability for this is :

$$\frac{2\pi - \theta}{2\pi}, \quad \text{if } 0 < \theta < \pi$$

$$\frac{\theta}{2\pi}, \quad \text{if } \pi < \theta < 2\pi$$

Combining these with the probability that B falls within a small arc of angle $d\theta$ which is $d\theta/2\pi$, we have

$$\text{the required probability} = \int_0^{\pi} \frac{2\pi-\theta}{2\pi} \cdot \frac{d\theta}{2\pi} + \int_{\pi}^{2\pi} \frac{\theta}{2\pi} \cdot \frac{d\theta}{2\pi} = \frac{3}{4}.$$

Example 7·18. Bertrand Problem. *If a chord is taken at random in a circle what is the chance that its length is not less than the radius of the circle ?*

Sol. 1. If we *assume* that all distances of the chord from the centre of the circle are equally likely, then for all the distances less than $\sqrt{r^2-\left(\frac{r}{2}\right)^2}$, *i.e.*, $\frac{\sqrt{3}}{2}r$, the chord is greater than r where r is the radius of the circle.

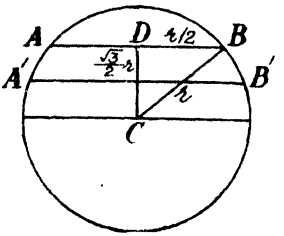


Fig. 22.

Hence the required probability

$$= \frac{\sqrt{3}}{2}\, r/r = \frac{\sqrt{3}}{2} = \cdot 866 \text{ approx.}$$

Sol. 2. If we *assume* that all angles between the chord and a diameter are equally likely, then for all values of angle $\theta < \pi/3$, the chord is greater than r where θ is the angle between the chord and a diameter.

Hence the required probability

$$= \frac{2\pi/3}{2\pi/2} = \frac{2}{3} = \cdot 666 \text{ approx}$$

Note. We get two apparently contradictory results, since we have two different hypotheses. In the first solution it was assumed that the distance of the chord from the centre has uniform distribution while in the second solution the distribution of the angle θ was taken as uniform.

Example 7·19. *Find the mean distance of two points taken at random on the circumference of a circle.*

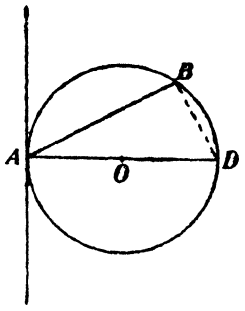


Fig. 23.

Let A be any point and AOD be the diameter. Also let B be the second random point and AB make an angle θ with AOD. Then $\angle ABD$ is a right angle and $AB = AD \cos\theta = 2r\cos\theta$, r being the radius of the circle and $-\pi/2 \leqslant \theta \leqslant \pi/2$.

Mean distance of two points

$$= \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} 2r\cos\theta\, d\theta = \frac{4r}{\pi}.$$

Example 7·20. Buffon's Needle Problem. *A smooth table is ruled with parallel lines at distance a apart. A needle of length $l<a$ is dropped on the table. What is the probability that it will cross one of the lines ?* (*M.A., Bombay, 1953 ; M A., Banaras, 1943*)

Take one of the parallel lines for x-axis and any line perpendicular to it for y-axis. The probability that the centre of the needle has an ordinate lying between the limits y and $y+dy$ is dy/a ; and the probability that the inclination of the needle to OY should be between θ and $\theta+d\theta$ is $d\theta/\pi$. Hence the probability that the needle will cross OX is

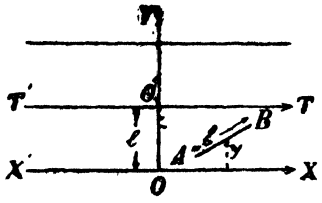


Fig. 24.

$$\int \int \frac{dy\, d\theta}{a\pi}$$

where the double integral is taken over the range of value of y and θ for which the needle will cross OX. The possible values of y are clearly given by $|\, y \,| \leqslant l/2\cos\theta$ and θ lies in the range $-\pi/2 \leqslant \theta \leqslant \pi/2$.

$$\text{Hence} \qquad \text{probability} = 2\int_{-\pi/2}^{\pi/2} \int_{0}^{l/2\cos\theta} \frac{dy\, d\theta}{2\pi} = \frac{2l}{\pi}$$

Ex. 1. A line of given length a is divided into three parts. What is the probability that these form the sides of the triangle ? [**Ans.** ¼]

[**Hint.** Let AB be the straight line of length a and let the three parts be of lengths x, y and $a-(x+y)$. Then we require

$$x+y>a-(x+y)$$
$$x+a-(x+y)>y$$
$$y+a-(x+y)>x.$$

These conditions are equivalent to

$$x+y>\frac{a}{2},\quad x<\frac{a}{2},\quad y<\frac{a}{2},$$

so that the point x, y lies in the area WVS]

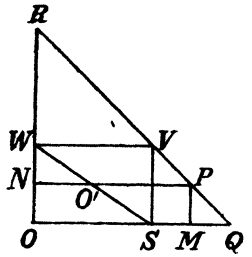


Fig. 25.

Ex. 2. A st. line of length a is divided at random into to parts. Find the mean value of the ractangle contained by the two parts [**Ans.** $a^2/6$]

7·4. Chebychev's Lemma. Law of Large Numbers.

Chebychev's Lemma. *If X is a non-negative stochastic variable with mean $\bar{x}$, then $P(x<\bar{x}t^2) \geqslant 1-\frac{1}{t^2}$, t being a real parameter.*

We have $$\bar{x} = \int_0^\infty xf(x)dx = \int_0^{\bar{x}t^2} xf(x)dx + \int_{\bar{x}t^2}^\infty xf(x)\,dx$$

or $$\bar{x} \geqslant \int_{\bar{x}t^2}^{\infty} xf(x)dx \geqslant \bar{x}t^2 \int_{\bar{x}t^2}^{\infty} f(x)dx = \bar{x}t^2 P(x \geqslant \bar{x}t^2)$$

$$\therefore \quad \frac{1}{t^2} \geqslant P(x \geqslant \bar{x}t^2) = 1 - P(x < \bar{x}t^2).$$

Hence $$P(x < \bar{x}t^2) \geqslant 1 - \frac{1}{t^2}$$

Cor. 1. Let $X = [y - \bar{y}]^2$, so that $\bar{x} = E(x) = Var\ y = \sigma_y^2$. Hence the above lemma gives

$$P(\mid y - \bar{y} \mid < t\sigma_y) \geqslant 1 - \frac{1}{t^2}$$

or $$P(\mid y - \bar{y} \mid \geqslant t\sigma_y) \leqslant \frac{1}{t^2}$$

This inequality is known as *Chebychev's Inequality.*

Also $P(\mid y - \bar{y} \mid \geqslant k) \leqslant \dfrac{\text{var } y}{k^2}$ by putting $t\sigma_y = k$.

Alternative Proof : Theorem. *If y is a continuous random variable with mean $\bar{y}$ and variance σ^2 then for any positive real parameter t, we have*

$$P(\mid y - \bar{y} \mid \geqslant t\sigma_y) \leqslant \frac{1}{t^2}$$

Proof. We have

$$\sigma_y^2 = E[Y - \bar{Y}]^2 = \int_{-\infty}^{+\infty} (y - \bar{y})^2 f(y)dy$$

$$= \int_{-\infty}^{\bar{y} - t\sigma_y} (y - \bar{y})^2 f(y)dy + \int_{\bar{y} - t\sigma_y}^{\bar{y} + t\sigma_y} (y - \bar{y})^2 f(y)dy + \int_{\bar{y} + t\sigma_y}^{\infty} (y - \bar{y})^2 f(y)dy$$

$$\geqslant \int_{-\infty}^{\bar{y} - t\sigma_y} (y - \bar{y})^2 f(y)dy + \int_{\bar{y} + t\sigma_y}^{\infty} (y - \bar{y})^2 f(y)dy$$

$$\geqslant k^2\sigma_y^2 \int_{-\infty}^{\bar{y} - t\sigma_y} f(y)dy + k^2\sigma_y^2 \int_{\bar{y} + t\sigma_y}^{\infty} f(y)dy$$

Since for the first integral $y \leqslant \bar{y} t\sigma_y \Rightarrow (y-\bar{y})^2 \geqslant t^2\sigma_y^2$ and for the second integral $y \geqslant \bar{y} + t\sigma_y \Rightarrow (y-\bar{y})^2 \geqslant t^2\sigma^2{}_y$

Thus $$\sigma_y^2 \geqslant t^2\sigma_y^2[P(y-\bar{y} \leqslant -t\sigma_y) + P(y-\bar{y} \geqslant t\sigma_y)]$$

$$= t^2\sigma_y^2 \; P[\,|\, y-\bar{y} \,|\, \geqslant t\sigma_y]$$

$$\Rightarrow \qquad P[\,|\, y-\bar{y} \,|\, \geqslant t\sigma_y] \leqslant \frac{1}{t^2}$$

Note. This inequality fixes an upper limit to the probability that the value of the variate chosen at random will differ from the mean by more than t times the standard deviation.

Remark 2. It can be proved that

If the variables $x_1, x_2, \ldots\ldots, x_n$ are uniformly bounded then the condition

$$\lim_{n \to 0} \frac{B_n}{n^2} = 0$$

is necessary as well as sufficient.

Example 7·21. *Use Tchebycheff's inequality to find how many times a true coin must be tossed in order that the probability will be at least 0·90 that X, the proportion of the number of heads will be between 0·4 and 0·6* (*B.Sc. Kerala, 1968*)

By Bernoulli's Law of Large Numbers, we have

$$P\left[\left|\frac{X}{n} - p\right| \leqslant t\right] 1 - \frac{1}{4nt^2}, \; t > 0 \qquad \ldots(1)$$

Since the coin is fair, we have $p = 0{\cdot}5$, we want

$$P\left[\left|\frac{X}{n} - p\right| \leqslant 0{\cdot}1\right] = 0{\cdot}9$$

Putting $t = 0{\cdot}1$ in (1), we have

$$P\left[\left|\frac{X}{n} - p\right| \leqslant 0{\cdot}1\right] \geqslant 1 - \frac{1}{4n(0{\cdot}1)^2}$$

$$= 1 - \frac{1}{0{\cdot}04n}$$

Hence $$1 - \frac{1}{0{\cdot}04n} = 0{\cdot}9 \Rightarrow n = 250$$

Thus the number of losses of the true coin = 250.

Example 7·22. *A random variable X has the density function* e^{-x} *for* $x \geqslant 0$. *Show that Tchebycheff's inequality gives*

$$P[\,|\, X-1 \,|\, > 2] < \tfrac{1}{4}$$

and show that the actual probability is e^{-3}. (*B.Sc., Bombay, 1968*)

We have the distribution of the random variable X as

$$dP = e^{-x}dx, \; x \geqslant 0$$

$$E(x)=\int_0^\infty xe^{-x}dx=1$$

$$E(x^2)=\int_0^\infty x^2e^{-x}dx=\Gamma 3=2$$

$\therefore$ Var $X=E(X^2)-[E(X)]^2=2-1=1$

Now $$P\left[\left|\frac{X-E(X)}{\sigma_x}\right|\geqslant t\right]\leqslant\frac{1}{t^2}$$

by Tchebycheff's inequality.

$\Rightarrow$ $$P[\,|X-1|\geqslant 2]\leqslant\tfrac{1}{4}$$

Now $$P[\,|X-1|\geqslant 2]$$

$$\geqslant 1-P[-1\leqslant x\leqslant 3],\text{ since } X \text{ cannot be negative}$$

$$=1-P(0<x<3)$$

$$=1-\int_0^3 e^{-x}dx=e^{-3}.$$

Theorem. *If $g(x)$ is a non-negative function of a random variable X, then for every $k\geqslant 0$,*

$$P[g(x)\geqslant k]\leqslant\frac{E[g(x)]}{k}.$$

Proof. Let A denote the set of all points satisfying the relation $g(x)\geqslant k$ and let $\bar{A}$ be the complement of A. Then if $f(x)$ is the probability of k,

$$P[g(x)\geqslant k)=\int_A f(x)dx \text{ and}$$

$$E[g(x)]=\int_A g(x)\,f(x)dx+\int_{\bar{A}} g(x)\,f(x)dx$$

$$\geqslant k\int_A f(x)dx+\int_{\bar{A}} g(x)\,f(x)dx$$

$\Rightarrow$ $$E[g(x)]\geqslant k\int_A f(x)dx=k\;P[g(x)\geqslant k]$$

$\Rightarrow$ $$P[g(x)\geqslant k]\leqslant\frac{E[g(x)]}{k}$$

Cor. 1. Put $g(x)=(x-\mu)^2$, $k=t\sigma_x$ and we get Tchebychev's inequality.

Cor. 2. Put $g(x)=x$ and we get $P(x \geqslant k) \leqslant E(x)/k$ which is **Markoff's inequality.**

Cor. 3. Let $g(x)=|x|$. Then we have

$$P[|x| \geqslant k] \leqslant \frac{E[|x|]}{k}.$$

Example 7·23. *Two good dice whose faces are numbered 1 to 6 are thrown. If X is the sum of the numbers showing up, prove that*

$$P[|X-7| \geqslant 3] \leqslant \frac{35}{54}$$

(B.Sc., Calcutta, 1972)

Values of X and the corresponding probabilities are :

X :	2	3	4	5	6	7	8	9	10	11	12
$P(x)$:	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Then $EX=2.\frac{1}{36}+3.\frac{2}{36}+\ldots\ldots+12.\frac{1}{36}=7$

$$EX^2=2^2\frac{1}{36}+3^2.\frac{2}{36}+\ldots\ldots+12^2\frac{1}{36}=\frac{1974}{36}$$

$$\Rightarrow \quad \text{Var } X=\frac{210}{36}$$

By Chebychev's inequality, we have

$$P[|X-7| \geqslant 3] \leqslant \frac{\sigma^2}{9}=\frac{210}{9\times 36}=\frac{35}{54}$$

Exact probability of $|X-7| \geqslant 3$

$$=\frac{1}{36}+\frac{2}{36}+\frac{3}{36}+\frac{3}{36}+\frac{2}{36}+\frac{1}{36}=\frac{1}{3}$$

Remark. Bound given by Chebychev's inequality is

$$>\frac{35}{54}-\frac{1}{3}=\frac{17}{54}.$$

Cor. 2. Let $y=\frac{y_1+y_2+\ldots+y_n}{n}$, $\bar{y}=\frac{\bar{y}_1+\bar{y}_2+\ldots+\bar{y}_n}{n}$ and $t\sqrt{\frac{B_n}{n^2}}=\varepsilon$ where ε is an arbitrary positive number and

$$B_n=E(u)=(y_1+y_2+\ldots+y_n-\bar{y}_1-\bar{y}_2-\ldots-\bar{y}_n)^2$$

Then from Cor. 1, page 185 we get

$$P\left[\left|\frac{y_1+y_2+\cdots+y_n}{n}=\frac{\bar{y}_1+\bar{y}_2+\cdots+\bar{y}_n}{n}\right|<\varepsilon\right]\geqslant 1-\frac{B_n}{n^2\varepsilon^2}\geqslant 1-\delta$$

provided $\frac{B_n}{n^2\varepsilon^2}<\delta$.

This is known as *Weak Law of Large Numbers.* It may be stated as follows :

With the probability approaching unity or certainty as near as we please we expect that the arithmetic mean of the values actually assumed by n stochastic variates will differ from the mean of their expectations by less than any given number, however small provided the number of variates can be taken sufficiently large and provided the condition $B_n/n^2\to 0$ *as* $n\to\infty$ *is fulfilled.*

Remark. The weak law of large numbers holds good if

(*i*) $E(x_i)$ exists for all i

(*ii*) $B_n=\text{Var}\,(x_1+x_2+\cdots+x_n)$ exists, and

(*iii*) $\frac{B_n}{n^2}\to 0$ as $n\to\infty$.

Ex. 1. Prove Chebychev's inequality for a non-negative discrete stochastic variate. Prove it also for a continuous variate from first principles

Ex. 2. If x is a stochastic variable such that $E(x)=\mu$ and $Var\,(x)=\sigma^2$ then

$$P(\,|\,x-\mu\,|\,<\delta)\geqslant 1-\frac{\sigma^2}{\delta^2}$$

where δ is any assigned quantity.

[**Hint.** $\sigma^2=\int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx\geqslant\int_{-\infty}^{\mu-\delta}(x-\mu)^2 f(x)dx+\int_{\mu+\delta}^{\infty}(x-\mu)^2(x)dx$

$$\geqslant\delta^2\left[\int_{-\infty}^{\mu-\delta}f(x)dx+\int_{\mu+\delta}^{\infty}f(x)dx=\delta^2\,P\,|\,(x-\mu\,|\,>\delta)\right.$$

Ex. 3. For the number of points X on a dice, prove that Chebychev's inequality gives

$$P[\,|\,X-E(X)\,|\,>2{\cdot}5]<{\cdot}47,$$

while the actual probability is nearly zero.

Ex. 4. For the geometric distribution $p_k=1/2^k$ ($k=1, 2, 3\ldots\ldots$), prove that Chebychev's inequality gives

$$P[\,|\,x-2)\,|\,>2]<\tfrac{1}{4}$$

while the actual probability is $\frac{1}{16}$.

Ex. 5. For the distribution

$$dP=y_0e^{-x}x^k\,dx,\ 0\leqslant x<\infty\ .\ y_0 \text{ being constant, } k\geqslant 0,$$

determine $P[\,|\,x-E(x)\,|\,>2\sigma]$

and compare this with the value given by Chebychev's inequality.

(B.Sc., Bombay, 1963)

Ex. 6. Let x_i assume values i and $-i$, each with probability $\frac{1}{2}$. Show that the law of large numbers need not hold as B_n/n^2 does not approach zero.

Ex. 7. If x_i can have only two values i^α and $-i^\alpha$ with equal probabilities show that the law of large number can be applied to $x_1, x_2, \ldots\ldots, x_n$ if $\alpha<\frac{1}{2}$.

(I.C.A.R. Delhi, 1947)

Ex. 8. Prove that if $|\,x_i-E(x_i)\,|>C$ where C is a constant independent of n, then if the law of large number holds, $B_n/n^2\to 0$ as $n\to\infty$.

Ex. 9. If x_n takes the values 1 and 0 with corresponding probabilities p_n and $1-p_n$, examine whether the weak law of large numbers can be applied to the sequence $\{x_n\}$ where x_n are independent.

Ex. 10. $\{x_k\}$, $k=1, 2,$ is a sequence of independent random variables each taking the value $-1, 0, 1$. Given that $P(x_k=1)=\frac{1}{k}=P(x_k=-1)$;

$P(x_k=0)=1-\frac{2}{k}$, examine if the law of large numbers holds for this sequence.

(B.Sc., Mysore, 1968)

7·4·1. Bernoulli's Theorem. *Let m be the number of successes in n independent trials with the constant probability p of success. Then given ε and δ, two positive numbers, however small, we can find a number N depending on ε and δ such that the probability*

$$\left|\frac{m}{n}-p\right|<\varepsilon$$

will hold is greater than $1-\delta$, provided $n>N$.

Proof. Associate with trials, 1, 2, 3,..., n stochastic variates $x_1, x_2, \ldots\ldots, x_n$ as follows :

$x_i=1$ in case of success in the ith trial

$=0$ in case of failure in the ith trial,

so that $m=x_1+x_2+\ldots\ldots+x_n$.

Since trials are independent, these variates are independent.

Also $E(x_i)=p$, $E(x_i^2)=p$, $Var\ (x_i)=pq$, $Var\ (m)=npq$, $q=1-p$

$$E\left(\frac{m}{n}\right)=p,\ Var\left(\frac{m}{n}\right)=\frac{pq}{n}=\sigma^2 \text{ say. Also let } \varepsilon=t\sigma.$$

Applying Chebychev's inequality to the variate m/n, we get

$$E\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)\geqslant 1-\frac{pq}{n\varepsilon^2}$$

$$>1-\delta \text{ when } n>\frac{pq}{n\varepsilon^2}=N.$$

Stochastic Convergence or Convergence in Probability. A variable t_n is said to *converge stochastically* (or in probability sense) to parameter θ if given ε and δ, two positive numbers however small, we can find a number $N(\varepsilon, \delta)$ such that

$$P(|t_n-\theta|<\varepsilon)>1-\delta \text{ when } n>N$$

i.e.,
$$\lim_{n\to\infty} P(|t_n-\theta|\geqslant\varepsilon)=0$$

In other words, t_n is said to approach θ in probability sense if we can find a number N such that when n is greater than N we can make a probability t_n being as near to θ as we like as near to unity as we please. It may be noted that when $n>N$, $|t_n-\theta|$ is not necessarily less than ε though the probability of its being so is very nearly unity. In this sense the concept of stochastic convergence is different from the concept of convergence in Mathematics.

Notation. In symbols, we write as

$$t_n\to\theta \text{ as } n\to\infty.$$

Example 7·24. *A symmetric die is thrown 360 times. Determine the lower bound for the probability of getting 50 to 70 ones.*

Let m be the total number of successes. Then

$$m=X_1+X_2+\ldots\ldots+X_{360}$$

$$\Rightarrow \quad E(m)=np=360\times\tfrac{1}{6}=60$$

$$\text{Var } m=npq=360\times\tfrac{1}{6}\times\tfrac{5}{6}=50$$

Chebychev's inequality gives

$$P[|m-60|<t\sqrt{50}]\geqslant 1-\frac{1}{t^2}$$

$$\Rightarrow \quad P[60-t\sqrt{50}<m<60+t\sqrt{50}]\geqslant 1-\frac{1}{t^2}$$

Putting $t=\sqrt{2}$,

$$P[60-10<m<60+10]\geqslant 1-\tfrac{1}{2}$$

$$\Rightarrow \quad P[50<m<70]\geqslant\tfrac{1}{2}.$$

Bernoulli's theorem may now be stated as follows :

In an infinite series of Bernoullian trials, the relative frequency of success converge stochastically to the value of the probability in each trial as the number of trials increases indefinitely.

The theorem shows how the mathematical and statistical definitions of probability can be reconciled.

Ex. 1. If, in Bernoulli's theorem, Q is the probability of the inequality

$$| m-np | \geqslant n\epsilon,$$

prove that $$Q < \frac{pq}{n\epsilon^2},$$

[**Hint.** In the identity $\sum_{m=0}^{n} T_m(m-np)^2 = npq$ where $T_m = {}^nC_m p^m q^{n-m}$, drop all the terms in which $| m-np | < n\epsilon$ and in the remaining terms, replace $(m-np)^2$ by $n^2\epsilon^2$. The resulting inequality is

$$\sum_{| m-np | \geqslant n\epsilon} T_m > \frac{pq}{n\epsilon^2}.$$

Ex. 2. Prove Poisson's theorem : If m represents the number of successes in n independent trials, the probability of success at the ith trial being p_i then $\frac{m}{n}$ converges in probability of $\frac{1}{n}\sum_{i=1}^{n} p_i$. *(M.A., Delhi, 1956)*

7·5. Continuous Bivariate Probability Distributions. By analogy it is easy to extend the concept of probability distribution of a single variate to that of two or more variates varying simultaneously. Let X and Y be two stochastic variates and x, y be their particular values. The values x and y can be represented by a point with co-ordinates (x, y) in the X-Y plane. The domain S of all the possible values of X and Y will be represented by a portion (finite or infinite) of the x-y plane with a definite boundary unless his domain coincides with the whole plane. The probability that the point (x, y) will lie in the infinitesimal rectangular region, of area $dxdy$, and defined by

$$P[x-\tfrac{1}{2}dx \leqslant X \leqslant x+\tfrac{1}{2}dx,\ y-\tfrac{1}{2}dy \leqslant Y \leqslant y+\tfrac{1}{2}dy]$$

is denoted by $f_{XY}(x, y)\,dxdy$, where the function $f_{XY}(x, y)$ is called the probability density function at the point (x, y) of the joint distribution for x and y. If there is no confusion, we prefer to write $f(x, y)$ instead of $f_{XY}(x, y)$.

The *requirement for the continuous function* $f(x, y)$ to be a joint probability density function for x and y are :

$$\left.\begin{aligned} &(i)\ f(x, y) \geqslant 0 \text{ everywhere in the } x\text{-}y \text{ plane.} \\ &(ii)\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dx\,dy = 1 \\ &(iii)\ P[a \leqslant X \leqslant b,\ c \leqslant Y \leqslant d] = \int_{x=a}^{x=b}\int_{y=c}^{y=d} f(x, y)dxdy \end{aligned}\right\} \ldots(7\cdot15)$$

Marginal and Conditional Distributions and Density Functions. If $P[y-\frac{1}{2}dy \leqslant Y \leqslant y+\frac{1}{2}dy,\ -\infty < X < \infty]$

$$= dy \int_{-\infty}^{\infty} f(x, y)dx$$

$$= h_Y(y)dy. \text{ where } h_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y)dx$$

then $dP = h(y)dy$ is called the *marginal distribution of* Y and $h(y)$ is called the *marginal distribution function of* Y.

Similarly, if $P[x-\frac{1}{2}dx \leqslant X \leqslant x+\frac{1}{2}dx,\ -\infty < Y < \infty]$

$$= dx \int_{-\infty}^{\infty} f(x, y)dy$$

$$= g(x)dx, \text{ where } g_x(x) = \int_{-\infty}^{+\infty} f(x, y)dy$$

then $dP = g(x)dx$ is called the *marginal distribution of* X and $g_x(x)$ is called the marginal density function of X.

The conditional density function of Y, given X, for two random variables X, Y is defined as follows :

$$f(y \mid x) = \frac{f(x, y)}{g(x)}$$

where x and y are two real numbers.

Similarly, the conditional density function of X_1 given Y, is

$$f(x \mid y) = \frac{f(x, y)}{h(y)}.$$

provided that $g(x) > 0$ and $h(y) > 0$.

If the conditional density function $f(x \mid y)$ does not depend on Y and if the range of the conditional distribution does not depend on Y, then X is independent of Y in probability sense. It is easily seen that if X and Y are independent in probability sense, then their joint distribution is equal to the product of their marginal distributions, *i.e.*, $F_{X,\,Y}(x, y) = F_X(x) \times F_Y(y)$. The converse of this is also true, *i.e.*, if $f(x, y)$ can be factorised into two functions $g(x)$ and $h(y)$, *i.e.*, $f(x, y) = g(x) \times h(y)$ and *if the range of* X *and* Y *do not depend on each other*, then X and Y are independent of each other.

Independent Random Variables. Let X and Y be two random variables jointly distributed with $F_{XY}(x, y)$ as their joint distribution function and $f_{XY}(x, y)$ as their probability density function. Further, let $F_X(x)$, $F_Y(y)$ be their marginal distribution functions and $f_x(x)$, $f_y(y)$ be their marginal density functions. The random variables X, Y are said to be independent if

$$F_{XY}(x, y) = F_X(x)\ F_Y(y).$$

Differentiating partially with respect to x and then with respect to y, we have

$$\frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) = \frac{\partial}{\partial x} F_X(x) \frac{\partial}{\partial y} F_Y(y)$$

$$\Rightarrow \quad f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$$

$$\Rightarrow \quad f(x, y) = g(x)\ h(y).$$

Theorem. *A necessary and sufficient condition for the independence of two random variables X and Y is that their joint density function should be expressible as the product of a function of X alone and a function of Y alone.*

Proof. The condition is clearly necessary.

To establish the sufficiency, we have to prove

$$F_{XY}(x, y) = F_X(x)\ F_Y(y)$$

$$\Rightarrow \quad F_{XY}(\infty, \infty) = F_X(\infty)\ F_Y(\infty) = 1$$

Hence $$F_{XY}(x, y) = \frac{F_X(x)}{F_X(\infty)} \cdot \frac{F_Y(y)}{F_Y(\infty)}$$

Also $$F_X(x) = F_{XY}(x, \infty) = \frac{F_X(x)}{F_X(\infty)}$$

and $$F_Y(y) = F_{XY}(\infty, y) = \frac{F_Y(y)}{F_Y(\infty)}$$

Hence $$F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$$

which was to be proved.

Remark 1. If X and Y are independent random variables then $P(X = x_1, Y = y_1) = P(X = x_1)\ P(Y = y_1)$.

2. If X and Y are independent random variables, and

$$A = \{x : a < x < b\},\ B = \{y : c < y < d\}$$

be any two events, then

$$P(A \cap B) = P(A)\ P(B).$$

Proof. We shall assume that X and Y are continuous random variables.

$$P(A \cap B) = P[(x : a < x < b) \cap (y : c < y < d)]$$

$$= \int_a^b \int_c^d f_{XY}(x, y)\, dxdy$$

$$= \int_a^b \int_c^d f_X(x)\, f_Y(y)\, dxdy,$$

since A and B are independent

$$= \left[\int_a^b f_X(x)dx\right]\left[\int_c^d f_Y(y)dy\right]$$

$$= P(A)\, P(B)$$

which proves the result.

The mathematical expectation of any function $\phi(x, y)$ is defined by

$$E[\phi(x, y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x, y)\, f(x, y) dxdy. \qquad \ldots(7{\cdot}16)$$

In particular, the moment μ_{rs}' about the origin (0, 0) of order r in X and s in Y is defined as

$$\mu_{rs}' = E(X^r Y^s)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s f(x, y)\, dxdy. \qquad \ldots(7{\cdot}17)$$

Also $$\mu_{rs} = E[(x-\bar{x})^r (y-\bar{y})^s]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\bar{x})^r (y-\bar{y})^s f(x, y)\, dxdy \qquad \ldots(7{\cdot}18)$$

$$M(t_1, t_2) = E\left[e^{t_1 x + t_2 y}\right]$$

where t_1 and t_2 are real.

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 x + t_2 y} f(x, y)\, dxdy. \qquad \ldots(7{\cdot}19)$$

When this integral is convergent the exponential may be expanded in powers of t_1 and t_2 and

$$\mu_{rs}' = \text{coefficient of } \frac{t_1^r}{r!}\frac{t_2^s}{s!} \text{ in } M(t_1, t_2) \qquad \ldots(7{\cdot}20)$$

$$\mu_{10}' = \iint x f(x, y)\, dx\, dy = \int_{-\infty}^{\infty} x g(x)\, dx.$$

Similarly

$$\mu_{01}' = \iint y f(x, y)\, dxdy = \int_{-\infty}^{\infty} y h(y)\, dy$$

$$\mu_{20} = \iint (x-\bar{x})^2 f(x, y)\, dxdy = \int_{-\infty}^{\infty} (x-\bar{x})^2 g(x) dx = \sigma_x^2$$

$$\mu_{02} = \iint (y-\bar{y})^2 f(x, y) dxdy = \int_{-\infty}^{\infty} (y-\bar{y})^2 h(y) dy = \sigma \bar{y}^2$$

$$\mu_{11} = \iint (x-\bar{x})(y-\bar{y})\, f(x, y)\, dxdy = \text{Cov}\ (x, y).$$

The correlation coefficient r_{xy} between x and y is defined as

$$r_{xy} = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}}. \qquad \ldots(7{\cdot}21)$$

The function $F_{XY}(x, y)$ which is defined by

$$F_{XY}(x, y) = P(-\infty < X \leqslant x,\ -\infty < Y \geqslant y)$$

$$= \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\, dxdy \qquad \ldots(7{\cdot}22)$$

is called the *distribution function* of the two dimensional random variables x and y. Unless there is confusion, we shall prefer to denote $F_{XY}(x, y)$ by $F(x, y)$.

It has the following *properties* :

(*i*) $0 \leqslant F(x, y) \leqslant 1$.

(*ii*) $\dfrac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y)$.

(*iii*) $F(x, y)$ is a non decreasing function in the two dimen sional random variables X and Y.

(*iv*) $F(x, -\infty)$ 0, $F(-\infty, y)-0$, $F(-\infty, \infty)=1$.

(*v*) $P(x_1 \leqslant X \leqslant x_2,\ y_1 \leqslant Y \leqslant y_2) = P(X \leqslant x_2,\ Y \leqslant y_2)$
$-P(X \leqslant x_2,\ Y \leqslant y_1) - P(X \leqslant x_1,\ Y \leqslant y_2) + P(X \leqslant x_1,\ Y \leqslant y_1)$
$= F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$ ...(7·23)

We shall now show by an example that conditions (*i*), (*ii*) and (*iii*) are not *sufficient* to ensure that $F(x, y)$ is a distribution function.

Let $F(x, y)=0$ for $x+y \leqslant 0$
$=1$ for $x+y>0$

i.e., the function $F(x, y)$ takes the value 0 at points on and below the straight line $y=-x$ and the value 1 at points above this line. The function $F(x, y)$ so defined statisfies (*i*), (*ii*), (*iii*) and (*iv*).

However, it does not satisfy (*v*) for if we consider a rectangle in the $x-y$ plane whose three vertices lie below and one above the straight line $y=-x$, *e.g.*, the rectangle with vertices (3, 3) ; (3, −1) ; (−1, 3) ; (−1, −1), then on applying (7·23), we have

$$P(-1 \leqslant X \leqslant 3,\ -1 \leqslant Y \leqslant 3) = F(3, 3) - F(3, -1) - F(-1, 3) + F(-1, -1)$$
$$= 1-1-1+0 = -1,$$

which is impossible since probability cannot be negative.

Example 7·23. *Show that the conditions that the function*

$$f(x, y) = k \exp\{Ax^2 + 2Hxy + By^2\}, \quad -\infty < x, y < \infty$$

may be a density function are

$$(a)\ A \leqslant 0,\ (b)\ B \leqslant 0,\ (c)\ AB - H^2 \geqslant 0.$$

Show further that if these conditions are satisfied and the integral of $f(x, y)$ between $-\infty$ and ∞ for both variates is unity, then

$$k = \frac{1}{\pi}\sqrt{AB - H^2}$$

Clearly $f(x, y)$ will represent a density function if and only if

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k \exp\{Ax^2 + 2Hxy + By^2\}\, dxdy = 1$$

Now $Ax^2+2Hxy+By^2$ can be written in either of the two forms :

$$A\left[\left(x + \frac{H}{A}y\right)^2 + \frac{AB-H^2}{A^2}y^2\right]$$
$$B\left[\left(y + \frac{H}{B}x\right)^2 + \frac{AB-H^2}{B^2}x^2\right]$$

Hence clearly we should have, for the convergence of the integrals,

$$A \leqslant 0,\ B \leqslant 0, \text{ and } AB - H^2 \geqslant 0.$$

Let $A = -a,\ B = -b,\ H = h \quad \therefore \quad AB - H^2 = ab - h^2,\ a > 0,\ b > 0$

Also $$k \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(ax^2+2hxy+by^2)}\, dx\, dy = 1$$

or $$k \int_{-\infty}^{\infty} e^{-\frac{(ab-h^2)y^2}{a}}\, dy \times \int_{-\infty}^{\infty} e^{-\frac{(ax+hy)^2}{a}}\, dx = 1$$

or $$k \sqrt{\frac{\pi a}{ab-h^2}} \times \sqrt{\frac{\pi}{a}} = 1 \qquad \left[\because \int_{-\infty}^{+\infty} e^{-c^2u^2}\, du = \frac{\sqrt{\pi}}{c} \right]$$

or $$k = \frac{1}{\pi} \sqrt{ab-h^2} = \frac{1}{\pi} \begin{vmatrix} -A & H \\ H & -B \end{vmatrix}^{\frac{1}{2}}.$$

Example 7·24. *If* $f(x, y) = 2 - x - y \qquad (0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1)$
$= 0$ *eleswhere,*

find (a) the marginal probability functions (b) the conditional probability functions (c) μ_{20}' *and* μ_{02}' *and (d) the coefficient of correlation between x and y.* (*M.A., Delhi, 1963* ; *M.A., Statistics, Delhi, 1959* ; *M.Sc. Gauhati, 1968*)

We have

(*a*) $$g(x) = \int_0^1 (2 - x - y)\, dy = \frac{3}{2} - x,$$

$$h(y) = \int_0^1 (2 - x - y)\, dx = \frac{3}{2} - y.$$

(*b*) $$f(y \mid x) = \frac{f(x, y)}{g(x)} = \frac{2 - x - y}{\frac{3}{2} - y},$$

$$f(x \mid y) = \frac{f(x, y)}{h(y)} = \frac{2 - x - y}{\frac{3}{2} - y},$$

(*c*) $$\bar{x} = \mu_{10}' = \int_0^1 x \left(\frac{3}{2} - x \right) dx = \frac{5}{12}$$

$$\bar{y} = \mu_{01}' = \int_0^1 y \left(\frac{3}{2} - y \right) dy = \frac{5}{12}$$

Hence $$\mu_{20}=\int_0^1\left(x-\frac{5}{12}\right)^2\left(\frac{3}{2}-x\right)dx=\frac{11}{144}$$

$$\mu_{02}=\int_0^1\left(y-\frac{5}{12}\right)^2\left(\frac{3}{2}-y\right)dy=\frac{11}{144}$$

(*d*) $$\mu_{11}=\int_0^1\int_0^1\left(x-\frac{5}{12}\right)\left(y-\frac{5}{12}\right)(2-x-y)\,dx\,dy$$

or $$\mu_{11}=\int_0^1\left(x-\frac{5}{12}\right)dx\int_0^1\left(y-\frac{1}{12}\right)(2-x-y)\,dy$$

$$=\int_0^1\left(x-\frac{5}{12}\right)\left(\frac{1}{24}-\frac{x}{12}\right)dx=\frac{-1}{144}$$

Hence $$r=\frac{\mu_{11}}{\sigma_x\sigma_y}=\frac{\frac{-1}{144}}{\frac{11}{144}}=-\frac{1}{11}$$

Example 7·25. *Two discrete random variables X and Y have the joint probability density function*

$$p(\lambda, y)=\frac{\lambda^x e^{-\lambda}p^y(1-p)^{x-y}}{y\,!\,(x-y)\,!},$$

$$y=0, 1,\ldots\ldots, x\ ;\ x=0, 1, 2,\ldots\ldots$$

where λ, p *are constants with* $\lambda>0$ *and* $0<p<1$. *Find* (*i*) *the marginal probability density functions of X and Y.*

(*ii*) *The conditional distributions of Y for a given X and of X for a given Y.* *(M.Sc. Gauhati, 1969 ; I.S.S. 1970)*

We have

$$p_X(x)=\sum_{y=0}^{x}p(x, y)=\sum_{y=0}^{x}\frac{\lambda^x\ e^{-\lambda}p^y(1-p)^{x-y}}{y\,!\,(x-y)\,!}$$

$$=\frac{\lambda^x e^{-\lambda}}{x\,!}\sum_{y=0}^{x}x!\,\frac{p^y\,(1-p)^{x-y}}{y\,!\,(x-y)\,!}$$

$$= \frac{\lambda^x e^{-\lambda}}{x\,!}(1-p+p)^x$$

$$= \frac{\lambda^x e^{-\lambda}}{x\,!}$$

Similarly $p_y(y)$ can be obtained.

$$(ii)\ p_{Y \mid x}(y \mid x) = \frac{p_{xy}(x, y)}{p_x(x)}$$

$$= \frac{[\lambda^x e^{-\lambda} p^y (1-p)^{x-y}]}{y\,!\,(x-y)\,!} \Big/ \left(\frac{\lambda^x e^{-\lambda}}{x\,!} \right)$$

$$= \binom{x}{y} p^y (1-p)^{x-y}$$

Similarly conditional probability distribution of X given Y can be obtained.

Ex. 1. The joint density function of a bivariate distribution is given as follows :

$$f(x, y) = x+y \qquad 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1$$

$$= 0, \text{ otherwise}$$

Determine the marginal distributions and the covariance between x anp y. *(M.A , Punjab, 1963)*

[**Ans.** $g(x) = x + \frac{1}{2}$; $h(y) = y + \frac{1}{2}$]

Ex. 2. Let X and Y have the joint density function

$$f(x, y) = 2, \qquad 0 < x < y < 1$$

Find the marginal and conditional probability density functions.

Ans. $g(x) = 2(1-x),\ 0 < x < 1$ | $f(x/y) = \frac{1}{y},\ 0 < y < 1$

$= 0$, elsewhere | $= 0$, elsewhere

$h(y) = 2y,\ 0 < y < 1$ | $f(y/x) = \frac{1}{1-x},\ 0 < x < 1$

$= 0$, elsewhere | $= 0$, elsewhere.

Ex. 3. If X and Y are two random variables, having joint density functions

$$f(x, y) = \tfrac{1}{8}(6-x-y), \qquad 0 < x < 2,\ 2 < y < 4$$

$$= 0, \qquad \text{otherwise}$$

Find *(a)* $P(X<1, Y<3)$ *(b)* $P(X+Y<3)$ and *(c)* $P(X<1 \mid Y<3)$. *(B.Sc. Bombay, 1969)*

Hint. *(b)* $P(X+Y<3) = \int_0^1 \int_2^{3-x} \tfrac{1}{8}(6-x-y)\, dxdy = \text{etc.}$

(c) $P(X<1 \mid Y<3)=\dfrac{P(X<1 \cap Y<3)}{P(Y<3)}$

and $P(Y<3)=\int_0^2 \int_2^3 \dfrac{6-x-y}{8}\, dxdy=\dfrac{5}{8}$.

[Ans. (a) $\frac{3}{8}$ (b) $\frac{5}{24}$ (c) $\frac{3}{5}$]

7·6. Variate Transformations.

7·6·1. Univariate Distributions

Theorem. *If $x=h(y)$ is an increasing or decreasing function of y and if $f(x)$ is the density function of x then $g(y)$, the density function of y, is given by*

$$g(y)=f[h(y)] \mid h'(y) \mid$$

Proof. Let the variable x be continuous and have $f(x)$ as its probability function. Suppose $y=H(x)$ is new variable whose probability function is required. We shall assume that the function $H(x)$ is either an increasing or decreasing function throughout the range of x. The relation $y=H(x)$ can be written as

$$x=H^{-1}(y)=h(y) \text{ say.}$$

For $H(x)$ an increasing function, the variable y will assume a value less than y_0 where $y_0=H(x_0)$ if and only if x assumes a value less than x_0 ; hence the probability that $y<y_0$ must be equal to the probability that $x<x_0$.

Thus, we have

$$\int_{-\infty}^{y_0} g(t)\, dt=\int_{-\infty}^{x_0} f(t)\, dt$$

where $g(y)$ is the density function of y.

Since this result holds for every pair of corresponding values (x_0, y_0) we have, on dropping the subscript,

$$\int_{-\infty}^{y} g(t)\, dt=\int_{-\infty}^{x} f(t)\, dt. \qquad \ldots(7{\cdot}25)$$

Differentiating both sides of (7·25) with respect to y and using the fundamental theorem of Integral Calculus, we have

$$g(y)=f[h(y)]h'(y)$$

For $H(x)$ a decreasing function we shall obtain similarly

$$g(y) = -f[h(y)]h'(y)$$

Thus whether $H(x)$ is an increasing or decreasing function of x, we get

$$g(y) = f[h(y)] \mid h'(y) \mid .$$

Example 7·26. *(a) Given* $f(x) = e^{-x}$, $x \geqslant 0$, *find the probability density function of the variable* $Y = X^2$.

(b) If X *has a uniform distribution in* $(0, 1)$ *with* $f(x) = 1$, $0 \leqslant x \leqslant 1$; 0, *otherwise, find the density function of* $Y = -2 \log X$.

Since $y = x^2$ is an increasing function of x when $x > 0$ the above theorem can be applied.

Hence $\quad h(y) = \sqrt{y} \qquad \therefore\ h'(y) = \dfrac{1}{2\sqrt{y}}$

$$\therefore \qquad g(y) = e^{-\sqrt{y}} \frac{1}{2\sqrt{y}},\ y \geqslant 0$$

$$= 0, \text{ otherwise.}$$

(c) $G_y(y) = P(Y \leqslant y) = P(-2 \log X \leqslant y) = P(\log X \geqslant -\tfrac{1}{2}y)$

$$= P(X \geqslant e^{-\frac{1}{2}y}) = 1 - P(X \leqslant e^{-\frac{1}{2}y})$$

$$= 1 - \int_0^{e^{-\frac{1}{2}y}} 1 \, . \, dx = 1 - e^{-\frac{1}{2}y}$$

$$\Rightarrow \qquad g(y) = \frac{d}{dy} G(y) = \tfrac{1}{2} e^{-y/2}\ ,\ 0 < y < \infty.$$

Example 7·27. *If* X *is a random variate having a continuous probability density function and* $F(x)$ *denotes the probability that* X *is less than or equal to* x *then show that* $F(x)$ *is a random variate having a rectangular distribution in the range* $(0, 1)$.

(M.A. Delhi, 1955)

We have $P(X \leqslant x) = F(x) = \displaystyle\int_{-\infty}^{x} f(x)\, dx$

Then $\qquad dP = dF = f(x)\, dx$ and $F'(x) = f(x)$

Let $u = F(x)$ then

$$dP = f(x)\, dx = f(x) \frac{dx}{du}\, du = f(x) \, . \, \frac{1}{f(x)}\, du = du,\ 0 \leqslant u \leqslant 1$$

Hence u follows a rectangular distribution with unit range.

Note. The example shows that by letting $u=F(x)$ where $F(x)$ is the cumulative distribution of x, any density for a continuous variate x may be transformed to the uniform density

$$f(u)=1,\ 0\leqslant u\leqslant 1.$$

The transformation $u=F(x)$ is called the *probability transformation.*

Example 7·28. *What transformation will change the variate x to one which will have the uniform distribution over the unit interval if*

$$f(x)=\frac{x-1}{2},\ 1\leqslant x\leqslant 3\ ?$$

$$=0,\ x<1 \text{ and } x>3\ ?$$

We have $$u=F(x)=\int_1^x \frac{x-1}{2}\,dx=\frac{1}{4}(x-1)^2$$

which is the required transformation.

7·6·2. Bivariate Transformation. The transformations of the variates x and y for bivariate distributions follow the ordinary laws for the transformation of differentials.

For example, if $dF=f(x, y)dx\,dy$

$$x=x(u, v),$$

$$y=y(u, v),$$

we have $$dF=f[x(u, v), y(u, v)]\ |\ J\ |\ du\,dv$$

where J is the Jacobian given by

$$J=\frac{\partial(x, y)}{\partial(u, v)}=\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix}$$

If $$F(x)=\int_c^u f(t)\,dt$$

where c is a constant and u is a function of x, then

$$\frac{d}{dx}F(x)=f(u)\ \frac{du}{dx}.$$

Example 7·29. *If the probability density function $f(x, y)$ of two variates X and Y is given by*

$$f(x, y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\cdot exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2}-\frac{2\rho xy}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}\right)\right\}$$

show that the variates u and v defined as follows are independent variates :

(*i*) $\quad u=\dfrac{x}{\sigma_x}-\dfrac{\rho y}{\sigma_y}, \qquad v=(1-\rho^2)^{\frac{1}{2}}\ \dfrac{y}{\sigma_y}.$

(*ii*) $\quad u=\dfrac{1}{\sqrt{1-\rho^2}}\left(\dfrac{y}{\sigma_y}-\dfrac{\rho x}{\sigma_x}\right), \quad v=\dfrac{x}{\sigma_x}.$

(*B.A. Hons. Calcutta, 1968*)

(*i*) Here $\quad 1/J=\begin{vmatrix} \dfrac{1}{\sigma_x} & \dfrac{-\rho}{\sigma_y} \\ 0 & \dfrac{(1-\rho^2)^{\frac{1}{2}}}{\sigma_y} \end{vmatrix}=\dfrac{(1-\rho^2)^{\frac{1}{2}}}{\sigma_x\sigma_y}$

and $$\frac{x^2}{\sigma_x^2}-\frac{2\rho xy}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}=u^2+v^2.$$

The distribution then becomes

$$dF=\left[\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\Big/\frac{\sqrt{1-\rho^2}}{\sigma_x\sigma_y}\right] exp\left\{\frac{-1}{2(1-\rho^2)}(u^2+v^2)\right\} dudv$$

$$=\left[\frac{1}{\sqrt{2\pi(1-\rho^2)}}\ exp\left\{\frac{-u^2}{2(1-\rho)^2}\right\} du\right]$$

$$\times\left[\frac{1}{\sqrt{2\pi(1-\rho^2)}}\ exp\left\{\frac{-v^2}{2(1-\rho^2)}\right\} dv\right]$$

$$=f(u)\ du\ f(v)\ dv, \qquad -\infty<u<\infty,\ -\infty<v<\infty.$$

Thus the transformed variates u and v are independent.

(*ii*) Proceed as in (*i*).

Example 7·30 *Prove that if the joint density function for the random variables X and Y is $f(x, y)$, then the density function of the sum $Z=X+Y$ is given by*

$$\int_{-\infty}^{+\infty} f(x, z-x)dx$$

In the particular case, when

$$f(x, y)=\frac{1}{4a^2}\left[1+xy\left(x^2-y^2\right)\right], \quad |x|, |y| \leqslant a, \ a>0$$

$$=0, \qquad \text{, elsewhere}$$

show that the distribution of Z is a triangular distribution.

The joint distribution of X and Y is given by

$$dP=f(x, y)\,dx\,dy, -\infty<x, y<\infty.$$

Put $x=x, z-x=y$

then $dP=f(x, z-x)dx\,dz,$

Integrate out for x, then the density function of the random variable Z is given by

$$\int_{-\infty}^{+\infty} f(x, z-x)dx$$

Now $$f(x, y)=\frac{1}{4a^2}[1+xy(x^2-y^2)]$$

$$\therefore \quad f(x, z-x)=\frac{1}{4a^2}[1+x(z-x)\{x^2-(z-x)^2\}]$$

$$=\frac{1}{4a^2}[1+3z^2x^2-2zx^3-z^3x]$$

Hence the distribution of Z is given by

$$\theta(z)=\int \frac{1}{4a^2}(1+3z^2x^2-2zx^3-z^3x)dx. \qquad \ldots(1)$$

In order to find the limits of integration which depend on z we note that in introducing the variables (x, z) instead of (x, y) ; the

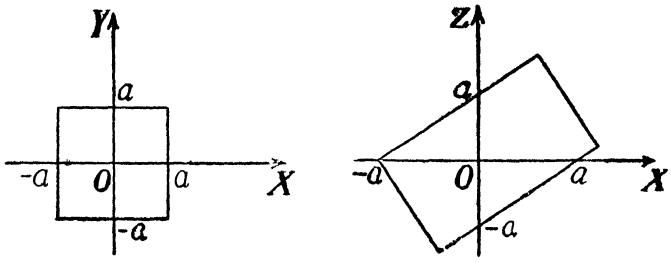


Fig 24.

square domain $|x| \leqslant a, |y| \leqslant a$ changes into a domain defined by the inequalities :

$$\begin{array}{lll} |x| \leqslant a, & x-a \leqslant z \leqslant x+a \quad \text{or} & z-a \leqslant x \leqslant z+a. \\ \text{Thus} \quad z-a \leqslant -a, & z+a \leqslant a & \text{for } z \leqslant 0 \\ z-a > -a, & z+a > -a. & \text{for } z > 0. \end{array}$$

Hence integrating (1) from $-a$ to $z+a$ for $x \leqslant 0$ and $z-a$ to a for $z>0$.

$$\theta(z) = \begin{cases} \displaystyle\int_{-a}^{z+a} \frac{1}{4a^2}(1+3z^2x^2-2zx^3-z^3x)\,dx = \frac{2a+z}{4a^2}, & -2a \leqslant z \leqslant 0 \\ \displaystyle\int_{z-a}^{a} \frac{1}{4a^2}(1+3z^2x^2-2zx^3-z^3x)\,dx = \frac{2a-z}{4a^2}, & 0 < z \geqslant 2a \\ 0 & |z| > 2a \end{cases}$$

Thus the distribution of Z is triangular.

Ex. 7 31. *Prove the following result for the example 7·30, page 204.*

$$\text{C.F. for } X = \frac{\sin at}{at}$$

$$\text{C.F. for } Y = \frac{\sin at}{at}$$

$$\text{C.F. for } Z = \left(\frac{\sin at}{at}\right)^2.$$

Hence show that

$$\varphi_z(t) = \varphi_x(t)\varphi_y(t)$$

though the variables X and Y are not independent. (*M.A. Lucknow 1963*)

we have $g(x) = \displaystyle\int_{-a}^{+a} \frac{1}{4a^2}\left[\,[1+xy(x^2-y^2)\,\right]dx = \frac{1}{4a^2}\left[\, y + \frac{x^3y^2}{2} - \frac{xy^4}{4}\,\right]_{-a}^{+a} = \frac{1}{2a},$

$$h(y) = \int_{-a}^{+a} \frac{1}{4a^2}\left[\, 1+xy(x^2-y^2)\,\right]dx = \frac{1}{4a^2}\left[\, x + \frac{x^4y}{4} - \frac{x^2y^3}{2}\,\right]_{-a}^{+a} = \frac{1}{2a}$$

$\therefore$ $g(x)h(y) = \dfrac{1}{4a^2} \neq f(x, y)$, so that variables X and Y are not independent.

$$\phi_x(t) = \int g(x)e^{itx}dx = \frac{1}{2a}\int_{-a}^{+a} e^{itx}dx = \frac{e^{iat}-e^{-iat}}{2iat} = \frac{\sin at}{at}$$

Similarly, $\phi_y(t) = \dfrac{\sin at}{at}$.

Also $$\varphi_z(t)=\int_{-2a}^{0}\frac{2a+z}{4a^2}e^{itz}dz=\int_{0}^{2a}\frac{2a-z}{4a^2}e^{itz}dz$$

$$=\frac{2-e^{2iat}-e^{-2iat}}{4a^2t^2}=\frac{1}{2a^2t^2}\quad(1-\cos 2at)=\left(\frac{\sin at}{at}\right)^2$$

$\therefore$, $\varphi_z(t)=\varphi_x(t)\varphi_y(t)$.

EXERCISE 7

1. (*a*) Explain the terms (*i*) probability differential, (*ii*) probability density function, and (*iii*) distribution function.

(*B.Sc. Bombay, 1967, 68* ; *B.Sc. Hons. Delhi, 1967* ; *M.A. Delhi, 1960*)

(*b*) If X is a random variable, then $|E(x)| \leqslant E(|x|)$

(*c*) If X and Y are two random variables such that $x \leqslant y$, then $E(X) \leqslant E(Y)$

(*d*) If $g(x) \geqslant 0$ everywhere then $E[g(x)] \geqslant 0$.

(*e*) Show that whatever the distribution function $F(x)$ of a random variable X, $P[a \leqslant F(a) \leqslant (b)].=b-a$, $0<a$, $b<1$.

2. (*a*) If $f(x)=2x$ when $0 \leqslant x \leqslant 1$ and zero otherwise, find the probability that (*i*) $x<\frac{1}{2}$ (*ii*) $\frac{1}{4}<x<\frac{1}{2}$ (*iii*) $x>\frac{3}{4}$ given $x<\frac{1}{2}$.

(*b*) If $f(x)=x \qquad 0 \leqslant x<1$

$\quad=2-x, \qquad 1 \leqslant x<2$

find $P\frac{1}{2}(\leqslant x \leqslant \frac{3}{2}.)$

(*c*) A random variable x takes values 0, 1, 2,...with probability proportional to $(x+1)(\frac{1}{5})^x$. Find the probability that $x \leqslant 5$. (*B.A. Hons., Delhi, 1966*)

[**Hint.** Use $\int_0^\infty k(x+1)\left(\frac{1}{5}\right)^x dx=1$]

[**Ans.** (*a*) $\frac{1}{4}$, $\frac{3}{16}$, $\frac{7}{12}$, (*b*) $\frac{3}{4}$]

3. If $f(x)=3x^2$ when $0 \leqslant x \leqslant 1$ and zero otherwise, find the number a such that x is equally likely to be greater than or less than a. Find the number b such that the probability that x will exceed b is equal to ·05.

Ans. (*i*) $a=(\frac{1}{2})^{\frac{1}{3}}$, (*ii*) $(b)=(0{\cdot}95)^{\frac{1}{3}}$]

4. If $f(x)=1$, $0 \leqslant x \leqslant 1$ (*a*) what is the probability that at least two out of the three sample values will exceed ·6 ? (*b*) What value of x is such that the probability is $\frac{1}{2}$ that at least two out of 3 sample values will exceed it ?

[**Ans.** (*a*)=·352, (*b*) $x=\frac{1}{2}$]

5. (*a*) Suppose the life in hours of a certain kind of radio tube has the density $f(x)=\frac{100}{x^2}$ when $x \geqslant 100$ and zero when $x<100$. What is the probability that none of three such tubes in a given radio set will have to be replaced during the first 150 hours of operation ? What is the probability that all three of the original tubes will have been replaced during the first 150 hours ?

(*M.A. Bombay, 1958* ; *M.A. Punjab, 1958* ; *M.A. Delhi, 1960*)

(*b*) A bomb plane carrying three bombs flies directly above a rail road track. If a bomb falls within 40 feet of the track, the track will be sufficiently damaged to disrupt traffic. With a certain bomb sight, the density of points of impact of a bomb is

$$f(x)=\frac{100+x}{10,000} \qquad -100\leqslant x<0$$

$$=\frac{100-x}{10,000} \qquad 0\leqslant x\leqslant 100$$

$$=0 \qquad \text{eleswhere}$$

If all three bombs are used what is the probability that the track will be damaged where x represents the vertical deviation from the aiming point which is the track in this case. (*B.Sc. Bombay, 1970* ; *B.A. Hons., Delhi, 1957*)

(*c*) A country filling station is supplied with gasoline once a week. If its weekly volume x of sales in thousands of gallons is distributed by $f(x)=5\,(1-x)^4$, $0<x<1$, what must be the capacity of its tank in order that the probability that its supply will be exhausted in a given week shall be ·01 ?

[**Ans.** (*a*) $\frac{1}{27}$, $\frac{8}{27}$ (*b*) Required Probability $=1-\left(\frac{9}{25}\right)^3$

(*c*) Volume $=1000[1-90)^{\frac{1}{5}}$]

6. The probability of a shot hitting at a point distant r from the centre C of the target follows the law :

$$p(r)=\frac{2}{\pi(1+r^2)}.$$

The target is divided into four regions as follows :

S_1 : Within the circle with centre C and radius $\frac{1}{\sqrt{3}}$.

S_2 : Without S_1 but within the concentric circle of radius 1

S_3 : Without S_2 but within the concentric circle of radius $\sqrt{3}$

S_4 : Without S_3.

The scores for hits in the regions S_1, S_2, S_3 and S_4 are 4, 3, 2 and 0 respectively. Let X_1, X_2, X_3, and X_4 be the number of hits registered in the region S_1, S_2, S_3, and S_4 out of a total of N shots. Find

(*i*) the probability distribution of X_1, X_2, X_3, X_4

(*ii*) the expected score after 5 shots, and

(*iii*) the probability that the score exceeds 15.

(*M.Sc. Poona, 1965* ; *B.Sc. Hons. Cal., 1967*)

[**Hint.** Let Y denote the score of a hit. Then

$$(i)\ P(X=X_1)=P(Y=4)=\int_0^{1\sqrt{3}} p(r)dr=\tfrac{1}{8},\ P(X=X_2)=P(Y=3)=\tfrac{1}{8}$$

$$P(X=X_3)=P(Y=2)=\frac{1}{6},\quad P(X=X_4)=P(Y=0)$$

$$=1-\left(\frac{1}{3}+\frac{1}{6}+\frac{1}{6}\right)=\frac{1}{3}.$$

$$E(Y)=0.\ \frac{1}{3}+2.\ \frac{1}{6}+3.\ \frac{1}{6}+4.\ \frac{1}{3}=\frac{13}{6}$$

(*ii*) $E(Y_1+Y_2+Y_3+Y_4+Y_5)=\sum_{i=1}^{5}\frac{13}{6}=\frac{65}{6}$

(*iii*) Reqd. probability $=P(\sum_{i=1}^{5} Y_i>15)=P(\Sigma Y_i=16)+P(\Sigma Y_i=17)$

$$+P(\Sigma Y_i=18)+P(\Sigma Y_i=19)+P(\Sigma Y_i=20)=\text{etc.}$$

7. The probability density function of coded measurements of pitch diameter of a fitting is given by

$$f(x)=\frac{1}{(1+x)^2},\ 0\leqslant x<\infty$$

Find the distribution function of the distribution. Hence obtain (*i*) $P(x>2)$ (*ii*) median and the quartiles of the distribution. Investigate whether the mean and the variance of the distribution exist. (*M.A. Delhi, 1962*)

8. Find the mode and the median for the distribution

$$dF=\sin x\,dx,\ 0\leqslant x\leqslant \pi/2.$$

[Ans. Modal value $=\pi/2$, Median $=\pi/3$]

9. Find the Arithmetic Mean and the Median for the distribution

$$dF=2x\,dx,\ 0\leqslant x\leqslant 1.$$

$$\left[\text{Ans. Arithmetic Mean}=\frac{2}{3},\quad \text{Median}=\frac{1}{\sqrt{2}}\right]$$

10. Prove that the Geometric Mean, G, of the distribution

$$dF=6(2-x)(x-1)dx,\ 1\leqslant x\leqslant 2$$

is given by $6\log(16G)=19$. [*B.A. Hons. Delhi, 1955*]

11. For the continuous distribution

$$dF=y_0x(2-x)dx,\ 0\leqslant x\leqslant 2,\ y_0 \text{ being a constant}$$

find the mean, mean deviation about mean, μ_2', μ_3', μ_4', μ_2, μ_3, μ_4 and also skewness. Also show that for this distribution $\mu_{2n+1}=0$. Deduce β_2.
(*B.Sc. Delhi, 1971, B.A. Punjab, 1961 ; B.A. Hons., Delhi, 1959, 58, 60*)

$$\left[\text{Ans. } y_0=\frac{3}{4},\ \text{Mean}=1,\quad \mu_2'=\frac{6}{5},\ \mu_2=\frac{1}{5},\quad \mu_3'=\frac{3}{5},\ \mu_3=0.\right.$$

$$\left.\text{skewness}=0,\ \text{Mean deviation}=\frac{3}{8}\right]$$

12. (*a*) For the continuous distribution

$$dF=y_0(x-x^2)dx,\ 0\leqslant x\leqslant 1,\ y_0 \text{ being constant.}$$

find the Arithmetic Mean, the Harmonic Mean, the Mode and the Median.
(*B.Sc. Delhi, 1971, B.A. Hons., Delhi, 1949 ; B.Sc. Hons. Sub. Statistics, Delhi, 1965*)

Compute $P(\mu-2\sigma<x<\mu+2\sigma)$. *(B.Sc. Sardar Patel, 1969)*

(b) Find μ_2, μ_3 and μ_4 for the distribution given by

$$dF=\frac{dx}{2a}, \quad -a\leqslant x\leqslant a. \qquad \textit{(B.Sc. Agra, 1959)}$$

[Ans. *(a)* $y_0=6$, $A.M.=\frac{1}{2}$, $H.M.=\frac{1}{3}$, Modal value$=\frac{1}{2}$, Median$=\frac{1}{2}$

(b) $\mu_2=\frac{1}{3}a^2$, $\mu_3=0$, $\mu_4=\frac{1}{5}a^4$]

13. *(a)* Prove that for the distribution

$$dF=\frac{1}{B(m,\,n)}\cdot\frac{x^{m-1}}{(1+x)^{m+n}}dx,\ 0\leqslant x\leqslant\infty,\ n>2$$

variance is $\dfrac{m(m+n-1)}{(n-1)^2(n-2)}$.

Find also the mode and moment of rth order about the origin. Show also that the harmonic mean is $(m-1)/n$. *(B.A. Hons. Delhi, 1954)*

(b) Calculate the coefficient of variation for the rectangular distribution in $(0, b)$, given that the probability law for the rectangular distribution is

$$P(x\leqslant t)=\frac{1}{b}. \qquad \textit{(B.A. Madras, 1954)}$$

[Ans. *(a)* Mode is at $x=\dfrac{m-1}{n+1}$, $m>1$, $\mu_r'=\dfrac{m(m+1)\ldots(m+r-1)}{(n-1)(n-2)\ldots(n-r)}$, $r<n$.

(b) $100\times\dfrac{1}{\sqrt{3}}$]

14. Show that for the continuous distribution

$$dF=\frac{1}{B(m,\,n)}(1-x)^{m-1}x^{n-1}\,dx,\ 0\leqslant x\leqslant 1$$

Arithmetic Mean$=\dfrac{m}{m+n}$, Harmonic Mean$=\dfrac{n-1}{m+n-1}$, $m>1$. Verify that $A.M.>H.M.$ Also show that *(M.A. Banaras, 1949)*

$$\text{Variance}=\frac{mn}{(m+n)^2(m+n+1)} \qquad \textit{(M.Sc. Agra, 1954)}$$

Prove also that

$$\log G=\frac{1}{B(m,\,n)}\int_0^1(1-x)^{m-1}x^{n-1}\log x\,dx$$

$$=\frac{1}{B(m,\,n)}\cdot\frac{\partial}{\partial n}B(m,\,n)=\frac{\partial}{\partial n}[\log\Gamma(n)-\log\Gamma(m+n)].$$

15. Show that for the Gamma distribution

$$dF=\frac{1}{\Gamma(m)}e^{-x}x^{m-1},0\leqslant x\leqslant\infty,\ m>0$$

mean is m, variance$=m$, harmonic mean$=m-1$, $m.g.f.=(1-t)^{-m}$, $|\,t\,|<1$.

μ_r' (about origin)$=m(m+1)\ldots(m+r-1)$.

16. (*a*) **Express the constants** y_0, a **and** p $(0<p<1)$ **of the distribution**

$$dF=y_0\left(1-\frac{x^2}{a^2}\right)^{-p}dx,\ -a<x<a$$

in terms of σ **and** β_2. *(B.A. Hons. Delhi, 1962)*

(*b*) Show that for the distribution

$$dF=y_0\left(1-\frac{x^2}{a^2}\right)^{-p}dx,\ 0<p<1,\ -a\leqslant x\leqslant a,\ y_0 \text{ being a constant}$$

$$\mu'_{r+1}=\frac{ra^2}{r+2-2p}\ \mu'_{r-1}$$

[**Hint.** Integrate by parts to get the recurrence relation]

17. Find y_0, μ_1', μ_2, μ_3 and μ_4 for the distribution

$$dF=y_0\left(1+\frac{\alpha}{2}\right)^{\left(\frac{4}{\alpha^2}-1\right)}e^{\frac{-2x}{\alpha}},\ \frac{-2}{\alpha}\leqslant x<\infty$$

[**Ans.** $y_0=\beta^{2\beta^2-1}/e^{\beta^2}\sqrt{\beta^2}$, where $\beta=\frac{2}{\alpha}$, $\mu_1'=0$, $\mu_2=1$, $\mu_3=\alpha$, $\mu_4=\frac{3}{2}\alpha^2+3$]

18. The probability law of a continuous random variable x is

$$f(x)=y_0e^{-b(x-a)},\ a\leqslant x<\infty,$$

where a, b, y_0 are constants. Show that $y_0=b=\sigma-1$ and $a=m-\sigma$ where m and σ are the mean and standard deviation of the distribution Show also that $\beta_1=3$, $\beta_2=9$.

(B.Sc. Hons. Calcutta, 1964 ; M.A. Delhi, 1957 ; B. A. Hons. Delhi, 1950, 60)

19. If x is chosen at random in the interval (a, b), what are $E(x)$ and $Var\,(x)$? [**Ans.** $\frac{a+b}{2}$, $\frac{(b-a)^2}{12}$]

20. A random variate x is known to have the distribution

$$dP=C\left(1+\frac{x}{a}\right)^{m-1}e^{-\frac{mx}{a}}dx,\ -a\leqslant x<\infty.$$

Find the constant C and the first four moments about the mean. Determine the linear relation between β_1 and β_2 of this distribution. Also find the *M.G.F.* about the origin. *(M.A. Statistics Delhi, 1968)*

[**Ans.** $C=m^m.e^{-m}/a\lfloor\overline{m}$, $\mu_1'=0$, $\mu_2=\frac{a^2}{m}$, $\mu_3=\frac{2a^3}{m^2}$, $\mu_4=\frac{3a^4(m+2)}{m^3}$

$2\beta_2-3\beta_1=6$. $M.G.F.=e^{-at}\left(1+\frac{at}{m}\right)^{-m}$]

21. Pareto's distribution is defined by the density function

$$f(x)=\frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1}\ \text{for } x>\beta$$

$$=0,\qquad\qquad \text{for } x\leqslant\beta.$$

Show that a variable with this distribution has a moment of order r if and only if $\alpha > r$. In particular, if $\alpha > 1$, show that

$$E(x) = \alpha\beta/(\alpha-1), \; Var\, x = \alpha\beta^2/[(\alpha-1)^2(\alpha-2)].$$

22. Find the position of the mode and the median of the curve

$$f(x) = \frac{abx^{a-1}}{(1+bx^a)^2}, \; b>0, \; a>1, \; 0 \leqslant x \leqslant \infty.$$

$$\left[\text{Ans.} \quad \text{Mode} = \left[\frac{a-1}{b(b+1)}\right]^{1/a}, \quad \text{Median} = \left(\frac{1}{b}\right)^{1/a}\right]$$

23. Can the function $E[(1-tx)^{-1}]$ be used to generate the moments of a variate x ? [Ans. Yes]

24. For the distribution

$$dP = \frac{x}{a^2} \; e^{-x^2/2a} \; dx, \; 0 \leqslant x < \infty.$$

find the distance between the quartiles and show that the ratio of this distance to the standard deviation is independent of the parameter a. Find the ratio.

$$\left[\text{Ans.} \quad \text{Distance} = a\sqrt{2}[\sqrt{\log 4} - \sqrt{\log \tfrac{4}{3}}], \; \text{Ratio} = \frac{\sqrt{\log 4} - \sqrt{\log \frac{4}{3}}}{\sqrt{1-\pi/2}}\right]$$

25. Laplace's double exponential distribution is defined by the density function

$$f(x) = \tfrac{1}{2}e^{-|x|}, \; -\infty < x < \infty$$

Check that $\int_{-\infty}^{\infty} f(x)\, dx = 1$

and find (*a*) arithmetic mean (*b*) standard deviation (*c*) mean deviation about the mean (*d*) characteristic function for the distribution.

(*B.Sc. Hons., Calcutta, 1967*)

[**Ans.** (*a*) mean $= 0$, (*b*) $s.d. = \sqrt{2}$,

(*c*) m.d. about mean $= \tfrac{1}{2}\int_{-\infty}^{+\infty} |x| e^{-|x|}\, dx$

$$= \int_0^{\infty} |x| e^{-|x|}\, dx = \int_0^{\infty} xe^{-x} = 1,$$

(*d*) $c.f. = \dfrac{1}{1+t^2}$

26. For the probability density function

$$f(x) = \frac{2(a+x)}{b(a+b)}, \; -b \leqslant x \leqslant 0$$

$$= \frac{2(a-x)}{a(a+b)}, \; 0 \leqslant x \leqslant a$$

find the mean, variance and median. *(B.Sc. Bombay, 1968)*

[**Ans.** mean$=\frac{a-b}{3}$, variance$=\frac{1}{18}(a^2+ab+b)$,

median$=a-\sqrt{\frac{1}{2}a(a+b)}$]

27. A continuous cumulative distribution $F(x)$ is defined as follows :

$$(a) \qquad F(x)=0 \qquad\qquad x\leqslant 1$$
$$=\frac{1}{16}(x-1)^4 \qquad 1<x\leqslant 3$$
$$=1 \qquad\qquad x>3.$$

Find the probability density function $f(x)$. Find the mean of x and the median. *[B.Sc. (S) Bombay, 1964]*

[**Ans.** $f(x)=\frac{1}{8}(x-1)^3$, $1\leqslant x\leqslant 3$, Mean$=2{\cdot}65$, Median$=1+\sqrt[3]{8}$.

$=0$, otherwise.

(*b*) The *p.d.f.* of a random variable X is

$$f(x)=\begin{cases} 0 & \text{for } x\leqslant -a \\ \frac{a+x}{a^2} & \text{for } -a<x\leqslant 0 \\ \frac{a-x}{a^2} & \text{for } 0<x\leqslant a \\ 0 & \text{for } x\geqslant 0 \end{cases}$$

Compute cumulative distribution function $F(x)$.

(I.S.I. 1969, B.Sc. Hons. Poona, 1966)

[**Ans.** $F(x)=0$, $x\leqslant -a$

$$= \frac{ax+\frac{1}{2}x^2+\frac{1}{2}a^2}{a^2}, \quad -a<x<0$$
$$= \frac{\frac{1}{2}+ax-\frac{1}{2}x^2}{a}, \quad 0<x\leqslant a\,;\ =1,\ x\geqslant a.]$$

(*c*) A random variable X has a continuous distributions F, where

$$F(x)=0, \qquad \text{if } x<0$$
$$=k_x, \qquad \text{if } 0\leqslant x<1$$
$$=1, \qquad \text{if } x>1$$

Determine k and *p.d.f.*

28. Two independent variates are each uniformly distributed within the range $-a$ to $+a$. Show that their sum x has a probability density given by

$$f(x)=\frac{2a+x}{4a^2}, \quad -2a\leqslant x\leqslant 0$$
$$=\frac{2a-x}{4a^2}, \quad 0\leqslant x\leqslant 2a.$$

Verify that the *m.g.f.* calculated from this value of $f(x)$ is equal to

$$\left(\frac{1}{at}\sinh at\right)^2. \qquad \textit{(M.A. Punjab, 1952)}$$

29. If $f(x)$ is the elementary probability law of continuous random variable x in the interval $a \leqslant x \leqslant b$, so that

$$f(x)=0 \text{ for } x<a \text{ and } x>b$$

and

$$\int_a^b f(x)dx=1,$$

show that the variable $P=\int_a^x f(x)\,dx$ has a rectangular distribution.

30. (*a*) Explain the concepts of (*i*) conditional probability distribution, (*ii*) independence of random variables and (*iii*) marginal probability density function. Given the joint density function

$$f(x, y)=3x^2y+3y^2x \qquad 0 \leqslant x \leqslant 1, \qquad 0 \leqslant y \leqslant 1$$
$$=0, \qquad \text{otherwise}$$

show that the marginal density functions are

$$x+\frac{3}{2}x^2,\ y+\frac{3}{2}y^2$$

Find the conditional probability $P(\frac{1}{2} \leqslant x \leqslant \frac{3}{4} \mid \frac{1}{3} \leqslant y \leqslant \frac{2}{3})$.

[Ans. $\frac{311}{1033}$]

31. (*a*) In each of the following problems a joint density (probability) function $f(x, y)$ is given. It is understood that $f(x, y)=0$ outside the region specified. Determine whether the variables are dependent or independent.

$f(x, y)=4xy \qquad (0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 1)$ [Ans. Independent]

$f(x, y)=\frac{1}{4\pi^2}\,[1-\sin\,(x+y)], \quad -\pi \leqslant x \leqslant \pi, \quad -\pi \leqslant y \leqslant \pi$

[Ans. Dependent]

$f(x, y)=\frac{1}{2}[\cos(x+y)+\cos(x-y)]=\sin x \sin y, \quad (0 \leqslant x \leqslant \pi/2, \ 0 \leqslant y \leqslant \pi/2)$

[Ans. Independent]

(*b*) The joint density of two discrete random variables x and y is given by

y \ x	0	1	2	3
0	1/48	5/48	5/48	1/48
1	5/48	15/48	5/48	0
2	5/48	5/48	—	0
3	1/48	0	0	0

Find the marginal densities of x and y

(*ii*) Find the conditional density of x given $y=3$

(*iii*) Find $E(y \mid x=0)$

[**Ans.** (*i*) $\frac{1}{4}, \frac{25}{28}, \frac{5}{24}, \frac{1}{48}$ (*ii*) $\frac{1}{48}$ (*iii*) 1·5.]

32. (*a*) Given $f(x, y)=8xy$ $(0 \leqslant x<1,\ 0 \leqslant y \leqslant x)$; show that x and y are not independent random variables in spite of the fact that $f(x, y)$ is the product of a function of x and a function of y alone.

(*b*) If $f(x)=e^{-x}$, $0 \leqslant x < \infty$ and zero for $x<0$, show that $u=x_1+x_2$ and $v=\frac{x_1}{x_1+x_2}$ are stochastically independent variates.

(*c*) If x, y and z are random variables with *p.d.f.*

$f(x, y, z)=1-\sin 2\pi x \sin 2\pi y \sin 2\pi z$, $0 \leqslant x, y, z \leqslant 1$

show that any two of the variables are independent. (*M. Sc. Poona, 1957*)

(*d*) A random observation on a bivariate population (x, y) can yield one of the following pairs of values with probabilities noted against them.

For each observation pair	Probability
(1, 1) ; (2, 1) ; (3, 3) ; (4, 3)	$\frac{1}{20}$
(3, 1) ; (4, 1) ; (1, 2) ; (2, 2) (3, 2) ; (4, 2) ; (1, 3) ; (2, 3)	$\frac{1}{10}$

Find the probability that $y=2$ given $x=4$ and also the probability that $y=2$. Examine if the two events $x=4$ and $y=2$ are independent.

(*B.Sc. Madras, 1959*)

[**Hint.** $P(y=2)=\frac{4}{10}$, $P(x=4)=\frac{1}{4}$, $P(x=4,\ y=2)=\frac{1}{10}$

$(Py=2,\ x=4)=\frac{\frac{1}{10}}{\frac{1}{4}}=\frac{2}{5}$

Hence $P(x=4)P(y=2)=P(x=4,\ y=2)$].

33. (*a*) If $f(x, y)=x+y$ $\quad 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1$

$=0$, elsewhere

find the density function for (*i*) $u=X+Y$ (*ii*) $v=XY$

(*b*) Let X and Y have the joint density function

$f(x, y)=\frac{3}{2}x$, $0<x<1$, $-x<y<x$

$=0$, otherwise.

find the density function of $X-Y$. (*M.Sc. Bombay 1967, 68*)

[**Ans.** (*i*) $h(u)=\begin{cases} u^2 & \text{for } 0 \leqslant u \leqslant 1 \\ 2u-u^2 & \text{for } 1 \leqslant u \leqslant 2,\ 0, \text{ otherwise} \end{cases}$

(*ii*) $h=2-2v$, $\quad 0<v \leqslant 1$, $=0$ otherwise]

34. Two random variables X, Y have

$$P(x=0, y=0)=\frac{2}{9}, \quad P(x=0, y=1)=\frac{1}{9}$$

$$P(x=1, y=0)=\frac{1}{9}, \quad P(x=0, y=1)=\frac{5}{9}$$

Examine whether (*i*) X and Y are identical variables
(*ii*) X and Y are independent random variables.

35. (*i*) If $f(x, y)=1$ when $0\leqslant X\leqslant 1$, $0\leqslant Y\leqslant 1$ and zero otherwise, find the probability that

(*a*) $x<\frac{1}{2}$, $y<\frac{1}{2}$, (*b*) $x+y<1$, (*c*) $x+y>1$, (*d*) $x>2y$, (*e*) $x>\frac{1}{2}$
(*f*) $x^2+y^2<\frac{1}{4}$, (*g*) $x=y$ (*h*) $x>\frac{1}{2}$, given $y<\frac{1}{2}$, (*i*) $x>y$ given $y>\frac{1}{2}$.

[**Ans.** (*a*) $\frac{1}{4}$, (*b*) $\frac{1}{2}$, (*c*) $\frac{1}{2}$, (*d*) $\frac{1}{4}$, (*e*) $\frac{3}{8}$, (*f*) $\frac{1}{16}\pi$, (*g*) $\frac{1}{2}$, (*h*) $\frac{1}{2}$ (*i*) $\frac{1}{4}$]

(*ii*) Given $f(x, y)=1$ over the square $(0\leqslant x, y\leqslant 1)$ show that
$P(xy>u)=1-u+u\log u$.

[**Hint.** (*ii*) $P(xy>u)=\int_{x=u}^{1}\int_{y=\frac{u}{x}}^{1} 1\,dxdy=$etc.]

36. (*a*) If $f(x, y)=2$; $0\leqslant x\leqslant 1$, $0\leqslant y\leqslant x$ find the marginal and conditional density functions.

(*b*) If $f(x, y)=xe^{-x(y+1)}$ $(x\geqslant 0, y\geqslant 0)$ find the marginal and conditional density functions.

[**Ans.** (*a*) $g(x)=2x$, $h(y)=2$, (*b*) $f(x\mid y)=\frac{1}{y}$, $f(y\mid x)=\frac{1}{x}$ (*b*) $g(x)=e^{x}$,
$h(y)=\frac{1}{(y+1)^2}$, $f_1(x\mid y)=(1+y)^2xe^{-x(1+y)}$, $f_2(y\mid x)=x\,e^{-xy}$, $x, y>0$]

37. Let X have the density function $f(x)=e^{-x}$, $x>0$ and zero for $x<0$. Find the density function for (*i*) $u=\sqrt{x}$, (*ii*) $u=x^{-1}$.

[**Ans.** (*i*) $g(u)=2ue^{-u^2}$; (*ii*) $g(u)=u^{-2}e^{-\frac{1}{u}}$]

38. If X is normally distributed with zero mean and unit variance find the density function for $u=e^{x}$.

Ans. [$g(u)=\frac{1}{u\sqrt{2\pi}}\,e^{-\frac{1}{2}(\log u)^2}$, $u\geqslant 1$]
$=0$, otherwise.

39. (*a*) If $f(x)=1$, $0\leqslant x\leqslant 1$, find the density function of $u=3x+1$.

[**Hint.** First find the cumulative distribution of u and then differentiate it.

(*b*) If $f(x)=\frac{1}{2}e^{-\frac{x}{2}}$, $0\leqslant x<\infty$ and zero for $x<0$, find the marginal density function of $u=\frac{1}{2}(x_1-x_2)$.

[**Ans.** $g(u)=\frac{1}{2}e^{-|u|}$, $-\infty<u<\infty$]

40. (*a*) If $f(x, y)=e^{-(x+y)}$: $x>0$, $y>0$ find

(*i*) the probability of the events (*a*) $x<1$, (*b*) $x>y$ (*c*) $x+y<1$ (*d*) $x<2y$

(*ii*) the density function of $u=\frac{1}{2}(x+y)$

and show that $\mu_{10}'=1=\mu_{01}'$, $\mu_{20}'=2=\mu_{01}'$.

(*M.Sc. Bombay, 1958*; *B Sc Bombay, 1968*)

[Ans. (*i*) $1-\frac{1}{e}$, $\frac{1}{2}$, $1-\frac{1}{e}$, $\frac{2}{3}$ (*ii*) $g(u)=4ne^{-2u}$]

(*b*) Let the random variables X and Y have their density function

$$f(x, y)=\frac{1}{2\pi}\exp\left[-\frac{1}{2}\left(x^2-2xy+2y^2\right)\right]$$

show that

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\,dxdy=1,\ \mu_{10}=0,\ \mu_{01}=0,\ \mu_{11}=1,\ \mu_{20}=2,$$

$$\mu_{02}=1,\ \rho=1/\sqrt{2},\ g(x)=\int_{-\infty}^{+\infty} f(x, y)dy=\frac{1}{2\sqrt{\pi}}\ e^{-\frac{1}{4}x^2},\ h(y)=\frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}y^2}$$

$$f(y\mid x)=\frac{1}{\sqrt{\pi}}\ e^{-\left(y-\frac{x}{2}\right)^2},\quad f(x\mid y)=\frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}(x-y)^2}$$

41. Let X and Y be non-negative continuous chance variables having the joint density

$f(x, y)=4xy\exp[-(x^2+y^2)]$; $0<x<\infty$, $0<y<\infty$.

Prove that the density function of $u=\sqrt{x^2+y^2}$ is

$g(u)=2u^3\exp[-u^2]$; $0<u<\infty$.

(*M.A. Aligarh, 1960* ; *B.A. Hons. Delhi, 1952*)

42. Let $f(x, y)=\frac{2}{a^3}$, $0\leqslant x\leqslant y$, $0\leqslant y\leqslant a$, show that

(*i*) $g(x)=\frac{2}{a^2}(a-x)$, (*ii*) $h(y)=\frac{2}{a^2}\cdot y$, (*iii*) $f(x\mid y)=\frac{1}{y}$ (*iv*) $\mu_{10}'=\frac{a}{3}$,

(*v*) $\mu_{01}'=\frac{2a}{3}$, (*vi*) $r_{xy}=\frac{1}{2}$.

43. (*a*) A point P is taken at random in a line AB of length $2a$, als positions of the point being equally likely. Show that the expected value of the area of the rectangle $AP.PB$ is $2a^2/3$, and that the probability of the area exceeding $\frac{1}{2}a^2$ is $1/\sqrt{2}$. (*B.Sc. Agra, 1956, 58*)

(*b*) Two points are taken at random on a given straight line of length a. Prove that the probability of their distance exceeding a given length $c(<a)$ is equal to $\left(1-\frac{c}{a}\right)^2$.

44. From a point on the circumference of a circle of radius a, a chord is drawn in a random direction (*i.e.*, all directions are equally likely). Show tha the expected value of the length of the chord is $4a/\pi$ and the variance of th

length is $2a^2\left(1-\frac{8}{\pi^2}\right)$. Also show that the chance is $\frac{1}{3}$ that the length of the chord will exceed the length of the side of an equilateral triangle inscribed in the circle. *(B.Sc. Agra, 1962)*

45. Two points X_1 and X_2 are taken randomly on a line of finite length l. If X_1 and X_2 are at distances x_1 and x_2 respectively from one given end, find $E[\,|x_1-x_2|^n]$.

[Ans. $\frac{2l^n}{(n+1)(n+2)}$]

46. *(a)* What is the expected value of the square of the distance between two points P and Q selected at random in a square whose side is of length l?

(b) What is the expected distance between two points selected at random in a circle of radius a? [Ans. *(a)* $l^2/3$. *(b)* $128a/45\pi$]

47. A point is taken at random within the ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1,\ (a<b)$$

What is the chance that its distance from the centre exceeds a?

[Ans. $(b-a)/b$]

48. *(a)* Find the mean value of the ordinates of a semi-circle of radius r supposing the ordinates drawn through equidistant points on the circumference.

(b) From a point A on the circumference of a circle of radius r two st. lines are drawn to two other points B and C taken at random on the circumference. Find the mean difference in length between the lines AB and AC.

(M.A. Delhi, 1961)

[Ans. *(a)* $\frac{2r}{\pi}$]

49. If the roots of the quadratic equation $x^2-ax+b=0$ are real and b is positive but otherwise unknown, what are the mean values of the roots?

[Ans. $\frac{5a}{6}$, $\frac{a}{6}$]

50. Three points are taken at random on the circumference of a circle Find the probability that they lie in the semi-circle.

(Assume that in the limit elementary intervals of one are equally probable). [Ans. $\frac{3}{4}$]

51. The sides of a rectangle are taken at random each less than an inch and all lengths are equally likely. Find the chance that the diagonal is less than an inch. [Ans. $\frac{\pi}{4}$]

52. A point P is taken at random on the circumference of a circle of radius r. Find the expected value of the distance AP, between P and a given point A on the circumference. *(M.A. Punjab, 1959)*

[Ans. $\frac{4r}{\pi}$]

53. Bertrand Paradox. A chord is drawn at random in a given circle. What is the probability that it is greater than the side of the equilateral triangle inscribed in that circle ?

Sol 1. Assuming that the distance of the chord from the centre has uniform distribution, the required probability$=\frac{1}{2}r/r=\frac{1}{2}$, r being the radius of the circle.

Sol. 2. Assuming that the distribution of the angle between the chord and the tangent at one end of the chord is uniform, the required probability is

$$\left(\frac{2\pi}{3}-\frac{\pi}{3}\right)\Big/\pi=\frac{1}{3}.$$

54. A straight line AB is divided by a point C into two parts $AC=a$, $CB=b$. Point X and Y are taken at random on AC and CB respectively. What is the probability that AX, XY, BY can form a triangle ? *(I.S.I. Calcutta, 1953)*

[**Ans.** $\frac{1}{2}\frac{b}{a}$]

[**Hint.** In order that AX, XY, BY can form a triangle, the following inequalities must be fulfilled :

$$x<(a+b-x-y)+y \quad \text{or} \quad x<a+b-x$$
$$y<(a+b-x-y)+x \quad \text{or} \quad y<a+b-y$$
$$a+b-x-y<x+y$$

These inequalities are equivalent to

$$x<\frac{a+b}{2},\quad y<\frac{a+b}{2},\quad x+y<\frac{a+b}{2}\Big]$$

55. (*a*) A board is ruled with equidistant parallel lines, the width of the strip between two consecutive parallel lines being d. A needle so fine that it can be likened to a rectilinear segment of length $l<d$ is thrown on the board.

(*i*) What is the probability that needle will intersect one of the lines ?
(M.A. Bombay, 1956 ; M.A. Nagpur, 1953)

(*ii*) What happens when $l<d$?

[**Ans.** $2l/d\pi$, (*ii*) $\frac{l}{ad}(1-\cos d)$ where $\sin d=\frac{a}{l}$]

(*b*) The radius X of a circle is rectangularly distributed between 1 and 2. Show that the area y of the circle is distributed between π and 4π with probability distribution $f(y)\,dy=\frac{1}{2}\,\pi^{-\frac{1}{2}}\,y^{-\frac{1}{2}}\,dy$.

Show that the mean and variance of the area of the circle is $\frac{7\pi}{3}$ and $\frac{34\pi^2}{45}$ respectively.

(*c*) The two equal sides of an isosceles triangle are each of unit length and the angle θ between them is rectangularly distributed between 0 and $\pi/6$.

Show that the area of the triangle is described between 0 and 1/4 with probability distribution

$$f(y)dy=\frac{12}{\pi}\left(1-4y^2\right)^{-\frac{1}{2}}dy.$$

Show that the mean of the area of the triangle is $\frac{3}{2\pi}(2-\sqrt{3})$.

56. (*a*) If $f(x)$ and its derivatives vanish at the end points of the range prove by integration by parts that for the distribution $dF=f(x)dx$,

$$\int_{-\infty}^{+\infty} e^{it\infty} f^r(x)dx=(-1)^r(it)^r\psi(t)$$

where $\psi(t)$ is the *C.F.* and $f^r(x)=\frac{d^r}{dx^r}f(x)$.

(*b*) In particular, show that the distribution

$$dF=\exp\left\{\frac{a}{|q}\left(-\frac{d}{dx}\right)^q\right\}f(x)dx,$$

the qth cumulant exceeds the qth cumulant for the distribution $dF=f(x)dx$ by a and that all other cumulants of the two distributions are equal.

(M.A. Delhi, 1968)

57. For (*i*) $f(x)=1,\ 0\leqslant x\leqslant 1$

(*ii*) $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2},\ -\infty<x<\infty$

determine $P[\,|\,x-E(x)\,|\,\geqslant 1{\cdot}5\sigma]$. Compare this value with the value given by Chebychev's inequality.

[**Ans.** (*i*) 0·134 (*ii*) 0·134 Chebychev's inequality gives 0·144]

58. (*a*) If X be a random variable such that $E(X)=3$ and $E(X^2)=0$, then using Tchebycheff's lemma, show that

$$P(-2<X<8)\geqslant\frac{21}{25}.$$

(B Sc. Hons. Gauhati, 1971)

(*b*) A random variable X takes the values -1, 1, 3, 5 with associated probabilities 1/6, 1/6, 1/6 and 1/2. Find by direct computation $P[\,|\,X-3\,|\,\geqslant 1]$. Find an upper bound for the probability by applying Tchebycheff's inequality.

(B.Sc. Mysore, 1967)

(*c*) Find k if $P[\,|\,X-E(X)\,|\,>k\sigma_x)\leqslant\frac{1}{4}$. Let X be a random variable such that $P(X\leqslant 0)=0$ and let $\mu=Ex$ exists show that

$$P(X\geqslant 2\mu)\leqslant\tfrac{1}{2}.$$ *(I.S.I. 1971)*

(*d*) Two good dice whose faces are numbered 1 to 6 are thrown. If X is the sum of the numbers showing up, prove that

$$P\{\,|\,X-7\,|\,\geqslant 3\}<\frac{35}{54}$$

Compare with the exact probability. *(B.Sc. Calicut, 1972)*

(*e*) A sample of size n is drawn from a population whose mean is 5 and s.d. 1. Prove that

$$P\{ \mid X-5 \mid \leqslant 0{\cdot}00\} \geqslant 1-\frac{10^6}{n}. \qquad (B.Sc.\ Shivaji,\ 1971)$$

(*f*) Compare the upper bound on the probability

$$P\{ \mid X-E(X) \mid \geqslant 2\sqrt{\text{Var } X}\}$$

obtained from Chebychev's inequality with exact probability if X is uniformly distributed over $(-1, 3)$. (*B.Sc Poona, 1971*)

(*g*) Suppose X has a rectangular distribution on $(-1, 1)$. Compute

$$P\left|\frac{X-E(X)}{\sigma_x}\right| \geqslant 2$$

and compare it with the upper bound given by Tchebycheff's inequality.

59. (*a*) **Chebychev's inequality.** Show that for any distribution

$$\int_{-\infty}^{+\infty}\left(1-\frac{x^2}{t^2}\right)dF(x) \leqslant \int_{-t}^{+t} dF(x)$$

and hence or otherwise deduce that if variance σ^2 of x exists,

$$P\left(\mid x-E(x) \mid \geqslant k\sigma\right) \geqslant \frac{1}{k^2},\ k>0. \qquad (I.C.A.R.,\ 1959,\ 61)$$

60. If x is a stochastic variable with zero mean and s.d σ, show that

$$P(t) \leqslant \frac{\sigma^2}{\sigma^2+t^2} \qquad \text{for } t<0$$

and

$$1-P(t) \leqslant \frac{\sigma^2}{\sigma^2+t^2} \qquad \text{for } t>0$$

where $P(t)$ denotes the probability of the inequality $x \leqslant t$.

(*M.A., Statistics Delhi, 1968*)

[**Hint.** Let $t<0$ then

$$-t = \int_{-\infty}^{\infty}(x-t)dF(x) < \int_{x}^{\infty}(x-t)dF(x)$$

Apply Schwartz inequality *viz.*

$$\left[\int g(x)h(x)dF(x)\right]^2 \leqslant \int g^2\,dF(x)\int h^2\,dF(x)$$

so that

$$t^2 \leqslant \left[\int_{x}^{\infty}(x-1)dF\,x)\right]^2 \leqslant \int_{x}^{\infty} dF(x)\int_{x}^{\infty}(x-t)^2\,dF(x)$$

$$=[(1-P)(t)](\sigma^2t+\sigma^2)$$

61. If $\gamma_2=\beta_2-3$, then show that it is possible to improve Chebychev inequality in the form,

$$P\{ \mid x-m \mid \geqslant k\sigma\} \leqslant \frac{\gamma_2+2}{(k^2-1)^2+\gamma_2+2}.$$

Thus the equations to determine the best values of θ_1, θ_2 are given by

$$0=\frac{\partial E}{\partial \theta_1}=-\Sigma(x_i-\theta_1)+\Sigma(z_i-180+\theta_1+\theta_2),$$

and

$$0=\frac{\partial E}{\partial \theta_2}=-\Sigma(y_i-\theta_2)+\Sigma(z_i-180+\theta_1+\theta_2).$$

These give $\Sigma x+3\theta_1+\Sigma z-540+3\theta_1+3\theta_2=0$

or $6\theta_1+3\theta_2-417{\cdot}6=0$ ...(1)

and $\Sigma x+3\theta_2+\Sigma z-540+3\theta_1+3\theta_2=0$

or $3\theta_1+6\theta_1-481{\cdot}8=0$ ...(2)

Solving (1) and (2) for θ_1 and θ_2, we get

$$\hat{\theta}_1=39{\cdot}27,\quad \hat{\theta}_2=60{\cdot}66,\quad \hat{\theta}_3=180-\hat{\theta}_1-\hat{\theta}_2=80{\cdot}07$$

9·2. Curve Fitting. The fitting of a curve to a given data is very important both from the point of view of theoretical and practical Statistics. In theoretical Statistics, we shall find that the study of regression and correlation (whether total, partial or multiple) can be regarded as fitting of linear curves to the given bivariate or multivariate frequency or probability distributions. In practical Statistics, it enables us to get a close functional relation between x and y. The relation is, in general, expressed by a polynomial but other types of relationship, algebraic, exponential or logarithmic can also be fitted by using the principle of least squares.

9·21. Fitting of Polynomials or Parabolic Curves. Let it be required to fit the curve whose equation is

$$Y=a_0+a_1x+a_2x^2+\ldots\ldots+a_px^p, \qquad \ldots(9{\cdot}2)$$

the coefficients a_0, a_1, ..., a_p, being constants and $a_p\neq 0$, to a number (say $>p+1$) of given values (x_1, y_1) ; (x_2, y_2) ; ..... ; (x_n, y_n) of the two variables so that we have to determine the $(p+1)$ constants a_0, a_1, ..., a_p. The curve whose equation is (9·2), is called the *parabola of degree p*. Substituting x_i for x in (9·2), we get

$$Y_i=a_0+a_1x_i+a_2x_i^2+\ldots\ldots+a_px_i^p \ (i=1, 2, \ldots\ldots, n).$$

The quantity Y_i is generally called the *expected values* of y corresponding to $x=x_i$, contrasted with the value y_i which is called the *observed value* of y corresponding to the same value x_i of x. These two values are, in general, different as the observed points do not necessarily lie on a parabolic curve. The difference y_i-Y_i is called

the *residual* corresponding to $x=x_i$ and, according to the *principle of least squares*, we choose the constants $a_0, a_1, \ldots\ldots, a_p$ so that sum of squares of residuals is minimum.

$$\text{Now} \quad S=\sum_{i=1}^{n}(y_i-Y_i)^2=\sum_i(y_i-a_0-a_1x_i-\ldots-a_px_i^p)^2. \qquad \ldots(9\cdot3)$$

Differentiating (9·3) partially, we get $(p+1)$ normal equations. On simplification, these equations reduce to

$$\left.\begin{aligned}
\Sigma y &= na_0 + a_1\Sigma x + \ldots\ldots + a_p\Sigma x^p \\
\Sigma xy &= a_0\Sigma x + a_1\Sigma x^2 + \ldots\ldots + a_p\Sigma x^{p+1} \\
\Sigma x^2y &= a_0\Sigma x^2 + a_1\Sigma x^3 + \ldots\ldots + a_p\Sigma x^{p+2} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\Sigma x^py &= a_0\Sigma x^p + a_1\Sigma x^{p+1} + \ldots\ldots + a_p\Sigma x^{2p}
\end{aligned}\right\} \qquad \ldots(9\cdot4)$$

These equations are called the *normal equations* for fitting (9·2). They are $(p+1)$ in number and when solved as simultaneous equations give the values of the $(p+1)$ unknowns $a_0, a_1, \ldots\ldots, a_p$.

In particular, when $p=1$, the normal equations for fitting the straight line $y=a_0+a_1x$ are

$$\left.\begin{aligned}
\Sigma y &= a_0n + a_1\Sigma x \\
\Sigma xy &= a_0\Sigma x + a_1\Sigma x^2
\end{aligned}\right\} \qquad \ldots(9\cdot5)$$

Thus the best fitting linear function for the points (x_i, y_i), $i=1, 2, \ldots\ldots, n$ may be expressed in the form

$$\begin{vmatrix} x & y & 1 \\ \Sigma x & \Sigma y & n \\ \Sigma x^2 & \Sigma xy & \Sigma x \end{vmatrix} = 0.$$

When $p=2$, the normal equations for fitting the second degree parabola $y=a_0+a_1x+a_2x^2$ are

$$\left.\begin{aligned}
\Sigma y &= a_0n + a_1\Sigma x + a_2\Sigma x^2 \\
\Sigma xy &= a_0\Sigma x + a_1\Sigma x^2 + a_2\Sigma x^3 \\
\Sigma x^2y &= a_0\Sigma x^2 + a_1\Sigma x^3 + a_2\Sigma x^4
\end{aligned}\right\} \qquad \ldots(9\cdot6)$$

Note. Change of origin. When the values $x_i (i=1, 2, \ldots\ldots, n)$ correspond to equal increment h nnd n is *odd* say $=2m+1$, the equations (9·4) can be simplified by taking the origin of x at the middle of the value and the common increment h as unit of measurement. Thus if a be the middle value, $u=(x-a)/h$ takes the values $-m, \ldots\ldots, -1, 0, 1, \ldots\ldots, m$. Hence, we get

$$\Sigma u=0=\Sigma u^3=\ldots\ldots$$

If n is *even*, say $2m$ we take the origin of x at the mean of the middle pair of values and $h/2$ as the new unit *i.e.*,

$$u = \frac{x - (\text{mean of two suitable terms})}{\frac{1}{2}h}.$$

The values of x then become $-(2m-1)$, $-(2m-3)$, ..., -3, -1, 1 3, ..., $(2m-3)$, $(2m-1)$, and again $\Sigma u = 0 = \Sigma u^3 = ...$

Example 9·3. *The weight of a calf taken at weekly intervals are given below. Fit a straight line using the method of least squares, and calculate the average rate of growth per week.*

Age (x)	1	2	3	4	5	6	7	8	9	10
Weight (y)	52·5	58·7	65·0	70·2	75·4	81·1	87·2	95·5	102·2	106·4

(*B.A., Hons., Delhi, 1956* ; *B.A. Madras, 1945*)

Since n is even, we take the origin of x at the mean of the middle pair of values, and $\frac{1}{2}$ as the new unit that is we take $u = \frac{x - 5·5}{\frac{1}{2}} = 2x - 11$. The values of u, then are

$$-9, -7, -5, -3, -1, 1, 3, 5, 7, 9$$

We find that $\Sigma u = 0$, $\Sigma u^2 = 330$, $\Sigma y = 796·2$ and $\Sigma uy = 1016·8$

The normal equations for fitting the straight line

$$y = a + bu$$

are $$\Sigma y = na + b\Sigma u, \quad \Sigma uy = a\Sigma u + b\Sigma u^2.$$

On substituting the values of Σy, etc., we have

$$796·2 = 10a, \qquad 1016·8 = 330b$$

$$\therefore \quad a = 79·62, \qquad b = 3·08 \text{ approx.}$$

The regression equation is, therefore,

$$y = 79·62 + 3·08u = 79·62 + 3·08\,(2x - 11)$$

or $$y = 45·74 + 6·16x$$

The average rate of growth per week is 6·16 units.

Example 9·4. *Fit a second degree parabola to the following data taking x as the independent variable.*

x :	1	2	3	4	5	6	7	8	9
y :	2	6	7	8	10	11	11	10	9

(*M.Sc., Agra, 1953*)

Introduce two new variables u and v connected by the relations $u = x - 5$, $v = y - 7$. Since n is odd and the values of x are equidistant we take the origin for the x series at the middle value 5. This makes $\Sigma u = 0 = \Sigma u^3$.

Let us fit the curve

$$v=a_0+a_1u+a_2u^2.$$

The calculations of Σuv, Σu^2v, etc., are shown in the following table :

	u	v	uv	u^2	u^2v	u^3	u^4
	−4	−5	20	16	−80		256
	−3	−1	3	9	−9		81
	−2	0	0	4	0		16
	−1	1	−1	1	1		1
	0	3	3	0	0		0
	1	4	4	1	4		1
	2	4	8	4	16		16
	3	3	9	9	27		81
	4	2	8	16	32		256
Total	0	11	51	60	−9	0	703

Substituting the values of Σu, etc., in the normal equations (9·6), we get

$$11=9a_0+0+60a_2$$
$$51=0+60a_1+0$$
$$-9=60a_0+0+708a_2.$$

Solving these simultaneous equations we obtain

$$a_0=3{\cdot}0,\ a_1=0{\cdot}85, \text{ and } a_2=-0{\cdot}27.$$

Hence the equation of the curve is

$$v=3+0{\cdot}85u-{\cdot}027u^2$$

or
$$y-7=3+0{\cdot}85(x-5)-0{\cdot}27(x-5)^2$$

or
$$y=-1+3{\cdot}55x-0{\cdot}27x^2$$

Ex. Show that if the pair (x_i, y_i), $i=1, 2, \ldots\ldots n$ occurs with the frequency f_i where $\Sigma f_i=N$, the normal equations, for fitting the parabola of the second degree $y=a_0+a_1x+a_2x^2$ are

$$\Sigma f_iy_i = Na_0+a_1\Sigma f_ix_i+a_2\Sigma f_ix_i^2$$
$$\Sigma f_ix_iy_i = a_0\Sigma f_ix_i+a_1\Sigma f_ix_i^2+a_2\Sigma f_ix_i^3$$
$$\Sigma f_ix_i^2y_i = a_0\Sigma f_ix_i^2+a_1\Sigma f_ix_i^3+a_2\Sigma f_ix_i^4$$

9·3. Curves reducible to st. lines. Fitting of the Curves of the type $y=ab^x$ and $y=ax^b$. Whenever it is required to fit curves of these types to the data, the procedure is to take common logarithms of both sides of the equation and then form normal equations to find the best values of the constants. For example, to fit the curve $y=ab^x$ we proceed as follows :

We have $\log y=A+Bx$

where $A=\log a,\ B=\log b.$

The normal equations are

$$\Sigma \log y=nA+B\Sigma x$$

$$\Sigma x \log y=A\Sigma x+B\Sigma x^2$$

The equations can be solved for A and B and then by taking the antilogs, a and b can be found out.

Example 9·5. *Fit the curve $y=ae^{bx}$ to the following data, e being Napierian base, 2·71828* :

x :	0	2	4
y :	5·012	10	31·62

The equation to the curve can be written as

$$\log y=\log a+(b \log e)x$$

or $Y=A+Bx$

where $Y=\log_{10} y,\ A=\log_{10} a,\ B=b \log_{10} e$

The normal equations are

$$nA+B\Sigma x=\Sigma Y$$

$$A\Sigma x+B\Sigma x^2=\Sigma xY$$

The calculations of Σx, ΣY, Σx^2 and ΣxY are shown in the following table :

	x	y	$Y(=\log y)$	x^2	xY
	0	5·012	0·7	0	0
	2	10·00	1·0	4	2
	4	31·62	1·5	16	6
Total	6		3·2	20	8

Thus we have

$$3A+6B=3·2$$

and $$6A+20B=8$$

Solving, we get

$$A=0{\cdot}666, \qquad B=0{\cdot}2$$

$$\therefore \qquad a=4{\cdot}642, \qquad b=0{\cdot}46 \text{ approx.}$$

Hence equation of the curve fitted to the data is

$$y=4{\cdot}642e^{0{\cdot}46x}.$$

9·4. Fitting of Logistic curve of the form

$$y=\frac{a}{1+bc^x}$$

The given equations can be written as

$$\frac{1}{y}=\frac{1}{a}+\frac{b}{a}\,c^x \quad \Rightarrow \quad \frac{1}{y}-\frac{1}{a}=\frac{b}{a}\,c^x$$

$$\Rightarrow \qquad Y=AB^x$$

where $$Y=\frac{1}{y}-\frac{1}{a},\ A=\frac{b}{a},\ B=c.$$

Such curves are widely used in the study of growth.

EXERCISES 9

1. (*a*) Write a short note on 'Method of Least Squares'.

(*b*) Given a set of m linear equations involving n unknowns $(m>n)$, show how the most probable values of the unknown are determined.

2. Find the most plausible values of x, y and z from the following equations :—

(*a*) $x-y+2z=3,\ 3x+2y-5z=5,\ 4x+y+4z=21,\ -x+3y+3z=14.$

(*B.A. Panjab 1956*)

(*b*) $x+2y+z=1,\ 2x+y+z=4,\ -x+y+2z=4,\ 4x+2y-5z=-7.$

(*M.A. Punjab 1945*)

[**Ans.** (*a*) $x=2{\cdot}47$ (*b*) $x=1{\cdot}16$
$y=3{\cdot}55$ $y=-0{\cdot}76$
$z=1{\cdot}92$ $z=2{\cdot}80$]

3. Fit a straight line to the following points :—

x :	0	1	2	3	4
y :	1	1·8	3·3	4·5	6·3

(*M.Sc., Agra, 1959*)

[**Ans.** $y=0{\cdot}72+1{\cdot}33x$]

4. Fit a parabola of second degree to the following data :

x :	1·0	1·5	2·0	2·5	3·0	3·5	4·0
y :	1·1	1·3	1·6	2·6	2·7	3·4	4·1

[**Hint.** Take 2·5 as the working origin for x series.]

[**Ans.** $y=1{\cdot}04-0{\cdot}20x+0{\cdot}24x^2$]

5. The profit of a certain company in the xth year of its life are given by

x :	1	2	3	4	5
y :	1250	1400	1650	1950	2300

Taking $u=x-3$ and $v=(y-1650)/50$ show that the parabola of the second degree of v on u is

$$v+0{\cdot}086=5{\cdot}30u+0{\cdot}64u^2$$

and deduce that the parabola of second degree of y on x is

$$y=1140\times72x+32\ 15x^2.$$

(*B.A. Hons., Delhi, 1954* ; *M.Sc., Agra*)

6. Fit a parabola of the second degree to the following data taking x as the independent variable ;—

x :	0	1	2	3	4
y :	1	1·8	1·3	2·5	6·3

Find out the difference between the actual value of y and the value of y obtained from the fitted curve when $x=2$.

(*M.Sc., Agra, 1956* ; *B.Sc., Lucknow, 1952*)

[**Ans.** $y=1{\cdot}48+1{\cdot}13\ (x-2)+0{\cdot}55\ (x-2)^2$, Difference$=-{\cdot}18$]

7. The pressure (p) and volume (v) of a gas are known to be related by an equation of the form $pv^k=$constant.

Indicate how you would proceed to find the value of k by fitting a straight line to set of observed values of p and v, taking p to be the independent variable.

(*B.A. Hons., Delhi, 1949*)

8. (*a*) Derive the least square equations for fitting a curve of the type (*i*) $y=ax+bx^{-1}$ (*ii*) $y=ax^b$ to a set of n points. (*B.A. Hons., Delhi, 1958*)

(*b*) It is known that the readings for x and y given below should follow a law of the form $y=ax+bx^{-1}$ when a and b are constants.

x :	1	2	3	4	5	6	7	8
y :	5·43	6·28	8·23	10·32	12·63	14·86	17·27	19·51

Use the method of least squares to find the least values of a and b.

(*I.S.I., 1968*)

[**Ans.** $a=2{\cdot}40$, $b=3{\cdot}03$]

9. Write down normal equations for fitting $y=ae^{bx}$ to a set of points.

Fit the curve $y=ae^{bx}$ to the following data :

x :	1	2	3	4	5	6	7	8
y :	15·3	20·5	27·4	36·6	49·1	65·6	87·8	117·6

[**Ans.** $y=11{\cdot}58\ e^{-{\cdot}2898x}$]

10. Use the method of least squares to determine a and b in the formula $y=ax+bx^2$ for the following observations :

x :	1	2	3	4	5
y :	1·8	5·1	8·9	14·1	19·8

Calculate the value of y when $x=2$.

[**Ans.** $a=1{\cdot}521$, $b=0{\cdot}49$; $y=5{\cdot}006$ when $x=2$]

11. For a given bivariate distributions, find the straight line for which the sum of squares of the normal deviations is a minimum.

[**Hint.** Let the line be $x\cos\alpha+y\sin\alpha-p=0$. Then

$$S=\Sigma f_i(x_i\cos\alpha+y_i\sin\alpha-p)^2.$$

Normal equations are $\frac{\partial S}{\partial \alpha}=0$ and $\frac{\partial S}{\partial p}=0$. These determine p and α. The first of these equations shows that the line passes through the mean of the distribution. Taking the mean as origin, the second equation gives α where

$$\tan 2\alpha=2\mu_{11}/(\sigma_x^2-\sigma_y^2)]$$

12. Show that the parabola of best fit for the points

$$(x_1, y_1)\ ;\ (x_2, y_2)\ldots\ldots(x_{2n+1}, y_{2n+1})$$

where the values of x are in A.P. with common difference unity and $\bar{x}=0$, can be expressed in the form

$$\begin{vmatrix} y & x^2 & x & 1 \\ \Sigma y_i & \frac{n(n+1)(2n+1)}{3} & 0 & 2n+1 \\ \Sigma x_i y_i & 0 & \frac{n(n+1)(2n+1)}{3} & 0 \\ \Sigma x_i^2 y_i & \frac{n^2(n+1)^2}{2} & 0 & \frac{n(n+1)(2n+1)}{3} \end{vmatrix} =0$$

13. When do we prefer logarithmic curve to ordinary curve ?

10

Correlation and Regression

10·1. Bivariate Universes. Correlation Coefficient. In earlier chapters, we have been mainly concerned with univariate universes only. We shall now discuss bivariate universes and in particular study the simultaneous variation of two variables, say X and Y. The variables may, for example, be heights and weights of students in a class, ages of husbands and wives, or records of rainfall and yields of crops. The relationship between the variates may be of any type but the relationship which is of great interest in Statistics is of the linear type.

Just as $E(x-\bar{x})^2$ gives us a measure for the variation in x and $E(y-\bar{y}^2)$ gives us a similar measure for the variation in y, we may expect $E(x-\bar{x})(y-\bar{y})$ to give us a measure for the simultaneous variations of x and y. But this will depend on the units used for x and y. To make it independent of these units, we should divide it by a quantity having the same dimension as $E(x-\bar{x})(y-\bar{y})$. Keeping this in view, Karl Pearson defined in 1890 the measure r_{xy} defined by

$$r_{xy}=\frac{E(x-\bar{x})(y-\bar{y})}{\sqrt{E(x-\bar{x})^2}\sqrt{E(y-\bar{y})^2}}=\frac{\text{Cov }(x,\ y)}{\sqrt{\text{var }(x).\ \text{var }(y)}}=\frac{\mu_{11}}{\sigma_x\sigma_y} \quad \ldots(10\cdot1)$$

is called the *product moment correlation coefficient* or simply the *correlation coefficient.* By symmetry it is clear that $r_{xy}=r_{yx}$. Further since σ_x and σ_y are essentially positive the correlation coefficient between x and y has the same sign as the covariance μ_{11} of the two variates.

Example 10·1. *Two independent variates x_1 and x_2 have means 5 and 10 and variances 4 and 9 respectively. Obtain the correlation coefficient between $y_1=3x_1+4x_2$ and $y_2=3x_1-x_2$.*

(B.A., Punjab, 1953, 54)

We have

$$E(y_1)=3E(x_1)+4E(x_2),\ E(y_2)=3E(x_1)-E(x_2)$$

Hence $y_1-E(y_1)=3[x_1-E(x_1)]+4[x_2-E(x_2)]$

and $y_2-E(y_2)=3[x_1-E(x_1)]-[x_2-E(x_2)]$

$$\therefore\quad E[y_1-\bar{y}_1][y_2-\bar{y}_2]=9E[(x_1-\bar{x}_1)^2]-4E[(x_2-\bar{x}_2)^2]+9E[(x_1-\bar{x}_1)(x_2-\bar{x}_2)] \qquad \ldots(1)$$

$$=9\ Var\ (x_1)-4\ Var\ (x_2)=9\times4-4\times9=0$$

The last term on the right of (1) vanishes since x_1 and x_2 are independent variates.

Hence $r_{y_1y_2}=\dfrac{E(y_1-\bar{y}_1)(y_2-\bar{y}_2)}{\sigma_{y_1}\sigma_{y_2}}=\dfrac{0}{\sigma_{y_1}\sigma_{y_2}}=0.$

Note. It is obvious that the answer is independent of the values which the means $\bar{x}_1$ and $\bar{x}_2$ take.

10·1·1. Change of Origin and Scale. Let u and v be two new variates defined by $u=\dfrac{x-a}{h},\quad v=\dfrac{y-b}{k}$

so that $x=a+hu\ ;\ \bar{x}=a+h\bar{u}\ ;\ y=b+kv\ ;\ \bar{y}=b+k\bar{v}.$

Substituting in the expression for r_{xy}, we get

$$r_{xy}=\frac{E(x-\bar{x})(y-\bar{y})}{\sqrt{E(x-\bar{x})^2}\sqrt{E(y-\bar{y})^2}}=\frac{hkE(u-\bar{u})(v-\bar{v})}{\sqrt{h^2k^2}\sqrt{E(u-\bar{u})^2}\sqrt{E(v-\bar{v})^2}}$$

$$=\frac{hk}{\sqrt{h^2k^2}}r_{uv}$$

so that if h and k are both positive (they are positive if we change the units of measurement only), we have

$$r_{xy}=r_{uv} \qquad \ldots(10\cdot2)$$

Thus r_{uv} is independent of a, b (corresponding to change of origin) and h, k (corresponding to change of scale).

Thus the value of the correlation coefficient is independent of the origin of reference and the unit of measurement. In other words, r is a pure number. This property is quite useful in numerical evaluation of r since due to this property, we can choose any convenient origin and scale.

Ex. If h and k are of opposite sign, show that $r_{uv}=-r_{xy}$.

10·1·2. Theorem. *If x and y are independent variates, they are uncorrelated.*

Since x and y are independent, we have

$$\text{Cov } (x, y)=E(x-\bar{x})(y-\bar{y})=E(x-\bar{x})E(y-\bar{y})$$
$$=[E(x)-E(\bar{x})][E(y)-E(\bar{y})]=0$$

$\therefore$ $r=0$. Hence the theorem.

The converse of this theorem is not necessarily true

i.e., variates may be uncorrelated but dependent.

For example, if x is a variate with a constant density function, say

$$f(x)=\tfrac{1}{2}, \qquad -1\leqslant x\leqslant 1$$

and if $y=x^2$,

then $$E(x)=\int_{-1}^{+1} x\,.\,f(x)dx=\int_{-1}^{+1}\tfrac{1}{2}x dx=0.$$

so that $E(x)E(y)=0.E(y)=0$

Furthermore, $$E(xy)=E(x^3)=\int_{-1}^{+1} x^3\,.\,\tfrac{1}{2}dx=0$$

Hence $$\text{Cov } (x, y)=E(x-\bar{x})(y-\bar{y})=E(xy)-E(x)E(y)$$
$$=0-0=0.$$

Thus x and y are uncorrelated. However, for each value of x, there is only one possible value for y ; and for each value of y, there are only two possible values of x. Therefore, x and y are far from being independent.

Note. For other examples of this type, see **Ex.** 3, page 307.

10·1·3. Limits for the Coefficient of Correlation. We shall show that $|\,r\,|\leqslant 1$, that is, $-1\leqslant r\leqslant 1$.

Lemma. (*Schwarz's Inequality*). *For any stochastic variates x and y,*

$$[E(xy)]^2\leqslant E(x^2)E(y^2).$$

Proof of the Lemma. For every real constant a, the variate $(ax-y)^2$ is always non-negative and therefore has a non-negative expectation.

That is $$E(ax-y)^2\geqslant 0$$

or $$a^2E(x^2)-2aE(xy)+E(y^2)\geqslant 0 \text{ for every real } a.$$

In particular this holds for $a=\dfrac{E(xy)}{E(x^2)}$. Substituting this value of a, we have

$$\frac{[E(xy)]^2}{E(x^2)}-2\frac{[E(xy)]^2}{E(x^2)}+E(y^2)\geqslant 0.$$

The result now follows if we multiply by $E(x^2)$ and transpose.

Proof of theorem. If x' , y' denote deviations of the variates x and y from their respective means, we have by the Lemma,

$$[E(x'y')]^2 \leqslant E(x'^2)E(y'^2)$$

Dividing both sides by $E(x'^2)E(y'^2)$, we get

$$r^2 \leqslant 1 \qquad \text{or} \qquad -1 \leqslant r \leqslant 1.$$

Note. For other proofs of this important result, see Ex. 1, page 306 and § 10·4, page 291.

10·2. Bivariate Frequency Distribution. Let the variate value pair (x_i, y_i) occur with frequency f_i, $i=1, 2.....n$, $\Sigma f_i = N$.

Then for a bivariate frequency distribution we define

$$r_{xy} = \frac{\Sigma f_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma f_i(x_i-\bar{x})^2}\sqrt{\Sigma f_i(y_i-\bar{y})^2}}$$

$$= \frac{\Sigma f_i x_i y_i - (1/N)(\Sigma f_i x_i)(\Sigma f_i y_i)}{\sqrt{\Sigma f_i x_i^2 - (1/N)(\Sigma f_i x_i)^2}\sqrt{\Sigma f_i y_i^2 - (1/N)(\Sigma f_i y_i)^2}} \quad ...(10·3)$$

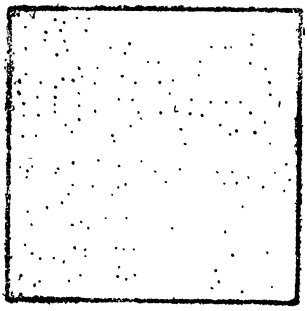
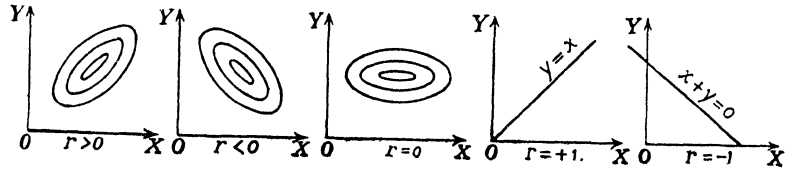


Fig. 29. Scatter diagram.

The points (x_i, y_i) can be represented in xy-plane to give what is called a *scatter* or *dot diagram* (see Fig. 29). This diagram is useful in giving an intuitive idea of the relationship between x and y but for a quantitative precise measure, we use the measure r.

Scatter diagrams for various values of r



Example 10·2 *Calculate from the data reproduced below pertaining to 66 selected villages in Meerut district, the value of r, between 'total cultivable area' and 'the area under wheat'.*

Total Cultivable Area (*in Bighas*)

Area under wheat (in Bighas)	0–500	500–1000	1000–1500	1500–2000	2000–2500	Total
0–200	12	6	—	—	—	18
200–400	2	18	4	2	1	27
400–600	—	4	7	3	—	14
600–800	—	1	—	2	1	4
800–1000	—	—	—	1	2	3
Total	14	29	11	8	4	66

(*B.A., Punjab, I.A.S. 1949*)

Let us introduce two new variables u and v given by the relations

$$u=\frac{x-1250}{500}, \qquad v=\frac{y-500}{200}.$$

x:			250	750	1250	1750	2250				
			−1000	−500	0	+500	+1000	Total f	fv	fv^2	fuv
y:		u / v	−2	−1	0	+1	+2				
100	−400	−2	12 (48)	6 (12)	—	—	—	18	−36	72	60
300	−200	−1	2 (4)	18 (18)	4 (0)	2 (−2)	−2	27	−27	27	18
500	0	0	—	4 (0)	7 (0)	3 (0)	—	14	0	0	0
700	200	1	—	1 (−1)	—	2 (2)	1 (2)	4	4	4	3
900	400	2	—	—	—	1 (2)	2 (2)	3	6	12	10
Total f			14	29	11	8	4	66	−53	115	91
fu			−28	−29	0	8	8	−41			
fu^2			56	29	0	8	16	109			
fuv			52	29	0	2	8	91			

We have $\Sigma fuv = 12\times(-2)\times(-2)+6\times(-1)(-2)+\ldots\ldots$
$= 48+12+\ldots\ldots = 91$

Now
$$r_{xy}=r_{uv}=\frac{\Sigma fuv-\frac{1}{n}(\Sigma fu)(\Sigma fv)}{\sqrt{\Sigma fu^2-\frac{1}{n}(\Sigma fu)^2}\sqrt{\Sigma fv^2-\frac{1}{n}(\Sigma fv)^2}}$$
$$=\frac{91-\frac{1}{66}(-41)(-53)}{\sqrt{109-\frac{(-41)^2}{66}}\sqrt{115-\frac{(-53)^2}{66}}}$$
$$=\frac{91\times66-41\times53}{\sqrt{109\times66-(-41)^2}\sqrt{115\times66-(-53)^2}}$$
$$=\frac{6006-2173}{\sqrt{7194-1681}\sqrt{7590-2809}}=+\cdot749.$$

Thus the coefficient of correlation between 'total cultivable area' and the 'area under wheat' is $+\cdot749$.

10 3 1. Linear Regression. In case there exists association or relationship between two variates x and y, the dots of the scatter diagram will be more or less concentrated round a curve which may be called the *curve* of *regression* and the relationship is said to be expressed by means of *curvilinear regression*. In the particular case when the curve is a straight line, it is called the *line of regression* and the regression is said to be linear. More precisely, the line of regression is the straight line which gives the 'best' fit in the least square sense to the given frequency or probability distribution. If the straight line is so chosen that the sum of squares of deviations parallel to the axis of y is minimized it is called the line of regression of y on x and it gives the *best estimates of y for any given value of x*. If, on the other hand, the sum of squares of the deviations parallel to the x-axis is minimized, the resulting straight line is the line of regression of x on y and it gives the best estimate of x for any given value of y. The two lines of regression would obviously, in general, be different.

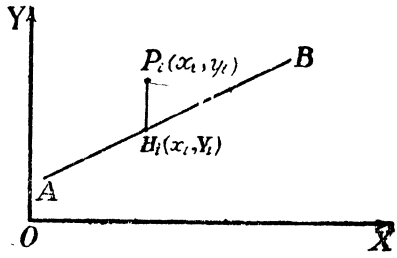


Fig. 30 (*a*)

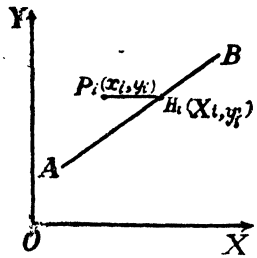


Fig. 30

10·3·2. Equations to the Lines of Regression. Regression Coefficients.

For any value of Y, say y, the distribution of X is called a x-array which is, of course, conditional distribution of X given that $Y=y$. The mean of the conditional distribution is

$$\bar{x}_y=E(X \mid Y=y) \qquad \ldots(A)$$

Similarly, for any value of X, say x the distribution of Y is called the conditional distribution of Y given that $X=x$. The mean of the distribution is given by

$$\bar{y}_x=E(Y \mid X), \qquad \ldots(B)$$

The equations (A) and (B) are called the regression curves of X on Y and of Y on X respectively. However, if the regression of Y on X (and of X on Y) is linear, then we have

$$E(X \mid Y=y)=a_1+b_1y,$$

where we are to determine the constants a_1 and b_1.

Taking expectation on both sides with respect to y, we have

$$EE(X \mid Y=y)=a_1+b_1E(Y)$$

or
$$E(X)=a_1+b_1E(Y)$$

$$\therefore E(X \mid Y=y)-E(X)=b_1[Y-E(Y)]$$

which is the linear regression of Y on X. The constant b_1 is determined as follows :

Multiply both sides by $Y-E(Y)$ and take expectations again. We have

$$E[\{E(X \mid Y=y)-E(X)\}\{Y-E(Y)\}]=b_1E[Y-E(Y)]^2$$

Hence
$$b_1=\frac{\mu_{11}}{\text{var } y}$$

Similarly, the linear regression equation of X on Y is

$$E(y \mid x)-E(y)=b_2[x-E(x)]$$

where
$$b_2=\frac{\mu_{11}}{\text{var } x}$$

b_1, b_2 are called the regression coefficients of Y on X and of X on Y respectively.

Alternatively. Let $Y=a+bx$ be the equation of the line of regression of Y on X, so that corresponding to $x=x_i$, the expected

value of Y is Y_i $(=a+bx_i)$ and S, the sum of the squares of deviations of observed value y_i from the expected value Y_i, is given by

$$S=\sum_{i=1}^{n} f_i(y_i-Y_i)^2=\sum_{i=1}^{n} f_i(y_i-a-bx_i)^2.$$

According to the principle of least squares, we have to choose a and b so that S is minimum. The normal equations are :

$$\frac{\partial S}{\partial a}=0, \quad \frac{\partial S}{\partial b}=0,$$

that is, $$\Sigma f_i(y_i-a-bx_i)=0 \qquad ...(10\cdot4)$$

and $$\Sigma f_i x_i(y_i-a-bx_i)=0 \qquad ...(10\cdot5)$$

Since $\Sigma f_i=N$, $\Sigma f_i x_i=N\bar{x}$, $\Sigma f_i y_i=N\bar{y}$, $\Sigma f_i x_i^2=N(\sigma_x^2+\bar{x}^2)$,

$$\Sigma f_i^2=N(\sigma_y^2+\bar{y}_i^2),\ \Sigma f_i x_i y_i=N(\mu_{11}+\bar{x}\bar{y}), \qquad ...(10\cdot6)$$

the equations (10·4) and (10·5) reduce respectively to

$$\bar{y}=a+b\bar{x} \qquad ...(10\cdot7)$$

and $$\mu_{11}+\bar{x}\bar{y}=a\bar{x}+b(\bar{x}^2+\sigma_x^2) \qquad ...(10\cdot8)$$

Multiplying (10·7) by $\bar{x}$ and subtracting from (10·8), we get

$$b=\frac{\mu_{11}}{\sigma_x^2} \qquad ...(10\cdot9)$$

and then (10·7) gives

$$a=\bar{y}-\frac{\mu_{11}}{\sigma_x^2}\,\bar{x},$$

so that the equation to the line of regression of y on x is

$$y-\bar{y}=\frac{\mu_{11}}{\sigma_x^2}(x-\bar{x}) \qquad ...(10\cdot10)$$

Similarly, the equation to the line of regression of x on y is

$$x-\bar{x}=\frac{\mu_{11}}{\sigma_y^2}(y-\bar{y}) \qquad ...(10\cdot11)$$

It is obvious that both the straight lines pass through the mean point $(\bar{x}, \bar{y})$. The coefficients μ_{11}/σ_x^2 and μ_{11}/σ_y^2 are called the *regression coefficients* of y on x, and of x on y respectively and give estimate of changes of y and x corresponding to unit changes in x and y respectively. They are usually denoted by b_{yx} and b_{xy} respectively, and unlike the correlation coefficient these are not symmetrical in x and y. From (10·9), using the definition of r the correlation coefficient, we get

$$b_{yx}=\frac{\mu_{11}}{\sigma_x^2}=\frac{r\sigma_x\sigma_y}{\sigma_x^2}=\frac{r\sigma_y}{\sigma_x}. \qquad ...(10\cdot12)$$

Similarly $$b_{xy}=\frac{r\sigma_x}{\sigma_y} \qquad \ldots(10\cdot13)$$

Hence $$b_{yx}\times b_{xy}=r^2$$

$$\Rightarrow \qquad r=\pm\sqrt{b_{yx}\times b_{xy}}$$

Thus, *the correlation coefficient is the geometric mean between the regression coefficients. The sign of r is the same as that of any regression coefficient.*

Notes. 1. Since σ_x, σ_y *are always positive,* b_{yx}, b_{xy}, μ_{11} *and* r *have the same sign.*

2. A slightly different way of deriving (10·10) is outlined in Ex. 5, page 292.

3. *One of the two regression coefficients must be less than unity and the other greater than unity or both will be unity.*

For $$b_{yx}\,.\,b_{xy}=r^2\leqslant 1 \text{ or } b_{yx}\leqslant\frac{1}{b_{yx}}$$

4. *The regression coefficients are independent of the change of origin, but not of scale.*

For if $u=\dfrac{x-a}{h}$ and $v=\dfrac{y-b}{k}$ then

$$b_{yx}=\frac{r\sigma_y}{\sigma_x}=r\,.\,\frac{k\sigma_v}{h\sigma_u}=\frac{k}{h}\left(\frac{r\sigma_v}{\sigma_u}\right)=\frac{k}{h}\,b_{vu}$$

Similarly $$b_{xy}=\frac{k}{h}b_{uv}$$

5. *Arithmetic mean of the regression coefficients is greater than the correlation coefficient provided that correlation coefficient is positive.*

For if $$\tfrac{1}{2}\left(\frac{r\sigma_y}{\sigma_x}+\frac{r\sigma_x}{\sigma_y}\right)\geqslant r \quad \text{or } \sigma_y^2+\sigma_x^2\geqslant 2\sigma_x\sigma_y,\ r>0$$

then $(\sigma_y-\sigma_x)^2\geqslant 0$ which is always true.

Example 10·3. *The following marks have been obtained by a class of students in statistics (out of 100).*

Paper I :	45	55	56	58	60	65	68	70	75	80	85
Paper II :	56	50	48	60	62	64	65	70	74	82	90

Compute the coefficient of correlation for the above data. Find also the equations of the lines of regression.

(B.Sc., Agra ; Indian Audit, 1945)

Calculation of the coefficient of correlation between marks obtained by a class of students in statistics paper I and statistics paper II.

Paper I			*Paper II*			
Marks out of 100 x	*Deviation from assumed average 65* u	*Square of deviation* u^2	*Marks out of 100* y	*Deviation from assumed average 70* v	*Square of deviation* v^2	*Product of deviation* uv
45	−20	400	56	−14	196	280
55	−10	100	50	−20	400	200
56	−9	81	48	−22	484	198
58	−7	49	60	−10	100	70
60	−5	25	62	− 8	64	40
65	0	0	64	− 6	36	0
68	3	9	65	− 5	25	−15
70	5	25	70	0	0	0
75	10	100	74	4	16	40
80	15	225	82	12	144	180
85	20	400	90	20	400	400
$n=11$	$\Sigma u=+2$	$\Sigma u^2=1414$	$n=11$	$\Sigma v=-49$	$\Sigma v^2=1865$	$\Sigma uv=1393$

Substituting the above values in Pearson's formula, we get

$$r=\frac{1393-\frac{2\times-49}{11}}{\sqrt{1414-\frac{(2)^2}{11}}\sqrt{1865-\frac{(49)^2}{11}}}$$

$$=\frac{1393\times11+2\times49}{\sqrt{1414\times11+2^2}\sqrt{1865\times11-(-40)^2}}$$

$$=\frac{15323+98}{\sqrt{15554-4}\sqrt{20515-2401}}=\frac{15421}{\sqrt{15550\times18114}}$$

$$=\cdot918.$$

Regression coefficient of x on $y = b_{xy} = r\frac{\sigma_x}{\sigma_y} = r\frac{\sigma_u}{\sigma_v}$

$$= \frac{\Sigma uv - \frac{(\Sigma u)(\Sigma v)}{n}}{\Sigma v^2 - \frac{(\Sigma v)^2}{n}}$$

$$= \frac{1393 - \frac{2 \times 49}{11}}{1865 - \frac{(-49)^2}{11}}$$

$$= \frac{1393 + 8{\cdot}9}{1865 - 218{\cdot}3} = \frac{1401{\cdot}9}{1646{\cdot}7} = {\cdot}85$$

Regression coefficient of y on $x = b_{yx} = r\frac{\sigma_y}{\sigma_x} = r\frac{\sigma_v}{\sigma_u}$

$$= \frac{\Sigma uv - \frac{(\Sigma u)(\Sigma v)}{n}}{\Sigma u^2 - \frac{(\Sigma u)^2}{n}}$$

$$= \frac{1393 - \frac{2 \times -49}{11}}{1414 - \frac{(2)^2}{11}}$$

$$= \frac{1393 + 8{\cdot}9}{1414 - 0{\cdot}4} = \frac{1401{\cdot}9}{1413{\cdot}6} = 0{\cdot}99$$

Arithmetic average of the marks of the 1st paper.

$$\bar{x} = \text{assumed average} + \frac{\Sigma u}{n}$$

$$= 65 + \frac{2}{11} = 65{\cdot}2 \text{ marks.}$$

Arithmetic average of the marks of the 2nd paper.

$$\bar{y} = \text{assumed average} + \frac{\Sigma v}{n}$$

$$= 70 + \frac{-49}{11} = 65{\cdot}55 \text{ marks.}$$

Substituting the above values in the following regression equation of x and y.

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

we get $(x - 65{\cdot}2) = {\cdot}85(y - 65{\cdot}55)$

or $x = {\cdot}85y - 55{\cdot}72 + 65{\cdot}2 = {\cdot}85y + 9{\cdot}48$

Similarly, the equation of the line of regression of y on x is

$$y-65{\cdot}55=0{\cdot}99(x-65{\cdot}2)$$

or
$$y={\cdot}99x+1.$$

Example 10·4. *Show that θ, the acute angle between the two lines of regression, is given by*

$$\tan\theta=\frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}.$$

Interpret the case when $r=0$, $r=\pm 1$.

(M.Sc., Agra, 1955, 59 ; B.A. Hons. Delhi, 1948)

From (10·10) and (10·11) we find that the slopes of the two lines of regression are respectively $\frac{r\sigma_y}{\sigma_x}$ and $\frac{\sigma_y}{r\sigma_x}$.

Hence
$$\tan\theta=\frac{\frac{r\sigma_y}{\sigma_x}\sim\frac{\sigma_y}{r\sigma_x}}{1+\frac{r\sigma_y}{\sigma_x}\cdot\frac{\sigma_y}{r\sigma_x}}$$

$$=\frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}, \qquad (\because r^2\leqslant 1)$$

When $r=\pm 1$, $\tan\theta=0$ so that $\theta=0$ or π.

Hence the two lines of regression coincide and there is a perfect correlation between the two variates x and y.

When $r=0$, the two lines of regression are perpendicular. The estimated value of y (or x) is the same for all values of x (or y).

Ex. 1. Obtain (10·11) from the *first principles.*

Ex. 2. Construct three examples in which $r=+1$, 0, -1 respectively.

Ex. 3. For a continuous probability distribution with density function $f(x, y)$, choose a and b so as to minimize

$$\iint (y-a-bx)^2 f(x, y)dxdy.$$

Deduce the equations of the two lines of regression.

Ex. 4. Obtain the equations of the lines of regression for a discrete probability distribution.

Ex. 5. If the origin is taken at the meanpoint $(\bar{x}, \bar{y})$, show that (10·7) and (10·8) give $a=0$, $b=\mu_{11}/\sigma_x^2$. Hence deduce (10·10).

Ex. 6. If $r=0$, show that the two lines of regression are parallel to the axes.

Ex. 7. Prove the following results :

(*i*) $\Sigma f_i Y_i(y_i-Y_i)=0$

(*ii*) $\Sigma f_i(y_i-Y_i)(Y_i-\bar{y})=0$

(*iii*) $\sigma_y=|r|\sigma_y$

where y_i is the observed value and Y_i is the corresponding least square estimate for any given x.

10·4. Standard Error of Estimate. The sum of the squares of the deviations of points from the line of regression of y on x is by using (10·7) and (10·8)

$$NS_y^2=\Sigma f_i(y_i-a-bx_i)^2=\Sigma f_i\left(y_i'-\frac{r\sigma_y}{\sigma_x}x_i'\right)^2$$

$$=\Sigma f_i y_i'^2-2r\frac{\sigma_y}{\sigma_x}.\Sigma f_i x_i'\,y_i'+\frac{r^2\sigma_y^2}{\sigma_x^2}.\Sigma f_i x_i'^2$$

$$=N\left(\sigma_y^2-2r\frac{\sigma_y}{\sigma_x}.r\sigma_x\sigma_y+r^2\sigma_y^2\right)$$

or $$S_y^2=\sigma_y^2(1-r^2)$$

or $$S_y=\sigma_y(1-r^2)^{\frac{1}{2}} \qquad ...(10·14)$$

This value of S_y is sometimes called the root-mean square error of estimate of y or more briefly the standard error of estimate of y. Similarly, the standard error of estimate of x is given by

$$S_x=\sigma_x(1-r^2)^{\frac{1}{2}} \qquad ...(10·15)$$

Since the sum of squares of deviations is non-negative, it follows that $r^2\leqslant 1$ or $-1\leqslant r\leqslant 1$.

If $r=1$ or -1, the sum of squares of deviations from either line of regression is zero. Consequently each deviation is zero, and all the points lie on both the lines of regression. These two lines then coincide and there is a linear functional relation between the variates x and y, giving perfect correlation. As r^2 comes nearer to unity, S_x^2 and S_y^2 approach zero so that the points are closer to the lines of regression which are nearer to coincidence. Thus *the departure of the value of r^2 from unity is a measure of departure of the relationship between two variates from linearity.*

Ex. 1. For continuous probability distribution, prove that

$$\int\int\left[(y-\bar{y})-\frac{\mu_{11}}{\sigma_x^2}(x-\bar{x})\right]^2 f(x,y)dxdy=\sigma_y^2(1-r^2).$$

Ex. 2. Find the correlation coefficient, r, given

x:	1	2	3	4
y:	1	4	9	16

Explain why r^2 is not unity though there is a functional relation $y=x^3$ between x and y. [Ans. $r=+·98$]

Example 10·5. *Establish the formula* $nr\sigma_x\sigma_y=n_1r_1\sigma_{x_1}\sigma_{y_1}+n_2r_2\sigma_{x_2}\sigma_{y_2}+n_1d_{x_1}d_{y_1}+n_2d_{x_2}d_{y_2}$ *where* n_1, n_2 *and* n *are respectively the*

sizes of the first, second and combined sample, $(\bar{x}_1, \bar{y}_1)$, $(\bar{x}_2, \bar{y}_2)$, $(\bar{x}, \bar{y})$ *their means,* r_1, r_2 *and* r *their coefficients of correlation,* $(\sigma_{x_1}, \sigma_{y_1})$, $(\sigma_{x_2}, \sigma_{y_2})$, (σ_x, σ_y) *their standard deviations and* $d_{x_1} = \bar{x}_1 - \bar{x}, dy_1 = \bar{y}_1 - \bar{y}$, $d_{x_2} = \bar{x}_2 - \bar{x}$, $d_{y_2} = \bar{y}_2 - \bar{y}$. *(B.Sc., Agra 1964)*

By Ex. 3·4, page 33 we have

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}, \quad \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

$$n\sigma_x^2 = n_1(\sigma_{x_1}^2 + d_{x_1}^2) + n_2(\sigma_{x_2}^2 + d_{x_2}^2)$$

$$n\sigma_y^2 = n_2(\sigma_{y_1}^2 + d_{y_1}^2) + n_2(\sigma_2^2 + dy_2^2)$$

$$r_1 = \frac{\text{Cov}\,(x_1, y_1)}{n_1\sigma_{x_1}\sigma_{y_1}}, \quad r_2 = \frac{\text{Cov}\,(x_2, y_2)}{n_2\sigma_{x_2}\sigma_{y_2}}$$

For the pooled sample, we have

$$r = \frac{\text{Cov}\,(x, y)}{n\sigma_x\sigma_y}$$

$$\Rightarrow \quad nr\sigma_x\sigma_y = \sum_1^n (x-\bar{x})(y-\bar{y}) = \sum_1^{n_1} (x-\bar{x})(y-\bar{y}) + \sum_{n_1+1}^{n} (x-\bar{x})(y-\bar{y})$$

Now $\sum_1^{n_1} (x-\bar{x})(y-\bar{y}) = \sum_1^{n_1} [(x-\bar{x}_1) + (\bar{x}_1-\bar{x})][y-\bar{y}_1) + (\bar{y}_1-\bar{y})]$

$$= \sum_1^{n_1} (x-\bar{x}_1)(y-\bar{y}_1) + n_1(\bar{x}_1-\bar{x})(\bar{y}_1-\bar{y})$$

$$= n_1 r_1 \sigma_{x_1}\sigma_{y_1} + n_1 d_{x_1} d_{y_1} \text{ since } \sum_1^n (x-\bar{x}_1) = \sum^n dx_1 - x$$

Similarly $\sum_{n_1+1}^{n} (x-\bar{x})(y-\bar{y}) = n_2 r_2 \sigma_{x_2}\sigma_{y_2} + n_2 d_{x_2} d_{y_2}$

Hence on substituting these values in (1) we get the required result.

10·5. Correlation of Ranks. If we have a group of individuals ranked according to different characters, it is natural to enquire whether the ranks can be made to give us the measure of degree of relation between the two characters. Suppose we have n individuals whose ranks according to character A are $x_1, x_2, \ldots, x_n$, and according to character B are $y_1, y_2, \ldots\ldots, y_n$ where x's and y's are merely the permutations of n numbers 1 to n. Assuming that no two individuals are bracketed or tied in either classifications, both x and y take all integral values from 1 to n.

$$\therefore \qquad \sum_{i=1}^{n} x_i = \frac{n(n+1)}{2} = \sum_{i=1}^{n} y_i.$$

Hence $\bar{x}=\bar{y}$ and $\sum_{i=1}^{n} (x_i-\bar{x})^2 = \frac{1}{12}(n^3-n) = \sum_{i=1}^{n} (y_i-\bar{y}^2)$.

If d_i stands for the difference in ranks of the ith individual, we have

$$d_i = x_i - y_i = (x_i-\bar{x}) - (y_i-\bar{y}) = x_i' - y_i',$$

where x_i', y_i' are the deviations of x_i and y_i from the means $\bar{x}$ and $\bar{y}$.

$$\therefore \qquad \sum_{i=1}^{n} d_i^2 = \Sigma x_i'^2 + \Sigma y_i'^2 - 2\Sigma x_i' y_i'$$

$$\Rightarrow \qquad \Sigma x_i' y_i' = \tfrac{1}{2}\left[\frac{n^3-n}{6} - \sum_{i=1}^{n} d_i^2\right]$$

The coefficient of correlation between x and y is given by

$$r = \frac{\Sigma x_i' y_i'}{\sqrt{\Sigma x_i'^2}\ \sqrt{\Sigma y_i'^2}} = \frac{\frac{1}{2}\left[\frac{n^3-n}{6} - \Sigma d_i^2\right]}{\frac{1}{12}(n^3-n)}$$

$$= 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \ldots(10{\cdot}16)$$

This formula for rank correlation coefficient is due to Spearman.

Note. In case of a 'tie' we add the factor $\frac{1}{12}\, m(m^2-1)$, known as correction factor, to Σd^2 where m is the number of times an item has repeated value.

Example 10·6. *The rankings of ten students in two subjects A and B are as follows :*

A :	3	5	8	4	7	10	2	1	6	9
B :	6	4	9	8	1	2	3	10	5	7

What is the coefficient of rank correlation ? (*M.A., Delhi, 1951*)

The differences d_i are : $-3, 1, -1, -4, 6, 8, -1, -9, 1, 2$.

Hence $\Sigma d_i = 0$, $\Sigma d_i^2 = 214$

$$\therefore \qquad r = 1 - \frac{6 \times 214}{10(100-1)} = -\frac{147}{495} = -0{\cdot}3 \text{ approx.}$$

Ex. 1. If $d_i=0$ for all i, show that $r=1$ and conversely if $r=1$, then $d_i=0$ for all i.

Ex. 2. If $x_i+y_i=n+1$, show that $r=-1$ and conversely. Explain the result.

[**Hint.** $x_i-y_i=d_i$, $x_i+y_i=n+1$ $\quad\therefore\quad d_i=2x_i-(n+1)$,

$\therefore \quad \Sigma d_i^2 = 4\Sigma x_i^2 - 4(n+1)\,\Sigma x_i + n(n+1)^2 = \frac{1}{3}n(n^2-1)$].

Ex. Show that the maximum value of Σd_i^2 is $\frac{1}{3}\left(\frac{n^3-n}{n}\right)$. Hence or otherwise show that the rank correlation coefficient lies between -1 and $+1$.

10·6. Correlatien Ratio. So far we have assumed that there is a single observed value y_i corresponding to any given value x_i, but in general there may be more than one such value of y. Let the value y_{ij} occur with frequency f_{ij}, then the definitions given earlier can be extended to get

$$\sum_i \sum_j f_{ij}\, y_{ij} = \sum_i n_i \bar{y}_i = N\bar{y},$$

where
$$\bar{y}_1 = \frac{\sum_j f_{ij}\, y_{ij}}{\sum_j f_{ij}} = \frac{\sum_j f_{ij}\, y_{ij}}{n_i},$$

$$\sum_i \sum_j f_{ij} = N, \quad \sum_i n_i = N,$$

and
$$\bar{y} = \frac{\sum_i \sum_j f_{ij}\, \bar{y}_{ij}}{\sum_i \sum_j f_{ij}} = \frac{\sum_i n_i\, \bar{y}_i}{\sum_i n_i}$$

so that $\bar{y}_i$ is the mean of the ith array consisting of all y's corresponding to x_i and $\bar{y}$ is the general mean that is the weighted mean of the array means, the weights being the array frequencies. Let (x_{ij}, y_{ij}) the typical pair of values in the ith array, be represented by the point P_{ij} and the mean $(x_i, \bar{y}_i)$ of the array by the point G_i (see Fig. 31).

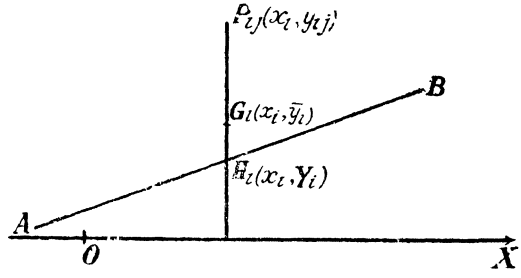


Fig. 31.

Let $y = a + bx$ be the equation of the line of regression, then, by the principle of least squares, a and b are determined by minimizing

$$S = \sum_i \sum_j f_{ij}(y_{ij} - a - bx_i)^2$$

The normal equations are

$$\sum_i \sum_j f_{ij}\,(y_{ij} - a - bx_i) = 0$$

and
$$\sum_i \sum_j f_{ij}\, x_i\,(y_{ij} - a - bx_i) = 0$$

These equations give for the equation of the line of regression

$$y - \bar{y} = \frac{\mu_{11}}{\sigma_x^2}(x - \bar{x}),$$

where $\mu_{11}=\frac{1}{N}\sum_i\sum_j f_{ij}(y_{ij}-\bar{y})(x_i-\bar{x})=\frac{1}{N}\sum_i n_i(y_i-\bar{y})(x_i-\bar{x})$,

$$\sigma_x^2=\frac{1}{N}\sum_i\sum_j f_{ij}(x_i-\bar{x})^2=\frac{1}{N}\sum_i n_i(x_i-\bar{x})^2,$$

$$\sigma_x^2=\frac{1}{N}\sum_i\sum_j f_{ij}(y_{ij}-\bar{y})^2=\frac{1}{N}\sum_i n_i(y_i-\bar{y})^2.$$

Thus we get the same equations for the lines of regression and in the same way, we get

$$\sum_i\sum_j f_{ij}(y_{ij}-Y_i)^2=N\sigma_y^2(1-r^2).$$

On analogy with the relation (10·14), we define correlation ratio η_{yx} of y on x by the relation

$$NS_y'^2=\sum_i\sum_j f_{ij}(y_{ij}-\bar{y}_i^2)=N\sigma_y^2(1-\eta_{yx}^2) \qquad ...(10·17)$$

where $S_y'^2$ is the mean square of deviations from the means of arrays.

From (10·17), it follows that

$$0\leqslant\eta_{yx}^2\leqslant 1 \qquad ...(10·18)$$

Also since the sum of squares of deviations in any array is least when they are measured from the mean of the array, we have

$$S_y'^2\leqslant S_y^2.$$

This gives

$$\eta^2_{yx}\geqslant r^2 \qquad ...(10·19)$$

Combining (10·18) and (10·19) we have

$$0\leqslant r^2\leqslant\eta_{yx}^2\leqslant 1. \qquad ...(10·20)$$

When the regression of y on x is linear, the means of the arrays will lie on the line of regression and then we have

$$\eta^2_{yx}=r^2.$$

A non-zero value of $\eta^2_{yx}-r^2$ is associated with the non-linearity of the regression of y on x.

If $\eta_{xy}^2=1$, then $S_y'^2=0$, so that all the deviations are zero, and all the values in each array coincide with the mean of the array : that is all points lie on the curve of the means of the arrays. The nearer is η_{yx}^2 to unity, the smaller is $S_y'^2$ and, therefore, the closer are the points to the curve of the means. We may thus regard *the correlation ratio η_{yx} as a measure of the degree to which the association between the variables approaches a function relationship* expressible in the form $y=f(x)$, when $f(x)$ is a single-valued function of x.

An analogous discussion for the arrays of x's obviously leads to

$$\eta^2_{xy}=1-\frac{S_x'^2}{\sigma_x^2} \qquad \ldots(10{\cdot}21)$$

giving η_{xy}, the correlation ratio of x on y.

For some purposes, it is convenient to express the correlation ratio in a form involving the standard deviation of the means of arrays. The result is proved below.

10·6·1. Theorem. *The correlation ratio of y on x is the ratio of the standard deviation of the weighted means of the arrays of y's (weighted by the corresponding array frequencies) to the standard deviation of all y's of the distribution.*

Proof. *We have*

$$N\sigma_y^2=\sum_i \sum_j f_{ij}(y_{ij}-\bar{y})^2. \qquad \ldots(10{\cdot}22)$$

Writing $y_{ij}-\bar{y}$ as $(y_{ij}-\bar{y}_i)+(\bar{y}_i-\bar{y})$ and observing that

$$\sum_i \sum_j f_{ij}(y_{ij}-\bar{y}_i)=\sum_i \left(\sum_j f_{ij}y_{ij}-\bar{y}_i \sum_i f_{ij}\right)$$

$$=\sum_i n_i\bar{y}_i-\sum_i \bar{y}_i n_i=0$$

(10·22) can be expressed in the form

$$N\sigma_y^2=\sum_i \sum_j f_{ij}(y_{ij}-\bar{y}_i)^2+\sum_i \sum_j f_{ij}(\bar{y}_i-\bar{y})^2$$

$$=\sum_i \sum_j f_{ij}(y_{ij}-\bar{y}_i)^2+\sum_i n_i(\bar{y}_i-\bar{y})^2$$

$$=NS_y'^2+N\sigma_{my}^2$$

where $$N\sigma_{my}^2=\sum_i n_i(\bar{y}_i-\bar{y})^2$$

so that σ_{my} is the s.d. of the means of arrays of y's, each mean being weighted by the frequency n_i of that array.

Therefore, using (10·17), $\eta_{yx}^2 = \dfrac{\sigma_y^2 - S_y'^2}{\sigma_y^2} = \dfrac{\sigma_{my}^2}{\sigma_y^2}$ ...(10·23)

Hence the theorem is established.

Example 10·6. *Calculate η_{xy} from the following correlation table :*

x \ y / u \ v		10	15	20	25	*Total* f	*Mean of the rows*
		−2	−1	0	1		
7	−2	3	2	—	—	5	12
9	−1	—	1	4	6	11	22·3
11	0	—	3	4	2	9	19·4
13	1	2	1	5	—	8	16·9
15	2	—	6	—	1	7	16·4
Total f		5	13	13	9	40	
Mean of the columns		9·4	12·2	11 2	10·1		

We calculate $\bar{x}$, σ_x and σ_{xy}

$$\bar{x} = a + \frac{h\Sigma u f_x}{\Sigma f_x} \text{ where } u = \frac{x-20}{5}$$

or
$$\bar{x} = 20 + \frac{(-14)}{40} \times 5 = 18{\cdot}25$$

$$\sigma_x^2 = h \sqrt{\frac{\Sigma u^2 f_x}{\Sigma f_x} - \left(\frac{\Sigma u f_x}{\Sigma f_x}\right)^2}$$

$$= 5 \times \sqrt{\frac{42}{40} - \left(\frac{-14}{40}\right)^2} = 5 \times {\cdot}968 = 4{\cdot}84.$$

The table given on the next pages shows the calculation of σ_{my}.

Mean of the rows	$f\bar{x}$	u' (=Deviation from 19)	$f\bar{x}\times u'$	$fx\times u'^2$
12	5	−7	−35	245·00
22·3	11	3·3	36·3	119·79
19·4	9	0·4	3·6	1·44
16·9	8	−2·1	−16·8	35·28
16·4	7	−2·0	−18·2	47·32
Total	40		−30·1	448·83

$$\bar{x}=19+\frac{(-30\cdot1)}{40}=18\cdot25.$$

This is same as already obtained.

$$\sigma_{m_y}^2=\frac{448\cdot83}{40}-\left(\frac{30\cdot1}{42}\right)^2=10\cdot6583$$

$$\therefore \qquad \sigma_{m_y}=3\cdot26$$

Hence $$\eta_{xy}=\frac{\sigma_{m_y}}{\sigma_x}=\frac{3\cdot26}{4\cdot84}=0\cdot68.$$

Ex. 1. Show that η_{xy} calculated for the distribution given in table of the above example is 0·37.

Ex. 2. Show that

$$\sum_i n_i(\bar{y}_i-y)^2=\sum_i\left(\frac{T_i^2}{n_i}\right)-\frac{T^2}{N}$$

where T_i is the sum of the y's in the ith vertical array and T is the sum of the y's for the whole distribution.

[**Hint.** $n_i\,\bar{y}_i=\sum_i f_{ij}y_{ij}=T_i$, $\sum_i n_i\,\bar{y}_i=\Sigma T_i=N\bar{y}=T$.]

Ex. 3. Show that η_{yx} can be calculated from the formula

$$\eta^2_{yx}=\left(\sum_i\frac{T_i^2}{n_i}-\frac{T^2}{N}\right)\Big/N\sigma_y^2$$

show that the value of η^2_{yx} calculated from this formula is unaffected by the change of origin and unit. State the corresponding formula and result for η_{xy}.

Ex. 4. Apply the above formula to calculate η_{yx} and η_{xy} for the data of Example 10·6.

Ex. 5. Prove from first principles that

(*i*) $1 \geqslant \eta^2_{xy} \geqslant r^2$ *(B.A. Hons., Delhi 1960)*

(*ii*) $\eta_{xy} = \frac{\sigma_{xm}}{\sigma_x}$. *(M.A. Statistics, Delhi 1968)*

Ex. 6. Prove the relations ;

(*i*) $\sum_i n_i(y_i - Y_i) = N\sigma^2_y(\eta^2_{yx} - r^2)$

(*ii*) $\sum_i n_i(y_i - \bar{y})^2 = \Sigma\, n_i(\bar{y}_i - Y_i)^2 + \Sigma\, n_i(Y_i - \bar{y})^2$.

Ex. 7. Prove the above theorem for a continuous probability distribution.

10·7. Bivariate Normal Distribution. The bivariate normal distribution is a generalization of the normal distribution for a single variate. It is derived briefly as follows :

Assume first that one of the variates, say x, is distributed normally with mean m_1 and *s.d.* σ_x so that the probability that the value of x taken at random will lie in the interval dx is

$$\frac{1}{\sigma_x\sqrt{2\pi}} \exp\left\{\frac{-(x-m_1)^2}{2\sigma_x^2}\right\}dx$$

Assume next that each array of y corresponding to an assigned value of x has a normal distribution with the standard deviation of an array given by $\sigma_y\sqrt{(1-\rho^2)}$; independent of x, ρ is the coefficient of correlation between the variates.

Finally, assume that the regression of y on x is linear. Since the mean of each array is on the line of regression $Y = m_2 + \frac{\rho\sigma_y}{\sigma_x}(x - m_1)$, the probability that a value of y taken at random from an assigned array of y will be in the interval dy is

$$\frac{1}{\sigma_y\sqrt{2\pi(1-\rho)^2}} \exp\left[\frac{-1}{2\sigma_y^2(1-\rho^2)}\left\{y - m_2 - \frac{\rho\sigma_y}{\sigma_x}\left(x - m_1\right)\right\}^2\right]dy$$

By the theorem of compound probability, the joint probability that x will fall in the interval dx and y in the interval dy is

$$dP = f(x, y)\, dxdy, \quad -\infty < x, (y < \infty, \qquad \ldots(10{\cdot}23)$$

where

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot$$

$$\exp\left[-\frac{1}{2(1-\rho^2)} \cdot \left\{\frac{(x-m_1)^2}{\sigma_x^2} - \frac{2\rho(x-m_1)(y-m_2)}{\sigma_x\sigma_y} + \frac{(y-m_2)^2}{\sigma_y^2}\right\}\right]$$

$$\ldots(10{\cdot}23')$$

This distribution is called the bivariate normal distribution (see figure 32 below) and the variates are said to be *normally correlated.* This surface

$$z=f(x, y)$$

is called the *normal correlation surface.*

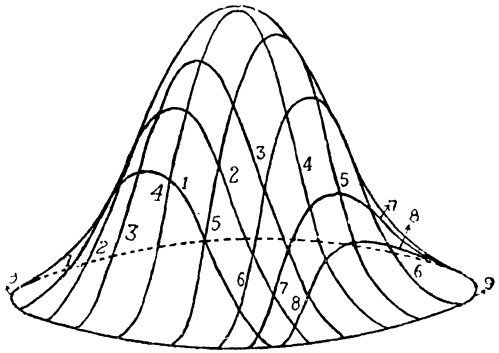


Fig. 32. Bivariate Normal Surface.

The function $f(x, y)$ actually represents a density function is proved by showing that its integral over the whole plane is unity *i.e.*,

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)dxdy=1 \qquad \ldots(10{\cdot}24)$$

Put $\quad u=\dfrac{x-m_1}{\sigma_x}\ ,\ v=\dfrac{y-m_2}{\sigma_y}$

and then L.H.S. of (10·24) becomes

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}(u^2-2\rho uv+v^2)\right]dudv$$

$\ldots(10{\cdot}25)$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\{(u-\rho v)^2+(1-\rho^2)v^2\}\right]dudv$$

Put $\quad w=\dfrac{u-\rho v}{\sqrt{1-\rho^2}}\ ,\ v=v$

and the integral may be written as the product of two simple integrals

$$\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}dw\times\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}dv$$

both of which are equal to unity. Thus equation (10·24) is verified.

Note. (10·24) can alternatively be proved by using

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\left[-(ax^2+2hxy+by^2)dxdy = \pi/\sqrt{ab-h^2}.\right.$$

$$a>0,\ b>0,\ ab-h^2>0.$$

10·7·1. *Properties of the Normal Correlation Distribution.*

1. *The m.g.f. of the distribution about the mean* (m_1, m_2) is given by $M(t_1, t_2)=E[\exp\{t_1(x-m_1)+t_2(y-m_2)\}]$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\{t_1(x-m_1)+t_2(y-m_2)\}\, f(x, y)dxdy$$

Making the substitutions

$$\xi=x-m_1-\sigma_x^2 t_1-\rho\sigma_x\sigma_y t_2$$
$$\eta=y-m_2-\rho\sigma_x\sigma_y t_1-\sigma_y^2 t_2$$

we find that

$$M(t_1, t_2)=\exp[\tfrac{1}{2}(t_1^2\sigma_x^2+2t_1t_2\sigma_x\sigma_y\rho+t_2^2\sigma_y^2)]\times\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

$$\times\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{\xi^2}{\sigma_x^2}-\frac{2\rho\xi\eta}{\sigma_x\sigma_y}+\frac{\eta^2}{\sigma_y^2}\right)\right]d\xi d\eta$$

$$=\exp[\tfrac{1}{2}(t_1^2\sigma_x^2+2t_1t_2\sigma_x\sigma_y\rho+t_2^2\sigma_y^2)] \qquad \ldots(10\cdot26)$$

Now μ_{rs}=coefficient of $\frac{t_1^r t_2^s}{r!\ s!}$ in the expansion of $M(t_1, t_2)$.

We find that

$$\mu_{20}=\sigma_x^2,\ \mu_{11}=\rho\sigma_x\sigma_y,\ \mu_{02}=\sigma_y^2$$
$$\mu_{30}=\mu_{21}=\mu_{12}=\mu_{03}=0$$
$$\mu_{40}=3\sigma_x^4,\ \mu_{31}=3\rho\sigma_x^3\sigma_y,\ \mu_{04}=3\sigma_y^4$$
$$\mu_{22}=(1+2\sigma^2)\sigma_x^2\sigma_y^2,\ \mu_{13}=3\rho\sigma_x\sigma_y^3.$$

2. *The marginal density of* x is given by

$$g(x)=\int_{-\infty}^{+\infty} f(x, y)dy \qquad \text{Put } v=\frac{y-m_2}{\sigma_y}$$

$$=\int_{-\infty}^{+\infty}\frac{1}{2\pi\sigma_x\sqrt{1-\rho^2}}\exp\left[-\tfrac{1}{2}\frac{(x-m_1)^2}{\sigma_x^2}-\frac{1}{2(1-\rho^2)}\left(v-\frac{\rho(x-m_1)}{\sigma_x}\right)^2\right]dv$$

Put $w=v-\rho\left(\frac{x-m_1}{\sigma_x}\right)\Big/\sqrt{1-\rho^2}$

$$\therefore \quad g(x)=\frac{1}{\sigma_x\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{(x-m_1)^2}{\sigma_x^2}}$$

Similarly, marginal density of y is given by

$$h(y)=\frac{1}{\sigma_y\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{(y-m_2)^2}{\sigma_y^2}}$$

Hence $X\sim N(m_1, \sigma_x^2)$ and $Y\sim N(m_2, \sigma_y^2)$.

3. *The conditional density of x for given value of y is*

$$f(x \mid y)=\frac{f(x, y)}{h(y)}$$

$$=\frac{1}{\sigma_x\sqrt{2\pi(1-\rho^2)}}\times\exp\left[\frac{-1}{2\sigma_x^2(1-\rho^2)}\left\{(x-m_1)-\frac{\rho\sigma_x}{\sigma_y}(y-m_2)\right\}^2\right]$$

which is a univariate normal density with mean

$$E(x \mid y)=m_1-\frac{\rho\sigma_x}{\sigma_y}(y-m_2)$$

and with variance Var $(x \mid y)$ given by $\sigma_x^2(1-\rho^2)$. Interchanging x and y throughout, we get the conditional distribution of y for a given value of x.

The exponent in the bivariate normal distribution can be written as

$$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-m_1}{\sigma_x}-\frac{\rho(y-m_2)}{\sigma_y}\right)^2+\frac{(y-m_2)^2}{\sigma_y^2}(1-\rho^2)\right] \qquad \ldots(10{\cdot}27)$$

or $$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-m_2}{\sigma_y}-\frac{\rho(x-m_1)}{\sigma_x}\right)^2+\frac{(x-m_1)^2}{\sigma_x^2}(1-\rho^2)\right]$$

From (10·27), $$E(x \mid y)=m_1-\frac{\rho\sigma_x}{\sigma_y}(y-m_2)]$$

and $$Var(x \mid y)=\sigma_x^2(1-\rho^2)$$

The conditional mean of either variate, given the values of the other, is a linear function of that value and that its conditional variance is constant, independent of that value. Thus the regression of x on y (or of y on x) is *linear* and *homoscedastic* (Greek word meaning equal scattering).

Ex. 1. (*a*) Write down the probability density function of bivariate normal distribution in the variates x and y, and show that

(*i*) x is distributed normally with variance σ_x^2.

(*ii*) y is distributed normally with variance σ_y^2.

(*iii*) For any given x, y is distributed normally about mean $(\rho\sigma_y/\sigma_x)\,x$ with variance $\sigma_y^2\,(1-\rho^2)$.

(*M.A. Punjab, 1958*; *M.A., Delhi, 1958, 60, 65*)

(*iv*) For any given y, x is distributed normally about mean $(\rho\sigma_x/\sigma_y)\,y$ with variance $\sigma_x^2\,(1-\rho^2)$.

(*v*) Both the regressions of y on x and of x on y are linear.

(*M.A. Punjab 1960*)

(*vi*) The variances of all the y-arrays are same (*i.e.*, regression of y on x is *homoscedastic*).

(*b*) Find the probability of the simultaneous materialization of the inequalities $x > E(x)$ and $y > E(y)$.

(*M.A. Delhi, 1967*; *M.Sc. Bombay, 1969*; *M.A. Patna, 1957*)

Hint. (*b*) $\int_0^\infty \int_0^\infty e^{-(ax^2+2hxy+by^2)}\,dxdy = \frac{1}{2\sqrt{ab-h^2}}\cot^{-1}\frac{h}{\sqrt{ab-h^2}}$

$a>0,\ b>0,\ h>0,\ ab-h^2>0$

Ex. 2. By differentiating $\int_{-\infty}^{+\infty} dP=1$ (as given by (10·23) with respect to ρ, show that $E(x-m_1)(y-m_2)=\rho\sigma_x\sigma_{xy}$. Find also the mean and variance of x and y by the same method.

Ex. 3. Obtain the regression curve for the mean of y for a given x for the distribution

$$f(x, y) = \frac{9(1+x+y)}{2(1+x)^4(1+y)^4},\quad 0 \leqslant x, y < \infty.$$

(*M.A. Delhi, 1962*)

Also show that the marginal distribution of x is

$$dF = \frac{3}{4}\cdot\frac{2x+3}{(1+x)^4}\,dx,\ 0 \leqslant x < \infty. \qquad \text{Ans. } y = \frac{x+3}{2x+3}$$

Ex. 4. Prove that the curve of regression of y on x for the distribution

$$f(x, y) = \begin{cases} 2 & x-y,\ 0<x<1,\ 0<y<1 \\ 0 & \text{elsewhere} \end{cases}$$

is a hyperbola.

(*M.A. Statistics, Delhi, 1969*)

Ex. 5. For the bivariate normal distribution given by

$$z = \text{const exp.}\left[\frac{-1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right]$$

(*a*) Prove that

(*i*) $\dfrac{\partial^2 z}{\partial x \partial y} = \dfrac{\partial z}{\partial \rho}$

(*ii*) $M(t_1, t_2) = \exp\left[\tfrac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right]$

(*b*) Obtain the values of $\dfrac{\partial^2 M}{\partial t_1 \partial t_2}$, $t_1 \dfrac{\partial M}{\partial t_1}$ and $t_2 \dfrac{\partial M}{\partial t_2}$.

(*c*) Show that $\dfrac{\partial^2 M}{\partial t_1 \partial t_2} - \rho \left\{ t_1 \dfrac{\partial M}{\partial t_1} + t_2 \dfrac{\partial M}{\partial t_2} \right\} = \rho(1-\rho^2) t_1 t_2 M.$

Get a relation between these quantities and M and identifying coefficient show that

$$\mu_{r2} = (r+s-1)\mu_{r-1,\, s-1} + (r-1)(s-1)(1-\rho^2)\mu_{r-2,\, s-2}.$$

Hence or otherwise show that $\mu_{rs} = 0$ if $r+s$ is odd. (*M.A. Madras, 1962*)

10·8. Intra-class Correlation Coefficient. Sometimes specially in biological situations, it is desired to study the correlation between some characteristic of members of one or more families. Thus we may be interested in the correlation heights of brothers or weights of sisters or lengths of leaves in trees. The correlation coefficients between heights of tallest and shortest brothers or between the weights of eldest and youngest sisters would not serve our purpose, as here we are interested in correlations in which all the members of a family have the same status.

Let there be h families containing $k_1, k_2, \ldots\ldots, k_h$ members and let x_{ij} $[i = 1, 2, \ldots\ldots, h\ ;\ j = 1, 2, \ldots\ldots, k_i]$ denote the value of the jth member of the ith family.

The product-moment correlation coefficient obtained from the

$$N = \sum_{i=1}^{h} k_i(k_i - 1)$$

pairs x_{ij}, x_{il} $[i = 1, 2, \ldots\ldots, h\ ;\ j, l = 1, 2, \ldots\ldots, k_i\ ;\ j \neq l]$ is called the *intra-class correlation coefficient* and the table giving all these variate values is called *intra-class correlation table.* There is an obvious symmetry in the table, since the value of each member of the ith family occur $(k_i - 1)$ times as a value of the x-variate and the same number of times as value of the y-variate. Therefore,

$$\bar{x} = \bar{y} = \frac{\sum_i \sum_j (k_i - 1) x_{ij}}{\sum_i k_i(k_i - 1)} = \frac{1}{N}\left[\sum_i (k_i - 1) \sum_j x_{ij}\right]$$

Similarly,

$$\sigma_x^2 = \sigma_y^2 = \frac{1}{N} \sum_i (k_i - 1) \sum_j (x_{ij} - \bar{x})^2,$$

and

$$\text{Cov } (x, y)=\frac{1}{N}\sum_i\sum_j\sum_l{}'(x_{ij}-\bar{x})(x_{il}-\bar{x}),\ i\neq l$$

$$=\frac{1}{N}\sum_i\sum_j(x_{ij}-\bar{x})\sum_l{}'(x_{il}-\bar{x})$$

Σ' indicating that the summation over l has to exclude the value j, so that

$$\sum_l{}'(x_{il}-\bar{x})=\sum_l(x_{il}-\bar{x})-(x_{ij}-\bar{x})=k_i(\bar{x}_i-\bar{x})-(x_{ij}-\bar{x}),$$

where $\bar{x}_i$ denotes the mean of the ith class *i.e.*, $\bar{x}_i=\frac{1}{k_i}\sum_j x_{ij}$

$$\therefore\quad \text{Cov } (x, y)=\frac{1}{N}\left[\sum_i k_i(\bar{x}_i-\bar{x})\sum_j(x_{ij}-\bar{x})-\sum_i\sum_j(x_{ij}-\bar{x})^2\right]$$

$$=\frac{1}{N}\left[\sum_i k_i^2(\bar{x}_i-\bar{x})^2-\sum_i\sum_j(x_{ij}-\bar{x})^2\right]$$

$$\therefore\quad \rho_{xy}=\rho=\frac{\sum_i k_i^2(\bar{x}_i-\bar{x})^2-\sum_i\sum_j(x_{ij}-\bar{x})^2}{\sum_i(k_i-1)\sum_j(x_{ij}-\bar{x})^2}$$

Case of equal number of members in each family.

Let $\quad k_1=k_2=\ldots\ldots=k_h=k$, then

$$\rho=\frac{k^2\,\Sigma(\bar{x}_i-\bar{x})^2-\Sigma\,\Sigma(x_{ij}-\bar{x})^2}{(k-1)\Sigma\,\Sigma(\bar{x}_{ij}-\bar{x})^2}$$

Let σ^2 denote the variance of x and σ_m^2 the variance of means of families (about the mean x), so that

$$\sigma^2=\frac{1}{hk}\sum_i\sum_j(x_{ij}-\bar{x})^2\ ;\ \sigma_m^2=\frac{1}{h}\sum_i(\bar{x}_i-\bar{x})^2$$

$$\therefore\quad \rho=\frac{1}{k-1}\left(\frac{k\sigma_m^2}{\sigma^2}-1\right)=\frac{k\sigma_m^2-\sigma^2}{(k-1)\sigma^2} \qquad \ldots(10{\cdot}28)$$

Notes. 1. For distinction, the ordinary correlation coefficient *e.g.*, the correlation between the heights of eldest and youngest brothers would be called an *intra-class correlation coefficient.*

2. Since $\sigma_m^2\geqslant 0$, the intra-class correlation coefficient cannot be less than $-1/(k-1)$, though it may attain the value $+1$. This is thus a skew coefficient in the sense that a negative value for it has not the same significance as an equivalent positive value. In fact if k is very large, ρ is almost always positive.

3. ρ will attain the value $-1/(k-1)$ if $\rho_m^2=0$ *i.e.*, if there is no variation between family mean. In particular, if there is only one family of N members, the intra-class correlation coefficient has the value $-1/(N-1)$. In §12·12, we shall deduce a similar result for sampling from a finite population.

4. ρ is obviously independent of the origin and the scale.

Ex. 1. In five families of 3, the heights of brothers are in inches :
69, 70, 71 ; 70, 71, 72 ; 71, 72, 73 ; 72, 73, 74 ; 73, 74, 75.

Find the intra class coefficient of correlation.

(*M.A. Statistics Delhi, 1966*)

Ex. 2. From the intra-class correlation table for the above heights calculate without using the formula.

Ex. 3. In Ex. 1, find the correlation coefficient between the heights of the tallest and shortest brothers and compare with the value of ρ.

EXERCISES 10

1. Show that, if x', y' are the deviations of the variables x, y from their means,

$$r=1-\frac{1}{2N}\sum_i f_i\left(\frac{x_i'}{\sigma_x}-\frac{y_i'}{\sigma_y}\right)^2,\ \Sigma f_i=N$$

and

$$r=-1+\frac{1}{2N}\sum_i f_i\left(\frac{x_i'}{\sigma_x}+\frac{y_i'}{\sigma_y}\right)^2.$$

Deduce that $-1\leqslant r\leqslant 1$.

2. Let X and Y be continuous chance variables having the joint density $f(x, y)$. Define r, the coefficient of correlation between X and Y, in terms of $f(x, y)$. Prove that if X and Y are independent, then $r=0$. Is the converse also true ?

(*B.A. Hons., Delhi, 1952*)

3. (*a*) Give examples to illustrate that two uncorrelated variables may not be independent.

(*b*) Let x have the marginal density $f_1(x)=1$, $-\frac{1}{2}<x<1$, and let the conditional density of y be

$$f(y\mid x)=1,\ x<y<x+1,\ -\tfrac{1}{2}<x<0$$
$$=1,\ -x<y<1-x,\ 0<x<\tfrac{1}{2}$$
$$=0,\ \text{otherwise.}$$

Show that $E(x)=0$, $E(x\,y)=0$ and hence deduce that the correlation between x and y is zero.

[**Hint.** (*a*) Let

x:	-1	-2	-3	1	2	3
y:	1	4	9	1	4	9.

It can be easily proved that coefficient of correlation between x and y is zero but at a glance we find that x and y are connected by the relation $y=x^2$.

Another example :

Let x be a normal variate with zero mean and unit variance and let $y=x^2$, then

$$\text{Cov}\,(x, y)=E(x-\bar{x})(y-\bar{y})=E(x-0)(x^2-\bar{y})=E(x^3)-\bar{y}E(x)$$
$$=0-\bar{y}\,.\,0=0].$$

4. (*a*) Calculate the coefficient of correlation and obtain the lines of regression for the following data :

x :	1	2	3	4	5	6	7	8	9
y :	9	8	10	12	11	13	14	16	15

Obtain an estimate of y which should correspond on the average to $x=6{\cdot}2$.

(*I.A.S. 1955*)

[Ans. $r={\cdot}95$, the lines of regression are $y-12={\cdot}95(x-5)$ and $x-5=95(y-12)$. Required estimate is 13·14]

[Hint. Take $u=x-5$, $v=y-12$, then $\Sigma u=0=\Sigma v$
$\Sigma u^2=60$, $\Sigma v^2=60$, $\Sigma uv=57$. $\therefore$ $r=$etc.]

(*b*) Calculate the coefficient of correlation between the values of x and y from the following table :

x :	78	89	97	69	59	79	61	61
y :	125	137	156	112	107	136	123	108

(You may use 69 as working mean for x and 112 as that for y).

[Ans. $r=+{\cdot}957$]

5. Calculate the coefficient of correlation from the following table giving the ages of 100 husbands and their wives in years.

Age of husbands

Ages of wives	20–30	30–40	40–50	50–60	60–70	*Total*
15–20	5	9	3	...	...	17
25—35	...	10	25	2	...	37
35–45	...	1	12	2	...	15
45—55	...	...	4	16	5	25
55–65	...	...	...	4	2	6
Total	5	20	44	24	7	100

[Ans. $r={\cdot}796$]

6. The variables x and y are correlated by the equation $ax+by+c=0$. Show that the correlation between them is -1 if the signs of a and b are alike, and $+1$ if they are different.

7. (*a*) If σ_x and σ_y are the standard deviations of two uncorrelated statistical variates x and y, prove that the standard deviation of $ax+by$ is

$$\sqrt{a^2\sigma_x^2+b^2\sigma_y^2}.$$ (*B.A., Madras, 1945*)

(*b*) Show that if a and b are constants and r is the correlation between x and y, then the correlation between ax and by is equal to r if the signs of a and

b are alike and to $-r$ if they are different. Also show that, if the constants a, b, c are positive, correlation between $ax+by$ and cy is equal to

$$(ar\sigma_x+b\sigma_y)/\sqrt{a^2\sigma_x^2+b^2\sigma_y^2+2abr\sigma_x\sigma_y}.$$

8. Let x_1, x_2 and x_3 be uncorrelated variables each having the same standard deviation. Obtain the coefficient of correlation between (x_1+x_2) and (x_2+x_3). *(I.A.S., 1949)*

[Ans. $r=\frac{1}{2}$]

9. If $u=ax+by$ and $v=bx-ay$ where x and y are measured from their respective means, and r is the coefficient of correlation between x and y and if u and v are uncorrelated, show that

(*i*) $\sigma_u\sigma_v=(a^2+b^2)\sigma_x\sigma_y(1-r^2)^{\frac{1}{2}}$ *(B.Sc., Agra, 1967; B.A., Delhi 1967)*

(*ii*) $ab(\sigma_x^2-\sigma_y^2)=r\sigma_x\sigma_y(a^2-b^2)$

(*iii*) $\sigma_u^2+\sigma_v^2=(a^2+b^2)(\sigma_x^2+\sigma_y^2)$ *(B.A., Delhi, 1971)*

[**Hint.** (*i*) Use the result. $\begin{vmatrix} Var\ u & Cov(u,v) \\ Cov(u,v) & Var\ v \end{vmatrix} = \begin{vmatrix} Var\ x & Cov(x,y) \\ Cov(x,y) & Var\ y \end{vmatrix} \times \begin{vmatrix} a & b \\ b & -a \end{vmatrix}^2$]

10. If x and y are independent random variables, show that

$$r_{x+y\ :\ x-y}=r^2x\ :\ x+y-r^2y\ :\ x+y$$

where $r_{x+y\ :\ x-y}$ denotes the coefficient of correlation between $x+y$ and $x-y$.

11. (*a*) Show that the coefficient of correlation r between two variables x and y is given by

$$r=(\sigma_x^2+\sigma_y^2-\sigma_{x-y}^2)/2\sigma_x\sigma_y$$

where σ_x^2, σ_y^2 and σ_{x-y}^2 are the variances of x, y and $x-y$ respectively.

(B.A., Madras, 1947)

(*b*) X_1, X_2, X_3 are three variables each with variance σ^2 and the correlation coefficient between any two of them is r. If $\bar{X}=\frac{1}{3}(X_1+X_2+X_3)$, show that

$$\text{Var}\ \bar{X}=\tfrac{1}{3}(1+2r)$$

Deduce that $r\geqslant-\frac{1}{2}$. *(B.Sc. Gujarat, 1967)*

12. (*a*) The variates x and y have zero means, the same variance σ^2 and zero correlation. Show that the variates

$$u=(x\cos\alpha+y\sin\alpha) \text{ and } v=(x\sin\alpha-y\cos\alpha)$$

have the same variance σ^2 and zero correlation. *(B.A. Hons., Delhi 1952)*

(*b*) If the variances σ_x^2 and σ_y^2 of x and y are different, show that the correlation coefficient between the two variates u and v where $u=x\cos\alpha+y\sin\alpha$ and $v=x\sin\alpha-y\cos\alpha$ is given by

$$r_{uv}=\frac{\sigma_x^2+\sigma_y^2}{[\{\sigma_x^2-\sigma_y^2)^2+4\sigma_x^2\sigma_y^2\ \text{cosec}^2\ 2\alpha\}]^{\frac{1}{2}}}$$

(M.A. Statistics, Delhi, 1965, 68)

Deduce that

$$-\frac{\sigma_x^2\sim\sigma_y^2}{\sigma_x^2+\sigma_y^2}\leqslant r_{uv}\leqslant\frac{\sigma_x^2\sim\sigma_y^2}{\sigma_x^2+\sigma_y^2}.$$

(*c*) Prove that $1 \geqslant n_{yx}^2 > \rho^2$ where ρ is the coefficient of correlation between x and y. How are the values of n_{yx} interpreted ? Bring out the difference in the interpretation of the values of n_{yx} and those of ρ. (*M.A. Delhi, 1968*)

13. x_1 and x_2 are two variates with variances σ_1^2 and σ_2^2 respectively and r is the correlation between them. Determine the value of the constant k such that $x_1 + kx_2$ and $x_1 + \frac{\sigma_1}{\sigma_2} x_2$ are uncorrelated.

(*B.Sc., Hons , Sub. 1965 ; B.A., Hons., Delhi, 1960 ; B.A., Madras, 1969*)

[Ans. $k = -\sigma_1/\sigma_2$]

14. (*a*) If x_i $(i=1, 2, \ldots, n)$ are the values of the variable x and w_i are the corresponding weights, then show that the weighted arithmetic mean

$$\bar{x}_w \left(= \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i \right)$$

is greater or less than the unweighted mean

$$\bar{x} \left(= \sum_{i=1}^{n} x_i / n \right)$$

according as the weights and their values of the variable are positively or negatively correlated. (*B.A. Hons., Delhi, 1949*)

(*b*) If x and y are standardized random variables, and

$$r(ax+by, bx+ay) = (1+2ab)/(a^2+b^2)$$

find $r(x, y)$. (*B.A. Delhi, 1971*)

[Ans. $(a^2+b^2)/[(a^2-b^2)^2-2ab]$]

15. (*i*) If two random variates X and Y assume only two values each, and Cov $(X, Y)=0$, show that X and Y are independent.

(*ii*) Carefully examine each of the following five statements (*a*) to (*e*) and say whether the statement is true or false adding in each case a few lines to substantiate your conclusion.

If the correlation coefficient between the random variables X and Y be negative, then it means that

(*a*) the correlation coefficient between $-X$ and $-Y$ is positive.

(*b*) $E(XY) < E(X)E(Y)$.

(*c*) the regression coefficient of X on Y and that of Y on X are both negative.

(*d*) The regression of Y on X is linear and the line is downward sloping.

(*e*) Y^2 and X^2 are dependent random variables.

(*Certificate Exam., Calcutta, Sept. 1956*)

16. (*a*) Find the most likely price in Bombay (x) corresponding to the price of Rs. 70 at Calcutta (y) from the following data :

Average price at Calcutta = 65
Average price at Bombay = 67
S.D. at Calcutta = 2·5
S.D. at Bombay = 3·5

Coefficient of correlation is +·8 between the two prices of the commodity in two towns. (*B.A. Punjab, 1961 S*)

(*b*) Two random variables have the least square regression lines with equations $3x+2y-26=0$ and $6x+y-31=0$. Find the mean values and the correlation coefficient between x and y.

(*B.Sc., Lucknow ; B.A. Hons., Delhi, 1957, 36 ; M.A., Delhi, 1960 ; M.A. Bombay, 1968*)

(*c*) In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible :

Variance of $x=9$.

Regression equations : $8x-10y+66=0$; $40x-18y=214$.

What were (*a*) the mean values of x and y, (*b*) the standard deviation of y, and (*c*) the coefficient of correlation between x and y ?

(*B.A., Punjab, 1961 ; I.A.S., 1947*)

[Ans. (*a*) The equation to the line of regression of x on y is
$x=1{\cdot}12y-5{\cdot}8$, $x_{70}\equiv$ Rs. 72·6

(*b*) $\bar{x}=4$, $y=7$, $r=-{\cdot}5$ (*c*) $\bar{x}=13$, $\bar{y}=17$, $\sigma_y=4$, $r_{xy}=+{\cdot}6$]

17. (*a*) A computor while calculating the correlation coefficient between two variates x and y from 25 pairs of observations obtained the following constants

$$n=25,\ \Sigma x=125,\ \Sigma x^2=650,\ \Sigma y=100,\ \Sigma y^2=460,\ \Sigma xy=50.$$

It was, however, later discovered at the time of checking that he had copied down two pairs as $\begin{array}{c|c} x & y \\ \hline 6 & 14 \\ 8 & 6 \end{array}$ while the correct values were $\begin{array}{c|c} x & y \\ \hline 8 & 12 \\ 6 & 8 \end{array}$

Obtain the correct value of the correlation coefficient.

(*B.A., Punjab, 1952 ; B.A. Hons., Delhi, 1952*)

(*b*) In two sets of variables x and y with 50 observations each, the following data were observed :

$$\bar{x}=10,\ \sigma_x=3,\ \bar{y}=6,\ \sigma_y=2,\ r=\cdot3$$

But on subsequent verification, it was found that one value of x (=10) and one value of y (=6) were found inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of r affected.

(*B.A., Punjab, 1961 ; I.A.S., 1947*)

[Ans. (*a*) $\frac{1}{3}$ (*b*) $\bar{x}=13$, $\bar{y}=17$, $\sigma_y=4$, $r_{xy}=+{\cdot}6$]

(*c*) If x_1, x_2 and x_3 are three variates with means $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$ respectively and each variate occurs the same number of times, show that

$$r_{12}+r_{23}+r_{31}>-\tfrac{3}{2}$$

where r_{12} is the coefficient of correlation between x_1 and x_2.

[**Hint.** Let $A=\frac{1}{x}\Sigma\left(\frac{x_1-\bar{x}_1}{\sigma_1}+\frac{x_2-\bar{x}_2}{\sigma_2}+\frac{x_3-\bar{x}_3}{\sigma_3}\right)^2$

$=\frac{\sigma_1^2}{\sigma_1^2}+\frac{\sigma_2^2}{\sigma_2^2}+\frac{\sigma_3^2}{\sigma_3^2}+\frac{2}{\sigma_1\sigma_2}+r_{12}\sigma_1\sigma_2+1=3+2(r_{12}+r_{23}+r_{31})$]

(*M.A. Delhi, 1955 ; I.S.I., Calcutta*)

18. (*a*) If θ is the acute angle between the two regression lines with coefficient of correlation r, then $\tan\theta=\frac{1-r^2}{r}\cdot\frac{2\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$ and hence deduce that $\sin\theta\leqslant 1-r^2$.

(*b*) If the tangent of the angle between the lines of regression of y on x and x on y is 0·6 and the s.d. of y is twice the s.d. of x, find the value of the correlation between the two variables. [Ans. $r=\frac{1}{2}$]

19. (*a*) Two variables X and Y are known to be related to each other by the relation $Y=X/(aX+b)$. How is the theory of linear regression to be employed to estimate the constants a and b from a set of n pairs of observations (x_i, y_i), $i=1, 2, \ldots\ldots, n$. (*B.Sc. Sub. Delhi, 1960*)

[Hint. $\frac{1}{Y}=a+\frac{b}{X}\Rightarrow u=a+bY$ where $\frac{1}{X}=u$, $\frac{1}{Y}=v$.]

(*b*) Explain why there are two lines of regression. Write down two regression equations that may be associated with the following pairs of values :

x :	152	114	138	158	144	153	141	117	136	154
y :	198	300	414	594	676	549	320	483	481	659

(*I.A.S., 1951*)

(*c*) Obtain the equation of the line of regression of yield of rice on water from the data given in the following table which shows the amount of water applied in inches and the yield of rice in tons per acre in an experimental farm. Estimate the most probable yield of rice for 40 inches of water.

Water x	12	18	24	30	36	42	48
Yield y	5·27	5·68	6·25	7·21	8·02	8·71	8·42

(*B.A. Madras, 1953*)

[Ans. (*b*) $x=0{\cdot}03y+12{\cdot}693$, $y=3{\cdot}575x-34{\cdot}67$. (*c*) $y=0{\cdot}10x+4{\cdot}08$, 808 tons.]

20. If x_1, x_2 and x_3 are three variates measured from their respective means as origin and of equal variances, find the coefficient of correlation between x_1+x_3 and x_2+x_3 in terms of r_{12}, r_{13} and r_{23} and show that it is equal to

(*i*) $\frac{r_{12}+1}{2}$ if $r_{13}=r_{23}=0$ or (*ii*) $\frac{r_{12}+3}{4}$ if $r_{13}=r_{23}=1$.

(*B.Sc., Allahabad, 1968 ; I.S.I., Calcutta, 1956*)

21. Two judges in a beauty contest rank the ten competitors in the following order :

6	4	3	1	2	7	9	8	10	5
4	1	6	7	5	8	10	9	3	2

Do the two judges appear to agree in their standard ?

(*B.A., Punjab, 1952*)

[Ans. Rank correlation=0·224]

This indicates some sort of concordance between the standards of two judges.

22. The ranks of the same 16 students in two subjects A and B were as follows. Two numbers within brackets denote the ranks of the students in A and B respectively.

(1, 1), (2, 10), (3, 3), (4, 4), (5, 5), (6, 7), (7, 2), (8, 6), (9, 8), (10, 11), (11, 15), (12, 9), (13, 14), (14, 12), (15, 16), (16, 13).

Calculate the rank correlation for proficiencies of this group in subjects A and B. *(B.A., Punjab; M.Sc., Agra, 1952)*

[Ans. 0·8]

23. *(a)* If the variables $x_1, x_2, \ldots, x_{2n}$ all have the same variance σ^2 and the correlation coefficient between x_i and x_j $(i \neq j)$ has the same value, show that the correlation between $\sum_1^n x_i$ and $\sum_{n+1}^{2n} x_j$ is given by $n\rho/\{1+(n-1)\rho\}$.

(b) The independent random variables x and y are defined by

$$f(x)=4ax \qquad 0 \leqslant x \leqslant r$$
$$=0 \text{ elsewhere}$$
$$f(y)=4by \qquad 0 \leqslant y \leqslant s$$
$$=0 \text{ elsewhere.}$$

Find the coefficient of correlation between u and v in terms of a and b where $u=x+y$ and $v=x-y$.

$$\left[\text{Ans.} \quad (b)\ r_{uv}=\frac{b-a}{b+a}\right]$$

24. The n variates y_i satisfy the regression relations

$$y_i=\alpha+\beta(x_i-\bar{x})+e_i, \quad i=1, 2, \ldots\ldots, n$$

where α, β are population parameters, the x's are fixed and the e_i are independently distributed with mean zero and variance σ^2. Prove from the first principles that

(i) the least square estimate b is an unbiased estimate of β with variance $\sigma^2/\Sigma(x_i-\bar{x})^2$.

(ii) If $\eta_x=\alpha+\beta[x-\bar{x}]$ and $y_x=a+b(x-\bar{x})$, denote the population and sample regression lines, show that y_x is an unbiased estimate of η_x and calculate its variance.

The variate $u=\Sigma m_i y_i$ where m's are fixed numbers. If u is to be an unbiased estimate of β for all possible values of α and β, find the restrictions that must be satisfied by the m's. By choosing the m's so as to minimize the variance of u subject to the restrictions, prove that the smallest variance is reached where $u=b$. *(I.C.A.R., Delhi, 1947)*

$$\left[\text{Ans.} \quad Var\,(y_x)=\frac{\sigma^2}{n}+\frac{\sigma^2(x-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right.$$

Restrictions to be satisfied by m's are $\Sigma m_i=0$ $\Sigma m_i x_i=1$. $\left.\right]$

25. By minimizing $\Sigma f_i (x_i \cos \alpha+y_i \sin \alpha-p)^2$ for variations in α and p, show that there are two straight lines passing through the mean of the distri-

bution for which the sum of squares of normal deviations has an extreme value. Prove also that their slopes are given by

$$\tan 2\alpha=\frac{2\mu_{11}}{\sigma_x^2-\sigma_y^2}$$ *(M.A. and B.A. Hons., Madras)*

26. *(a)* The variates x and y are normally correlated and u, v are defined by $u=x\cos\alpha+y\sin\alpha$, $v=y\cos a-x\sin\alpha$.

Show that u, v will be uncorrelated if

$$\tan 2\alpha=\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}=\frac{2\mu_{11}}{\sigma_x^2-\sigma_y^2}.$$

(M.A. Punjab, 1958 ; M.A. Statistics, Delhi, 1958)

(b) Show further that, if in this case σ_u^2, σ_v^2 are their variances, then

$$\sigma_u\sigma_v=\sigma_x\sigma_y\sqrt{1-\rho^2},\ \sigma_u^2+\sigma_v^2=\sigma_x^2+\sigma_y^2.$$

27. If x and y are standard normal variates, with coefficient of correlation ρ, show that

(*i*) regression of y on x is linear,

(*ii*) $x+y$ and $x-y$ are independently distributed,

(*iii*) $Q=\dfrac{x^2-2\rho xy+y^2}{(1-\rho^2)}$ is distributed like a chi-square, *i.e.*, as that of the sum of squares of standard normal variates.

(M.A. Statistics, Delhi, 1959)

28. Show that, if x and y are independent normal variates with zero means, the points of inflexion of the curve of intersection of the normal correlation surface by planes through the z-axis, lie on the elliptic cylinder

$$\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}=1.$$

29. The bivariate random variables x_1, x_2 have a bivariate normal distribution with means 50 and 75 and s.d. 6 and 12 with a correlation coefficient 0·55. Find the following probabilities :

(*i*) $P(65\leqslant x_1\leqslant 75)$ (*ii*) $P(71\leqslant x_2<80/x_1=55)$

and (*iii*) $P(|x_1-x_2|\geqslant 15)$ *(B.Sc. Poona, 1971)*

30. From a standard bivariate normal population a sample of n values (x_1, y_1) ; $(x_2, y_2),\ldots\ldots,(x_n, y_n)$ is drawn. Show that the distribution of

$$Z_1=\frac{1}{n}\Sigma x^2 \text{ and } Z_2=\frac{1}{n}\Sigma y^2$$

has the moment generating function

$$\text{const.}\times\left\{\left(1-\frac{2t_1}{n}\right)\left(1-\frac{2t_2}{n}\right)-\frac{4\rho^2t_1t_2}{n^2}\right\}^{-\frac{n}{2}}$$ *(M.A. Patna, 1956)*

$$\left[\text{ Use : } \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\left[-(ax^2+2hxy+by^2)\right]dxdy=\pi/\sqrt{ab-h^2}\right.$$

and *m.g.f.* of $\frac{1}{n}x^2$ and $\frac{1}{n}y^2$ is $M(t_1, t_2)=E\left(\frac{t_1x^2}{n}+\frac{t_2y^2}{n}\right)$=etc. $\left.\right]$

31. Show that the curve along which the density function is constant for a bivariate normal distribution are the ellipses

$$\frac{x^2}{\sigma_x^2}-\frac{2\rho xy}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}=\lambda^2.$$

For the ellipse, show that *(B.A. Hons. Cal. 1968)*

(*i*) the line of regression of y on x is conjugate to y-axis.

(*ii*) the line of regression of x on y is conjugate to x-axis.

(*iii*) the principal axes are inclined to the x-axis at an angle θ where

$$\tan 2\theta=\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}.$$

32. In a bivariate normal population, the standard deviation of an x-array is k times that of x variate as a whole. Show that the correlation is $\sqrt{1-k^2}$.

[**Hint.** $\sigma_x\sqrt{1-\rho^2}=k\sigma_x$ $\quad\therefore\quad \rho=\sqrt{1-k^2}$]

33. If x and y follow the bivariate normal law with zero means and variances σ_1^2 and σ_2^2 respectively and correlation coefficient ρ, show that

(*i*) $u=\frac{x}{\sigma_1}+\frac{y}{\sigma_2}$ and $v=\frac{x}{\sigma_1}-\frac{y}{\sigma_2}$

are independently distributed with zero means and variance $2(1+\rho)$ and $2(1-\rho)$.

(M.Sc. Patna, 1963, M. Sc., Gujarat, 1969 ; B. Sc. Poona, 1971)

(*ii*) $E(x^2y^2)=\sigma_1^2\sigma_2^2(2\rho^2+1)$ *(M.A. Agra 1967)*

[**Hint.** $E(x_1^2x_2^2)$=coeff. of $t_1^2t_2^2$ on $M(t_1, t_2)$]

34. Show that x and y are standard normal variates with correlation coefficient ρ between them, then the correlation coefficient between X_1^2 and X_2^2 is ρ^2. *(M.Sc. Patna 1963 ; M Sc. Gujarat 1969)*

$$\left[\text{Hint. } E(X_1^2X_2^2)=2\rho^2+1,\ EX_1^4=\text{coeff. of } \frac{t_1^4}{45}=3=EX_2^4.\right.$$

$$\left.\rho_{x_1^2,\,x_2^2}=\frac{2\rho^2+1-1}{\sqrt{(3-1)(3-1)}}=\rho^2\right]$$

35. A bivariate distribution in two discrete variables, x and y is defined by probability generating function

$$\text{Exp } [a(u-1)+b\,(v-1)+c(u-1)\,(v-1)$$

Simultaneous probability of $x=r$, $y=s$ where r and s are integers being the coefficient of $u^r\,v^s$. Find the correlation coefficient of x and y.

(M. Sc., Poona, 1964)

[**Hint.** Put $u=e^{t_1}$, $v=e^{t_2}$ in p.g.f. and the result will be the M.G.F. so that $E(x)=\left(\frac{\partial M}{\partial t_1}\right)_{t_1=t_2=0}=a, E(x^2)=\left(\frac{\partial^2 M}{\partial t_1,\,\partial t_2}\right)_{t_1=t_2=0}=ab+c$

11

Multiple and Partial Correlation

11·1. Introduction. So far we have considered correlation between two variates only. But often it is necessary to find correlation between three or more variates. For example, the statures of men are influenced by those of all their ancestors, and the yield of grain is affected by the amounts of different fertilizers used.

Whenever we are interested in the combined influence of a group of variates upon a variate not included in the group, our study is that of *multiple regression* and *multiple correlation*.

In a multivariate population, the different variates may be mutually correlated and this correlation will, in general, be influenced by the other variates of the population. To study the relationship between any two variates, we have two methods. Firstly, we may consider only those members of the observed data in which the other members have specified values. Secondly, we may eliminate mathematically the effect of other variates on the two variates under study. The first method has the disadvantage that it limits the size of the data and also the results of this will be applicable to only those data in which the other variates have assigned values. In the second method it may not be possible to eliminate the entire influence of the other variates but we can easily eliminate the linear effects. The correlation between two variates when the linear effect of the other variates in them has been eliminated from both is called *partial correlation*.

11·2. Distribution of three variates. The theory of partial and multiple correlation was developed by Karl Pearson (1896) for three variates and then generalized by G. Udny Yule

(1897). In what follows, we shall use Yule's notation. Also for the sake of simplicity, we shall study the distribution of three variates only though the *arguments will apply to the case of n variates also.*

Suppose we are given N sets of corresponding values of three variates x_1, x_2 and x_3. Let these variates be *measured from their respective means* and the quantities so obtained be denoted by x_1, x_2 and x_3. The regression equation of x_1 say, on x_2 and x_3 may be written in the form

$$X_1 = a + b_{12\cdot3}x_2 + b_{13\cdot2}x_3 \qquad \ldots(11\cdot1)$$

where the constants a and b's are such as to give on the average the 'best' estimate of x_1 corresponding to any assigned values of x_2 and x_3. Thus we are to find a and b's such that

$$u = \Sigma(x_1 - X_1)^2 = \Sigma(x_1 - a - b_{12\cdot3}x_2 - b_{13\cdot2}x_3)^2 = \Sigma x^2_{1\cdot23}$$

where

$$x_{1\cdot23} = x_1 - a - b_{12\cdot3}x_2 \quad b_{13\cdot2}x_3$$

is a minimum, and the summation is taken place over all sets of values of x_2 and x_3.

The three normal equations for determining a and b's are

$$\left.\begin{aligned} \Sigma(x_1 - a - b_{12\cdot3}\,x_2 - b_{13\cdot2}\,x_3) &= 0 \\ \Sigma x_2(x_1 - a - b_{12\cdot3}\,x_2 - b_{13\cdot2}\,x_3) &= 0 \\ \Sigma x_3(x_1 - a - b_{12\cdot3}\,x_2 - b_{13\cdot2}\,x_3) &= 0 \end{aligned}\right\}$$

or

$$\left.\begin{aligned} \Sigma x_{1\,23} &= 0 \\ \Sigma x_2 x_{1\cdot23} &= 0 \\ \Sigma x_3 x_{1\cdot23} &= 0 \end{aligned}\right\} \qquad \ldots(11\cdot2)$$

The first of these equations gives $a = 0$ and the last two equations may be written in the form

$$\Sigma x_1 x_2 - b_{12\,3}\,\Sigma x_2^2 - b_{13\cdot2}\,\Sigma x_2 x_3 = 0,$$
$$\Sigma x_1 x_3 - b_{13\cdot3}\,\Sigma x_2 x_3 - b_{13\cdot2}\,\Sigma x_3^2 = 0$$

Expressing the summation in terms of standard deviations and correlation coefficients, we have

$$N r_{12}\sigma_1\sigma_2 = N b_{12\cdot3}\sigma_2^2 + N b_{13\cdot2} r_{23}\sigma_2\sigma_3$$
$$N r_{13}\sigma_1\sigma_3 = N b_{12\cdot3} r_{23}\sigma_2\sigma_3 + N b_{13\cdot2}\sigma_3^2$$

or

$$\left.\begin{aligned} r_{12}\sigma_1 &= b_{12\cdot3}\sigma_2 + b_{13\cdot2} r_{23}\sigma_3 \\ r_{13}\sigma_1 &= b_{12\cdot3} r_{23}\sigma_2 + b_{13\cdot2}\sigma_3 \end{aligned}\right\} \qquad \ldots(11\cdot3)$$

where r_{ij} is the coefficient of correlation between x_i and x_j for the N given pairs of values and σ_j is the standard deviation of x_j for the N values.

11·3. Determination of Regression Coefficient. Equation to Regression Plane. Solving equations (11·3) for $b_{12\cdot3}$ and $b_{13\cdot2}$, we obtain

$$b_{12\cdot3}=\begin{vmatrix} r_{12}\sigma_1 & r_{13}\sigma_3 \\ r_{13}\sigma_1 & \sigma_3 \end{vmatrix} \div \begin{vmatrix} \sigma_2 & r_{23}\sigma_3 \\ r_{23}\sigma_2 & \sigma_3 \end{vmatrix}$$

or
$$b_{12\cdot3}=-\frac{\begin{vmatrix} -\sigma_1 & r_{12} & r_{23} \\ \sigma_2 & r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}}=-\frac{\sigma_1}{\sigma_2}\frac{\triangle_{12}}{\triangle_{11}} \qquad \ldots(11\cdot4)$$

and
$$b_{13\cdot2}=-\frac{\begin{vmatrix} -\sigma_1 & 1 & r_{12} \\ \sigma_3 & r_{23} & r_{13} \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}}=-\frac{\sigma_1}{\sigma_3}\frac{\triangle_{13}}{\triangle_{11}} \qquad \ldots(11\cdot5)$$

where $\triangle_{ij}$ is the co-factor of the element in the ith row and jth column in the determinant

$$\triangle=\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

Hence on substituting the values of $b_{12\cdot3}$ and $b_{13\cdot2}$ the equation to the regression plane of x_1 on x_2 and x_3 is

$$X_1=\left[-\frac{\sigma_1}{\sigma_2}\frac{\triangle_{12}}{\triangle_{11}}\right]x_2+\left[-\frac{\sigma_1}{\sigma_3}\frac{\triangle_{13}}{\triangle_{11}}\right]x_3 \qquad \ldots(11\cdot6)$$

Note. The quantity $x_{1\cdot23}=x_1-b_{12\cdot3}x_2-b_{13\cdot2}x_3$ is called the *residual* of second order. The order of the residual is the number of subscripts after the point. The quantities b's are called *partial regression coefficients.* The first subscript attached to the b's is the subscript of the letter on the left (the dependent variate), the second subscript is that of the x to which it is attached. These are called *primary* subscripts. The subscripts separated from the primary subscripts by a point are called the *secondary* subscripts.

A standard deviation denoted by a symbol with p secondary subscripts will be termed as *standard deviation of the pth order* the standard deviations σ_1, σ_2, etc., being regarded as of order zero ; the standard deviations $\sigma_{1\cdot23}$, $\sigma_{2\cdot13}$ etc., of the second order and so on.

Ex. 1. Show that, if the origin is not necessarily at the mean, the equation of the plane of regression is

$$\begin{vmatrix} \frac{x_1-\bar{x}_1}{\sigma_1} & \frac{x_2-\bar{x}_2}{\sigma_2} & \frac{x_3-\bar{x}_3}{\sigma_3} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 0$$

or
$$\frac{x_1-\bar{x}_1}{\sigma_1}\Delta_{11} + \frac{x_2-\bar{x}_2}{\sigma_2}\Delta_{12} + \frac{x_3-\bar{x}_3}{\sigma_3}\Delta_{13} = 0$$

Ex. 2. Write the equations of the plane of regression of x_2 on x_1 and x_3, and of x_4 on x_1 and x_2.

Ex. 3. For p variates, show that the equation of the regression hyperplane of x_i on the rest of the variates is

$$\frac{x_1-\bar{x}_1}{\sigma_1}\Delta_{i1} + \frac{x_2-\bar{x}_2}{\sigma_2}\Delta_{i2} + \cdots\cdots + \frac{x_p-\bar{x}_p}{\sigma_p}\Delta_{ip} = 0.$$

11·4. Properties of the Residuals.

I. Let the equation to the plane of regression of x_1 on x_2 and x_3 be

$$X_1 = b_{12\cdot 3}x_2 + b_{13\cdot 2}x_3$$

Then the normal equations for determining the b's are

$$\Sigma x_2 x_{1\cdot 23} = 0 = \Sigma x_3 x_{1\cdot 23}$$

Similarly equations for the regression of x_2 on x_1 and x_3 and on x_1 and x_2 are

$$\Sigma x_1 x_{2\cdot 13} = 0 = \Sigma x_3 x_{2\cdot 13}$$

$$\Sigma x_2 x_{3\cdot 12} = 0 = \Sigma x_1 x_{3\cdot 12}$$

These normal equations express the following result :

The sum of the products of corresponding values of a variate and a residual is zero (or the variance of the variate and the residual is zero), provided the subscript of the variate occurs among the secondary subscripts of the residual.

II. Writing $x_{1\cdot 2} = x_1 - b_{12}x_2$, we get by virtue of normal equations,

$$\Sigma x_{1\cdot 23}x_{1\cdot 2} = \Sigma x_{1\cdot 23}(x_1 - b_{12}y_2) = \Sigma x_{1\cdot 23}x_1.$$

and
$$\Sigma x_{1\cdot 23}x_{1\cdot 23} = \Sigma x_{1\cdot 23}(x_1 - b_{12\cdot 3}x_2 - b_{13\cdot 2}x_3) = \Sigma x_{1\cdot 23}x_1.$$

Thus *the sum of the products (or covariance) of two residuals is unaltered by omitting from one residual any or all of the secondary subscripts which are common to both.*

III. By virtue of normal equations, we have

$$\Sigma x_{3\cdot 2}x_{1\cdot 23}=\Sigma(x_3-b_{32}x_2)x_{1\cdot 23}=0$$

Similarly, $$\Sigma x_{2\cdot 3}x_{1\cdot 23}=0$$

Thus it follows that *the sum of the products (or covariance) of two residuals is zero provided all the subscripts or one residual occur among the secondary subscrips of the second.*

Ex. 1. Prove that (*i*) $\Sigma x_2 x_{1\cdot 23}=0$, (*ii*) $\Sigma x_3 x_{1\cdot 23}=0$

Ex. 2. For the distribution of p variates, prove the following results :

(*i*) $\Sigma x_k x_{1\cdot 23\cdots p}=0, \qquad k=2, 3,\ldots\ldots p.$

(*ii*) $\Sigma x_{1\cdot 34\cdots p}x_{2\cdot 34\cdots p}=\Sigma x_{1\cdot 34\cdots p}x_2=\Sigma x_1 x_{2\cdot 34\cdots p}$

Ex. 3. Prove that

(*i*) $\Sigma x_{2\cdot 34\cdots p}x_{1\cdot 23\cdots p}=0$

(*ii*) $\Sigma x_{2\cdot 34\cdots p}(x_1-b_{12\cdot 4\cdots p}x_2-)=0$

Deduce that $b_{2\cdot 24\cdots p}=(\Sigma x_{2\cdot 34\cdots}\, p^{x}{}_{1\cdot 24\cdots p})\ \Sigma(x_{2\cdot 34\cdots p})^2$ so that if q stands for the group of suffixes 3, 4,..., p

$$b_{12\cdot q}=\frac{\text{cov}\,(x_{1\cdot q},\,(x_{2\cdot q})}{\text{var}\,(x_{2\cdot q})}\,.$$

Similarly, show that

$$b_{31\cdot q}=\frac{\text{cov}\,(x_{1\cdot q},\,x_{3\cdot q})}{\text{var}\,(x_{1\cdot q})}.$$

11·5. Variance of a Residual. We shall now obtain formula for the variance $\sigma^2{}_{1\cdot 23}$ of the residual $x_{1\cdot 23}$ (the deviation of the observed value x_1 from its corresponding computed value on the regression plane), in terms of $\sigma_1{}^2$ and the correlation coefficients.

We have

$$\begin{aligned}N\sigma^2{}_{1\cdot 23}&=\Sigma x^2{}_{1\cdot 23}=\Sigma x_1 x_{1\cdot 23}\\&=\Sigma x_1(x_1-b_{12\cdot 3}x_2-b_{13\cdot 2}x_3)\\&=N\sigma_1{}^2-Nb_{12\cdot 3}\sigma_1\sigma_2 r_{12}-Nb_{13\cdot 2}\sigma_1\sigma_2 r_{13}\end{aligned}$$

or $$\sigma_1\left(1-\frac{\sigma^2{}_{1\cdot 23}}{\sigma_1{}^2}\right)=b_{12\cdot 3}\sigma_2 r_{12}+b_{13\cdot 2}\sigma_3 r_{13}$$

Eliminating $b_{12\cdot 3}$ and $b_{13\cdot 2}$ between this equation and equations (11·3), we have

$$\begin{vmatrix}1-\dfrac{\sigma^2{}_{1\cdot 23}}{\sigma_1{}^2} & r_{12} & r_{13}\\ r_{21} & 1 & r_{23}\\ r_{31} & r_{32} & 1\end{vmatrix}=0$$

or $$\triangle - \frac{\sigma^2_{1\cdot 23}}{\sigma_1^2}\triangle_{11} = 0$$

Hence $$\sigma^2_{1\cdot 23} = \sigma^2_1 \cdot \frac{\triangle}{\triangle_{11}} \qquad \ldots(11\cdot 7)$$

Thus the variance of a residual of order 2 is expressible in terms of variance of zero order and correlations of zero order.

Show that the standard error of estimating x_1 from assigned terms of value of x_2 and x_3 is given by

$$\sigma_1 \left(\frac{\triangle}{\triangle_{11}}\right)^{1/2}$$

where $\triangle_{11}$ is the cofactor of r_{11} in

$$\triangle = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix}$$

Ex. Show that, for the distribution of p variates,

$$\sigma^2_{1\cdot 23\ldots p} = \frac{\triangle}{\triangle_{11}}\sigma^2_1$$

where $$\triangle = \begin{vmatrix} 1 & r_{12} & r_{13}\cdots\cdots r_{1p} \\ r_{21} & 1 & r_{23}\cdots\cdots r_{2p} \\ \cdots & \cdots & \cdots \quad \cdots \\ r_{p1} & r_{p2} & r_{p3}\cdots\cdots 1 \end{vmatrix}$$

and $\triangle_{11}$ is the minor of the element in the first row and first column of $\triangle$.

11·6. Multiple Correlation Coefficient. Consider the regression equation of x_1 on x_2 and x_3 *viz.*,

$$X_1 = b_{12\cdot 3}x_2 + b_{13\cdot 2}x_3 = x_1 - x_{1\cdot 23}$$

Next the correlation between x_1 (observed value of the variate) and X_1 (expected value of the variate) is given by

$$R_{1(23)} = \frac{\Sigma x_1 X_1}{\sqrt{(\Sigma x_1^2)(\Sigma X_1^2)}}$$

We have $\Sigma x_1 X_1 = \Sigma\{x_1(x_1 - x_{1\cdot 23})\} = \Sigma x_1^2 - \Sigma x_1 x_{1\cdot 23}$

$= \Sigma x_1^2 - \Sigma x^2_{1\cdot 23}$ by property II

$= N(\sigma_1^2 - \sigma^2_{1\cdot 23})$

Also $\Sigma X_1^2 = \Sigma(x_1 - x_{1\cdot 23})^2 = \Sigma x_1^2 - 2\Sigma x_1 x_{1\cdot 23} + \Sigma x^2_{1\cdot 23}$

$= \Sigma x_1^2 - \Sigma x^2_{1\cdot 23}$

$= N(\sigma_1^2 - \sigma^2_{1\cdot 23})$ $[\because \Sigma x_1 x_{1\cdot 23} = x^2_{1\cdot 23}]$

$$\therefore \qquad R_{1(23)}=\frac{\sigma_1^2-\sigma^2_{1\cdot 23}}{\sigma_1\sqrt{\sigma_1^2-\sigma^2_{1\cdot 23}}}=\left(1-\frac{\sigma^2_{1\cdot 23}}{\sigma_1^2}\right)^{\frac{1}{2}} \qquad ...(11\cdot 7')$$

or $$1-R^2_{1(23)}=\frac{\sigma^2_{1\cdot 23}}{\sigma_1^2}=\frac{\Delta}{\Delta_{11}} \quad \Rightarrow \quad R^2_{1(23)}=\sigma_1^2-\sigma^2_{1\cdot 23} \qquad ...(11\cdot 8)$$

From (11·7), $\sigma_1^2 \geqslant \sigma^2_{1\cdot 23}$ and therefore from (11·7'), $R_{1(23)} \geqslant 0$.

Note 1. $R_{1(23)}$ is necessarily positive or zero.

2. If $R_{1(23)}=1$, then from (11·8) we have $\sigma^2_{1\cdot 23}=0$, *i.e.*, all the residuals $x_{1\cdot 23}$ are zero, the observed and expected value of x_1 coincide, and thus the observed x_1 is a linear function of x_2 and x_3.

Ex. 1. Establish the following alternative form for $R_{1(23\cdots p)}$' the multiple correlation coefficient between x_1 and $x_2, x_3, \ldots\ldots, x_p$.

(*i*) $$R^2_{1(23\cdots p)}=1-\frac{\sigma^2_{1\cdot 23\cdots p}}{\sigma_1^2}$$

(*ii*) $$1-R^2_{1(23\cdots p)}=\frac{\Delta}{\Delta_{11}}$$

where Δ and Δ_{11} have the usual meanings.

(*iii*) $$R_{1(23)}=b_{12\cdot 3}r_{12}\frac{\sigma_2}{\sigma_1}+b_{13\cdot 2}r_{13}\frac{\sigma_3}{\sigma_1}$$

Ex. 2. For a trivariate population prove that the coeff. of multiple correlation $R_{1(23)}$ satisfies the following relation :

$$R^2_{1(23)}=\frac{r_{12}^2+r_{13}^2-2r_{12}r_{23}r_{31}}{1-r_{23}^2} \geqslant r_{12}^2.$$

(*M.A. Statistics, Delhi, 1962, 65, 68*)

11·7. Partial Correlation Coefficient. Let us find the partial correlation coefficient $r_{12\cdot 3}$ which may be defined as the correlation coefficient between $x_{1\cdot 3}$ and $x_{2\cdot 3}$ *i.e.*, between x_1 and x_2 after the linear effect of the third variate x_3 has been eliminated from both x_1 and x_2.

We have, by property III,

$$\begin{aligned} 0=\Sigma x_{2\cdot 3}x_{1\cdot 23}&=\Sigma x_{2\cdot 3}(x_1-b_{12\cdot 3}x_2-b_{13\cdot 2}x_3)\\ &=\Sigma x_1x_{2\cdot 3}-b_{12\cdot 3}\Sigma x_{2\cdot 3}x_2\\ &=\Sigma x_{1\cdot 3}x_{2\cdot 3}-b_{12\cdot 3}\Sigma x^2_{2\cdot 3}\end{aligned}$$

This gives $$b_{12\cdot 3}=\frac{\Sigma x_{1\cdot 3}x_{2\cdot 3}}{\Sigma x^2_{2\cdot 3}} \qquad ...(11\cdot 9)$$

This shows that $b_{12\cdot 3}$ is the coefficient of regression of $x_{1\cdot 3}$ on $x_{2\cdot 3}$. Similarly $b_{21\cdot 3}$ is the coefficient of regression of $x_{2\cdot 3}$ on $x_{1\cdot 3}$ so that $r_{12\cdot 3}$, the coefficient of correlation between $x_{1\cdot 3}$ and $x_{2\cdot 3}$ is given by

$$r^2_{12\cdot 3} = b_{12\cdot 3} b_{21\cdot 3}$$

$$= \frac{\Delta_{12}}{\Delta_{11}} \cdot \frac{\Delta_{21}}{\Delta_{22}} = \frac{\Delta_{12}^2}{\Delta_{11}\Delta_{22}}$$

Since $r_{12\cdot 3}$ has the same sign as $b_{12\cdot 3}$ which is the sign of $-\Delta_{12}$ by (11·4), we have

$$r_{12\cdot 3} = -\frac{\Delta_{12}}{\sqrt{\Delta_{11}\Delta_{22}}} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \qquad \ldots(11\cdot 10)$$

Note. Formula for $r_{23\cdot 1}$ and $r_{31\cdot 2}$ can be obtained similarly. We observe that the partial correlation coefficients of the first order can be expressed in terms of the total correlation coefficients.

Cor. Since $r^2_{1\cdot 23} \leqslant 1$, we have

$$(r_{12} - r_{13} r_{23})^2 \leqslant (1 - r_{13}^2)(1 - r_{23}^2)$$

that is $$r_{12}^2 + r_{13}^2 + r_{23}^2 - 2 r_{12} r_{13} r_{23} \leqslant 1$$

If we take r_{12}, r_{13} as known, this gives as limits for r_{23} :

$$r_{12} r_{13} \pm \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2} \qquad \ldots(11\cdot 11)$$

Note. Writing down similar equations for $b_{lm\cdot qp}$ and $b_{ml\cdot qp}$ multiplying and then taking square root, we can show that

$$r_{lm\cdot qp} = \frac{r_{lmq} - r_{lp\cdot q} r_{mp\cdot q}}{\sqrt{(1 - r_{lp\cdot q}^2)(1 - r^2_{mp\cdot q})}} \qquad \ldots(11\cdot 12)$$

l, m, q, p are unequal.

given the correlation coefficient in terms of those of lower order.

Ex. 1. Defining $r_{12\cdot q}$, the product moment coefficient of correlation between $x_{1\cdot q}$ and $x_{2\cdot q}$ as

$$r_{12\cdot q} = (b_{12\cdot q} b_{21\cdot q})^{\frac{1}{2}}$$

and using Ex. 3, page 321, show that

$$r_{12\cdot q} = \frac{\text{cov}\,(x_{1\cdot q}, x_{2\cdot q})}{[\text{var}(x_{1\cdot q})\ \text{var}(x_{2\cdot q})^{\frac{1}{2}}]}.$$

Ex. 2. Prove that for a trivariate distribution,

(*i*) $b_{12\cdot 3} = \dfrac{b_{12} - b_{13} b_{32}}{1 - b_{23} b_{32}}$ *(M.A. Bombay, 1956)*

(*ii*) $r_{12\cdot 3} \dfrac{\sigma_{1\cdot 3}}{\sigma_{2\cdot 3}} = \dfrac{\sigma_1}{\sigma_2} \cdot \left(\dfrac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \right)$.

Ex. 11·1. *If* $r_{12}=+0{\cdot}80$, $r_{13}=-0{\cdot}40$, $r_{23}=-0{\cdot}56$, *find the values of* $r_{12\cdot3}$, $r_{13\cdot2}$ *and* $r_{23\cdot1}$. *(M.A. Statistics, Delhi, 1965)*

We have
$$r_{12\cdot3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$
$$=\frac{0{\cdot}80-(-0{\cdot}40)(-0{\cdot}56)}{\sqrt{\{1-(0{\cdot}40)^2\}\{1-(0{\cdot}56)^2\}}}$$
$$=0{\cdot}759$$

Similarly,
$$r_{13\cdot2}=\frac{r_{13}-r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}}$$
$$=0{\cdot}097$$

and
$$r_{23\cdot1}=-0{\cdot}436.$$

Ex. 1. If $r_{12}=0{\cdot}86$, $r_{13}=0{\cdot}65$, $r_{23}=0{\cdot}72$, show that $r_{12\cdot3}=0{\cdot}743$.

Ex. 2. Show that the values $r_{12}=\cdot6$, $r_{13}=-\cdot4$, $r_{23}=\cdot7$ are inconsistent.

Ex. 11·2. *If the correlation coefficient of zero order in a set of* p *variates were equal to* ρ, *show that*

(a) every partial correlation of the sth order is $\rho/(1+s\rho)$; ***and***

(b) the coefficients of multiple correlation R *of a variate with the other* $(p-1)$ *variates is given by*

$$1-R^2=(1-\rho)\,\frac{[1+(p-1)\rho]}{1+(p-2)\rho}.$$

(M.A., Delhi, 1953, 55, 57 ; I.C.A·R., Delhi, 1947 ; M.A., Bombay, 1958 ; M.A., Nagpur, 1953)

(a) We are given that
$$r_{ij}=\rho,\ i,j=1,2,\ldots\ldots,p,\ i\neq j$$

We have
$$r_{lm\cdot h}=\frac{r_{lm}-r_{lh}r_{mh}}{\sqrt{(1-r^2_{mh})(1-r^2_{lh})}}\quad l,\ m,\ h=1,2,\ldots,p,\ l\neq m\neq h$$
$$=\frac{\rho-\rho.\rho}{\sqrt{(1-\rho^2)(1-\rho^2)}}=\frac{\rho}{1+\rho}$$

Thus every partial correlation of the 1st order is $\rho/(1+\rho)$.

$$r_{lm\cdot hk}=\frac{r_{lm\cdot k}-r_{lh\cdot k}r_{mh\cdot k}}{\sqrt{(1-r^2_{lh\cdot k})(1-r^2_{mh\cdot k})}}$$

$$=\frac{\frac{\rho}{1+\rho}-\frac{\rho}{1+\rho}\cdot\frac{\rho}{1+\rho}}{\sqrt{\left[1-\left(\frac{\rho}{1+\rho}\right)^2\right]\left[1-\left(\frac{\rho}{1+\rho}\right)^2\right]}}=\frac{\rho}{1+2\rho}$$

that is every partial correlation coefficient of the 2nd order is $\rho/(1+2\rho)$.

Now assuming that every partial correlation coefficient of orders is $\rho/(1+s\rho)$, the partial correlation coefficient of order $s+1$ would be

$$=\frac{\frac{\rho}{1+s\rho}-\frac{\rho}{1+s\rho}\cdot\frac{\rho}{1+s\rho}}{\sqrt{1-\left(\frac{\rho}{1+s\rho}\right)^2}\sqrt{1-\left(\frac{\rho}{1+s\rho}\right)^2}}=\frac{\rho}{1+(s+1)\rho},$$

Thus the result is proved by *Induction.*

(*b*) Since all the arguments hold good for a distribution of n variates, we have

$$1-R^2=\frac{\Delta}{\Delta_{11}}$$

where Δ is the determinant which has its jth element of the ith row equal to $r_{ij}(i, j=1, 2,\ldots\ldots,p)$ $r_{ij}=1$ and Δ_{11} is the determinant obtained by omitting the first row and the first column of Δ.

Since all the correlation coefficients of zero order are equal to ρ, we have

$$\Delta=\begin{vmatrix}1 & \rho\ldots\ldots\rho\\ \rho & 1\ldots\ldots\rho\\ \ldots\ldots & \ldots\ldots\\ \rho & \rho\ldots\ldots 1\end{vmatrix}=[1+(p-1)\rho](1-\rho)^{p-1}$$

and then clearly $\Delta_{11}=[1+(p-2)\rho](1-\rho)^{p-2}$.

$$\text{Hence } 1-R^2=(1-\rho)\frac{[1+(p-1)\rho]}{1+(p-2)\rho}.$$

[For another proof of this result, see Ex. 11, page 327.]

EXERCISES 11

1. Prove the identity :

$$b_{12\cdot3}\,b_{23\cdot1}\,b_{31\cdot2}=r_{12\cdot3}\,r_{23\cdot1}\,r_{31\cdot2} \qquad (B.A.\ Hons.,\ Delhi,\ 1954)$$

[**Hint.** $b_{12\cdot3}=\frac{\Sigma x_{1\cdot3}x_{2\cdot3}}{\Sigma x_{2\cdot3}^2}=\frac{r_{12\cdot3}\sigma_{1\cdot3}}{\sigma_{2\cdot3}}$. Similarly for $b_{23\cdot1}$ and $b_{31\cdot2}$]

2. Show that $\sigma_{1\cdot23}\,\sigma_{2\cdot13}/r_{12\cdot3}=-\Delta\sigma_1\sigma_3/\Delta_{12}$.

3. It is possible to get the following from a set of experimental data ?

$r_{12}=0{\cdot}6$, $r_{23}=0{\cdot}8$, $r_{31}=-0{\cdot}5$. [**Ans.** No]

4. If the relation $ax_1+bx_2+cx_3=0$ holds for all sets of values of x_1, x_2 and x_3, what must be the partial correlations ? (*M.A. Calcutta*)

[**Ans.** $r_{12\cdot3}=r_{31\cdot2}=r_{23\cdot1}=-1$]

5. Show that if $x_3=ax_1+bx_2$, the three partial correlations are numerically equal to unity, $r_{13\cdot2}$ having the sign of a, $r_{23\cdot1}$ the sign of b and $r_{12\cdot3}$ the opposite sign of a/b. (*M.A., Bombay, 1956 ; B.A. Hons., Delhi, 1953*)

6. X, Y, Z are independent random variables with the same variance.

If $$X_1=\frac{1}{\sqrt{2}}(X-Z)$$

$$X_2=\frac{1}{\sqrt{3}}(X+Y+Z)$$

$$X_3=\frac{1}{\sqrt{6}}(X+2Y+Z)$$

show that X_1, X_2, X_3 have equal variances. Calculate $R_{12\cdot3}$ and $R_{12(3)}$.

(M.Sc., Rajasthan, 1966)

7. Show that correlation coefficient between the residuals $x_{1\cdot23}$ and $x_{2\cdot13}$ is equal and opposite to that between $x_{1\cdot3}$ and $x_{2\cdot3}$.

(M.A., Delhi, 1956, 59 ; I.C.A.R., Delhi, 1951)

Hint. Coeff. of correlation between $x_{2\cdot13}$ and $x_{1\cdot23}$

$=(1/N)\,\Sigma x_{2\cdot13}x_{1\cdot23}/\sigma_{2\cdot13}\sigma_{1\cdot23}=(1/N)\,\Sigma x_{2\cdot13}\,(x_1-b_{12\cdot3}x_2-b_{13\cdot2}x_3)/\sigma_{2\cdot13}\sigma_{1\cdot23}$
$=0-b_{12\cdot3}\Sigma x_{2\cdot13}\div(1/N)\sigma_{2\cdot13}\sigma_{1\cdot23}=$etc.

8. *(a)* Show that for given r_{12} and r_{13}, r_{23} must lie in the range

$$r_{12}r_{13}\pm(1-r_{12}^2-r_{13}^2+r_{12}^2r_{13}^2)^{\frac{1}{2}}$$

(I.C.A.R., Delhi, 1951 ; B.A. Hons., Calcutta, 1954 ; M.A., Delhi, 1954)

(b) If $r_{12}=k$, $r_{23}=-k$, show that r_{13} will lie between -1 and $1-2k^2$.

(I.A.S. Calcutta, 1953)

[Hint. $r^2_{12\cdot3}\leqslant1$ gives on simplification $r_{13}^2+2k^2r_{13}+(2k^2-1)\leqslant0$.**]**

9. Show that $R_1(_{23})\geqslant r_{12}$. What is the relation between $R_1(_{23})$ and r, the coefficient of correlation between two variables ? *(M.A. Delhi, 1971)*

[Hint. $N\sigma^2_{1\cdot23}=\Sigma x^2_{1\cdot23}=\Sigma x_{1\cdot23}(x_1-b_{12\cdot3}x_2-b_{13\cdot2}x_3)$
$=\Sigma x^2_{1\cdot2}-b_{13\cdot2}\Sigma x_{1\cdot2}x_{3\cdot2}$ using Property II
$=N\sigma^2_{1\cdot2}(1-b_{13\cdot2}b_{31\cdot2})$
$\sigma^2_{1\cdot23}=\sigma_1^2(1-r_{12}^2)(1-r^2_{13\cdot2})$

From equation 11·8 $\sigma^2_{1\cdot23}=\sigma_1^2[1-R_1^2(_{23})]$

$\therefore$ $1-R_1^2(_{23})=(1-r_{12}^2)(1-r^2_{13\cdot2})\leqslant1-r_{12}^2$. Hence etc.**]**

10. Establish the following results:

(i) $\sigma^2_{1\cdot23\cdots p}=\sigma_1^2(1-r_{12}^2)(1-r^2_{13\cdot2})\ldots\ldots(1-r^2_{1p.23\ldots\overline{p-1}})$

(ii) $1-R_1^2(_{23\cdots p})=(1-r_{12}^2)(1-r^2_{13\cdot2})\ldots\ldots(1-r^2_{1p.23\ldots\overline{p-1}})$

(B.A. Hons., Calcutta, 1974)

Hence show that if the multiple correlation coefficient $R_1(_{23\cdots p})$ is equal to zero x_1 is uncorrelated with any of the other variables. *(M.A., Madras 1979)*

11. If all the correlation coefficients of zero order in a set of p variables were equal to ρ, then

(i) every partial correlation of the sth order is $\rho/(1+s\rho)$

and *(ii)* $1-R_1^2(_{23\cdots p})=\dfrac{(1-\rho)[1+(p-1)\rho]}{1+(p-2)\rho}$.

[Hint. *(i)* In Ex. 10 *(ii)* above put $r_{12}=\rho$, $r_{13\cdot2}=\dfrac{\rho}{1+\rho}$, etc. and simplify**]**

12. Show that

$$1-r^2_{1p.23\ldots\overline{p-1}} = [1-R^2_{1(23\ldots\underline{p})}]/[1-R^2_{1\ (23\ldots\overline{p-1})}]$$

which show how a partial correlation coefficient of order $p-2$ may be expressed in terms of multiple correlation coefficient of order $p-1$ and $p-2$.

(B.A. Hons., Calcutta, 1954)

13. Show that for p variate there are ${}^pC_2$ correlation coefficients of order zero and $({}^{p-2}C_2)({}^pC_2)$ of order s. Show further that there are $({}^pC_2)2^{p-2}$ correlation coefficients and altogether $({}^pC_2)2^{p-1}$ regression coefficients.

14. Show that if r_{12} is zero, $r_{12.3}$ will not be zero unless one at least of r_{13} and r_{23} is zero.

15. Prove the formulae:

(*a*) $$b_{12}=\frac{b_{12.3}+b_{13.2}b_{21.3}b_{32.1}}{1-b_{13.2}b_{23.1}}$$

(*b*) $$r_{12}=\frac{r_{12.3}+r_{13.2}\,r_{23.1}}{\sqrt{(1-r^2_{13.2})(1-r^2_{23.1})}}$$

16. If x_1, x_2 and x_3 are three variates measured from their respective means as origin and if e_1 is the expected value of x_1 for given values of x_2 and x_3 from the linear regression of x_2 and x_3, prove that

$$\text{Cov}\,(x_1, e_1)=\text{Var}\,(e_1)=\text{Var}\,(x_1)=\text{Var}(x_1-e_1).$$

(I.S.I. Calcutta 1956)

17. Prove that the necessary and sufficient condition for the coincidence of the three regression planes for a trivariate distribution is

$$r_{12}^2+r_{13}^2+r_{23}^2-2r_{12}r_{13}r_{23}=1$$

18. If $r_{23}=1$, prove that

(*i*) $r_{12}^2=r_{13}^2$.

(*ii*) $\sigma^2_{1.23}=\sigma_1^2(1-r_{12}^2)$.

19. (*a*) If $r_{23}=0$, prove that

(*i*) $R^2_{1(23)}=r_{12}^2+r_{13}^2$.

(*ii*) $\sigma^2_{1.23}=\sigma_1^2(1-r_{12}^2-r_{13}^2)$.

(*b*) If $R_{1.23}=0$, prove that

$$r_{12}=r_{13}=0$$ *(B.Sc. Poona, 1971)*

Hint. $1-R^2_{1(23)}=(1-r^2_{12})(1-r^2_{13.2})$

20. If $x_1=y_1+y_2$, $x_2=y_2+y_3$, $x_3=y_3+y_1$,

where y_1, y_2, y_3 are uncorrelated variables and each of which has zero mean and unit *s.d.*, find the multiple correlation coefficient between x_1 and two variables x_2 and x_3. *(M.A. Delhi, 1961)*

Ans. $$R_{1.23}=\left(1-\frac{\Delta}{\Delta_1}\right)^{\frac{1}{2}}=\frac{1}{\sqrt{3}}$$

21. $x_0, x_1, \cdots, x_n$ are $(n+1)$ variates. Obtain a linear function of $x_1, x_2, \cdots, x_n$ which have a maximum correlation with x_0. Show that the correlation R of x_0 with the linear function is given by

$$R=\left(1-\frac{\Delta}{\Delta_{11}}.\right)^{\frac{1}{2}}$$

where Δ_1, Δ_{11} have usual significance. *(I.S.I., 1967)*

22. For a trivariate distribution, show that

$$1+2r_{12}\,r_{13}\,r_{23} \geqslant r_{12}^2+r_{23}^2+r_{13}^2$$

where r_{12} is correlation coefficient between the random variables x_1 and x_2, etc.

(B.Sc. Hons. Poona, 1960, I.S.I. 1968)

23. Five hundred students were examined in three subjects I, II and III, each subject carrying 100 marks. A student getting 120 or more but less than 150 marks was put in pass class, a student getting 150 or more but less than 180 was put in second class and a student getting 180 or more marks was put in first class. The following results were obtained :

	I	*II*	*II*
Mean	35·8	52·4	48·8
S.D.	4·2	5·3	6·1
Corr. Coeff.	$r_{12}=0{\cdot}6$	$r_{13}=0{\cdot}7$	$r_{23}=0{\cdot}8$

(*i*) Find the number of students in each of three classes.

(*ii*) Estimate the number of students with total marks lying between 180 and 190.

(*iii*) Calculate the proportion of students who failed.

(*iv*) If r_{23} is not known, obtain the limits within which it may lie when r_{12} and r_{13} have the given values.

(I.C.A.R., Delhi, 1947, M.A. Statistics. Delhi, 1967)

Solution. Let Z denote the total marks of the students in the three subjects and X_1, X_2, X_3 the total marks of the students in subjects I, II and III respectively.

$$Z=X_1+X_2+X_3$$

$$\therefore \quad EZ=EX_1+EX_2+EX_3=35{\cdot}8+52{\cdot}4+48{\cdot}8=137$$

$$\text{Var } Z=\text{Var } X_1+\text{Var } X_2+\text{Var } X_3+2\,\text{Cov}(X_1, X_2)+2\,\text{Cov}(X_2, X_3)+2\,\text{Cov}(X_3, X_1)$$

$$=17{\cdot}64+28{\cdot}09+37{\cdot}21+26{\cdot}712+35{\cdot}868+51{\cdot}728$$

$$=197{\cdot}248$$

$$\therefore \quad \sigma_z=14{\cdot}045 \quad \therefore \quad t=\frac{Z-E(Z)}{\sigma_2}\sim N(0,1)$$

Z	t	$p=\int_{-\infty}^{t} p(t)\,dt$	*Class*	*Area under the curve in class* (A)	*Frequency* $500\times(A)$
120	1·21050	0·11314	120—150	0·70937	354·685
150	0·92567	0·82251	150—180	0·17639	88·195
180	3·06180	0·99890	180—190	0·00102	0·510
190	3·77400	0·99992	190—240	0·88678	443·390
240	7·33410	1·00000	240—	0·00000	0·000

(*i*) The number of students in the first, second and third class are approx. 355, 88 and 0.

(*ii*) Total number of students with total marks between 120–190 is 443.

(*iii*) Probability that student gets more than 240 marks is zero.

(*iv*) The corr. coeff. between marks in subjects I and II of the students who secured equal marks in subjects III is $r_{12\cdot3}$.

Hence $$r_{12\cdot3}=\frac{r_{12}-r_{23}r_{13}}{\sqrt{(1-r_{23}^2)(1-r_{13}^2)}}=\text{etc.}$$

(*v*) By (11·11), we can find the limits for r_{23}.

24. In a certain examination, there are three papers I, II and III, each of 100 marks. The distribution of marks in each paper is normal with the following means, s.d. and intercorrelation.

	I	*II*	*III*
Mean	53	43	48
S.D.	11	8	9

$r_{12}=0{\cdot}8,\ r_{13}=0{\cdot}3,\ r_{23}=0{\cdot}4$

The first class, second class and pass class marks are, on the total 60%, 50% and 40%. Find the percentage of students who will pass in each class.

(*M. A. Delhi, 1962*)

12

Theory of Sampling

12·1. Sampling. Before we consider sampling theory, we first define the term population. The word '*population*' or '*universe*' in Statistics is used to refer to any collection of individuals or of their ıttributes or of results of operations which can be numerically speci-ied. Thus we may speak of the populations of weights, heights, ›rices of wheat, mileages of automobile tyres, etc.

A population containing a finite number of individuals or mem-›ers is called a *finite population*. Such, for instance, is the popula-ion of ages of twenty boys in a class. A population with infinite ıumber of members is known as *infinite population*. The population ›f pressures at various points in the atmosphere is an example of this ›ype of population. We call the population of concrete individuals an *existent population* while the collection of all possible ways in which an event can materialize as the *hypothetical population*. The population of heads and tails obtained by tossing a coin an infinite number of times is a hypothetical population. A part or small section selected from the population is called a *sample* and the proces: of such selection is called *sampling*. Sampling is resorted to wher either it is impossible to enumerate the whole population or wher it is too costly to enumerate in terms of time and money or wher the uncertainty inherent in sampling is more than compensated by the possibilities of errors in complete enumeration. To serve a usefu purpose sampling should be unbiased or representative.

The aim of the theory of sampling is to get as much informa tion as possible, ideally the whole of the information about th population from which the sample has been drawn. In particular given the form of the parent population we would like to *estimat*

the parameters of the population or specify the limits within which the population parameters are expected to lie with a specified degree of confidence. It is, however, to be clearly understood that the logic of the theory of sampling is the logic of induction, that is we pass from particular (*i.e.*, sample) to general (*i.e.*, population) and hence all results will have to be expressed in terms of probability.

The fundamental *assumption* underlying most of the theory of sampling is *random sampling* which consists in selecting the individuals from the population in such a way that each individual of the population has the same chance of being selected. The selection of random samples is very much facilitated by the use of *table of random numbers* due to Tippet, Kendall, Smith, Fisher and Yates or Rand Corporation (U.S.A.).

Tippet's Random Numbers. The random number table as given by L.H.C. Tippet consists of 10400 four-digit numbers so that there are 41600 digits. These numbers were taken from the census report in U.K. Any number taken either from row or column or diagonal of these random number table which is random in character may be taken to constitute the sample.

Types of Sampling. Some of the important types of sampling are :—

(*i*) Purposive sampling.
(*ii*) Random sampling.
(*iii*) Simple sampling.
(*iv*) Stratified sampling.

12·2. Sampling of Attributes. The sampling of attributes may be regarded as the drawing of samples from a population whose members possess the attribute A or not $-A$. For example in sampling from a population of men, we may have persons who are smokers and non-smokers. The choosing or drawing of an individual in sampling may be called an 'event' or 'trial' and the possession of the specified attribute A by the individual selected a 'success'.

By simple sampling we mean random sampling in which each event has the same probability p of success and the probability of an event is independent of the success or failure of events in the preceding trials. Thus simple sampling is a special case of random sampling in which trials are independent and probability of success is

constant. For example, counting the number of success in the throwing of a dice or tossing of a coin is a case of simple sampling since the probability of getting heads with a coin is unaffected by the previous trials and remains constant irrespective of the number of trials made provided the coin remains unbiased. Following is the illustration of sampling which is random but not simple.

Suppose in an urn there are 2 white balls and 3 black balls. Now the probability of drawing a white ball is $\frac{2}{5}$. Suppose at the first trial we draw a white ball. The probability of getting a white ball at the second trial, when the ball is not replaced, is $\frac{1}{4}$ which is clearly not the same as the probability in the first trial. Thus the sampling though random is not simple. However, if the ball drawn at the first trial was put back in the urn before the next trial, success or failure having been noted, the random sampling would become simple sampling. The two cases are sometimes distinguished as *sampling without replacement* and *sampling with replacement.*

It should be clear that random sampling from an infinite population would always be *simple* since drawing of one individual cannot effect the probability distribution but when sampling from a finite population the sampling may or may not be simple according as the members drawn are replaced or not.

12·2 1. Mean and Standard Deviation in Simple Sampling of Attributes. Suppose we take a simple sample of n members. This is clearly identical with that of a series of n independent trials, with constant probability p of success. The probabilities of 0, 1, 2,..., n successes are the terms in the binomial expansion of $(q+p)^n$, $q=1-p$. The distribution so obtained is called *sampling distribution* of the number of successes in the sample. The expected value, or the mean value, of the number of successes is, therefore, np and the standard deviation [also called the *standard error* (*s. e.*)] of the number of successes is $\sqrt{npq}$. By the *proportion of successes* in a sample, we mean the number of successes divided by the number of members in the sample. The mean and the standard error of the proportion of successes can be obtained by dividing the corresponding results for the number of successes by n. Thus

mean of the proportion of successes $=p$

and $e=s.e.$ of the proportion of successes $=\sqrt{pq/n}$

Hence *precision* (reciprocal of *s.e.*) of the proportion of successes $=\sqrt{n/pq}$ which is proportional to $\sqrt{n}$ since p and (therefore) q are constants. In other words, *the precision of the proportion of the successes varies as the square root of the number of members in the simple sample.*

12·3. Large Samples. Testing the Significance for a single Proportion. Suppose large number n of independent Bernoullian trials is performed and x successes are obtained. We wish to test the hypothesis that the probability of success in each trial is p. Assuming the hypothesis to be correct, the mean and variance of the sampling distribution of the number of successes are np and npq respectively, $q=1-p$.

It was shown in §8·1 that, for large n,

$$z=\frac{x-np}{\sqrt{npq}}$$

is distributed as a standard normal variate. Hence from table 2, we have

$$P(|z|>3)=0{\cdot}0027.$$

Therefore, in random sampling such a result is extremely unlikely. We conclude that the truth of the hypothesis is itself very improbable and we say that the difference between the observed number of successes x and the expected number of successes np is highly *significant.*

If $|z|<1{\cdot}96$, we say that the data is consistent with the hypothesis. If $|z|>1{\cdot}96$, we say that the difference or deviation is significant at 5% *level of significance* and if $|z|>2{\cdot}58$, the difference is said to be significant at 1% level of significance.

Cor. 1. Let the observed proportion of successes in a sample of n observation be p, then $E\left(\frac{x}{n}\right)=p$, $\text{Var}\left(\frac{x}{n}\right)=\frac{1}{n^2}\text{Var}\,x=\frac{pq}{n}$. Thus

$$z=\frac{\frac{x}{n}-p}{\sqrt{\frac{pq}{n}}}\sim N(0,\,1).$$

Cor. 2. The probability for the proportion in the population is

$$p\pm 3\sqrt{\frac{pq}{n}}.$$

Notes. 1. The above argument holds if instead of the number of successes we use the proportion of successes.

2. The above test is valid for *large samples* only since the binomial distribution may not be very near the normal for small samples. For small samples, binomial distribution is to be used.

3. It may be remembered that while the above test may furnish evidence against the hypothesis, it cannot prove the hypothesis to be correct. It can at the most prove no evidence against it.

4. The hypothesis tested above is a typical example of a *statistical hypothesis by which we understand some assumption about the density function of a random variate.* In the above case, the variate is the number of successes and it is not being tested that it follows the binomial distribution. In fact, this is assumed. The statistical hypothesis here is that the population parameter, the probability of success in a single trial, has the value p.

5. By the *testing of a statistical hypothesis is meant a procedure for deciding whether to accept or reject the hypothesis.* The procedure usually consists in assuming or accepting the hypothesis as correct and then calculating the probability of getting the observed or *more extreme* sample. If this probability is less than a certain pre-assigned small value, the hypothesis is rejected, since samples with small probabilities should be rare and we assume that a rare event has not happened.

6. The probability level below which we reject the hypothesis is called the *level of significance.* Usually, two levels are used, *viz.*, the 5% and 1%, but in any particular problem the level would depend on the risk, rejecting the hypothesis when it is true, would entail.

7. The tests are usually called *tests of significance* since here we test whether the differences between the sample values and the population values or between the values given by two samples are so large that they *signify* evidence against the hypothesis, or these differences are small enough to be accounted for as due to fluctuations of sampling, *i.e.*, they may be regarded as due only to the fact that we are dealing with a sample and not with the whole population.

8. According to Prof. Fisher, the hypothesis, a definite statement about the population parameter, tested for possible rejection

under the assumption that it is true is usually called the *null hypothesis* and is usually denoted by H_0. To test whether there is a difference between two populations we shall assume that there is no difference. Similarly, to test whether there is relationship between two variates, we assume there is no relationship. The reason is that a hypothesis can be rejected but cannot be proved. Rejection of no difference will mean a difference, while rejection of no relationship will imply a relationship.

9. A region corresponding to a statistic in the sample space which leads to the rejection of H, is called a *critical region or region of rejection.*

Example 12·1. *A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be an unbiased one.*

On the assumption of an unbiased coin, the expected number of heads in 400 tosses $=400\times\frac{1}{2}=200$.

Deviation of the actual number of heads from expected is, therefore, $216-200=16$.

Now $e=s.e.$ of the number of heads $=\sqrt{400\times\frac{1}{2}\times\frac{1}{2}}=10$.

$\therefore$ Deviation is only 1·6 times the $s.e.$ and hence it is likely to appear as a result of fluctuations of simple sampling. We conclude that the coin may be an unbiased one.

Example 12·2. *A die is thrown 9,000 times and a throw of 3 or 4 is reckoned as a success. Suppose that 3,240 throws of a 3 or 4 have been made out. Do the data indicate an unbiased die ?*

Suppose the die is unbiased. Then the chance of success, that is of throwing a 3 or 4 with one die in one throw $=\frac{2}{6}=\frac{1}{3}=p$.

$\therefore$ The expected number of successes $=9000\times\frac{1}{3}=3000$

$e=$standard deviation of simple sampling

$=\sqrt{npq}=\sqrt{9000\times\frac{1}{3}\times\frac{2}{3}}=44{\cdot}72$.

The deviation $3240-3000$, *i.e.*, 240 is nearly 5·4 e. Therefore, it is most improbable that it is due to fluctuation of sampling, that is, the die is almost certainly biased and that $p\neq\frac{1}{3}$.

The estimate of p obtained from the given sample is

$$\frac{3240}{9000}=0{\cdot}36.$$

Now $s.e.$ of the proportion of success $=\sqrt{\frac{0{\cdot}36\times0{\cdot}64}{9000}}=0{\cdot}005$.

Hence the true value of p almost certainly lies within the range $0{\cdot}36 \pm 3 \times {\cdot}005$, *i.e.*, it lies between 0·345 and 0·375.

Example 12·3. *In a locality of 18,000 families a sample of 840 families was selected. Of these 840 families, 206 families were found to have a monthly income of Rs 50 or less. It is desired to estimate how many out of the 18,000 families have a monthly income of Rs. 50 or less. Within what limits would you place your estimate ?*

p = proportion of families having monthly income of Rs. 50 or less $= \frac{206}{840} = {\cdot}245$ so that $q = {\cdot}755$.

Assume that the condition of this problem will give a simple sample.

S.E. of the proportion of families having monthly income of Rs. 50 or less $= \sqrt{\frac{pq}{n}} = \sqrt{\frac{{\cdot}245 \times {\cdot}755}{840}} = 0{\cdot}015 = 1{\cdot}5\%$.

Hence taking ·245 to be the estimate of families having a monthly income of Rs. 50 or less the limits are 24·5 per cent $\pm(3 \times 1{\cdot}5)$ per cent, *i.e.*, 20 per cent and 29 per cent, *i.e.*, between 3,600 and 5,220 families.

Ex 1. All the following statistical hypothesis

(*i*) There is life on planet Mars ;

(*ii*) The orbit of a newly discovered comet is a parabola :

(*iii*) A new instrument can take measurements with less than ½% error.

Ex. 2. A coin is thrown 10 times. Assuming it to be unbiased find

(*i*) the probability of 2 successes or less ;

(*ii*) the probability of the number of successes differing from 5 by more than 3.

At 5% level of significance, what number of successes will enable you to reject the hypothesis of an unbiased coin ?

12·4. The Testing of Significance for Difference of Proportions. Two large simple samples of n_1 and n_2 members respectively are taken from two populations. The samples give proportions of attribute A's as p_1 and p_2 respectively. We are to find :

(*a*) *Is the difference, $p_1 \sim p_2$ due to fluctuations of simple sampling significant, the two populations being similar in respect of the proportion of A's therein ?*

(*b*) *If the difference in population proportions is real, could it be hidden in sampling, i.e., could it vanish or change sign owing to the fluctuations of sampling in samples taken from the two populations ?*

(*a*) On the hypothesis that the populations are similar as regards the given attribute, we can combine the two samples to give an estimate of the common value of the proportion of A's in the populations. This estimate is given by

$$p=\frac{n_1p_1+n_2p_2}{n_1+n_2} \qquad \ldots(12\cdot1)$$

If e_1 and e_2 be the standard errors in the proportions of successes in independent samples, the estimate of the standard error, *i.e.*, of the proportion of successes in the parent universe is obtained from

$$e^2=e_1^2+e_2^2=\frac{pq}{n_1}+\frac{pq}{n_2}$$

or

$$e^2=pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right). \qquad \ldots(12\cdot2)$$

On the hypothesis that the two populations are identical,

$$E(p_1-p_2)=E(p_1)-E(p_2)=0.$$

If n_1 and n_2 be large then

$$z=\frac{p_1\sim p_2}{e} \qquad \ldots(12\cdot3)$$

can be treated as a standard normal variate. If the difference $p_1\sim p_2$ is less than $2e$, it may be due to fluctuations of sampling, but if it is greater than $3e$, there is likely to be real difference in population proportions but if the difference lies between $2e$ and $3e$, then it is significant at 5% level of significance.

Example 12·4. *A machine puts 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 ?*

Here $p_1=\frac{16}{500}=\cdot032$, $p_2=\frac{3}{100}=\cdot03$, so that $p_1-p_2=\cdot002$.

On the assumption that the machine has not been improved, we have on combining together as our estimate of proportions,

$$p=\frac{16+3}{500+100}=\frac{19}{600}$$

$$\therefore\ e=\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=\sqrt{\frac{19}{600}\times\frac{581}{600}\left(\frac{1}{500}+\frac{1}{100}\right)}$$
$$=0{\cdot}0192$$

Since the difference in proportions, *viz.*, ·002 is less than $2e$, the difference is not significant and we conclude that the machine has not been improved.

(*b*) Suppose that the two samples of large sizes n_1 and n_2 have been drawn from two populations in which the proportions are respectively p_1 and p_2 where $p_2 < p_1$. The question arises :

Is it likely that the proportions $\frac{x_1}{n_1}$ and $\frac{x_2}{n_2}$ in the samples will be such that $\frac{x_1}{n_2}-\frac{x_2}{n_2}\leqslant 0$, in other words : *Is the real difference between the populations likely to be hidden in simple sampling* ?

Since $$E\left(\frac{x_1}{n_1}\right)=p_1,\ E\left(\frac{x_2}{n_2}\right)=p_2$$

$$\text{Var}\left(\frac{x_1}{n_1}\right)=\frac{p_1q_1}{n_1},\ \text{Var}\left(\frac{x_2}{n_2}\right)=\frac{p_2q_2}{n_2}$$

$$\therefore\qquad z=\frac{\left(\frac{x_1}{n_1}-p_1\right)-\left(\frac{x_2}{n_2}-p_2\right)}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \qquad \ldots(12{\cdot}4)$$

can be treated for large n_1, n_2 as an asymptotically standard normal variate.

We require $$P\left[\frac{x_1}{n_1}-\frac{x_2}{n_2}\leqslant 0\right]$$

or $$P[ze+p_1-p_2\leqslant 0]$$

where $$e=\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$$

i.e., we require $$P\left[z\leqslant-\left(\frac{p_1-p_2}{e}\right)\right].$$

If $p_1-p_2>1{\cdot}645e$, then from table 2, this probability is less than ·05, in other words the probability of the difference being hidden is less than 5% and we say that it is unlikely that the difference will be hidden in sampling.

Example 12·5. *In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be*

hidden in samples of 1,200 and 900 respectively from the two populations ? *(B.Sc. Agra, 1958 ; B.A. Hons., Delhi, 1954)*

Here $p_1=0{\cdot}30$, $p_2=0{\cdot}25$ so that $p_1-p_2=0{\cdot}05$

$$e^2=\frac{(\cdot3)(\cdot7)}{1200}+\frac{(\cdot25)(\cdot75)}{900}=\frac{138}{3,60,000}, \text{ so that } e=0{\cdot}0195$$

The difference p_1-p_2 is about $2\frac{1}{2}$ times the standard error.

Hence it is unlikely that the real difference will be hidden.

Ex. 1. If $p_1-p_2>2{\cdot}327e$, prove that the difference is unlikely to be hidden at 1% level of significance.

Ex. 2. What is level of significance of $p_1-p_2>1{\cdot}96e$?

12·5. Sampling of values of a Variable. We now consider the sampling of values of a variable such as age, height, weight, etc. Each individual or member of the population provides a value of the variable and the population is a frequency distribution of the variable. From the population, we can draw a random sample of n members and this is the same as choosing n values of the variable from those of the distribution.

Parameters and Statistics. The statistical constants of the population, *viz.*, the mean, the variance, etc. are known as the parameters. The statistical concepts of the sample computed from the members or observations of the sample to determine or estimate the parameters of the population from which the sample has been drawn, are known as statistics.

12·5·1. Sampling Distribution. Any function of the observed sample values calculated to estimate some parameters of the population is called a *statistic*. We can draw N_{C_n} different samples of size n from a finite population of size N and an infinite member of samples of size n from an infinite population. Each sample gives its own values for the statistic under consideration. The assemblage or aggregate of all these values of the statistics, together with their relative frequencies or the probabilities with which they occur, constitutes the sampling distribution of the corresponding statistic. Thus we can talk of the sampling distribution of the sample mean $\bar{x}$, of the sampling variance s^2, etc.

Let the density function for the sampling distribution of the statistic x be $f(x)$, then the probability that any sample chosen at

random from all possible samples will give a value of X between x_1 and x_2 is given by

$$P(x_1<X<x_2)\Rightarrow\int_{x_1}^{x_2} f(x)\,dx \div \int_{-\infty}^{\infty} f(x)\,dx$$

$$=\int_{x_1}^{x_2} f(x)\,dx \qquad \because \int_{-\infty}^{\infty} f(x)\,dx=1.$$

In particular, the probability that the sample will give a value of x greater than x_0 is given by

$$P(x>x_0)=\int_{x_0}^{\infty} f(x)\,dx \div \int_{-\infty}^{\infty} f(x)\,dx=\int_{x_0}^{\infty} f(x)dx$$

Ex. From a rectangular population with 6 members 1, 2, 3, 4, 5, 6 samples of sizes 3 are drawn. Find the sampling distribution of the mean and variance and deduce the probability that

$$(i)\ \bar{x}>4 \qquad (ii)\ \left|\bar{x}-\frac{7}{2}\right|>1 \qquad (iii)\ s^2>1 \qquad (iv)\ 1<s^2<2$$

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

12·6. Sample Distribution of the Mean. Central Limit Theorem. We shall find the distribution of the mean of random samples of given size drawn from a specified population.

Case I. *Parent population is normal.*

Let the population from which the random sample is drawn be a normal population with mean μ and variance σ^2. By the definition of a random sample, variate values $x_1, x_2, \ldots, x_n$ of the sample are independent and from §8·2 it follows the *sample mean is distributed normally about mean μ and variance σ^2/n.* The result shows how the *precision* of a sample mean increases as the sample size increases.

Case II. *Parent population is non-normal.*

Theorem. *If the variable X has a non-normal distribution with mean μ and finite standard deviation σ and if the moment generating function exists, then the variate*

$$Z=\frac{(\bar{x}-\mu)\sqrt{n}}{\sigma}$$

has a distribution that approaches the standard normal distribution as n tends to infinity.

Proof. $M_Z(t)=M_{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}(t)$

$$M_z(t)=e^{-\frac{\mu\sqrt{n}}{\sigma}t}\, M_{\bar{x}}\left(\frac{\sqrt{n}}{\sigma}t\right)$$

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

$$\therefore \log_e M_z(t)= -\frac{\mu\sqrt{n}}{\sigma}t+n \log M_e\left(\frac{t}{\sigma\sqrt{n}}\right)$$

$$= -\frac{\mu\sqrt{n}}{\sigma}t+n \log\left[1+\mu_1'\frac{t}{\sigma\sqrt{n}}+\mu_2'\frac{t^2}{2\sigma^2 n}+\cdots\right]$$

$$= -\frac{\mu\sqrt{n}}{\sigma}t+n\left[\left(\mu_1'\frac{t}{\sigma\sqrt{n}}+\mu_2'\frac{t^2}{2\sigma^2 n}+\cdots\cdots\right)\right.$$

$$\left.-\tfrac{1}{2}\left(\frac{\mu_1' t}{\sigma\sqrt{n}}+\cdots\cdots\right)^2+\cdots\cdots\right]$$

$$=\frac{t^2}{2}+O\left(\frac{1}{\sqrt{n}}\right) \qquad [\because\ \mu_1'=\mu]$$

$\therefore$ As n tends to infinity, we have

$$\log_e M_z(t)\to\frac{t^2}{2} \text{ or } \quad M_z\ (t)\to\exp\left(\frac{t^2}{2}\right).$$

Now $\exp(\frac{1}{2}t^2)$ is the *m.g.f.* of a standard normal variate.

Hence the theorem is established.

Note. This theorem is known as the *Central Limit Theorem* and can be proved to hold for all populations with *finite* variance. The existence of a moment generating function is not necessary. The condition that the variance is finite is, however, not very restrictive since most of the distributions have a finite range and, therefore, a finite variance or an infinite range but a finite variance. The degree of approximation to normal distribution depends on (*i*) the degree of departure of the parent population from the normal and (*ii*) the sample size n. *For a parent normal population, the distribution of the sample mean is always exactly normal i.e., it is normal for all n.*

Ex. 1. Verify the theorem for the following distributions:—

(*i*) $f(x)=p^x(1-p)^{1-x}$ $[x=0, 1]$

(*ii*) $f(x)=\dfrac{e^{-m}m^x}{x!}$ $[x=0, 1, 2,\ldots\ldots]$

(*iii*) $f(x)=1$ $[0\leqslant x\leqslant 1]$

(*iv*) $f(x)=\dfrac{1}{\lfloor l}e^{-x}x^{l-1}, l>0,$ $[x\geqslant 0]$

(*v*) $f(x)=e^{-x}$ $[x\geqslant 0]$

Ex. 2. Prove that the moment generating function for a Cauchy distribution is the same as that of the parent population. Explain the failure of the central limit theorem in this case.

Ex. 3. Show that the rth cumulant for the sample mean is k_r/n^{r-1} where k_r is the rth cumulant for the parent population.

Deduce that for the sampling distribution of the sample mean,

(*i*) mean is the population mean,

(*ii*) variance is σ^2/n where σ^2 is the variance of the population,

(*iii*) third moment about mean is μ_3/n^2,

(*iv*) skewness is $1/\sqrt{n}$ of the skewness of the population,

(*v*) excess of kurtosis is $1/n$ of the excess of kurtosis for the population.

Deduce that for large samples from a population of moderate skewness and excess of kurtosis, the distribution of sample mean is approximately normal.

Ex. 4. State and establish the central limit theorem for the case of identically distributed random variables. Under what conditions does it hold for non-identical random variables ? *(M.A. Delhi, 1967)*

Example 12·6. *A sample of 400 male students is found to have a mean height of 67·47 inches. Can it be reasonably regarded as a sample from a large population with mean height 67·39 inches and s.d. 1·30 inches ?*

$\sigma_{\bar{x}}$, the standard error (*i.e.*) of the mean of the simple sample of 400 from such a population is $1{\cdot}30/\sqrt{400}=0{\cdot}065$ inch. The deviation of the mean of the sample from that of the population is ·08 inch and this is less than twice the *s.e.* of the mean and is, therefore, not significant. We conclude that the given sample might have been drawn from the population specified.

12·7. Confidence or Fiducial Limits for Unknown Mean. Let the population from which a random sample of size n is drawn, have mean μ and *s.d.* σ. If σ be known but not μ there is a range of possible value of μ for which the observed mean $\bar{x}$ of the sample is not significant at any specified level of probability. The relative deviation of $\bar{x}$ from μ is the ratio of $\bar{x}\sim\mu$ to the standard errors of $\bar{x}$ *i.e.*, $(\bar{x}\sim\mu)\sqrt{n}/\sigma$. If $\bar{x}$ is not significant at 5% level of probability, this relative deviation must be less than that which (in a normal distribution) is exceeded numerically with a probability of 5%. Such a relative deviation is 1·96.

Hence the observed mean will not be significant at 5% level of probability if

$$\left|\left(\frac{\bar{x}\sim\mu}{\sigma}\right)\sqrt{n}\right| < 1{\cdot}96$$

i.e., if $$\bar{x}-1{\cdot}96\,\sigma/\sqrt{n}<\mu<\bar{x}+1{\cdot}96\sigma/n \qquad \ldots(12{\cdot}5)$$

The values $\bar{x}\mp 1{\cdot}96\sigma/\sqrt{n}$ are called 95% *Fiducial Limits* or *Confidence Limits*, for the mean of the population corresponding to the given sample. They give the limits for μ in order that the observed sample mean $\bar{x}$ should not be significant at the prescribed level of probability.

Note. For normal population with known variance, the above test of significance is valid for all sample size.

Example 12·7. *Suppose that distribution of the status of men is a normal distribution with s.d. 2·48 inches. One hundred male students in a large university are measured and their average height is found to be 68·52 inches. Determine the 98 per cent confidence limits for the mean height of the men of the university.*

We have $\sigma_{\bar{x}}=\dfrac{2{\cdot}48}{\sqrt{100}}=0{\cdot}248$ inch.

The 98 per cent fiducial limits for the mean of the population are $\bar{x}\mp 2{\cdot}33\ \sigma/\sqrt{n}$, *i.e.*, 68·52 inches $\mp$·578 inch *i.e.*, 67·942 inches and 69·098 inches. Thus at 2% level of confidence the confidence limits are 67·942 and 69·098 inches.

Ex. 1. If p is the observed proportion of success in n independent Bernoullian trials, prove that the 95% fiducial limits for the population proportion p' are, for the large samples, given by

$$p\pm 1{\cdot}96\sqrt{\frac{p-p^2}{n}}$$

Ex. 2. Prove in the above exercise that the 99% fiducial limits are the roots of the quadratic

$$(p-p^2)(2{\cdot}58)^2=n(p'-p)^2 \qquad (M.A.\ Patna,\ 1956)$$

Ex. 3. An unbiased coin is thrown n times. It is desired that the relative frequency of the appearance of heads should lie between ·49 and ·51. Find the smallest value of n that will ensure this result with 90% confidence.

[Ans. 6745]

Ex. 4. p_1, p_2 are the observed proportions of successes in two independent sets of trials of large sizes n_1, n_2. Prove that the 95% fiducial limits for the difference $p_1'-p_2'$ of the proportions of successes in the population are

$$p_1 - p_2 \pm 1{\cdot}96 \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}.$$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

12·8. Testing the Significance of the Difference between the Means of two Large Samples. Suppose two random samples of n_1 and n_2 members respectively have been drawn from the *same* population of *s.d.* σ. We would like to know if the difference $\bar{x}_1 \sim \bar{x}_2$ of their means is due to fluctuations of sampling. On the assumption that the samples are independent and drawn from the same population, the standard error, e, of the difference of their means is given by

$$e^2 = e_1^2 + e_2^2$$

where e_1 and e_2 are the *s.e.* of the means of the two samples and are $\sigma/\sqrt{n_1}$ and $\sigma/\sqrt{n_2}$ respectively so that

$$e = \sigma \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}} \qquad \ldots(12{\cdot}6)$$

If n_1 and n_2 be sufficiently large then $x_1 \sim \bar{x}_2$ is asymptotically normally distributed with mean zero and *s.d.e.* Consequently if the difference $\bar{x}_1 \sim \bar{x}_2$ exceeds $3e$ the difference can hardly be accounted for by the fluctuations of sampling and our assumption is unlikely to be correct while if the difference exceeds $2e$, it is regarded as significant at the 5% level of probability.

If two independent samples of n_1 and n_2 members respectively be drawn from different populations with variances σ_1^2 and σ_1^2 respectively we can examine whether the two populations from which samples have been drawn differ in mean apart from the difference in dispersion. Since the samples are independent the *s.e.*, e, of the difference of their means is given by

$$e^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \qquad \ldots(12{\cdot}7)$$

Assuming that n_1 and n_2 are large and the two populations have the same mean the difference of the means of the samples will be normally distributed with mean zero and *s.d.e.* given by (12·7). If the difference of the means of the samples exceeds $3e$, it can hardly be accounted for on the basis of fluctuations of sampling and our assumption that the two populations have the same mean is almost certainly wrong.

Note. 1. In the above discussion, we have assumed that the population variance σ_1^2, σ_2^2 are known. In practice this is hardly the case and accordingly in the expressions for e, these have to be replaced by their estimates obtained for the samples, *viz.*, by the sample variances s_1^2, s_2^2 respectively

where $$s_1^2=\frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-\bar{x}_1)^2.$$

2. The above tests are valid for only large samples for two reasons :

(*i*) the parent populations may not be normal, though we are assuming that they do not depart strikingly from it. In particular we assume that the populations have finite variances. For populations like Cauchy's where the variance is not finite, the tests would break down completely even for 'infinitely large samples'.

(*ii*) The population variances are not known and have to be replaced by their estimates.

3. For *normal* populations with *known* variances, the above tests are *valid for all sample sizes.*

4. If the hypothesis to be tested is that the population means are μ and μ', we can carry out the test of significance as above, but in this case

$$z=\frac{(\bar{x}_1-\bar{x}_2)\sim(\mu-\mu')}{\sqrt{\frac{s_1^2}{n_2}+\frac{s_2^2}{n_2}}} \qquad \ldots(12\cdot8)$$

will be asymptotically a standard normal variate for large n_1, n_2.

Example 12·8. *If 60 new entrants in a given university are found to have a mean height of 68·60 inches and 50 seniors a mean height of 69·51 inches ; is the evidence conclusive that the mean height of the seniors is greater than that of the new entrants. Assume the standard deviation of heights to be 2·48 inches.*

Since the two samples are independent and come from the same population, the *s.e.*, e, of the difference of the two mean heights is

$$e=\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}=2\cdot48\sqrt{\frac{1}{60}+\frac{1}{50}}=0\cdot47 \text{ inch}$$

The observed difference of the two mean heights is 0·91 inch which is less than $2e$ and so could be due to fluctuations of sampling.

Hence it cannot be said that the mean height of the seniors is necessarily greater than that of the new entrants.

Example 12·9. *A random sample of 100 farms in a certain year gives an average yield of barley of 2,000 lbs. per acre with a s.d. of 192 lbs. A random sample of 100 farms in the following year gives an average yield of 2,100 lbs. per acre with a s.d. of 224 lbs. Show that the data are inconsistent with the hypothesis that the average yield in the country as a whole were same in the two years.*

The *s.e.* of the average yields of the two samples are $\frac{192}{\sqrt{1000}}$ and $\frac{224}{\sqrt{1000}}$ respectively.

Since the two samples are independent, the *s.e.* of the difference of their mean yields is given by

$$e=\sqrt{\frac{(192)^2}{1000}+\frac{(224)^2}{1000}}=\sqrt{87{\cdot}04}=9{\cdot}33 \text{ per acre.}$$

The observed difference between the means of the samples is 100 lbs. per acre which is about 10 times its *s.e.* and so cannot be due to fluctuations of sampling. Hence the data are inconsistent with the assumption that the average yield in the country as a whole were same in the two years.

Ex. 1. 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, prove that the 98% fiducial limits for the difference between population means are

$$\bar{x}_1-\bar{x}_2\pm 2{\cdot}326\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

Ex. 2. How will the above limits be effected if

(*i*) the populations are not known to be normal,

(*ii*) the variances are not known,

(*iii*) confidence coefficient is 95% ?

12·8·2. Testing the Significance of Difference between Standard Deviations of two Large Samples. Let two random samples of large sizes n_1 and n_2 having standard deviations s_1 and s_2 be drawn from two normal populations with standard deviations σ_1 and σ_2 respectively. We would like to know if the difference between s_1 and s_2, *i.e.*, s_1-s_2 is due to fluctuations of sampling.

We set up the the null hypothesis that the samples have been drawn from normal populations having the same standard deviation *i.e.*, $\sigma_1-\sigma_2=\sigma$. Under the null hypothesis, the statistic

$$z=\frac{s_1\sim s_2}{\sqrt{\text{Var}\,(s_1-s_2)}}\sim N(0,\,1)$$

since the samples are of large sizes

and $$E(s_1-s_2)=E(s_1)-E(s_2)=\sigma-\sigma=0$$

Now $$\text{Var}\,(s_1-s_2)=\text{Var}\,s_1+\text{Var}\,s_2$$

$$=\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2},$$

since samples are independent

$$\therefore \qquad z=\frac{s_1\sim s_2}{\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}}\sim N(0,\,1)$$

If the value of $z > 2$ but < 3, then the deviation is significant at 5% level.

Example 12·10. *Random samples drawn from two countries gave the following data relating to the heights of adult males.*

	Country A	*Country B*
Mean height in inches	*67·42*	*67·25*
Standard deviation	*2·58*	*2·50*
Number in samples	*1000*	*1200*

Is the difference between the standard deviations significant ?

(B.Sc. Hons., Sub., Delhi, 1968)

We set up the null hypothesis that there is no significant difference between sample standard deviations. Under the null hypothesis,

$$z=\frac{s_1-s_2}{\sqrt{\frac{s_1^2}{2n_1}+\frac{s_2^2}{2n_2}}}\sim N(0,\,1)$$

Now $$z=\frac{2\cdot58-2\cdot50}{\sqrt{\frac{(2\cdot58)^2}{2\times1000}+\frac{(2\cdot50)^2}{2\times1200}}}=1\cdot03,\text{ on simplification}$$

Since $z<1\cdot96$, we conclude that data do not provide any evidence against the null hypothesis at 5% level.

Hence the sample standard deviations do not differ significantly.

STANDARD ERRORS OF STATISTICS

12·9·1. Introduction. We have seen earlier that for large samples, the distribution of most of the statistics tends to normality.

Thus if to estimate a population parameter we construct from the sample values the values of the parameter in the sub-population, composed by the sample, its distribution will, in general, tend to normality about the true population parameter as mean and with a sampling variance which will in general be of order $1/n$ for large samples. The positive square root of the sampling variance is called the *standard error* of the estimate and for large samples, it is a valid measure of precision for such statistics of which the distribution tends to normality. Now we shall obtain the standard errors of important statistics.

12·9·2. Standard Error of Sample Moments about the arbitrary values. It is the usual convention to denote the parameters of the population by Greek letters and the corresponding statistics of the sample by Roman letters. Thus, if μ'_r and μ_r denote respectively the rth moment of the population about the arbitrary value and about the mean, then m'_r and m_r denote the corresponding moments of the sample. If n be the size of the sample, we have

$$m_r' = \frac{1}{n} \sum_{i=1}^{n} x_i^r, \quad m_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_1')^r \qquad \ldots(12\cdot9)$$

$$\therefore \quad E(m_r') = E\left(\frac{1}{n} \Sigma x_i^r \right) = \frac{1}{n} \Sigma E(x_i^r) = E(x^r) = \mu_r'$$

$$\text{Var } (m_r') = E(m_r' - \mu_r')^2 = E\left(\frac{1}{n} \Sigma x_i^r - \mu_r \right)^2$$

$$= \frac{1}{n^2} E[(\Sigma x_i^r)^2 - 2n\mu_r'(\Sigma x_i^r) + n^2 \mu_r'^2]$$

$$= \frac{1}{n^2} E[(\Sigma x_i^r)^2 - \mu_r^2] = \frac{1}{n^2} E[\Sigma x_i^{2r} + \sum_{i \neq j} x_i^r x_j^r] - \mu_r'^2$$

$$= \frac{1}{n} [nE(x_i^{2r}) + n(n-1)\ E(x_i^r)\ E(x_j^r)] - \mu_r'^2,$$

since x_i's are independent

$$= \frac{1}{n^2} [n\mu_{2r}' + n(n-1)\mu_r'^2] - \mu_r'^2$$

$$\therefore \text{ Var } (m_r') = \frac{1}{n} (\mu_{2r}' - \mu_r'^2). \qquad \ldots(12\cdot10)$$

$$\text{Cov}(m_r', m_s') = E[(m_r' - \mu_r')(m_s' \quad \mu_s')]$$

$$= E\left[\left(\frac{1}{n}\Sigma x_i^r - \mu_r'\right)\left(\frac{1}{n}\Sigma x_i^s - \mu_s'\right)\right]$$

$$= \frac{1}{n}E(\Sigma x_i^{r+s}) + \frac{1}{n^2}E\left(\sum_{i\neq j} x_i^r x_j^s\right) - \frac{1}{n}\mu_r' E(\Sigma x_i^s)$$

$$- \frac{1}{n}\ \mu_s' E(\Sigma x_i^r) + E(\Sigma \mu_r' \mu_s')$$

$$= \frac{1}{n^2}\cdot \mu'_{r+s} + \frac{1}{n^2}\, n(n-1)\mu_r'\mu_s' - \frac{1}{n^2}\,\mu_r' n\mu_s'$$

$$- \frac{1}{n}\cdot \mu_s' n\mu_r' + \mu_r'\ \mu_s'$$

$$= \frac{1}{n}\ (\mu_{r+s}' - \mu_r'\mu_s'). \qquad \ldots(12{\cdot}11)$$

$$E(m_2) = E\left[\frac{1}{n}\Sigma(x_i - m_1')^2\right] = \frac{1}{n}E[\Sigma x_i^2 - 2m_1'\Sigma x_i$$

$$+ nm_1'^2]$$

$$= \frac{1}{n}\ E[\Sigma x_i^2 - \frac{2}{n}(\Sigma x_j)(\Sigma x_i) + \frac{1}{n}(\Sigma x_i)^2]$$

$$= \frac{1}{n}\ E\left[\Sigma x_i^2 - \frac{1}{n}\ \Sigma x_i^2 - \frac{2}{n}\sum_{i\neq j} x_i x_j\right]$$

$$= \mu'_2 - \frac{1}{n}\ \mu_2' - \frac{n(n-1)}{n^2}\mu_1'^2 = \frac{n-1}{n}\ (\mu'_2 - \mu_1'^2)$$

$$= \left(1 - \frac{1}{n}\right)\mu_2. \qquad \ldots(12{\cdot}12)$$

$$\text{Var}\,(m_2) = E\left(m_2 - \frac{n-1}{n}\mu_2\right)^2$$

$$= E\left[m_2^2 - 2\frac{(n-1)}{n}m_2\mu_2 + \left(\frac{n-1}{n}\right)^2\mu_2^2\right]$$

$$= E(m_2^2) - \left(\frac{n-1}{n}\right)^2\mu_2^2 \qquad \text{using } (12{\cdot}12)$$

Now $$m_2^2 = \left[\frac{1}{n}\ \Sigma x_i^2 - \frac{1}{n^2}(\Sigma x_i)^2\right]^2$$

$$= \frac{1}{n^2}\ (\Sigma x_i^2)^2 - \frac{2}{n^3}\ (\Sigma x_i^2)^2(\Sigma x_i)^2 + \frac{1}{n^4}(\Sigma x_i)^4$$

$$= \frac{1}{n^2}\left(\Sigma x_i^4 + 2\sum_{i\neq j} x_i^2 x_j^2\right) - \frac{2}{n^3}\left(\Sigma x_i^2\right)$$

$$\times\left(\Sigma x_i^2 + 2\sum_{i\neq j} x_i x_j\right) + \frac{1}{n^4}\left(\Sigma x_i^2 + 2\sum_{i\neq j} x_i x_j\right)\left(\Sigma x_i^2 + 2\sum_{i\neq j} x_i x_j\right)$$

$$=\frac{1}{n^2}\left(\Sigma x_i^4+2\Sigma x_i^2x_j^2\right)-\frac{2}{n^2}\left(\Sigma x_i^4+2\sum_{i\neq j}x_i^2x_j^2\right.$$

$$\left.+2\sum_{i\neq j}x_i^2x_j+2\sum_{i\neq j\neq k}x_i^2x_jx_k\right)+\frac{1}{n^4}\left(\Sigma x_i^4+6\sum_{i\neq j}x_i^2x_j^2\right)$$

$$\therefore\quad E(m_2^2)=\frac{1}{n}\mu_4+\frac{n-1}{n}\mu_2^2-\frac{2}{n^2}\mu_4-\frac{2(n-1)}{n^2}.\mu_2^2+\frac{1}{n^3}\mu_4$$

$$+\frac{3(n-1)}{n^3}\mu_2^2,\text{ taking }\mu_1'=0$$

$$=\left(\frac{1}{n}-\frac{1}{n^2}+\frac{1}{n^3}\right)\mu_4+\frac{n-1}{n}\left(1-\frac{2}{n}+\frac{3}{n^2}\right)\mu_2^2$$

$$=\frac{(n-1)^2}{n^3}\mu_4+\frac{(n-1)(n+1)(n-3)}{n^3}\mu_2^2$$

$$\therefore\quad \text{Var }(m_2)=\frac{(n-1)^2}{n^3}\mu_4+\frac{(n-1)(n+1)(n-3)}{n^3}\mu_2^2-\frac{(n-1)^2}{n^2}\mu_2^2$$

$$=\frac{(n-1)^2}{n^3}\mu_4-\frac{n-1}{n^3}\left(n-1-2\right)\mu_2^2$$

$$=\frac{(n-1)^2}{n^3}(\mu_4-\mu_2^2)+\frac{2(n-1)}{n^3}\mu_2^2$$

Hence $\text{Var }(m_2)=\dfrac{\mu_4-\mu_2^2}{n}$ to order $1/n$, and that for normal parent

$$\text{Var }(m_2)=\frac{3\sigma^4-\sigma^4}{n}=\frac{2\sigma^4}{n}$$

Cor. $$\text{Var }(s^2)=\text{Var}\left(\frac{n}{n-1}m_2\right)=\frac{n^2}{(n-1)^2}\text{Var }(m_2)$$

$$=\frac{1}{n}(\mu_4-\mu_2^2)+\frac{2}{n(n-1)}\mu_2^2$$

$$=\frac{1}{n}(\mu_4-3\mu_2^2)+\frac{2}{n-1}\mu_2^2$$

$$=\frac{2\sigma^4}{(n-1)}\text{ for normal parent.}$$

12·10. Theorem. *If y_i's, $i=1, 2, \ldots, k$ are k statistics calculated from a sample of size n and if each y_i, is distributed about mean ξ_i with variance of order $1/n$, then*

$$Var\ (f)\equiv\sum_{i=1}^{k}\left(\frac{\partial f}{\partial\xi_i}\right)^2 Var(y_i)+\sum_{i\neq j}\left(\frac{\partial f}{\partial\xi_i}\right)\left(\frac{\partial f}{\partial\xi_j}\right)Cov\ (y_i.\ y_j)$$

where $f\equiv f(y_1\ y_2, \ldots\ldots, y_k)$ is any function of y's and possesses partial derivatives.

Proof. By Taylor's theorem, we have

$$f(y_1, y_2 \cdots\cdots, y_k) = f(\xi_1, \xi_2, \cdots\cdots, \xi_k) + \Sigma\left[(y_i - \xi_i)\frac{\partial f}{\partial \xi_i}\right] + \frac{1}{2!}\left[\Sigma(y_i - \xi_i)^2 \frac{\partial^2 f}{\partial \xi_i^2} + \sum_{i \neq j}(y_i - \xi_i)(y_j - \xi_j)\frac{\partial^2 f}{\partial \xi_i \partial \xi_j}\right] + \cdots\cdots$$

since Var (y_i) is of order $1/n$, we have $(y_i - \xi_i) \propto 1/\sqrt{n}$. Hence neglecting terms of order higher than $1/n$, we have

$$f(y_1, y_2, \cdots, y_k) = f(\xi_1, \xi_2, \ldots, \xi_k) + \Sigma\left[(y_i - \xi_i)\frac{\partial f}{\partial \xi_i}\right]$$

$$\therefore \quad \text{Var}\,(f) = E[f(y_1, y_2, \ldots, y_k) - f(\xi_1, \xi_2, \ldots, \xi_k)]^2$$

$$= E\left[\Sigma(y_i - \xi_i)\frac{\partial f}{\partial \xi_i}\right]^2$$

$$= E\left[\Sigma\left\{\left(y_i - \xi_i\right)^2\left(\frac{\partial f}{\partial \xi_i}\right)^2\right\} + \sum_{i \neq j}\left\{(y_i - \xi_i)(y_j - \xi_j)\frac{\partial f}{\partial \xi_i}\frac{\partial f}{\partial \xi_j}\right\}\right]$$

$$= E\left[\left(\frac{\partial f}{\partial \xi_i}\right)^2 \text{Var}\,(y_i)\right] + \sum_{i \neq j}\left[\frac{\partial f}{\partial \xi_i}\frac{\partial f}{\partial \xi_j}\,\text{Cov}\,(y_i, y_j)\right]$$

Note. If $f_1 = f_1(y_1, y_2, \ldots, y_k)$ and $f_2 = f_2(y_1, y_2, \ldots, y_k)$ be two functions of y's, then

$$f_1(y_1, y_2, \ldots, y_k) = f_1(\xi_1, \xi_2, \ldots, \xi_k) + \Sigma\left[(y_i - \xi_i)\frac{\partial f_1}{\partial \xi_i}\right]$$

$$f_2(y_1, y_2, \ldots, y_k) = f_2(\xi_1, \xi_2, \ldots, \xi_k) + \Sigma\left[(y_i - \xi_i)\frac{\partial f_2}{\partial \xi_i}\right]$$

$$\therefore \text{Cov}(f_1, f_2) = E[f_1(y_1, y_2, \cdots y_k) - f_1(\xi_1, \xi_2, \ldots \xi_k)][f_2(y_1 y_2, \ldots, y_k) - f_2(\xi_1, \xi_2, \ldots, \xi_k)]$$

$$= E\left[\Sigma\left\{\left(y_i - \xi_i\right)\frac{\partial f_1}{\partial \xi_i}\right\}\right]\left[\Sigma\left\{\left(y_i - \xi_i\right)\frac{\partial f_2}{\partial \xi_i}\right\}\right]$$

$$= \sum_{i \neq j}^{k}\left[\frac{\partial f_1}{\partial \xi_i}\frac{\partial f_2}{\partial \xi_i}\,\text{Var}\,(y_i)\right] + \sum_{i \neq j}\left[\frac{\partial f_1}{\partial \xi_i}\frac{\partial f_2}{\partial \xi_i}\,\text{Cov}\,(y_i, y_j)\right]$$

Cor. *Standard Error of Sample Moments about Mean.*

$$m_r = \frac{1}{n}\Sigma(x_i - m_1')^r$$

$$= m_r' - {}^rC_1 m_{r-1}' m_1' + \ldots + (-1)^i\, {}^rC_i m_{r-i}' m'^i - \ldots + (-1)^r m'^r$$

$$= f(m_r', m_{r-1}', \ldots, m_1')$$

$$=f(\mu_r, \mu_{r-1}, \ldots, \mu_1)+\sum_{i=1}^{r}\left[(m_i'-\mu_i)\frac{\partial f}{\partial \mu_i}\right]$$

since m_r is distributed about the mean μ_r with variance $\propto 1/n$.

$$=f\mu_r, \mu_{r-1}, \ldots, \mu_1)+(m_1'-\mu_r)-(m_1'-\mu_r)^r c_1\mu_{r-1}$$

$\therefore$ Var $(m_r)=E[m_r-f\{\mu_r, \mu_{r-1}, \ldots, \mu_1)]^2$

$$=E[(m_r'-\mu_r)-r(m_1'-\mu_1)\mu_{r-1}]^2$$

$$=(m_r'-\mu_r)^2+r^2\mu^2{}_{r-1}(m_1'-\mu_1)^2-2r\mu_{r-1}(m_1'-\mu_1)\times(m_r'-\mu_r')]$$

$$=\text{Var}\,(m_r')+r^2\mu^2{}_{r-1}\,\text{Var}\,(m_1')-2r\mu_{r-1}\,\text{Cov}\,(m_1', m_r')$$

$$=\frac{1}{n}(\mu_{2r}-\mu_r^2)+r^2\mu^2{}_{r-1}(\mu_2-\mu_1{}^2)-2r\mu_{r-1}(\mu_{r+1}+\mu_r\mu_1)]$$

on using 12·8 and $\mu_1'=0$ ...(12·13)

or Var $(m_r)=\frac{1}{n}(\mu_{2r}-\mu_r{}^2+r^2\mu_2\mu_{r-1}{}^2-2r\mu_{r-1}\mu_{r+1})$

Cov $(m_r, m_s)=E[\{m_r-f(\mu_r, \ldots, \mu_1)\}\{m_s-f(\mu_s, \ldots, \mu_1)\}]$

$$=E[\{(m_r'-\mu_r)-r(m_1'-\mu_1)\mu_{r-1}\}\{(m_s-\mu_s)-s(m_1'-\mu_1)\mu_{s-1}\}]$$

$$=E[(m_r'-\mu_r)(m_s'-\mu_s)-r\mu_{r+1}(m_1'-\mu_1)(m_s'-\mu_s)-s\mu_{s-1}(m_1'-\mu_1)(m_r'-\mu_r)+rs\mu_{r-1}\mu_{s-1}(m_1'-\mu_1)^2]$$

$$=\text{Cov}\,(m_r', m_s')-r\mu_{r-1}\,\text{Cov}\,(m_s', m_1')-s\mu_{s-1}\,\text{Cov}\,(m_r', m_1')+rs\mu_{r-1}\,\mu_{s-1}\,\text{Var}\,(m_1')$$

$$=\frac{1}{n}[\mu_{r+s}-\mu_r\mu_s-r\mu_{r-1}(\mu_{s+1}+\mu_s\mu_1)-s\mu_{s-1}(\mu_{r+1}-\mu_r\mu_1)+rs\mu_{r-1}\mu_{r-1}(\mu_2-\mu_1')^2]$$

$$=\frac{1}{n}(\mu_{r+s}-\mu_r\mu_s-r\mu_{r-1}\mu_{s-1}-s\mu_{s-1}\mu_{r+1}+rs\mu_{r-1}\mu_{s-1}\mu_2)$$

...(12·14)

Covariance between the Sample Mean and any Sample Mean moment

$$\text{Cov}\,(m_s, m_r')=E[\{m_r-f(\mu_r, \ldots, \mu_1)\}\{m_s'-\mu_s\}]$$

$$=E[\{m_r'-\mu_r)-r(m_1'-\mu_1)\mu_{r-1}\}\{m_s'-\mu_s\}]$$

$$=\text{Cov}\,(m_r', m_s')-r\mu_{r-1}\,\text{Cov}\,(m_1', m_3')$$

$$=\frac{1}{n}(\mu_{r+s}-\mu_r\mu_s-r\mu_{r-1}\mu_{s+1}) \quad \because \mu_1'=0 \quad ...(12·15)$$

Example 12·11. *Prove that in samples of size n from a symmetrical population, the first product-moment between the mean and any moment of even order vanishes to order 1/n.*

To order $1/n$, we have from (12·13)

$$\text{Cov } (m_1', m_r) = \text{Cov } (m_r, m_1')$$

$$= \frac{1}{n} (\mu_{r+1} - r\mu_2\mu_{r-1})$$

Now if r is even, μ_{r+1} and μ_{r-1} are moments of odd order and, therefore, vanish for a symmetrical population.

$$\text{Cov } (m_1', m_r) = 0.$$

Example 12·12. *Prove that in samples of size n from a symmetrical population, the variance of the even moments of samples about sample mean is the same as the variance of the corresponding even moments about the population mean.*

For a symmetrical distribution, μ_{r+1} and μ_{r-1} vanish and therefore, from (12·11) we have

$$\text{Var } (m_{2r}) = \frac{1}{n} (\mu_{4r} - \mu_{2r}^2) = \text{Var } (m_{2r}') \text{ from (12·10)}$$

Example 12·13. *Find the sampling variance of the following statistics : (i) Standard deviation (ii) Coefficient of variance (iii)* β_1.

(i) $\qquad m_2 = S^2$

$\therefore$ $\qquad \delta m_2 = 2S\delta S$

Hence squaring and taking mean values, we have

$$\text{Var } m_2 = 4\mu_2 \text{ Var } S$$

or

$$\text{Var } S = \frac{1}{4\mu_2} \text{ Var } m_2$$

$$= \frac{1}{4\mu_2} \cdot \frac{1}{n} (\mu_4 - \mu_2^2)$$

$$= \frac{\sigma^2}{2n} \text{ for a normal population.}$$

Hence for a normal population standard error of standard deviation is $\sigma/\sqrt{2}$.

(ii) The coefficient of variation V, is given by

$$V = \frac{100\sqrt{m_2}}{m_1'}$$

Taking logarithmic differentiations, we have

$$\frac{dV}{V} = \frac{dm_2}{2m_2} - \frac{dm_1'}{m_1'}$$

Hence, squaring and taking mean values, we get

$$\frac{\text{Var } V}{V^2}=\frac{1}{4m_2^2}\text{ Var } m_2-\frac{1}{m_2m_1'}\text{ Cov }(m_2, m_1')+\frac{1}{m_1'^2}\text{ Var } m_1'$$

$$=\frac{1}{n}\left(\frac{\mu_4-\mu_2^2}{4m_2^2}-\frac{\mu_3}{m_2m_1'}+\frac{\mu_2}{\mu_1'^2}\right)$$

Taking $m_r=\mu_r$, we have

$$\text{Var } V=\frac{V^2}{n}\left(\frac{\mu_4-\mu_2^2}{4m_2^2}-\frac{\mu_3}{\mu_2\mu_1'}+\frac{\mu_2}{\mu_1'^2}\right)$$

For a normal population, this reduces to

$$\text{Var } V=\frac{V^2}{n}\left(\frac{1}{2}+\frac{\mu_2}{\mu_1'^2}\right)=\frac{V^2}{2n}\left(1+\frac{2V^2}{10^2}\right)$$

$$\text{Var } V=V^2/2n \text{ approx.}$$

(*iii*) Let b_1 be the sample estimate of the coefficient β_1, then

$$\beta_1=\frac{\mu_3^2}{\mu_2^3}=\mu_3^2\times\frac{1}{\mu_2^3}$$

$$\therefore\quad d\beta_1=\frac{2\mu_3}{\mu_2^3}d\mu_3-\frac{3\mu_3^2}{\mu_2^4}d\mu_2$$

$$\therefore \text{Var}(\beta_1)=\left(\frac{2\mu_3}{\mu_2^3}\right)^2\text{ Var }\mu_3+\left(\frac{3\mu_3^2}{\mu_2^4}\right)^2\text{ Var }\mu_2$$

$$+2\left(\frac{6\mu_2^3}{\mu_2^7}\right)\text{Cov }(\mu_3, \mu_2)$$

$$\therefore \text{Var }(b_1)=\left(\frac{2\mu_3}{\mu_2^3}\right)^2\cdot\frac{\mu_6-\mu_3^2+9\mu_2^3-6\mu_2\mu_4}{n}$$

$$+\left(\frac{-3\mu_3^2}{\mu_2^4}\right)\cdot\frac{\mu_4-\mu_2^2+6\mu_2\mu_1^2-4\mu_1\mu_3}{n}$$

$$+2\left(\frac{-6\mu_3^3}{\mu_2^7}\right)\cdot\frac{\mu_5-\mu_3\mu_2-2\mu_1\mu_2-3\mu_2\mu_3-6\mu_0\mu_1\mu_2}{n}$$

$$=4\beta_1\cdot\left(\frac{\beta_4-\beta_1+9-6\beta_2}{n}\right)+2\beta_1^2\left(\frac{\beta_2-1}{n}\right)-12\beta_1\left(\frac{\beta_3-4\beta_1}{n}\right)$$

where $\beta_{2n+1}=\mu_3\mu_{2n+3}/\mu_2^{n+3}$, $\beta_{2n}=\mu_{2n+2}/\mu_2^{n+1}$ and $\mu_1=0$.

Hence $\text{Var }(b_1)=\beta_1(4\beta_4-24\beta_2+36+9\beta_1\beta_2-12\beta_3+35\beta_1)/n$

$\text{Var }(b_1)=6\beta_1/n$ for a normal parent.

Ex. 1. Show that for a symmetrical distribution :

$$\text{Cov }(m_{2r}, m_{2s})=\text{Cov }(m_{2r}', m_{2s}').$$

Ex. 2. Prove the following results :

(*i*) $$\text{Var }(m_3)=\frac{1}{n}(\mu_6-\mu_3^2-6\mu_4\mu_2+9\mu_2^3)$$

(*ii*) $$\text{Var }(m_4)=\frac{1}{n}(\mu_8-\mu_4^2-8\mu_5\mu_3+16\mu_2\mu_3^2)$$

(*iii*) $$\text{Var}\,(k_4)=\text{Var}\,\mu_4-12\mu_2\,\text{Cov}\,(\mu_1,\mu_2)+36\mu_2^2\,\text{Var}\,(\mu_2)$$

$$=\frac{1}{n}(\mu_8-12\mu_6\mu_2-8\mu_5\mu_3-\mu_4^2+48\mu_4\mu_2^2+64\mu^2{}_3\mu_2-36\mu^4{}_2)$$

(*iv*) $$\text{Var}\,(b_2)=\frac{1}{n}(\beta_6-4\beta_4\beta_2+4\beta_2{}^3-\beta_2{}^2+16\beta_4\beta_1-8\beta_3+16\beta_1)$$

where $$\beta_{2n+1}=\mu_3\mu_{2n+3}/\mu_2{}^{n+3},\ \beta_{2n}=\mu_{2n+2}/\mu_2{}^{n+1}.$$

Deduce that, for a normal parent, Var $(m_3)=6\sigma^6/n$, Var $(m_4)=96\sigma^8/n$, Var $(k_4)=24\sigma^8/n$.

Ex. 3. Given a distribution of values x_i, $i=1, 2,\ldots\ldots, n$, show that

(*i*) $$E\mid x_i-\mu_1'\mid=\delta$$

(*ii*) $$E\mid x_i-\mu_1'\mid=\sigma^2$$

where μ_1', δ and σ^2 are the parent mean deviation and variance respectively.

Hence show that, if d is the sample mean deviation,

$$\text{Var}\,(d)=\frac{1}{n^2}(\sigma^2-\delta^2)\ \text{approx.}$$

Deduce that, for a normal population,

$$\text{Var}\,(d)=\frac{\sigma^2}{n}\left(1-\frac{2}{\pi}\right)\ \text{approx.}$$

[**Hint.** $$d\alpha=\frac{1}{n}\sum_{i=1}^{n}\mid x_i-\alpha\mid$$

where $d\alpha$=sample mean deviation measured from the fixed point α.

Now $$E\,(d\alpha)=E\mid x_i-\alpha\mid=\delta\alpha$$

where δx is corresponding population mean deviation.

$$\text{Var}\,d\alpha=\frac{1}{n}\,\text{Var}\,(\mid x-\alpha\mid),$$

since the n sample observations are independent.

$$=\frac{1}{n}\{(E\mid x-\alpha\mid^2)-E^2(\mid x-\alpha\mid)\}$$

$$=\frac{1}{n}\{\sigma^2+(\alpha-\mu_1')^2-\delta\alpha^2\}$$

Let $\alpha=\mu_1'$, then Var $d\mu_1'=\frac{1}{n}[(\sigma^2-\delta^2\mu_1')]$

For a normal parent, $d\mu_1'=\sigma\sqrt{\frac{2}{\pi}}$]

Ex. 4. If $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ and $m_2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ are the sample mean and the sample variance, show that

$$\text{Cov}\,(\bar{x},\ m_2)=\frac{n-1}{n^2}\mu_3,$$

where μ_3 is the the third central moment of the parent population.

Show that if the population is symmetrical, $\bar{x}$ and m_2 are uncorrelated.

(*M.A., Eco., Statistics, Delhi, 1959*)

Ex. 5. (*a*) Show that

$$E(m_r')=\mu_r',\quad \text{Var}(m_r')=\frac{1}{n}(\mu_{2r}'-\mu_r'^2)$$

where m_r' is the rth sample moment about origin for a random sample $x_1, x_2\ldots\ldots$ x_n from a population. (*M.A. Madras, 1960*)

(*b*) Find the sampling variances (or standard error) of the coefficient of variance for samples of size n from a normal population with variance σ^2.

How will you use these variances to test the significance of m_r' and the coefficient of variation ?

(*M.A. Statistics, Delhi, 1960, 61, 65 ; M. A , Eco. Statistics, Delhi, 1958*)

(*c*) Show that $\frac{1}{n-1}\Sigma(x-\bar{x})^2$ is an unbiased estimate of the population variance and that the sampling variance of this estimate is given by

$$\frac{1}{n}\left(\mu_4-\frac{n-3}{n-1}\mu_2^2\right). \qquad (M.A.\ Madras,\ 1950)$$

(*d*) Obtain the formula for the variance of a function of a statistics in large samples when the variance of the statistic is known. Hence calculate the variance of b_1 where b_1 is the sample estimate of β_1.

(*M.A. Delhi, 1968*)

Ex. 6. A random sample of N individuals is drawn from an infinite population and out of this sample, a sub-sample of n is randomly drawn. Then mean of a character $\bar{x}$, in the sample is $\bar{X}$ and in the sub-sample is $\bar{x}$. Show that if the process of sampling and sub-sampling is repeated again and again, the standard errors of $d=\bar{x}-X$ is

$$\sigma_d=\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

where σ is the standard deviation of x in population.

If a sample of $4n$ be drawn, how large a sub-sample do you consider must be taken to make it practically certain that $\bar{x}$ will differ from $\bar{X}$ by less than $\sigma/5$?

(*M.A., Bombay, 1950*)

Ex. 7. M_r is the rth central moment in a sample of size n. If n is large, show that the covariance between M_r and M_s is

$$[\mu_{r+s}-s\mu_{r-1}\mu_{s+1}-r\mu_{s-1}\mu_{r+1}+rs\mu_{r-1}\mu_{s-1}\mu_2]/n$$

μ_r being the rth central moment in the population.

Show that in large samples from symmetrical distribution, the correlation between two consecutive central moments is zero. (*M.A., Bombay, 1950*)

12·14. Sampling from a Finite Population (without replacement). Let the values $y_1, y_2, \ldots, y_N$, constitute a finite population with mean μ and variance σ^2, so that

$$\mu=\frac{1}{N}(y_1+y_2+\ldots+y_N)$$

$$\sigma^2=\frac{1}{N}[(y_1-\mu)^2+(y_2-\mu)^2+\ldots+(y_N-\mu)^2]$$

$$=\frac{1}{N}[y_1^2+y_2^2+\ldots+y_N^2]-\mu^2$$

Let $x_1, x_2 \ldots\ldots, x_n$ be a random sample drawn without replacement from the population and let $\bar{x}$ be the sample mean ; then

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

and
$$E(\bar{x}) = \frac{1}{N}[E(x_1) + E(x_2) + \ldots + E(x_n)]$$

Now for a fixed value of i, the variate x_i takes values $y_1, y_2, \ldots, y_N$ each with probability $1/N$ so that

$$E(x_i) = \frac{1}{N}(y_1 + y_2 + \ldots + y_N) = \mu \quad (i = 1, 2, \ldots\ldots, n)$$

$$\therefore \quad E(\bar{x}) = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \mu. \qquad \ldots(12{\cdot}16)$$

Thus *the expected value of the sample mean in random sampling is the population mean* μ.

The variance $\sigma_{\bar{x}}^2$ of the sample mean $\bar{x}$ is

$$\sigma_{\bar{x}}^2 = E[\bar{x} - E(\bar{x})]^2 = E(\bar{x} - \mu)^2$$

$$= \frac{1}{n^2} E[(x_1 - \mu) + (x_2 - \mu) + \ldots\ldots + (x_n - \mu)]^2$$

$$\sigma_{\bar{x}}^2 = \frac{1}{n^2}\Big[\sum_{i=1}^{n} E(x_j - \mu)^2 + 2 \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{i=1}^{n} (x_i - \mu)(x_j - \mu)\Big]$$

Since $x_1, x_2, \ldots\ldots, x_n$ is a random sample, the n expected values in the first sum are the same, and for the same reason, the $n(n-1)$ expected values in the second sum are identical. Therefore

$$\sigma_{\bar{x}}^2 = \frac{1}{n^2}[nE(x_1 - \mu)^2 + n(n-1)E(x_1 - \mu)(x_2 - \mu)]$$

$$= \frac{1}{n} E(x_1 - \mu)^2 + \frac{n-1}{n} E(x_1 - \mu)(x_2 - \mu)$$

$\ldots(12{\cdot}15)$

Now

$$E(x_1 - \mu)^2 = \frac{1}{N}[(y_1 - \mu)^2 + (y_2 - \mu)^2 + \ldots + (y_N - \mu)^2] = \sigma^2$$

and *the covariance between any two members of the sample is*

$$E[(x_1 - \mu)(x_2 - \mu)] = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq 1}}^{N} (y_i - \mu)(y_j - \mu)$$

$$= \frac{1}{N(N-1)} \sum_{i=1}^{N} (y_i-\mu)[(y_1-\mu)+(y_2-\mu)$$

$$+\ldots+(y_N-\mu)]\,[-(y_i-\mu)]$$

$$= \frac{1}{N(N-1)}\left[\sum_{i=1}^{N} (y_i-\mu)\{0-(y_i-\mu)\}\right]$$

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

$$\therefore \quad \text{Var}\,\bar{x} = \frac{1}{n}\sigma^2 + \frac{1}{n}(n-1)\left(-\frac{\sigma^2}{N-1}\right) = \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)$$

$$= \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}. \qquad \ldots(12{\cdot}17)$$

Thus *the variance of the sample mean for random sampling from finite population is*

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

In the same way, can we find μ_3 μ_4, for the distribution of sample mean $\bar{x}$. The method is similar, though obviously the algebra is a little more complicated.

Ex. 1. Show that covariance between any two members tends to zero as N tends to infinity. Explain this.

Ex. 2. Deduce the $\sigma_{\bar{x}}^2$ for a random sample from (*i*) a finite population when each member is replaced before the next drawing (*ii*) an infinite population is, in each case, σ^2/n.

Ex. 3. Show that, without any loss of generality, the population mean can be assumed to be zero. Rewrite the above proof for Var $\bar{x}$ with the simplification.

Ex. 4. A box contains N tickets bearing numbers $y_1, y_2, \ldots, y_N$, n tickets are drawn at random. If s denotes the sum of the numbers drawn, prove that

$$E(s)=nm,\ \text{Var}\,(s)=n\sigma^2\cdot\frac{N-n}{N-1},$$

where m is the mean of the numbers appearing on all tickets in the box.

Ex. 5. Show that the correlation coefficient between any two members of the sample is $\frac{1}{N-1}$ and explain why it is negative.

Ex. 6. (*a*) Verify formulae (12·7) and (12·8) for the set 2, 4, 6 taking samples of two.

(*b*) Let the population consist of 6 values 1, 2, 3, 4, 5, 6 ;

(*i*) write the means of all the 15 possible samples of size 2;

(*ii*) find the mean of all these 15 means and verify that it is the same as the population mean *viz.*, 3·5 ;

(*iii*) find the variance of these 15 values and verify that it satisfies the formula deduced above ;

(*iv*) find $\mu_3(\bar{x})$ for these 15 values and verify that it satisfies the formula

$$\mu_3(\bar{x}) = \frac{(N-n)(N-2n)}{n^2(N-1)(N-2)}\,\mu_3$$

where μ_3 is the population value ;

(*v*) find s^2 for each sample and show that the average value of the 15 values of s^2 is $\frac{n-1}{n}\,\mu_2$.

Ex. 7. Show that the sampling variance of the proportion of males in a random sample of n people drawn from a population of N is given by

$$\frac{N-n}{N-1}\cdot\frac{p(1-p)}{n}, \qquad \textit{(M.A., Bombay, 1952)}$$

where p is the proportion of males in the population. Obtain also an unbiased estimate of this sampling variance in terms of sample estimates.

(I.A.S , 1948, 49 ; B.U., 1957 ; M.A., Delhi, 1968)

[If each male is assigned a number unity and each female a number zero, then $\sigma^2 = p(1-p)$.]

Ex. 8. From the formula for the sampling variance of the sample mean, deduce an expression for the variance of the sample proportion in sampling from a binomial population of size N.

Ex. 9. Deduce the result, of the two exercises, from the expression for the variance of the hypergeometric distribution.

Ex. 10. From a finite population of size N and variance

$$\sigma^2 = \frac{1}{N-1}\sum_{1}^{N}(x_i-\bar{x})^2,$$

a random sample of n is drawn without replacement. Show the arithmetic mean and variance given by

$$\frac{1}{n}\sum_{1}^{n}x_i \text{ and } \frac{1}{n-1}\sum_{1}^{n}(x_i-\bar{x})^2,$$

are unbiased estimates of the corresponding population values.

Ex. 11. Show that if N is large the error in writing $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is approx $50(n-1)/N$ per cent of the value of $\sigma_{\bar{x}}$.

EXERCISE 12

1. (*a*) Explain the terms :

(*i*) Confidence limits, (*ii*) Critical region,
(*iii*) Null hypothesis, (*iv*) Statistic and parameter.

(*b*) Suggest a possible source of bias in the following :

(*i*) The mean income per family in a certain town is sought to be estimated by sampling from motor owners.

(*ii*) Readers of newspapers are sampled by printing in it an invitation to them to send up their observations on some typical events.

2. (*a*) A coin is tossed 10,000 times and it turns up head 5,195 times. Is it reasonable to think that the coin is unbiased ?

(*b*) A coin is tossed 1,400 times and it turns up head 216 times. Discuss whether the coin may be unbiased one. (*M.Sc., Agra, 1949*)

[**Ans.** (*a*) No Coin is biased. (*b*) The coin may be regarded as an unbiased one]

3. Certain crosses of the peas gave 5,321 yellow and 1,804 green seeds. The expectation is 25 per cent green seeds on Mendelian hypothesis. Is the divergence significant or might have occurred as due to fluctuations of simple sampling ?

[**Hint.** Total number of seeds is 7,125 and the expected green seeds $=1781$.

The deviation 23 is about 0·6 *s.d.* which is 36·6 and hence it might have occurred as due to fluctuations of simple sampling.]

4. (*a*) A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that *S.E.* of the proportion of bad one's in a sample of this size is 0·015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8·5 and 17·5. (*I.A.S , 1954*)

(*b*) 400 oranges are taken from a large consignment and 50 are found to be bad. Estimate the percentage of bad oranges in the consignment and assign the limits within which the percentage lies.

[**Hint.** (*a*) S.E. of the proportion of bad pineapples

$$=\sqrt{\frac{65}{500}\times\frac{435}{500}\times\frac{1}{500}}=0{\cdot}015 \text{ or } 1{\cdot}5\%$$

Limits of the bad pineapples in the consignment would be

$$\left(\frac{65}{500}\times 100\right)+3\times 1{\cdot}5=\text{etc.}$$

(*b*) Percentage of bad oranges=12·5. Probable limits are 7·7% and 17·4%.]

5. (*a*) In a simple sample of 600 men from a certain large city, 450 are found to be smokers. In one of 900 from another large city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to prevalence of smoking among men ?

[*B.Sc., Meerut, 1969 ; B.Sc., Hons., Sub., 1965, 67 (S)*]

[**Ans.** The hypothesis that the populations are similar is almost certainly wrong.]

(*b*) A sample of 900 days is taken from meteorological records of a certain district and 100 of them are found to be foggy. What are the probable limits to the percentage of foggy days in the district ? (*B.Sc., Kurkshetra, 1971*)

[**Ans.** Limits are $\left(\frac{1}{9}\times 100\right)\pm 3\times 1{\cdot}05$ per cent *i.e.*, 8% and 14·25%]

(*c*) In a year there are 956 births in a town *A*, of which 52·5% were males while in towns *A* and *B* combined, this proportion in a total of 1,406 births was 0·496. Is there any significant difference in the proportion of male births in the two towns? (*B.A. Hons., Delhi, 1971* ; *I.A.S. 1961*)

[**Hint.** $0{\cdot}496=\frac{n_1p_1+n_2p_2}{n_1+n_2}=\frac{956\times{\cdot}525+450\times p_2}{1406}$

This gives $p_2=0{\cdot}432$, $z=\frac{p_1\sim p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$.]

[**Ans.** There is significant difference in the proportion of male births in the towns *A* and *B*.]

6. In a random sample of 500 persons from town *A*, 200 are found to be consumers of rice. In a sample of 400 from town *B*, 200 are found to be consumers of rice. Discuss the question whether the data reveal a significant difference between *A* and *B* so far as the proportion of rice consumers is concerned.

[**Ans.** Yes, the deviation is significant.]

7. One thousand articles from a factory are examined and found to be 3 per cent defective. Fifteen hundred similar articles from a second factory are found to have only 2 per cent defective. Can it reasonably be concluded that the products of the first factory are inferior to the second ?

[**Ans.** It cannot be concluded that the products of the first factory are inferior to the second since the deviation which is 1 per cent is about 1·6 times the S.E. of the proportion of defectives which is $\left[2{\cdot}4\times97{\cdot}6\left(\frac{1}{1000}\times\frac{1}{1000}\right)\right]^2$ *i.e.*, 0·625 per cent approx.]

8. The subject under investigation is the measure of dependence of Tamil on words of Sanskrit origin. One newspaper article reporting the proceeding of the Constituent Assembly contained 2,025 words of which 729 were declared by a literary critic to be of Sanskrit origin. A second article by the same author describing atomic research contained 1,600 words of which 640 were declared by the same critic to be of Sanskrit origin. Assuming that the simple sampling conditions hold, estimate the limits for the proportion of Sanskrit words in the writer's vocabulary and examine whether there is any significant difference in the dependence of the writer on words of Sanskrit origin in writing on these two subjects. (*I.A.S., 1947*)

[**Hint.** The estimate of percentage of Sanskrit terms in the writer's vocabulary in both the articles taken together is given by

$$p=\frac{(729+640)\times100}{2025+1600}=\frac{1369}{3625}\times100=37{\cdot}77$$

S.E. of the difference between the percentages

$$=\sqrt{37{\cdot}77\times62{\cdot}33\left(\frac{1}{2025}+\frac{1}{1600}\right)}=1{\cdot}16 \text{ per cent}$$

$\therefore$ Observed difference is about 3·4 times the S.E. and so could not have arisen through fluctuations of sampling,

Limits are $\frac{1369}{3625}$ per cent $\pm 3\sqrt{\frac{1369}{3625}\times\frac{2256}{3625}\times\frac{1}{3625}}$ per cent

i.e., 35·34 and 40·20 per cent approx.

9. A sample of 900 members is found to have a mean of 3·4 cms. Can it be reasonably regarded as a simple sample from a large population with mean 3·2 cms. and S.D. 2·3 cms. ?

[Ans. No. The deviation 0·2 cm. is 2·6 times the S.E. of the mean which is ·077 cm.]

10. The mean of a certain normal distribution is equal to the standard error 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.

(*M.A., Punjab, 1952*)

$$\left[\text{Ans. } \frac{1}{\sqrt{2\pi}}\int_{\frac{1}{2}}^{\infty} e^{-\frac{1}{2}t^2}\,dt \text{ or } 0{\cdot}3085\right]$$

11. A normal population has mean 0·1 and S.D. 2·1. Find the probability that the mean of a simple sample of 900 members will be negative. [Ans. 0·0764]

[Hint. $z=\frac{\bar{x}-0{\cdot}1}{\left(\frac{2{\cdot}1}{30}\right)}$ $\therefore$ $\bar{x}=0{\cdot}1+{\cdot}07z$ $\therefore$ $x<0$ gives $z<-\frac{10}{7}$

$\therefore$ $P(\bar{x}<0)=P(z>1{\cdot}43)$

$=0{\cdot}5-P(0<z<1{\cdot}43)=$etc.]

12. The sampling distribution of a statistic is a normal population with mean 9 units and S.D. 2 units. What is the probability that the sample will give a value of the statistic greater than 12 units ? [Ans. 0·0668]

13. The means of simple samples of 1,000 and 2,000 are 67·5 and 68·8 inches respectively. Can the samples be regarded as drawn from the same population of S.D. 2·5 inches ? (*B.A. Hons. Delhi, 1952 ; M Sc., Agra, 1954*)

[Ans. No. The difference 0·5 is more than 5 times the S.E. of the difference which is 0·097 inch nearly.]

14. 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 ?

(*B.Sc., All., 1960 ; B.Sc., Agra, 1961*)

[Ans. 62 or more]

15. The guaranteed average life of a certain type of electric light bulbs is 1,000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90 per cent of the bulbs do not fall short of the guaranteed average by more than 2·5 per cent. What must be minimum size of the sample ?

[Ans. 41]

[**Hint.** We wish $\bar{x} > 100 - 2{\cdot}5$ of $1000 = 975$.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{975 - 1000}{\frac{125}{\sqrt{n}}} \text{ or } z = \frac{-\sqrt{n}}{5}$$

We have $P\left(z > -\frac{\sqrt{n}}{5}\right) = 0{\cdot}90$ or $P(0 < z < \sqrt{n}/5) = 0{\cdot}40$

$\therefore$ $\sqrt{n}/5 = 1{\cdot}28$ or $n = 41$ approx.]

16. Given that on the average 4% of insured men of age 65 die within a year, and that 60 of a particular group of 1,000 such men died within a year. Can this group be regarded as a representative sample ?

[**Ans.** No, since the deviation of a proportion of deaths *i.e.* ($\cdot 06 - \cdot 04$) or $\cdot 02$ is more than three times the S.E. of the proportion which is

$$\sqrt{\frac{04 \times \cdot 96}{1000}} \text{ } i.e., \cdot 0062.]$$

17. A random sample of 1,000 men from Northern India gives their mean wage to be Rs. 2·50 per day with a S.D. of Rs. 1·50. A sample of 1,500 men from Southern India gives a mean wage of Rs. 2·70 per day with a S.D. of Rs. 2. Discuss whether the mean rate of wages varies as between the two regions.

[**Ans.** No. The difference of the means is less than three times the S.E. of the difference.]

18. A simple sample of heights 6,400 Englishmen has a mean of 67·85 inches and a S.D. of 2·56 inches while a simple sample of heights of 1,600 Australians has a mean of 68·55 inches and a S.D. of 2·52 inches. Do the data indicate that the Australians are on the average taller than Englishmen ?

[**Ans.** Australians are on the average taller than Englishmen. The difference of the means *viz.*, 0·70 in. is ten times the S.E. which is

$$\sqrt{\frac{(2{\cdot}56)^2}{6400} + \frac{(2{\cdot}52)^2}{1600}} \text{ } i.e., 0{\cdot}07 \text{ inch.}$$

19. A potential buyer of light bulbs bought 30 bulbs of each of 2 brands. Upon testing these bulbs, he found that brand *A* had a mean life of 1,282 hours with a S.D. of 80 hours, whereas brand *B* had a mean life of 1,208 hours with a S.D. of 94 hours. Can the buyer be quite certain that the two brands differ in quality ? (*B.A. Punjab, 1961 S.*)

[**Ans.** Brand *A* is to be preferred. The difference of the means, *i.e.*, 74 hours is greater than 3 times its S.E. which is

$$\sqrt{\frac{(80)^2}{50} + \frac{(94)^2}{50}} = 17{\cdot}5 \text{ hours.}$$

20. The mean yield of two sets of plots and their variability are as given. Examine :

(*i*) whether the difference in the mean yields of the two sets of plots is significant, and

(*ii*) whether the difference in the variability in yields is significant.

	Set of 40 plots	*Set of 60 plots*
Mean yield per plot	1258 lb.	1243 lb.
S.D per plot	34	28

(*I.A.S., 1965*)

[**Ans.** (*i*) significant at 5% level,
(*ii*) Not significant at 5% level.]

21. Two populations have their means equal, but S.D. of one is twice the other. Show that in the samples of size 2000 from each drawn under simple sampling conditions, the difference of means will in all probability, not exceed $0{\cdot}15\sigma$ where σ is the smaller S.D. What is the probability that the difference will exceed half this amount ?

(*B.Sc. Nagpur 1963*)

[**Hint.** $z = \dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{(2\sigma)^2}{n_1}}} = \dfrac{|\bar{x}_1 - \bar{x}_2|}{0{\cdot}05\,\sigma}$, $n_1 = 2000 = n_2$

We should have $|z| < 3$. Hence etc.

We want $P[|\bar{x}_1 - \bar{x}_2| > \frac{1}{2} \times 0{\cdot}15\,\sigma] = P[|z| > \frac{3}{2}]$

$= 1 - P\left[|z| < \frac{3}{2}\right] =$ etc.

22. A sample of n objects is taken from an infinite population in which the proportion $p = (1-q)$ of objects possess a certain attribute A. Prove that $\bar{p}\bar{q}/(n-1)$ is an unbiased estimator of pq/n where $\bar{p}(=1-q)$ is the proportion of objects possessing attribute A in the sample.

Instead of taking a sample in the above manner, the objects of the sample are selected at random in succession until a fixed number m of objects possessing A is obtained in the sample. Find the probability distribution of the size n of the sample and prove that $(m-1)/(n-1)$ is an unbiased estimate of p, but m/n is not an unbiased estimate of p. (*M.A., Delhi, 1956 ; I.C.A.R., Delhi, 1949*)

23. A bag contains N balls numbered 1 to N. Let X denote the larger number in n drawings when random sampling with replacement is used, each drawing consisting of one ball and noting the number on the ball. Find the probability distribution of X and show that the expectation of X is

$$[N^{n+1} - \sum_{k=1}^{N} (k-1)^n]/N^n$$

and show that an approximation for the latter for large N is $nN/(n+1)$.

(*M.A. Delhi, 1955, 57*)

24. Suppose x takes values 0 and 1 with probability $(1-p)$ and p respectively. If n independent observations $x_1, x_2, \ldots\ldots, x_n$ are taken and r denotes the number of observations which are as large as x_1, find the probability distribution of r. (*M.A., Bombay, 1955*)

[**Ans.** Reqd. distribution of r is $p \times {}^{n-1}c_{r-1} p^{r-1} q^{n-r} + q \times {}^{n-1}c_{r-1} q^{r-1} p^{n-r}$]

25. It is known that 40% of a population die within a year, 30 persons are vaccinated of which only 5 die in a year. The inventor of the vaccine claims that the vaccine has reduced the mortality by 20%. Examine the following:

(*a*) Justify his claim.

(*b*) Discuss whether the vaccine is effective at all.

(*c*) Find the least value of n such that if $n/6$ die after vaccination, his claim is definitely accepted. (*I.C.A.R., Delhi, 1947*)

[**Hint.** (*a*) $z=\dfrac{0{\cdot}32-0{\cdot}96}{\sqrt{\dfrac{0{\cdot}32\times{\cdot}68}{30}}}>1{\cdot}96$. Hence the claim is rejected.

(*b*) $z=\dfrac{{\cdot}4-{\cdot}166}{\sqrt{\dfrac{4\times{\cdot}6}{36}}}>1{\cdot}96$. Hence the vaccine is effective.

(*c*) n is the least integer given by $\dfrac{{\cdot}32-\dfrac{n}{6}}{\sqrt{\dfrac{{\cdot}32\times{\cdot}68}{n}}}>1{\cdot}96$.]

26. Experience has shown that 40% of the young of a certain breed of animals die of a certain disease during the first year of their life. The discoverer of a new vaccine claims that its use at least haves mortality and in support of his claim cites an experiment in which out of 40 youngs treated with the vaccine only 8 died. Examine how far his claim is justified ? (*I.C.A.R., Delhi, 1947*)

27. If for one half of n events, the chance of success is p and the chance of failure is q, whilst for the other half, the chance of successes is q and the chance of failure is p. Show that the standard deviation of the number of successes is the same as if the chance of successes were p in all cases but the mean of the number of successes is $\frac{1}{2}n$ and not np.

(*B. Sc. Delhi 1976* ; *B.Sc., Hons., Cal., 1964* ; *B.Sc., Gauhati, 1964*)

[**Hint.** If x_1 and x_2 denote the number of successes in the first and the second half of the n events respectively, then we have

$$E(n_1)=np/2,\ E(n_2)=nq/2,\ \text{Var}\,(n_1)=\tfrac{1}{4}npq,\ \text{Var}\,(n_2)=\tfrac{1}{4}npq$$

$$\therefore \quad E(n_1+n_2)=E(n_1)+E(n_2)=n/2$$

and $$\text{Var}\,(n_1+n_2)=\text{Var}\,n_1+\text{Var}\,n_2=\tfrac{1}{4}npq=\tfrac{1}{4}npq=npq.]$$

28. In samples of size n from an infinite population find the covariance between the arithmetic mean and variance. (*M.A., Madras, 1956*)

29. The moment statistic m_r is defined as $\dfrac{1}{n}\Sigma\,(x-\bar{x})^r$. Show that in normal samples

$$\text{Cov}\,(m_r, m_n)=\frac{\sqrt{3(n-1)}}{\sqrt{4n^2-9n+6}}$$

30. Show that the coefficient of correlation between the mean of a sample of n individuals and its square *s.d.* is given by

$$\frac{\sqrt{\beta_1(n_1-1)}}{\sqrt{\beta_2(n-1)-n+3}}$$

(*M.A., Bombay, 1954* ; *B.A., Hons., Delhi 1962*)

13

Exact Sampling Distributions

13·1. Introduction. The sampling distributions play an important role in statistical inference. We construct from the observations different statistics to estimate the population parameters. The sampling distributions of these statistics will, in general (but not always, see Distribution Free Methods, Chapter 18) depend on the form and the parameters of the parent population. The probability of the observed value of the statistics then permits the making of statements about the parameters and hence we are able to draw conclusions about the parent population. The sampling distributions are thus fundamental to the whole subject and we shall obtain the sampling distributions of important statistics when the population from which the sample is drawn is specified. We shall assume that the sampling is simple and random. Thus, if the parent distribution is $dF(x)$, the simultaneous distribution of n values $x_1, x_2 \ldots\ldots x_n$ is

$$dF\ (x_1)\ dF\ (x_2), \ldots\ldots dF\ (x_n)$$

and if u is the statistic

$$u = u(x_1, x_2 \ldots\ldots, x_n)$$

the distribution function of u is given by

$$F(u) = \int\int \ldots \int dF(x_1)\ dF(x_2), \ldots, dF(x_n) \qquad \ldots(13\cdot1)$$

the integration being taken over the domain of the x's such that $u(x_1, x_2 \ldots\ldots, x_n) \leqslant u_0$ where u_0 is the fixed value of u.

The evaluation of these integrals is generally a complicated task and either of the following four techniques may be used in simplifying the task.

(*a*) *Geometrical Method* : It involves the use of geometrical terminology *i.e.*, the geometry of Euclidean Hyperspace.

(*b*) *Analytical Method* : The evaluation of integrals is done by the changes of variables.

(*c*) *Method of characteristic Functions.*

(*d*) *Method of Induction.*

In this chapter we shall use only the first method ; the other three methods will be explained in the next chapter.

If $x_1, x_2, \ldots\ldots, x_n$ is a random sample drawn from a standard normal population, we shall first derive the sampling distributions of

$$(i)\ \chi^2 = x_1^2 + x_2^2 + \ldots + x_n^2 = \sum_{i=1}^{n} x_i^2,$$

$$(ii)\ \bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ and}$$

$$(iii)\ s^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

From the sampling distribution of χ^2, $\bar{x}$ and s^2 and the distributions of sum and ratio of two independent χ^2 variates shall deduce the exact sampling distributions of the statistics '*t*', '*F*' and '*z*'. The distributions of these three statistics together with those of χ^2, $\bar{x}$ and s^2 constitute the most important exact sampling distributions Statistics.

13·2. Distributions of the sum of squares of independent standard normal variates [The Chi-Square Distribution]

The following geometrical proof is due to Prof. R.A. Fisher.

Let $x_1, x_2, \ldots\ldots, x_n$ be n independent normal variates, each of which is distributed normally with mean zero and variance unity. We require the sampling distribution of

$$\chi^2 = x_1^2 + x_2^2 + \ldots + x_n^2 = \sum_{i=1}^{n} x_i^2$$

The probability that random value of the variable x_i will fall in the interval dx_i is

$$\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x_i^2}\, dx_i$$

---

*The distribution was discovered by Helmert in 1875 and rediscovered independently in 1900 by Karl Pearson who devised by means of it the χ^2 test of goodness of fit which we shall discuss in Chapter 16.

Since x_i's are independent variates the probability that the values of the variates will fall simultaneously in the respective intervals dx_i, $i=1, 2, \ldots, n$ is

$$dP=\frac{1}{(2\pi)^{n/2}}\ e^{-\frac{1}{2}\chi^2}\ dx_1\, dx_2 \ldots dx_n \qquad \ldots(13\cdot2)$$

Let us represent the sample by the point E with co-ordinates $(x_1, x_2, \ldots, x_n)$ in Euclidean hyperspace of n dimensions. Then dP is the probability that E will fall in the element of the volume $dx_1 dx_2 \ldots dx_n$.

To obtain the distribution of χ^2 or χ we find out how the volume element $\prod_{i=1}^{n} dx_i$ transforms when χ lies between χ and $\chi+dx$.

Now when χ is constant, $\sum_{i=1}^{n} x_i^2=\chi^2$ represents a hypersphere with radius χ and centre at the origin. As $\sqrt{\Sigma x_i^2}$ lies between χ and $\chi+d\chi$ the sample point E lies between two concentric hyperspheres of radii χ and $\chi+d\chi$ and the volume of the annulus between the two concentric hypersphere is proportional to $d(\chi^n)$ *i.e.*, to $\chi^{n-1}\, d\chi$. Hence the probability that the value of χ from the random sample will fall in the interval $d\chi$ is from (13·2) proportional to

$e^{-\frac{1}{2}\chi^2}\ \chi^{n-1}\, d\chi$ *i.e.*, proportional to $e^{-\frac{1}{2}\chi^2}\ (\frac{1}{2}\chi^2)^{\frac{n-2}{2}}\, d(\frac{1}{2}\chi^2)$.

Since χ^2 must lie between 0 and ∞ and the integral of the probability density function over the whole range must be unity, the constant factor is $1/\Gamma(n/2)$

Hence the probability that χ^2 lies in the interval $d\chi^2$ is given by

$$dP=\frac{1}{\Gamma(n/2)}\ e^{-\frac{1}{2}\chi^2}\ (\tfrac{1}{2}\chi^2)^{\frac{n-2}{2}}\, d(\chi^2)$$

or

$$dP=\frac{1}{2^{n/2}\Gamma(n/2)}\cdot e^{-\frac{1}{2}\chi^2}\ (\chi^2)^{\frac{n-2}{2}}\, d(\chi^2),\ 0\leqslant\chi^2<\infty \qquad \ldots(13\cdot3)$$

The distribution (13·3) is called the χ^2 distribution with n *degree of freedom* (*d.f.*). The number of independent variates is called the number of degrees of freedom. (The variate having the distribution (13·3) is called the χ^2, chi-square, variate with n *d.f.*)

Cor. 1. If x_i's are n independent normal variates with mean μ_i, and *s.d.* σ_i $(i=1, 2, \ldots\ldots, n)$, then

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$$

is a χ^2 variate with n *d.f.* since $(x_i-\mu_i)/\sigma_i$ are n independent standard normal variates.

Cor. 2. Since for random sampling from a normal population with mean μ and *s.d.* σ, $\bar{x}$ is distributed normally about mean μ with *s.d.* $\frac{\sigma}{\sqrt{n}}$ (§8·2, Cor. 1, it follows that $\left[\frac{\sqrt{n}\,(\bar{x}-\mu)}{\sigma}\right]^2$ is a χ^2 variate with 1 *d.f.*

Cor. 3. Let $x_1, x_2, \ldots\ldots, x_n$ be n standard normal variates but instead of being independent, let these be subject to p *independent linear constraints* of the type

$$a_{i1}x+a_{i2}x+\ldots\ldots+a_{in}x_n=b_i, \quad (i=1, 2, \ldots\ldots, p).$$

This means that in Euclidean hyperspace, the sample point is restricted to lie upon p hyperplanes. The first hyperplane cuts the hypersphere of constant density in a hypersphere of one lower dimension, the second hyperplane cuts this hypersphere in hypersphere of one dimension further less and so on. The result of p intersections will be a hypersphere of p lower dimensions and the distributions of χ^2 subject to p independent linear constraints would be obtained by replacing n by $\nu=n-p$ in (13·3) so that

$$dP=\frac{1}{2^{\frac{1}{2}\nu}\Gamma(\nu/2)}\cdot e^{-\frac{1}{2}\chi^2}(\chi^2)^{\frac{1}{2}(\nu-2)}d(\chi^2),\ 0\leqslant\chi^2<\infty \qquad \ldots(13\cdot4)$$

Thus we observe that *each independent linear constraint reduces the number of degrees of freedom by one.*

13·2·1. Chief features of the Chi-square Probability Curve. The equation to the Chi-square probability curve is

$$y=\frac{1}{2^{n/2}\,\Gamma n/2}\ e^{-\frac{1}{2}\chi^2}\ (\chi^2)^{(n-2)/2} \qquad \ldots(13\cdot5)$$

In Fig. 33, the curve has been drawn for $\nu=1, 2, \ldots\ldots, 6$ where $\nu=n-1$.

The probability that value of χ^2 from a random sample will not exceed a fixed value χ_0^2 is obtained by integrating (13·5) with

respect to χ^2 from 0 to χ_0^2. Similarly, if P be the probability that the value of χ^2 from a random sample will exceed χ_0^2, we have

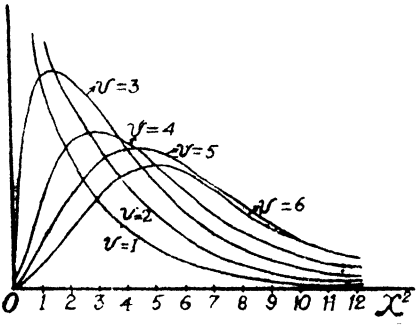


Fig. 33. χ^2 Probability Curve.[1]

Table 3. Values of χ^2 with probability P of being exceeded in random sampling
ν = number of degrees of freedom

ν \ P	0·99	0·95	0·50	0·30	0·20	0·10	0·05	0·01
1	0·0002	0·004	0·46	1·07	1·64	2·71	3·84	6·64
2	0·020	0·103	1·39	2·41	3·22	4·00	5·99	9·21
3	0·115	0·35	2·37	3·66	4·64	6·25	7·82	11·34
4	0·30	0·71	3·36	4·88	5·99	7·78	9·49	13·28
5	0·55	1·14	4·35	6·06	7·29	9·24	11·07	15·09
6	0·87	1·64	5·35	7·23	8·56	10·64	12·59	16·81
7	1·24	2·17	6·35	8·38	9·80	12·02	14·07	18·48
8	1·65	2·73	7·34	9·52	11·03	13·36	15·51	20·09
9	2·09	3·32	8·34	10·66	12·24	14·68	16·92	21·67
10	2·56	3·94	9·34	11·78	13·44	15·99	18·31	23·21
11	3·05	4·58	10·34	12·90	14·63	17·28	19·68	24·72
12	3·57	5·23	11·34	14·01	15·81	18·55	21·03	26·22
13	4·11	5·89	12·34	15·12	16·98	19·81	22·36	27·69
14	4·66	6·57	13·34	16·22	18·15	21·06	23·68	29·14
15	5·23	7·26	14·34	17·32	19·31	22·31	25·00	30·58
16	5·81	7·96	15·34	18·42	20·46	23·54	26·30	32·00
17	6·41	8·67	16·34	19·51	21·62	24·77	27·59	33·41
18	7·02	9·39	17·34	20·60	22·76	25·99	28·87	34·80
19	7·63	10·12	18·34	21·69	23·90	27·20	30·14	36·19
20	8·26	10·85	19·34	22·78	25·04	28·41	31·41	37·57
21	8·90	11·59	20·34	23·86	26·17	29·62	32·67	38·93
22	9·54	12·34	21·34	24·94	27·30	30·81	33·92	40·29
23	10·20	13·09	22·34	26·02	28·43	32·01	35·17	41·64
24	10·86	13·85	23·34	27·10	29·55	33·20	36·42	42·98
25	11·52	14·61	24·34	28·17	30·68	34·38	37·65	44·31
26	12·20	15·38	25·34	29·25	31·80	35·56	38·88	45·64
27	12·88	16·15	26·34	30·32	32·91	36·74	40·11	46·96
28	13·56	16·93	27·34	31·39	34·03	37·92	41·34	48·28
29	14·26	17·71	28·34	32·46	35·14	39·09	42·56	49·59
30	14·95	18·49	29·34	33·53	36·25	40·26	43·77	50·89

Reproduced by permission of the author, Professor R.A. Fisher, from his book on *Statistical Methods for Research Workers.*

$$P=\frac{1}{2^{n/2}\Gamma n/2}\int_{\chi_0^2}^{\infty} e^{-\frac{1}{2}\chi^2}(\chi^2)^{\frac{1}{2}n-1}\,d(\chi^2).$$

The value of χ_0^2 have been tabulated for various fixed value of P and for values of n from 1 to 30. For values of $n>30$ $\sqrt{2\chi^2}-\sqrt{2n-1}$ is asymptotically a standard normal variate. [See Ex. 4 and 7 at the end of this chapter]. This property is made use of for finding χ_0^2 for a given P and a given $n>30$. If $P=\alpha/100$, then χ^2 is called the α per cent value of the Chi-square distribution with n $d.f.$

13·2·2. Some Properties of the Chi-square Distribution. The moment generating function of χ^2 distribution with respect to origin is :

$$M_0(t)=\int_0^{\infty} e^{t\chi^2}\frac{1}{2^{n/2}\Gamma n/2}\,e^{-\frac{1}{2}\chi^2}(\chi^2)^{\frac{1}{2}n-1}\,d(\chi^2)$$

$$=(1-2t)^{-\frac{n}{2}},\ |\,2t\,|<1 \qquad \ldots(13\cdot6)$$

$$=1+\frac{n}{2}(2t)+\frac{\left(\frac{n}{2}\right)\left(\frac{n}{2}+1\right)}{2\,!}(2t)^2+\ldots\ldots$$

$$+\frac{\frac{n}{2}\left(\frac{n}{2}+1\right)\ldots\ldots\left(\frac{n}{2}+r-1\right)}{r\,!}(2t)^r+\ldots$$

$$\therefore\ \mu_r'=2^r\cdot\frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}+2\right)\ldots\ldots\left(\frac{n}{2}+r-1\right)$$

Consequently the mean value of χ^2 is n and variance is $n(n+2)-n^2=2n$ and thus $(\chi^2-n)/\sqrt{2n}$ is a standard variate.

Taking logarithms of both sides of (13·6) and writing down the expansion of the right side ; we have

$$\log M_0(t)=\frac{n}{}\left[\,2t+\frac{(2t)^2}{2}+\ldots+\frac{(2t)^r}{r}+\ldots\ldots\,\right]$$

Hence $k_r=n\,.\,2^{r-1}\,.\,(r-1)\,!$

Cor. Additive Property of Independent Chi-square Variates.

If χ_1^2 and χ_2^2 are independent χ^2 variates with n_1 and n_2 d.f., then $\chi_1^2+\chi_2^2$ is a χ^2 variate with (n_1+n_2) d.f.

Proof. *m.g.f.* of $(\chi_1^2+\chi_2^2)$

$=m.g.f.$ of $\chi_1^2 \times m.g.f.$ of χ_2^2 (§ 5·6·2)

$=(1-2t)^{-n_1/2} \times (1-2t)^{-n_2/2} = (1-2t)^{-(n_1+n_2)/2}$

and comparing with (13·6), the result follows.

Cor. 2. The Chi-square distribution tends to normal distribution as n tends to infinity.

Pooof. By § 5·7·1, we have

$$M(t) = e^{-\frac{nt}{\sqrt{2n}}} M_0\left(\frac{t}{\sqrt{2n}}\right) = e^{-\frac{nt}{\sqrt{2n}}}\left(1-\frac{2t}{\sqrt{2n}}\right)^{-\frac{n}{2}}$$

$$\therefore \quad \log_e M(t) = \frac{-nt}{\sqrt{2n}} - \frac{n}{2}\log_e\left(1-\frac{2t}{\sqrt{2n}}\right)$$

$$\Rightarrow \quad \log_e M(t) = -\frac{nt}{\sqrt{2n}} + \frac{n}{2}\left[\frac{2t}{\sqrt{2n}} + \frac{1}{2}\left(\frac{2t}{\sqrt{2n}}\right)^2 + \ldots\right]$$

$$\Rightarrow \quad \log_e M(t) = \tfrac{1}{2}t^2 + 0\left(\frac{1}{n}\right)$$

Hence as $n\to\infty$, $\log_e M(t)\to\frac{1}{2}t^2$ $\Rightarrow$ $M(t)\to e^{\frac{1}{2}t^2}$

Hence the result follows by § 8·1·2.

Ex. 1. (*a*) Show that if x is a standard normal variate, $\frac{1}{2}\chi^2$ is a Gamma variate with parameter $\frac{1}{2}$.

(*b*) Use Ex. 1 (*a*) above and the additive property of Gamma variates to show that $\frac{1}{2}\chi^2$ is a Gamma variate with parameter $n/2$. Hence deduce the χ^2 distribution.

Ex. 2. From the table, find χ_0^2 given

$$P[\chi^2 \geqslant \chi_0^2] = \cdot 5 \text{ for } \nu = 1, 4, 7, 10.$$

Ex. 3. If the sum of two independent positive variates is a x^2 variate with $(m+n)$ *d.f.* and if one of them is a χ^2 variate with *m.d.f.* show that the other is a χ^2 variate with *n. d. f.*

Ex. 4. Let x_1, x_2 be independent random Chi-square variates each having 2 *d. f.* then the *m. g. f.* of the variate Y defined by $Y=\frac{1}{2}(x_1-x_2)$ is $(1-t^2)^{-1}$, $-1<t<1$. Deduce that the probability density function of Y is

$$g(y) = \tfrac{1}{2}e^{-|y|}, \quad -\infty < y < \infty$$

Ex. 5. A linear transformation $\xi_i = \sum_{j=1}^{n} c_{ij}\, x_j$, $(i=1, 2, \ldots\ldots, n)$ is said to be orthogonal if

$$\sum_{i=1}^{n} c_{ij}\, c_{ik} = 1 \text{ if } j \neq k$$

$$= 0 \text{ if } j = k.$$

Show that for such a transformation

(*i*) $\sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} x_i^2$.

(*ii*) If $x_1, x_2, \ldots\ldots, x_n$ are independent standard normal variates, so are $\xi_1, \xi_2 \ldots\ldots, \xi_n$.

Ex. 6. Derive the sampling distribution of $\chi^2 = \sum_{i=1}^{n} x_i^2$ by using the integral

$$\int_{-\infty}^{+\infty} \ldots\ldots \int_{\Sigma \chi_i^2 = u} f\ (\Sigma \chi_i^2) d\chi_i^2 = \tfrac{1}{2} c(n)\, u^{\frac{n-2}{2}} f(u) \text{ where } c(n) = (2\pi)^{\frac{1}{2}n} / \sqrt{n}.$$

13·3. Simultaneous Sampling Distribution of the Mean and Variance in Random Samples drawn from a Normal Population. The following geometrical proof is due to Prof. Fisher.

Let $x_1, x_2, \ldots, x_n$ be the n values of the random sample drawn from a normal population with mean μ and variance σ^2. Let $\bar{x}$ and s^2 be the mean and variance of the sample. Then we require the simultaneous distribution of $\bar{x}$ and s^2.

First Step. Joint distribution of the values $x_1, x_2, \ldots\ldots, x_n$.

Since the values $x_1, x_2, \ldots\ldots, x_n$ are independent, the probability that these n values of the sample will fall in the respective intervals $dx_1, dx_2, \ldots\ldots, dx_n$ is, by the theorem of compound probability,

$$dp = \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2} dx_1\, dx_2 \ldots\ldots dx_n. \qquad \ldots(13{\cdot}7)$$

Second step. Transformation of the density function.

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} [(x_i - \bar{x}) + (\bar{x} - \mu)]^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2,$$

the sum of the products vanishes since $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.

Thus $\sum_{i=1}^{n} (x_i - \mu)^2 = ns^2 + n(\bar{x} - \mu)^2$. $\qquad \ldots(13{\cdot}8)$

Third step. Transformation of the volume element.

Let us represent the sample by a point P with co-ordinates $(x_1, x_2, \ldots, x_n)$ in Euclidean hyperspace of n dimensions. Let OE be the unit vector in the n dimensional space. The direction cosines (*d.c.*) of the unit vector are proportional to 1, 1,...,1 so that the co-ordinates of any point on this line are $(\alpha, \alpha, \ldots, \alpha)$.

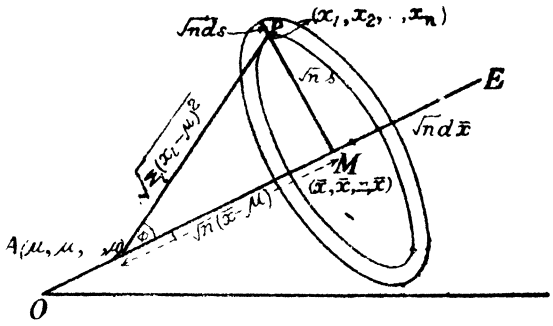


Fig. 34.

Let PM be the perpendicular from P on OE. If M be $(\alpha, \alpha, \ldots, \alpha)$ then since (the *d.c.*'s of PM are proportional to $x_1-\alpha$, $x_2-\alpha$, ..., $x_n-\alpha$ and that of OM are proportional to $\alpha, \alpha, \ldots\ldots, \alpha$ and these are at right angles, we have

$$\alpha(x_1-\alpha)+\alpha(x_2-\alpha)+\ldots\ldots+\alpha(x_n-\alpha)=0.$$

Since $\alpha \neq 0$, this gives $\alpha = \dfrac{x_1+x_2+\ldots+x_n}{n} = \bar{x}$

so that the co-ordinates of M are $(\bar{x}, \bar{x}, \ldots\ldots, \bar{x})$.

Then $OM^2 = \bar{x}^2+\bar{x}^2+\ldots\ldots+\bar{x}^2 = n\bar{x}^2$, *i.e.*, $OM = \bar{x}\sqrt{n}$,

and $PM^2 = (x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots\ldots+(x_n-\bar{x})^2 = ns^2$ *i.e.*, $PM = s\sqrt{n}$.

The probability density factor in (13·7) has been expressed in terms of $\bar{x}$ and s. We are now to find out the transformation of the volume element $dx_1, dx_2, \ldots\ldots, dx_n$ as $\bar{x}$ and s are allowed to take values between the limits $(\bar{x}, \bar{x}+d\bar{x})$ and $(s, s+ds)$ respectively. When $\bar{x}$ and s are kept constant, OM and PM become constant so that the sample point P moves on the surface of a hypersphere of radius $s\sqrt{n}$ and centre at M in the $(n-1)$ dimensional space which is absolutely orthogonal to the line OE. The spherical shell in which P moves has a thickness $d(PM)$ *i.e.*, $\sqrt{n}\,.\,ds$ and suffers a displacement of length $d(OM)$ *i.e.*, $\sqrt{n}\,d\bar{x}$ along the unit vector OE. Therefore, the volume element described by P, as $\bar{x}$ lies in $(\bar{x}, \bar{x}+d\bar{x})$ and s in $(s, s+ds)$, is proportional to $(s\sqrt{n})^{n-2}.\ \sqrt{n}\,ds.\ \sqrt{n}\,d\bar{x}$, that is volume element $=$ const. $s^{n-2}\,ds\,d\bar{x}$.

Fourth step. The required probability, therefore, is

$$dp = \text{const. } e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2}\, e^{-\frac{ns^2}{2\sigma^2}}\, s^{n-2}\, ds d\bar{x}.$$

Since dp is expressible in the form

$$f_1(\bar{x})\, d\bar{x} \,.\, f_2(s)ds$$

we conclude that *the distributions of* $\bar{x}$ *and* s *are independent.*

Alternative Proof. The sampling distributions of sample mean, $\bar{x}$, and sampling variance, s^2, in random samples drawn from a normal population can be easily obtained as follows using the following Dirichlet's integral given in many textbooks on advanced calculus Dirichlet's integral is

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots\cdots\cdots \int_{\substack{\Sigma x_i^2 = u. \\ \Sigma b_i x_i = v}} f(\Sigma x_i^2) F(\Sigma b_i x_i) dx_1 dx_2 \ldots dx_n$$

$$= \frac{1}{2}\, c(n-1) \,.\, f(u) F(v) \left[n - \frac{v^2}{\Sigma b_i^2} \right]^{\frac{1}{2}(n-3)} \left(\Sigma b_i^2 \right)^{-\frac{1}{2}}$$

where $c(n) = 2(\pi)^{\frac{1}{2}n} / \Gamma(\frac{1}{2}n)$.

To obtain the distribution of $\bar{x}$ and s^2, we have

$$\left(\frac{1}{2\pi} \right)^{\frac{1}{2}n} \int_{-\infty}^{+\infty} \cdots\cdots\cdots \int_{\substack{\Sigma x_i^2 = u \\ \frac{1}{n}\Sigma x_i = \bar{x}}} e^{-\frac{1}{2}\Sigma x_i^2} dx_1 dx_2 \ldots\ldots dx_n$$

$$= \left(\frac{1}{2\pi} \right)^{\frac{1}{2}n} \cdot \frac{1}{2}\, c(n-1) \,.\, e^{-\frac{1}{2}u} \left(u - n\bar{x}^2 \right)^{\frac{1}{2}(n-3)} n^{\frac{1}{2}} du dx$$

which is the joint distribution of u and $\bar{x}$.

Put $u = ns^2 + n\bar{x}^2$ so that the joint distribution of ns^2 and $\bar{x}$ is

$$\left(\frac{1}{2\pi} \right)^{\frac{1}{2}n} \cdot \frac{1}{2} c(n-1) \,.\, e^{-\frac{1}{2}ns^2} \left(ns^2 \right)^{\frac{1}{2}(n-3)} n^{\frac{1}{2}} e^{-\frac{1}{2}n\bar{x}^2} ds^2 d\bar{x}$$

Thus the distribution of the sample mean $\bar{x}$ is

$$dp_1 = c_1 e^{-\frac{n}{2}\bar{x}^2} d\bar{x}, \quad -\infty \leqslant \bar{x} < \infty$$

and the distribution of the sample variance s^2 is

$$dp_2 = c_2 e^{-\frac{n}{2}s^2} (s^2)^{\frac{1}{2}(n-3)} ds^2, \quad 0 < s^2 < \infty.$$

Cor 1. The distribution of sample mean $\bar{x}$ is

$$dp_1 = c_1 e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} d\bar{x}, \quad -\infty < \bar{x} < \infty \qquad \ldots(13{\cdot}8)$$

where the constant c_1 is given by

$$c_1 \int_{-\infty}^{\infty} e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} d\bar{x} = 1, \text{ so that } c_1 = \frac{1}{\sigma}\sqrt{\frac{n}{2\pi}}.$$

Thus $\bar{x}$ *is distributed normally with mean* μ *and variance* σ^2/n. In other words, $(\bar{x}-\mu)\sqrt{n}/\sigma$ *is a standard normal variate and hence* $n(\bar{x}-\mu)^2/\sigma^2$ *is a* χ^2*-variate with 1 d.f.*

Cor. 2. The distribution of s^2 is

$$dp_2 = c_2 e^{-\frac{ns^2}{2\sigma^2}} (s^2)^{\frac{n-3}{2}} ds^2, \quad 0 < s^2 < \infty \qquad \ldots(13{\cdot}9)$$

where $\int_0^{\infty} c_2 e^{-\frac{ns^2}{2\sigma^2}} (s^2)^{\frac{n-3}{2}} ds^2 = 1$ so that $c_2 = (n/2\sigma^2)^{(n-1)/2} \div \left|\frac{n-1}{2}\right.$

Writing (13·9) as

$$dp_2 = \frac{1}{2^{\frac{n-1}{2}} \left|\frac{n-1}{2}\right.} e^{-\frac{ns^2}{2\sigma^2}} \left(\frac{ns^2}{\sigma}\right)^{\frac{n-3}{2}} d\left(\frac{ns^2}{\sigma^2}\right). \qquad \ldots(13{\cdot}9)$$

we conclude that ns^2/σ^2 is a χ^2*-variate with* $(n-1)$ *d.f.* Further we observe that $\frac{1}{2}ns^2/\sigma^2$ *is a Gamma variate with parameter* $\frac{1}{2}(n-1)$.
i.e., $\frac{1}{2}ns^2/\sigma^2$ is $\gamma\{\frac{1}{2}(n-1)\}$.

Also
$$E(s^2) = \frac{1}{\left|\frac{n-1}{2}\right.} \left(\frac{n}{2\sigma^2}\right)^{\frac{n-1}{2}} \int_0^{\infty} s^2 . e^{-\frac{ns^2}{2\sigma^2}} (s^2)^{\frac{n-3}{2}} ds^2$$

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

$$\Rightarrow \quad E(s^2) = \frac{n-1}{n}\sigma^2 \quad \therefore E\left(\frac{ns^2}{n-1}\right) = \sigma^2.$$

In other words, $\frac{ns^2}{n-1}\left[= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = S^2\right]$ *is an unbiased estimate of* σ^2 *and is based on* $(n-1)$ *d.f.* This is expected since the n variates $x_i-\bar{x}$ are not all independent but are connected by the linear relation $\sum_{i=1}^{n}(x_i-\bar{x}) = 0$.

[**Definition.** A statistic $t = t(x_1, x_2, \ldots\ldots, x_n)$ is said to be unbiased estimate of the parameter θ if $E(t) = \theta$].

Note. The three main results of this section hold for normal population only and in a sense characterise the normal population. It can be proved that for no other distribution

(*i*) the sample mean and sample variance are independently distributed,

(*ii*) the sample mean is exactly normally distributed.

or (*iii*) the sum of squares of deviations from either the population of sample mean is exactly distributed by the Chi-square distribution.

Ex. 1. In Fig. 33, using $PA^2=PM^2+MA^2$, deduce that

$$\sum_{i=1}^{n}(x_i-\mu^2)=ns^2+n(\bar{x}-\mu)^2$$

Ex 2. Deduce the mean, variance and characteristic function of the Chi-square distribution from the mean, variance and characteristic function of a Gamma variate.

Ex. 3. Show that $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$ is distributed as χ^2 with *n.d.f.* Explain why $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$ is distributed as χ^2 with $(n-1)$ *d.f.*

Ex. 4. Write down the density function for $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ and show that it is maximum when $S^2=\sigma^2$.

Ex. 5. Prove the following result for a random sample $x_1, x_2, \ldots\ldots, x_n$ drawn from a population with finite variance σ^2 :

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\frac{n-1}{n}\sigma^2.$$

Ex. 6. Show that *m.g.f.* of the distribution of s^2 is $\left(1-\frac{2t\sigma^2}{n}\right)^{-\frac{n-1}{2}}$

Hence $\mu_1'=\frac{n-1}{n}\sigma^2$, $\mu'_2=\frac{(n^2-1)}{n^2}\sigma^4$ and $Var\ (s^2)=2\frac{(n-1)^4}{n}\sigma^4$. Deduce for large sample, $Var\ (s^2)=\frac{2\sigma^4}{n}$.

Ex. 7. For what value of k, will $E(ks^2-\sigma^2)$ be minimum ?

13·4. Distribution of Sum and Ratio of two independent Chi-square Variates. *If* χ_1^2 *and* χ_2^2 *are independent* χ^2 *variates with* n_1 *and* n_2 *degrees of freedom respectively, then*

(*a*) $\chi^2=\chi_1^2+\chi_2^2$ *is a* χ^2 *variate with* (n_1+n_2) *d.f. and*

(*b*) $T^2=\frac{\chi_1^2}{\chi_2^2}$ *is a* $\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ *variate*

Proof. The joint probability that a random value χ_1^2 will fall in the interval $d(\chi_1^2)$ and, at the same time, a random value of χ_2^2 will fall in the interval $d(\chi_2^2)$ is, by the theorem of compound probability,

$$dP = \frac{1}{2^{\frac{1}{2}(n_1+n_2)}\overline{|n_1/2}\ \overline{|n_2/2}} \cdot e^{-\frac{1}{2}(\chi_1^2+\chi_2^2)} (\chi_1^2)^{\frac{1}{2}(n_1-2)} (\chi_2^2)^{\frac{1}{2}(n_2-2)} d(\chi_1^2) d(\chi_2^2) \quad \ldots(13{\cdot}9)$$

In other words, the probability that χ_1 and χ_2 (the positive square roots of χ_1^2 and χ_2^2) will fall simultaneously in the intervals $d\chi_1$ and $d\chi_2$ respectively is

$$dP = \frac{2}{2^{(n_1+n_2)/2}\overline{|n_1/2}\ \overline{|n_2/2}} e^{-\frac{1}{2}(\chi_1^2+\chi_2^2)} \chi_1^{n_1-1} \chi_1^{n_2-2} d\chi_1 d\chi_2.$$

Let us introduce the new variates χ and θ given by

$$\chi_1 = \chi \sin\theta,\ \chi_2 = \chi\cos\theta \ i.e.,\ \chi = \sqrt{\chi_1^2+\chi_2^2},\quad \theta = \tan^{-1}\left(\frac{\chi_1}{\chi_2}\right)$$

so that $$d\chi_1 d\chi_2 = \left|\frac{\partial(\chi_1, \chi_2)}{\partial(\chi, \theta)}\right| d\chi d\theta = \chi d\chi d\theta,$$

and as χ_1^2 and χ_2^2 range from 0 to ∞, χ^2 ranges from 0 to ∞ and θ from 0 to $\pi/2$.

Hence the probability that simultaneously χ will fall in the interval $d\chi$ and θ in the interval $d\theta$ is

$$dp = \frac{4}{2^{\frac{n_1+n_2}{2}}\left|\frac{n_1}{2}\right|\frac{n_2}{2}} e^{-\frac{1}{2}\chi^2} (\chi\sin\theta)^{n_1-1} (\chi\cos\theta)^{n_2-1} \chi d\chi d\theta$$

$$= \frac{1}{2^{\frac{n_1+n_2}{2}}\left|\frac{n_1+n_2}{2}\right.} e^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{1}{2}(n_1+n_2-2)} d(\chi^2) \times \frac{2}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} (\sin\theta)^{n_1-1} (\cos\theta)^{n_2-1} d\theta.$$

Thus it follows that

(*i*) the variates χ^2 and θ are independently distributed ;

(*ii*) $\chi^2 = \chi_1^2 + \chi_2^2$ is distributed like χ^2 with (n_1+n_2) degrees of freedom.

The variate θ is distributed as

$$dp_1 = \frac{2}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \sin^{n_1-1}\theta \cos^{n_2-1}\theta\, d\theta,\ 0 \leqslant \theta \leqslant \pi/2 \quad \ldots(13{\cdot}9')$$

Now $\theta=\tan^{-1}\left(\frac{\chi_1}{\chi_2}\right)=\tan^{-1}T$, so that $d\theta=\frac{dT}{1+T^2}$.

$\therefore \sin\theta=\frac{T}{\sqrt{1+T^2}}$, $\cos\theta=\frac{1}{\sqrt{1+T^2}}$.

As θ ranges from 0 to $\pi/2$ T ranges from 0 to ∞ and the variate T is distributed as

$$dp_2=\frac{2}{B\left(\frac{n_1}{2},\ \frac{n_2}{2}\right)}\cdot\frac{T^{n_1-1}}{(1+T^2)^{\frac{1}{2}(n_1+n_2)}}dT$$

$$=\frac{1}{B\left(\frac{n_1}{2},\ \frac{n_2}{2}\right)}\frac{(T^2)^{\frac{n_1}{2}-1}}{(1+T^2)^{\frac{1}{2}(n_1+n_2)}}d(T^2),\ 0<T^2<\infty \qquad \ldots(13\cdot10)$$

Hence T^2 is a $\beta_2\left(\frac{n_1}{2},\ \frac{n_2}{2}\right)$ variate.

Example 13·1. *Show that if* χ_1^2 *and* χ_2^2 *are independently distributed as* χ^2 *with* n_1 *and* n_2 *d.f. respectively, then*

$$u=\frac{\chi_1^2}{\chi_1^2+\chi_2^2} \text{ is a } \beta_1\left(\frac{n_1}{2},\ \frac{n_2}{2}\right) \text{ variate.}$$

First Method. $u=\frac{\chi_1^2}{\chi_1^2+\chi_2^2}=1\Big/\left(1+\frac{\chi_1^2}{\chi_2^2}\right)=\frac{1}{1+1/T^2}$

so that $1/T^2=\frac{1}{u}-1=\frac{1-u}{u}$, $T^2=\frac{u}{1-u}$ and $d(T^2)=\frac{du}{(1-u)^2}$.

The range of T^2 is from 0 to ∞ and of u is from 0 to 1. The probability differential of T^2 is (13·10). Consequently the probability differential of u is

$$dp=\frac{1}{B\left(\frac{n_1}{2},\ \frac{n_2}{2}\right)}\left(\frac{u}{1-u}\right)^{\frac{n_1}{2}-1}(1-u)^{\frac{n_1+n_2}{2}}\frac{du}{(1-u)^2}$$

$$=\frac{u^{\frac{n_1}{2}-1}(1-u)^{\frac{n_2}{2}-1}}{B\left(\frac{n_1}{2},\ \frac{n_2}{2}\right)}du,\ 0\leqslant u\leqslant 1.$$

Hence u is a $\beta_1\left(\frac{n_1}{2},\ \frac{n_2}{2}\right)$ variate.

Second Method. We require the distribution of

$$u=\frac{\chi_1^2}{\chi_1^2+\chi_2^2}=\sin^2\theta.$$

Now $\theta=\sin^{-1}\sqrt{u}$, where $\chi_1=\chi\sin\theta$ and $\chi_2=\chi\cos\theta$ and by (13·9″) distribution of θ is

$$dp=\frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}(\sin\theta)^{n_1-1}(\cos\theta)^{n_2-1}\,d\theta$$

We easily find that the probability differential of u is

$$dp=\frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}u^{n_1/2-1}(1-u)^{n_2/2-1}du,\quad 0<u<1$$

so that u is a $\beta_1\left(\frac{n_1}{2},\frac{n_2}{2}\right)$ variate.

Example 13·2. *Show that if X and Y are independent normal variates with means m_1, m_2 and variances σ_1^2, σ_2^2 respectively, then quotient*

$$Z=\frac{X-m_1}{Y-m_2}$$

follows Cauchy's distribution

$$dp=\frac{\sigma_1\sigma_2\,dz}{\pi(\sigma_1^2+\sigma_2^2z^2)},\quad -\infty<z<\infty.$$

(M.A., Madras, 1960 ; M.A., Bombay, 1952, 58 ; M.A., Statistics, Delhi 1958, 60, 65 ; M.A., Punjab, 1959 ; M.A., Delhi 1959 ; M.A., Agra, 1967 ; I.C.A.R., Delhi, 1947)

We have

$$\frac{\sigma_2^2z^2}{\sigma_1^2}=\left(\frac{x-m_1}{\sigma_1}\right)^2\Big/\left(\frac{y-m_2}{\sigma_2}\right)^2=\frac{u}{\nu}$$

where, by Cor. 1 page 370, $u=\left(\frac{x-m_1}{\sigma_1}\right)^2$ is a χ^2 variate with 1 *d.f.* and so is ν and hence $\sigma_2^2z^2/\sigma_1^2$ is the quotient of two independent χ^2 variates. Therefore, by § 13·4, $\sigma_2^2z^2/\sigma_1^2$ is a $\beta_2\left(\frac{1}{2},\frac{1}{2}\right)$ variate with probability differential

$$dp=\frac{\left(\frac{\sigma_2^2}{\sigma_1^2}\right)^{\frac{1}{2}-1}d\left(\frac{\sigma_2^2}{\sigma_1^2}z^2\right)}{B\left(\frac{1}{2},\frac{1}{2}\right)\left(1+\frac{\sigma_2^2}{\sigma_1^2}z^2\right)}=\frac{\sigma_1\sigma_2\,d(z^2)}{\pi(\sigma_1^2+\sigma_2^2z^2)\sqrt{z^2}}$$

The range of z^2 is from 0 to ∞ and of z is from $-\infty$ to $+\infty$, the probability differential of z is

$$dp=\frac{\sigma_1\sigma_2\,dz}{\pi(\sigma_1^2+\sigma_2^2\,z^2)},\quad -\infty<z<\infty,$$

the factor 2 disappearing since the integral from $-\infty$ to ∞ must be equal to unity.

Ex. 1. Putting $\frac{\chi_1^2}{\chi_2^2}=y$, $\chi_1^2+\chi_2^2=z$ in (13·9′) and using

$$J(\chi_1^2,\chi_2^2)=\frac{z}{(1+y)^2},$$

show that the joint probability density of y and z is given by

$$\frac{1}{2\left|\frac{n_1}{2}\right|\frac{n_2}{2}}\frac{y^{\frac{n_1-1}{2}}}{(1+y)^{\frac{1}{2}(n_1+n_2)}}\left(\frac{z}{2}\right)^{\frac{(n_1+n_2-2)}{2}}e^{-\frac{z}{2}},\quad 0\leqslant z\leqslant\infty,\ 0\leqslant y<\infty$$

Hence show that the density function of y is

$$\frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\cdot\frac{y^{\frac{n_1-2}{2}}}{(1+y)^{\frac{1}{2}(n_1+n_2)}}$$

and that of z is

$$e^{-\frac{z}{2}}z^{\frac{n_1+n_2-2}{2}}\Big/\,2^{\frac{1}{2}(n_1+n_2)}\left|\tfrac{1}{2}(n_1+n_2)\right.$$

Ex. 2. Deduce, from Ex. 1, the distribution of the statistic F defined by

$$F=\frac{n_2}{n_1}\,\frac{\chi_1^2}{\chi_2^2}.$$

[Hint. Put $y=\frac{n_1}{n_2}F$ in the distribution of y.]

13·5. *Student's t Distribution. Let $x_1, x_2,\ldots\ldots, x_n$ be the members of a random sample drawn from a normal population with mean μ and variance σ^2.

---

*The distribution of the statistic t was first found by W.S. Gosset who wrote under the *nom de plume* of 'Student'. His paper entitled 'the probable error of the mean', published in 1908 ushered the modern era of Statistics. Student had found the distribution of $(\bar{x}-\mu)/S$. The correct definition of t and its distribution were rigorously established by Fisher in 1926.

The t distribution is independent of population parameters μ and σ^2. The definition of t is independent of σ^2 and, therefore, it will not be necessary to know σ^2 before using this distribution. This is the chief merit of the t distribution. Before its discovery it was usual to replace the unknown σ by S in

$$t=\frac{x-\mu}{\sigma/\sqrt{n}}\ \text{to get}\ t=\frac{\bar{x}-\mu}{S/\sqrt{n}}$$

and still to regard t as normally distributed. While for large samples the distributions of t is asymptotically normal it is far from normal in small samples. The discovery of its distribution supplied an exact distribution applicable to small samples and in fact to samples of all sizes.

Then we define the statistic t as

$$t=\frac{(\bar{x}-\mu)\sqrt{n}}{S} \qquad \ldots(13\cdot11)$$

where $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ *i.e.*, $(n-1)S^2=\sum_{i=1}^{n}(x_i-\bar{x})^2=ns^2$,

Hence $\frac{t^2}{\nu}=\frac{(x-\mu)^2\div\frac{\sigma^2}{n}}{\frac{ns^2}{\sigma^2}}$ where $\nu=n-1$.

We proved in § 13·3 that $n(\bar{x}-\mu)^2/\sigma^2$ is distributed as a χ^2 variate with 1 *d.f.* and ns^2/σ^2 is a χ^2 variate with $(n-1)$ *d.f.*

Thus t^2/ν is the ratio of two independent χ^2 variates, distributed with 1 and ν *d.f.* respectively and, therefore, by § 13·4, t^2/ν is a $\beta_2[\frac{1}{2}, \nu/2]$ variate and its distribution is

$$dp=\frac{\left(\frac{t^2}{\nu}\right)^{-\frac{1}{2}}\,d\left(\frac{t^2}{\nu}\right)}{B\left(\frac{1}{2},\frac{\nu}{2}\right)\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}},\quad 0<t^2<\infty$$

or

$$dP=\frac{dt}{\sqrt{\nu}\,.\,B\left(\frac{1}{2},\frac{\nu}{2}\right)\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}},\quad -\infty<t<\infty \qquad \ldots(13\cdot12)$$

the factor 2 disappearing since the integral $-\infty$ to $+\infty$ must be equal to unity. This distribution is known as the t distribution with ν *d.f.*

Alternative Proof. We can deduce the t distribution from the distributions of $\bar{x}$ and s as follows :

The probability differential for $\bar{x}$ is

$$dp=\frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\,e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2}\,d\bar{x}$$

We require the probability differential of

$$t=\frac{(\bar{x}-\mu)\sqrt{n}}{S}=\frac{(\bar{x}-\mu)\sqrt{\nu}}{s}$$

where $\nu S^2=\sum_{i=1}^{n}(x_i-\bar{x})^2=ns^2$, $\nu=n-1$

Hence, for a fixed value of s in the interval ds, the probability that t will fall in the interval dt is

$$dp=\frac{\sqrt{n}}{\nu\sqrt{2\pi}}\ e^{-\frac{ns^2t^2}{2\nu\sigma^2}}\ s\ dt.$$

The probability that a random value of s will fall in the interval ds is

$$dp'=2\left(\frac{n}{2\sigma^2}\right)^{\frac{\nu}{2}}\frac{e^{-\frac{ns^2}{2\sigma^2}}}{\left|\frac{\nu}{2}\right.}\ s^{\nu-2}\ ds,\quad 0<s<\infty$$

The probability that simultaneously t will lie in the interval dt and s in the interval ds is the product $dp\,.\,dp'$. Integrating this product for s from 0 to ∞ we get the probability that t will lie in the interval dt as

$$dP=\frac{2\sqrt{n}}{\sigma\sqrt{2\pi}\left|\frac{\nu}{2}\right.}\cdot\frac{n^{\nu/2}}{(2\sigma^2)^{\nu/2}}\,dt\times\int_0^\infty e^{-\frac{ns^2}{2\sigma^2}\left(1+\frac{t^2}{\nu}\right)}\ s^{n-2}\ ds$$

$$=\frac{dt}{\sqrt{\nu}B\left(\frac{1}{2},\ \frac{\nu}{2}\right)\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}.$$

Ex. 1. Show that, in Fig. 33, $t=\sqrt{n-1}\cot\phi$ so that t is constant over the cone obtained by rotating AP about the unit vector OE.

Show that the distribution of Φ is given by

$$dF=\text{Const. }\sin^{n-2}\phi\,d\phi,\ 0\leqslant\phi\leqslant\pi,$$

Hence deduce the distribution of t.

Ex. 2. Show that in both the cases discussed above, t can be regarded as the ratio of a statistic distributed normally with zero means to an unbiased estimate of the variance of that statistic in random sampling.

Ex. 3. If $I_x(p, q)$ represents the incomplete Beta function defined by

$$I_x(p,q)=\frac{1}{B(p,q)}\int_0^x t^{p-1}(1-t)^{q-1}\,dt,\quad p>0,\ q>0$$

show that the distribution $P(t)$, given by (13·12), is given by

$$P(t)=1-\tfrac{1}{2}I_x\left(\frac{n}{2},\ \frac{1}{2}\right),\ \text{where}\ x=1\Big/\left(1+\frac{1}{\nu}t^2\right),$$

(M.A., Poona, 195)

13·5·1. Chief Features of the t Probability Curve.

The equation to the t probability curve is

$$y=\frac{1}{\sqrt{\nu}B\left(\frac{1}{2},\frac{\nu}{2}\right)}\cdot\frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}.$$

The curve is symmetrical about the line $t=0$ since only even powers of t appear in the equation of the curve. Further since $\frac{1}{1+t^2/\nu}$ decreases rapidly as $|t|$ increases, the t-curve will tail off to zero on either side of the origin. In other words, the curve is asymptotic to the t-axis at each end. The curve has a maximum ordinate at $t=0$.

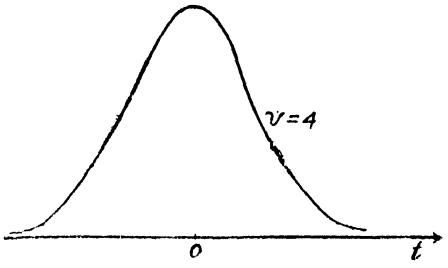


Fig. 34. t Probability Curve.

The probability that the value of t from a random sample will be between two fixed values t_1 and t_2 is given by integrating y with respect to t from t_1 to t_2. Similarly, if P be the probability that the value of $|t|$ from a random will exceed t_0 then we have

$$P=2\int_{t_0}^{\infty}\frac{1}{\sqrt{\nu}B\left(\frac{1}{2},\frac{\nu}{2}\right)}\cdot\frac{dt}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}.$$

Values of t_0 have been tabulated for various fixed values of P for values of ν from 1 to 60. For values of $\nu>60$, the fact t is asymptotically a standard normal variate can be used.

13·5·2. Some Properties of t Distribution. All the moments of odd order about the origin vanish. The moment of order $2r$ about the origin (which is also the mean) is

$$\mu_{2r}'=2\int_0^{\infty}\frac{t^{2r}}{\sqrt{\nu}B\left(\frac{1}{2},\frac{\nu}{2}\right)}\cdot\frac{dt}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}$$

$$=\int_0^{\infty}\frac{t^{2r}\left(\frac{t^2}{\nu}\right)^{-\frac{1}{2}}d\left(\frac{t^2}{\nu}\right)}{B\left(\frac{1}{2},\frac{\nu}{2}\right)\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}$$

Table 4

Values of mod. t with a probability of being exceeded in random sampling

ν=number of degrees of freedom

ν \ P	0·50	0·10	0·05	0·02	0·01
1	1·000	6·34	12·71	31·82	63·66
2	0·816	2·92	4·30	6·96	9·92
3	0·765	2·35	3·18	4·54	5·84
4	0·741	2·13	2·78	3·75	4·60
5	0·727	2·02	2·57	3·36	4·03
6	0·718	1·94	2·45	3·14	3·71
7	0·711	1·90	2·36	3·00	3·50
8	0·706	1·86	2·31	2·90	3·36
9	0·703	1·83	2·26	2·82	3·25
10	0·700	1·81	2·23	2·76	3·17
11	0·697	1·80	2·20	2·72	3·11
12	0·695	1·78	2·18	2·68	3·06
13	0·694	1·77	2·16	2·65	3·01
14	0·692	1·76	2·14	2·62	2·98
15	0·691	1·75	2·13	2·60	2·95
16	0·690	1·75	2·12	2·58	2·92
17	0·689	1·74	2·11	2·57	2·90
18	0·688	1·73	2·10	2·55	2·88
19	0·688	1·73	2·09	2·54	2·86
20	0·687	1·72	2·09	2·53	2·84
21	0·686	1·72	2·08	2·52	2·83
22	0·686	1·72	2·07	2·51	2·82
23	0·685	1·71	2·07	2·50	2·81
24	0·685	1·71	2·06	2·49	2·80
25	0·684	1·71	2·06	2·48	2·79
26	0·684	1·71	2·06	2·48	2·78
27	0·684	1·70	2·05	2·47	2·77
28	0·683	1·70	2·05	2·47	2·76
29	0·683	1·70	2·04	2·46	2·76
30	0·683	1·70	2·04	2·46	2·75
35	0·682	1·69	2·03	2·44	2·72
40	0·681	1·68	2·02	2·42	2·71
45	0·680	1·68	2·02	2·41	2·69
50	0·679	1·68	2·01	2·40	2·68
60	0·678	1·67	2·00	2·39	2·66
∞	0·674	1·64	1·96	2·33	2·58

Reproduced by permission of the author. Professor R.A. Fisher, from his book on *Statistical Methods for Research Worker.*

$$\mu_{2r}' = \frac{\nu^r}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \int_0^\infty \frac{\left(\frac{t^2}{\nu}\right)^{r+\frac{1}{2}-1}}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}} \, d\left(\frac{t^2}{\nu}\right)$$

$$= \frac{\nu^r}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} B\left(r+\frac{1}{2}, \frac{\nu}{2}-r\right), \quad r<\frac{\nu}{2}.$$

$$\left[\text{On letting } 1+\frac{t^2}{\nu}=\frac{1}{y} \text{ i.e., } \frac{t^2}{\nu}=\frac{1-y}{y}\right]$$

Hence $\mu_{2r}=\dfrac{(2r-1)(2r-3).....1}{(\nu-2)(\nu-4).....(\nu-2r)}\nu^r$

Ex. 1. Show that, as $\nu\to\infty$

$$\frac{1}{\sqrt{\pi}\sqrt{\nu}}\cdot\frac{\left|\overline{\frac{\nu+1}{2}}\right.}{\left|\overline{\frac{\nu}{2}}\right.}\frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}\to\frac{1}{\sqrt{2\pi}}\quad e^{-\frac{1}{2}t^2}$$

Hence deduce that the t distribution approaches normality as the number of degrees of freedom increases indefinitely.

(M.A. Bombay, 1950 ; M.A. Poona, 1959 ; M.A. Delhi, 1959)

[**Hint.** Use $x\,!=\sqrt{2\pi}\;\;e^{-x}\;e^{x+\frac{1}{2}}$]

Ex. 2. Show that standardised t variate is $t\sqrt{(\nu-2)/\nu}$ and that its distribution also approaches normality as the number of degrees of freedom increases.

Ex. 3. From the table, find t_0 such that

$$P[-t_0\leqslant t\leqslant t_0]=\cdot 95$$

for $\nu=5$, 10, 15, 20, 25 and 30 and show that this illustrates the approach of the distribution to normality.

Ex. 4. Show that for $\nu=1$ the student's distribution reduces to Cauchy distribution.

13·6. Distribution of t for Comparison of two Sample Means. Let $x_{1i}(i=1, 2,....., n_1)$ and $x_{2j}(j=1, 2,....., n_2)$ be two independent random samples drawn from two normal populations with mean μ and μ' respectively but with *same* variance σ^2. Let $\bar{x}_1$, $\bar{x}_2$ be sample means and let

$$S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 \text{ and } S_2^2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2$$

Then we define the statistic t by the equation

$$t=\frac{(\bar{x}_1-\bar{x}_2)-(\mu-\mu')}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad ...(13·13)$$

where

$$S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

$$=\frac{\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2+\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2}{n_1+n_2-2} \qquad ...(13·14)$$

and is called the pooled estimate of the population variance.

Hence if $\nu=n_1+n_2-2$ we have, from (13·13) and (13·14)

$$\frac{t^2}{\nu}=\frac{[(\bar{x}_1-\mu)-(\bar{x}_2-\mu')]^2/\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}{[(n_1-1)S_1^2+(n_2-1)S_2^2]/\sigma^2} \qquad ...(13·15)$$

Since sample have been drawn from normal populations with means μ and μ' respectively and same variance σ^2, $\bar{x}_1-\mu$ and $\bar{x}_2-\mu'$ are distributed normally about mean zero with variance σ^2/n_1 and σ^2/n_2 respectivery. Further, since the samples are independent, the difference $(\bar{x}_1-\mu)-(\bar{x}_2-\mu')$ is distributed normally about zero mean with variance $\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$. Hence the numerator of second member of (13·15) is a χ^2 variate with 1 *d. f.* We have already proved in § 13·3, Cor. 2, page 377 that $(n_1-1)\ S_1^2/\sigma^2$ and $(n_2-1)\ S_2^2/\sigma^2$ are Chi-square variates with (n_1-1) and (n_2-1) *d. f.* respectively and, therefore, by virtue of the additive property of χ^2, the sum

$$\frac{1}{\sigma^2}[(n_1-1)S_1^2+(n_2-1)S_2^2]$$

is a χ^2 variate with $\nu(=n_1+n_2-2)$ *d. f.* Also

$$E(S^2)=E\left[\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right]$$

$$=\frac{(n_1-1)\sigma^2+(n_2-1)\sigma^2}{n_1+n_2-2}=\sigma^2 \qquad \ldots(13\cdot16)$$

Thus the second member of (13·15) is the ratio of two independent χ^2 variates with 1 and ν *d. f.* respectively. Therefore, by § 13·4, $\frac{t^2}{\nu}$ is a $\beta_2\left(\frac{1}{2}, \frac{\nu}{2}\right)$ variate so that t defined by (13·13) conforms to the t distribution with *v.d. f.*

Ex. 1. Can we use $S'^2=\left[\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2+\sum_{j=1}^{n_2}(x_{2j}-\bar{x}^2)\right]\Big/(n_1+n_2-1)$ instead of S^2 where

$$\bar{x}=\frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}?$$

[Ans. No]

Ex. 2. For three independent samples of sizes 5, 6, 7, the sample variances are 16, 25, 36. Find an unbiased estimate of σ^2 by pooling the three samples.

13·7. *Snedecor's *F* Distribution. *Distribution of the ratio of independent estimates of the population variance.*

Let $x_{1i}(i=1, 2,\ldots\ldots,n_1)$ and $x_{2j}\ (j=1, 2,\ldots\ldots,n_2)$ be the values of two independent random samples drawn from the same normal population with variance σ^2. Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means

*Prof. R.A. Fisher had found the distribution of $z=\frac{1}{2}\log_e F$ and published (in 1925) tables giving the values of z that will be exceeded with probability 0·05 and 0·01 respectively corresponding to specified values of *d. f.* From these, Prof. G.W. Snedecor (1934) prepared a table for the variance ratio F and called it 'F' in honour of Fisher.

and let $S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2$ and $S_2^2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2$.

Then we define the statistic F by the relation

$$F=\frac{S_1^2}{S_2^2}. \qquad ...(13\cdot17)$$

Hence $\frac{\nu_1 F}{\nu_2}=\frac{(n_1-1)S_1^2/\sigma^2}{(n_2-1)S_2^2/\sigma^2}$, where $\nu_1=n_1-1$, $\nu_2=n_2-1$.

The numerator and denominator of the second member are independent χ^2 variates with ν_1 and ν_2 *d. f.* respectively. Hence $\nu_1 F/\nu_2$ is a $\beta_2(\nu_1/2, \nu_2/2)$ variate so that the probability that a random value of F will fall in the interval dF is

$$dp=\frac{\nu_1^{\frac{\nu_1}{2}}\nu_2^{\frac{\nu_2}{2}}F^{\frac{\nu_1-2}{2}}dF}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)(\nu_1F+\nu_2)^{\frac{1}{2}(\nu_1+\nu_2)}},\ 0<F<\infty \qquad ...(13\cdot18)$$

This distribution is called the distribution of the variance ratio F with ν_1 and ν_2 *d. f.*

Ex. 1. If $\nu_1=\nu_2=n-1$, show that the distribution of the $F'=1/F$ is the same as the distribution of F with $n-1$ and $n-1$ *d. f.* and that the curve $F'=1/F$ has to be traced from right to the left. Explain the practical utility of this.

(B.A., Vikram, 1961)

Ex. 2. If two independent samples are drawn from normal populations with variances σ_1^2 and σ_2^2 respectively show that the distribution of $F=S_1^2/S_2^2$ is

$$dp=\frac{\sigma_1^{\nu_1}\sigma_2^{\nu_2}\nu_1^{\frac{\nu_1}{2}}\nu_2^{\frac{\nu_2}{2}}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\cdot\frac{F^{\frac{\nu_1-2}{2}}}{\left(\frac{\nu_1}{\sigma_1^2}F+\frac{\nu_2}{\sigma_2^2}\right)^{\frac{1}{2}(\nu_1+\nu_2)}}dF.$$

Ex. 3. Prove that $u=\frac{\frac{\nu_1}{\nu_2}F}{1+\frac{\nu_1}{\nu_2}F}$ is a $\beta_1\left(\frac{\nu_1}{2}-1, \frac{\nu_2}{2}-1\right)$ variate.

Hence show that the tables of incomplete Beta distribution can be used to tabulate the distribution function of F.

13·7·1. Some Properties of the F Distribution and the F Curve.

The rth moment about the origin is given by

$$\mu_r'=E(F^r)=\frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\int_0^\infty F^{r+\frac{\nu_1}{2}-1}\left(1+\frac{\nu_1}{\nu_2}F\right)^{-\frac{\nu_1+\nu_2}{2}}dF$$

$$= \frac{\left(\frac{\nu_1}{\nu_2}\right)^{-r}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^1 y^{\frac{\nu_2}{2}-r-1} (1-y)^{r+\frac{\nu_1}{2}-1} dy,$$

where $1+1\frac{\nu_1}{\nu_2} F = \frac{1}{y}$ and $dF = -\frac{\nu_2}{\nu_1} \cdot \frac{dy}{y^2}$.

$$\therefore \quad \mu_r' = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{-r}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{B\left(\frac{\nu_1}{2}+r, \frac{\nu_2}{2}-r\right)}{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}+r}}, \quad r < \frac{\nu_2}{2}$$

or $$\mu_r' = \frac{\left|\overline{\frac{\nu_1}{2}+r}\right| \left|\overline{\frac{\nu_2}{2}-r}\right|}{\left|\overline{\frac{\nu_1}{2}}\right| \left|\overline{\frac{\nu_2}{2}}\right|} \left(\frac{\nu_2}{\nu_1}\right)^r, \quad r < \frac{\nu_2}{2}.$$

In particular,

$$\mu_1' = \frac{\left|\overline{\frac{\nu_1}{2}+1}\right| \left|\overline{\frac{\nu_2}{2}-1}\right|}{\left|\overline{\frac{\nu_1}{2}}\right| \left|\overline{\frac{\nu_2}{2}}\right|} \left(\frac{\nu_2}{\nu_1}\right) = \frac{\nu_2}{\nu_1-2}.$$

which is *independent* of ν_1 and is always greater than unity.

$$\mu_2' = \frac{\left|\overline{\frac{\nu_1}{2}+2}\right| \left|\overline{\frac{\nu_2}{2}-2}\right|}{\left|\overline{\frac{\nu_1}{2}}\right| \left|\overline{\frac{\nu_2}{2}}\right|} \left(\frac{\nu_2}{\nu_1}\right)^2 = \frac{(\nu_1+2)\nu_2^2}{(\nu_2-2)(\nu_2-4)\nu_1}.$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \sim 2\left(\frac{1}{\nu_1}+\frac{1}{\nu_2}\right).$$

The distribution of F is independent of the population variance σ^2 and depends on ν_1 and ν_2 only. The F curve is J shaped, if $\nu_2 \leqslant 2$ and bell shaped for $\nu_1 > 2$. For $\nu_1 > 4$ the shape of the F curve is as shown in Fig. 35.

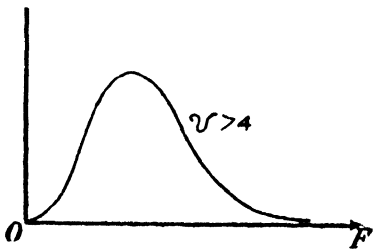


Fig. 35. F Probability Curve.

Ex. 1. For $v_1=v_2=n-1$, prove that (*i*) A.M.$=$H.M.$=\frac{n-1}{n-3}$, (*ii*) The median is at $F=1$. (*iii*) $Q_1Q_3=1$, (*iv*) $D_1D_3=1$.

Ex. 2. Find Pearson's measure of skewness *viz.*, $\frac{\text{mean}-\text{mode}}{\text{standard deviation}}$ for the F distribution and show that it is always positively skew.

Also show that for $v_1=v_2=n-1$, $\frac{\text{mean}-\text{mode}}{\text{mean}-\text{median}} \to 3$ as $n \to \infty$.

Ex. 3. If $F(v_1, v_2)$ represents a F variate with v_1, v_2 *d.f.* prove that $F(v_2, v_1)$ is distributed as $1/F(v_1, v_2)$ variate. Deduce that if $c>0$, then

$$P\left[F(v_1, v_2) \geqslant c\right] = P\left[F(v_2, v_1) \leqslant \frac{1}{c}\right]$$

Show how you can make use of this property to deduce the values of F which will be exceeded with probability ·95 and ·99 from the table values of F exceeded with probability ·05 and ·01.

Ex. 4. Show how the probability points of $F(v_2, v_1)$ can be obtained from those of $F(v_1, v_2)$. (*M.A. Agra, 1967 ; M.A. Delhi, 1960*)

Ex. 5. If F has a F distribution with $n_1=5$, $n_2=10$, find a and b so that $P(F \leqslant a)=·05$ and $P(F \leqslant b)=·95$ and accordingly $P(a<F<b)=90$.

[Hint. If $P(F \leqslant a)=P\left(\frac{1}{F} \geqslant \frac{1}{a}\right) = 1-P\left(\frac{1}{F} \leqslant \frac{1}{a}\right)$

$=1-P[F(n_2, n_1) \leqslant \frac{1}{a}] \Rightarrow P[F(n_2, n_1) \geqslant \frac{1}{a}]=·05$]

[Ans. $a=\frac{1}{4·74}$, $b=3·33$]

Ex. 6. Prove that if x has the F-distribution with (m, n) *d.f.* and y has the F-distribution with (n, m) *d.f.*, then for every $a>0$

$$P(x \leqslant a)+P\left(y \leqslant \frac{1}{a}\right)=1. \qquad (B.Sc., Bombay, 1968)$$

Example 13·3. (*a*) *Let* χ_1^2, χ_2^2 *be independently distributed variates each having a chi-square distribution with* n_1 *and* n_2, *d.f. Derive the distribution of* $F=\frac{\chi_1^2/n_1}{\chi_2^2/n_2}$.

(*I.S.I., Calcutta, 1953, 56 ; M.A. Delhi, 1954*)

(*c*) *If* s_1^2, s_2^2 *are the sample variances based on n independent observations each from two normal populations with variances* σ_1^2, σ_2^2; *calculate 90% confidence interval for* σ_1^2/σ_2^2 *in terms of c where the probability of F exceeding c based on* $n-1$, $n-1$ *d.f. is 0·5.*

(*M.A., Delhi, 1954*)

(*a*) The joint distribution of χ_1^2 and χ_2^2 is

$$\frac{1}{2^{\frac{n_1}{2}}\left|\frac{n_1}{2}\right.} e^{-\frac{1}{2}\chi_1^2} (\chi_1^2)^{\frac{n_1}{2}-1} \, d\chi_1^2 \times \frac{1}{2^{\frac{n_2}{2}}\left|\frac{n_2}{2}\right.} e^{-\frac{1}{2}\chi_2^2} (\chi_2^2)^{\frac{n_2}{2}-1} \, d\chi_2^2$$

Let us make the change of variates

$$F=\frac{\chi_1^2/n_1}{\chi_2^2/n_2},\ v=\chi_2^2,\ 0\leqslant F<\infty,\ 0\leqslant v<\infty$$

Then $\chi_1^2=\frac{vn_1}{n_2}F,\ \chi_2^2=v.$

The Jacobian of transformation is

$$J=\begin{vmatrix} \frac{n_1 v}{n_2} & \frac{n_1}{n_1}F \\ 0 & 1 \end{vmatrix}=\frac{n_1}{n_2}v$$

The distribution of the transformed variates is

$$\frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{2^{\frac{1}{2}(n_1+n_2)}\left|\frac{n_1}{2}\right|\frac{n_2}{2}}F^{\frac{n_1}{2}-1}\ v^{\frac{n_1+n_2}{2}-1}\ e^{-\frac{1}{2}\left(\frac{n_1}{n_2}F+1\right)v}\ dF\,dv$$

Integrating out for variable v throughout the whole range, we get the distribution of F as

$$dP=\frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{\left|\frac{n_1}{2}\right|\frac{n_2}{2}}F^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}F\right)^{-\frac{n_1+n_2}{2}}dF$$

$$\times\int_0^\infty\left[\left(1+\frac{n_1}{n_2}F\right)\left(\frac{v}{2}\right)\right]^{\frac{n_1+n_2}{2}-1}e^{-\frac{v}{2}\left(1+\frac{n_1}{n_2}F\right)}dv$$

$$=\frac{\left|\frac{n_1+n_2}{2}\right.}{\left|\frac{n_1}{2}\right|\frac{n_2}{2}}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}F^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}F\right)^{-\frac{n_1+n_2}{2}}dF,\ 0<F<\infty$$

This is F distribution with n_1 and n_2 $d.f.$

We have $F=\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\times\frac{n-1}{n-1}$. ($\because\ n_1=n_2=n$)

$$\therefore\quad \frac{1}{F}=\frac{s_2^2/\sigma_2^2}{s_1^2/\sigma_1^2}\times\frac{n-1}{n-1}.$$

By Ex. 1, page 389, the distribution of F and $1/F$ are same, each is based on $n-1$, $n-1$ $d.f.$

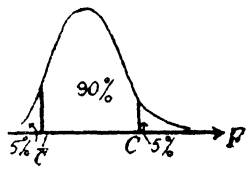


Fig. 36.

Now $P(F>c)=\cdot 05$ (given) ...(1)

$$\therefore \qquad P\left(\frac{1}{F} < \frac{1}{c}\right) = \cdot 05 \qquad \ldots(2)$$

or $$P\left(F < \frac{1}{c}\right) = \cdot 05 \qquad \ldots(3)$$

Combining (1) and (3), we have

$$P\left[\frac{1}{c} \leqslant F \leqslant c\right] = \cdot 90$$

or $$P\left[\frac{1}{c} \leqslant \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \leqslant c\right] = \cdot 90$$

or $$P\left[\frac{1}{c}\frac{s_2^2}{s_1^2} \leqslant \frac{\sigma_2^2}{\sigma_1^2} \leqslant c\frac{s_1^2}{s_2^2}\right] = \cdot 90$$

Similarly, $$P\left[\frac{1}{c}\frac{s_1^2}{s_2^2} \leqslant \frac{\sigma_1^2}{\sigma_2^2} \leqslant c\frac{s_1^2}{s_2^2}\right] = \cdot 90$$

$$\therefore \qquad P\left[c\frac{s_1^2}{s_2^2} \geqslant \frac{\sigma_1^2}{\sigma_2^2} \geqslant \frac{1}{c}\frac{s_1^2}{s_2^2}\right] = \cdot 90$$

or $$P\left[\frac{1}{c}\frac{s_1^2}{s_2^2} \leqslant \frac{\sigma_1^2}{\sigma_2^2} \leqslant c\frac{s_1^2}{s_2^2}\right] = \cdot 90$$

Hence 90% confidence limits for $\frac{\sigma_1^2}{\sigma_2^2}$ are $\frac{1}{c}\frac{s_1^2}{s_2^2}$ and $c\frac{s_1^2}{s_2^2}$.

Ex. Let $\bar{x}_i, s_i^2, n_i$ $(i=1, 2, \ldots\ldots, k)$ be means, variances and sizes of sample drawn from k normal populations with means $\mu_i (i=1, 2, \ldots\ldots, k)$ and same variance σ^2, show that

(*i*) $$\frac{\frac{\sum_1^k n_i(\bar{x}_i-\mu_i)^2}{k}}{\frac{\sum_1^k (n_i-1)s_i^2}{}} \cdot \frac{n_1+n_2+\ldots+n_k-k}{k}$$

is distributed as F with k and $(n_1+n_2+\ldots+n_k-k)$ *d.f.*

(*ii*) $$\frac{\frac{\sum_1^k n_i(\bar{x}_i-\bar{x})^2}{k}}{\frac{\sum_1^k (n_i-1)s_i^2}{}} \cdot \frac{n_1+n_2+\ldots\ldots+n_k-k}{k-1}$$

is distributed as F with $k-1$, $n-k$ *d. f.* where $n=\sum_{i=1}^{k} n_i$.

13·8. Fisher's z distribution. Making the change of variable *viz.*,

$$z = \tfrac{1}{2}\log_e F \qquad \text{or} \qquad F = e^{2z} \qquad \ldots(13\cdot19)$$

in the F distribution, we get

Table 5

The variance ratio, 5 *and* 1% *'points' of F*

ν_1 is the number of degrees of freedom for the greater estimate of variance, and ν_2 for the smaller.

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
2	18·51	19·00	19·16	19·25	19·30	19·33	19·37	19·41	19·45	19·50
	98·49	99·00	99·17	99·25	99·30	99·33	99·36	99·42	99·46	99·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·84	8·74	8·64	8·53
	34·12	30·82	29·46	28·71	28·24	27·91	27·49	27·05	26·60	26·12
4	7·71	6·94	6·59	6·39	6·26	6·16	6·04	5·91	5·77	5·63
	21·20	18·00	16·69	15·98	15·52	15·21	14·80	14·37	13·93	13·46
5	6·61	5·79	5·41	5·19	5·05	4·95	4·82	4·68	4·53	4·36
	16·26	13·27	12·06	11·39	10·97	10·67	10·27	9·89	9·47	9·02
6	5·99	5·14	4·76	4·53	4·39	4·28	4·15	4·00	3·84	3·67
	13·74	10·92	9·78	9·15	8·75	8·47	8·10	7·72	7·31	6·88
7	5·59	4·74	4·35	4·12	3·97	3·87	3·73	3·57	3·41	3·23
	12·25	9·55	8·45	7·85	7·46	7·19	6·84	6·47	6·07	5·65
8	5·32	4·46	4·07	3·84	3·69	3·58	3·44	3·28	3·12	2·93
	11·26	8·65	7·59	7·01	6·63	6·37	6·03	5·67	5·28	4·86
9	5·12	4·26	3·86	3·63	3·48	3·37	3·23	3·07	2·90	2·71
	10·56	8·02	6·99	6·42	6·06	5·80	5·47	5·11	4·73	4·31
10	4·96	4·10	3·71	3·48	3·33	3·22	3·07	2·91	2·74	2·54
	10·04	7·56	6·55	5·99	5·64	5·39	5·06	4·71	4·33	3·91
12	4·75	3·88	3·49	3·26	3·11	3·00	2·85	2·69	2·50	2·30
	9·33	6·93	5·95	5·41	5·06	4·82	4·50	4·16	3·78	3·36
14	4·60	3·74	3·34	3·11	2·96	2·85	2·70	2·53	2·35	2·13
	8·86	6·51	5·56	5·03	4·69	4·46	4·14	3·80	3·43	3·00
16	4·49	3·63	3·24	3·01	2·85	2·74	2·59	2·42	2·24	2·01
	8·53	6·23	5·29	4·77	4·44	4·20	3·89	3·55	3·18	2·75
18	4·41	3·55	3·16	2·93	2·77	2·66	2·51	2·34	2·15	1·92
	8·28	6·01	5·09	4·58	4·25	4·01	3·71	3·37	3·01	2·57
20	4·35	3·49	3·10	2·87	2·71	2·60	2·45	2·28	2·08	1·84
	8·10	5·85	4·94	4·43	4·10	3·87	3·56	3·23	2·86	2·42
25	4·24	3·38	2·99	2·76	2·60	2·49	2·34	2·16	1·96	1·71
	7·77	5·57	4·68	4·18	3·86	3·63	3·32	2·99	2·62	2·17
30	4·17	3·32	2·92	2·69	2·53	2·42	2·27	2·09	1·89	1·62
	7·56	5·39	4·51	4·02	3·70	3·47	3·17	2·84	2·47	2·01
40	4·08	3·23	2·84	2·61	2·45	2·34	2·18	2·00	1·79	1·51
	7·31	5·18	4·31	3·83	3·51	3·29	2·99	2·66	2·29	1·81
60	4·00	3·15	2·76	2·52	2·37	2·25	2·10	1·92	1·70	1·39
	7·08	4·98	4·13	3·65	3·34	3·12	2·82	2·50	2·12	1·60
80	3·96	3·11	2·72	2·49	2·33	2·21	2·06	1·88	1·65	1·32
	6·96	4·88	4·04	3·56	3·25	3·04	2·74	2·41	2·03	1·49

Extracted from G.W. Snedecor's *Statistical Methods*, pp. 184—87, by courtesy of the author and the Iowa Collegiate Press

$$dP=\frac{2\,\nu_1^{\frac{\nu_1}{2}}\,\nu_2^{\frac{\nu_2}{2}}\,e^{\nu_1 z}\,dz}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)\left(\nu_1 e^{2z}+\nu_2\right)^{\frac{1}{2}(\nu_1+\nu_2)}},\quad -\infty<z<\infty \qquad \ldots(13{\cdot}20)$$

This distribution is known as Fisher's z distribution and is more nearly symmetrical than the distribution of F. It is symmetrical function of n_1 and n_2. Table has been prepared showing, for various values of ν_1 and ν_2, the values of z that will be exceeded in simple sampling with probabilities 0·05 and 0·01 respectively.

Ex. 1. Writing $x=\frac{\nu_1 e^z}{\nu_1 e^z+\nu_2}$, show that (13·20) reduces to Pearson Type 1

$$dP=\frac{1}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\,x^{\frac{1}{2}(\nu_1-2)}\,(1-x)^{\frac{1}{2}(\nu_2-2)}\,dx,\ 0\leqslant x\leqslant 1.$$

Ex. 2. Show that the characteristic function of z is

$$\phi(t)=\left(\frac{\nu_2}{\nu_1}\right)^{\frac{\theta}{2}}\left|\frac{\overline{\frac{\nu_2-\theta}{2}}}{\overline{\frac{\nu_2}{2}}}\right|\left|\frac{\overline{\frac{\nu_1-\theta}{2}}}{\overline{\frac{\nu_1}{2}}}\right| \text{ where } \theta=it.$$

Taking logarithm and using Stirling expansion

$$\log\left|\overline{1+x}\right|=\tfrac{1}{2}\log 2\pi+(x+\tfrac{1}{2})\log x-x+\frac{1}{12x}-\ldots$$

show that for large ν_1 and ν_2

$$\log\phi(t)=\frac{\theta}{2}\left(\frac{1}{\nu_2}+\frac{1}{\nu_1}\right)+\frac{\theta^2}{4}\left(\frac{1}{\nu_1}+\frac{1}{\nu_2}\right)\ldots\ldots$$

Hence, deduce that for large ν_1 and ν_2, z is distributed approximately normally with mean$=\frac{1}{2}\left(\frac{1}{\nu_2}+\frac{1}{\nu_1}\right)$ and variance$=\frac{1}{2}\left(\frac{1}{\nu_2}+\frac{1}{\nu_1}\right)$.

(M.A., Bombay, 1950)

13·9. Inter-relations between t, χ^2 and F Distributions.

13·9·1. Theorem. *The square of t variate with n.d. f. is distributed as F with 1 and n.d. f.*

Putting $F=t^2$, $n_1=1$, $n_2=n$ in the distribution of F with n_1, n_2 *d. f.*, we get

$$dP=\frac{\left(\frac{1}{n}\right)^{\frac{1}{2}}\,(t^2)^{\frac{1-2}{2}}\,d(t^2)}{B\left[\frac{1}{2},\frac{n}{2}\right]\left[1+\frac{t^2}{n}\right]^{(\frac{1}{2}+\frac{1}{2}n)}}$$

$$=\frac{1}{\sqrt{n}}\cdot\frac{1}{B\left[\frac{1}{2},\frac{n}{2}\right]}\cdot\frac{dt}{\left[1+\frac{t^2}{n}\right]^{\frac{1}{2}(n+1)}}$$

which is t distribution with n*.d. f.*

13·9·2. Theorem. *As n_2 tends to infinity, n_1 F tends to be distributed as a χ^2 variate with n_1 d. f.*

The distribution of F with n_1, n_2 $d.f.$ can be written as

$$dP = \frac{n_1^{\frac{n_1}{2}} F^{\frac{n_1}{2}-1}}{\left|\frac{n_1}{2}\right| \left[1+\frac{n_1}{n_2}F\right]^{\frac{n_1+n_2}{2}}} dF \times \frac{\left|\frac{n_1+n_2}{2}\right|}{n_2^{\frac{n_1}{2}} \left|\frac{n_2}{2}\right|}$$

Now using Stirling's formula, we get

$$\frac{\left|\frac{n_1+n_2}{2}\right|}{n_2^{\frac{n_1}{2}}\left|\frac{n_2}{2}\right|} = \frac{e^{-\frac{n_1+n_2}{2}-1}\left[\frac{n_1+n_2}{2}-1\right]^{\frac{n_1+n_2}{2}-\frac{1}{2}}\sqrt{2\pi}}{n_2^{\frac{n_1}{2}} e^{-\frac{n_2}{2}-1}\left[\frac{n_2}{2}-1\right]^{\frac{n_2}{2}-\frac{1}{2}}\sqrt{2\pi}}$$

$$= \frac{e^{-\frac{n_1}{2}}\left[\frac{n_1+n_2-2}{2}\right]^{\frac{n_1+n_2-1}{2}}}{n_2^{\frac{n_1}{2}}\left[\frac{n_1-2}{2}\right]^{\frac{n_2-1}{2}}}$$

$$\sim \frac{e^{-\frac{n_1}{2}}\left[1+\frac{n_1-2}{n_2}\right]^{\frac{n_1-1}{2}}\left[1+\frac{n_1-2}{n_2}\right]^{\frac{n_2}{2}}}{2^{\frac{n_1}{2}}\left(1-\frac{2}{n_2}\right)^{\frac{n_2-1}{2}}}$$

$$\therefore \quad \frac{\left|\frac{n_1+n_2}{2}\right|}{n_2^{\frac{n_1}{2}}\left|\frac{n_2}{2}\right|} \longrightarrow \frac{e^{-\frac{n_1}{2}} e^{\frac{n_1}{2}}}{2^{\frac{n_1}{2}}} \quad \text{or } 2^{-\frac{1}{2}n_1} \text{ as } n_2 \to \infty$$

Thus, when n_2 tends to infinity, we get

$$dP = \frac{n_1^{\frac{n_1}{2}} F^{\frac{n_1}{2}-1} dF}{2^{\frac{n_1}{2}}\left|\frac{n_1}{2}\right| e^{\frac{n_1}{2}F}}$$

Putting $F = \frac{\chi^2}{n_1}$, we have

$$dP = \frac{n_1^{\frac{n_1}{2}}\left(\frac{\chi^2}{n_1}\right)^{\frac{n_1}{2}-1} d\left[\frac{\chi^2}{n_1}\right]}{2^{\frac{n_1}{2}}\left|\frac{n_1}{2}\right| e^{\frac{1}{2}\chi^2}}$$

or $$dP=\frac{1}{2^{\frac{1}{2}n_1}\left|\frac{1}{2}n_1\right.} e^{-\frac{1}{2}\chi^2}(\chi^2)^{\frac{1}{2}(n_1-2)}\,d(\chi^2)$$

which is χ^2 distribution based on n_1 *d. f.*

Note. In Example 1·3·3, we deduced the F distribution from χ^2 distribution. In the present example, we have deduced the χ^2 distribution as a limiting form of the F distribution.

Ex. Show how the F distribution is related to the t distribution and the Chi-square distribution. *(M.A. Bombay, 1958 ; B.A., Hons. Delhi, 1958)*

EXERCISES 13

1. Show that for the χ^2 distribution with *v.d. f.*, the moments about the mean obey the relation $\mu_{j+1}=2j(\mu_j+\nu\mu_{j-1})$, $j\geqslant 1$. Obtain also the values of β_1 and β_2.

2. (*a*) Show that the mode of χ^2 distribution is $n-2$ and that of the χ distribution is given by $\chi^2=n-1$.

(*b*) Also show that for χ^2 distribution with *n.d. f.*

$$\gamma_1=\sqrt{\frac{8}{n}},\quad \gamma_2=\frac{12}{n} \text{ for large } n.$$

3. Show that, for large n, $\sqrt{2\chi^2}$ where χ^2 is a Chi-square variate with *xd. f.* is normally distributed about mean $\sqrt{2n-1}$ with variance unity.
(M.A. Delhi, 1973 ; I.C.A.R., Delhi, 1950)

First Proof. Using Stirling's formula for $n!$ we find, after some reduction moments of the χ-distribution about origin are given by

$$\mu_1'=\sqrt{n}\left(1-\frac{1}{4n}+\frac{1}{32n^2}-\frac{17}{192n^3}+\ldots\ldots\right)$$

whence
$$\mu_1'^2=n\left(1-\frac{1}{2n}+\frac{1}{8n^2}-\frac{37}{192n^2}+\ldots\ldots\right)$$

$$\left.\begin{aligned}\mu_2'&=n\\ \mu_3'&=(n+1)\mu_1'\\ \mu_4'&=(n+2)n\end{aligned}\right\}\quad\text{and}\quad\left.\begin{aligned}\mu_2&=\frac{1}{2}-\frac{1}{8n}+\ldots\ldots\\ \mu_3&=\frac{1}{4\sqrt{n}}+\ldots\ldots\\ \mu_4&=\frac{3}{4}+\frac{31}{48n}+\ldots\ldots\end{aligned}\right.$$

Hence for the moments of $\sqrt{2\chi^2}$ we have

$$\mu_2=1-\frac{1}{4n},\ \gamma_1=\frac{1}{\sqrt{2n}}+\ldots,\ \gamma_2=\frac{3}{12n}+\ldots\ldots$$

Comparing with the results of Ex. 2 (*b*) we conclude that $\sqrt{2\chi^2}$ tends to normality with considerably greater rapidity than χ^2.

Moreover, the expression for μ_1' of χ is equal to $\sqrt{n-\frac{1}{2}}$ to order $n^{-3/2}$ and hence $\sqrt{2\chi^2}$ is distributed about mean $\sqrt{2n-1}$ to that order with unit variance to order n^{-1}.

Second Proof. $\sqrt{2\chi^2}\leqslant\sqrt{2n-1}+z$ is equivalent to $\chi^2\leqslant(n-\frac{1}{2})+z\sqrt{2n-1}+\frac{1}{2}z^2$ *i.e.* $\dfrac{\chi^2-1}{\sqrt{n}}<z+\dfrac{z^2}{2\sqrt{2n}}$.

As $n \to \infty$, this expression is asymptotic to $n+z\sqrt{2n}$, and therefore

$$P\left[\frac{\chi^2-n}{\sqrt{2n}} \leqslant z\right] \sim P\left[\sqrt{2\chi^2} \leqslant \sqrt{2n-1}+z\right]$$

and we know by the central limit theorem that as the number n of degrees of freedom increases the distribution of $(\chi^2-n)/\sqrt{2n}$ tends to normal.

4. (*a*) Show that for 2 *d. f.* the probability P of a value of χ^2 greater than χ_0^2 is $e^{-\frac{1}{2}\chi_0^2}$ and hence that $\chi_0^2 = 2 \log_e (1/P)$. Deduce the value of χ_0^2 when $P=0{\cdot}5$.

(*M.A., Bombay, 1950* ; *M.A. Patna, 1947* ; *B.A., Hons , Delhi, 1956, 59*)

(*b*) The variates $x_1, x_2, \ldots\ldots, x_n$ are independently distributed in the rectangular form

$$dF = dx, \qquad 0 \leqslant x \leqslant 1.$$

Show that $-2 \log (x_1, x_2 \ldots\ldots x_n)$ is distributed as χ^2 with $2n$ *d. f.*

(*B.Sc., Bombay, 1969* ; *B.Sc., Sardar Patel, 1969*)

[**Hint.** (*b*) Show that $y=-2 \log x$ or $x=e^{-\frac{1}{2}y}$ is distributed as χ^2 variate with 2 *d. f.* and then use additive property of χ^2-distribution]

5. (*a*) Integrating by parts, show that if r is a positive integer,

$$\int_{\beta}^{\infty} \frac{1}{r\,!}\, e^{-t}\, t^r\, dt = \frac{1}{r\,!}\, \beta^r\, e^{-\beta} + \int_{\beta}^{\infty} \frac{1}{r-1)\,!}\, e^{-t}\, t^{r-1}\, dt$$

and by continuing the process, show that the value of the intergal is

$$e^{-\beta}\left(1+\beta+\frac{\beta^2}{2\,!}+\ldots\ldots+\frac{\beta^r}{r\,!}\right)$$

(*M.A., Madras, 1959*)

Hence show that the probability, in random sampling that the value of χ^2 with ν (even) *d. f.* will exceed χ_0^2 is obtained from this expression by putting

$$\beta = \tfrac{1}{2}\chi_0^2 \text{ and } r = \tfrac{1}{2}(\nu-2).$$

Hence or otherwise show that the values of $P(\chi^2 \geqslant \chi_0^2)$ for a given value of χ^2 can be obtained from tables of the Poisson exponential limit.

(*b*) For χ^2 distribution of $(2\nu+2)$ *d. f.* prove that

$$P(\chi^2 \geqslant \chi_0^2) = \frac{1}{\nu\,!}\int_{\lambda}^{\infty} e^{-\nu}\, \nu^{\nu}\, d\nu = e^{-\lambda} \sum_{0}^{\nu} \frac{\lambda^r}{r\,!}$$

where $\lambda = \frac{1}{2}\chi_0^2$. Explain the uses of the result. (*M.A., Statistics, Delhi, 1960*)

6. (*a*) If x is a normal variate with zero mean and s.d. unity, obtain the distribution of $t = \dfrac{x\sqrt{n}}{\chi} = \dfrac{x}{\sqrt{\dfrac{1}{n}\sum_{1}^{n} x_i^2}}$. (*I.C.A.R , Delhi, 1947*)

Find also the mean and variance of the distribution

(*b*) Let u be a normal variate with zero mean and unit variance and ν^2 have a χ^2 distribution with n *d. f.* and be distributed independently of u. Derive the distribution of $\dfrac{u\sqrt{n}}{\nu}$. (*M.A., Delhi, 1951*)

(*c*) Show that the mean deviation about origin of t distribution with n*d. f.* is

$$\sqrt{n}\left|\frac{n-1}{2}\right| \Big/ \sqrt{\pi}\left|\frac{n}{2}\right|$$

[**Hint.** (*a*) $\frac{t^2}{n}$ is a $\beta_2\left(\frac{1}{2}, \frac{n}{2}\right)$ variate

(*b*) Required distribution is t distribution with n *d.f.*].

7. Express the constants y_0, a and m of the distribution

$$f(x)=y_0\left(1-\frac{x^2}{a^2}\right)^m, \quad -a\leqslant x\leqslant a$$

in terms of its μ_2 and β_2. Show that if x is related to a variable t by the following equation

$$x=\frac{at}{[2(m+1)+t^2]^{\frac{1}{2}}}$$

then t is distributed in Student's distribution with $2(m+1)$ degrees of freedom. Use the transformation to calculate the probability that $t\geqslant 2$ when the degrees of freedom are (*i*) 2 and (*ii*) 4.

(*B.A., Hons., Delhi, 1962*; *M.A. Statistics, Delhi, 1960*; *M.A., Bombay, 1956*; *M.A. Patna, 1956, 60*)

[**Ans.** $y_0=1/aB(\frac{1}{2}, m+1)$, $a^2=(2m+3)\mu_2$, $m=(9-5\beta_2)/2(\beta_2-3)$, ·092, ·5$-\frac{16\sqrt{5}}{64}$]

8. If x_1, x_2, x_3, x_4 are independent random readings from a normal population with zero mean and unit variance obtain the sampling distributions of the following statistics :

(*a*) $u=\dfrac{x_1^2+x_2^2}{x_1^2+x_2^2+x_3^2}$ (*b*) $v=\dfrac{x_1+x_2+x_3}{[\frac{1}{2}(x_1-x_3)^2+\frac{1}{3}(x_1-2x_2+x_3)^2+x_4^2)]^{\frac{1}{2}}}$

(*M.A., Calcutta 1942*)

(you may assume the additive property of x^2)

[**Hint.** (*b*) $x_1+x_2+x_3$ is normally distributed about zero mean with variance 3.

Hence $\left(\frac{x_1+x_2+x_3}{\sqrt{3}}\right)^2$ is distributed like χ^2 with 1 *d.f.* Similarly each of $\left(\frac{x_1-x_3}{\sqrt{2}}\right)^2$, $\left(\frac{x_1-2x_2+x_3}{\sqrt{6}}\right)^2$, $(x_4)^2$ is distributed independently (to be proved) like χ^2 with 1 *d.f.* and hence their sum is distributed like χ^2 with 3 *d.f.* Hence v^2 is the ratio of two variates independently distributed like χ^2 with 1 and 3 *d.f.* respectively. Now apply § 13·4].

9. If x_1, x_2, x_3 obey the joint distribution

$$(x_1\ x_2\ x_3)^{-\frac{1}{2}} f(x_1+x_2+x_3)\,dx_1\,dx_2\,dx_3$$

in the range $(0, \infty)$, obtain the sampling distribution of

(*i*) $u=x_1+x_2+x_3$. (*ii*) $v=\dfrac{x_2}{x_1}$, (*iii*) $w=\dfrac{x_3}{x_1+x_2}$.

(*M.A., Calcutta, 1946*)

[Put $U=x_1+x_2+x_3$, $UV=x_2+x_3$, $UVW=x_3$, $dx_1\,dx_2\,dx_3=U^2V\,dU\,dV\,dW$]

10. Show that the probability density of s^2 is maximum when $s^2=\sigma^2$ as $n\to\infty$ but the probability density of s is not maximum when $s=\sigma$ as $n\to\infty$.

11. Show that the mode of the F distribution with $n_1(\geqslant 2)$, n_2 *d.f.* is given by

$$F=\frac{n_2(n_1-2)}{n_1(n_2+2)}$$

and is always less than unity.

Show also that the two points of inflexion which exist for $n_1 > 4$ are equidistant from the mode.

12. (*a*) If u, v possess independent χ^2 distributions with n_1 and n_2 *d.f.* respectively, then $F = \frac{u/n_1}{v/n_2}$ has the F distribution with n_1 and n_2 *d.f.*

(*M.A., Stat. Delhi 1968*)

Show that by the change of variable

$$x = \frac{\frac{n_1}{n_2} F}{1 + \frac{n_1}{n_2} F} \quad \text{or} \quad F = \frac{n_2}{n_1} \cdot \frac{x}{1-x}$$

the F distribution can be transformed to Pearson Type I.

13. If x is distributed as F with 2 and *nd.f.* show that

$$P(x \geq k) = \left(1 + \frac{2k}{n}\right)^{\frac{-n}{2}}$$

14. Show that the distribution of $F = \frac{s_1^2 \sigma_2^2}{s_2^2 \sigma_1^2}$ where s_1^2, s_2^2 are the independent estimates of σ_1^2 and σ_2^2 respectively based on n_1, n_2 degrees of freedom depends only on n_1 and n_2.

(*b*) Show also that the mean of the F distribution does not involve n_1.

(*M A., Delhi, 1958*)

(*c*) When $n_1 = 2$, show that the significance level of F corresponding to a significance probability p is

$$F = \frac{n_2}{2}\left(p^{\frac{-2}{n_2}} - 1\right).$$

(*M.A., Statistics, Delhi, 1965* ; *M.A., Delhi, 1957* ; *I.C.A.R. Delhi, 1947*)

(*d*) Find the ·95 probability level of E for 2 and 4 degrees of freedom by direct integration of the distribution function.

15. (*a*) Let x_1, x_2 be a random sample from a distribution having probability density function $f(x) = e^{-x}$, $0 < x < \infty$ and 0 elsewhere. Show that (*i*) $z = x_1/x_2$ has an F distribution ; (*ii*) $x_1 + x_2$ and x_1/x_2 are independently distributed. (*M.A. Aligarh, 1960*)

(*b*) Show that the distribution of the quotient of two independent standard normal variates follows Cauchy's distribution. (*M.A., Stat. Delhi, 1968*)

16. Show that the mean of Fisher's z distribution is the value of

$$\tfrac{1}{2} \log\left(\frac{n_2}{n_1}\right) - \sum_{r=0}^{\infty}\left(\frac{1}{n_1+t+2r} - \frac{1}{n_2-t+2r}\right)$$

when t is put equal to zero and variance is the differential coefficient of this expression with respect to t with t put equal to zero. (*I.C.A.R., Delhi, 1946*)

17. If s_1^2 and s_2^2 are the variances in two independent samples of the same size taken from a common normal population, determine the distribution of s_1^2/s_2^2. (*M.A., Eco. Statistics Delhi, 1958*)

[**Hint.** It can be shown that *m.g.f.* of s^2 in samples of n from a normal population with zero mean and variance σ^2 is

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

$$\therefore \quad E[e^{t(s_1^2+s_2^2)^2}] = [M(t)]^2 = \left(1-\frac{2t\sigma^2}{n}\right)^{-(n-1)}$$

Hence the probability differential for $s_1^2+s_2^2$ is

$$dp = \frac{\left(\frac{n}{2\sigma^2}\right)^{n-1}}{n-1} \exp\left\{-n\,\frac{(s_1^2+s_2^2)}{2\sigma^2}\right\} \times (s_1^2+s_2^2)^{n-2} d(s_1^2+s_2^2)]$$

18. Show that if χ_1^2, χ_2^2 are two independent χ^2 variates with n_1, n_2 *d.f.* and if t follows Student's distribution, then the statistics

$$\frac{\chi_1^2}{\chi_1^2+\chi_2^2},\ \chi_1^2,\ \frac{1}{\chi_1^2},\ \frac{\chi_1^2}{\chi_2^2},\ t$$

follow respectively Pearson Types I, III, V, VI, and VII. (*I.C.A.R., Delhi 1947*)

19. x_i $(i=1, 2, \ldots\ldots, n)$ are independently and normally distributed with zero mean and common variance σ^2. Let

$$\xi_i = \sum_{j=1}^{n} c_{ij} \qquad i=1, 2, \ldots\ldots, p<n$$

where $$\sum_j c_{ij}c_{i'j} = \delta_{ii'}$$

where $\delta_{ii'}$ is Kronecker delta, show that

$$\left[\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{p} \xi_i^2\right] \Big/ \sigma^2$$

is distributed as χ^2 with $(n-p)$ degrees of freedom (R.A. Fisher)

(*M.A., Delhi, 1956*)

Solution. Since we are given that the coefficient c_{ij} satisfy the relations

$$\left.\begin{aligned} \sum_i c^2{}_{ij} = 1 = \sum_i c^2{}_{ij'} \qquad & \ldots(i) \\ \sum_i c_{ij}c_{ij'} = 0 = \sum_j c_{ji}c_{ji'} \qquad & (i' \neq i) \end{aligned}\right]$$

we immediately have

$$\sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} x_i^2$$

so that $$\left[\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{p} \xi_i^2\right] \Big/ \sigma^2 = \sum_{i=p+1}^{n} \left(\frac{\xi_i^2}{\sigma^2}\right)$$

Also like x_i, ξ_i are normally distributed with zero mean and common variance σ^2. For the probability that simultaneously the n values of the independent variates x_i will fall in the respective intervals dx_i, is by the theorem of compound probability,

$$dp = (2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2\sigma^2}\Sigma x_i^2} dx_1 dx_2 \ldots dx_n = (2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2}\Sigma \frac{\xi_1^2}{\sigma^2}} dx_1 dx_2 \ldots dx_n$$

But $$dx_1 dx_2 \ldots dx_n = J\left(\begin{matrix} x_1, x_2, \ldots, x_n \\ \xi_1, \xi_2, \ldots, \xi_n \end{matrix}\right) d\xi_1\, d\xi_2 \ldots\ldots d\xi_n$$

$$= |\, c_{ij} \,|\, d\xi_1 d\xi_2 \ldots\ldots d\xi_n = d\xi_1 d\xi_2 \ldots\ldots d\xi_n$$

since $|\, c_{ij}^2 \,| = 1$ by virtue of the condition (1).

Hence the probability that ξ_i will fall simultaneously in their respective intervals $d\xi_i$ is

$$dp=\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}\xi_1^2/\sigma^2}\,d\xi_1.\ \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}\xi_2^2/\sigma^2}d\xi_2\cdots\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}\xi_n^2/\sigma^2}\,d\xi_n$$

It follows that ξ_i*'s are independently and normally distributed about zero mean with same variance* σ^2.

Hence $\sum_{p+1}^{n}\frac{\xi_i^2}{\sigma^2}$ is the sum of the squares of $(n-p)$ independently and normally distributed variates and therefore by § 13·2 cor. 1 it is distributed like χ^2 with $(n-p)$ *d.f.*

20. Use Exercise 19 to show that x_i are object to p linear constraints the distribution of the sum of the squares is that of χ^2 with $(n-p)$ *d.f.*

21. $x_1, x_2,\ldots, x_n$ is a sample of n independent observations from a normal population with mean μ and variance σ^2, show by means of an orthogonal transformation

$$y_1=\frac{x_1}{\sqrt{n}}+\frac{x_2}{\sqrt{n}}+\ldots+\frac{x_n}{\sqrt{n}}=\bar{x}\sqrt{n}$$

and $$y_i=c_{i1}x_1+c_{i2}x_2+\ldots+c_{in}x_n,\ i=1,2,\ldots,n$$

or otherwise, that the distribution of

$$u=\sqrt{n}\,(\bar{x}-\mu) \text{ and } v=\sum_{1}^{n}(x_i-\bar{x})^2$$

are independent and derive their distributions. Hence obtain the mean and variance of

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

22. If $\bar{x}$ and s^2 are the mean and the variance of a random sample of n members from an infinite population, derive the formulae for :

(*i*) Var $(\bar{x})$.

(*ii*) Var (s^2), for large n.

(*iii*) Var (s), for large n.

If $a=\bar{x}+ks$ and $b=\bar{x}-ks$ where $k=2(\beta_2-1)^{-\frac{1}{2}}$, show that, in large samples, a and b are uncorrelated. (*M.A., Punjab, 1958*; *M.A., Patna 1956*)

23. If $x_i(i=1,\ldots,n)$ are independent random variables each distributed normally with mean zero and variance one, show that Σx_i and $\Sigma(x_i-\bar{x})^2$ are independent. Find the mean and variance of $(1/n)\ \Sigma(x_i-\bar{x})^2$ and show that they approach 1 and $2/n$ as n increases. (*M.A. Aligarh, 1960*)

24. Show that the distribution of the sum y of three independent observations from the distribution

$$f(x)=\tfrac{1}{2}e^{-|x|},\ -\infty<x<\infty$$

is given by $\frac{1}{16}(3+3|y|+y^2)\ e^{-|y|}\,dy$

[**Hint.** Distribution of y is obtained by integrating $e^{-ity}.(1+t^2)^{-3}$ as, contour integral, the characteristic function being $1/(1+t^2)$].

25. Show that the distribution of the standard deviation of samples of two from the rectangular population $dF=dx$, $0 \leqslant x \leqslant 1$, is given by

$$f(s)=4(1-2s),\ 0 \leqslant s \leqslant \tfrac{1}{2}.$$

[**Hint.** $dF=dx_1\ dx_2$, $s=\dfrac{x_1-x_2}{2}$, $u=\dfrac{x_1+x_2}{2}$, $dx_1dx_2=2\ dsdu$. Integrate for u from s to $1-s$ and double the result.]

26. The variables x and y with zero means and standard deviations σ_1 and σ_2 are normally correlated with coefficient ρ. Show that u and v, defined by

$$u=\frac{x}{\sigma_1}+\frac{y}{\sigma_2}; \qquad v=\frac{x}{\sigma_1}-\frac{y}{\sigma_2}$$

are independent normal variates with variances $2(1+\rho)$ and $2(1-\rho)$ respectively; and that, if R is the correlation between u and v in random samples of n pairs from the bivariate population R^2 is a $\beta_1(\frac{1}{2}, \frac{1}{2}n-1)$ variate.

(*M.A., Statistics, Delhi, 1959*)

[**Hint.** Since $-\dfrac{1}{2(1-\rho^2)}\left[\dfrac{x^2}{\sigma_1^2}-\dfrac{2\rho xy}{\sigma_1\sigma_2}+\dfrac{y^2}{\sigma_2^2}\right]$

$$=-\tfrac{1}{2}\left[\frac{1}{2(1+\rho)}\left\{\frac{x}{\sigma_1}+\frac{y}{\sigma_2}\right\}^2+\frac{1}{2(1-\rho)}\left\{\frac{x}{\sigma_1}-\frac{y}{\sigma_2}\right\}^2\right]$$

We conclude that $u=\dfrac{x}{\sigma_1}+\dfrac{y}{\sigma_2}$ and $v=\dfrac{x}{\sigma_1}-\dfrac{y}{\sigma_2}$ are independent normal variates with variances $2(1+\rho)$, $2(1-\rho)$, respectively. Hence etc.]

27. Derive the distribution of r and show that when $n=5$ and if $P(|Z| \geqslant c)=\alpha$, c is a root of the equation $c\sqrt{1-c^2}+\sin^{-1} c+\frac{1}{2}\pi(\alpha-1)=0$.

(*B.A. Hons. Calcutta*)

14

Exact Sampling Distributions (*Continued*)

14·1. Introduction. In the previous chapter we have obtained the sampling distributions of the statistics $\bar{x}$, s^2, χ^2, t, z, F, $\chi_1^2+\chi_2^2$ and χ_1^2/χ_2^2 for random samples from normal population, by using the geometrical method. In the present chapter we shall illustrate the use of direct analytic method by deriving the χ^2 distribution by its use. We next shall discuss the use of the method of characteristic function to derive the sampling distributions specially that of the sample mean and illustrate its use by deriving the distributions of the sample means from Binomial, Poisson, Normal and Cauchy distributions. The use of the method of induction is discussed in §14·4. In §14·5 we show how to derive the distribution of the sample range and in the next section of the various quantiles.

Up to §14·6 we consider samples from univariate universes only. In §14·7 we consider random samples from a bivariate normal population and derive the distributions of the sample means, variances and in particular the important distribution of the sample correlation coefficient.

14·2. The Analytic Method. We shall illustrate this method by finding the χ^2 distribution *viz.*, sampling distribution of the sum of squares of n independent variates each of which is distributed normally with zero mean and unit variance.

The distribution of the variate x_i is

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}x_i^2\right) dx_i.$$

Hence the joint distribution of n independent variate x_i is

$$dF=(2\pi)^{-\frac{1}{2}n} \exp\left[-\tfrac{1}{2}(x_1^2+x_2^2+\ldots+x_n^2)\right] dx_1 dx_2 \ldots dx_n,$$

The distribution function of

$$z=x_1^2+x_2^2+\ldots\ldots+x_n^2 \qquad \ldots(14\cdot1)$$

is given by

$$F(x)=\int\int\int\ldots\ldots\ldots\int(2\pi)^{-\frac{1}{2}n}\exp\left[-\tfrac{1}{2}\sum_{i=1}^{n}x_i^2\right]dx_1dx_2\ldots dx_n$$

where the domain of x's is conditioned by

$$x_1^2+x_2^2+\ldots\ldots+x_n^2\leqslant z.$$

Let us change the variates to $z, \theta_1, \theta_2,\ldots\ldots,\theta_{n-1}$

where $x_1=\sqrt{z}\cos\theta_1\cos\theta_2\ldots\ldots\cos\theta_{n-1}$,

$x_2=\sqrt{z}\cos\theta_1\ldots\ldots\cos\theta_{n-2}\sin\theta_{n-1}$,

.....

.....

$x_n=\sqrt{z}\sin\theta_1$

and hence $x_1^2+x_2^2+\ldots\ldots+x_n^2=z$.

It can be easily seen that

$$\frac{\partial(x_1, x_2,\ldots\ldots,x_n)}{\partial(z, \theta_1, \theta_2\ldots\ldots\theta_{n-1})}=z^{\frac{n-2}{2}}f(\theta_1, \theta_2,\ldots\ldots, \theta_{n-1})$$

where $f(\theta_1, \theta_2,\ldots, \theta_{n-1})$ is a function of the θ's only.

Hence

$$F(z)=(2\pi)^{-\frac{1}{2}n}\int\int\ldots\ldots\ldots\int z^{\frac{1}{2}(n-2)}\,e^{-\frac{1}{2}z}f(\theta_1, \theta_2,\ldots,\theta_{n-1})d\theta_1\ldots d\theta_{n-1}dz \qquad \ldots(14\cdot2)$$

in which z varies from 0 to z and θ's vary from 0 to π or 2π.

The right hand side of (14·2) is the product of n integrals of which those containing $\theta_1, \theta_2,\ldots\ldots,\theta_{n-1}$ are constants. Thus we get

$$F(z)=k\int_0^z z^{\frac{n-2}{2}}\,e^{-\frac{z}{2}}\,dz,$$

where k is a constant given by

$$1=k\int_0^\infty z^{\frac{n-2}{2}}\,e^{-\frac{1}{2}z}\,dz=k2^{\frac{n}{2}}\Gamma(n/2)$$

Thus the required distribution of z is

$$dF=2^{-\frac{1}{2}n}\left[\Gamma(\tfrac{1}{2}n)\right]^{-1}e^{-\frac{1}{2}z}\,z^{\frac{1}{2}(n-2)}\,dz, \qquad 0\leqslant z\leqslant\infty.$$

Ex. 1. Obtain from first principles the sampling distribution of

$$z=x_1^2+x_2^2+x_3^2$$

where x's are each distributed independently and normally with unit variance and zero mean.

Ex. 2. Derive the χ^2 distribution from that of the sum of the squares of n standard normal variates subject to linear restriction $\sum_{i=1}^{n} a_i x_i = 0$ by the substitution $x_i = u_i \sqrt{\chi}$ subject to $\sum_{i=1}^{n} u_i^2 = 1$.

14·3. The Method of Characteristic Functions. We shall make use of the following properties of characteristic functions :

If $x_1, x_2, \ldots, x_n$ is a random sample of size n so that $x_1, x_2, \ldots, x_n$ are independent and $\bar{x}$ be the sample mean, then

$$(i)\ \phi_{x_1+x_2+\ldots+x_n}(t) = \phi_{x_1}(t) \times \phi_{x_2}(t) \times \ldots \times \phi_{x_n}(t) = [\phi_x(t)]^n$$

$$(ii)\ \phi_{\bar{x}}(t) = \phi_{x_1}\left(\frac{t}{n}\right) \phi_{x_2}\left(\frac{t}{n}\right) \ldots \phi_{x_n}\left[\frac{t}{n}\right] = \left[\phi_x\left(\frac{t}{n}\right)\right]^n$$

We illustrate the method by solving a few examples.

14·3·1. Distribution of the Mean for the Poisson Distribution. The characteristic function (*c.f.*) of the Poisson distribution with parameter λ is

$$\exp\left[(\lambda e^{it} - 1)\right]$$

Therefore, by the property (*ii*) of *c.f.*, the *c.f.* of the mean is

$$\left[\exp\left\{e^{\frac{it}{n}} - 1\right\}\right]^n = \exp\left[n\lambda\left\{e^{\frac{it}{n}} - 1\right\}\right]$$

Hence the distribution of the mean is the Poisson distribution with parameter $n\lambda$, the interval being $1/n$ instead of unity.

14·3·2. Distribution of the Mean of the Normal Population. The *c.f.* of the normal distribution with mean μ and *s.d.* σ is

$$\exp\left[-\tfrac{1}{2}t^2\sigma^2 + it\mu\right]$$

The *c.f.* of the distribution of the mean of n values is then

$$\exp\left\{n\left(-\tfrac{1}{2}\frac{t^2\sigma^2}{n^2} + \frac{it\mu}{n}\right)\right\} = \exp\left\{-\tfrac{1}{2}\frac{t^2\sigma^2}{n} + it\mu\right\}$$

This is the *c.f.* of the normal distribution with mean μ and *s.d.* $\sigma/\sqrt{n}$ which is, therefore, the distribution required.

Ex. 1. Show that the distribution of the mean for the Binomial distribution $(q+p)^r$ is $(q+p)^{rn}$, the interval being $1/n$ instead of unity.

Ex. 2. By using characteristic function, obtain the sampling distribution of the mean in samples of size n from Cauchy's population.

$$dF=\frac{dx}{\pi(1+x^2)}, \quad -\infty \leqslant x \leqslant \infty. \qquad \left[\text{Ans. } \frac{1}{\pi}\cdot\frac{d\bar{x}}{1+\bar{x}^2}\right]$$

$$\left[\text{Hint. Use } \int_{-\infty}^{\infty}\frac{e^{itx}dx}{\pi(1+x^2)}=e^{-|t|}\right]$$

Ex. 3. Derive by the method of characteristic functions the sampling distribution of the sum of the squares of n independent standard normal variates.

14·4. The Method of Induction. If $x_1, x_2, \ldots\ldots, x_n$ are n independent variates having the same distribution the general form of the distribution of their sum may be suggested when the results for two or three variates have been worked out and then the validity may be tested by induction. We, therefore, give the method of obtaining the distribution of the sum of two independent variates whose individual distributions are known.

Let x_1 and x_2 be two independent variates with density function $f_1(x_1)$ and $f_2(x_2)$ respectively then the distribution of $z=x_1+x_2$ is given by

$$F(z)=\iint f_1(x_1)\, f_2(x_2)dx_1dx_2,$$

the range of integration being that for which $x_1+x_2 \leqslant z$.

$$F(z)=\int_{-\infty}^{\infty}\left[F_1(x_1)\right]_{-\infty}^{z-x_2} f_2(x_2)dx_2$$

where $$\int_{-\infty}^{x_1} f_1(x_1)dx=F_1(x_1)$$

Hence $$F(z)=\int_{-\infty}^{\infty} F_1(z-x_2)\, f_2(x_2)dx_2$$

Differentiating with respect to z, we get

$$f(z)=\int_{-\infty}^{\infty} f_1(z-x_2)f_2(x_2)dx_2$$

This is the density function of z.

Example 14·1. *Derive (by the method of Induction) the sampling distribution of the mean of a sample of n values $x_1, x_2, \ldots\ldots, x_n$ from Cauchy's distribution.*

$$dF=\frac{dx}{\pi(1+x^2)}, -\infty<x<\infty. \qquad (I.C.A.R., Delhi\ 1947)$$

We first obtain the distribution of the sum

$$z=x_1+x_2+\cdots\cdots+x_n.$$

The distribution of the sum $z=x_1+x_2$ is

$$\int_{-\infty}^{\infty} \pi^2\left[\frac{1}{1+(z-x_2)^2}\right]\left(\frac{1}{1+x_2^2}\right)dx_2=\frac{2}{\pi(2^2+z^2)}.$$

This suggests the general form $\frac{n}{\pi(n^2+z^2)}$ for the probability function of n variates.

If this is true, then the probability function of $(n+1)$ variates is

$$\int_{-\infty}^{\infty} \frac{n}{\pi^2}\left[\frac{1}{n^2+(z-x)^2}\right]\left(\frac{1}{1+x^2}\right)dx$$

$$=\frac{n}{\pi}\left(\frac{n+1}{n}\right)\Big/[(n+1)^2+z^2]$$

$$=\frac{1}{\pi}\cdot\frac{n+1}{[(n+1)^2+z^2]}$$

The result holds for $n=1$, 2 and is, therefore, true in general.

But $\bar{x}=z/n$. Hence the distribution of $\bar{x}$ is

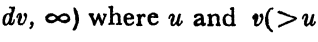
$$dP=\frac{1}{\pi}\cdot\frac{d\bar{x}}{1+\bar{x}^2},\ -\infty<\bar{x}<\infty.$$

Ex. Use the method of induction to obtain the sampling distribution of the sum of square of n independent variates each of which is normally distributed with zero mean and standard deviation.

14·5. Distribution of the Range. In industrial quality control work, the range is a useful measure of dispersion because of its ease in computation Also for small samples from a normal population it can be shown that in a certain sense, the range is as efficient for estimating σ as is the sample standard deviation. We, therefore, are interested in obtaining the sampling distribution of the range w. Since the range of a sample of values is the difference between the highest and lowest values we shall first obtain the joint distribution of these values.

Case 1. *Infinite range of population.* Consider a random sample of n values drawn from a continuous population whose density function is $f(x)$. Let these sample values be arranged in order of increasing magnitude, and denote the ordered set by $x_1, x_2, \cdots\cdots, x_n$.

Let the infinite range be divided into five intervals $(-\infty, u)$, $(u, u+du)$, $(u+du, v)$, $(v, v+dv)$, $(v+dv, \infty)$ where u and $v(>u)$ are any two values of x. Then the probability that a value of x will fall in the interval $(u, u+du)$ is $f(u)du$.



Similarly, the probability that a value of x will fall in the interval $(v, v+dv)$ is $f(v)dv$ and the probability that a value of x will fall in the interval $(u+du, v)$ is

$$\int_{u+du}^{v} f(x)dx = F(v) - F(u+du)$$

where $\int f(x)dx = F(x) + \text{constant}.$

Consequently the probability that in the drawing of n values, no value of x will fall in the first interval, one in the second interval du, $(n-2)$ values in the third interval, one in the fourth interval dv, and no value in the fifth interval is,

$$dP = \frac{n!}{(n-2)} [f(u)du][F(v) - F(u+du)]^{n-2}[f(v)dv]$$
$$= n(n-1) f(u) f(v)[F(v) - F(u)]^{n-2} dudv$$

neglecting infinitesimals of order higher than second in the R.H.S.

or
$$dP = n(n-1) f(u) f(v) \left[\int_{u}^{v} f(x)dx\right]^{n-2} dudv \qquad \ldots(14{\cdot}2)$$

Thus if u and v denote the lowest and the highest values respectively, (14·2) is the joint distribution of u and v.

To obtain the distribution of the range, w, let $w = v - u$, then $J\begin{pmatrix} u, & w \\ u, & v \end{pmatrix} = 1$ so that the probability that u will fall in the interval du and w in the interval dw is

$$dP' = n(n-1) f(u) f(u+w) \left[\int_{u}^{u+w} f(x)dx\right]^{n-2} dudw.$$

Hence the probability that w will fall in the interval dw, irrespective of the value of u, is

$$dP = n(n-1) \left[\int_{-\infty}^{\infty} \left\{ f(u) f(u+w) \left(\int_{u}^{u+w} f(x)dx \right)^{n-2} du \right\} dw \right] \qquad \ldots(14{\cdot}3)$$

This is the required probability differential of w. (14·3) can be written as follows :

$$dP=n(n-1)dw\int_{-\infty}^{+\infty}\left\{F(u+w)-F(u)\right\}^{n-2}.f(u+w)dF(u).$$

$$E(w)=n(n-1)\int_{-\infty}^{+\infty}dF_n\int_{-\infty}^{x_n}(F_n-F_1)^{n-2}(x_n-x_1)dF_1$$

$$=-\int_{-\infty}^{+\infty}dx_n\int_{-\infty}^{x_n}\frac{d^2}{dx_1\,dx_n}(F_n-F_1)^n(x_n-x_1)dx_1$$

On integration by parts under the integral sign, we have

$$E(w)=-\int_{-\infty}^{\infty}dx_n\left\{\left[\,(-x_n-x_1)\,\frac{d}{dx_n}F_n{}^n\,\right]_{x_1=-\infty}+\int_{-\infty}^{x_n}\frac{d}{dx_n}(F_n-F_n)^n dx_1\right\}$$

Since $(F_n-F_1)^n=0$ at the upper limit of integration, the order of integration and differentiation may be interchanged in this last integral so that

$$E(w)=\int_{-\infty}^{+\infty}\left[(x_n-x_1)\,\frac{d}{dx}F_1{}^n\,\right]_{x_1=-\infty}dx_n-\int_{-\infty}^{+\infty}(1-F_1)^n dx_1$$

$$=\int_{-\infty}^{+\infty}\left\{\frac{d}{dx_n}[(x_n-x_1)\,F_n{}^n]-F_n{}^n\right\}_{x_1=-\infty}-\int_{-\infty}^{+\infty}(1-F_1)^n dx_1$$

$$=\left[(x_n-x_1)F_n{}^n\right]_{-\infty,-\infty}^{-\infty,\infty}-\int_{-\infty}^{+\infty}F_n{}^n dx_n-\int_{-\infty}^{+\infty}(1-F_1)^n dx_1$$

$$=\int_{-\infty}^{+\infty}\{1-F^n-(1-F)^n\}dx,\text{ where } dF=f(x)dx.$$

This result is due to L.H.C. Tippett (1925).

Case II. *Finite range of the population.* Let the density function be defined in the finite interval (a, b). Proceeding as in Case I, it can be shown that the result (14·2) holds good. Since $u=v-w$ for w fixed, u cannot exceed the value $b-w$. Hence the probability that w will fall in the interval dw, irrespective of the value of u, is

$$dp=n(n-1)\left[\int_a^{b-w}\left\{f(u)\,f(u+w)\left(\int_u^{u+w}f(x)dx\right)^{n-2}\right\}du\right]dv \quad ...(14\cdot4)$$

This is the required probability differential of w.

The expected value of w is

$$E(w)=\int_0^{b-a} wh(w)\,dw$$

where $h(w)$ is the coefficient of dw in (14·4).

Example 14 2. *Suppose a random sample of n observations is taken from a population with the density function* $f(x)=e^{-x}, x\geqslant 0$. *Find the density function of the range w.*

(M.A. Delhi, 1949 ; I.C.A.R. Delhi, 1949)

Also show that the mean value of w is

$$1+\tfrac{1}{2}+\tfrac{1}{3}+\ldots\ldots+\frac{1}{n-1}.$$

Here $\int_u^{u+w} f(x)\,dx=e^{-u}-e^{-u-w}=e^{-u}(1-e^{-w})$

$$\therefore\quad \theta(w)=n(n-1)\int_0^{\infty}e^{-u}e^{-(u+w)}e^{-(n-2)u}(1-e^{-w})^{n-2}\,du, \text{ using (14·3)}$$

$$=(n-1)e^{-w}(1-e^{-w})^{n-2}$$

$$\therefore\quad E(w)=(n-1)\int_0^{\infty} we^{-w}(1-e^{-w})^{n-2}\,dw \qquad ...(1)$$

$$=(n-1)\int_0^{1}[-\log(1-z)]z^{n-2}\,dz, \text{ putting } 1-e^{-w}=z$$

$$=(n-1)\int_0^{1}\left(z+\frac{z^2}{2}+\frac{z^2}{3}+\ldots\right)z^{n-2}\,dz$$

[The series is uniformly convergent in (0,1)]

$$=(n-1)\left[\frac{1}{n}+\frac{1}{2}\cdot\frac{1}{n+1}+\frac{1}{3}\cdot\frac{1}{n+2}+\ldots\right]$$

$$=\left[\frac{1}{1}-\frac{1}{n}+\frac{1}{2}-\frac{1}{n+1}+\frac{1}{3}-\frac{1}{n+2}+\ldots\right]$$

$$=1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n-1}.$$

Ex. 1. In (1) Ex. 14·2 expand $(1-e^{-w})^{n-2}$ by binomial theorem and integrate term-by-term, using the result $\int_0^\infty we^{-rw}\,dw=\frac{1}{r^2}$.

Deduce the algebraic identity :

$$\frac{{}^{n-1}c_2}{1}-\frac{{}^{n-1}c_3}{2}+\frac{{}^{n-1}c_4}{3}-\ldots+\frac{(-1)^{n-1}\,{}^{n-1}c_{n-1}}{n-1}$$

$$=1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n-1}.$$

Ex. 2. Suppose that samples of size 4 are taken from the population with distribution function $f(x)=e^{-x}$, $x\geqslant 0$. Determine the limits w_1 and w_2 for w such that

$$P(w<w_1)=\cdot 025 \text{ and } P(w>w_2)=\cdot 025$$

Ex 3. For a sample of size n from a population with the uniform distribution $f(x)=\frac{1}{b}$, $0\leqslant x\leqslant b$, show that $\frac{w}{b}$ is a $\beta_1(n-1, 1)$ variate. Obtain the mean and variance of the distribution of the statistic $w'=wb^{-1}$.

[**Ans.** $E(w')=(n-1)/(n+1)$]

Ex. 4. Show that the probability that, in a sample of n random numbers drawn from the numbers between 0 and 1, the range will exceed 0·5 is $1-\left(\frac{n+1}{2n}\right)$. Show also that 8 is the least value of n for which the probability exceeds 0·95.

[Here $f(x)=1$, $P(w>0\cdot 5)]=\int_{\frac{1}{2}}^{1} n(n-1)w^{n-2}(1-w)dw=$ etc. Solve by trial for small value of n]

Ex. 5. Find how many random numbers between 0 and 1 would not be drawn in order that the probability will exceed 0·95 that the range will exceed 0·9.

[n is the least integer given by $\int_{0\cdot 9}^{1} n(n-1)w^{n-2}(1-w)dw>\cdot 95$]

14·6. Distribution of Quantiles. Consider a random sample of size n drawn from the continuous population whose density function is $f(x)$. The probability that one value of x will fall in the interval $(-\infty, u)$ is

$$\int_{-\infty}^{u} f(x)dx=F(u)=F, \text{ say,}$$

and the probability that one value of x will fall in the interval $(u+du, \infty)$ is

$$\int_{u+du}^{\infty} f(x)\,dx = 1 - \int_{-\infty}^{u+du} f(x)\,dx = 1 - F(u+du)$$

Consequently the probability that out of a sample of n, $l-1$ values will fall below the value u, one in the infinitesimal interval $(u, u+du)$ and the remaining $(n-l)$ fall above u, is

$$dp = \frac{n!\,F^{l-1} f(u)du\,[1-F(u+du)]^{n-l}}{(l-1)!\,(n-l)!}$$

$$= \frac{n!}{(l-1)!\,(n-l)!} F^{l-1}(1-F)^{n-l}\,dF \qquad \ldots(14{\cdot}5)$$

where $dF = f(u)\,du$ and infinitesimals of order higher than the first have been neglected on the right hand side.

This is the distribution of u, the number of the sample to which a proportion l/n of the members fall *i.e.*, of the lth quantile.

Example 14·3. *Show that in odd samples of n from a rectangular population of unit range the sampling mean and variance of the distribution of the median is given exactly by* $\frac{1}{2}$ *and* $\frac{1}{4}(n+2)$ *respectively.* *(M.A., Bombay, 1958)*

Let $n = 2m+1$, then the median is $(m+1)$th member of the sample so that in this case $l = m+1$.

Hence the distribution of the median is

$$dP = \frac{(2m+1)!}{(m!)^2} F^m(1-F)^m\,dF$$

where $dF = du, \qquad 0 \leqslant u \leqslant 1$

and $$F = \int_0^u f(x)\,dx = \int_0^u 1\,dx = u.$$

Therefore $$dP = \frac{(2m+1)!}{(m!)^2} u^m(1-u)^m\,du, \quad 0 \leqslant u \leqslant 1$$

$$= \frac{1}{B(m+1,\ m+1)} u^m(1-u)^m\,du$$

$$E(u) = \frac{1}{B(m+1,\ m+1)} \int_0^1 u^{m+1}(1-u)^m\,du$$

$$= \frac{B(m+2,\ m+1)}{B(m+1,\ m+1)} = \frac{1}{2}.$$

Similarly, $E(u^2)=\dfrac{B(m+3,\ m+1)}{B(m+1,\ m+1)}=\dfrac{m+2}{2(2m+3)}$

Therefore $Var(u)=E(u^2)-[E(u)]^2=\dfrac{1}{4(2m+3)}=\dfrac{1}{4(n+2)}$

Example 14·4. *Show that for large samples the distribution of the median tends to normality.*

Let a be the population median, then $F(a)=\int_{-\infty}^{a} f(x)dx=\frac{1}{2}$

Also let $f(a)=y_0$. Then, by Taylor's theorem,

$$F(u)=F(a)+(u-a)\ F'(a)+\frac{(u-a)^2}{2\,!}F''(a)+\ldots$$

$$=\tfrac{1}{2}+(u-a)y_0+\frac{(u-a)^2}{2\,!}f'(a)+\ldots$$

$$f(u)=y_0+(u-a)\,f'(a)+\ldots$$

For a symmetrical distribution $f'(a)=0$ and for a moderately asymmetrical distribution $f'(a)$ will be small. Let

$$Z=\frac{(2m+1)\,!}{(m\,!)^2}F^m(1-F)^m f(u) \text{ where } n=2m+1.$$

For large m, Stirling's formula for factorials,

$$Z\cong\frac{\sqrt{2\pi}e^{-(2m+1)}\ (2m+1)^{2m+3/2}}{(\sqrt{2\pi}\ m^{m+1/2}e^{-m})^2}[\tfrac{1}{2}+(u-a)y_0+\ldots]^m \times[\tfrac{1}{2}-(u-a)y_0+\ldots]^m y_0$$

$$\cong\frac{e^{-1}}{\sqrt{2\pi}}2^{2m+3/2}\,e^{1}\,\sqrt{m}(\tfrac{1}{4})^m[1-4m(u-a)\ y_0{}^2-\ldots]y_0$$

$$=\frac{1}{\sqrt{\pi}}\left(\frac{1}{2\sqrt{2m}y_0}\right)^{-1}e^{-4m(u-a)^2\,y_0{}^2}$$

$$\cong\frac{1}{\sqrt{\pi}}\cdot\frac{1}{\sigma}\ e^{-\frac{(u-a)^2}{2\sigma^2}} \text{ where } \sigma^2=\frac{1}{8my_0{}^2}\sim\frac{1}{4ny_0{}^2}$$

Thus for large sample size n the median is normally distributed about the population median at mean and variance approximately equal to $1/(4ny_0^2)$.

Ex. 1. For large samples from a normal distribution with variance σ^2 show that the variance of the sample median is $\pi\sigma^2/2n$ and for a Cauchy distribution its value is $\pi^2/4n$.

Ex. 2. Obtain the large sample distribution of the median in samples from any population, and deduce the form it takes when the population is

$$dP=\theta\ e^{-\theta x}\,dx,\ 0\leqslant x<\infty,\ \theta>0. \qquad (M.A.,\ Patna,\ 1957)$$

Ex. 3. Show that the distribution of the largest number x_n, of the sample drawn from the population

$$dP = \frac{2x}{2\theta+1}dx,\ 0 \leqslant x \leqslant \theta+1,$$

is $$\frac{2nx_n}{(2\theta+1)^n}.(x_n^2-\theta^2)^{n-1}dx_n.$$

Ex. 4. Show that the distribution of the mth value from the top is

$$\frac{n\,!}{(n-m)\,!\,(m-1)\,!}\,F^{n-m}(1-F)^{m-1}dF$$

when the sample of size n is drawn from a population with a continuous density function $f(x)$.

In samples from the rectangular distribution $dF=dx$, $0 \leqslant x \leqslant 1$ [$\because$ $F(x)=x$], the distribution of the mth value from top is $dP=\text{const.}\ x^{n-m}(1-x)^{m-1}\,dx$, $0 \leqslant x \leqslant 1$.

Ex. 5. Let x_1, x_2, x_3, x_4, x_5 be an order statistic random sample of size from a distribution having probability density function $f(x)=e^{-x}$, $0<x<\infty$, zero else where. Find *p.d. f.* of the median x_3 of the random sample.

[Ans. $g(x_3)=\frac{5\,!}{2\,!\,2\,!}\ e^{-3x_3}\ (1-e^{-x_3})^2$, $0<x_3<\infty$, $g(x_3)=0$, elsewhere]

14·7. *Distribution of the Correlation Coefficient. The sampling distribution of the correlation coefficient, ρ, for random samples of size n from a correlated bivariate normal population will be derived in four steps.

First Step. *Geometrical interpretation of r.*

Let $x_1, y_1\ ;\ x_2, y_2\ ;\ \ldots\ldots;\ x_n, y_n$ be a random sample of size n drawn from a correlated bivariate normal population

$$dP=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\ \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-m_1)^2}{\sigma_1^2}-\frac{2\rho(x-m_1)(y-m_2)^2}{\sigma_1\sigma_2}+\frac{(y-m_2)^2}{\sigma_2^2}\right\}\right]dxdy \quad \ldots(14\cdot6)$$

Let us take a sample space of $2n$ dimensions and in this space let $OX_1, OX_2, \ldots\ldots, OX_n$ and $OY_1, OY_2, \ldots\ldots, OY_n$ be the rectangular axes, O being the origin. Then these axes form two mutually orthogonal spaces, each of n dimensions, and the sample may be represented by the point P with co-ordinates $(x_1, x_2, \ldots\ldots, x_n\ ;\ y_1, y_2 \cdots\cdots, y_n)$.

Let $P_1(x_1, x_2, \ldots\ldots, x_n)$ and $P_2(y_1, y_2, \ldots\ldots, y_n)$ be the projections of P on the x-space and y-space respectively.

---

*The distribution of the correlation coefficient, r, in random samples of n pairs from a bivariate normal population in which the correlation is ρ was given by Prof. R.A. Fisher in 1915.

Further let OE_1, OE_2 be the equiangular lines in the x-space and y-space respectively and let M_1, M_2 be the feet of the perpendiculars from P_1 and P_2 on these. Then as in § 13·3

$$OM_1=\bar{x}\sqrt{n},\ P_1M_1=s_1\sqrt{n},\ OM_2=\bar{y}\sqrt{n},\ P_2M_2=s_2\sqrt{n},$$

where the variances s_1^2, s_2^2 of x and y in the sample space are given by

$$s_1^2=\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2,\ s_2^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2,$$

and the correlation, r ; in the sample is given by

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{ns_1s_2}$$

If we rotate the y-space round O so that it coincides with the x-space, *i.e.*, the axis OY_i falls along $OX_i(i=1, 2,\ldots\ldots, n)$ then P_2 will occupy a new position P_2' with coordinates $y_1, y_2,\ldots\ldots, y_n$ in the x-space and OE_2 will coincide with OE_1.

Let M_2' be the foot of the perpendicular from P_2' on OE_1' then $OM_2'=\bar{y}\sqrt{n}$ and $P_2'M_2'=s_2\sqrt{n}$, M_2 and M_2' are $(\bar{x}, \bar{x}_1-\bar{x})$ and $(\bar{y}, \bar{y}_1-\bar{y})$ respectively.

If θ be the angle between P_1M_1 and $P_2'M_2'$, we have

$$\cos\theta=\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}}=r.$$

Second Step. *Transformation of the density function.*

Now $\sum_{i=1}^{n}(x_i-m_1)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-m_1)^2=n[s_1^2+(\bar{x}-m_1)^2]$

Similarly, $\sum_{i=1}^{n}(y_i-m_2)^2=n[s_2^2+(\bar{y}-m_2)^2]$ and

$$\begin{aligned}\sum_i(x_i-m_1)(y_i-m_2)&=\sum_i(x_i-\bar{x}+\bar{x}-m_1)(y_i-\bar{y}+\bar{y}-m_2)\\&=\sum_i(x_i-\bar{x})(y_i-\bar{y})+n(\bar{x}-m_1)(\bar{y}-m_2)\\&=n[rs_1s_2+(\bar{x}-m_1)(\bar{y}-m_2)]\end{aligned}$$

Therefore the density factor in (14·6) reduces to

$$\text{const. exp}\left[\frac{-n}{2(1-\rho^2)}\left\{\frac{(\bar{x}-m_1)^2}{\sigma_1^2}-\frac{2\rho(\bar{x}-m_1)(\bar{y}-m_2)}{\sigma_1\sigma_2}+\frac{(\bar{y}-m_2)^2}{\sigma_2^2}\right\}\right]$$

$$\times\exp\left[\frac{-n}{2(1-\rho^2)}\left(\frac{s_1^2}{\sigma_1^2}-\frac{2\rho rs_1s_2}{\sigma_1\sigma_2}+\frac{s_2^2}{\sigma_2^2}\right)\right]$$

Third Step. *Transformation of the volume element.*

We will now find out how the volume element changes when $\bar{x}$, $\bar{y}$, s_1, s_2, and r (or θ) are allowed to vary within infinitesimal ranges. For this purpose let us first keep $\bar{x}$ and s_1 fixed, then P_1 will move on the surface of an $(n-1)$ dimensional hypersphere with centre at M_1 and radius $s_1\sqrt{n}$. Hence, if $\bar{x}$ lies between $\bar{x}$ and $\bar{x}+d\bar{x}$, s_1 between s_1 and s_1+ds_1, the element of volume described by P_1 is constant $\times s_1^{n-2}ds_1d\bar{x}$. Now keeping $\bar{x}$ and s_1 fixed, if we allow s_2, $\bar{y}$ and θ to lie in the intervals (s_2, s_2+ds_2), $(\bar{y}, \bar{y}+d\bar{y})$, $(\theta, \theta+d\theta)$, then the volume element described by P_2' in the x-space and hence P_2 in the y-space is const. $\times(s_2 \sin\theta)^{n-3}ds_2d\bar{y}$, $s_2d\theta$ *i.e.*, const. $s_2^{n-2}\sin^{n-3}\theta ds_2d\bar{y}d\theta$.

Now since P_1 and P_2 are the projections of P on the mutually orthogonal spaces, the volume element described by P as $\bar{x}$, $\bar{y}$, s_1, s_2, θ vary between their infinitesimal ranges is simply the product of these two volume elements and is thus

$$\text{Const.}\quad s_1^{n-2}\, s_2^{n-2}\sin^{n-3}\theta\; ds_1\; ds_2\; d\theta\; d\bar{x}\; d\bar{y}$$

i.e.,
$$\text{Const.}\quad s_1^{n-2}\, s_2^{n-2}(1-r^2)^{\frac{n-4}{2}}\; ds_1\; ds_2\; dr\; d\bar{x}d\bar{y}$$

Hence the joint sampling distribution of $\bar{x}$, $\bar{y}$, s_1, s_2 and r is

$$C_1 \exp\left[\frac{-n}{2(1-\rho^2)}\left\{\frac{(\bar{x}-m_1)^2}{\sigma_1^2}-\frac{2\rho(\bar{x}-m_1)(\bar{y}-m_2)}{\sigma_1\sigma_2}+\frac{(\bar{y}-m_2)^2}{\sigma_2^2}\right\}\right] d\bar{x}d\bar{y}$$

$$\times C_2 \exp\left[\frac{-n}{2(1-\rho^2)}\left(\frac{s_1^2}{\sigma_1^2}-\frac{2\rho r s_1 s_2}{\sigma_1\sigma_2}+\frac{s_2^2}{\sigma_2^2}\right)\right] s_1^{n-2}\; s_2^{n-2}(1-r^2)^{\frac{n-2}{2}}\; ds_1\; ds_2\; dr.$$

This shows that *joint distribution of $\bar{x}$ and $\bar{y}$ and the joint distribution of s_1, s_2 are mutually independent and hence in normal samples, the distributions of means is independent of that of variances and covariances.*

Fourth Step. In order to obtain the sampling distribution of r, we are to integrate out the second factor of the above expression for s_1 and s_2 from 0 to ∞. We shall consider two cases.

Case I. $\rho \neq 0$

Let us set $\xi = \dfrac{s_1 s_2}{\sigma_1\sigma_2}$, $e^z = \dfrac{s_1\sigma_2}{s_2\sigma_1}$

then $\dfrac{s_1^2}{\sigma_1^2}-\dfrac{2\rho r s_1 s_2}{\sigma_1\sigma_2}+\dfrac{s_2^2}{\sigma_2^2} = \xi e^z - 2\rho\xi r + \xi e^{-z} = 2\xi\,(\cosh z - \rho r)$

Also Jacobian of transformation

$$=\frac{\partial(\xi, z)}{\partial(s_1, s_2)}=\begin{vmatrix} \frac{s_2}{\sigma_1\sigma_2} & \frac{s_1}{\sigma_1\sigma_2} \\ \frac{1}{s_1} & -\frac{1}{s_2} \end{vmatrix}=\frac{-2}{\sigma_1\sigma_2}=\text{const.}$$

Hence the joint distribution of r, z and ξ is

$$\text{const. } \exp\left[\frac{-n\xi}{1-\rho^2}\left(\cosh z-\rho r\right)\right]\times\xi^{n-2}(1-r^2)^{\frac{n-4}{2}}\,d\xi\,dz\,dr$$

Thus limits for ξ and z are from 0 to ∞. Integrating out for ξ, we have joint distribution of z and r as

$$\text{const.}\frac{(1-\rho^2)^{\frac{n-1}{2}}}{(\cosh z-\rho r)^{n-1}}(1-r^2)^{\frac{n-4}{2}}\,dz\,dr$$

Hence the distribution of r is

$$\text{const. } (1-r^2)^{\frac{n-4}{2}}\,dr.\int_0^\infty\frac{dz}{(\cosh z-\rho r)^{n-1}}$$

Putting $-\rho r=\cos\theta$, evaluating the constant, distribution of r can be expressed in the form [see Ex. 4, page 420]

$$*dP=\frac{(1-\rho^2)^{\frac{n-1}{2}}}{\pi\,\Gamma(n-2)}(1-r^2)^{\frac{n-4}{2}}\frac{d^{n-2}}{d(r\rho)^{n-2}}\left[\frac{\cos^{-1}(-\rho r)}{\sqrt{1-\rho^2r^2}}\right]dr.$$

Case II. $\rho=0$ (*i.e.*, sample is drawn from an uncorrelated bivariate normal population). In this case the second factor reduces to

$$\text{Cons. exp}\left[-\frac{n}{2}\left(\frac{s_1^2}{\sigma_1^2}+\frac{s_2^2}{\sigma_2^2}\right)\right]s_1^{n-2}s_2^{n-2}(1-r^2)^{\frac{n-4}{2}}\,ds_1\,ds_2\,dr$$

This shows that the distribution of r is independent of the distribution of s_1 and s_2 and is given by

---

*Prof. Harold Hotteling (1953) made an extensive study of the distribution of r. He has obtained the following form of the distribution of r:

$$\frac{n-2\overline{|n-2}}{\sqrt{2\pi}\overline{|n-\frac{1}{2}}}(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}(1-\rho r)^{-n+\frac{3}{2}}F\left(\tfrac{1}{2}, \tfrac{1}{2}, n-\tfrac{1}{2}; \frac{1+\rho r}{2}\right),$$

where $$F(a, b, c\,x)=\sum_{j=0}^{\infty}\frac{\overline{|a+j}}{\overline{|b}}\cdot\frac{\overline{|b+j}}{\overline{|b}}\cdot\frac{\overline{|c}}{\overline{|c+j}}\cdot\frac{x^j}{j!}$$

is a hypergeometric function. Hotteling discusses methods of integrating the density function and also calculates moments of r.

$$dp = c(1-r^2)^{\frac{n-4}{2}}\, dr, \quad -1 \leqslant r \leqslant 1 \qquad \ldots(14{\cdot}7)$$

where $1 = \int dp = c\int_0^1 (r^2)^{-\frac{1}{2}}(1-r^2)^{\frac{n-4}{2}}\, d(r^2)$, *i.e.* $c = \dfrac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)}$.

14·7·1. Alternative Method of Derivation of the Distribution of r when $\rho=0$. In the case of an uncorrelated population *i.e.*, when $\rho=0$, the distribution of the correlation coefficient, r, can be obtained easily as follows :

Let the n values $y_i (i=1, 2.....n)$ be subjected to an orthogonal transformation yielding n independent variates $\eta_1, \eta_2, ..., \eta_n$ so that

$$\sum_{i=1}^{n} \eta_i^2 = \sum_{i=1}^{n} y_i^2.$$

Choosing $\eta_1 = \dfrac{1}{\sqrt{n}} . \sum_i y_i = \sqrt{n}\, \bar{y}$ (the sum of the squares of the coefficients of y_i's is unity), we get

$$\sum_{i=1}^{n} \eta_i^2 = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + n\bar{y}^2 = ns_2^2 + \eta_1^2$$

or $$ns_2^2 = \sum_{i=2}^{n} \eta_i^2$$

Also $$\sqrt{n}\, r s_3 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{n}\, s_1} = \frac{\Sigma(x_i - \bar{x})y_i}{\sqrt{n}\, s_1}$$

Since the sum of the squares of the coefficient of y_i is unity, we may take

$$\eta_2 = \frac{\Sigma(x_i - \bar{x})y_i}{\sqrt{n}\, s_1} = \sqrt{n}\, r\, s_2$$

Henc $\sum_{i=3}^{n} \eta_i^2 = n(1-r^2)s_2^2$.

Since η_i/σ^2 are independent standard normal variates we conclude that $\dfrac{nr^2 s_2^2}{\sigma_2^2}$ and $\dfrac{n(1-r^2)s_2^2}{\sigma_2^2}$ are distributed independently like χ^2 with 1 and $(n-2)$ *d. f.* respectively.

Now $r^2 = \dfrac{r^2 s_2^2}{s_2^2} = \dfrac{nr^2 s_2^2/\sigma_2^2}{n^2 s_2^2/\sigma_2^2 + n(1-r^2)s_2^2/\sigma_2^2} = \dfrac{\chi_1^2}{\chi_1^2 + \chi_2^2}$

where χ_1^2 and χ_2^2 are distributed like χ^2 with 1 and $(n-2)$ *d. f.* respectively and, therefore, by Ex. 13·1, page 380, r^2 is a $\beta_1\left[\frac{1}{2}, \frac{n-2}{2}\right]$ variate.

Hence the distribution of r^2 is

$$dp=\frac{(r^2)^{\frac{1}{2}-n}(1-r)^{\frac{n-4}{2}}}{B\left[\frac{1}{2},\ \frac{n-2}{2}\right]}d(r^2),\ 0\leqslant r^2\leqslant 1$$

Thus the distribution of r is

$$dp=\frac{1}{B\left(\frac{1}{2},\ \frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}dr,\ -1\leqslant r\leqslant 1.$$

Ex. 1. Show that $\bar{x},\bar{y}$ are distributed in a bivariate normal distribution with means m_1, m_2 ; variances σ_1^2/n_1, σ_2^2/n_2 and correlation coefficient ρ.

Ex. 2. Show that, if $\rho=0$, $\bar{x}$ and $\bar{y}$ are independently distributed and so are s_1^2 and s_2^2.

Ex. 3. Show that when $\rho=0$, $t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ is distributed as t with $(n-2)$ $d.f$

Ex. 4. Show that $\int_0^\infty \frac{dz}{\cosh z+\cos\theta}=\frac{\theta}{\sin\theta}$

Deduce that $\int_0^\infty \frac{d^{n-2}}{d(\cos\theta)^{n-2}}\left(\frac{\theta}{\sin\theta}\right)=\int_0^\infty \frac{dz}{(\cosh z+\cos\theta)^{n-1}}$.

Ex. 5. Show that in samples of size 4 from a bivariate formula population for which $\rho=0$, r is distributed as a rectangular variate.

Ex. 6. Show that if $\rho=0$, the $2m$th moment about origin of the distribution of r, is

$$\mu'_{2m}=\frac{\left|\frac{n-1}{2}\right|\left|m+\frac{1}{2}\right|}{\sqrt{\pi}\left|\frac{n-1}{2}+m\right|}$$

14·8. Fisher's z Transformation. The ordinates and the density function of the correlation coefficient are not expressible in terms of simple mathematical functions and have been tabulated by F.N. David (1938).

The distribution of r is far from normal and its probability curve is very skew in the neighbourhood of $\rho=\pm 1$ even for large n. We can not, therefore, use the standard error. Fisher (1921), however, showed that for the variate $z-\xi$ where

$$z=\tfrac{1}{2}\log_e\left[\frac{1+r}{1-r}\right],\ r=\tanh z$$

$$\xi=\tfrac{1}{2}\log_e\left[\frac{1+\rho}{1-\rho}\right],\ \rho=\tanh\xi$$

we have,

$$\mu_1' = \frac{\rho}{n-1} + \frac{5\rho+\rho^3}{4(n-1)^2}, \quad \gamma_1 = \frac{\rho}{(n-1)^{3/2}}\left[\rho^2 - \frac{9}{16}\right] + \ldots$$

$$\mu_2 = \frac{1}{n-1} + \frac{4-\rho^2}{(n-1)^2} + \ldots\ldots, \quad \gamma_2 = \frac{32-3\rho^4}{16(n-1)} + \ldots$$

so that $z-\xi$ is approximately normally distributed with mean zero and variance which is approximately $1/(n-3)$. For $n>50$, this approximation is quite good, but can be used even for smaller values of n for many practical purposes.

14·8·1. Fisher's z transformation when $\rho=0$. In this case, the distribution of z, by putting $r=\tanh z$ in (14·7), is

$$dp = \frac{\operatorname{sech}^{n-2} z\, dz}{\mathcal{B}\left[\frac{1}{2}, \frac{n-2}{2}\right]} \propto \exp\left[-\tfrac{1}{2}(n-2)z^2\right]dz \text{ approximately.}$$

Thus the distribution of z is approximately normal with variance $1/(n-2)$. Fisher has shown, however, that a better approximation to the variance is $1/(n-3)$.

14·9. Distribution of regression coefficients in normal samples.

Putting $b_2 = \dfrac{r\, s^2}{s_1}$ in the second factor of (14·7), we have the joint distribution of s_1, s_2 and b_2.

$$dF = c_2 \exp\left[\frac{-n}{2(1-\rho^2)}\left\{\frac{s_1^2}{\sigma_1^2} - \frac{2\rho s_1^2 b_2}{\sigma_1\sigma_2} + \frac{s_2^2}{\sigma_2^2}\right\}\right] s_1^{n-1} s_2^{n-3} \times \left(1 - \frac{s_1^2 b_2^2}{s_2^2}\right)^{(\frac{1}{2}n-4)} ds_1 ds_2 db_2$$

To obtain the distribution of s_1 and b_2, we integrate *w.r.t.* s_2

$$\therefore \quad dF = \text{const.} \exp\left[\frac{-n}{2(1-\rho^2)}\left\{\frac{s_1^2}{\sigma_1^2} - \frac{2\rho s_1^2 b_2}{\sigma_1\sigma_2} + \frac{s_1^2 b_2^2}{\sigma_2^2}\right\}\right] s_1 ds_1 db_2$$

Integrating *w.r.* to s_1, we obtain the distribution of b_2

$$dF = \text{const.}\left(1 - \frac{2\rho\sigma_1}{\sigma_2} b_2 + \frac{\sigma_1^2}{\sigma_2^2} b_2^2\right)^{-\frac{1}{2}n} db_2$$

$$\frac{\Gamma(n/2)\sigma_2^{n-1}(1-\rho^2)^{\frac{1}{2}(n-1)}}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)\sigma_1^{n-1}} \cdot \frac{db_2}{\left\{\frac{\sigma_2^2}{\sigma_1^2}(1-\rho^2) + \left(b_2 - \frac{\rho}{\sigma_1}\right)\right\}^{\frac{1}{2}n}}$$

The distribution of b_2 is symmetrical about the point $\beta_2 = \rho\sigma_2/\sigma_1$, the population regression coefficient. It tends to normality fairly rapidly and the use of the standard error for regression is, therefore,

valid for lower values of n than in the case of the correlation coefficient.

Putting $u=\dfrac{(b_2-\beta_2)\,s_1}{(\quad)}$, we can show that the distribution of u is $dF = \text{const.}\dfrac{du}{(1+u^2)^{\frac{1}{2}(n-1)}}$

It follows that $t=u\sqrt{n-2}=\dfrac{(b_2-\beta_2)s_1\sqrt{n-2}}{(s_2^2-b_2^2s_1^2)}$

is distributed in Student's t form with $(n-2)$ $d.f.$

EXERCISES 14

1. A transformation rectangular to polar coordinates is

$$y_1=u\sin\theta_1,$$
$$y_2=u\cos\theta_1\sin\theta_2,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y_{n-1}=u\cos\theta_1\cos\theta_2\ldots\cos\theta_{n-2}\sin\theta_{n-1},$$
$$y_n=u\cos\theta_1\cos\theta_2\ldots\cos\theta_{n-2}\cos\theta_{n-1},$$

where $-\dfrac{\pi}{2} < \theta_i \leqslant \dfrac{\pi}{2}$, $i=1, 2, \ldots\ldots, n-2$ and $\pi < \theta_{n-1} \leqslant \pi$

Prove that (*a*) $u^2=\sum_1^n y_i^2$

(*b*) Jacobian of transformation is

$$u^{n-1}\cos^{n-2}\theta_1\cos^{n-3}\theta_2\ldots\ldots\cos\theta_{n-2}.$$

Deduce that the surface of a sphere of unit radius in n dimensions is

$$(2\pi)^{\frac{n}{2}}\Big/\left|\overline{\frac{n}{2}}\right.$$

2. Derive the distribution of section 14·2 by using the transformations

$$x_1^2+x_2^2+\ldots\ldots+x_n^2=u_1$$
$$x_1^2+x_2^2+\ldots\ldots+x_{n-1}^2=u_1u_2$$
$$x_1^2+x_2^2+\ldots\ldots+x_{n-2}^2=u_1u_2u_3$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x_1^2 \qquad\qquad\qquad\qquad =u_1u_2u_3\ldots u_n$$

and integrating over $u_2, u_3\ldots\ldots, u_n$ from 0 to 1 and over u_1 from 0 to z.

3. If a variable possesses a continuous density function and if z denotes the proportion of the population, that lies between the extreme values of a random sample of size n from the population, then the density function of z is given by

$$f(z)=n(n-1)z^{n-2}(1-z).$$

4. Show that the distribution of the range in samples of n from rectangular population $dF=\frac{1}{2}dx$, $-1\leqslant x\leqslant 1$ is given by

$$\phi(w)=\frac{n(n-1)}{2n}\,w^{n-2}(2-w) \qquad (M.A., \textit{Bombay})$$

5. Show that the distribution of range in samples of n from the population

$$dF=2(1-x)dx, \qquad 0\leqslant x\leqslant 1$$

is $\phi(w)=\dfrac{n(n-1)w^{n-2}}{2}\left[\dfrac{2w^{n+1}}{n^2-1}+\dfrac{(2-w)^{n+1}}{n+1}-\dfrac{w^2(2-w)^{n-1}}{n-1}\right]$.

6. Show that the distribution of the range w in samples of three from a normal population of unit variance is given by

$$\phi(w)=\frac{6}{\pi\sqrt{2}}\int_0^{w^2/\sqrt{6}} e^{-\frac{1}{4}w^2}\, e^{-\frac{1}{2}z^2}\, dz.$$

(M.A., Bombay, 1952)

7. If two independent variates x and y have the density functions $f_1(x)$ and $f_2(y)$ respectively, show that the density function for the ratio $v=\dfrac{x}{y}$ where v is non-negative is given by

$$f(v)=\int_{-\infty}^{\infty} yf_1(vy)f_2(y)dy.$$

Use this formula to derive the distribution of t and F.

8. Show that in samples of n from the distribution

$$dF=e^{-x}dx,\ 0\leqslant x\leqslant \infty$$

the variance of the smallest sample value is $1/n^2$.

9. Use the method of characteristic function to show that the distribution of the sample mean $\bar{x}$ from Pearson Type III population

$$dp=\frac{1}{\Gamma l}\, e^{-\frac{x}{a}}\left(\frac{x}{a}\right)^{l-1}\frac{dx}{a},\ a>0,\ x\geqslant 0$$

is given by $\quad dp=\dfrac{1}{\Gamma(nl)}\, e^{-\frac{nx}{a}}\left(\dfrac{nx}{a}\right)^{nl-1}\dfrac{n}{a}\, d\bar{x},$

Hence or otherwise obtain Var $(\bar{x})$. *(M.A. Madras 1959)*

10. If x, y, z have the joint density

$$f(x, y, z)=\frac{6}{(1+x+y+z)^3},\ x, y, z>0.$$

find the distribution of $u=x+y+z$. [Ans. $dP=\dfrac{3u^2}{(1+u)^4}\,du$]

[Hint. Put $u=x+y+z$, $uv=y+z$, $uvw=z$ and use that $\left|J\left(\dfrac{x, y, z}{u, v, w}\right)\right|=u^2v$]

11. (*a*) Use the method of Induction to derive the χ^2 distribution.

(*b*) By using characteristic function obtain the distribution of the mean of normal and rectangular distributions the rectangular distribution having the range 0 to 1. *(I.C.A.R. Delhi 1948)*

12. Obtain the characteristic functions of the sampling distribution of the mean in samples of size n from the rectangular population

$$dF=dx,\ 0\leqslant x\leqslant 1$$

Show that $\beta_1=0,\ \beta_2=3-\frac{6}{5n}$.

(M.A., Delhi, 1958 ; M.A., Bombay, 1953)

13. (*a*) Show that the charactertistic function for $y=\log x$ is

$$\phi(t)=\int_{-\infty}^{\infty} e^{it \log x} f(x)dx=\int_{-\infty}^{\infty} x^{it} f(x)dx$$

and the characteristic function of $\log G$ where G is geometric mean of a random sample of size n is $\left[\phi\left(\frac{t}{n}\right)\right]^n$. Hence deduce that the density function for G, by inversion theorem, is

$$f(y)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it \log G}\left[\phi\left(\frac{t}{n}\right)\right]^n dt.$$

(*b*) If $x_1, x_2 \ldots\ldots, x_n$ is a random sample of size n drawn from the rectanuglar population

$$dF=dx,\ 0\leqslant x<1$$

show that the density function of the geometric mean of the sample is

$$f(y)=\frac{n^n y^{n-1}}{\lfloor n}(-\log y)^{n-1},\ 0\leqslant y\leqslant 1.$$

14. Write a brief explanatory note on transformation of variates. Show that if X is an exponential variate, then e^{-x} is a rectangular variate.

Also prove that the sum of a number of independent identical exponential variates is a Chi-square variate and deduce the additive property of Chi-square variates. *(M.A. Agra, 1967)*

15. A frequency function $f(x)$ has derivative $f'(x)$ and a single mode at $x=m$ such that $f'(x)>a$ or $<a$ according as $x<m$ or $>m$. Show that for a given β the range from u to $u+v$ determined by $\beta=\int_u^{u+v} f(x)dx$ more is shortest when $f(u)=f(u+v),\ u<m<u+v$. *(B.A. Hons. Calcutta)*

15

Tests of Significance Based on t, F and z Distributions

15·1. Introduction. In Chapter 12 the following tests of significance were discussed :

(*a*) In a *large* number n of independent Bernoullian trials, x successes are obtained, to test the hypothesis that the constant probability of success in each trial is p. [§ 12·3].

(*b*) Two independent sets of n_1, n_2 (*large*) independent Bernoullian trials are performed and x_1, x_2 successes respectively are obtained, to test the hypothesis that the constant probability of success is the same for both sets. [§ 12·4]

(*c*) Given a random sample $x_1, x_2, \ldots\ldots, x_n$ of *large size* n from a *normal* population with *known* variance σ^2, to test the hypothesis that the population mean is μ. [§ 12·7]

(*d*) Given a random sample $x_1, x_2, \ldots\ldots x_n$ of *large size* n from a population (which is not necessarily normal) with a finite (but unknown) variance, to test the hypothesis that the population mean is μ. [§12·8]

(*e*) Given two independent random samples of *any sizes* n_1, n_2 from *normal* populations with *known* variances σ_1^2, to test the hypothesis that the means of the populations are the same. [§12·8]

(*f*) Given two independent random samples of large sizes n_1, n_2, from two populations (not necessarily normal) with unknown but finite variance, to test the hypothesis that the population means are the same. [§ 12·8].

It will be noted that tests in cases (*c*) and (*e*) were applicable to samples of any size since we were able to find the *exact* sampling dis-

tributions of the sample mean and of the difference between the sample means. On the other hand, the other four tests were applicable to large samples only and were based on the fact that in each case we could find statistics which were normally distributed for large samples. The asymptotically standard normal statistics were :

$$z_1 = \frac{x - np}{\sqrt{npq}} = \frac{\frac{x}{n} - p}{\sqrt{\frac{pq}{n}}} \qquad \text{for case } (a)$$

$$z_2 = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{for case } (b)$$

where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$, $q = 1 - \bar{p}$

$$z_3 = \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \qquad \text{for case } (d)$$

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

$$z_4 = \frac{\bar{x}_1 - \bar{x}_2}{s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{for case } (f)$$

where $s^2 = \left\{ \sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 + \sum_{j=1}^{n_2} (x_{j2} - \bar{x}_2)^2 \right\} / (n_1 + n_2)$

To develop tests of significance applicable to all sample sizes we have to use the exact sampling distributions of those or other suitable statistics. We have accordingly in Chapters 13 and 14 obtained the exact sampling distribution of t (due to Student) and of F (due to Fisher) and of other statistics like z, s^2, χ^2, etc., to carry out different tests of significance for samples of all sizes. We discuss tests of significance based on exact sampling distributions of t, F and r in the present chapter. Tests based on χ^2 distribution will be discussed in Chapter 16.

For all the tests developed below, it is assumed that the parent population is normal. Also while the tests in the next two sections are applicable to all sample sizes, they are, in practice, applied to samples of small or moderate sizes only, since for large sample the tests based on the normal approximation are easier and quicker to apply and give fairly satisfactory results.

15·2. Tests of significance based on t distribution. We consider five tests of significance based on the t distribution.

15·2·1. Testing the significance of the mean of a random sample from a normal population. Given a random sample $x_1, x_2, \ldots\ldots, x_n$ from a normal population we have to test the hypothesis that the mean of the population is μ. For this purpose we assume the hypothesis to be correct and then see whether the data is consistent with the hypothesis or whether it provides some evidence against the hypothesis. To carry out the test we calculate the statistic

$$t = \frac{(\bar{x}-\mu)\sqrt{n}}{S} \qquad \ldots(15\cdot1)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

t-distribution has been derived in § 13·5 and a table based on it is given on page 386. For different values of the degrees of freedom $\nu(=n-1)$ the table gives the values of $t_{0\cdot05}$ and $t_{0\cdot01}$, which are defined by the relations

$$P[\,|\,t\,|\,\geqslant t_{0\cdot05}] = \cdot05$$
$$P[\,|\,t\,|\,\geqslant t_{0\cdot01}] = \cdot01.$$

If our calculated value of $|\,t\,|$ exceeds $t_{0\cdot05}$, we say that the difference between $\bar{x}$ and μ is significant at 5% level of significance; if it exceeds $t_{0\cdot01}$, the difference is said to be significant at 1% level or just highly significant. If $|\,t\,| < t_{0\cdot05}$, we say that the data is consistent with the hypothesis that μ is the mean of the population.

Example 15·1. *Show that 95% fiducial limits for the mean μ of the population are $\bar{x} \mp St_{0\,05}/\sqrt{n}$. Deduce that for a random sample of 16 values with mean 41 5 inches and the sum of the squares of the deviations from the mean 135 (inches)^2 and drawn from a normal population 95% fiducial limits for the mean of population are 39·9 and 43·1 inches.*

We have $P\left[\left|\frac{(\bar{x}-\mu)\sqrt{n}}{S}\right| \leqslant t_{0\cdot05}\right] = \cdot95$

or $$P\left[\bar{x} - \frac{St_{0\,05}}{\sqrt{n}} \leqslant \mu \leqslant \bar{x} + \frac{St_{0\cdot05}}{\sqrt{n}}\right] = \cdot95$$

so that we can say with a confidence coefficient ·95 that the confidence interval $\bar{x} \mp \frac{S}{\sqrt{n}} t_{0\,05}$ contains the population mean μ. The

limits of this confidence interval give us the 95% confidence or fiducial limits for μ.

In the particular case,

$$n=16,\ \nu=n-1=15,\ S=\sqrt{\tfrac{1}{15}\times 135}=3$$

Also from the table, $t_{0\cdot05}=2\cdot13$

$$\therefore\quad \frac{\sqrt{n}}{S}\,t_{0\cdot05}=\tfrac{3}{4}\times 2\cdot13=1\cdot6 \text{ approx.}$$

Therefore, the required fiducial limits are $41\cdot5\mp1\cdot6$, *i.e.*, 39·9 and 41·3 inches.

Example 15·2. *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, 71 and 71 in the light of these data, discuss the suggestion that the mean height in the population is 66 inches.*

(M.Sc. Agra, 1954)

Here $\bar{x}$ = sample mean = 67·8 inches

$$S^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\frac{1}{n-1}\left[\sum_{i=1}^{n}d_i'^2-\frac{(\Sigma d_i')^2}{n}\right.$$

where d_i' is the deviation from the assumed mean, say 67·8.

$$\therefore\quad S=\sqrt{\tfrac{1}{9}(8-6\cdot4)}=3\cdot011 \text{ inches}$$

$$=\frac{67\cdot8-66}{3\cdot011/\sqrt{10}}=1\cdot89 \text{ approx.}$$

Also v = number of degrees of freedom $=n-1=9$.

For 9 *d.f.*, we get from the table, $t_{0\cdot05}=2\cdot26$.

As the calculated value of $|t|$ is less than $t_{0\cdot05}$, the difference between $\bar{x}$ and μ may be due to fluctuations of random sampling. In other words the data does not provide any significant evidence against the hypothesis that the population mean is 66".

Ex. 1. Obtain the 99% fiducial limits in Example 15·1.

Ex. 2. Obtain the 95% fiducial limits in Example 15·2.

Ex. 3. For a given sample show that the length of the 'symmetrical' confidence interval increases with the confidence coefficient. Comment on this result.

Ex. 4. Show that the length of the evidence interval $\frac{2S}{\sqrt{n}}\,t_{0\cdot05}$ need not necessarily decrease with n. Show that the expected value of its square in random sampling is $\frac{4\sigma^2}{n}(t_{0\cdot05})^2$ and definitely decreases with n.

15·2·2. Testing the significance of difference between two sample means. Here we consider three tests of significance.

(*a*) Given two independent random samples $x_{11}, x_{12}, \ldots\ldots x_{1n_1}$ and $x_{21}, x_{22} \ldots x_{2n_2}$, with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations S_1 and S_2 from normal populations with *the same variance*, we have to test the hypothesis that the population means are the same. In other words, since a normal distribution is completely specified by its mean and variance, we have to test the hypothesis that the two independent samples come from the *same* normal population.

To carry out the test, we calculate the statistic

$$t = \frac{\bar{x}_1 \sim \bar{x}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \qquad \ldots(15{\cdot}2)$$

where $$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i},\ \bar{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \bar{x}_{2j},$$

and $$S^2 = \frac{1}{n_1+n_2-2}[(n_1-1)S_1^2+(n_2-1)S_2^2]$$

$$= \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2+\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2\right]$$

It has been shown in § 13·6 that this statistic follows the t-distribution with n_1+n_2-2 degrees of freedom. If the calculated value of $|t|$ be $> t_{0\cdot05}$ $(t_{0\cdot01})$, the difference between the sample means is said to be significant at 5% (1%) level of significance; otherwise the data are said to be consistent with the hypothesis.

Note. If the *size of the two samples is the same* say equal to n, and the data are paired [see (*b*), page 431] when we calculate the statistic t defined by

$$t = \frac{\bar{d}-0}{\left(\frac{S}{\sqrt{n}}\right)} \qquad \ldots(15{\cdot}3)$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i-\bar{d})^2$, d_i = difference between the ith members of the samples and $\bar{d}$ = mean of the difference $= \Sigma d_i/n$, and the number of *d.f.* on which t is based $= n-1$. For non-paired data, we continue to use (15·2).

Example 15·3. *For a random sample of 10 pigs fed on diet A, the increases in weight in pounds in a certain period were*

10, 6, 16, 17, 13, 12, 8, 14, 15, 9 *lbs.*

For another random sample of 12 *pigs, fed on diet B, the increases in the same period were*

7, 13, 22, 15, 12, 14, 18, 8, 21, 23 10, 17 *lbs.*

Test whether diets A and B differ significantly as regard the effect on increases in weight (or test whether the mean increases in the two samples are significantly different). You may use the fact that 5% value of t for 20 degrees of freedom is 2·09.

(M.A., Bombay 1958 ; B.A. Hons., Delhi, 1951, 57, 71 ; I.A.S., 1950)

We first calculate the mean and standard deviations for the two samples.

Mean gain in weight on diet $A = \bar{x}_1 = \frac{\Sigma x_{1i}}{n_1} = \frac{120}{10} = 12$ pounds.

Mean gain in weight on diet $B = \bar{x}_2 = \frac{\Sigma x_{2j}}{n_2} = \frac{180}{12} = 15$ pounds.

The difference in the sample means = 3 pounds.

Again $S = \sqrt{[\sum_i (x_{1i} - \bar{x}_1)^2 + \sum_j (x_{2j} - \bar{x}_2)^2]/(n_1 + n_2 - 2)}$

$$= \sqrt{\frac{120 + 314}{10 + 12 - 2}} = \sqrt{\frac{434}{20}} = \sqrt{21{\cdot}1} = 4{\cdot}65.$$

Diet A			*Diet B*		
Gain in weight (x_{1i})	*Deviation from mean* $(x_{1i} - \bar{x}_1)$	*Square of deviation* $(x_{1i} - \bar{x}_1)^2$	*Gain in weight* x_{2j}	*Deviation from mean* $(x_{2j} - \bar{x}_2)$	*Square of deviation* $(x_{2j} - \bar{x}_2)^2$
10	−2	4	7	−8	64
6	−6	36	13	−2	4
16	+4	16	22	+7	49
17	+5	25	15	0	0
13	+1	1	12	−3	9
12	0	0	14	−1	1
8	−4	16	18	+3	9
14	+2	4	8	−7	49
15	+3	9	21	+6	36
9	−3	9	23	+8	64
			10	−5	25
			17	+2	4
120	0	120	180		314

Substituting these values in the formula (15·2) we get

$$|t| = \frac{3}{4{\cdot}65\sqrt{\frac{1}{10}+\frac{1}{12}}} = 1{\cdot}6.$$

The number of degrees of freedom $= \nu = n_1 + n_2 - 2 = 20$.

For 20 *d. f.* we are given that $t_{0{\cdot}05} = 2{\cdot}09$.

The calculated value of $|t|$ is less than $t_{0{\cdot}04}$ and hence the difference between the sample means is not significant.

(*b*) In the above example the difference in the increases in weight are regarded as due to differences in diets *A* and *B*, apart from the random differences due to sampling but it is possible that the pigs in the two samples may differ in their ages, sexes, initial weights, earlier nutritional standards etc., and part of the differences in increases in weights may be due to these non-random factors. To get a more precise test the data are sometimes paired. Thus in the above example instead of taking 22 pigs at random, we may choose 11 pairs, each pair being as homogeneous as possible *e g.*, a pair may be taken from a litter. Similarly in the case of dieting experiments on children we may use twins. If we take n pairs, the variate to be considered is the difference between the values for a pair and the hypothesis to be tested is that the samples of these differences comes from a normal population with mean zero so that t test with $(n-1)$ *d.f.* would be required.

Example 15·4. *The yields of two types 'Type 17' and 'Type 51' of grains in pounds per acre in 6 replications are given below. What comments would you make on the differences in the mean yields ? You may assume that if there be 5 degrees of freedom and $P = 0{\cdot}2$, t is approximately 1·476.*

Replication	*Yield in pounds* Type 17	*Yield in pounds* Type 51
1	20·50	24·86
2	24·60	26·39
3	23·06	28·19
4	29·98	30·75
5	30·37	29·97
6	23·83	22·04

(*I.A.S., 1951*)

We calculate the mean and standard deviation of the differences in the yields of the two types of grains.

Replication	*Yield in pounds* Type 17	*Yield in pounds* Type 51	*Difference* d	*Deviation from mean* $d-\bar{d}$	*Square of deviation* $(d-\bar{d})^2$
1	20·50	24·86	4·36	+2·717	7·38
2	24·60	26·39	1·79	+0·147	0·02
3	23·06	28·19	5·13	+3·487	12·15
4	29·98	30·75	0·77	−0·873	0·76
5	30·37	29·97	−0·40	−2·043	4·17
6	23·83	22·04	−1·79	−3·433	11·79
			9·86		36 27

$\bar{d}$ = mean difference of the yields $= \frac{\Sigma d}{n} = \frac{9{\cdot}86}{6} = 1{\cdot}643$ pounds

$S = S.D.$ of difference of yields $= \sqrt{\frac{(\Sigma d - \bar{d})^2}{n-1}}$

$= \sqrt{\frac{36{\cdot}7}{5}} = 2{\cdot}69$ pounds.

We set up the null hypothesis that the difference in types has no effect on yields, *i.e.*, the population mean of the difference is zero then

$$t = \frac{1{\cdot}643 - 0}{2{\cdot}69}\sqrt{6} = 1{\cdot}489.$$

This value of t is less than $t_{0{\cdot}05}$ for 5 $d.f.$ and, therefore, the difference is not significant at 5% level but is significant at 20% level.

(*c*) Given two independent random samples from normal populations with the same variance we have to test the hypothesis that the populations are μ_1 and μ_2 respectively.

For this case

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \ldots(15.4)$$

with $(n_1 + n_2 - 2)$ $d.f.$ and the test of significance is carried out as in case (*a*).

Note about assumptions for the t test.

For testing the significance of a sample mean, t test has a greater range of validity and precision than the normal approximation test of § 12·6, but this is achieved by placing an important restriction *viz.*, that the parent population must be normal. If the parent population is not normal the distribution of the statistics $\sqrt{n}(\bar{x}-\mu)/S$ may be quite different from the t-distribution since, in that case, the distribution of $\bar{x}$ and s^2 will not be independent and also the distribution will depend on parameters giving the departure of the parent distribution from normality.

For testing the significance of the difference between two sample means, we make an additional assumption that the variances of the two populations are the same. Before applying the t test it may be desirable to test this assumption by applying the F test as discussed in §15·2·1. If the two variances are different, t-test is no longer valid and we can apply* Behren 'd' test based on fiducial intervals. We put

$$d=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{S_1^2+S_2^2}}$$

then we carry out the test of significance by using the tables of Sukhatme and Fisher for various values of n_1, n_2 and and S_1^2/S_2^2, the values of d corresponding to various probability levels.

Ex. 1. Show that the 95% fiducial limits for the difference of means of the populations are $(\bar{x}_1-x_2)\pm St_{0\cdot05}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ where $t_{0\cdot05}$ is the 5% value of t for (n_1+n_2-2) *d.f.* Deduce that if zero lies outside this range for $\mu-\mu'$ the difference between the sample means is significant at 5% level.

Ex. 2. Eleven school boys were given a test in Drawing. They were given a month's further tuition and a second test of equal difficulty was held at the end of it. Do the marks give evidence that students have benefited by the extra coaching ?

Roll No.	1	2	3	4	5	6	7	8	9	10	11
Marks, 1st test	23	20	19	21	18	20	18	17	23	16	19
Marks, 2nd test	24	19	22	18	20	22	20	20	23	20	17

[**Hint.** $d=1$, $S=2\cdot24$, $t=1\cdot482$, $\nu=10$, $t_{0\cdot05}=2\cdot228$, difference is not significant. Hence the students do not seem to gain by extra-coaching].

15·2·3. Testing the significance of an observed correlation coefficient. Given a random sample (x_1, y_1), $(x_2, y_2),\ldots,(x_n, y_n)$, from a bivariate normal population we have to test the hypothesis that the correction coefficient of the population is zero.

*See Statistical Inference by H.C. Saxena.

If the hypothesis is true, it was shown in §14·7·1 that the statistic

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \ldots(15\cdot5)$$

is a t variate with $(n-2)$ *d. f.* If our calculated value of $|t|$ exceeds $t_{0\cdot05}$ for $(n-2)$ *d.f.* we say that the value of r is significant at 5% level of significance. If $|t|<t_{0\cdot05}$ the sample is consistent with the hypothesis of an uncorrrelated population.

Example 15·5. *A random sample of 15 from a normal universe gives a correlation coefficient of $-0\cdot5$. Is this significant of the existence of correlation in the population ?*

Here $n=15, \qquad r=-0\cdot5,$

$$|t|=\frac{0\cdot5\sqrt{15-2}}{\sqrt{1-(-0\cdot5)^2}}=2\cdot07.$$

Also for 13 *d. f.*, $t_{0\cdot05}=2\cdot16$.

$\therefore$ Sample correlation coefficient is not significant, to warrant the existence of a correlation in the population.

Example 15·6. *Find the least value of r in a sample of 18 pairs from a bivariate normal population, significant at 5% level.*

We have $$t=\frac{r\sqrt{18-2}}{\sqrt{1-r^2}}=\frac{4r}{\sqrt{1-r^2}}$$

Also for 16 *d. f.*, $t_{0\cdot05}=2\cdot06$

$\therefore$ For significance at 5% level, we should have

$$\left|\frac{4r}{\sqrt{1-r^2}}\right|>2\cdot06$$

This gives $|r|>\cdot47$ approximately.

Thus ·47 is the least value of $|r|$ required.

Ex. 1. Is a value of $r=0\cdot5$ significant if obtained from a random sample of 11 pairs of values from a normal population ?

Ex. 2. Verify the following table giving the least value of $|r|$ that are significant at 5% level for sample of size n.

n	6	10	15	20	27	42	52	82	102
$\|r\|$	·811	·032	·514	·444	·381	·384	·273	·217	·195

15·2·4. Testing the significance of an observed partial correlation coefficient. Fisher has shown that the sampling distribution of a partial correlation coefficient of order k (*i.e.*, with k secondary subscripts) is of the same form as that of correlation coefficient from a bivariate normal population with the sample size n

reduced by k. In particular, if we are given a random sample from a multivariate normal distribution and we have to test the hypothesis that a particular partial correlation coefficient of order k in the population is zero, we make use of the fact that

$$t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-k-2} \qquad \ldots(15\cdot6)$$

is a t variate with $(n-k-2)$ *d. f.*

Example 15·7. *Show that in a sample of size 20 from a normal population, a correlation coefficient* $r_{12\cdot34}=0\cdot5$ *is significant at 5% level.*

Here $r=0\cdot5, \quad n=20, \quad k=2.$

$$\therefore \quad t=\frac{0\cdot5}{\sqrt{1-\cdot25}}\sqrt{16}=\frac{2}{\sqrt{\cdot75}}=\frac{4\sqrt{3}}{3}=\frac{6\cdot928}{3}$$

$=2\cdot31$ approx.

As for 16 *d. f.*, $t_{0\cdot05}=2\cdot12$.

Also $|t|<t_{0\cdot06}$, the value of $r_{12\cdot34}$ is significant at 5% level.

Ex. Are the following values of r significant at 5% level.

(*i*) $r_{12\cdot345}=0\cdot40$ for $n=21$? Ans. No.

(*ii*) $r_{12\cdot345}=0\cdot457$ for $n=32$? Ans. Yes.

15·2·5. Testing the significance of an observed regression coefficient. We are given a random sample (x_1, y_1), (x_2, y_2),... (x_n, y_n) from a bivariate normal population. Let the regression equation of y on x obtained from the sample be

$$y-\bar{y}=b(x-\bar{x})$$

so that the estimated value of y corresponding to a given x_i is

$$Y_i=\bar{y}+b(x_i-\bar{x}).$$

We have to test the hypothesis that the regression coefficient of y on x in the population is β.

If the hypothesis is correct. We have shown in § 14·2 page 404 that the statistic

$$t=(b-\beta)\,[(n-2)\sum_i (x_i-\bar{x})^2/\Sigma(y_i-Y_i)^2]^{\frac{1}{2}} \qquad \ldots(15\cdot7)$$

conforms to t distribution with $(n-2)$ *d. f.*

Example 15·8. *For a sample of size 30 where* x *takes the values 1, 2, 3,..., 30 it is found that* $\Sigma(x_i-\bar{x})(y_i-\bar{y})=599\cdot62$ *and* $\Sigma(y_i-\bar{y})^2=1020\cdot6$. *Test the significance of regression coefficient of* y *on* x.

Here $\Sigma(x_i-\bar{x})^2=\frac{n(n^2-1)}{12}=\frac{30(899)}{12}=2247\cdot5$

$$b=\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}=\frac{599\cdot62}{2247\cdot5}=0\cdot2668.$$

Also $$\Sigma(y-Y_i)^2=\Sigma(y_i-\bar{y})^2-b^2\Sigma(x_i-\bar{x})^2$$
$$=1020{\cdot}6-159{\cdot}97=860{\cdot}59$$

Also the hypothesis gives $\beta=0$

Substituting in (15·7)

$$t=2{\cdot}28 \text{ with } \nu=28$$

This value of t is significant at 5% level but not at 1% level.

Ex. Show that 95% fiducial limits for the regression coefficient are

$$b\pm t_{0{\cdot}05}[(n-2)\Sigma(x_i-\bar{x})^2/\Sigma(y_i-Y_i)^2]^{-\frac{1}{2}}$$

where $t_{0{\cdot}05}$ is based on $n-2$ *d. f.*

15·2·6. Testing the significance of a rank correlation coefficient. For samples of moderate size it can be shown that if r is the rank correlation coefficient from a sample of size n, and the hypothesis that the population rank correlation coefficient is zero is correct then

$$t=r\left(\frac{n-2}{1-r^2}\right)^{1/2}$$

conforms to the t distribution with $(n-2)$ *d. f.*

This is otherwise obvious since the rank correlation coefficient can be regarded as the correlation between two variates.

Ex. Test the significance of the following rank correlation coefficient at 5% level.

(*a*) $r=0{\cdot}139$, $n=10$, [$t={\cdot}397$, not significant]

(*b*) $r=0{\cdot}618$, $n=20$, [$t=3{\cdot}34$ highly significant].

15·3. Tests based on the F distribution. We consider five tests of significance based on the F distribution.

15·3·1. Testing the significance of the ratio of two independent estimates of the population variance. Given two independent random samples from normal populations, we have to test the hypothesis that the population variances are the same. If the hypothesis is correct, it has been shown in § 13·7, that the statistic S_1^2/S_2^2, where

$$S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2,\quad S_2^2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2$$

and $S_1^2<S_2^2$, conforms to the F distribution with n_1-1, n_2-1 *d. f.* For carrying out the test of singificance, therefore, we calculate

$$F=\frac{S_1^2}{S_2^2}. \qquad \ldots(15{\cdot}8)$$

If the value exceeds $F_{0\cdot05}$ for n_1-1, n_2-1 *d. f.* we say that ratio is significant at 5% level. If $F<F_{0\cdot05}$ we say that the same could have come from two normal populations with the same variance.

Example 15·9. *Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 inches square and 91 inches square respectively. Can they be regarded as drawn from the same normal population ?*

Here $$S_1^2=\frac{160}{8}=20, S_2^2=\frac{91}{7}=13$$

$$\therefore \quad F=\frac{20}{13}=1\cdot54 \text{ approx.}$$

Also $F_{0\cdot05}$ for 8 and 7 *d. f.* is 3·37 and, therefore, the calculated values of F is not at all significant. Hence the two samples can be regarded as drawn from two normal populations with the same variance. If the populations are to be the same their means should also be the same and whether this is so can be tested by the t test.

Ex. 1. In two groups of ten children each the increases in weight due to two diffierent diets in the same period were in pounds,

8,	5,	7,	8,	3,	2,	7,	6,	5,	7
3,	7,	5,	6,	5,	4,	4,	5,	3,	6

Find whether the variances are significantly different.

[**Ans.** Not significant]

Ex. 2. Prove that

$$\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{0\cdot25}}, \quad \frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{0\cdot975}}$$

are 95% fiducial limits for the ratiod σ_1^2/σ_2^2 of the population variances, when the samples are known to have been drawn from populations.

15·3·2. Testing the significance of correlation ratio. Let (x_i, y_{ij}) $(i=1, 2,\ldots\ldots, h\,;\, j=1, 2,\ldots\ldots, n_i)$ denote a random sample of size $N(=\Sigma n_i)$ from a bivariate normal population and let η be the correlation ratio of y on x defined by

$$\sum_i \sum_j (y_{ij}-\bar{y}_i)^2=(1-\eta^2)\sum_i \sum_j (y_{ij}-\bar{y})^2$$

To calculate the significance of η, it can be shown that the ratio

$$F=\frac{\eta^2}{1-\eta^2}\times\frac{N-h}{h-1} \qquad \ldots(15\cdot9)$$

confoms to the F distribution with $v_1=h-1$, $v_2=N-h$ *d. f.*

Example 15·10. *A random sample of 100 pairs of values from a normal population grouped in 10 arrays of y's gives a correlation ratio $\eta_{yx}=0\cdot3$. Is it significant of association between the variates ?*

Here $n=100$, $h=10$, $\eta=0{\cdot}3$,

$$F=\frac{0{\cdot}09}{1-{\cdot}09}\times\frac{90}{9}=\frac{0{\cdot}90}{0\,91}=\frac{90}{91},\ \nu_1=9,\ \nu_2=90.$$

It is obvious that this value is not significant even at 5% level.

Ex. The significance of correlation ratio of Example 10·6.

15·3·3. Testing for the sinificance of a regression function. The appropriate statistic for testing whether the sample indicates any degree of association as represented by the regression equation is

$$F=\frac{r^2(n-2)}{1-r^2} \qquad \ldots(15{\cdot}10)$$

which is a F variate with 1 and $n-2$ *d. f.* This is equivalent to testing the significance of r and it is obvious that F is the square of the statistic t so that the two tests are equivalent.

Ex. Solve examples 15·5 and 15·6 by the use of the F table and show that these give the same results as obtained there.

15·3·4. Testing for non-linearity of regression. For a sample of size N arranged in h arrays from a normal sample the appropriate statistic for testing the hypothesis of linearity of regression is

$$F=\frac{\eta^2-r^2}{1-\eta^2}\times\frac{N-h}{h-2}, \qquad \ldots(15{\cdot}11)$$

with $h-2$, $N-h$ *d.f.*

Example 15·11. *A random sample of 150 pairs from a bivariate normal population when grouped in 15 arrays of y's gives* $r=0{\cdot}4$, $\eta=0{\cdot}5$. *Are these results consistent with the assumption of linearity of regression ?*

Here $N=150$, $h=15$, $r=0{\cdot}4$, $\eta=0{\cdot}5$

$$F=\frac{{\cdot}25-{\cdot}16}{1-{\cdot}25}\times\frac{135}{13}=1{\cdot}25,\ \nu_1=13,\ \nu_2=135$$

which is not significant at 5% level. Hence the results are consistent with the hypothesis of linearity of regression.

Ex. From 823 observations arranged in 9 arrays,

$$\Sigma n_i(\bar{y}_i-\bar{\bar{y}})^2=396,\ \Sigma\Sigma(y_{ij}-\bar{y}_i)=3832.$$

Is the regression linear ? [Ans. Not significant]

15·3·5. Testing the significance of an observed multiple correlation. Let R be the multiple correlation coefficient between a variate and p other variates in a random sample of size N from a $(p+1)$ variate normal distribution. Fisher has found the distribution

of R and he has shown that if the population multiple correlation coefficient is zero, then the statistic

$$F = \frac{R^2}{1-R^2} \cdot \frac{N-p-1}{p} \qquad \ldots(15{\cdot}12)$$

conforms to the F distribution with p, $N-p-1$ $d.f.$

Ex. 1. For a sample of 30 sets of values from a normal population $R_{2(31)}$ is found to be 0·5. Show that this is significant of correlation in the population between x_2 and x_1, x_3. ($F=4{\cdot}5$, $\nu_1=2$, $\nu_2=27$

Ex. 2. From a sample of size 20, $R_{1(23)}$ is found to be ·5433. Show that this is significant at 1% level.

15·4. Further applications of the F distribution.

15·4·1. Testing homogeneity of means. Analysis of variance. The t test enables us to test the significance of the difference between two sample means but if we have a number of means and we have to test the hypothesis that means are homogeneous or that they do not differ significantly, then the technique known as *Analysis of Variance* developed by Fisher is useful.

Consider a set of N values of the variate x arranged into h rows and k columns. Let x_{ij} represent the value of the member belonging to ith row and jth columns, $\bar{x}_{.j}$ the mean of the jth column and $\bar{x}_{i.}$ the mean of the ith row. The data is then arranged as below :

	1st col.	2nd col.	jth col.	kth col.	Row totals	Row means
1st row	x_{11}	x_{12}	..... x_{1j}	x_{1k}	$T_{1.}$	$\bar{x}_{1.}$
2nd row	x_{21}	x_{22}	..... x_{2j}	x_{2k}	$T_{2.}$	$\bar{x}_{2.}$
ith row	x_{i1}	x_{i2}	..... x_{ij}	x_{ik}	$T_{i.}$	$\bar{x}_{i.}$
hth row	x_{h1}	x_{h2}	... x_{hj} ...	x_{hk}	$T_{h.}$	$\bar{x}_{h.}$
Column totals	$T_{.1}$	$T_{.2}$	... $T_{.j}$ ...	$T_{.k}$	T	$\bar{x}_{..}$
Column means	$\bar{x}_{.1}$	$\bar{x}_{.2}$	... $\bar{x}_{.j}$ ...	$\bar{x}_{.k}$	$\bar{x}_{..}$	

We regard the data as classified into h classes, each row corresponding to a class according to some factors of classification. Thus each row may represent the price of commodity obtained from k shops in a district and the different rows may correspond to different districts or the data may represent marks obtained by h groups of students, each group consisting of k students.

The various values x_{ij} differ among themselves and this variability may be due to :

(i) variation within classes or groups :

(ii) variation from class to class.

The object of the analysis of variance would be to break up the total variation into components due to each of the factors and then compare them by the F test.

If we assume that each group is homogeneous within itself, then the variation within itself is due to chance or random causes. If the data is homogeneous with respect to the factor of classification, *i.e.*, if there is no effect due to this factor then the variation between classes would also be due to random causes and the two variations would be of the same order.

We shall set up the hypothesis that all x_{ij}'s ($i=1, 2.....,\ h$; $j=1, 2,....., k$) are drawn from a normal population with mean m and variance σ^2, and there is no real difference between classes.

From the above table, we have

$$\sum_{ij}\sum x_{ij}=\sum_{i}T_{i.}=T=\sum_{j}T_{.j}=\sum_{ij}\sum x_{ij} \text{ where } T=\text{grand total},$$

$$\bar{x}_{i.}=\frac{T_{i.}}{k},\quad \bar{x}_{.j}=\frac{T_{.j}}{h},\quad \bar{x}_{..}=\frac{T}{hk} \text{ where } \bar{x}_{..}=\text{general mean.}$$

Now we have the following useful algebraic relation :

$$\begin{aligned}\sum_{ij}\sum(x_{ij}-\bar{x}_{..})^2&=\sum_{ij}\sum(x_{ij}-\bar{x}_{i.}+\bar{x}_{i.}-\bar{x}_{..})^2\\&=\sum_{ij}\sum(x_{ij}-\bar{x}_{i.})^2+\sum_{ij}\sum(\bar{x}_{i.}-\bar{x}_{..})^2+2\sum_{ij}\sum(x_{ij}-\bar{x}_{i.})(\bar{x}_{i.}-\bar{x}_{..})\\&=\sum_{ij}\sum(x_{ij}-\bar{x}_{i.})^2+k\sum_{i}(\bar{x}_{i.}-\bar{x}_{..})^2+2k\sum_{i}(\bar{x}_{i.}-\bar{x}_{i}\,)(\bar{x}_{i.}-\bar{x}_{.})\\&=\sum_{ij}\sum(x_{ij}-\bar{x}_{i.})^2+k\sum_{i}(\bar{x}_{i}-\bar{x}_{..})^2 \qquad \ldots(15{\cdot}13)\end{aligned}$$

The sum of the squares (s.s.) on the L.H.S. is called *total sum of squares*. Now $\sum_{j}(x_{ij}-\bar{x}_{i.})^2$ represents sum of squares of deviations of all values of the ith class from their class mean. This is S.S. ***within***

the ith class. The first S.S. on the R.H.S. represents the S.S. *within* each class for all the classes. Accordingly is is called the S.S. within classes. Similarly, the second S.S. on the R.H.S. arises due to variation *between* class means and it is accordingly called S.S. *between classes.*

Now we make use of the fact assumed above that x_{ij}'s are normally distributed about the same mean m and variance σ^2.

Therefore, as proved in § 13·2,

$$\frac{1}{\sigma^2}\sum_{ij}\sum(x_{ij}-\bar{x}_{..})^2$$

is distributed as χ^2 with $hk-1$ *d.f.* and its expected value is $(hk-1)$.

Similarly, since $\bar{x}_{i.}$ is the mean of k members x_{ij}, $j=1, 2,\ldots, k$

$$\sum_{i}\frac{(\bar{x}_{i.}-\bar{x}_{..})^2}{\sigma^2/k}$$

is distributed as χ^2 with $h-1$ *d.f.* and its expected value is $h-1$. In the same way

$$\sum_{ij}\sum\frac{(x_{ij}-\bar{x}_{i.})^2}{\sigma^2}$$

is distributed as χ^2 with $\sum_{i}(k-1)=h(k-1)=(N-h)$ *d.f.* We thus get

$$E\left[\sum_{ij}\sum\frac{(x_{ij}-\bar{x}_{..})^2}{N-1}\right]=\sigma^2,\ E\left[\sum_{i}\frac{k(\bar{x}_{i.}-\bar{x}_{..})^2}{h-1}\right]=\sigma^2,$$

$$E\left[\sum_{ij}\sum\frac{(x_{ij}-\bar{x}_{i.})^2}{N-h}\right]=\sigma^2, \text{ where } N=hk.$$

In view of (15·13), it is obvious that the three unbiased estimates of σ^2 are *not independent.* However, in view of the independence of the distribution of sample mean and variance in random sampling from a normal distribution established in § 13·3, it can be shown that the second and third estimates are independent and their ratio can be treated by the F test

where $$F=\frac{\sum k(\bar{x}_{i.}-\bar{x}_{..})^2}{\sum\sum(x_{ij}-\bar{x}_{i.})^2}\times\frac{N-h}{h-1}$$

with $$v_1=h-1,\ v_2=N-h.$$

The above results are expressed conveniently in the following *analysis of variance table.*

Source of variation	*D.F.*	Sum of squares	Mean square
Between class means	$h-1$	$\underset{i}{\Sigma} k(\bar{x}_{i.}-\bar{x}_{..})^2$	$\underset{i}{\Sigma} k(\bar{x}_{i.}-\bar{x}_{..})/(h-1)$
Within classes	$N-h$	$\underset{i\,j}{\Sigma\Sigma}(x_{ij}-\bar{x}_{i.})^2$	$\underset{i\,j}{\Sigma\Sigma}(x_{ij}-\bar{x}_{i.})^2/(N-h)$
Total	$N-1$	$\underset{i\,j}{\Sigma\Sigma}(x_{ij}-\bar{x}_{..})^2$	

It will be noted that the mean squares are obtained by dividing the S.S. by the corresponding *d.f.* For the practical calculation of S.S. we may note first that all the S.S. are independent of change of origin and any convenient origin may, therefore, be chosen. Secondly, all the S.S. can be expressed in terms of row totals $T_{i.}$, column total $T_{.j}$ and grand total T. thus

$$\underset{i\,j}{\Sigma\Sigma}(x_{ij}-\bar{x}_{..})^2=\underset{i\,j}{\Sigma\Sigma}x_{ij}^2-2\bar{x}_{..}\underset{i\,j}{\Sigma\Sigma}x_{ij}+N\bar{x}_{..}^2=\underset{i\,j}{\Sigma\Sigma}x_{ij}^2-\frac{T^2}{N}$$

$$\underset{i}{\Sigma}k(\bar{x}_{i.}-\bar{x}_{..})^2=k\underset{i}{\Sigma}\left(\frac{T_i}{k}-\frac{T}{hk}\right)^2=\underset{i}{\Sigma}\frac{T_i^2.}{k}-\frac{T^2}{N}.$$

Now $\Sigma\Sigma\ (x_{ij}-\bar{x}_{i.})^2$ can be obtained by subtracting on making use of (15·13). The term T^2/N is usually called the *correction factor* on account of measurements being taken from the mean.

Example 15·12. *10 varieties of wheat are given in 3 plots each and following yields in maunds per ace obtained :*

Plots/Variety	1	2	3	4	5	6	7	8	9	10	*Total*
I	7	7	14	11	9	6	9	8	12	9	
II	8	9	13	10	9	7	13	13	11	11	
III	7	6	16	11	12	5	12	11	11	11	
Total	22	22	43	32	30	18	34	32	34	31	298

Test the significance of the differences between variety yields.

Since the S.S. are unaffected by change of origin, we subtract 10 maunds from each yield.

Now $h=10$, $k=3$, $T_{.1}=-8$, $T_{.2}=-8$....., $T=-2$, $N=30$

$$\Sigma\Sigma x_{ij}^2=(-3)^2+(-2)^2+(-3)^2+(-3)^2.....=204$$

$$\Sigma\frac{T_{.j}^2}{3}=\tfrac{1}{3}[(-8)^2+(-8)^2+(13)^2+.....+(1)^2]=160{\cdot}67.$$

$$\text{Total S.S.}=\Sigma\Sigma x_{ij}^2-\frac{T^2}{N}=204-\frac{4}{30}=203{\cdot}87.$$

$$\text{S.S. between classes}=\tfrac{1}{3}\Sigma T_{.j}^2-\frac{T^2}{N}=160{\cdot}67-\frac{4}{30}=160{\cdot}54$$

S.S. within classes $=203{\cdot}87-160{\cdot}54=43{\cdot}33$

Thus we get

Analysis of Variance Table

Source of variation	*D.F*	*S.S.*	*M.S.*
Between varieties	9	160·54	17·84
Within varieties	20	43·33	2·17
Total	29	203·87	

$$F=\frac{17{\cdot}84}{21{\cdot}7}=8{\cdot}22,\ \nu_1=9,\ \nu_2=20$$

This is highly significant at 1% level.

15·4·2. Analysis of variance for two factors of classification. If in table given on page 440 we consider the data divided according to both rows and columns, we break up the total S.S. as follows :

$$\underset{i\,j}{\Sigma\Sigma}(x_{ij}-\bar{x}_{..})^2=k\underset{i}{\Sigma}(\bar{x}_{i.}-\bar{x}_{..})^2+h\underset{j}{\Sigma}(\bar{x}_{.j}-\bar{x}_{..})^2+\underset{i\,j}{\Sigma\Sigma}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})^2 \quad ...(15{\cdot}14)$$

since the sum of products are easily seen to vanish.

As in the last sub-section, on the basis of normality of distribution of the variates and the null hypothesis of the homogeneity of both the row and column factors of classification the above four sums on being divided respectively by $hk-1$, $h-1$, $k-1$ and $(h-1)(k-1)$ give unbiased estimates of the population variance σ^2 and by using the theorem due to Cochran it can be shown that estimates given by S.S. the between row 'means' and also those given by 'between column means' are independent of the estimate given by the last error, *i.e.*, S.S. The analysis of variance table is :

Source of variation	D.F.	S.S.	M.S.	F
Between row	$h-1$	$\Sigma_i k(\bar{x}_{i.}-\bar{x}_{..})^2$	S.S./D.F.	
Between columns	$k-1$	$\Sigma_i h(\bar{x}_{.j}-\bar{x}_{..})^2$		
Error	$(h-1)(k-1)$	$\Sigma_i\Sigma_j(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})^2$		
Total	$hk-1$	$\Sigma_i\Sigma_j(\bar{x}_{ij}-\bar{x}_{..})^2$	—	

Example 15·13. *Yields of 4 varieties of wheat in 3 blocks are given below :*

Varieties ↓ \ *Blocks* →	1	2	3
I	10	9	8
II	7	7	6
III	8	5	4
IV	5	4	4

(*M.A., Punjab, 1949*)

Is the difference between varieties significant ?

Subtracting 5 from each element, we get

Varieties	*Blocks*			*Total*
I	5	4	3	12
II	2	2	1	5
III	3	0	−1	2
IV	0	−1	−1	−2
Total	10	5	2	17

$\Sigma\Sigma x_{ij}^2=25+16+9+4+4+1+9+0+1+0+1+1=71,$

$\frac{1}{3}T_{i.}^2=\frac{1}{3}[12^2+5^2+2^2+(-2)^2]=\frac{1}{3}(177) \quad =59,$

$\frac{1}{4}T_{.j}^2=\frac{1}{4}[10^2+5^2+2^2] \quad =\frac{1}{4}(129) \quad =32{\cdot}25.$

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Varieties	3	59−24·1 =34·9	11·63	18·1
Blocks	2	32·25−24·1=8·15	4·07	6·4
Error	6	3·85	0·64	—
Total	11	71−24·1=46·9	—	—

The difference between varieties is highly significant.

The difference between blocks is significant at 5% level but not at 1% level.

Ex. 1. Prove the following :

(*i*) $\sum_i \sum_j (\bar{x}_{i.}-\bar{x}_{..})\,(\bar{x}_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})=0$

(*ii*) $\sum_j \sum_i (\bar{x}_{..}-\bar{x}_{.j})\,(\bar{x}_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})=0$

Ex. 2. For the data of example 15·12 establish the following table.

Due to	*D.F.*	*S.S.*	*M.S.*	*F*
Between rows	2	8·27	4·13	2·12
Between columns	9	160·54	17·84	—
Error	18	35·06	1·94	—
Total	29	203·87	—	—

15·5. Tests of significance based on Fisher's z transformation. Fisher's z transformation

$$z=\tfrac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)=1\cdot1513\log_{10}\left(\frac{1+r}{1-r}\right)$$

have been discussed in §14·8. Fisher has shown that for samples of a large size n, z is distributed asymptotically normally about mean ξ, where

$$\xi=\tfrac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)=1\cdot1513\log_{10}\left(\frac{1+\rho}{1-\rho}\right)$$

with variance $1/(n-3)$. This transformation forms the basis of the tests of significance discussed below, but will be obvious that unlike the tests of the previous two sections the present tests will hold for large samples only.

15·5·1. To test whether r differs significantly from ρ. Given a random sample $(x_1, y_1), \ldots, (x_n, y_n)$ from a bivariate normal population, we have to test the hypothesis that the population correlation coefficient is ρ. To do this we calculate z and ξ by applying Fisher's transformation and then calculate the asymptotically standard normal variate

$$\frac{z-\xi}{1/\sqrt{n-3}} \qquad \ldots(15\cdot15)$$

If the absolute value of this statistic exceeds 1·96 the difference is significant at 5% level.

Example 15·14. *A correlation coefficient of 0·7 is discovered in a sample of 28 pairs. Apply z transformation to find out if this differs significantly (a) from 0 (b) from 0·5.*

(*a*) $\rho=0 \qquad \xi=0$

Also $z=1\cdot1513 \log_{10}\left(\frac{1+\cdot7}{1-\cdot7}\right)=0\ 87$

$\therefore \qquad \frac{0\cdot87-0}{1/\sqrt{28-3}}=4\cdot35>1\cdot96$

The hypothesis of zero correlation is discredited and we conclude that there is a correlation in the population

(*a*) $\rho=0\cdot5,\ \xi=1\cdot1513\ \log_{10}\left(\frac{1+0\cdot5}{1-0\cdot5}\right)=0\cdot55$

$$\frac{0\cdot87-0\cdot55}{1/\sqrt{28-3}}=0\cdot6<1\cdot96$$

The data is consistent with the hypothesis of $\rho=0\cdot5$

Ex. 1. Show that the 95% fiducial limits for ζ are $z\pm1\cdot96\sqrt{n-3}$.

Ex. 2. Prove that the 95% fiducial limits for ρ in the above example are 0·45 and 0·85.

15·5·2. To test the significance of the difference between two independent correlation coefficients. We have to test the hypothesis that the two samples of sizes n_1, n_2 giving correlation coefficients r_1, r_2 may be regarded as drawn from the sample population or from two populations with the same correlation coefficient. If the hypothesis is correct, the statistic

$$\frac{z_1-z_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}} \qquad \ldots(15\cdot16)$$

where $z_1=\frac{1}{2}\log_e\left(\frac{1+r_1}{1-r_1}\right),\ z_2=\frac{1}{2}\log_e\left(\frac{1+r_2}{1-r_2}\right)$

is asymptotically a standard normal variate. If absolute value of this statistic is greater than 1·96, the difference will be significant at 5% level.

Example 15·15. *The correlation coefficients between temperatures of rice and breakage percentage calculated from two samples of 12 and 16 are ·8912 and ·8482 respectively. Do the two estimates differ significantly ?*

Here $n_1=12$, $n_2=16$, $r_1=·8912$, $r_2=·8482$

.. $z_1=1·4276$, $z_2=1·2496$

Now $$\frac{z_1-z_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}=\frac{1·4276-1·2496}{\sqrt{0·1880}}=0·41$$

which is not at all significant at 5% level.

Ex. Two samples of sizes 23, 28 give correlation coefficients 0·5 and 0·8 respectively. Show that these are not significantly different.

15·5·3. Combination of estimates of a correlation coefficient. Correlation coefficients $r_1, r_2,\ldots\ldots,r_k$ are obtained from samples of sizes $n_1, n_2,\ldots\ldots, n_k$ from the same bivariate normal population with correlation coefficient ; ρ, then $z_1, z_2,\ldots\ldots,z_k$ are independent estimates of the same quantity $\zeta=\frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$. To get a pooled estimate for ζ, we take weighted average of $z_1, z_2\ldots\ldots,z_k$ and to get an estimate with the minimum variance we use weights inversely proportional to the variance (*cf.* Ex. 18, page 75).

The estimate is accordingly

$$\bar{z}=\frac{\Sigma(n_i-3)z_i}{\Sigma(n_i-3)}$$

and from this we can get an estimate for ρ from the relation $\bar{\rho}=\tanh\bar{z}$. Then since the variate z_i are approximately normally distributed about ξ with variance $1/(n_i-3)$, the sum $\Sigma(n_i-3)(z_i-\bar{z})^2$ is distributed approximately as χ^2 with $(k-1)$ *d.f.* The significance can be ascertained from the table of χ^2. Not significance shows that the estimate r_i of correlation in the population may be regarded as homogeneous.

Example 15·16. *Combine the two estimates of r in Example 15·15 to get a pooled estimate*

$$\bar{z}=\frac{(n_1-3)z_1+(n_2-3)z_2}{(n_1-3)+(n_2-3)}=\frac{12·8484+16·2448}{9+13}=\frac{29·0932}{22}=1·3224.$$

The corresponding estimate of ρ is 0 8674.

Ex. 1. Independent samples of 21, 30, 39, 26, 35 pair of values give correlation coefficients 0·39, 0·61, 0·43, 0·54, 0·48. Prove that the pooled estimate for ρ is 0·5 nearly and the value of χ^2 is 1·4 approx.

Ex. 2. Two independent samples of sizes 20 and 25 give correlation coefficients ·60 and ·80. Show that these are not significantly different. Show also that the weighted average value of r=·7267.

Ex. 3. Show that in exercise of § 15·4·2 the 95% fiducial limits for r_1 from the given value of are r_2 0·47 and 0·93.

15·5·4. Tests of significance for partial correlation coefficients. The results of the last three sections hold equally for partial correlation coefficients, only in each case the sample number should be reduced by k, the order of the partial correlation coefficients indicated by the number of secondary subscripts.

Ex. From independent sample of 30 and 20 sets of values, partial correlation of order three are found to be ·4 and ·6 respectively. Examine

(*i*) whether the first sample could have come from a population with the corresponding correlation coefficient equal to ·6 ;

(*ii*) whether the two samples could have come from the same normal population and if this is true find the pooled estimate.

EXERCISES 15

1. Obtain tests of significance for testing the following hypothesis :

(*a*) σ_0 is the hypothetical value of σ in a normal population $N(\mu, \sigma^2)$ with known μ.

(*b*) σ_0 is the hypothetical value of σ in a normal population $N(\mu, \sigma^2)$ with unknown μ.

(*c*) $\mu_1=\mu_2$ in two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ where μ_1 and μ_2 are unknown but σ_1^2 and σ_2^2 are known.

(*M. A. Patna, 1957* ; *B. A. Hons. Delhi, 1958*)

[**Hint.** (*a*) $\Sigma \frac{(x_i-\mu)^2}{\sigma_0^2}$ is distributed as χ^2 with *n.d.f.*

(*b*) $\Sigma \frac{(x_i-\bar{x})^2}{\sigma_0^2}$ is distributed as χ^2 with $(n-1)$ *d.f.*

(*c*) $\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ is a standard normal variate.]

2. Find Student's t for the following variate values in a sample of eight,

$-4, -2, -2, 0, 2, 2, 3, 3$

taking the mean of the universe to be zero. (*M. Sc., Agra, 1948*)

[**Ans.** =0·29 nearly]

3. A machine which produces mica insulating washers for use in electric devices is said to turn out washers having a thickness of 10 mils (1 mil=0·001 inch). A sample of 10 washers has an average thickness of 9·52 mils with a standard deviation of 0·60 mil. Find out the significance of such a deviation.

[**Ans.** t=2·528, Deviation is significant]

4. A certain stimulus administered to each of 12 patients resulted in the following increases of blood pressures.

5, 2, 8, −1, 3, 0, 6, −2, 1, 5, 0, 4.

Can it be concluded that the stimulus will be, in general, accompanied by an increase in blood pressure ?

(M. A., Eco., Delhi, 1957 ; I.C.A.R., Delhi)

[**Ans.** $t=2{\cdot}347$. The value is significant and therefore it can be concluded that stimulus will be in general accompanied by blood pressure.]

5. A sample of 11 rats from a central population had an average blood viscosity of 3·92 with a *s. d.* of 0·61. On the basis of this sample, establish 95% limits of μ, the mean blood viscosity of control population.

[**Ans.** Limits are $3{\cdot}92=\dfrac{2{\cdot}228(0{\cdot}61)}{\sqrt{11}}$ *i.e.*, 3·51 and 4·33]

6. Suppose X is a continuous chance variate such that all values of x in the interval $(0, \theta)$ are equally likely where θ is unknown. Suppose a sample of 20 values of X is drawn and the sample mean $\bar{x}$ has the value 3·2. What are the 90% confidence limits of θ ?

$$\left[\ dF=\frac{dx}{\theta},\ 0\leqslant x\leqslant\theta\ ;\ \mu_1'=\tfrac{1}{2}\theta,\ \mu_2'=\frac{1}{12}\cdot\theta^2\Rightarrow\ \sigma=\frac{\theta}{\sqrt{12}}\ .\right.$$

From table, t for 11 *d. f.* at 1% level $=1{\cdot}73$.

$$\therefore\quad \frac{\bar{x}-\theta/2}{\dfrac{\theta}{\sqrt{12}}\cdot\dfrac{1}{\sqrt{n}}}=\pm 1{\cdot}73.\ \text{Since}\ \bar{x}=3{\cdot}2\quad\therefore\ \theta=\text{etc.}]$$

7. The sleep of 10 patients was measured for the effects of the soporific drugs referred to in the following table as Drug A and Drug B. From the data given below show that there is a significant difference between the effects of two drugs, on the assumption that different random samples of patients were used to test the two drugs A and B.

Additional hours of sleep gained by use of soporific drugs.

Patient	*Drug A*	*Drug B*
1	+0·7	1·9
2	−1·6	0·8
3	−0·2	1·1
4	−1·2	0·1
5	−0·1	0·1
6	3·4	4·4
7	3·7	5·5
8	0·8	1·6
9	0	4·6
10	2·0	3·6

(I.C.A.R. Delhi, 1957 ; M.A., Punjab, 1954)

[**Hint.** This is an example of paired samples. Using the formula (15·3)

$$t=\frac{1{\cdot}6\times 3{\cdot}1}{1{\cdot}23}=3{\cdot}78$$

This value is significant and hence there is a significant difference between the effect of the two drugs which is in favour of the second.]

8. The following table shows the mean number of bacterial colonies per plate obtained by four slightly different methods from soil samples taken at 4 P.M. and 8 P.M. respectively.

	Method A	*Method B*	*Method C*	*Method D*
4 P.M.	29·75	27·50	30·25	27·80
8 P.M.	39·20	40·60	36·20	42·40

Are there significantly more bacteria at 8 P.M. than at 4 P.M.

(*B.A. Hons., Delhi, 1959*)

[Ans. Treating as paired samples, $t=5{\cdot}560$, $n=3$. There are more bacteria at 8 P.M. than at 4 P.M.]

9. In a certain experiment to compare two types of pig-foods *A* and *B*, the following results of increase in weights were observed in pigs :

Pig Number		1	2	3	4	5	6	7	8	Total
Increase in weight in lbs.	*Food A*	49	53	51	52	47	50	52	53	407
	Food B	52	55	52	53	50	54	54	53	423

Assuming that the two samples of pigs are independent, can it be concluded that food *B* is better than food *A*. Examine the case when the same set of eight pigs were used in both the foods ? [Ans. Food *B* is better than Food *A*]

(*B.A., Hons. Delhi, 1955, 71 ; Bombay, 1952*)

10. A farmer grows crops on two fields *A* and *B*. On *A* he puts Rs. 10 worth of manure per acre and on *B* Rs. 20 worth. The net returns per acre, exclusive of the cost of manure, on the two fields are :

Year A	*Yield A, Rs. per acre*	*Yield B, Rs. per acre*
1	34	36
2	28	33
3	42	48
4	37	38
5	44	50

Other things being equal, discuss the question whether it is likely to pay the farmer to continue the more expensive dressing ?

[Ans. $\bar{d}=$Rs. 4, $S=$Rs. $2{\cdot}345$, $t=3{\cdot}814$, $\nu=4$. It is likely to pay the farmer to continue the more expensive dressing]

11. The heights of six randomly chosen sailors are in inches 63, 65, 68, 69, 71, and 72. Those of the randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72, 73.

Discuss in the light that the data throw that the soldiers are, on the average, taller than the sailors.

[Ans. The soldiers on the average are not likely to be taller than the sailors].

12. Eight pots growing three wheat plants each were exposed to a high tension discharge while nine similar pots were enclosed in an earthen wire case. The number of tillers in each pot were as follows :—

Caged	17	26	18	25	27	28	26	23	17
Electrified	16	16	22	16	21	18	15	20	

Discuss whether electrification exercises any real effect on tillering.

(I.C.A.R., Delhi, 1956)

[Ans. The electrification does exert some effect on tillering]

13. Mitchell conducted a paired feeding experiment with pigs on the relative value of limestone and bone meal for bone development. The results are :

Ash content in percentages of scapulas of pairs of pigs fed on limestone and bone meal.

Pair	*Limestone*	*Bone meal*
1	49·2	51·5
2	53·3	54·9
3	50·6	52·2
4	52·0	53·3
5	46·8	51·6
6	50·5	54·1
7	52·1	54·2
8	53·0	54·3

Determine the significance of the difference between the means in two ways (1) By assuming that the values are paired and (2) by assuming that the values are not paired. *(I.C.A.R., Delhi, 1956)*

[Ans. It can be reasonably concluded that the difference between meals is significant]

14. The means of two random samples of sizes 9 and 7 respectively, are 196·40 and 198·82 respectively. The sum of the squares of the deviation from the means are 26·94 and 18·73 respectively. Can the samples be considered to have been drawn from the same normal population, it being given that the value of t for 14 *d.f.* at 5% level of significance is 2·145 and at 1% level of significance is 2·977. *(B.A., and B.Sc., Lucknow, 1960)*

[Ans. The sample cannot be considered to have been drawn from the same normal population.]

15. Two horses *A* and *B* were tested according to the time (in seconds) to run a particular track with the following results :

Horse *A*	28	30	32	33	33	29	and	34
Horse *B*	29	30	30	24	27	and	29	

Analyse, as completely and in as many ways as you can, the above data and report whether or not you can discriminate between the two horses, and if so, with what degree of confidence. You may use the following extracts from statistical tables. Probability that t will exceed the tabular entry is absolute.

Degrees of freedom	*Percentage*		*Probability*	
	5	2·50	0·05	0·1
11	2·20	2·59	3·50	4·44
12	2·18	2·56	3·43	4·32
13	2·16	2·53	3·37	4·22
18	2·10	2·44	3·20	3·92

(*I A.S., 1963*)

[Hint. $\bar{x}_1=\frac{219}{7}=31{\cdot}3$ approx., $\bar{x}_2=\frac{169}{6}=28{\cdot}1$ approx.

Taking 32 and 19 as the working means for series A and B respectively, we get $S^2=\frac{31{\cdot}43+26{\cdot}86}{11}=5{\cdot}3$ nearly.

Hence $t=\frac{3{\cdot}2}{\sqrt{5{\cdot}3}\sqrt{\frac{1}{6}+\frac{1}{7}}}=2{\cdot}5$ approx.

As 5% level of significance the value of t for 11 $d.f.=2{\cdot}20$. Hence we can discriminate between the two horses. We can discriminate between the horses only with 97·5 degree of confidence. F test can also be applied].

16. A set of 15 observations gives mean=68·57, standard deviation=2·40, another set of 7 observations gives mean=65·14, standard deviation=27·70. Use the t test to find whether the two sets of data were drawn from populations with the same mean it being assumed that the standard deviations in the two populations were (1) equal (2) unequal. You may base your answer on the following extracts from t table. Significance limits of t probability of a value numerically greater than t.

Degrees of freedom	*Percentage*		*Probability*
	0·5	·02	·01
6	2·447	3·143	3·707
14	2·145	2·624	2·977
20	2·086	2·528	2·845

(*I.A.S. 1952*)

[Hint. $t=\frac{68{\cdot}57-65{\cdot}14}{\sqrt{\frac{(2{\cdot}40)^2}{15}+\frac{(2{\cdot}70)^2}{7}}}=2{\cdot}336$]

17. 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, 42, 35, 32, 34, 38, 28, 41, 43, 30 and 37.

Obtain the estimates of the variances of the population and test whether the two populations have the same variance. (*I.A.S , 1956*)

[Sol. $\bar{x}_1=22$, $\bar{x}_2=35$; $\sum_{i=1}^{10}(x_{1i}-\bar{x}_1)^2=1{\cdot}20$, $\sum_{j=1}^{12}(x_{2j}-\bar{x}_2)=314$

The estimates for the population variances furnished by the samples are 120/9 and 314/11 corresponding to 9 and 11 $d.f.$ respectively. The second estimate is larger, so that $F=\frac{314/11}{120/9}=2{\cdot}14$.

This value of F is much below the 5% point as indicated by the F table. It is, therefore, not at all significant, so that the samples may very well be regarded as drawn from the same population.

18. Show how you would use Student's t-test and Fisher's z-test to decide whether the two sets of observations

17, 27, 18, 25, 27, 29, 27, 23, 17

and 16, 16, 20, 16, 20, 17, 15, 21

indicate samples drawn from the same universe. (*M.Sc., Agra, 1949*

[**Hint.** $t=3{\cdot}05$, $z=\frac{1}{2}\log_e \frac{S_1^2}{S_2^2}=\frac{1}{2}\log_e \frac{23}{5{\cdot}4107}=0{\cdot}724.$]

19. Test whether correlation is significant if $r={\cdot}6$ and $n=38$.

(*B.A. Hons., Delhi, 1959*)

[Ans. Correlation is significant]

20. In a random sample of 28 pairs of values from a bivariate normal population the correlation was found to be 0·7. Is this value consistent with the assumption that the correlation in the population is 0·5 ? Find the 95% fiducial limits for ρ, giving that $r=0{\cdot}7$, $z=0{\cdot}87$ and for $z=0{\cdot}5$, $\xi=0{\cdot}55$.

[Ans. $t=\dfrac{z-\xi}{\sqrt{\frac{1}{n-3}}}=\dfrac{0{\cdot}87-0{\cdot}55}{\sqrt{\frac{1}{25}}}=1{\cdot}6<2$. Hence the value is not significant so that the correlation in the population might very well be 0·5, 95% fiducial limits are 0·45 and 0·85 approx.]

21. What is the probability that a correlation coefficient of 0·75 or less, can arise in a sample of 30 from a normal population in which the true correlation is +0·9 ?

[**Hint.** $t=\dfrac{z-\xi}{\sqrt{\frac{1}{n-3}}}=$ etc. $=2{\cdot}6$] [**Ans.** ·048]

22. The first of two samples consists of 23 pairs, and gives a correlation of 0·5 while the second of 28 pairs has a correlation of 0·8. Are these values significantly different ?

[**Hint.** Hypothesis is that they are drawn from the normal population

$z_1=1{\cdot}1513\ \log_{10}\left(\frac{1+{\cdot}5}{1-{\cdot}5}\right)=0{\cdot}55$, $z_2=1{\cdot}1513\ \log_{10}\left(\frac{1+{\cdot}8}{1-{\cdot}8}\right)=1{\cdot}10$

$m_1=23$, $n_2=28$, $t=\dfrac{z_1-z_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}=\dfrac{0{\cdot}55}{\sqrt{\frac{1}{20}+\frac{1}{25}}}$ $\dfrac{0{\cdot}55}{0{\cdot}3}=1{\cdot}83<2$

and is not quite significant at 5% level. The hypothesis is not discredited.]

23. Four breeds of cattle B_1, B_2, B_3, B_4 were fed on three different rations R_1, R_2, R_3. Gain in weight in pounds over a given period were recorded.

	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

Is there a significant difference : (*a*) between breeds ; (*b*) between rations ?

[Ans. No significant difference between breeds or between rations]

24. 12 pictures submitted in a competition were ranked by two judges with results as shown in the table below :

TABLE

Picture	A	B	C	D	E	F	G	H	I	J	K	L
Rank assigned by 1st judge	5	9	6	7	1	3	4	12	2	11	10	8
Rank assigned by 2nd judge	8	8	9	11	3	1	2	10	4	12	7	6

Calculate ρ, the rank correlation coefficient. Is there a lack of independence in these ranking ? Assume that on the hpothesis of independence of two sets on n rankings $t=\rho\left(\frac{n-2}{1-\rho^2}\right)^{\frac{1}{2}}$ follows the t distribution with $(n-2)$ degrees of freedom.

Table of t distribution

Degrees of freedom	10	11	12
Values of t significance at 5% level of probability	2·23	2·20	2·18

(I.A.S., 1956)

[Ans. $\rho=\cdot 79$, $t=4\cdot089$ approx. Yes ranking is significant]

25. *(a)* Find out the approximate probability that a correlation coefficient of 0·75 or less can arise in a sample of 18 from a normal population in which the true coefficient is +0·9.

(b) Test on the 5% level the significance of the following :

(i) A multiple correlation coefficient of +0·65 of y as 4 variates, x_1 x_2, x_3 and x_4 based on a sample of size 25.

(ii) A partial correlation −0·45 between x_1 and x_2 (eliminating effect of x_3 and x_4).

(iii) The correlation coefficient between mathematical aptitude and linguistic aptitude for a group of 42 boys is 0·72 and for a group of 275 girls is 0·75. Is the difference significant ?

(I.C.A.R., Delhi, 1947)

(c) In a sample of 25 sets of values from a normal population $r_{1\cdot234}$ was found out to be 0·4. Find whether there is significance of correlation in the population between x_1 and the variables x_2, x_3, x_4.

(M.A., Statistics, Delhi, 1959)

26. In a feeding experiment on swins, three rations R_1, R_2, R_3 were tried. The animals put into three classes of three each according to litter and initial body weight. The following table gives the gains in body weight in lbs. in a certain period.

	Class I	Class II	Class III
R_1	4	16	10
R_2	14	18	19
R_3	3	14	7

Analyse the data and state your conclusion. Has division into classes proved effective ? (5 per cent of F for $n_1=2$, $n_2=4$ is 6·94 and for $n_1=2$, $n_2=6$, is 5·14). *(M.A., Delhi, 1953)*

[**Sol.** Total sum of squares $=(221+776+510)-\frac{(105)^2}{9}=282$

Between class means $=\frac{(21)^2+(48)^2+(36)^2}{3}-1225=122$

Between ration means $=\frac{(30)^2+(51)^2+(24)^2}{3}-1225=134$

Within sum of squares $=282-(122+134)=26$.

Analysis of Variance Table

Source of variance	*S.S.*	*D.F.*	*M.S.*	*Variance ratio*
Between classes	122	2	61·0	$F_1=\frac{61}{6\cdot5}=9\cdot38$
Between rations	134	2	67·0	$F_2=\frac{67}{6\cdot5}=10\cdot30$
Errors	26	4	6·5	
Total	282	8		

If we do not consider variation between rations, then the table is

S.V.	S.S.	D.F.	M.S.	V.R.
Between classes	122	2	67·0	$F=2\cdot5$
Error	160	6	26·7	
Total	282	8		

F_1, F_2 are significant at 5% level ; F is not significant at 5% level.

28. The varieties A, B, 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 yield and state your conclusions.

A	*C*	*A*	*B*
6	5	8	9
C	*A*	*B*	*C*
8	4	6	9
B	*B*	*C*	*A*
7	6	10	6

(5 per cent value of F for $n_1=2$, $n_2=9$ is 1·236).

(M.A., Delhi, 1959 ; I.A.S. 1947, 1950)

[**Ans.** The difference in varieties of the given crop do not have any significant effect on the yield].

29. Set up a table of analysis of variance for :

Plots	Varieties			
	a	*b*	*c*	*d*
1	200	230	250	300
2	190	270	300	270
3	240	150	145	180

(Punjab Certificate, 1946)

Ans. TABLE

Source of variation	*S.S.*	*D.F.*	*M.S.*
Within varieties	29184	8	3648
Between varieties	2829	2	1414·5
Total	32013	10	

The variance ratio is not significant at 5% level.

30. 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 pound per week are shown below :—

District	Salesman			
	A	*B*	*C*	*D*
K	30	70	30	30
O	80	50	40	70
S	100	60	80	80

Carry out an analysis of variance and interpret the results stating the limitation under which your results are valid. *(I.C.A.R. Delhi, 1956)*

31. To test the significance of the variation of the retail price of a certain commodity among four large cities *A*, *B*, *C*, and *D*, seven shops were chosen at random in each city, and the prices observed in pence were as follows :

A : 82, 79, 73, 69, 69, 63, 61.
B : 84, 82, 80, 79, 76, 68, 62.
C : 88, 84, 80, 68, 68, 62, 66.
D : 79, 77, 76, 74, 72, 68, 64.

Do the data indicate that the prices for the four cities are significantly different ? *(B.Sc., Agra, 1956)*

Ans.

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Between cities	3	94·96	31·65	10
Within cities	24	1446·00	60·25	—
Total	27	1540·96	—	—

The value of F is not significant and the data is consistent with the hypothesis that the prices are the same in the four cities.

32. The following table gives the weights of 12 individuals drawn from four different sources situated in 3 different districts. Carry out an analysis of variance.

Sources	Districts						
	A	*B*	*C*				Total
S_1	61	62	63	−3	−2	−1	−6
S_2	63	64	63	−1	0	−1	−2
S_3	66	67	68	2	3	4	9
S_4	68	69	69	4	5	5	14
				2	6	7	15

[Hint. Change the origin to 64.

Total S.S. $= [3^2+2^2+\ldots+5^2 - \frac{[13]^2}{12} = 111 - 18{\cdot}75 = 92{\cdot}25$

Between S.S. (Rows) $= \frac{6^2+2^2+14^2}{3} - 18{\cdot}75 = 86{\cdot}92$

Between S.S. (columns) $= \frac{2^2+6^2+7^2}{2} - 18{\cdot}75 = 3{\cdot}5$

Error (within S.S.) = Total S.S. − B.S.S. rows − B.S.S. column = 1·83

Source of variation	*D.F.*	*S.S.*	*M.S.*	*V.R.*
B.S.S. Rows	3	86·92	28·97	$\frac{28{\cdot}97}{{\cdot}305} > F(3, 6)$ 1%
B.S.S. Column	2	3·5	1·25	
W.S.S.	6	18·3	·305	$\frac{1{\cdot}75}{{\cdot}305} > F(3,6)$ 5%
Total	11	92·25		

33. Carry out the analysis of the following Latin Square design :

50 A	70 B	70 C	80 D	90 E
70 B	90 C	80 D	80 E	50 A
60 C	50 D	90 E	80 A	90 B
50 D	60 E	80 A	50 B	70 C
80 E	90 A	50 B	70 C	60 D

(*B.A. Vikram, 1961*)

34. In § 15·5 3, prove that if w_i is the weight of z_i and $z = \Sigma_i w_i z_i / \Sigma_i w_i$ then $E(z) = \xi$. Show that Var z is minimum for variation in w_i if $\propto 1/\text{Var } z_i = (n_i - 3)$ and the minimum variance of z is $1/\Sigma_i (n_i - 3)$.

16

Tests of Significance Based on Chi-Square Distribution

16·1. Introduction. The chi-square distribution has a number of applications in statistics. We shall now consider these.

16·2. χ^2-Test as a Test of Goodness of Fit. An important problem of statistical inference is to test the hypothesis that the given data has been obtained by random sampling from a specified population with definite values for its parameters. The data etc. usually given can be arranged in the form of a frequency distribution, wherein, we are given the observed frequencies for the various 'classes' or 'cells'. The corresponding theoretical frequencies are obtained from the knowledge of the population and our problem is to test the compatibility of observed and theoretical frequencies or to determine whether the deviations of the observed frequencies from the theoretical frequencies are small enough to be regarded as due to fluctuations of random sampling or whether they indicated that the data could not have possibly come from a population giving rise to theoretical frequencies.

We have already come across a problem of this type in Chapter 12 when for sampling from a binomial population, we had two classes—success and failure. In n independent trials the observed frequencies were x and $n-x$ respectively and the expected frequencies on the hypothesis, that the probability of success in each trial is p, were np and nq, $q=1-p$ respectively. The deviation of the observed frequencies from theoretical frequencies were thus

$$x-np \text{ and } n-x-nq=np-x$$

respectively. The statistic used for testing the hypothesis was

$$z=\frac{x-np}{\sqrt{npq}}$$

so that we can write

$$z^2=\frac{(x-np)^2}{npq}=\frac{(x-np)^2}{np}+\frac{(n-x-np)^2}{nq} \qquad \ldots(16\cdot1)$$

For large samples, z was found to be an asymptotic standard normal variate so that z^2 is a χ^2 variate with 1 *d. f.*

If O_1, O_2 represent the observed frequencies in the two classes and e_1, e_2 the expected frequencies then (16·1) can be written as

$$z^2 = \frac{(O_1-e_1)^2}{e_1} + \frac{(O_2-e_2)^2}{e_2} \qquad \ldots(16\cdot2)$$

(16·2) suggests a very natural generalization to the case where we have k classes, instead of the two above, with $O_1, O_2, \ldots\ldots, O_k$ as the observed frequencies and $e_1, e_2, \ldots\ldots, e_k$ as the expected frequencies (with $\sum_i O_i = \sum_i e_i$). We form the statistic

$$\chi^2 = \frac{(O_1-e_1)^2}{e_1} + \frac{(O_2-e_2)^2}{e_2} + \ldots\ldots + \frac{(O_k-e_k)^2}{e_k} \qquad \ldots(16\cdot3)$$

and on the basis of the distribution of z^2 we expect that for large samples χ^2 will be distributed as a χ^2 variate with $k-1$ *d. f.* We proceed to establish this result in the next sub-section. Just as the distribution of z^2 was obtained from the binomial distribution, the distribution of χ^2 will be obtained from a consideration of the multinomial distribution.

16·2·1. Proof of the distribution of χ^2. We consider a sample of n members divided into k classes or cells and suppose that the members are distributed at random in these classes. If the probabilities for different cells are $p_1, p_2, \ldots\ldots, p_k$ then the probability p of there being O_1 members in the first cell, O_2 in the second,.....and O_k in the kth cell is the term $p_1^{O_1} p_2^{O_2} \ldots\ldots p_k^{O_k}$ in the multinomial expansion of $(p_1+p_2+\ldots\ldots+p_k)^n$ so that

$$P = \frac{n!}{O_1!\, O_2! \ldots O_k!} p_1^{O_1} p_2^{O_2} \ldots\ldots p_k^{O_k}$$

where $\sum_{i=1}^{k} O_i = n, \quad \sum_{i=1}^{k} p_i = 1.$

If the sample is sufficiently large, so that $O_1, O_2, \ldots\ldots, O_k$ are not small then using Stirling's approximation for factorials, we have

$$p \sim p_1^{O_1} p_2^{O_2} \ldots\ldots p_k^{O_k} \cdot \frac{\sqrt{2\pi} e^{-n} n^{n+1/2}}{\prod_{i=1}^{k} \left[\sqrt{2\pi}\, e^{-O_i} (O_i)^{O_i+\frac{1}{2}}\right]}$$

$$\sim C \left(\frac{np_1}{O_1}\right)^{O_1+\frac{1}{2}} \left(\frac{np_2}{O_2}\right)^{O_2+\frac{1}{2}} \ldots\ldots \left(\frac{np_k}{O_k}\right)^{O_k+\frac{1}{2}} \qquad \ldots(16\cdot4)$$

which C is a constant independent of $O_1, O_2, \ldots\ldots, O_k$. Now the expected frequency in the ith class is $e_i = np_i$

Let $$\xi_1=\frac{O_i-e_i}{\sqrt{e_i}}=\frac{O_i-np_i}{\sqrt{np_i}}$$

so that $$\chi^2=\sum_{i=1}^{n}\frac{(O_i-e_i)^2}{e_i}=\sum_{i=1}^{n}\xi_i^2.$$

Taking logarithms of both sides of (16·4), we have

$$\log_e\left(\frac{P}{C}\right)\sim\sum_i(O_i+\tfrac{1}{2})\log_e\left(\frac{e_i}{O_i}\right)$$

$$\sim-\sum_i\left[(e_i+\tfrac{1}{2}+\xi_i\sqrt{e_i})\log_e\left(1+\frac{\xi_i}{\sqrt{e_i}}\right)\right].$$

If e_i is large ξ_i will be small compared to $\sqrt{e_i}$ and to first order we have on expanding the logarithms on the right

$$\log_e\left(\frac{P}{C}\right)\sim-\sum_i(e_i+\tfrac{1}{2}+\xi_i\sqrt{e_i})\left(\frac{\xi_i}{\sqrt{e_i}}-\tfrac{1}{2}\frac{\xi_i^2}{e_i}+\ldots\ldots\right)$$

$$=-\tfrac{1}{2}\sum_i\xi_i^2-\sum_i\xi\sqrt{e_i}-O\left(\frac{1}{\sqrt{e_i}}\right),$$

where $O\left(\frac{1}{\sqrt{e_i}}\right)$ denotes quantities of the order of $1/\sqrt{e_i}$, or smaller.

Now $\sum_i\xi_i\sqrt{e_i}=\sum_i O_i-\sum_i e_i=n-n=0.$

Hence to order $1/\sqrt{e_i}$, we have

$$P\sim C\exp.\left\{-\tfrac{1}{2}\sum_{i=1}^{k}\xi_i^2\right\} \qquad \ldots(16\cdot6)$$

which shows that the probability varies as that of the sum of squares of k normal variates of unit variance which are subject to the linear constraint $\Sigma\xi_i\sqrt{e_i}=0$ but otherwise independent.

It thus appears that χ^2 as defined in (16·3) conforms to the χ^2 distribution with $(k-1)$ *d. f.* and if we write

$$\chi^2=\sum_{i=1}^{k}\frac{(O_i-e_i)^2}{e_i} \qquad \ldots(16\cdot7)$$

the variate follows the chi-square distribution

$$dP=\frac{1}{2^{\frac{1}{2}(k-1)}\left|\frac{k-1}{2}\right.}e^{-\frac{1}{2}\chi^2}(\chi^2)^{\frac{1}{2}(k-3)}\,d(\chi^2) \qquad \ldots(16\cdot8)$$

16·2·2. Remarks on the above distribution.

1. In chapter 13, we proved that sum of squares of n standard normal variates is distributed as a χ^2 variate with $n-p$ *d. f.* where p denotes the number of independent linear constraints on the variates. In the above case χ^2 is the sum of the squares of k variates with one linear constraint imposed by the fact that we require the total of the observed frequencies to be the same as the total of the expected frequencies. In our further discussion, we shall find it convenient to impose further linear constraints and the number of degrees of freedom for χ^2 as defined in (16·7) would be correspondingly reduced. Thus if m of the parameters of population are not specified, but have to be estimated from the data itself so that the observed and expected frequencies have not only to agree in the totals but also in m other statistics like mean, variance, etc., the number of degrees of freedom will be $k-1-m$, provided the constraints imposed by the process are linear.

2. In order that χ^2 test for *compatibility of observed and theoretical frequencies or of 'goodness of fit'* may be applicable, it is clear from the proof of the above distribution and further discussion of the same by Fisher that the following conditions must be satisfied :

(*i*) *The members of the sample should be independent.*

(*ii*) *Constraints on the cell frequencies, if any, should be linear.*

(*iii*) *n should be reasonably large, say, greater than 50.*

(*iv*) *The cell frequencies should not be small.*

(*v*) *In case parameter values have to be estimated from the data the method of estimation should be such as to minimize* χ^2. *This is equivalent in large sample to choosing a maximum likelihood estimator.**

Both experience and theoretical considerations show that the χ^2 approximation will be satisfactory provided all $e_i \geqslant 5$ and $k \geqslant 5$. If k less than 5 it is best to have e_i somewhat larger than 5. If some of the cell frequencies do not exceed 5, they may be combined with other cell frequencies so that the condition is satisfied.

3. If $\chi^2=0$, all the observed and the expected frequencies coincide. On the other hand, if observed and expected frequencies differ greatly, χ^2 will be large. Thus χ^2 affords a measure of the correspondence between theory and observation. We may note, however, that in the expression for χ^2 all the deviations of observed

*See §17·4·3

from theoretical frequencies are squared so that their signs are ignored. If the signs show some marked peculiarities it is desirable to supplement the χ^2 test with other tests that may be available.

4. The χ^2 table gives for any number of degrees of freedom $\chi^2_{0\cdot05}$ and $\chi_{0\cdot01}$ which will be exceeded in random sampling with probabilities 0·05 and 0·01 respectively. If the observed value of χ^2 exceeds $\chi^2_{0\ 05}$, we say that the difference between observed and theoretical frequencies is significant at 5% level.

Alternatively, we may determine the probability P of getting a value of χ^2 equal to or larger than the actual observed value. If this is small, we are justified in suspecting a significant divergence between theory and observation. If the probability is not small it does not prove the hypothesis, that is, it shows that the data are consistent with the hypothesis.

5. Not only small values of P lead us to suspect the hypothesis but a value P very near to unity may also lead to a similar result, for a large value of P will imply in general a small value of χ^2 showing too good an agreement between theory and observation to be true. Thus if $\chi^2=0$, $P=1$ but getting such a remarkable agreement can only be a rare event.

6. Since the density function of χ^2 depends only on v, the number of degrees of freedom and not on any population parameter the χ^2 test is considered sometimes as a non-parametric test, though it is often applied to parametric problems. The χ^2 distribution is concerned with the values of e_i but not with the form of the parent distributions. This property of χ^2 distribution explains why χ^2 distribution is found useful in such a large variety of cases.

7. An alternative expression for χ^2 as defined in (16·7) is

$$\chi^2=\sum_{i=1}^{k}\left[\frac{O_i^2}{e_i}\right]-2\sum_{i=1}^{k}O_i+\sum_{i=1}^{k}e_i=\sum_{i=1}^{k}\left[\frac{O_i^2}{e_i}\right]-N \quad \ldots(16\cdot9)$$

where $$N=\sum_{i=1}^{k}O_i=\sum_{i=1}^{k}e_i.$$

If the expected frequencies are not integers, (16·7) may be easier to use than (16·9).

Example 16·1. *12 dice were thrown 4096 times and a throw of 6 was reckoned as a success, the observed frequencies were as given below.*

Number of successes	0	1	2	3	4	5	6	7	& over	total
Frequencies	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* s^0 *far as the* χ^2 *test is concerned.* (*B.A. Hons., Delhi, 1954*)

On the hypothesis of unbiased dice the theoretical frequencies in 4096 throws are the terms in the binomial expansion of $4096\left(\frac{5}{6}+\frac{1}{6}\right)^{12}$ and are given below.

Number of successes	0	1	2	3	4	5	6	7	& over	total
Theoretical	495	1102	1212	808	364	116	27		8	4096

$$\therefore \chi^2=\frac{(447-459)^2}{459}+\frac{(1145-1102)^2}{1102}+\ldots+\frac{(24-27)^2}{27}+\frac{(8-8)^2}{8}$$

$$=5{\cdot}811$$

The number of classes is 8 but the total of the observed and theoretical frequencies agree. Therefore, v = number of degrees of freedom $=8-1=7$. From χ^2 table, $\chi^2_{0{\cdot}05}$ for 7 *d. f.* $=14{\cdot}07$ The calculated value of χ^2 is therefore, not significant and thus the observed frequency distribution is consistent with the hypothesis.

Example 16·2. *May the data in Example 6, page 476 be regarded as those of a random sampling from the Poisson distribution ?*

The parameter m of the Poisson distribution is not specified by the hypothesis and since the mean of Poisson distribution is m, an estimate of the population parameter is given by the sample mean *viz.*, by

$$\frac{1}{1096}[0\times 364+1\times 376+2\times 218+3\times 89+4\times 33+5\times 13+6\times 2+7\times 1]$$

$$=\frac{1}{1096}[1295]=1{\cdot}18.$$

The theoretical frequencies, with this value of the parameter, are

$$1096\times\frac{e^{-1{\cdot}18}(1{\cdot}18)^r}{r\ !},\ r=0,\ 1,\ 2,\ldots\ldots,\ 7\ ;\ e^{-1{\cdot}18}={\cdot}3075$$

and are given below :

No. of deaths per day	0	1	2	3	4	5	6	7	Total
Observed frequency	364	376	218	89	33	13	2	1	1096
Expected frequency	336·5	397·1	234·6	92·5	27·3	6·5	1·3	0·2	1096

Since the expected frequeucies in the last two classes are less than five, we group together the last three classes, then

$$\chi^2=\frac{(364-336\cdot5)^2}{336\cdot5}+\frac{(376-397\cdot1)^2}{397\cdot1}+\ldots+\frac{(16-8)^2}{8}=9\cdot99.$$

Here the number of the classes is 6 (since the last three classes have been grouped) but observed and the expected frequencies agree both in their totals as well as their means so that

$$\nu=\text{number of degrees of freedom}=6-1-1=4.$$

From the table, the value of χ^2 for for 4 *d. f.* at 5% level is 9·49 The calculated value is, therefore, significant at 5% level, but is barely significant there.

Ex. 1. A die was thrown 60 times and the following frequency distribution was observed :

Face	1	2	3	4	5	6
Observed frequency	15	6	4	7	11	17

Can you regard the die as honest ? [$x^2=13\cdot6, \nu=5$, significant at 5% level.]

16·3. χ^2 Test as a Test of Independence. A special and important class of classes where compatibility of observed and expected frequencies has to be tested is testing the independence of two attributes when for convenience the data is arranged in two way tables usually called *contingency tables.*

16·3·1. Contingency table. Let the data be classified into classes $A_1, A_2, \ldots, A_s$ according to attribute A and into t classes $B_1, B_2, \ldots, B_t$ according to attribute B. Let O_{ij} denote the observed frequency of the cell belonging to both the classes A_i and B_j [$i=1, 2, \ldots s$; $j=1, 2, \ldots, t$.]

Let (A_i) and (B_j) denote the totals of all the frequencies belonging to classes A_i and B_j respectively, then the data can be set into a $s \times t$ contingency table of s rows and t columns as follows :

Classes	B_1	B_2 ..... B_j ..... B_{t-1}	B_t	Totals
A_1	O_{11}	O_{12} ..... O_{1j} ..... O_{1t-1}	O_{1t}	(A_1)
A_2	O_{21}	O_{22} ..... O_{2j} ..... O_{2t-1}	O_{2t}	(A_2)
⋮	.....	.....	.....	⋮
A_i	O_{i1}	O_{i2} ..... O_{ij} ..... O_{it-1}	O_{it}	(A_i)
⋮	.....	.....	.....	⋮
A_{s1}	O_{s-11}	O_{s-12} ..... O_{s-1j} ..... O_{s-1t-1}	O_{s-1t}	(A_{s-1})
A_s	O_{s1}	O_{s2} ..... O_{sj} ..... O_{st-1}	O_{st}	(A_s)
Totals	(B_1)	(B_2) ..... (B_j) ..... (B_{t-1})	(B_t)	N

16·3·2. Calculation of χ^2 to test the hypothesis of independence. To test whether there is any relationship between the attributes A and B, we set up the null hypothesis of no relationship or of independence between A and B, and calculate the expected frequencies in all the cells on the hypothesis. Since we wish to test independence only and not any hypothesis as to the total number belonging to different classes the expected frequencies are calculated by keeping the class totals fixed. This is equivalent to imposing linear constraints that the class totals of the observed and expected frequencies must agree.

Let e_{ij} denote the expected frequency in the cell common to A_i and B_j. Then since A and B are independent, in any class the proportions of members belonging to $A_1, A_2, \ldots\ldots, A_i, \ldots\ldots, A_s$ should be same and equal to the proportions in the total. Thus

$$\frac{e_{1j}}{(A_1)}=\frac{e_{2j}}{(A_2)}=\ldots=\frac{e_{ij}}{(A_i)}=\ldots=\frac{e_{sj}}{(A_s)}=\frac{\Sigma e_{ij}}{\Sigma(A_i)}=\frac{(B_j)}{N}$$

so that
$$e_{ij}=\frac{(A_i)\times(B_j)}{N} \qquad \ldots(16\cdot10)$$

Thus on the basis of independence the expected frequency in any cell is equal to the product of the class totals of the two classes to which the cell belongs divided by the total number of observations.

Hence
$$\chi^2=\sum_{j=1}^{t}\sum_{i=1}^{s}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right]$$

$$=\sum_{j=1}^{t}\sum_{i=1}^{s}\frac{\left[O_{ij}-\frac{(A_i)\times(B_j)}{N}\right]^2}{\frac{(A_i)\times(B_j)}{N}} \qquad \ldots(16\cdot11)$$

16·3·3. Number of degrees of freedom associated with a $s\times t$ contingency table. There are in all $s\times t$ cells but since the marginal totals are fixed there are $s+t$ constraints. These constraints are, however, not independent since the sum of the border column frequencies must be equal to that of the border row frequencies and thus there are only $s+t-1$ independent linear constraints. Hence ν, the number of degrees of freedom, associated with a $s\times t$ contingency table is

$$\nu=st-(s+t-1)=(s-1)(t-1).$$

This is the number of the cells whose frequencies can be arbitrarily assigned. This is obvious from the above table where the frequencies of $(s-1)(t-1)$ cells enclosed by the dotted line can be fixed arbitrarily and when the frequencies of the remaining $s+t-1$ cells will be determined from the class totals.

Example 16·3. *In a locality 100 persons were randomly selected and asked about their educational achievements. The results are given as follows :*

		Education			
		Middle	*High School*	*College*	*Totals*
Sex	*Male*	10	15	25	50
	Female	25	10	15	50
	Total	35	25	40	100

Can you say that education depends on sex ?

(M.A., Delhi, 1954)

We set up the null hypothesis that there is no association between education and sex. On this hypothesis the expected frequencies are tabulated as follows :

	Middle	*High School*	*College*	*Total*
Male	17·5	12·5	20	50
Female	17·5	12·5	20	50
Total	35	25·0	40	100

$$\chi^2=\frac{(10-17\cdot5)^2}{17\cdot5}+\frac{(15-12\cdot5)^2}{12\cdot5}+\frac{(25-20)^2}{20}+\frac{(25-17\cdot5)^2}{17\cdot5}+\frac{(10-12\cdot5)^2}{12\cdot5}+\frac{(15-20)^2}{20}=9\cdot93 \text{ approximately.}$$

ν=the number of degrees of freedom$=(s-1)(t-1)=(3-1)(2-1)=2$.

From the table for 2 *d. f.*, $\chi^2_{0\cdot05}=5\cdot99$ and $\chi^2_{0\cdot01}=9\cdot21$.

Thus the calculated value of χ^2 is highly significant and the hypothesis that the education does not depend on sex is discredited.

Example 16·4. *Show that in a 2×2 contingency table wherein the frequencies are $\frac{a}{c}\Big|\frac{b}{d}$, χ^2 calculated from independent frequencies is*

$$\chi^2=\frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)} \qquad \ldots(16\cdot12)$$

(M.A., Punjab, 1958 ; M.Sc., B.Sc., Agra, 1956, 59 ; M.A., Patna 1957 ; B.A. Hons., Delhi, 1947, 49, 52, 59, 61, 66 ; M.A., Delhi, 1951)

Since the marginal totals are fixed the probability for a member to belong to class A is $(a+c)/N$ and for it to belong to class B is $(a+b)/N$.

Classes	A	$\overline{A}$	Totals
B	a	b	$a+b$
$\overline{B}$	c	d	$c+d$
Totals	$a+c$	$b+d$	$a+b+c+d=N$

Further the attributes being independent, the probability for it to belong to both classes A and B is $\left(\frac{a+b}{N}\right)\left(\frac{a+c}{N}\right)$, and therefore, the expected frequency of this class, denoted by $E(a)$, is given by

$$E(a)=N\times\frac{a+b}{N}\times\frac{a+c}{N}=\frac{(a+b)(a+c)}{N}$$

Similarly, $E(b)=\dfrac{(a+b)(b+d)}{N}$, $E(c)=\dfrac{(c+d)(a+c)}{N}$

$$E(d)=\frac{(c+d)(b+d)}{N}.$$

By definition,

$$\chi^2=\frac{[a-E(a)]^2}{E(a)}+\frac{[b-E(b)]^2}{E(b)}+\frac{[c-E(c)]^2}{E(c)}+\frac{[d-E(d)]^2}{E(d)}$$

$$=\frac{\left[a-\dfrac{(a+b)(a+c)}{a+b+c+d}\right]^2}{\dfrac{(a+b)(a+c)}{a+b+c+d}}+\text{three similar terms.}$$

$$=\frac{(ad-bc)^2}{(a+b+c+d)}\left[\frac{1}{(a+b)(a+c)}+\frac{1}{(a+b)(b+d)}+\frac{1}{(c+d)(a+c)}+\frac{1}{(c+d)(d+b)}\right]$$

$$=\frac{(ad-bc)^2}{(a+b+c+d)}\left[\frac{a+b+c+d}{(a+b)(a+c)(b+d)}+\frac{a+b+c+d}{(a+c)(b+d)(c+d)}\right]$$

$$=\frac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(b+d)(c+d)}.$$

Ex. 1. Show that in § 16·3·3 the linear constraints on expected frequencies are

$$\sum_{j=1}^{t}e_{ij}=(A_i),\quad \sum_{i=1}^{s}e_{ij}=(B_j),$$

$$\sum_{i=1}^{s}\sum_{j=1}^{t}e_{ij}=\sum_{i=1}^{s}(A_i)=\sum_{j=1}^{t}\sum_{i=1}^{s}e_{ij}=\sum_{j=1}^{t}(B_j)=N$$

Show that only $s+t-2$ constraints are independent.

Ex. 2. Prove that the marginal totals are held fixed,

P (a member belonging to class A_i) $=\dfrac{(A_i)}{N}$.

P (a member belonging to class B_j) $=\dfrac{(B_j)}{N}$.

Hence on the basis of independence of A and B, prove that

$$e_{ij}=\frac{(A_i)\times(B_j)}{N}.$$

Ex. 3. Deduce the value of χ^2 for a 2×2 table from that of a $2\times n$ table.

Ex. 4. Prove that for a $s\times t$ table, χ^2 is less than or equal to $N(s-1)$ or $N(t-1)$ whichever is less. *(M.A., Delhi, 1946)*

[**Hint.** Using the facts that $e_{ij}=\dfrac{(A_i)\times(B_j)}{N}$, $O_{ij}\leqslant A_i$, $O_{ij}\leqslant B_j$

we have $\chi^2=\Sigma\Sigma\dfrac{(O_{ij}-e_{ij})^2}{e_{ij}}=\Sigma\Sigma\dfrac{O_{ij}^2}{e_{ij}}-N=N\Sigma\Sigma\left[\dfrac{O_{ij}}{(A_i)}\times\dfrac{O_{ij}}{(B_j)}\right]-N$

$$\leqslant N\left\{\sum_{i=1}^{s}\sum_{j=1}^{t}\left(\frac{O_{ij}}{(A_i)}\right)-1\right\}=N\left\{\sum_{i=1}^{s}\left(\frac{A_i}{A_i}\right)-1\right\}$$

$\therefore\ \chi^2\leqslant N(s-1)$. Similarly, $\chi^2\leqslant N(t-1)$]

16·4. Further Consideration of 2×2 Contingency Tables.

16·4·1. *Yates Correction for Continuity. The distribution of χ^2 as obtained in chapter 13 is a continuous distribution while the distribution of frequencies is by its very nature discontinuous. The continuous χ^2 distribution may be regarded as the limit to which the true discontinuous distribution tends as the sample size increases. One of the reasons why we postulated earlier that the expected cell frequencies should not be less than 4 was to avoid irregularities due to this discontinuity.

In a 2×2 table there is only one *d. f.* and the frequency of only one cell can be assigned arbitrarily and if the frequency of a cell is small, say 3, the sum of the probabilities for 3 or more extreme (*i.e.*, 2, 1, 0) observations in the discontinuous case can be approximated by the tail area of the continuous χ^2 curve corresponding to $3\frac{1}{2}$ frequency. F. Yates, in 1934, therefore, suggested the following adjustment to the 2×2 contingency table.

Calculate ad and bc and add $\frac{1}{2}$ *to both factors of the smaller and subtract* $\frac{1}{2}$ *from both factors of the larger.* This does not change the marginal totals. The correction is known as *Yates correction for continuity.*

Example 16·5 *Two batches each of 13 experimental animals are inoculated and the other not inoculated were exposed to the infection of a disease. The following frequencies of dead and surviving animals were noted in the two cases. Can the inoculation be regarded as effective against the disease ?*

	Dead	*Survived*	*Total*
Inoculated	2	10	12
Not Inoculated	8	4	12
Total	10	14	24

After applying Yates correction the adjusted table is :

	Dead	*Survived*	*Total*
Inoculated	2·5	9·5	12
Not Inoculated	7·5	4·5	12
Total	10	14	24

Applying formula of Example 16·4 or otherwise

$$\chi^2=\frac{(\frac{3}{2}\times\frac{8}{2}-\frac{12}{2}\times\frac{7}{2})^2\times 24}{12\times 12\times 10\times 14}=4\cdot 286$$

For 1 *d. f.* the value of χ^2 at 5% level is the square of the standard normal variate for 5% *i.e.* of 1·96. Since 4·286 is greater than 3·84, χ^2 is significant at 5% level. Hence at this level, the inoculation may be said to be effective against the disease.

16·4·2. Fisher Exact Method for 2×2 contingency table. The testing of independence by χ^2 test in a 2×2 table is an approximation which is useful for the comparative simplicity of the calculations. The exact method is somewhat more laborious though necessary in case of doubt and valuable as displaying the true nature of the inference which the method of χ^2 is designed to draw.

2×2 table

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d=n$

The number of ways in which the body of this table can be completed is

$$\frac{n\,!}{a\,!\,b\,!\,c\,!\,d\,!},\quad n=a+b+c+d$$

and assuming independence of the two attributes the total number of ways in which marginal totals can be obtained is

$$\frac{n\,!}{(a+c)(b+d)\,!}\times\frac{n\,!}{(a+b)(c+d)\,!}$$

Consequently the probability of the distribution of the table is

$$\frac{(a+c)(b+d)\,!\,(a+b)(c+d)\,!}{a\,!\,b\,!\,c\,!\,d\,!\,(a+b+c+d)\,!}$$

and from here we can find the probabilities for $a=0, 1, 2,\ldots\ldots$

Example 16·6. *For the data of Example* 16·5 *compare the result of the exact treatment with that obtained from the* χ^2 *approximation.*

The probability for the number of dead-inoculated to be, or less corresponds to $a=0$, 1 and

$$=\frac{12\,!\,12\,!\,10\,!\,14\,!}{24\,!\,0\,!\,12\,!\,10\,!\,2\,!}+\frac{12\,!\,12\,!\,10\,!\,14\,!}{24\,!\,1\,!\,11\,!\,9\,!\,3\,!}+\frac{12\,!\,12\,!\,10\,!\,14\,!}{24\,!\,21\,!\,10\,!\,8\,!\,4\,!}$$

$$=\frac{12\,!\,14\,!}{24\,!\,2\,!}\left[1+\frac{12\times 10}{1\times 3}+\frac{12\times 11\times 10\times 9}{1\times 2\times 3\times 4}\right]=\cdot 018.$$

Without correction $\chi^2=6\cdot 172$ and the corresponding standard normal variate is 2·48 giving $P=\cdot 007$ with correction $\chi^2=4\cdot 286$ corresponding to $P=\cdot 02$ nearly. Thus Yates correction gives a great improvement. It may be noted here that since in the exact test we

have found the probabilities of a being, or less, to find the probability for a given χ^2 we use either one tail of the normal curve or half the probability obtained from the χ^2 table.

Ex. 1. (*a*) If p is the probability of A and $q(=1-p)$ that of $\bar{A}$, prove that if the two attributes are independent, the joint probability of a and c belonging to A out of $a+b+c+d$ respectively is

$$\frac{(a+b)\,!\,(c+d)\,!}{a\,!\,b\,!\,c\,!\,d\,!}p^{a+c}q^{b+d}$$

(*b*) Summing over possible samples with same marginal totals show that

$$\Sigma\frac{1}{a\,!\,b\,!\,c\,!\,d\,!}=\frac{n\,!}{(a+b)\,!\,(c+d)\,!\,(a+c)\,!\,(b+d)\,!}$$

(*c*) Deduce the result of Ex. 16·4·2.

Ex. 2. For Hellman's data $\begin{array}{c|c}4&16\\1&21\end{array}$ show that the exact test gives $P=$·1435, χ^2 uncorrected gives $P=0$·0612 and χ^2 after Yates correction gives $P=0$·1427.

Ex. 3. For Langes data $\begin{array}{c|c}10&8\\2&15\end{array}$, show that

$[P]$ exact	=·00046
$[P]$ χ^2 uncorrected	=·00015
$[P]$ χ^2 corrected	=·00061

Ex. 4. Show that for a 2×2 table the value of χ^2 after applying Yates correction for a continuity is

$$\frac{\left(ad-bc-\frac{N}{2}\right)^2N}{(a+b)(a+c)(b+d)(c+d)}\text{ or }\frac{\left(ad-bc+\frac{N}{2}\right)^2N}{(a+b)(a+c)(b+d)(c+d)}$$

according as $ad-bc\geqslant0$ or $\leqslant0$.

16·5. Other Tests of Significance Based on χ^2 Distribution.

16·5·1. Test for variance of a normal population. Given a random sample $x_1, x_2, \ldots\ldots, x_n$ from a normal population, we have to test the hypothesis that the variance of the population is σ^2.

If the hypothesis is correct, the statistic

$$\chi^2=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma^2}=\frac{ns^2}{\sigma^2}.$$

is distributed as a χ^2 variate with $(n-1)$ $d.f.$ If the calculated value of χ^2 is less than $\chi^2_{0\cdot05}$ for $(n-1)$ $d.f.$, the data will be consistent with the hypothesis.

Example 16·7. *Explain how you will find confidence intervals for the variance of a normal population.*

From χ^2 distribution with $n-1$ $d.f.$ we can find two numbers a and b corresponding to any given confidence coefficient α such that

$$P\,[a<\chi^2<b]=\int_a^b \frac{1}{2^{\frac{1}{2}(n-1)}\left|\frac{n-1}{2}\right.} e^{-\frac{1}{2}\chi^2}\,(\chi^2)^{\frac{1}{2}(n-3)}\,d(\chi^2)=\alpha,$$

so that
$$P\left[\frac{\Sigma(x_i-\bar{x})^2}{b}<\sigma^2<\frac{\Sigma(x_i-\bar{x})^2}{a}\right]=\alpha$$

giving a confidence interval σ^2 of length $\Sigma(x_i-\bar{x})^2\left(\frac{1}{a}-\frac{1}{b}\right)$.

In practice we may choose a and b so that the two tail areas are equal. Thus for a 95 per cent confidence interval, $a=\chi^2_{\cdot 975}$, $b=\chi^2_{\cdot 025}$.

Ex. A random sample of 15 gives an unbiased estimate of sample variance equal to 1·05 cm^2. Find a 90 per cent confidence interval for σ^2.

16.5.2. Combination of probabilities from independent tests of significance. Sometimes we have a number of independent tests of significance giving different probabilities, some of which may give significant and others non-significant results. Our problem is to combine the various probabilities, to get a single probability from which to decide on the significance of the aggregate.

$$\text{For 2 } d.f.,\ p_1=P[\chi^2\geqslant\chi_1^2]=\tfrac{1}{2}\int_{\chi^2}^{\infty} e^{-\frac{1}{2}\chi^2}\,d(\chi^2)=e^{-\frac{1}{2}\chi_1^2},$$

so that $\chi_1^2=-2\log_e p_1=2\log_e(1/p_1)$

Similarly for the other probabilities $p_2, p_3, \ldots\ldots, p_k$ we have

$$\chi_2^2=-2\log_e p_2$$
$$\ldots\ldots\ldots\ldots$$
$$\chi_k^2=-2\log_e p_k$$

so that $\chi^2=\chi_1^2+\chi_2^2+\ldots\ldots+\chi_k^2=-2\log_e(p_1p_2\cdots p_k)$.

But the sum of independent χ^2 variates is again a χ^2 variate.

$$\therefore\quad \log_e\left(\frac{1}{p_1p_2\cdots p_k}\right)^2=2{\cdot}3026.\ \log_{10}\left(\frac{1}{p_1p_2\cdots p_k}\right)^2$$

is a χ^2 variate with $2k$ $d.f.$ and if this value exceeds $\chi^2_{0\cdot 05}$ for $2k$ $d.f.$ the combined result will be said to be significant at 5% level.

Example 16·8. *Three independent tests of significance give the probabilities ·15, ·26, ·11 ; test whether the aggregates of these tests is significant.*

We have

P	$-2\log_e P$	$D.F.$
·15	3·79424	2
·26	2·69414	2
·11	4·41454	2
	10·90292	6

The 5% and 1% values of χ^2 for 6 *d. f.* are 12·592 and 10·645 respectively. Therefore, the probability of the combined test is slightly less than 0·10.

Ex. Four tests of significance have yielded probabilities ·145, ·263, ·087 and ·075. Test whether the aggregate of these four tests should be regarded as significant. (*M.A , Bombay, 1952*)

16·5·3. Testing homogeneity of several independent estimates of population variance. Bartlett Test. Let k independent samples give unbiased estimates $S_1^2, S_2^2, \ldots\ldots, S_k^2$ of the population variance based on $v_1, v_2, \ldots\ldots, v_k$ *d. f.*, then on the hypothesis that the samples are from the same population,

$$S^2 = \frac{\Sigma v_i S_i^2}{\Sigma v_i} = \frac{\Sigma v_i S_i^2}{v}$$

is also an unbiased estimate of the population variance σ^2 since

$$E(S^2) = \frac{1}{v}\sum_{i=1}^{n} v_i[E(S_i^2)] = \frac{1}{v}\sum_{i=1}^{n} v_i\sigma^2 = \sigma^2.$$

Bartlett has shown that on the hypothesis that the various samples come from the same population, the statistic

$$\sum_i \left(v_i \log_e \frac{S^2}{S_i^2} \right) \Big/ \left[1 + \frac{1}{3(k-1)} \left\{ \left(\sum_{i=1}^{n} \frac{1}{y} \right) \frac{1}{y} \right\} \right]$$

is distributed approximately as χ^2 with $k-1$ *d. f.*

The value of χ^2 calculated from the data would thus enable us to test the homogeneity of the several estimates.

It is clear, however, that if the largest and the smallest variance of a set do not differ significantly as tested by the F test, then those lying between these cannot differ significantly either, and the whole set can be reasonably regarded as coming from a single population. In these circumstances there is no need to employ the Bartlett test.

Ex. Three independent samples give the following results :—

Size	*Observations*
5	34, 40, 47, 60, 84
9	40, 59, 60, 67, 86, 92, 95, 93, 108
3	46, 93, 100

Prove that $S_1^2 = 324$, $S_2^2 = 512{\cdot}25$, $S_3^2 = 862{\cdot}833$, $S^2 = 528{\cdot}476$, $\chi^2 = 0{\cdot}391$, $v = 2$, Deduce that the apparent difference in variance is not significant.

16·5·4. Testing homogeneity of several independent estimates of an ordinary or partial correlation coefficient.

Let $r_1, r_2, \ldots\ldots, r_k$ be k independent estimates of the correlation coefficient. We have to test whether these can be regarded as having come from the same bivariate population.

Let z_i, $i=1, 2,\ldots, k$, and $\bar{z}$ be defined by

$$z_i=\tfrac{1}{2}\log_e\left(\frac{1+r_i}{1-r_i}\right)$$

$$\bar{z}=\frac{\Sigma(n_i-3)z_i}{\Sigma(n_i-3)}.$$

Then since for large samples $z_1, z_2,\ldots\ldots, z_k$ are all distributed normally about the same mean, the sum

$$\Sigma(n_i-3)\,(z_i-\bar{z})^2$$

is distributed approximately as χ^2 with $(k-1)$ *d. f.* and the test of independence can be carried out accordingly.

For partial coefficient of order k, n_i will be replaced by n_i-k.

Example 16·9. *The correlation coefficient between daily ration of green grass and rate of growing calves on the basis of observations taken on 10, 14, 16, 20, 25 and 28 cows at six farms were found to be 0·318, 0·106, 0·253, 0·340, 0·116 and 0·112. Can these be considered homogeneous ? If so, estimate the common correlation coefficient.* (*M.A. Delhi, 1967*)

We set up the hypothesis that the given values are homogeneous *or* the samples are drawn from equally correlated population.

We have

$$z_i=\frac{1}{2}\log_e\left(\frac{1+r_i}{1-r_i}\right)$$

$\therefore$ $$z_1=\frac{1}{2}\log_e\left(\frac{1+0\cdot318}{1-0\cdot318}\right)$$

$$=\frac{1}{2}\log_e\times\log\left(\frac{1\cdot318}{\cdot682}\right)$$

$$=1\cdot1513\ (\log 1\cdot318-\log\cdot682)$$

$$=\cdot3294 \text{ on simplification}$$

Similarly $z_2=\cdot1063$, $z_3=\cdot2586$, $z_4=\cdot354$

$z_5=\cdot1165$, $z_6=\cdot1125$

or $$\bar{z}=\frac{\Sigma(n_i-3)z_i}{\Sigma(n_i-3)}$$

$=0\cdot1919$, on simplification

Also $\chi^2_{6-1}=\Sigma(n_i-3)\,(z_i-\bar{z})^2=1\cdot008$,

on simplification.

But tabulated χ^2_5 at 5% level of significance $=11\cdot07$

Hence the null hypothesis appears to be correct.

In other words, the sample correlation coefficents are homogeneous.

Again $$\bar{z}=\frac{1}{2}\log_e\left(\frac{1+\rho}{1-p}\right)$$

where $$\bar{\rho}=\frac{e^{2\bar{z}}-1}{e^{2\bar{z}}+1}=\frac{1\cdot468-1}{1\cdot468+1}=0\cdot1894$$

Hence the pooled estimate of population correlation coeffi is 0·1894.

Ex. From examples of 24, 29, 33, 38, 42 a partial correlation coefficient of order 3 is found to be 0·38, 0·54, 0·60, 0·50, 0·42.

Test their homogeneity and find the pooled estimate.

16·6. Association of Attributes and Contingency

16·6·1. Measures of association for contingency tables.

For measuring the degree of association between two attributes A and B, the data may be arranged in a 2×2 table.

	B's	not-B's or β's	Totals
A's	$(AB)=a$	$(A\beta)=b$	$a+b$
not-A's or α's	$(\alpha B)=c$	$(\alpha\beta)=d$	$c+d$
Totals	$a+c$	$b+d$	$a+b+c+d=N$

The three measures of association that have to be proposed are :

$$Q=\frac{ad-bc}{ad+bc} \qquad [\textit{Yule's coefficient of association}]$$

$$Y=\frac{1-\sqrt{\frac{bc}{ad}}}{1+\sqrt{\frac{bc}{ad}}} \qquad [\textit{Yule's coefficient of colligation}]$$

$$V=\frac{ad-bc}{+[(a+b)(a+c)(b+d)(c+d)]^{\frac{1}{2}}}$$

For a $s\times t$ contingency table, we may use the *square contingency* χ^2 or the *mean-square contingency* χ^2/N. Karl Pearson proposed another measure

$$C=\sqrt{\frac{\chi^2}{N+\chi^2}} \qquad \ldots(16\cdot13)$$

called Pearson's *coefficient of contingency.*

Tschuprow proposed the coefficient

$$T=\left[\frac{\chi^2}{N(s-1)(t-1)}\right]^{\frac{1}{2}} \qquad \ldots(16\cdot14)$$

for the same purpose. The value C lies between 0 and 1.

Ex. 1. In a 2×2 table the association is called positive or negative according as

$$\delta = a - \frac{(a+b)(a+c)}{N} > \text{or} < 0$$

Show that Q, Y, V are positive or negative, according as the association is positive or negative.

Ex. 2. Show that Q, Y, V all vanish when the two attributes are independent.

Ex. 3. Show that in the table $\left|\begin{array}{c|cc} A & a & b \\ \hline \text{not } A & c & d \end{array}\right|$ Q and Y become $+1$ if $bc=0$ in which case there is *complete association*, and these become -1 if $ad=0$ in which case there is *complete disassociation*. (*B. Sc. Bombay, 1968*)

Ex. 4. Show that $V=1$ if $b=0$, $d=0$ (*i.e.*, if all A's are B's and *vice versa*) and $V=-1$ if $a=0$, $d=0$ (*i.e.* if no A's and B's are B's are A's). When $V=1$. the association is said to be *absolute*.

Ex. 5. Prove that $Q=\dfrac{2Y}{1+Y^2}$. Hence or otherwise show that the coefficient of association is greater in absolute value than the coefficient of the colligation except when both are zero or unity in absolute value. (*B. Sc. Kerala 1969 S*)

Ex. 6. (*a*) Prove that χ^2 vanishes if and only if the two attributes are independent.

(*b*) For a 2×2 table, prove that $V^2=\chi^2/N$.

Ex. 7. Prove that if $s<t$, then $T<\left(\dfrac{s-1}{t-1}\right)^{\frac{1}{2}}$ and hence that T can be unity only if $s=t$.

EXERCISES 16

1. 200 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
Frequencies	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 table from which these numbers were taken. (*M.Sc., Agra 1947*)

[Ans. $\chi^2=4{\cdot}3$, $v=9$. Hypothesis appears to be correct]

2. The following table gives the number of aircraft accidents that occurred during the various days of the week. Find whether the accidents are uniformly distributed over the week.

Days	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Total
No. of accidents	14	16	8	12	11	9	14	84

(*I.C.A.R., Delhi, 1951*)

[Ans. Accidents seem to be uniformly distributed over all days of the week]

3. Four coins were tossed at a time and this operation is repeated 160 times. It is found that 4 heads occur 6 times, 3 heads occur 43 times, 2 heads occur 69 times, one head occurs 34 times. Discuss whether the coins may be regarded as unbiased. (*Certificate Exam. Calcutta, 1952* ; *Audit & Acctts., 1945*)

[Ans. Coins are unbiased]

4. Five dice were thrown 96 times and the number of times 4, 5 or 6 were thrown are given below :

No. of dice throwing 4, 5 or 6	5	4	3	2	1	0
Frequency	7	19	35	24	8	3

Calculate the value of χ^2 up to two places of decimals. (*M.Sc., Agra 1946*)

[The expected frequencies are given by the various terms in the expansion of 96 $(\frac{1}{2}+\frac{1}{2})^5$. Combining the first two classes and the last two classes,

$$\chi^2=\frac{(26-18)^2}{18}+\frac{(35-30)^2}{30}+\frac{(24-30)^2}{30}+\frac{(11-18)^2}{18}=8\cdot31.]$$

5. In a biochemical experiment 20 insects were put each into 100 jars and were subjected to fumigant. After 3 hours, the number of living insects in each jar was counted. The distribution is shown below. Can this be fitted by a Binomial distribution ?

No. of insects alive x	0	1	2	3	4	5	6	7	8	9	Total
No. of jars f	3	8	11	15	16	14	12	11	9	1	100

[If every insect had an equal chance p of surviving them since the mean np of the population has to be estimated from the mean of the distribution.

$$np=\frac{\Sigma f_i x_i}{\Sigma f_i}=\frac{439}{100}$$

$\therefore$ $p=0\cdot2195$ and the theoretical frequencies obtained from $100(0\cdot7805+0\cdot2195)^{20}$ are 0·7, 4·0, 10·6, 17·8, 21·3, 19·2, 13·5, 7·6, 3·5, 1·3, 0·4, 0·1. Combining the first three and the last three classes $\chi^2=11\cdot36$, $v=4$.]

6. The following data shows suicides of women in 8 German States during 14 years :

State per year	0	1	2	3	4	5	6	7	Total
Observed frequency	364	376	218	89	33	13	2	1	1096

Fit a Poisson distribution to the data and show that the fit is not satisfactory.

7. Show that for a $2\times n$ table

$$\chi^2=\sum_r\left[N_1N_2\frac{\left(\frac{a_{1r}}{N_1}-\frac{a_{2r}}{N_2}\right)^2}{a_{1r}+a_{2r}}\right]$$

where a_{1r} and a_{2r} are the frequencies in the rth column and N_1 and N_2 are the border sums of the two rows.

(*B.A., Hons., Delhi, 1953* ; *M.Sc., Agra, 1949, 53*)

8. Show that for entries in the $2\times r$ contingency table.

							Totals
	a_1	a_2	.....	a_i	.....	a_r	a
	b_1	b_2	.....	b_i	.....	b_r	b
Totals	n_1	n_2	.....	n_i	.....	n_r	n

the value of χ^2 for testing homogeneity is

$$\chi^2 = \sum_{i=1} w_i(p_i - p)^2$$

where $p_i = \frac{a_i}{n_i}$, $p = \frac{a}{n}$, $w_i = \frac{n_i}{pq}$, $q = \frac{b}{n}$, $q_i = 1 - p_i$.

(M.A., Patna. M.A., Delhi, 1958 ; I.S.I., Calcutta, 1953 ; I.C.A.R., Delhi, 1947, 52)

$$\text{[Hint. } \chi^2 = \sum_{i=1}^{r} \left[\frac{\left(a_i - \frac{n_i a}{n} \right)^2}{\left(\frac{n_i a}{n} \right)} \right] + \sum_{i=1}^{r} \left[\frac{\left(b_i - \frac{n_i b}{n} \right)^2}{\left(\frac{n_i b}{n} \right)} \right]$$

$$= \Sigma \left[n_i \frac{(p_i - p)^2}{p} \right] + \Sigma \left[n_i \frac{(q_i - q)^2}{q} \right] = \text{etc.]}$$

9. Prove Brandt and Snedecor's formula for a $2 \times r$ contingency table with a, a' as any pair of observed frequency and n, n' the corresponding totals.

$2 \times r$ Table				Total
a_1,	a_2,...	a_i.....	a_r	n
a_1'	a_2'...	a_i'.....	a_r'	n'

$$\chi^2 = \frac{1}{\bar{p}\bar{q}} \left[\sum_{i=1}^{r} (a_i p_i) - n\bar{p} \right]$$

where $\bar{p} = \frac{n}{n+n'}$, $p = \frac{a_i}{a_i + a_i'}$, $\bar{q} = 1 - \bar{p}$, $q = 1 - p$.

(M.A. Delhi, 1959 ; B.A. Hons., Delhi, 1950)

Use this to prove that for the following 2×3 table, $\chi^2 = 1{\cdot}51$

				Total
	42	13	33	88
	20	8	25	53
Total	62	21	58	141

10. The following table occurs in a memoir of Karl Pearson :

		Eye colour in sons	
		No light	Light
Eye colour in fathers	No light	230	148
	Light	151	471

Test whether the colour of the son's eyes is associated with that of the father's ($\chi^2=3{\cdot}84$, $v=1$). *(I.A.S., 1942)*

[**Ans.** The colour of son's eye appears to be associated with that of the father's].

11. In an experiment on immunization of cattle from tuberculosis the following results were obtained :

	Affected	Unaffected
Inoculated	12	26
Not Inoculated	16	6

Examine the effect of vaccine in controlling susceptibility to tuberculosis. *(I.A.S., 1948, 63)*

[**Ans.** The vaccine is effective in controlling susceptibility to tuberculosis.]

12. In an experiment on the immunization of goats from anthrax the following results, were obtained. Derive your inference on the efficiency of the vaccine :

	Died	Survived	
Inoculated	2	10	12
Not Inoculated	6	6	12
Total	8	16	24

(I.A. & A.S., 1943)

[**Ans.** $\chi^2=3$ without Yates correction,

$\chi^2=1{\cdot}7$ with Yates correction, both are significant]

13. The normal rate of infection for a certain disease in cattle is known to be 50%. In an experiment with seven animals injected with a new vaccine, it was found that none of the animals caught infection. Can the evidence be regarded as conclusive at 1% level of significance to prove the value of the new vaccine ? *(I.A.S., 1949 ; I. Audit, 1942)*

[**Ans.** The evidence may be regarded as conclusive.]

14. An operation's research group studying the libraries of a certain university found that the number (x) of books withdrawn by individual users during any one visit to the science library was distributed with probability function $f(x)=(0{\cdot}4)(0{\cdot}6)^x$, $x=0, 1, 2.$

Test this hypothesis for goodness of fit using the following data, which represents an independent random sample of 100 observations.

Value of x :	0	1	2	3	4 or more
Frequency	37	26	17	5	15

(M.A., Delhi, 1969)

15. In experiments on pea breeding, Mendel obtained the following frequencies of seeds : 315 round and yellow, 101 wrinkled and yellow, 108 round and green, 32 wrinkled and green. Total 556. Theory predicts that the frequencies should be in the ratio 9 : 3 : 3 : 1. Find χ^2 and examine correspondence between theory and experiment. *(M.Sc., Agra, 1952)*

[**Ans.** The theoretical frequencies are approx. 313, 104, 104, 35 and $\chi^2=0{\cdot}51$ $\nu=3$. There appears to be correspondence between theory and experiment.]

16. (*a*) A certain hypothesis was tested by two similar experiments which gave $\chi^2=14{\cdot}7$ for $\nu=9$ and $\chi^2=14{\cdot}9$ for $\nu=11$. Show that the two experiments combined gave less reason for confidence in the hypothesis than either experiment alone.

(*b*) A certain hypothesis is tested by the similar experiments. These gave $\chi^2=11{\cdot}9$ for $\nu=6$, $\chi^2=14{\cdot}2$ for $\nu=8$ and $\chi^2=18{\cdot}3$ for $\nu=11$. Show that the three experiments together provide more justification for rejecting the hypothesis than any one experiment alone.

17. According to the genetic model the proportions in three groups should be $p^2 : 2pq : q^2$ where $p+q=1$. Are the data consistent with the sample 9, 51, 45 ?

18. According to the genetic model the proportions of individuals having the four blood types should be related by $O=q^2$, $A=p^2+2pq$; $B=r^2+2qr$, $AB=2pr$ with $p+q+r=1$. Given the sample $O=176$, $A=182$; $B=60$, $AB=17$, find if it is consistent with the model. [**Ans.** Consistent]

19. Show that the expected value of χ^2 for a $s\times t$ table is $\frac{n}{n-1}(s-1)(t-1)$.

[**Hint.** Let p_{ij} denote the probability that a randomly chosen individual belongs to the ith row and jth column of the table where n_{ij} is the number of observed individuals in the ith row and jth column. The hypothesis of independence is then equivalent to the hypothesis that there exist $s+t$ constraint $p_{i.}$ and $p_{.j}$ such that

$$p_{ij}=p_{i.}p_{.j}.$$

$$\sum_i p_{i.}=\sum_j p_{.j}=1$$

We have $$\chi^2=\sum_{i,\,j}\frac{(n_{ij}-np_{i.}p_{.j})^2}{np_{i.}p_{.j}}.$$

The estimates of $p_{i.}$ and $p_{.j}$ can be shown to be $\frac{n_{i.}}{n}$ and $\frac{n_{.j}}{n}$ respectively.

Thus $$\chi^2 = n \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{n_{i.} n_{.j}} = n \left(\sum_{i,j} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right)$$

Since we have here st groups and $s+t-2$ parameters determined from the sample, the limiting distribution of χ^2 has $st-(s+t-2)-1=(s-1)(t-1)$ *d.f.* of frequency. Assuming that the independence hypothesis is true, we have

$$E(\chi^2) = \frac{n}{n-1}(s-1)(t-1).]$$

20. x_{ij} $(i=1, 2, \ldots, m, j=1, 2, \ldots, n)$ are the cell frequencies in an $m \times u$ contingency table. The marginal totals of the rows and columns are denoted by r_i $(l=1, 2, \ldots. m)$ and c_j $(j=1, 2, \ldots, n)$ and the grand total by N. Show that the means, variances, and covariances of the cell-frcquencies are given by

$$E(x_{ij}) = \frac{r_i c_j}{N},$$

$$\text{Var}\,(x_{ij}) = r_i c_j (N-r_i)(N-c_j)/N^2(N-1)$$

$$\text{Cov}\,(x_i, x_j) = r_i c_{ij} r_k c_l / N(N-1), \; i=k, j=l$$

$$= r_i(N-r_i) c_j c_l / N^2 (N-1), \; i=k, j=l.$$

(*M.A., Bombay, 1953*)

21. (*a*) What is meant by a class frequency of (*i*) first order (*ii*) third order. How would you express a class frequency of first order in terms of class frequencies of third order. (*B.Sc., Agra, Univ., 1967*)

(*b*) What do you understand by consistency of given data ? How do you check it ? (*B.Sc., Allahabad, 1959*)

(*c*) What do you mean by independence of attributes ? Give a criteria of independence for attributes *A* and *B*. (*B.Sc., Nagpur. 1961*)

(*d*) When are two attributes said to be positively associated and negatively associated ? Also define complete association and disassociation of two attributes. (*B.Sc., Bombay 1968 ; B.Sc., Nagpur 1964*)

(*e*) What is association of attributes ? Write a note on the strength of association and how it is measured ? (*B.Sc. Madras 1965 S*)

(*f*) Define Yule's coefficient of association and the coefficient of colligation. Obtain the relation between the two. (*B.Sc., Shivaji, 1969*)

22. (*a*) Given that $(A)=(\alpha)=(B)=(\beta)=\frac{N}{2}$ show that $(A\,B)=(\alpha\,\beta)$, $(A\,\beta)=(\alpha\,B)$. (*B.Sc., Agra, 1964*)

(*b*) Given that $(A)=(\alpha)=(B)=(\beta)=(C)=(\gamma)=\frac{N}{2}$ and also that $(ABC)=(\alpha\beta\gamma)$ show that $2(ABC)=(AB)+(AB)+(BC)-\frac{N}{2}$ (*B.Sc., Hon., Delhi, 1970*)

23. Show that if

$$\frac{(A)}{N}=x, \; \frac{(B)}{N}=2x, \; \frac{(C)}{N}=3x \text{ and } \frac{(AB)}{N}=\frac{(BC)}{N}=\frac{(CA)}{N}=y$$

neither the values of x nor y can exceed $\frac{1}{4}$.

(*B.Sc. AQ, 1965 AQ, Madras, 1961 ; Punjab, 1961*)

24. (*b*) Given that $(A)=(B)=(C)=\frac{N}{2}$ and 80 per cent of A's are B's, 75 per cent of A's are C's, find the limits to the percentage of B's that are C's.

(*B.Sc., Agra, 1959 ; Jivajee, 1970*)

[**Ans.** 55% and 95%]

25. If $1000=N=1\frac{2}{3}(A)=2(B)=2\frac{1}{2}(C)=5(AB)$ and $(AC)=(BC)$ what should be the minimum value of (BC). (*B.Sc. Agra, 1968*)

[**Ans.** 150]

26. Among the adult population of a certain town 50 per cent are males, 60 per cent are wage-earners and 50 per cent are 45 years of age or over, 10 per cent of the males are not wage earners and 40 per cent of the males are under 45. Make the best possible inference about the limits within which the percentage of persons (male or female) of 45 years or over are wage earners.

(*B.Sc., Kanpur 1970*)

[**Hint.** $N=100$, $(A)=50$, $(B)=60$, $(C)=50$]

[**Ans.** $25 \leqslant (BC) \leqslant 45$]

27. In an university examination, 65% of the candidates passed in English, 90% passed in the second language and 60% passed in the optional subjects. Find how many at least should have passed the whole examination.

[**Ans.** 15%]

28. Show that if $\delta=(AB)-(AB)$, then with usual notations, prove that

(*i*) $[(A)-(\alpha)][(B)-(\beta)]+2N\delta=(AB)^2+(\alpha\beta)^2-(A\beta)^2-(\alpha\beta)^2$.

(*B.Sc., Poona, 1968*)

(*ii*) $\delta=\frac{(B)(\beta)}{N}\left\{\frac{(AB)}{(B)}-\frac{(A\beta)}{(\beta)}\right\}=\frac{(A)(\alpha)}{N}\left\{\frac{(AB)}{(A)}-\frac{(\alpha\beta)}{(\alpha)}\right\}$

(*B.Sc. Agra, 1970 ; Gujarat, 1968*)

17

*Statistical Theory of Point Estimation

17·1. Introduction. Given a random sample $x_1, x_2, \ldots\ldots, x_n$ from a population with density function of known form $f(x, \theta)$ but with unknown parameter θ, our problem is to find an estimate for θ in terms of the sample values. In general, we can construct a number of statistical functions of $x_1, x_2, \ldots\ldots, x_n$, which may claim to estimate θ. For example, in order to estimate the population mean we may use as estimators the sample mean or sample median or the geometric mean or the harmonic mean of the sample. Thus it is essential to discriminate between the various statistics so as to arrive at the *best* estimator, the word 'best' being used in Fisherian sense.

17·2. Fisher's Criteria for the Best Estimator. According to Fisher, an estimator is said to be the best if it is (*i*) unbiased (*ii*) consistent (*iii*) efficient and (*iv*) sufficient.

17·2·1. Consistent Estimator. An estimator t_n computed from a random sample of n values is said to be a consistent estimator for a population parameter θ if it converges in probability to θ as n tends to infinity, *i.e.*, if given two positive numbers ε and η, however small, we can find a positive integer N depending on ε and η such that

$$P[\, t_n - \theta \mid < \varepsilon] > 1 - \eta, \text{ for all } n > N.$$

i.e.,
$$P\{\lim_{n \to \infty} t_n = \theta\} = 1$$

Note. 1. The property of consistency ensures that the difference between t_n and θ would become smaller and smaller in probability sense as n increases indefinitely. In other words, it would give increasing accuracy with the increasing of the size of the sample.

*For a detailed discussion. see 'Statistical Inference' and for more solved problems, see 'Problems in Statistics' by H.C. Saxena, published by S. Chand&Co.

2. The property of consistency is a limiting property, *i.e.*, it refers to the behaviour of t_n as the sample size increases. Moreover, if there exists one consistent statistic t_n, we can construct an infinite number of consistent statistics since the statistics $\frac{an+b}{an+c}\, t_n$, $\frac{an^2+bn+c}{an^2+dn+e}\, t_n$, etc., converge in probability to θ.

3. A consistent estimator is unbiased in the limit but an unbiased estimator may or may not be a consistent estimator.

Example 17·1. *Show that, in random sampling from a normal population, the sample mean is a consistent estimator for the population mean.* (*M.A., Bombay, 1950*)

Let $x_1, x_2, \ldots\ldots, x_n$ be the sample drawn from the normal population with mean μ and finite variance σ^2. We have proved in §8·2, §12·6, §13·2 that the statistic

$$z = \frac{(\bar{x}-\mu)\sqrt{n}}{\sigma}$$

is a standard normal variate.

$$\therefore \quad P[\,|\,\bar{x}-\mu\,|\,<\varepsilon] = P[\,|\,z\,|\,<\varepsilon\sqrt{n}/\sigma] = \int_{-\frac{\varepsilon\sqrt{n}}{\sigma}}^{+\frac{\varepsilon\sqrt{n}}{\sigma}} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}z^2}\, dz$$

Hence given any number η, however small, we can choose n so that area under the standard normal curve between $+\varepsilon\sqrt{n}/\sigma$ becomes greater than $1-\eta$.

Example 17·2. *Show that for random sampling from Cauchy population with density function*

$$f(x, \mu) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\mu)^2}, \quad -\infty < x < \infty,$$

the sample mean is not a consistent estimator for the population mean but the sample median is a consistent estimator.

(*M.A., Delhi, 1957, 59 ; M.A., Bombay, 1950, 56*)

Let $\bar{x}$ be the mean of the sample $x_1, x_2, \ldots\ldots, x_n$ drawn from Cauchy population. We have proved in §14·3 and §14·4 that sample mean $\bar{x}$ is distributed as x and that its distribution is

$$dp = \frac{1}{\pi} \cdot \frac{1}{1+(\bar{x}-\mu)^2}\, d\bar{x}, \quad -\infty < \bar{x} < \infty$$

$$\therefore \quad P[\,|\,\bar{x}-\mu\,|\,<\varepsilon]=P[\,|\,x-\mu\,|\,<\varepsilon]=P[\,|\,z\,|\,<\varepsilon],\ z=x-\mu$$

$$=\frac{1}{\pi}\int_{-\varepsilon}^{+\varepsilon}\frac{1}{1+z^2}\,dz=\frac{2}{\pi}\tan^{-1}\varepsilon.$$

Since this does not involve n, $P[\,|\,\bar{x}-\mu\,|\,<\varepsilon]$ cannot always be greater than $1-\eta$, however large n may be. Hence sample mean is not a consistent estimator. In fact for estimating μ, a sample of 1000 observations has as much accuracy as a sample of 1 (one) observation only and thus by using the sample mean for estimating μ we would be wasting 999 observations.

In Ex. 14·4, we have shown that for large samples the statistic $z=(\widehat{x}-\mu)\sqrt{4n/\pi}$ is a standard normal variate (where x is read as x cap or x caret) is the sample median and population mean is same as population median Hence, for large samples,

$$P[\,|\,\widehat{x}-\mu\,|\,<\varepsilon]=P\left(\,|\,z\,|\,<\frac{2\varepsilon\sqrt{n}}{\pi}\right)=\int_{-\frac{2\varepsilon\sqrt{n}}{\pi}}^{+\frac{2\varepsilon\sqrt{n}}{\pi}}\frac{1}{\sqrt{2\pi}}\;e^{-\frac{1}{2}z^2}dz$$

By making n large, this can be made as near to unity as we please. Hence sample median is a consistent estimator for the population mean.

Ex. 1. Show that the sample variance is a consistent estimator for the population variance of a normal distribution.

[**Hint.** We have earlier proved that for a large sample, $E(s^2)$ is asympto tically equal to σ^2 and Var $(s)^2$ is equal to $2\sigma^4/n$.]

Ex. 2. Show that a statistic t_n whose mean value differs from θ by order $1/n^2$, whose variance is of order $1/n$ and which tends to normality as n increases, is a consistent estimator for θ.

Ex. 3 Show that the sample median is a consistent estimator for the population mean of a normal distribution.

[**Hint.** For a large sample, $\dfrac{\widehat{(x-}\ \sqrt{2n}}{\sigma\sqrt{\pi}}$ is a standard normal variate]

Ex. 4. Show that, for a random sampling from a normal population, in estimating for the mean, the estimator x_1 (=first observation of the sample) is unbiased but not consistent.

Ex. 5. If T is an unbiased estimator of θ, show that T^2 is a biased estimator of θ^2 but (*ii*) if T is a consistent estimator of θ then T^2 is also a consistent estimator of θ^2. (*M.A. Madras, 1961 ; M.A. Bombay, 1956*)

[**Hint.** (*i*) $0\neq\text{Var}\ T=E(T-\theta)^2=E(T^2)-2\theta E(T)+\theta^2=E(T^2)-\theta^2$. Hence $E(T^2)\neq\theta^2$

Ex. 6. (*a*) A statistic t_n for which $E(t_n)=\theta_n\to 0$ and Var $(t_n)\to 0$ as $n\to\infty$ is consistent. (*M.A., Bombay, 1956*)

(*b*) t_n is an unbiased estimate of θ_n with variance δ_n and $\theta_n \to \theta$, $\delta_n \to 0$ as $n \to \infty$. Show that t_n is a consistent estimator of θ. (*M.A. Statistics, Delhi, 1960*)

[**Hint.** (*b*) See Ex. 2, after Tchycheff Lemma].

Ex. 7. Show that the mean $\bar{x}$ of a random sample of size n from the distribution $f(x, \theta) = \theta^{-1} e^{-x/\theta}, 0 < x < \infty, 0 < \theta < \infty$, zero elsewhere, is an unbiased statistic for θ and has variance θ^2/n.

Ex. 8. (*a*) Let $x_1, x_2, \ldots\ldots, x_n$ be random sample from a normal distribution with mean zero and variance σ^2. Show that $\sum_1^n x_i^2/n$ is an unbiased statistic for σ^2 and has variance $2\sigma^4/n$. (*M.A., Statistics, Delhi, 1960*)

(*b*) $x_1, x_2, \ldots\ldots, x_n$ are random observations on a Bernoulli variable x taking the value with probability θ and the value 0 with probability $1-\theta$. Show that $T(T-1)/n(n-1)$ is an unbiased estimator of θ where $T = \sum_{i=1}^n x_i$. (*B.Sc., Gujarat, 1969*)

[**Hint.** $E(\chi_i) = 1-\theta+0-(1-\theta) = \theta$, Var $x_i = \theta(1-\theta)$, $E(T) = E(\sum_i^n x_i) = n\theta$,

Var $(T) = \text{Var}\,(\sum_i^n x_i) = n\theta(1-\theta) \Rightarrow ET^2 = x\theta(1-\theta+n\theta)$

Now $E[T(T-1)] = E(T^2-T) = E(T^2) - E(T) = $ etc.

Ex. 9. Let $y_1 < y_2 < y_3$ be the order statistics of a random sample of size 3 from the uniform distribution having density function

$$f(x, \theta) = \frac{1}{\theta}, \quad 0 < x < \theta, \; 0 < \theta < \infty,$$

$$= 0, \text{ elsewhere}$$

Show that $4y_1$, $2y_2$ and $(4/3)\,y_3$ are all unbiased statistics for θ. Find the variance of each of these unbiased statistics. [Ans. $3\theta^2/5$, $\theta^2/5$, $\theta^2/15$]

17·2·2. Efficient Estimator. Let for large samples, two consistent estimators t_n and t'_n be both distributed asymptotically normally about the true value of the parameter θ and with variance σ^2/n and σ'^2/n respectively. This will usually be the case in virtue of the central limit theorem. Then t_n is said to be more efficient estimator than t_n' if $\sigma^2 < \sigma'^2$. However, if var $(t_n) <$ var (t_n') for all n, then t_n will be said to be more efficient than t'_n for all n. The estimator with the smaller variance will be grouped more closely round the true value and on the average, will deviate less from the true value than the estimator with large variance and thus it may be reasonably regarded as more efficient than the other.

If we can find a consistent estimator t_n whose variance is less than that of all other consistent estimators for all n, then t_n will be said to be the most efficient and the *efficiency* E, or any other estimator is defined as the ratio of the variance of the most efficient

estimator to the variance of the given estimator. The efficiency of a statistic represents, in large samples, the fraction of the relevant information available in the sample which is utilized by the statistic in question.

Example 17·3. *Show that, in estimating for the mean for random sampling from a normal population sample mean is more efficient than the sample median. Discuss also the populations of the two estimators for Cauchy population.*

For a normal population with finite variance σ^2, the sample mean and sample variance are consistent estimators for the population mean and variance respectively.

For any n, Var (mean) $= \dfrac{\sigma^2}{n}$

and for large n, Var (median) $= \dfrac{\pi\sigma^2}{2n}$.

Since Var (mean) is smaller than Var (median), the mean is more efficient than the median for large n. For small n, by actual calculations, it can be proved that Var (mean) is smaller that Var (median). Also the efficiency of the median relative to the mean (which will be proved later to be most efficient) is $2/\pi$, *i.e.*, 0·637.

For Cauchy population, a direct comparison is not possible since the mean is not even a consistent estimator. However,

$$\text{Var (median)} = \pi^2/4n$$

and obviously in this case, median is a better estimator than mean.

Example 17·4. *If two most efficient estimators t_n and t_n', estimating the same parameter θ, tend in the limit to be distributed in the bivariate normal form*

$$dp = \frac{1}{2\pi v_n \sqrt{1-\rho^2}} \exp\left[-\frac{1}{2v_n(1-\rho^2)}\left\{(t_n-\theta)^2 - 2\rho(t_n-\theta)(t_n'-\theta)+(t_n'-\theta)^2\right\}\right] dt_n dt_n',$$

then show that $\rho = 1$, *where* Var $(t_n) = v_n = \text{Var}\ (t_n')$.

Let $u_n = \frac{1}{2}(t_n + t_n')$ ...(1), $u_n' = \frac{1}{2}(t_n - t_n')$ ...(2)

then clearly u_n is a consistent estimator for θ and $E(u_n') = 0$.

The joint distribution of u_n and u_n' is

$$dp = \text{const.}\ \exp\left[-\frac{1}{2v_n(1-\rho^2)}\{2(1-\rho)(u_n-\theta)^2 + 2(1+\rho)u_n'^2\}\right] du_n du'$$

Hence u_n and u_n' are independent variates and

$$\text{Var}\ (u_n) = v_n(1+\rho)/2$$

Since t_n and t_n' are most efficient, it follows from (1) that

$$\text{Var}\,(u_n) \geqslant \text{Var}\,t_n$$

or $$v_n(1+\rho)/2 \geqslant v_n$$

or $$1+\rho \geqslant 2 \text{ or } \rho \geqslant 1.$$

But ρ cannot be greater than 1, it follows that $\rho=1$ then

$$\text{Var}\,u_n = v_n$$

Thus in the limit, all most efficient estimators tend to equivalence.

Ex. 1. Given that in normal samples

$$\text{Var (mean deviation)} = \frac{2}{\pi}\cdot\frac{n-1}{n^2}\sigma^2\left(\frac{\pi}{2}+\sqrt{n(n-2)}-n+\sin^{-1}\frac{1}{n+1}\right)$$

$$\sim\frac{\sigma^2}{n}\left(-\frac{2}{n}\right) \text{ and Var (standard deviation)} = \frac{\sigma^2}{n},$$

compare their relative efficiencies.

[$E=\frac{1}{\pi-2}=0{\cdot}876$ and use of the mean deviation with 100 observations is equivalent to using the *s.d.* with 876 observations yet the mean deviation may be preferred on account of its ease in calculation].

Ex. 2. Show that for samples of size n from normal population with mean μ and variance σ^2, the statistic

$$\hat{\mu} = \frac{1}{n+1}\sum_{i=1}^{n} x_i$$

is most efficient for estimation μ, though it is biased. (*B.Sc. Bombay 1967, 71*)

$$\sim\left[E\left(\frac{1}{n+1}\Sigma x_i\right)=\frac{n\mu}{n+1}\right], \text{Var}\,\bar{x} = \sigma^2/n, \text{ Var}\,\hat{\mu} = \frac{n}{(n+1)^2}\sigma^2.$$

Ex. 3. If t is a most efficient estimator and t' a less efficient estimator with efficiency E and if the correlation coefficient between t and t' is ρ, show by considering estimator t'' defined by

$$(1+E-2\rho\sqrt{E})\,t'' = (1-\rho\sqrt{E})t+(E-\rho\sqrt{E})t'$$

that $$\rho=\sqrt{E}.$$

[**Hint.** If t and t' approach θ in probability sense so does t''. Moreover,

$$(1+E-2\rho\sqrt{E})^2\,\text{Var}\,t'' = (1-\rho\sqrt{E})^2\,\text{Var}\,(t)+(E-\rho\sqrt{E})^2\,\text{Var}\,t' + 2(1-\rho\sqrt{E})(E-\rho\sqrt{E})\,\text{Cov}\,(t, t')$$

Let $\text{Var}\,t=v$, then $\text{Var}\,t'=\frac{v}{E}$, $\text{Cov}\,(t, t')=\rho\sqrt{v}\sqrt{\frac{v}{E}}$

$$\therefore \text{Var}\,t'' = v[(1-\rho\sqrt{E})^2+(\sqrt{E}-\rho)^2+2(1-\rho\sqrt{E})(\sqrt{E}-\rho)\rho]/(1+E-2\rho\sqrt{E})^2$$

$$\frac{v(1-\rho^2)}{(1-\rho^2)+(\rho-\sqrt{E})^2} \leqslant v \qquad \ldots(1)$$

But $\text{Var}\,t'' \geqslant 0$ $\therefore$ $\text{Var}\,(t'')=v$, $\therefore$ from (1), $\rho=\sqrt{E}$.]

Ex. 4. Deduce from Exercise 3, the result of example 17·4.

Ex. 5. Show that in large sample the correlation between mean and median is ·7979.

Ex. 6. Prove that if $t''=\frac{1}{2}(t+t')$, where t and t' are both most efficient with variance v, then

$$\text{Var}\,t'' = v\frac{(1+\rho)}{2}$$

Hence deduce $\rho=1$.

Ex. 7. Prove the result of Ex. 3, by using the method of Example 17·3.

Ex. 8. Is the sample mean necessarily efficient estimator of the population mean for every population ?

Ex. 9. Deduce from Exercise 3 that Cov $(t, t-t')=0$.

[**Hint.** Cov $(t, t-t')=$Cov $(t, t')-$Var t.]

Ex. 10. Find the maximum likelihood estimator of θ based on a sample of size n from a rectangular population in the range θ to 2θ. Show that if ξ, η are the maximum and minimum of the observations the estimate $(2\xi+\eta)/5$ is more efficient than maximum likelihood estimate. *(M.A., Bombay, 1955)*

17·2·3. Sufficient Estimator. An estimator t_n is said to be sufficient for estimating a population parameter θ, if it contains all the information in the samples about the parameter. The criterion can be made more precise in terms of the concept of *likelihood function*. If $f(x, \theta)$ is the density function for a population, then the likelihood function $L(x_1, x_2,\ldots\ldots,x_n, \theta)$ for a random sample $(x_1, x_2,\ldots\ldots, x_n)$ is defined by

$$L(x_1, x_2,\ldots\ldots,x_n\theta)=\prod_{i=1}^{n} f(x_i, \theta)=f(x_1, \theta)\ f(x_2, \theta),\ldots\ldots f(x_n, \theta).$$

Now if it is possible to write

$$L(x_1, x_2, \ldots\ldots x_n, \theta)=L_1(t_n, \theta)\ L_2(x_1, x_2,\ldots\ldots,x_n)$$

where the second function on the right hand side does not contain θ, then t_n will be said to be *sufficient* for θ. In this form it is obvious that any other function t_n' cannot provide any additional information about θ ; for given t_n, the conditional distribution of t_n' cannot involve θ and, therefore, if t_n is known, t_n' cannot give us more information about θ. Thus the statistic t_n is said to be sufficient for the parameter θ if with respect to any other statistic t_n', the joint probability density $P(t_n, t_n')$ of t_n, t_n' is of the form

$$P(t_n, t_n')=P_1(t_n, \theta)\ P_2(t_n' \mid t_n)$$

where $P_2\ (t_n' \mid t_n)$ is the relative probability density of t_n' given t_n and is independent of θ.

If sufficient estimator exists, we need not use any other, for it is most desirable, but unfortunately sufficient estimators are the exceptions rather than the rule and, in practice, we have to be contented with less satisfactory estimators.

Example 17·5. *Show that, if, σ^2 is known in random sampling from normal population, $\bar{x}$ is a sufficient estimator for μ ; but if μ is*

known, $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is not a sufficient estimator for σ^2. Suggest a sufficient estimator for σ^2. (*M.A., Bombay, 1950*)

For the normal population

$$f(x,\mu)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

$$\therefore\ L(x_1, x_2, \ldots\ldots, x_n, \mu)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\tfrac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right]$$

$$=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{n}{2\sigma^2}\{(\bar{x}-\mu)^2+s^2\}\right]$$

$$=\exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right]\times\left\{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{ns^2}{2\sigma^2}\right]\right\}$$

Thus L has been broken up into two factors, the first factor contains only $\bar{x}$ (estimator) and μ (parameter) and the second factor does not contain μ. Therefore, knowing σ^2, $\bar{x}$ is sufficient for estimating μ.

On the other hand if μ is known, then while the second factor involves only s^2 and σ^2, the first factor *also* contains σ^2 and therefore, s^2 is not sufficient for σ^2.

However, if we define

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

then from (1), $L=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{nS^2}{2\sigma^2}}$

and S^2 is obviously sufficient for σ^2. In fact nS^2/σ^2 is distributed as χ^2 with n *d. f.* This example, of course, only illustrates that if μ is known there is no point in using s^2 based on deviations from the sample mean, when we can get a better estimator S^2 by taking deviations from the known population mean.

Example 17·6. *Show that a sufficient estimator is most efficient if a most efficient estimator exists.* (*M.A., Delhi, 1951*)

We assume that sufficient estimator t_n and any other estimator t_n' tend to joint normality for large n in the form

$$dP=\frac{1}{2\pi\sqrt{v_n v_n'}\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(t_n-\theta)^2}{v_n}-\frac{2\rho(t_n-\theta)(t_n'-\theta)}{\sqrt{v_n v_n'}}+\frac{(t_n'-\theta)^2}{v_n}\right\}\right]dt_n dt_n'.$$

Since t_n is sufficient, the distribution of t_n', given t_n cannot contain θ. The distribution of t_n is

$$dP_1 = \frac{1}{\sqrt{2\pi v_n}} \exp\left[-\tfrac{1}{2}\frac{(t_n-\theta)^2}{v_n}\right]dt_n$$

and, therefore, the conditional distribution of t_n', given t_n, is

$$dP_2 = \frac{1}{\sqrt{2\pi v_n}\sqrt{1-\rho^2}} \exp\left[\frac{-1}{2(1-\rho^2)}\left\{\frac{\rho(t_n-\theta)}{\sqrt{v_n}} - \frac{(t_n'-\theta)}{\sqrt{v_n}}\right\}^2\right]dt_n'.$$

Since this is not to depend on θ, we have

$$\frac{\rho}{\sqrt{v_n}} = \frac{1}{\sqrt{v_n'}} \quad \text{or} \quad \rho = \sqrt{\frac{v_n}{v_n'}}.$$

But $\rho \leqslant 1$ $\quad\therefore\quad$ $v_n \leqslant v_n'$

$\therefore$ Variance of the sufficient estimator is less than or equal to that of any other estimator—at least of all those estimators which tend to be distributed normally with the sufficient estimator. Accordingly in this class of estimators the sufficient estimator is most efficient.

Ex. 1. Show that if t_n is a sufficient estimator for θ, then $\partial/\partial\theta\,(\log L)$ is a function of t_n and θ only.

Ex. 2. Show that in estimating for the parameter μ in the Poisson distribution with general term $\frac{e^{-\mu}}{x!}\mu^x$, $\bar{x}$ is a sufficient estimator for μ.

Ex. 3. Show that for type III population

$$dp = -\frac{e^{-\frac{x}{\theta}}x^{p-1}}{\theta^p\lceil p}\,dx,\ 0 \leqslant x \leqslant \infty,\ p<0.$$

where p is known, $\bar{x}/p$ is a sufficient estimator for θ. *(M.A. Delhi, 1958)*

Ex. 4. Let x_1 and x_2 denote a random sample from the normal distribution with mean θ, $-\infty < \theta < \infty$, and s.d. unity. Show that $y_1 = x_1 + x_2$ is a sufficient statistic for θ. *(M.A., Delhi, 1966)*

[**Hint.** Put $y_2 = x_2$ $\quad\therefore\quad$ $x_1 = y_1 - y_2$, $x_2 = y_2$, $dx_1dx_2 = dy_1dy_2$]

Ex. 5. Examine whether it is possible to obtain a sufficient estimator for the parameter θ of the distribution

$$f(x,\theta) = \theta(\sin x)^{\theta-1}\cos x,\ 0 \leqslant x \leqslant 1. \qquad \textit{(M.A., Delhi, 1965)}$$

[Ans. $t_1 = \sin x_1 \sin x_2 - \sin x_n$ is a sufficient statistic for θ.]

Ex. 6. Show that if m is known, the estimator

$$S = \left[\frac{1}{n}\Sigma(x-m)^2\right]^{\frac{1}{2}}$$

is sufficient for σ in samples of n from a normal population with mean m and s.d. σ. *(M.A. Statistics, Delhi, 1959)*

Ex. 7. If t_1 and t_2 are two unbiased estimates of a parameter with variances σ_1^2, σ_2^2 and correlation ρ, what is the best unbiased linear combination of t_1 and t_2 and what is the variance of such a compound ?

[**Hint.** Let $t=l_1 t_1+l_2 t_2$ where l_1 and l_2 are constants.

Then $E(t)=l_1\theta+l_2\theta=\theta$ (given)

$\therefore$ $l_1+l_2=1$ ...(1)

Also $V(t)=V(l_1t_1+l_2t_2)=l_1^2\sigma_1^2+2\rho l_1l_2\sigma_1\sigma_2+l_2^2\sigma_2^2$ .. (2)

We are to minimize (2) subject to the condition (1)

$\therefore$ we have $1+\frac{dl_2}{dl_1}=0$

and $2l_1\sigma_1^2+2\rho\sigma_1\sigma_2\left(l_2+l_1\frac{dl_2}{dl_1}\right)+2\sigma_2^2 l_2\frac{dl_2}{dl_1}=0$

or $l_1\sigma_1^2+\rho\sigma_1\sigma_2 l_2-\rho\sigma_1\sigma_2 l_1-\sigma_1^2 l_2=0$

or $l_1(\sigma_1^2-\rho\sigma_1\sigma_2)=l_2(\sigma_2^2-\rho\sigma_1\sigma_2)$

which gives the ratio between l_1 and l_2.]

Ex. 8. Let t_1 be an unbiased estimator with minimum variance σ^2 and t_2 by any other unbiased estimator with variance σ^2/e. What is the correlation between t_1 and t_2 ?

[**Hint.** From Ex. 6 above we have the best linear compounds of t_1 and t_2 with coefficient l_1 and l_2 satisfying the conditions

$$l_1(e-\rho\sqrt{e})=l_2(1-\rho\sqrt{e}) \quad ...(3)$$

Now use the condition that the variance of the linear compound is not less than σ_2 or since t_1 cannot be improved upon, $l_2=0$ $\therefore$ from (3), $e=\rho\sqrt{e}$ or $\rho=\sqrt{e}$.]

Ex. 9. If t_1 and t_2 are two unbiased statistics having the same variance then the correlation is $\geqslant 2e-1$ where e is the ratio of the variance of the best estimate to the common variance of t_1 and t_2.

[**Hint** Let $t=\frac{1}{2}(t_1+t_2)$ and use that $V(t)\geqslant$ the least variance.]

***Ex. 10.** If e_1, e_2 are the efficients of two unbiased estimators A and B and ρ is coefficient of correlation between A and B, then

$$\sqrt{e_1e_2}-\sqrt{(1-e_1)(1-e_2)}<\rho<\sqrt{e_1e_2}+\sqrt{(1-e_1)(1-e_2)}.$$

(M.A., Bombay, 1958)

Solution. Let Var $A=V/e_1$, Var $B=V/e_2$

where V is the minimum variance.

Let $C=\lambda A+\mu B$

where λ and μ are const. such that $\lambda+\mu=1$

As A and B are unbiased so is C.

$\therefore$ Var $C=\left(\frac{\lambda^2}{e_1}+\frac{\mu^2}{e_2}+\frac{2\lambda\mu\rho}{\sqrt{e_1e_2}}\right)V$

But Var $C\geqslant V$

$\therefore$ $\frac{\lambda^2}{e_1}+\frac{\mu^2}{e_2}+\frac{2\lambda\mu\rho}{\sqrt{e_1e_2}}\geqslant 1=\lambda^2+\mu^2+2\lambda\mu,$

---

*This proof is due to Prof. V.D. Thawani.

or $$\lambda^2\left(\frac{1}{e_1}-1\right)+2\lambda\rho\left(\frac{\rho}{\sqrt{e_1e_2}}-1\right)+\mu^2\left(\frac{1}{e_2}-1\right)\geqslant 0$$

$$\therefore \left(\frac{\rho}{\sqrt{e_1e_2}}-1\right)^2-\left(\frac{1}{e_1}-1\right)\left(\frac{1}{e_2}-1\right)\geqslant 0$$

[discriminant of a quadratic expression]

[The conditions $\frac{1}{e_1}-1>0$ and $\frac{1}{e_2}-1>0$ are satisfied.]

$$\therefore \frac{\rho^2}{e_1e_2}-\frac{2\rho}{\sqrt{e_1e_2}}+\left(\frac{1}{e_1}+\frac{1}{e_2}-\frac{1}{e_1e_2}\right)\leqslant 0$$

or $$\rho^2-2\sqrt{e_1e_2}\rho+(e_1+e_2-1)\leqslant 0.$$

$\therefore$ ρ lies between the roots of the quadratic equation

$$x^2-2\sqrt{e_1e_2}x+(e_1+e_2-1)=0.$$

The roots are $\sqrt{e_1e_2}\pm\sqrt{e_1e_2-(e_1+e_2-1)}$

$$\therefore \sqrt{e_1e_2}-\sqrt{(1-e_1)(1-e_2)}<\rho<\sqrt{e_1e_2}+\sqrt{(1-e_1)(1-e_2)}.$$

Cor. 1. If $e_1=1$, $\rho=\sqrt{e_2}$

Cor. 2. If $e_1=e_2=e$, $e-(1-e)<\rho<e+1-e$

or $$2e-1<\rho<1$$

i.e., $\rho>(2e-1)$ which is due to Prof. C.R. Rao.

17·3. Method of Maximum Likelihood. We have discussed above the three criteria of consistency, efficiency and sufficiency which according to Fisher, optimum or best estimators should satisfy, but these criteria, by themselves do not give any method for getting these estimators. A number of methods for estimation have accordingly been proposed, the most important of which is the method of maximum likelihood due to Fisher.

According to the principle of maximum likelihood, we should choose the estimator which makes the likelihood function a maximum, *i.e.*, t_n will be the maximum likelihood estimator if

$$L(x_1, x_2, \ldots\ldots, x_n, t_n) > L(x_1, x_2, \ldots\ldots, x_n, t_n')$$

for any other estimator t_n'. If L is a differentiable function of θ, then $t_n(x_1, x_2, \ldots\ldots, x_n)$ is the solution (if any) of

$$\frac{\partial L}{\partial \theta}=0 \text{ with } \frac{\partial^2 L}{\partial \theta^2}<0.$$

Since L is positive, the first equation is equivalent to

$$\frac{\partial \log L}{\partial \theta}=0,$$

a form which is more convenient in practice.

The principle of maximum likelihood can also be applied to the case when the population is bivariate or multivariate and also for the simultaneous estimation of a number of parameters.

Note. It is a convention to denote the maximum likelihood estimator of parameter by placing $\widehat{\ }$ (caret) over the parmeter. Thus the maximum likelihood estimator of θ is denoted by $\widehat{\theta}$.

Example 17·7. *For random sampling from a normal population, find the maximum likelihood estimators for*

(*a*) *the population mean, when the population variance is known,*

(*b*) *the population variance, when the population mean is known,*

(*c*) *the simultaneous estimation of both the population mean and variance.*

Here the likelihood function is

$$L=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma_2}\sum_{i=1}^{n}(x-\mu)^2}$$

$$=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{n}{2\sigma^2}[(\bar{x}-\mu)^2+s^2]}$$

$$\log L=\text{Const.}-\frac{n}{2}\log\sigma^2-\frac{n}{2\sigma^2}[(\bar{x}-\mu)^2+s^2]$$

(*a*) If σ^2 is known, then in order to estimate for μ, we have

$$\frac{\partial \log L}{\partial\mu}=0 \text{ which gives } \bar{x}-\mu=0$$

$$\widehat{\mu}=\bar{x}$$

(*b*) If μ is known, then to estimate for σ^2, we have

$$0=\frac{\partial}{\partial\sigma^2}\log L=-\frac{n}{2}\cdot\frac{1}{\sigma^2}+\frac{1}{2}\frac{n}{\sigma^4}[(\bar{x}-\mu)^2+s^2]$$

$$\widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

(*c*) For simultaneous estimation of μ and σ^2,

$$\left.\begin{aligned}0&=\frac{\partial}{\partial\mu}\log L=\frac{n}{\sigma^2}(\bar{x}-\mu), \text{ and}\\ 0&=\frac{\partial}{\partial\sigma^2}\log L=-\frac{n}{2}\cdot\frac{1}{\sigma^2}+\frac{1}{2}\frac{n}{\sigma^4}[(\bar{x}-\mu)^2+s]^2\end{aligned}\right\}$$

$$\therefore\quad \widehat{\mu}=\bar{x}$$

and

$$\widehat{\sigma^2}=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Example 17·8. *A random sample $x_1, x_2, \ldots\ldots x_n$, is drawn from the exponential population with density function*

$f(x\ ;\ \alpha,\ \beta)=y_0 e^{-\beta(x-\alpha)}$, $\alpha \leqslant x < \infty$, $\beta > 0$,

y_0 being a constant, obtain the maximum likelihood estimators for α and β.

(I.C.A.R., 1946 ; M. A. Calcutta, 1946 ; M.A., Statistics, Delhi, 1959)

First we evaluate y_0 in terms of α and β.

We have $$y_0 \int_{\alpha}^{\infty} e^{-\beta(x-\alpha)}\, dx = 1.$$

This gives $$y_0 = \beta.$$

$$L = \beta^n \exp\left[-\beta \sum_{i=1}^{n} (x_i - \alpha)\right] \qquad \ldots(17{\cdot}1)$$

or $$\log L = n \log \beta - n\beta\,(\bar{x} - \alpha)$$

$$0 = \frac{\partial}{\partial \alpha} \log L = n\beta \qquad \ldots(17{\cdot}1\ a)$$

$$0 = \frac{\partial}{\partial \beta} \log L = \frac{n}{\beta} - n(\bar{x} - \alpha) \qquad \ldots(17{\cdot}1\ b)$$

The first equation gives $\beta = 0$ which is clearly inadmissible, and with this the second equation would fail to determine a finite α. In this case the method of differentation clearly fails. Our object, however, is to choose α and β so that L is maximum. From (17·1) it is obvious that whatever be $\beta(>0)$, L would be maximum when α takes its largest possible value but since from definition, the population range extends from α to ∞, it is obvious that α must be less than or equal to every member of the sample. Thus the greatest value which α can take, consistent with the sample, is the least value of the sample, *i.e.*, x_1 when the sample values are arranged in ascending order.

Thus $$\hat{\alpha} = x_1$$

and then from (17·16) $$\hat{\beta} = 1/(\bar{x} - x_1).$$

Ex. 1. Show that for the rectangular population with $f(x,\ \theta) = \theta^{-1}$, $(0 \leqslant x \leqslant \theta)$ the maximum likelihood estimator for θ is the largest number of the sample. *(M A., Delhi, 1951, 59 ; M.A., Calcutta, 1946)*

Ex. 2. Show that for the rectangular population with

$f(x) = \frac{1}{\beta - \alpha}$ $(0 < \alpha \leqslant x \leqslant \beta)$ the maximum likelihood estimators x_a and x_b for α and β are the smallest and largest members of the population respectively and that they form a sufficient set of statistics for α and β and their joint distribution is $n(n-1)(\beta - \alpha)^{-n}(x_a - x_b)^{n-2} dx_a dx_b$.

Ex. 3. Show that for binomial population with density function

$$f(x, p)=p^x q^{1-x},\ x=0, 1\ ;\ q=1-p$$

the maximum likelihood estimator for p is the sample mean. Show also that this is a sufficient estimator.

Ex. 4. Show that for the binomial population with density function $f(x, p)={}^nC_x p^x q^{n-x}$, $x=0, 1, 2, \ldots\ldots, n$ the maximum likelihood estimator for p is x/n and Var $(\hat{p})=pq/n$. *(I.C.A.R., Delhi, 1946)*

Ex. 5. Obtain the maximum likelihood estimator of the parameter θ of the distribution

$$dF=\frac{x^{p-1}e^{-x/\theta}}{\theta^p \overline{|p}}dx,\ 0\leqslant x<\infty$$

where p is known. *(M.A , Delhi, 1960)*

Ex. 6. Find the maximum likelihood estimator for α for $f(x)=\theta\alpha_x$, $0<x<1$. [Ans. $\hat{\theta}=-n/\log(x_1x_2\ldots x_n)$]

Ex. 7. What is the maximum likelihood estimator for α in the density $f(x)=(\alpha+1)x^\alpha$, $0<x<1$? [Ans. $\hat{\alpha}=-1-[n/\log(x_1x_2\ldots x_n)]$]

Ex. 8. Find the estimate for α in the density

$$f(x, \alpha)=\frac{2}{\alpha^2}(\alpha-x),\ 0<x<\alpha$$

for samples of size 2

Ex. 9. Same as **Ex. 10**, page 450 Let y be a random variable which depends on the parameter x where x is not a random variable. We assume that the random variable y has normal distribution with mean $ax+b$ and *s.d.* σ for every value x, where a, b and $\sigma>0$ are unknown constants. Let n pairs of observations (x_k, y_k) be given where $y_1, y_2, \ldots\ldots, y_n$ are independent. Determine the constants a, b and σ^2.

[**Hint.** $f(y, a, b, \sigma)=\dfrac{1}{\sigma\sqrt{2\pi}}\ \exp\left[\dfrac{(y-ax-b)^2}{2\sigma^2}\right]$

$$\therefore\ L=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k-ax_k-b)^2\right]$$

Equate the partial derivatives to zero.]

[**Ans.** $a=\dfrac{rs_2}{r_1}$, $b=\bar{y}-\dfrac{rs_2}{s_1}\bar{x}$, $\sigma^2=\dfrac{1}{n}\sum_{k=1}^{n}(y_k-ax_k-b)^2=s_2^2(1-r^2)$]

Ex. 10. Independent samples of sizes n_1 and n_2 are taken Rs. 134 from two normal populations with equal means μ and variances respectively equal to $\lambda\sigma^2$, σ^2. Find the *m.l.e.* of μ and show that its large sample variance is

$$\text{Var}\ (\hat{\mu})=\sigma^2/\left(\frac{n_1}{\lambda}+n_2\right)$$

Hence show that the unbiased estimator

$$t=(n_1\bar{x}_1+n_2\bar{x}_2)/(n_1+n_2)$$

has efficiency $\lambda(n_1+n_2)^2/[(n_1\lambda+n_2)(n_1+n_2\lambda)]$

which attains the value 1 if and only if $\lambda=1$.

Let x_{1i}, $i=1, 2, \ldots\ldots, n_1$ and x_{2j}, $j=1, 2, \ldots\ldots, n_2$ be two independent sample, from the given normal population respectively. Since the samples are independent,

$$L=\text{Likelihood of } (n_1+n_2) \text{ sample observations}$$

$$=\frac{1}{(2\pi\lambda\sigma^2)^{\frac{1}{2}n_1}} \exp\left[\frac{-1}{2\lambda\sigma^2}\sum_{i=1}^{n_1}(x_{1i}-\mu)^2\right]$$

$$\times\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}n_2}} \exp\left[\frac{-1}{2\sigma^2}\sum_{j=1}^{n_2}(x_{2j}-\mu^2)\right]$$

$$\log L=\text{const.}-\frac{1}{2\lambda\sigma^2}\sum_{i=1}^{n_1}(x_{1i}-\mu)^2-\frac{1}{2\sigma^2}\sum_{j=1}^{n_2}(x_{2j}-\mu)^2$$

$$0=\frac{\partial}{\partial\mu}\log L$$

gives $$\hat{\mu}=(n_1\bar{x}_1+n_2\bar{x}_2)\Big/\left(\frac{n_1}{\lambda}+n_2\right)$$

$$\text{Now Var } \hat{\mu}=\frac{1}{\left(\frac{n_1}{\lambda}+n_2\right)^2}\left[\frac{n_1^2}{\lambda^2}\text{ Var }\bar{x}_1+n_2^2\text{ Var }\bar{x}_2\right]=\text{etc.}$$

$$\text{Also}\quad E(t)=\frac{[n_1E(\bar{x}_1)+n_2E(\bar{x}_2)]}{n_1+n_2}$$

$$=\frac{n_1\mu+n_2\mu}{n_1+n_2}=\mu$$

Hence t is an unbiased estimate of μ

$$\text{Also Var } t=\frac{1}{(n_1+n_2)^2}[n_1^2\text{ Var }\bar{x}_1+n_2^2\text{ Var }\bar{x}_2]=\text{etc.}$$

$$\text{Also efficiency}=\frac{\text{Var }\hat{\mu}}{\text{Var } t}=\frac{\lambda(n_1+n_2)^2}{(n_1\lambda+n_2)(n_1+n_2\lambda)}$$

Now efficiency$=1$ gives $\lambda(n_1+n_2)^2=(n_1\lambda+n_2)(n_1+n_2\lambda)$. By comparing the coeff. of n_1n_2, we get

$$2\lambda=1+\lambda^2 \Rightarrow \lambda=1.$$

17·3·1. Independence tests in contingency tables. Let p_{ij} be the probability that an element selected at random from the universe belongs to the ith class with respect to the attribute A and jth closs with respect to the attribute B (Refer § 16·2·1) and let $p_{i\cdot}$ and $p_{\cdot j}$ be their respective boundary probabilities, then according to the hypothesis that the attributes A and B are independent in the universe whence the observed sample has been made.

$$p_{ij}=p_{i\cdot}.p_{\cdot j}$$

for every pair (i, j), where

$$\sum_{=1}^{s} p_{i\cdot}=\sum_{j=1}^{t} p_{\cdot j}=1. \qquad \ldots(17\cdot2)$$

The hypothesis H_0 formulated in this manner does not define the values of $s+t$ unknown parameters, *i.e.*, the values $p_{\cdot j}$ and $p_{i\cdot}$. Since we can determine two of these parameters from (17·2), there remains $s+t-2$ unknown parameters. To verify the hypothesis H_0 by applying a χ^2—test we would determine the unknown parameters from the sample by means of maximum likelihood principle. We have

$$L=\prod_{i=1}^{s}\prod_{j=1}^{t} p_{ij}^{o_{ij}}=\prod_{i=1}^{s}\prod_{j=1}^{t}(p_{i\cdot}\,p_{\cdot j})^{o_{ij}}=\prod_{i=1}^{s}\prod_{j=1}^{t} p_{i\cdot}^{o_{ij}}\,p_{\cdot j}^{o_{ij}}$$

$$=\prod_{i=1}^{s} p_{i\cdot}^{o_{i\cdot}}\prod_{j=1}^{t} p_{\cdot j}^{o_{\cdot j}}$$

$$=\left(1-\sum_{i=1}^{s-1} p_{i\cdot}\right)^{o_{s\cdot}}\left(1-\sum_{j=1}^{t-1} p_{\cdot j}\right)^{o_{\cdot t}}\prod_{i=1}^{s-1} p_{i\cdot}^{o_{i\cdot}}\prod_{j=1}^{t-1} p_{\cdot j}^{o_{\cdot j}}$$

since from (17·2), $p_{s\cdot}=1-\sum_{i=1}^{s-1} p_{i\cdot}$, $p_{\cdot t}=1-\sum_{j=1}^{t-1} p_{\cdot j}$.

and $\log L=O_{s\cdot}\log\left(1-\sum_{i=1}^{s-1} p_{i\cdot}\right)+O_{\cdot t}\log\left(1-\sum_{j=1}^{t-1} p_{\cdot j}\right)+\sum_{i=1}^{s-1} O_{i\cdot}\log p_{i\cdot}+\sum_{j=1}^{t-1} O_{\cdot j}\log p_{\cdot j}$

$$\frac{\partial}{\partial p_{i\cdot}}\log L=-\frac{O_{s\cdot}}{1-\sum_{i=1}^{s-1} p_{i\cdot}}+\frac{O_{i\cdot}}{p_{i\cdot}}=\frac{O_{i\cdot}}{p_{i\cdot}}-\frac{O_{s\cdot}}{p_{s\cdot}}=0,\ i=1,2,\ldots\ldots,s-1$$

$$\frac{\partial}{\partial p_{\cdot j}}\log L=-\frac{O_{\cdot t}}{1-\sum_{j=1}^{t-1} p_{\cdot j}}+\frac{O_{\cdot j}}{p_{\cdot j}}=\frac{O_{\cdot j}}{p_{\cdot j}}-\frac{O_{\cdot t}}{p_{\cdot t}}=0,\ j=1,2,\ldots\ldots,s-1$$

$$\therefore\quad p_{i\cdot}=\frac{O_{i\cdot}}{C}\ \text{where}\ C=\frac{O_{s\cdot}}{p_{s\cdot}},\ i=1,2,\ldots,s$$

and $$p_{\cdot j}=\frac{O_{\cdot j}}{D}\ \text{where}\ D=\frac{O_{\cdot t}}{p_{\cdot t}},\ j=1,2,\ldots,t$$

Hence $$\sum_{i=1}^{s} p_{i\cdot}=\frac{\sum_{i=1}^{s} O_{i\cdot}}{C}=\frac{N}{C}=1$$

and $$\sum_{j=1}^{t} p_{\cdot j}=\frac{\sum_{j=1}^{t} O_{\cdot j}}{D}=\frac{N}{D}=1$$

Thus $C=D=N$

$\therefore$ $p_{i.}=\frac{O_{i.}}{N}$

$p_{.j}=\frac{O_{.j}}{N}$

$\therefore$ $\chi^2=N\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{\left(O_{ij}-\frac{O_{i.}O_{.j}}{N}\right)^2}{O_{i.}O_{.j}}$

Since we have determined $s+t-2$ parameters from the sampling it follows that Pearson statistics χ^2 has approximately a χ^2 distribution with $st-(s+t-2)-1=(s-1)(t-1)$ $d.f.$ Therefore, we can apply a χ^2-test to vertify hypothesis H_o that the variables A and B are independent.

17·3·2. Properties of Maximum Likelihood Estimators. We state below, without proof, some of the important properties of maximum likelihood estimators, which give a strong *a posteriori* justification for the use of this method, whenever it can be applied.

(*a*) *Maximum likelihood estimators are consistent.* More precisely, if

(*i*) the density function $f(x, \theta)$ is continuous in x throughout its range and if

(*ii*) $f(x, \theta)$ is continuous and monotonic in θ in some θ-interval containing the true value θ_o and for all x, in some x-interval, then the maximum likelihood estimator, say $\widehat{\theta}$, is consistent.

(*b*) *The distribution of maximum likelihood estimator tends to normality for large samples.* More precisely, if

(*i*) $f(x, \theta)$ is continuous in x throughout its range, and if

(*ii*) in a θ-interval containing the true value θ_o, $\frac{\partial f}{\partial \theta}$ is continuous in θ for every x, x^2 $\frac{\partial f}{\partial \theta}$ approaches continuous function of θ as x tends to infinity, and $\frac{\partial f}{\partial \theta}$ does not vanish in some interval, then for large n, the maximum likelihood estimator θ will tend to be normally distributed with variance given by

$$\frac{1}{\text{var}\,\widehat{\theta}}=n\int_{-\infty}^{\infty}\frac{1}{f}\left(\frac{\partial f}{\partial \theta}\right)^2 dx=n\left(\int_{-\infty}^{\infty}\frac{\partial}{\partial \theta}\log f\right)^2$$

$$=nE\left(\frac{\partial}{\partial\theta}\log f\right)^2, \qquad \ldots(17{\cdot}2\,a)$$

where θ is to be put equal to θ_0 on the right.

If the range is independent of f, or if f and $\partial f/\partial\theta$ vanish at the extremity of the range which depends on θ, we have an alternative form for variance, *viz.*,

$$\frac{1}{\text{var }\widehat{\theta}}=-nE\left(\frac{\partial^2}{\partial\theta^2}\log f\right) \qquad \ldots(17{\cdot}2b)$$

(*c*) *Maximum likelihood estimators are most efficient, i.e.*, in the class of estimators which for large n tend to be normally distributed about population parameters as mean, the variance of the maximum likelihood estimators will be less than or equal to that of any other estimator, *i.e.*, if t is any other such estimator.

$$\text{var }(t)\geqslant\text{var }(\widehat{\theta}).$$

(*d*) ***Maximum likelihood estimators are sufficient, if sufficient estimators exist.*** In other words, if a sufficient estimator exists, it is a function of the maximum likelihood estimator.

(*e*) *Maximum likelihood estimators have the invariance property i.e.*, if $\widehat{\theta}$ is a maximum likelihood estimator for θ, then $f(\widehat{\theta})$ will be a maximum likelihood estimator for $f(\theta)$.

(*f*) *Maximum likelihood estimators are not necessarily unbiased.*

Example 17·9. *Find the efficiency of the median as an estimator for the population mean of the Cauchy distribution.*

We have seen (Ex. 4, page 356) that for large samples,

$$\text{var (median)}=\pi^2/4n.$$

To find the efficiency, we have to find the variance of the most efficient estimator, which would be maximum likelihood estimator, if it exists for Cauchy distribution

$$L=\left(\frac{1}{\pi}\right)^n\prod_{i=1}^{n}\frac{1}{1+(x_i-\mu)^2}$$

$$\therefore\qquad -\frac{\partial}{\partial\mu}\log L=\sum_{i=1}^{n}\frac{2(x_i-\mu)}{1+(x_i-\mu)^2}=0.$$

It is not easy to solve this equation for μ to get $\widehat{\mu}$ but the variance of $\widehat{\mu}$ can be easily found by using (17·2*b*).

$$\frac{1}{\text{var}\,(\widehat{\mu})} = -nE\left(\frac{\partial^2}{\partial\mu^2}\log f\right) = -\frac{n}{\pi}\int_{-\infty}^{\infty}\frac{2(x-\mu)^2-2}{[1+(x-)^2\mu]^3}\,dx = \frac{n}{2}.$$

Hence $$\text{var}\,(\widehat{\mu}) = 2/n$$

$\therefore$ Efficiency of the madian $= 8/\pi^2 \cong 0{\cdot}8$.

Ex. 1. Show that in estimating λ for a Poisson distribution

$$\text{var}\,(\widehat{\lambda}) = \lambda/n.$$

Ex. 2. Show that in estimating μ for normal population with known variance σ^2,

$$\text{var}\,(\widehat{x}) = \frac{\sigma^2}{n}.$$

Ex. 3. Show that if a most efficient estimator t of a parameter θ exists, the maximum likelihood equation will have a unique solution equal to t. Also prove that if a sufficient estimator exists, the maximum likelihood solution will be a function of the sufficient estimator. *(M.A. Delhi, 1959, 1973)*

Ex. 4. Deduce [17·2 (*b*), from 17·2 (*a*) page 499], using integration by parts.

17·4. Other Methods of Estimation. In order to estimate a population parameter, the theory of probability which is purely deductive has to be supplemented by some axioms to make inductive inference possible. That the estimator should make likelihood function a maximum is one such axiom. We study below five other axioms giving five other methods of estimation. *A priori* which axiom is more plausible can be a matter of opinion, but any method of estimation will be judged by the properties of the estimators it gives rise to.

17·4·1. Method of Moments. This method due to K. Pearson, is the oldest and is freely used in fitting distributions specially of the Pearson type. According to it to choose m parameters of a population, we equate in general the first m moments of the sample values to the first m moments of the population and solve from these m equations for the m estimators. Fisher, however, showed that these estimators are in general inefficient, *i.e.*, they do not have minimum variance in the class of estimators distributed normally for large n. For normal population, the method of moments gives sample mean and variance as estimators for population mean and variance and these estimators are efficient, but from this point of view the normal distribution is an exception and for all other Pearson distributions, moment-estimators are not efficient.

17·4·2. Method of Minimum Variance. The axiom for this method is that we choose those estimators which are unbiased and have minimum variance. Thus, if t_n is the estimator for θ, we should have

$$E(t_n)=\theta, \text{ Var } (t_n) \text{ should be minimum.}$$

In other words,

$$\int \cdots \iint t_n L(x_1, x_2 \ \ldots\ldots, x_n, \theta)\, dx_1\, dx_2 \ldots\, dx_n = \theta.$$

and $$\int \cdots \iint (t_n-\theta)^2 L(x_1, x_2, \ldots\ldots, x_n, \theta)\, dx_1\, dx_2 \ldots dx_n = \text{minimum.}$$

By using calculus of variations, it can be shown that if there exists an estimator t_n satisfying these conditions, we can express

$$\frac{\partial}{\partial \theta} \log L = \frac{t_n-\theta}{\lambda},$$

where λ may depend on θ but not on $x_1, x_2, \ldots\ldots, x_n$. In fact λ would be the variance of t_n.

Ex. 1. For the normal population, show that $\bar{x}$ is the minimum variance estimator and its variance is σ^2/n.

Ex. 2. For the Cauchy distribution, show that a minimum variance estimator does not exist.

Ex. 3. Show that for a minimum variance estimator t_n,

$$\frac{1}{\text{var } (t_n)} = -nE\left(\frac{\partial^2}{\partial \theta^2} \log f\right)$$

holds for all n.

Ex. 4. For the type III population,

$$e^{-\frac{x}{\theta}} x^{p-1} dx,\ 0 \leqslant x \leqslant \infty,\ p>1.$$

where p is known, show that the minimum variance estimator for θ is $\bar{x}/p$ and its variance is θ^2/np. *(M.A., Delhi, 1968 ; I.C.A.R. Delhi, 1950)*

Ex. 5. The random variate x follows the discrete frequency distribution qp^{x-1}, $x=1, 2, \ldots$, $q=1-p$. A sample of n values $x_1, x_2, \ldots, x_n$ is drawn from this distribution. Find the maximum likelihood estimate of p and calculate its variance. Discuss whether this estimate is unbiased when $n=1$.

(I.C.A.R. Delhi, 1947)

$\left[\text{Ans. } \hat{p}=1-\frac{1}{\bar{x}}\right]$

17·4·3. Method of Minimum Chi-square. The axiom of this method is that we choose estimators for the population parameters in such a way as to minimise

$$\chi^2 = \sum_i \frac{(o_i-e_i)^2}{e_i}$$

where e_i's are expected frequencies calculated from this population. In practice a modified version of this where we minimize

$$\chi'^2=\sum_i \frac{(o_i-e_i)^2}{o_i}$$

is used.

Now the likelihood function for getting o_1 members in the first cell, o_2 in the second cell,.....and o_r in the rth cell is obtained by using the multinomial distribution

$$L=\frac{n}{o_1!\,o_2!\ldots o_r!}\left(\frac{e_1}{o_1}\right)^{o_1}\left(\frac{e_2}{o_2}\right)^{o_2}\ldots\left(\frac{e_r}{o_r}\right)^{o_r}\cdot\left(\frac{o_1}{n}\right)^{o_1}\left(\frac{o_2}{n}\right)^{o_2}\ldots\left(\frac{o_r}{n}\right)^{o_r}$$

$$\therefore\quad \log L=\text{const.}+\sum_{i=1}^{r}\left[o_i\log\left(\frac{e_i}{o_i}\right)\right]$$

For large samples we may put

$$e_i=o_i+a_i n^{\frac{1}{2}},$$

where a_i is finite and, therefore, small compared with o_i ;

$$|\,a_i n^{\frac{1}{2}}\,|<o_i \text{ and } \sum_i a_i=0$$

$$\therefore\quad \log L=C+\sum o_i\left[\log\left(1+\frac{a_i n^{\frac{1}{2}}}{o_i}\right)\right], \text{ where } C \text{ is a constant}$$

$$=C-\tfrac{1}{2}\sum\frac{na_i^2}{o_i}+O(n^{-\frac{1}{2}})$$

$$=C-\tfrac{1}{2}\sum\chi'^2+O(n^{-\frac{1}{2}}) \qquad \ldots(17{\cdot}3)$$

Again

$$\chi^2-\chi'^2=\sum(o_i-e_i)^2\left(\frac{1}{e_i}-\frac{1}{o_i}\right)$$

$$=\sum\frac{(e_i-o_i)^2}{o_i}\left[\left(1+\frac{a_i n^{\frac{1}{2}}}{o_i}\right)^{-1}-1\right]$$

$$=O(n^{-\frac{1}{2}}) \qquad \ldots(17{\cdot}4)$$

From (17·3) and (17·4), we see that for large samples, up to order $n^{-\frac{1}{2}}$, the estimators obtained by minimising χ'^2 or χ^2 are equivalent to the estimators obtained by maximizing the likelihood function.

17·4·4. Method of Inverse Probability. The axiom of this method is that for a given '*a priori*' density function $g(\theta)$ of the parameter θ, we should choose our estimator θ so that the '*a posteriorie*' probability function for θ as obtained from the given random sample is a maximum. Now according to Bayes theorem.

"*A Posteriori*" probability $P \propto$ likelihood function $\times$ '*a priorie*' density function of θ

so that $$\frac{\partial}{\partial\theta}\log P\propto\frac{\partial}{\partial\theta}\log L+\frac{\partial}{\partial\theta}\log g(\theta).$$

Since on the R.H.S., the first factor has ordinarily a variance of order n, while the second term is independent of n, it appears that, for large n, the inverse probability estimators will approach the maximum likelihood estimators, whatever be the density function $g(\theta)$.

According to the relative frequency definition of probability, θ cannot have a density function (except the trivial one, *viz.*, that $P(\theta=\theta_0)=1$, $P(\theta\neq\theta_0)=0$ where θ_0 is the true value). According to Bayes postulate $g(\theta)$ is constant for a finite range and according to others there is a definite interpretation of $g(\theta)$ as denoting '*a priori*' measure of belief. The method has been subject to a great deal of controversy.

17·4·5. Method of Least Squares. The application of this method in the estimation of the parameters $b_0, b_1, \ldots\ldots, b_p$ for fitting a curve $Y=b_0+b_1x+\ldots\ldots+b_px^p$ has already been discussed in chapter 9.

In the more general case when we have n equations

$$\phi(x_i, \theta_0, \theta_1, \ldots\ldots, \theta_p)=0, i=1, 2, \ldots\ldots, n$$

to solve for $(p+1)$ parameters $\theta_0, \theta_1, \ldots\ldots, \theta_p$ and $n>p+1$, we apply the method of least squares and choose the parameters so as to minimize the sum of squares of residuals, *i.e.*, choose θ's so as to satisfy the $(p+1)$ equations

$$\frac{\partial}{\partial\theta_j}\Sigma\{\phi(x_i, \theta_0, \theta_1\ldots\ldots, \theta_p)\}^2=0, j=0, 1, 2, \ldots\ldots, p$$

In the case when the residuals are distributed normally with variance σ^2, the logarithm of likelihood function is given by

$$\log L=\text{const.}-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\phi^2(x_i, \theta_0, \theta_1, \ldots\ldots, \theta_p).$$

This is clearly examined by minimizing the sum of squares of residuals so that in this case the method of least squares will yield the same result of the method of maximum likelihood.

17·5. Interval Estimation. In the last three sections we have discussed the criteria which optimum '*point estimators*' should satisfy and the method by which such estimators may be obtained. A single estimator, however good it may be, is not expected to coincide with the population parameter and the distribution of errors is determined by the sampling distribution of the estimator. It is, therefore, desirable to give for any population parameter an

interval in which the population parameter may be expected to lie, with a specified degree of confidence. This is the problem of 'interval estimation' and the intervals for the parameters are called *confidence intervals.*

We have already determined confidence intervals for a number of cases. We first examine the argument behind these a little more closely.

Consider the estimation of the population mean from a random sample $x_1, x_2, \ldots\ldots, x_n$ drawn from a normal population with known variance σ^2. Since $\bar{x}$, in this case, is distributed normally about mean μ with variance σ^2/n, we have

$$P\left[-1{\cdot}96 \leqslant \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leqslant 1{\cdot}96\right] = 95 \qquad \ldots(17{\cdot}5)$$

or
$$P\left[\bar{x}-1{\cdot}96\,\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+1{\cdot}96\frac{\sigma}{\sqrt{n}}\right] = {\cdot}95 \qquad \ldots(17{\cdot}6)$$

and the interval $\bar{x}-1{\cdot}96\ \sigma/\sqrt{n}$, $\bar{x}+1{\cdot}96\sigma/\sqrt{n}$ was called the 95% confidence interval for the population mean μ. What does this statement mean ?

We note first that it does not mean that the probability for μ to lie in this interval is ·95, for this would imply that μ is a random variate, which is not the case, μ is a fixed, though unknown, constant and either the specified interval contains μ or it does not contain μ If it contains μ, the probability is 1, if it does not contain μ, the probability is zero.

Though μ is not a random variate the interval $\bar{x} \mp 1{\cdot}96\ \sigma/\sqrt{n}$ is a random interval, since its limits depend on chance and would be different for different samples. If we take a large number of random samples, we shall get a corresponding large number of random intervals and equation (17·6) is interpreted to mean that on an average, 95% of these intervals are expected to include the true population parameter μ.

To make it clear that in (17·6) μ is not to be interpreted as if μ is a random variate, we write it as

$$P_F\left[\bar{x}-1{\cdot}96\,\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+1{\cdot}96\,\frac{\sigma}{\sqrt{n}}\right] = {\cdot}95 \qquad \ldots(17{\cdot}7)$$

and call it the fiducial probability and this gives the probability even before drawing the sample that the resulting interval will

contain the true parameter value. Another alternative way of writing this is

$$P[\mu \ \varepsilon \ \delta_n]=\cdot 95$$

where δ_n is the interval $\bar{x}-1\cdot 96\ \sigma/\sqrt{n}$, $\bar{x}+1\cdot 96\ \sigma/\sqrt{n}$ and 'ε' is read as 'is contained in' or is 'covered by'.

It may be noted that even for the same confidence coefficient ·95, say we can have any number of confidence intervals. Thus from the normal distribution table, we find

$$P_F\left[\bar{x}-1\cdot 68\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+2\cdot 70\frac{\sigma}{\sqrt{n}}\right]=\cdot 95$$

$$P_F\left[\bar{x}-1\cdot 86\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+2\cdot 08\frac{\sigma}{\sqrt{n}}\right]=\cdot 95$$

But these intervals are of different lengths and for obvious reasons we prefer, for a given confidence coefficient, the intervals of shortest length.

Ex 1. Prove

$$P_F\left(x-\frac{2\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+\frac{2\sigma}{\sqrt{n}}\right)=P_F\left(\bar{x}-\frac{2\cdot 326\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+\frac{1\cdot 806\sigma}{\sqrt{n}}\right)=\cdot 9543.$$

Ex. 2. Find at least four confidence intervals for μ with confidence coefficient ·99.

Ex. 3. Show that for a symmetrical unimodal distribution, the confidence interval of shortest length is central.

17·5·1. 95 per cent confidence intervals for special cases. In earlier chapters we have found confidence intervals for a number of population parameters. In each case our success was due to our *finding a suitable function of sample and parameter whose distribution was independent of the population parameters.* We summarise the earlier results on the next page and then proceed to find confidence interval in a few more cases. For definiteness we give confidence limits corresponding to a confidence coefficient 95.

Example 17·10. *Show that for the rectangular population*

$$dP=\frac{dx}{\theta},\ 0 \leqslant x \leqslant \theta$$

and confidence coefficient α, confidence limits for θ are R and $R/\psi\alpha$, where R is the sample range and $\psi\alpha$ is given by

$$\psi_\alpha{}^{n-1}[n-(n-1)\psi\alpha]=1-\alpha.$$

(M.A. Patna 1958 ; M.A. Statistics Delhi, 1958)

The distribution of R has been obtained in § 14·5, as

$$dP=n(n-1)\theta^{-n}R^{n-2}\,(\theta-R)dR, \qquad 0 \leqslant R \leqslant \theta.$$

	Population or Populations	*Parameter*	*Confidence limits*	*Ref.*
1.	Normal with known variance σ^2	μ	$\bar{x} \mp 1{\cdot}96 \frac{\sigma}{\sqrt{n}}$	§12·7
2.	Normal with unknown variance	μ	$\bar{x} \mp \frac{t_{0{\cdot}05}}{\sqrt{n}} \sqrt{\Sigma \frac{(x_i-\bar{x})^2}{n-1}}$	§15·2
3.	Two normal with unknown variances σ_1^2, σ_2^2	$\mu_1-\mu_2$	$\bar{x}_1-x_2 \mp 1{\cdot}96 \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	§12·8
4.	Two normal with unknown but same variances	$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2 \mp t_{0{\cdot}05} S \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	§15·2
5.	Bivariate normal	ρ	$\tanh\left[\tanh^{-1} r_1 \mp 1{\cdot}96 \sqrt{\frac{1}{n-3}}\right]$	§15·5
6.	Bivariate normal	$\rho_1-\rho_2$	$\tanh\left[\tanh^{-1}r_1-\tanh^{-1}r_2 \mp 1\ 96\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}\right]$	§15·5
7.	Normal	σ^2	$\frac{\Sigma(x_i-\bar{x})^2}{\chi^2_{0{\cdot}025}}, \frac{\Sigma(x_i-\bar{x})^2}{\chi^2_{0\ 975}}$	§16·4
8.	Binomial	p	$\frac{x}{n} \mp 1{\cdot}96 \sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$	§12·3
9.	Binomial	p_1-p_2	$\frac{x_1}{n_1}-\frac{x_2}{n_2} \mp 1{\cdot}96 \sqrt{\frac{\frac{x_1}{n}\left(1-\frac{x_1}{n_1}\right)}{n_1}+\frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$	§12·4
10.	Normal	σ_1^2/σ_2^2	$\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{0\ 025}}, \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{0{\cdot}975}}$	§15·3
11.	Normal		$\beta b \mp t_{0{\cdot}05}\{[n-2)\Sigma_i(x_i-\bar{x})]^2/\Sigma_i(y_i-Y_i)^2\}^{-\frac{1}{2}}$	§15·2

Therefore the distribution of $\psi = R/\theta$

is $$dP = n(n-1)\psi^{n-2}(1-\psi)d\psi \qquad 0 \leqslant \psi \leqslant 1.$$

Thus we have found ψ, a function of the sample and parameter, whose distribution is independent of the parameter θ. Now let ψ_α be defined by

$$\int_{\psi_\alpha}^{1} n(n-1)\psi^{n-2}(1-\psi)d\psi = \alpha$$

i.e., $$\psi_\alpha{}^{n-1}[n-(n-1)\psi_\alpha] = 1-\alpha$$

Now $$P_F(\psi_\alpha \leqslant \psi \leqslant 1) = \alpha$$

or $$P_F\left[\psi_\alpha \leqslant \frac{R}{\theta} \leqslant 1\right] = \alpha$$

or $$P_F\left[R \leqslant \theta \leqslant \frac{R}{\psi_\alpha}\right] = \alpha$$

so that the confidence limits for θ are R and R/ψ_α.

Example 17·11. *Obtain the 95% confidence region for mean, variance of normal distribution.*

For obtaining the confidence limits for μ and σ^2 separately, we have used respectively the t and χ^2 distributions, but for the two together we cannot combine these since t and χ^2 are not independently distributed ; but in § 13·3 we saw that the mean and variance of a random sample from a normal distribution are independently distributed. Accordingly we find the three numbers a, b_1, b_2 such that

$$P\left[-a \leqslant \frac{\bar{x}-\mu}{(\sigma/\sqrt{n})} \leqslant a\right] = \cdot 95$$

and $$P\left[b_1 \leqslant \frac{\Sigma(x_i-\bar{x})^2}{\sigma^2} \leqslant b_2\right] = \cdot 95$$

so that $$P\left[-a \leqslant \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leqslant a\,;\; b_1 \leqslant \frac{\Sigma(x_i-\bar{x})^2}{\sigma^2} \leqslant b_2\right] = \cdot 95$$

because of the independence of the two distributions.

Thus the 95% confidence region is the area bounded by the straight lines

$$\sigma^2 = \frac{\Sigma(x_i-\bar{x})^2}{b_1}$$

$$\sigma^2 = \frac{\Sigma(x_i-\bar{x})^2}{b_2}$$

and the parabola

$$(\mu-\bar{x})^2 = \frac{a^2\sigma^2}{n}$$

in the parametric (μ, σ^2) space.

Ex. 1. The largest observation L of a sample of n from a rectangular distribution with density function $f(x)=\frac{1}{\theta}$, $(0\leqslant x\leqslant\theta)$ has the distribution

$$dP=n\left(\frac{L}{\theta}\right)^{n-1}\frac{dL}{\theta}. \qquad 0\leqslant L\leqslant\theta$$

so that the distribution of $v=\frac{L}{\theta}$ is

$$dP=n\,v^{n-1}\,dv \qquad 0\leqslant v\leqslant 1$$

and is independent of θ. Hence deduce that the confidence limits for confidence coefficient α are

$$L,\ L/(1-\alpha)^{\frac{1}{n}}.$$

Find also the shortest confidence interval. *(M.A., Aligarh, 1960)*

Ex. 2. Show that the length and the variance of the length of the t confidence interval approaches zero with increasing sample size.

Ex. 3. Show that the length of the confidence interval for σ^2 from a normal population approaches zero with sample size.

Ex. 4. Compute two different 95% confidence intervals for the range of a rectangular distribution, given the sample 1·5, 2·4, 3·8, 4·2, 2·8 given that the lower limit of the angle is zero.

Ex. 5. A random sample of two members from a rectangular population with range θ is 9, −1. Obtain 81% confidence limits for θ from the sample.
(M.A., Poona, 1959)

Ex. 6. Show that the 95% confidence limits for the parameter α in a random sample of size one from the population $f(x,\alpha)=\frac{2}{\alpha^2}(\alpha-x)$, $0<x<\alpha$ are $\frac{2x}{1-\sqrt{0\cdot25}}$, $\frac{2x}{1-\sqrt{\cdot975}}$ *(M.A. Statistics, Delhi, 1960)*

[**Hint.** $\frac{\alpha-x}{\alpha}$ is distributed independently of α.]

17·5·2. Confidence Intervals for Large Samples. We know that, under very general conditions, the maximum likelihood estimator $\hat{\theta}$ is distributed normally about mean θ with variance

$$\sigma^2(\hat{\theta})=\frac{-1}{nE\left[\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right]}$$

$\therefore$ For large samples

$$P\left[-1\cdot96\leqslant\frac{\hat{\theta}-\theta}{\sigma(\hat{\theta})}\leqslant1\cdot96\right]=\cdot95$$

For large samples, $\sigma(\theta)$ may be replaced by $\sigma(\hat{\theta})$ so that

$$P_F[\hat{\theta}-1\cdot96\ \sigma(\hat{\theta})\leqslant\theta\leqslant\hat{\theta}+1\cdot96\ \sigma(\hat{\theta})]=\cdot95$$

Similarly the confidence interval for any confidence coefficient α can be obtained.

Example 17·12. *Show that for the distribution*

$$dP=\theta\, e^{-x\theta}\, dx, \qquad 0\leqslant x\leqslant\infty$$

central confidence limits for large sample with $\alpha=0{\cdot}95$ *are given by*

$$\theta=\frac{1\pm\frac{1{\cdot}96}{\sqrt{n}}}{\bar{x}}$$

(*M.A. Statistics, Delhi, 1959*)

Here $L=\theta^n\, e^{-(x_1+x_2=\ldots\ldots+x_n)\theta}$

$\therefore$ $\log L=n\log\theta-n\bar{x}\theta$

Hence $\frac{\partial}{\partial\theta}\log L=\frac{n}{\theta}-n\bar{x}.$

$\therefore$ maximum likelihood estimator of θ is

$$\hat{\theta}=\frac{1}{\bar{x}}$$

Also $\text{Var}\,(\theta)=-\frac{1}{nE\left[\frac{\partial^2}{\partial\theta^2}\log f(x,\,\theta)\right]}=-\frac{1}{nE\left[-\frac{1}{\theta^2}\right]}=\frac{\theta^2}{n}.$

$$P\left\{-1{\cdot}96\leqslant\left[\frac{\frac{1}{\bar{x}}-\theta}{\frac{\theta}{\sqrt{n}}}\right]\leqslant 1{\cdot}96\right\}={\cdot}95$$

or $$P\left[1-\frac{1{\cdot}96}{\sqrt{n}}\leqslant\frac{1}{\bar{x}\theta}\leqslant 1+\frac{1{\cdot}96}{\sqrt{n}}\right]={\cdot}95$$

$$P_F\left\{\frac{1}{\bar{x}\left(1+\frac{1{\cdot}96}{\sqrt{n}}\right)}\leqslant\theta\leqslant\frac{1}{\bar{x}\left(1-\frac{1{\cdot}96}{\sqrt{n}}\right)}\right\}={\cdot}95$$

Note. The length of a confidence interval depends strongly on what estimator is used. Thus in estimating for μ from a normal population, the estimator x_t, where x_t is the smallest or largest or any other member of the sample given an interval $\sqrt{n}$ times as long as obtained by using the sample mean from sample of size n. In general large sample confidence intervals and regions based on maximum likelihood estimators will be smaller, on an average, than intervals and regions determined by other estimators of the parameters.

Ex. 1. Apply the above method for estimating μ from a normal distribution.

Ex. 2. Show that for large samples, the 95% confidence limits for the population parameter for a Poisson distribution are

$$\bar{x}+\frac{1{\cdot}96}{n}\pm\sqrt{3{\cdot}84\,\frac{\bar{x}}{n}+\frac{3{\cdot}69}{n^2}}$$

Show that, to order $n^{-\frac{1}{2}}$, this is equivalent to

$$\bar{x}+1{\cdot}96\sqrt{\bar{x}/n}.$$

EXERCISES 17

1. If in n trials of an event with probability x there are x successes, show that a maximum likelihood estimator of p is x/n. Find the sampling variance and show that it is sufficient.

2. Show that the distribution

$$dP=\tfrac{1}{2}\exp[-|x-\theta|]dx, \qquad -\infty\leqslant x\leqslant\infty$$

has a likelihood function for a sample of n which is maximum at the medium if n is odd and between the $\frac{n}{2}$ th and $\left(\frac{n}{2}+1\right)$th members if n is given.

[**Hint.** Use § 3·2·3.]

3. Show that the most general form of distribution for which the maximum likelihood estimator is

(*i*) the arithmetic mean of the sample is

$$f(x,\theta)=k\exp\left[(x-\theta)\frac{\partial\psi}{\partial\theta}+\psi(\theta)+\phi(x)\right]$$

(*ii*) the geometric mean of the sample is

$$f(x,\theta)=(x/\theta)^{\theta\frac{\partial\psi}{\partial\theta}}\exp[\psi(\theta)+\varphi(x)].$$

(*iii*) The harmonic mean of the sample is

$$f(x,\theta)=\exp\left[\frac{1}{x}\left\{\frac{\theta\partial\psi}{\partial\theta}-\psi(\theta)\right\}-\frac{\partial\psi}{\partial\theta}+\varphi(x)\right]$$

(*M.A. (Stat.) Delhi 1968*)

4. Find the maximum-likelihood estimator for ρ in samples of n from

$$dP=\frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right\}dxdy$$

and show that it is not the product moment correlation coefficient. Prove also that

$$\mathrm{Var}(\hat{\rho})=\frac{(1-\rho^2)}{n(1+\rho^2)}$$

5. Show that the maximum likelihood estimators for the five parameters frøm the bivariate normal population

$$dP=\frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\alpha}{\sigma_1}\right)^2-\frac{2\rho(x-\alpha)(y-\beta)}{\sigma_1\sigma_2}+\left(\frac{y-\beta}{\sigma^2}\right)^2\right\}\right]dxdy \qquad -\infty\leqslant x,\ y\leqslant\infty$$

are

$$\hat{\alpha}=\bar{x},\ \hat{\beta}=\bar{y},$$

$$\hat{\rho}_1^2=\frac{1}{n}\Sigma(x_i-\bar{x})^2,\ \hat{\rho}\,\hat{\sigma}_1\,\hat{\sigma}_2=\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y}),\ \hat{\sigma}_2^2=\frac{1}{2}\Sigma(\bar{y}_i-\bar{y})^2.$$

6. Show that if a most efficient estimator A and a less efficient estimator B tend to joint normality for large sample $B-A$ tend to zero correlation with A.

Show that the error in B may be regarded as composed for large samples of two parts which are independent, the error in A and the error in $B-A$. Interpret the first as the sampling error, necessarily inherent in the problem of estimation, the second as error due to inefficiency of the estimator.

(*M.A. Patna, 1958*)

7. Find the maximum likelihood estimator and variance for β for the Gamma distribution

$$f(x)=\frac{1}{\beta^{\alpha}\,\Gamma\alpha}\,e^{-\frac{x}{\beta}}\,x^{\alpha}, \qquad 0\leqslant x\leqslant\infty,\ \alpha \text{ being known.}$$

[Ans. $\widehat{\beta}=\bar{x}/\alpha$, Var $(\widehat{\beta})=\beta^2/n\alpha$]

8. Find the maximum likelihood estimator for p for

$$f(x)=^{2}c_x p^{x} q^{2-x}\ ;\ x=1,\ 2,\ q=1-p.$$

9. A sample of size n is drawn from each of four normal populations all of which have same variance σ^2. The means of the four populations are $a+b+c$, $a+b-c$, $a-b+c$, $a-b-c$. What are the maximum likelihood estimators for a, b, c and σ^2? (*M A. Statistics, Delhi, 1959* ; *M.A. Delhi, 1954*)

[**Hint.** The sample observations may be denoted by x_{ij}. $i=1, 2, 3, 4$ and $j=1, 2, 3, \ldots\ldots, n$.

$$L=\frac{1}{(2\sigma^2\pi)}\exp\left[-\frac{1}{2\sigma^2}\{\Sigma(x_{1j}-\overline{a+b+c})^2+\Sigma(x_{2j}-a+b-c)^2+\Sigma(\ )^2+\Sigma(\ )^2\right]$$

$$\therefore\ \frac{\partial}{\partial a}\log L=0 \text{ gives } \widehat{a}=\frac{1}{4}\,\Sigma\Sigma x_{1j}=\bar{x}, \text{ etc.}$$

10. Prove that a sufficient condition for t_n to be a consistent statistic for estimating θ is that $E(t_n)\to\theta$ and Var $(t_n)\to 0$ as $n\to\infty$.

11. Develop a method for finding the 95% confidence interval for the parameter of (*i*) a Binomial distribution (*ii*) a Piosson distribution.

(*I.C.A.R., Delhi, 1959*)

12. Show that for the rectangular distribution $f(x)=1/\theta$, $(0\leqslant x\leqslant\theta)$ the confidence limits for samples x_1, x_2 of two are

$$\frac{x_1+x_2}{1+\sqrt{1+\alpha}},\quad \frac{x_1+x_2}{1+\sqrt{1-\alpha}}.$$

13. For the type III distribution

$$dP=\frac{e^{-\frac{x}{\theta}}\,x^{p-1}}{\Gamma p\ \theta^{p}}\,dx \qquad \theta\leqslant x<\infty,\ p>1$$

where p is known, find the exact sampling distribution of the maximum likelihood estimator $\widehat{\theta}$ and show that $np\widehat{\theta}/\theta$ is distributed independently of θ. Hence find the confidence interval for θ with fiducial probability α.

14. Show that for the distribution

$$dP=\frac{1}{2(1-e^{-\theta})}\,e^{-|x|}\,dx,\ -\theta\leqslant x\leqslant 0,$$

the maximum likelihood estimate of θ is the largest member among the members

$$|x_1|,\ldots\ldots,|x_n|$$

15. For the double Poisson distribution

$$P(x=i)=\tfrac{1}{2}\frac{e^{-m_1}\,m_1^{i}}{i!}+\tfrac{1}{2}\,\frac{e^{-m_2}\,m_2^{i}}{i!},\ i=0,\ 1,\ 2,\ \ldots..$$

show that the moment method estimates for m_1, m_2 are

$$\mu_2'=\sqrt{\mu_2'-\mu_1'-\mu_1'^2}.$$

16. According to a certain genetic model, the proportions of insects in three classes should be $q^2 : 2pq : p^2$ where $p+q=1$. If n_1, n_2, n_3 are the observed frequencies in the classes, prove that

$$L=q^{2n_1}(2pq)^{n_2}\ p^{2n_3}$$

and that the maximum likelihood estimator for q is

$$\widehat{q}=\frac{2n_1+n_2}{2(n_1+n_2+n_3)}.$$

17. For the multinomial distribution, the observed frequencies in the k classes are $n_1, n_2, \ldots\ldots, n_k$. By maximizing the likelihood function

$$L=\frac{n!}{n_1!\,n_2!\ldots\ldots,n_k!}p_1^{n_1}p_2^{n_2}\ldots p_k^{n_k}$$

subject to $\sum_{1}^{n} p_i=1,\ \sum_{i=1}^{k} n_i=n,$

show that the maximum likelihood estimators are

$$\widehat{p_1}=\frac{n_1}{n},\ \widehat{p_2}=\frac{n_2}{n},\ldots\ldots,\widehat{p_k}=\frac{n_k}{n}.$$

18. If $x_1, x_2, \ldots\ldots, x_n$ is a random sample for a population with mean θ, prove that

$$a_1x_1+a_2x_2+\ldots\ldots+a_nx_n$$

will be an unbiased estimate for θ if $a_1+a_2+\ldots\ldots+a_n=1$. Show also that the variance of an unbiased linear estimate will be least if that estimate is sample mean.

[**Hint.** Var $T=\sigma^2(a_1^2+a_2^2+\ldots+a_n^2)$ $\therefore$ $a_1da_1+a_2da_2+\ldots+a_nda_n=0$

$\therefore$ $a_1=a_2=\ldots=a_n=p$, say $=\frac{1}{n}$ $\therefore$ Var $T=\frac{\sigma^2}{n}$.

19. Prove that for a random sample for a population with finite variance σ^2.

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

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

Deduce that $\frac{n-1}{n}s^2$ is an unbiased estimator for σ^2.

20. If T be an unbiased estimate of the parameter θ, prove that in large samples

$$Var_{F(T)}=Var_T\left\{\frac{d}{d\theta}F(\theta)\right\}^2$$

where $F(T)$ is any continuous function of T, *Var* T and *Var* $F(T)$ are the variances of T and $F(T)$ respectively. Also show that in large samples if θ be the expectation of T, then $F(\theta)$ is the expected value of $F(T)$. Show how the above result can be used in transforming variates so that its variance may be free from parameter value. *(I.C.A.R., Delhi, 1947)*

[**Hint.** In large samples, it is assumed that T is distributed θ normally about mean with variance $Var\ T \propto O\left(\frac{1}{\sqrt{n}}\right)$ and, therefore, $[T-\theta] \propto O\left(\frac{1}{\sqrt{n}}\right)$ θ being the size of sample. Assuming the derivatives of $F(T)$ are continuous at $T=\theta$, we have by expanding $F(T)$ by Taylor's theorem at $T=\theta$,

$$F(T)=F(\theta)+(T-\theta)F'(\theta)+O\left(\frac{1}{n}\right)\cdot \text{Suppose that } F'(\theta)\neq 0.$$

$$\therefore\ T-\theta=\frac{F(T)-F(\theta)}{F'(\theta)} \text{ and } dF(T)=F'(T)dT$$

$$\therefore\ dT=\frac{dF(T)}{F'(T)}=\frac{dF(T)}{F'(\theta)+(T-\theta)F''(\theta)}$$

or
$$dT=\frac{dF(T)}{F'(\theta)}\times\left[1+(T-\theta)\frac{F''(\theta)}{F(\theta)}\right]^{-1}$$

$$=\frac{dF(T)}{F'(\theta)}, \text{ neglecting terms higher than } \frac{1}{\sqrt{n}}.$$

The distribution of T is

$$dF=\frac{1}{\sqrt{2\pi\ Var\ (T)}}\cdot e^{-\frac{1}{2}\frac{(T-\theta)^2}{Var\ T}}$$

$$dT=\frac{1}{\sqrt{2\pi\ VarT}}e^{-\frac{1}{2}\left[\frac{F(T)-F(\theta)}{F'(\theta)}\right]^2\frac{1}{VarT}}\times\frac{dF(T)}{F(\theta)}$$

Thus $F(T)$ is distributed normally about the mean $F(\theta)$ with variance $[F(\theta)^2]\ Var\ T$.

Finally, let $Var_{F(T)}=\text{const.}=c^2$, then we see $\frac{dF(\theta)}{d\theta}=\frac{c^2}{Var\ (T)}$

and, therefore, $F(\theta)=c\int\frac{d\theta}{f(\theta)}$

where $Var(T)=[f(\theta)]^2$.

The Square Root Transformation for the Poisson variate.

If x is a Poisson variate, then $E(x)=\mu$, $Var\ (x)=\mu$.

The transformation is obtained by

$$F(\mu)=c\int\frac{d\mu}{\sqrt{\mu}}=\sqrt{\mu}$$

by choosing c, suitably.

Fisher's angular or sin^{-1} *transformation for the Binomial Proportion.*

The binomial proportion r/n has the mean value p and variance $p(1-p)/n$.

Transformation is obtained by

$$F(x)=c\int\frac{\sqrt{n}}{\sqrt{p(1-p)}}\,dp=\sin^{-1}\sqrt{p}$$

by choosing c suitably and $Var\left(\sin^{-1}\sqrt{\frac{r}{n}}\right)=\frac{1}{4n}$ approx.

(*M.A., Bombay, 1954*)

Show that if variance of x is approx. proportional to $(1-\alpha^2)^2$, then a suitable transformation for producing a variate with approx. constant variance is

$$Y=\frac{1}{x}\cdot\log_e\left(\frac{1+x}{1-x}\right).$$

21. Show that for Kapteyn's class of distributions

$$dp=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{G(x)-\mu}{2\sigma^2}\right]^2\left|\frac{dG'(x)}{dx}\right|dx$$

(*i*) If σ is known, the variance of the maximum likelihood estimator for μ is σ^2/n.

(*ii*) If μ is known, the variance of the maximum likelihood estimator for σ^2 is $2\sigma^4/n$.

(*iii*) If μ is known, the variance of the maximum likelihood estimator for σ is $\sigma^2/2n$.

22. Show that for the binomial and Poisson distributions the maximum likelihood estimator and moment estimators are identical.

23. For variable x following the law

$$f(x)=\frac{1}{2\sigma\sqrt{2\pi}}\left[e^{-\frac{1}{2\sigma^2}(x-a)^2}+e^{\frac{-1}{2\sigma^2}(x-b)^2}\right]$$

calculate the first four moments about the arithmetic mean.

Hence show that in terms of the sample moments, the best estimates of a, b and σ are

$$a=m_1+\left[\frac{3-\beta_2}{2}\,m_2^2\right]^{\frac{1}{4}},\ b=m_1-\left[\frac{3-\beta_2}{2}\,m_2^2\right]^{\frac{1}{4}},\ \sigma^2=m_1\left[1-\sqrt{\frac{3-\beta_2}{2}}\right]$$

where $m_1=\dfrac{a+b}{2}$. *(B.A., Hons., M.A., Madras, 1944)*

[Ans. $\mu_1'=m_1$, $\mu_2=\sigma_2^2+\left(\dfrac{a-b}{2}\right)^2$, $\mu_3=0$,

$\mu_4=3\sigma^4+\dfrac{3}{2}(a-b)^2\sigma^2+\dfrac{(a-b)^4}{16}$, $\beta_1=0$]

24. If x is a standard normal variate $N(0, 1)$, obtain the distribution of Z where $Z=e^{a+bx}$, $b<1$.

Prove that its rth moment is given by

$$\mu'_r=\exp\left\{ra+\frac{r^2b^2}{2}\right\}.$$

Hence obtain, using the method of fitting moments, the estimates of the parameters a and b written as $\hat{a}$ and $\hat{b}$ in terms of m_1' and m_2', the moments of the sample. *(M.A., Bombay, 1956)*

25. If e_1 and e_2 are the efficiencies of two unbiased estimates and ρ is the coefficient of correlation between them, show that

$$\sqrt{e_1e_2}-\sqrt{(1-e_1)(1-e_2)}<\rho<\sqrt{e_1e_2}+\sqrt{(1-e_1)(1-e_2)}.$$

(M.A., Bombay, 1958)

26. If T is the most efficient estimator and if T' is an unbiased estimator with efficiency e, obtain the efficiency of $lT+(1-l)T'$, where l is a constant.

(M.A., Poona, 1959)

[Ans. $\sqrt{e}$]

27. A variate x is such that its logarithm has a normal distribution with zero mean and unit variance. Show the efficiency of the sample mean $\bar{x}$ in estimating the true mean is $1/(e-1)$. *(M.A., Bombay, 1952)*

[**Hint.** The distribution of x is

$$g(x)dx \text{ where } g(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\log x)^2}.\frac{1}{x}$$

and the expected value of x is $\int_{-\infty}^{+\infty} xg(x)\,dx$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}y^2}e^{y}\,dx \text{ when } \log x=y$$

$$=\sqrt{e}$$

Similarly, $\mu_2'=e^2$ so that $\mu_2=e(e-1)$.

28. If T is the most efficient estimate of θ and T' an estimate of efficiency η, the variance of $T-T'$ is $\left(\frac{1}{\eta}-1\right)$ Var (T). *(M.A., Bombay, 1954)*

29. Estimate by the method of moments the parameters a, ξ, and σ in the distribution $p(x)=\frac{1}{\sqrt{2\pi[\sigma(x-a)]}}.e^{-\frac{1}{2\sigma^2}[\log(x-a)-\xi]^2}$, $x\geqslant a$.

(M.A., Bombay, 1954)

30. Find the maximum likelihood estimators of the parameters α and λ (λ being large) of the distribution

$$f(x,\alpha,\lambda)=\frac{1}{\Gamma\lambda}\left(\frac{\lambda}{\alpha}\right)^{\lambda}e^{-\lambda x}x^{\lambda-1}, \quad 0\leqslant x\leqslant\infty$$

You may assume that for large values of λ,

$$\psi(\lambda)=\frac{d}{d\lambda}\log\Gamma\lambda=\log\lambda-\frac{1}{2\lambda},\ \psi'(\lambda)=\frac{1}{\lambda}+\frac{1}{2\lambda^2}.$$

(M.A., Patna, 1962 ; M.A., Delhi, 1966 ; M.A., Aligarh, 1960 ; M. A. Bombay, 1956)

$\left[\text{Ans. } \bar{x},\ 1/\Sigma\left(2\log x-\frac{1}{n}\log x\right)\right]$.

[For more questions, see 'Problems in Statistics' by H.C. Saxena]

18

*Testing of Hypotheses, Sequential Analysis, Distribution-Free Methods, Statistical Decision Theory

18·1. Introduction. As remarked in Chapter 1, the present book is concerned mainly with problems of *Statistical Inference* and in the preceding seventeen chapters, we have given an account of this subject as developed by Karl Pearson and R.A. Fisher till about 1930. During the last two decades, a number of important developments, associated with the names of J. Neyman, E.S. Pearson, Abraham Wald, S.S. Wilks and others, have taken place in this subject. It is not possible within the scope of an introductory course to cover these completely. Nevertheless it is desirable that a student of Statistics should be aware of the developments. With this object, we give in this chapter some of the more elementary and basic results of these developments. Unlike the preceding chapters, however, proofs of a number of results have been omitted.

Statistical Inference may be regarded as consisting of—

(*i*) *Estimation of Parameters.*

(*ii*) *Testing of Hypothesis.*

We have discussed the first aspect in chapter 17. We shall discuss the *theory of testing hypotheses* as developed by Neyman and Pearson in section 2 of the present chapter. The testing of a hypothesis may be based on a sample whose size may be fixed in advance or the size itself may be determined during the course of the experiment by criteria which depend on the observations as they arise. The second aspect known as *Sequential analysis* was developed by Abraham Wald and would be studied in section 3. Again our tests

*For a delaited study of these topics, *see* the book 'Statistical Inference' by H.C. Saxena.

may require a specification of the form of the parent population or they may require no assumption about this form. The latter tests are called *non-parametric tests or distribution-free tests* and we shall discuss these in section 4. The statistical decision theory will be discussed in section 5.

18·2. Testing of Hypotheses.

18·2·1. Simple and Composite Hypotheses. Any statement about the population from which a given random sample $x_1, x_2, \ldots, x_n$ may have been drawn is called a *Statistical Hypothesis.* If it determines the population completely, the hypothesis is called a *Simple Hypothesis.* otherwise it is called a *Composite Hypothesis.* Thus in sampling from a bivariate normal population, characterised by five parameters, *viz.*, the two means, the two variances and a correlation coefficient, any hypothesis specifying one, two, three or four of the parameters is composite. The hypothesis would be simple only if all the five parameters are specified in addition to the normality of the population.

A hypothesis which specifies the population completely, except for ν parameters is said to be a *composite hypothesis with* ν *degrees of freedom.* Such a hypothesis may be regarded as an aggregate of (or a composite of) simple hypotheses, *e.g.*, the hypothesis that a population is normal with mean μ is composed of the aggregate, for all possible values of σ^2, of the simple hypotheses that the population is normal with mean μ and variance σ^2.

Ex. 1. "There is life on planet Mars." Is it a statistical hypothesis ?

Ex. 2. Give some examples of statistical hypotheses with 0, 1, 2, 3, 4 degrees of freedom.

Ex. 3. How many parameters there are in a multivariate normal distribution with p variates ? State some composite hypotheses about a population and explain how these can be regarded as aggregates of simple hypotheses.

[Ans. p means p variances and $\frac{p(p-1)}{2}$ covariances]

18·2·2. Two Types of Errors. Critical Region. Power Function. Operating Characteristic Function. A test of a statistical hypothesis H_o is a rule for determining a region w_o in the n-dimensional Euclidean sample space which is such that the hypothesis H_o will be rejected if the sample point falls within w_o and it will be accepted if the sample point lies outside this region. The region w_o so defined is called the *critical region* or *region of*

rejection. The rest of the sample space may be called the *region of acceptance*.

In testing a statistical hypotheses, we can make two possible types of error :

(*i*) we may reject the hypothesis when we ought to accept it, *i.e.*, when it is true, or

(*ii*) we may accept the hypothesis when we ought to reject it, *i.e.*, when it is false.

These are known as *errors of the first* and *second kind* respectively and we shall denote their probabilities by α and β respectively. We would like to carry out test, *i.e.*, choose our critical region so as to minimize both types of errors simultaneously, but this is, in general, not possible, for a sample of fixed size. In fact decreasing one type of error may very likely increase the other type. Thus by deciding to always accept the hypothesis, we can reduce type I error to zero, but in that case β would have its largest value. In practice we keep type I error fixed at a specified value and then, out of these critical regions all of which give this type I error, we choose that region which minimizes the type II error. The type I error which is the same for all these regions is sometimes called *size* of these regions.

The *power of a critical region* (or of a test) is defined as the probability of the sample point falling in this critical region when the given hypothesis H_0 is false and some other hypothesis, say H_1 is true. It is obvious that the power of a critical region will depend on the alternative hypothesis H_1 and will be different for different hypothesis H_1. Thus

Power of a critical region for testing H_0 against H_1
$= P$[rejecting H_0 when H_1 is true]
$= P$[rejecting H_0 when H_0 is false]
$= 1 - P$[accepting H_0 when H_0 is false]
$= 1 -$ type II error.

The power of a critical region refers to its 'power' of detecting the falsehood of H_0 when it is false.

In some fields of application it is more common to use the *operating characteristic function* (*OC* function) in place of the power function. The *OC* function of a test is the probability of accepting H_1 when some other hypothesis H_1 is true. The *OC* function is a

function of H_1 and can be plotted as such. It is obvious that for any H_1, the *OC* function is 'one minus the power function'.

For a given type I error, we choose our critical region so as to minimize the type II error or minimize the operating characteristic function or maximize the power function. If there exists a particular region w_0 with greater power than any other region giving same type I error, we say it is the *best critical region* and the test based on it will be called the *most powerful test.*

Ex. 1. "A judge can acquit a guilty person or conviet 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 cases. Give other examples also.

Ex. 2. Show that the *OC* function for testing H_1 against the H_0 is the power function for testing H_0 against H_1.

Ex. 3. For testing the hypothesis that a random sample comes from a normal population with mean zero, when its variance is known to be unity, a critical region of size ·05 is given by $|x| \geqslant 1{\cdot}96/\sqrt{n}$. Find its power for the hypothesis $\mu = -3, -2, -1, 0, 1, 2, 3$ and plot the power function as a function for μ.

Ex. 4. Explain the following terms.

(*i*) Errors of first and second kind
(*ii*) Best critical region
(*iii*) Power function
(*iv*) *O.C.* curve
(*v*) Simple and Composite hypotheses.

Ex. 5. Give a critical region of size ·05 for testing the hypothesis that a random sample comes from a normal population with variance σ_0^2 and determine its power function as a function of σ^2. If the test is based on tails of 'χ^2' distribution with equal areas show that the power function is less for $\frac{1}{2}\sigma_0^2$ than for σ_0^2.

18·2·3. Testing a Simple Hypothesis H_0 against an Alternative Simple Hypothesis H_1. Let $f(x, \theta)$ be the density function for the population. Let H_0 be the simple hypothesis $\theta = \theta_0$ and let H_1 be the simple hypothesis $\theta = \theta_1$. Also let L_0 and L_1 be the likelihood functions corresponding to the two hypothesis so that

$$L_0 = \prod_{i=1}^{n} f(x_i, \theta) \; ; \qquad L_1 = \prod_{i=1}^{n} f(x_i, \theta_1), \qquad \ldots(18{\cdot}1)$$

then the critical region w_0 of size α which minimizes the type II error is given by

$$L_1 \geqslant kL_0, \qquad \ldots(18{\cdot}2)$$

where the constant k is determined from the condition that

$$\int \cdots \int_{w_0} L_0 dx_1 dx_2 \ldots dx_n = \alpha. \qquad \ldots(18{\cdot}3)$$

Example 18·1. *For the population*

$$dp = \frac{1}{\sqrt{2\pi}} \; e^{-\frac{1}{2}(x-\mu)^2} dx, \quad -\infty < x < \infty$$

find w_0 *for testing* $H_o : \mu = \mu_0$ *against* $H_1 : \mu = \mu_1$.

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 - \tfrac{1}{2}n(\bar{x}-\mu_1)^2\right]$$

$$L_0 = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[=\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 - \tfrac{1}{2}n(\bar{x}-\mu_0)^2\right]$$

(18·1) then gives

$$n(2\bar{x}-\mu_0-\mu_1)(\mu_1-\mu_0) \geqslant \log_e k.$$

If $\mu_1 > \mu_0$ the critical region is given by $\bar{x} \geqslant A$, and
if $\mu_1 < \mu_0$ the critical region is given by $\bar{x} \leqslant B$,
where (18·3) gives

$$A = \mu_0 + 1{\cdot}645/\sqrt{n}, \; B = \mu_0 - 1{\cdot}645/\sqrt{n}.$$

Ex. 1. Show that if a random sample of size n is taken from the population.

$$dP = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx, \qquad -\infty \leqslant x \leqslant \infty$$

the critical region for testing the hypothesis $\sigma = \sigma_0$ against $\sigma = \sigma_1$ are given by

$$\sum_{i=1}^{n} x_i^2 \leqslant A \qquad \text{if } \sigma_1 > \sigma_0.$$

$$\sum_{i=1}^{n} x_i^2 \leqslant B \qquad \text{if } \sigma_1 < \sigma_0.$$

and determine A, B by using the χ^2 distribution.

Ex. 2. Suppose a random sample of n is taken from Poisson population $e^{-\lambda}\lambda^x/x!$ Give the most powerful critical region of size α test for testing the hypothesis $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda = \lambda_1, \lambda_1 > \lambda_0$. *(M.Sc., Madras, 1960)*

How can you use the above result to find a confidence interval for λ?

Write an expression for the power function of the test for the hypothesis $\lambda = \lambda_0$ against $\lambda > \lambda_0$. *(M.A. Statistics, Delhi, 1960)*

18·2·4. Testing a Simple Hypothesis against a Composite Hypothesis. A composite hypothesis may be regarded as an aggregate of simple hypothesis and as above, we may determine the best critical region for each of these hypothesis. In general this will vary from one simple alternative hypothesis to another. If there exists a region which is best for all the alternative hypothesis, it is called the *uniformly most powerful region* and the test based on it as the

uniformly most powerful test (*U.M.P.* test). Unfortunately, as the above example shows, a *U.M.P.* test does not, in general, exist but such a test will, in general, exist if the alternative hypothesis belong to a suitably restricted class. In general for testing $\theta=\theta_0$ a *U.M.P.* test exists if the alternative hypothesis are restricted to either $\theta>\theta_0$ or $\theta<\theta_0$.

18·2·5. Likelihood Ratio Test. Let the density function $f(x, \theta_1, \theta_2, \ldots\ldots, \theta_k)$ have k parameters and let the composite hypothesis H_0 to be tested specify $k-v$ of these, then the critical region for testing H_0 is given by

$$0<\lambda<A \qquad \ldots(18\cdot4)$$

where

$$\lambda=\frac{\text{max. likelihood for variation of unspecified parameters}}{\text{max. likelihood for variation of all parameters}} \qquad \ldots(18\cdot5)$$

λ is a function of the random sample values $x_1, x_2, \ldots\ldots, x_n$ and the constant A is to be determined from the sampling distribution of λ when H_0 is true, so that

$$P(0<\lambda<A)=a. \qquad \ldots(18\cdot6)$$

For large samples, it can be shown that when H_0 is true, $-2\log\lambda$ is asymptotically distributed as χ^2 with v degrees of freedom.

Example 18·2 *A random sample* $x_1, x_2, \ldots\ldots, x_n$ *is drawn from a normal population. Test the null hypothesis* $H_0 : \mu=\mu_0$.

Likelihood function

$$=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right]$$

Its maximum value when μ and σ are allowed to vary

$$=e^{-\frac{n}{2}}\left\{\frac{n}{2\pi\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\}^{\frac{n}{2}}$$

Its maximum value when $\mu=\mu_0$ and σ alone is allowed to vary

$$=e^{-\frac{n}{2}}\left\{\frac{n}{2\pi\sum_{i=1}^{n}(x_i-\mu_0)^2}\right\}^{\frac{n}{2}}$$

$$\therefore \quad \lambda = \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \mu_0)^2} \right]^{\frac{n}{2}} \left[\frac{1}{1 + \frac{t^2}{n-1}} \right]^{\frac{n}{2}}$$

where

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n(n-1)}}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x}^2) \right]}}.$$

Also λ is a monotonic function of t^2, so that the interval $0 < \lambda < A$ is equivalent to the critical region $t^2 > B$. For $\alpha = 0{\cdot}05$, the critical region is $| t | > t_{0{\cdot}05}$, where $t_{0{\cdot}05}$ is the 5% value of t for $(n-1)$ degrees of freedom. Thus the present method leads to the usual Student's t test for testing the significance of a mean.

Ex. 1. Explain the terms : (*i*) U.M.P.T. test (*ii*) likelihood ratio test.

Ex. 2. Develop the likelihood ratio test for testing the significance of the difference between two sample means assumed to come from normal population with same variance and show that it leads to the t-test of section 15·2·2.

Ex. 3. Test the null hypothesis that the variance of a normal population has a specified value σ_0^2 on the basis of a random sample of size n.

Ex. 4. Given random samples from each of two normal populations with means and variances (μ_1, σ_1^2) and (μ_2, σ_2^2), show that for testing the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$,

$$\lambda = \frac{\left(\frac{n_1 - 1}{n_2 - 1} F \right)^{n_1/2}}{\left(1 + \frac{n_1 - 1}{n_2 - 2} F \right)^{\frac{n_2 + n_1}{2}}}$$

Ex. 5. Given k independent random samples of sizes $n_1, n_2, \ldots\ldots, n_k$ rom k normal populations, find λ for testing the hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2.$$

18·2·6. Bias in Statistical Tests. Let $f(x, \theta)$ be the population density function and let it be desired to test the hypothesis $\theta = \theta_0$ against the other possible values of θ. For a given critical region of size α, we can find the power function for different values of θ and plot it. It happens sometimes that the power function does not have a minimum *at* $\theta = \theta_0$[*cf.* Ex. 5, page 522], but at some other value of θ, say θ_1. This would imply that we would have a greater probability of accepting the hypothesis when it is false than of rejecting it when it is true. We say that in this case the critical region is biased and we try to get unbiased critical region, *i.e.*, those for which

the power function has the minimum value at $\theta=\theta_0$. For this purpose Neyman and Pearson defined unbiased critical regions of several types A, A_1, B, C, etc., and proved their existence.

18·3. Sequential Analysis.

18·3·1. Sequential Likelihood Ratio Test. In the Neyman Pearson Theory of testing hypotheses discussed above, the sample size n and the type I error α are kept fixed and the test is so designed as to minimize the type II error β. In sequential test, on the other hand, α and β are fixed in advance and sample size n is a random variable which is determined by the *sequence* of observations as they arise. The sequential tests, lead on an average, to a saving of as many as 50 per cent observations as compared with fixed size samples for the same control of the errors.

The *sequential likelihood ratio* test for testing $H_0 : \theta=\theta_0$ against $H_1 : \theta=\theta_1$ employs the ratios

$$\lambda_m=\prod_{i=1}^{m}\frac{f(x_i,\theta_1)}{f(x_i,\theta_0)} \quad (m=1, 2, 3, \ldots\ldots) \qquad \ldots(18\cdot7)$$

and two numbers A and B, with $A>1$ and $B<1$. We go on taking the observations and after the mth observation, we take one of the following three decisions :—

(*i*) If $\lambda_m\leqslant B$, we accept H_0.

(*ii*) If $\lambda_m\geqslant A$, we accept H_1.

(*iii*) If $B<\lambda_m<A$, we continue the observations.

Thus the observations continue so long as λ_m lies between A and B and terminate as soon as either limit is crossed. *It can be shown that the probability is one that the process will terminate whatever be the distribution of n.*

To determine the numbers A and B, we compute the error α and β.

$$\alpha=P[H_0 \text{ is rejected when } H_1 \text{ is true}]$$
$$=P[\lambda_m \text{ exceeds } A \text{ before it becomes less than } B]$$
$$=P(\lambda_1\geqslant A)+P(B<\lambda_1<A,\ \lambda_2\geqslant A)+\ldots\ldots \qquad \ldots(18\cdot8)$$
$$\beta=P(\lambda_m\leqslant B \text{ before } \lambda_m \text{ exceeds } A)$$
$$=P(\lambda_1\leqslant B)+P(B<\lambda_1<A,\ \lambda_2\leqslant B)+\ldots\ldots \qquad \ldots(18\cdot9)$$

The determination of A and B from these equations would be extremely difficult. Fortunately this is not necessary as very simple and accurate approximations are available. These are

$$A\cong\frac{1-\beta}{\alpha},\quad B\cong\frac{\beta}{1-\alpha}. \qquad \ldots(18\cdot10)$$

These approximations, independent of the population distribution, make the task of carrying out a sequential test amazingly simple, since no sampling theory is needed.

18·3·2. Power Function and Average Sample Size. The power function $P(\theta)$ is the probability of rejecting $H_0 : \theta=\theta_0$ when θ is the true value and this can be obtained by adding the probabilities that H_0 will be rejected at each observation when θ is the true value so that

$$P(\theta)=P(\lambda_1>A)+P(B<\lambda_1<A,\ \lambda_2>A)+\ldots\ldots \quad \ldots(18\cdot11)$$

The exact computation would be tedious but again a very good approximation is available, *viz.*,

$$P(\theta)\cong\frac{1-B^h}{A^h-B^h}, \quad \ldots(18\cdot12)$$

where h is the non-zero number for which

$$\int_{-\infty}^{\infty}\left[\frac{f(x,\theta_1)}{f(x,\theta_0)}\right]^h f(x,\theta)dx=1. \quad \ldots(18\cdot13)$$

It can be shown that such a non-zero number h exists.

The probability $p(n)$ that the size of the sample would be n can be calculated for all n, *e.g.*,

$$p(1)=P(\lambda_1<B)+P(\lambda_1>A)$$
$$p(2)=P(B<\lambda_1<A,\ \lambda_2<B)+P(B<\lambda_1<A,\ \lambda_2>A)$$

and so on. The average sample size $E(n)$ can then be computed. A good approximation is given by

$$E(n)\cong\frac{P(\theta)\log A+[1-P(\theta)]\log B}{E[\log f(x,\theta_1)-\log f(x,\theta_0)]} \quad \ldots(18\cdot14)$$

where $P(\theta)$, A and B are determined from (18·10), (18·12) and (18·13).

$E(n)$ regarded as a function of θ is called the *average sample number function* or simply the *ASN* function. The *OC* function is easily found from (18·12) to be

$$OC \text{ function}\cong\frac{A^h-1}{A^h-B^h} \quad \ldots(18\cdot15)$$

The *ASN* function enables us to compare sequential tests with fixed size example tests.

Ex. 1. Write two more terms on the right of each of the equation (18·8), (18·9) and (18·11).

Ex. 2. Write expressions for $p(3)$, $p(4)$ and $p(5)$.

Ex. 3. Show that $P(\theta_0)=\alpha$, $P(\theta_1)=1-\beta$.

Ex. 4. For testing the null hypothesis that the mean of a normal population with known variance is μ_0 against the alternative that the mean is $\mu_1(\mu_1 > \mu_0)$, show that

$$h = \frac{\mu_1 + \mu_0 - 2\mu}{\mu_1 - \mu_0}$$

and find $P(0)$ and $E(n)$.

Ex. 5. In the above exercise, show that $E(n) \cong 6$, if

$$\mu_0 = \mu = 0,\ \mu_1 = 1,\ \alpha = \cdot 01,\ \beta = \cdot 01$$

Ex. 6. Show that (18·13) is satisfied by $h = 0$.

18·4. Distribution-Free Methods.

18·4·1. Comparison with Parametric Case. Most of the tests which we have considered in earlier chapters have been based on the assumption that the parent population is normal. Even if the parent population is not normal, we can often find transformations to reduce it to the normal form. In practice, however, our knowledge about the parent population may not be sufficient to enable us to find such a transformation and in such case we need tests which do not depend on any assumption about the form of the population. We shall consider some of these non-parametric or distribution-free tests in this section. It is obvious, however, that the collection of these results cannot be so comprehensive as in the normal case, but these tests are not only capable of wider application, but also are simpler to apply and do not need complicated sampling theory. Distribution-free methods are based on *order statistics* or *ordered samples, i.e.*, we suppose that the sample $x_1, x_2, \ldots\ldots, x_n$ is ordered so that the observations in it are in ascending order of magnitude. Also while in the parametric case the measures of location and dispersion which are most commonly used are the mean and the standard deviation respectively which do not depend on order in the sample ; in the non-parametric case, we prefer to use the median, quartiles, inter-quartile range, etc., for which an order sample is desirable.

18·4·2. Point Estimation and Confidence Intervals. The population median v is estimated by the sample median $\bar{x}$ which however is not an unbiased estimate, though the bias is not serious and tends to zero as n tends to infinity. Similarly to estimate the population quartiles or deciles we use the corresponding sample quartiles or deciles as estimators.

To construct a confidence interval for v the population median we use the fact that the probability of an observation, falling to the

left or right of ν is one-half in either case. The probability that x_r, rth order statistic, exceeds ν is given by

$$P(x_r > \nu) = \sum_{r=0}^{r-1} ({}^nc_i)(\tfrac{1}{2})^n.$$

Similarly, $$P(x_s < \nu) = \sum_{i=s}^{n} ({}^nc_i)(\tfrac{1}{2})^n$$

and $$P(x_r < \nu < x_s) = \sum_{i=r}^{s-1} ({}^nc_i)(\tfrac{1}{2})^n \qquad \ldots(18{\cdot}16)$$

(x_r, x_s) gives the confidence interval for the confidence coefficient given by the R.H S. of (18·16). It is obvious, however, that we cannot construct a confidence interval for every confidence coefficient as the R.H.S. of (18·16) can take only a certain set of values for different values of r and s. We can use linear interpolation, but usually we restrict ourselves to the confidence levels available with the simple order statistics. In fact the paucity of confidence levels for small samples constitutes one of drawbacks of distribution-free methods.

For moderate sample sizes, we can compute the R.H.S. of (18·16) either directly or by use of tables of incomplete beta functions. For large samples we can use the result that the number of successes will be asymptotically normally distributed with mean $n/2$ and variance $n/4$.

Ex. 1. Obtain estimates for Q_1, Q_3, D_2, D_7 from an ordered sample of size 15.

Ex. 2. For a sample of size 6, prove that

$$P(x_1 < \nu < x_6) \cong \cdot 97,$$

$$P(x_2 < \nu < x_5) \cong \cdot 78$$

$$P(x_1 < Q_1 < x_4) \cong \cdot 78.$$

Ex. 3. For an ordered sample of size 100 gives 99% confidence interval for ν.

Ex. 4. For an ordered sample of size n, list the various confidence levels available for estimating the upper quartile.

18·4·3. The Sign Test. Let $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots\ldots y_n)$, be two random samples of same size n from two populations and let the sample values be paired by pairing the ith member of one with the ith member of the other and consider the signs of the differences $x_i - y_i$ $(i = 1, 2, \ldots\ldots, n)$. If the two populations are continuous and *identical* the probability of a difference being positive or

negative is one half, so we can test this null hypothesis by treating the number of positive signs as a binomial variable with mean $n/2$ and variance $n/4$.

In case of 'ties' ($x_i = y_i$), we can either ignore them or decide to allot them positive or negative signs by tossing a coin or assign half of them positive signs and the other half negative signs. Usually the first choice of ignoring them is preferred and in that case the conditional probability of the sign being positive, given that the difference is non-zero, is one-half.

When applicable, a t test is generally more efficient than a sign test, but the sign test is much easier to apply and is applicable even when t-test is not applicable *viz.*, when the parent population is not necessarily normal.

Ex. For 20 paired observations without ties what number of positive difference would you consider significant at 5% level.

18·4.4. The Run Test. We have to test the null hypothesis that two ordered random samples $x_1, x_2, \ldots\ldots, x_n$, and $y_1, y_2, \ldots\ldots, y_n$, come from the same population. The two sets of observations are combined and arranged in order of magnitude as, say,

$$x_1, y_1, y_2, x_2, x_3, x_4, y_3 \cdots\cdots \qquad \ldots(18{\cdot}17)$$

and in this new arrangement we find the total number of *runs*, where a run is defined as a sequence of letters the same kind bounded by letters of the other kind. Thus (18·17) starts with a run of one x, followed by a run of two y's which is succeeded by a run of three x's and so on.

It is obvious that if the two samples are from the same population, x's and y's well be well mixed and r, the number of runs, would be large. If the two populations are so widely separated that they do not overlap, the number of runs would be only two. If the two populations have the same mean or median, but the first population has a larger variance, there will be long runs of x's at the end. Thus any difference in mean or variance tends to reduce r. The run test for testing the null hypothesis that the two populations are identical accordingly consists in counting the number of runs in the combined ordered sample and rejecting the null hypothesis if $r \leqslant r_0$ where r_0 is a number to be determined from the distribution of runs and depends on n_1, n_2 and the level of significance but not on the form of the distribution of the parent population.

The probability of exactly r runs is given by

$$p(r)=\begin{cases} 2.\ \dfrac{{}^{n_1-1}C_{k-1}\times{}^{n_2-1}C_{k-1}}{{}^{n_1+n_2}C_{n_1}}\ ;\ r=2k \\ \dfrac{{}^{n_1-1}C_{k}\times{}^{n_2-1}C_{k-1}+{}^{n_1-1}C_{k-1}\times{}^{n_2-1}C_{k}}{{}^{n_1+n_2}C_{n_1}} \\ r=2k+1 \end{cases} \qquad \ldots(18{\cdot}18)$$

The probability distribution of R, the total number of runs of $n=n_1+n_2$ objects, n_1 of type 1 and n_2 of type 2, in a random sample is

$$p(r)=f_n(r)=\begin{cases} 2\binom{n_1-1}{\frac{r}{2}-1}\binom{n_2-1}{\frac{r}{2}-1}\div\binom{n_1+n_2}{n_1}, \ r=2k \text{ i.e. if } r \text{ is even} \\ \left[\binom{n_1-1}{\frac{r-1}{2}}\binom{n_2-1}{\frac{r-3}{2}}+(\ )\div(\ )\right]\div\binom{n_1+n_2}{n_1} \\ \qquad\qquad \text{if } r \text{ is odd} \end{cases}$$

$$r=2,\ 3,\ \ldots\ldots,\ n_1+n_2$$

Proof For r even, there must be the same number of runs of both types. Thus the only values of r_1 and r_2 are $r_1=r_2=r/2$ and (2·1) is summed over this pair. If $r_1=r_2\pm1$, r is odd. In this case, (2·1) is summed over the two pairs of values

$$r_1=\frac{(r-1)}{2}$$

and

$$r_2=\frac{(r+1)}{2}\ ,\quad r_1=\frac{(r+1)}{2}$$

and

$$r_2=\frac{(r-1)}{2}\ ,$$

obtaining the given result.

$$E(R)=\sum_{i=1}^{n_1}4i\binom{n_1-1}{i-1}\binom{n_2-1}{i-1}+\sum_{i=1}^{n_1}(2i+1)\binom{n-1}{i}\binom{n_2-1}{i-1}$$
$$+\sum_{i=1}^{n_1}(2i+1)\binom{n_1-1}{i-1}\binom{n_2-1}{i}$$

To evaluate these three sums, the following lemmas are useful.

Lemma 1. $\sum_{r=0}^{c}\binom{m}{r}\binom{n}{r}=\binom{m+n}{m}$ where $c=\min(m, n)$

Proof. $(1+x)^{m+n}=(1+x)^m(1+x)^n$ for all x,

$$\sum_{i=0}^{m+n}\binom{m+n}{i}x^i=\sum_{j=0}^{m}\binom{m}{j}x^j\sum_{k=0}^{n}\binom{n}{k}x^k$$

Assume without loss generality that $c=m$ and equal the coeffi. of x^m on both sides of this$=n$, we have

$$\binom{m+n}{n}=\sum_{r=0}^{m}\binom{m}{m-r}\binom{n}{r}$$

Lemma 2. $$\sum_{r=0}^{c}\binom{m}{r}\binom{n}{r+1}=\binom{m+n}{m+1}$$

where $c=\min(m, n-1)$

[See also page 147 of Math. Statistics by S. Wilks

Proof. Proof follows as in Lemma 1, equal the coeffs. of x^{m+1}.

However it is difficult to find $E(R)$ and var R and is left as an exercise for the Reader. A simpler appproach however can be made with the help of *Indicator Fuction*. Consider R as the sum of indicator variables as follows for $n=n_1+n_2$. Let

$$R=1+I_2+I_3+\cdots\cdots+I_n$$

where in an ordered sequence of the two types of symbols, we define

$$I_k=\begin{cases}1 & \text{of the both element}\neq(k-1)\text{ st element}\\ 0 & \text{otherwise}\end{cases}$$

Then T_k is a Bernoullian random variable with parameter

$$p={}^{n_1}C_1\ {}^{n_2}C_2/{}^{n_1+n_2}C_2-n_1n_2\Big/\binom{n}{2},$$

so that

$$E\left(I_k\right)-E\left(I^2{}_k\right)=\frac{2n_1n_2}{n(n-1)}$$

$$E(R)=1+\sum_{k=2}^{n}E\left(I_k\right)=1+\frac{2n_1n_2}{n_1+n_2}$$

$$\text{Var } R=\text{Var}\left(\sum_{k=2}^{n}I_k\right)=(n-1)\text{ Var } I_k+\sum\sum_{2\leqslant j\neq k\leqslant n}\text{Cov}\left(I_j, I_k\right)$$

$$=(n-1)E\left(I^2{}_k\right)+\sum\sum_{2\leqslant j\neq k\leqslant n}$$

$$E\begin{pmatrix} I & I \\ j & k \end{pmatrix} - (n-1)^2\left[E\begin{pmatrix} I \\ k \end{pmatrix}\right]^2$$

To evaluate the $(n-1)(n-2)$ joint moments of the type

$$E\begin{pmatrix} I & I \\ j & k \end{pmatrix} \text{ for } j=1\ k,$$

the subscript choice can be classified as follows :

1. For the $2(n-2)$, selections where $j=k-1$ or $j=k+1$

$$E\begin{pmatrix} I_j & I \\ & k \end{pmatrix} = \frac{n_1n_2(n_1-1)+n_2\,n_1(n_2-1)}{n(n-1)(n-2)} = \frac{n_1n_2}{n(n-1)}$$

2. For the remaining $(n-1)(n-2)-2(n-2)=(n-2)(n-3)$ selections of jth

$$E(I_j,\ I_k) = \frac{4n_1n_2(n_1-1)(n_2-1)}{n(n-1)(n-2)(n-3)}$$

$$\therefore \text{ Var } R = \frac{2n_1n_2}{n} + \frac{2(n-2)n_1n_2}{n(n-1)} + \frac{4n_1n_2(n_1-1)(n_2-1)}{n(n-1)} - \frac{4n_1^2n_2^2}{n^2}$$

$$= \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

To test the null hypothesis at probability level P, we find r_0 from

$$\sum_{r=0}^{r_0} p(r) = P \qquad \ldots(18{\cdot}19)$$

as nearly as possible and reject H_0 if $r \leqslant r_0$.

For large n_1, n_2 the distribution of r is asymptotically normal with

$$E(r) = \bar{r} = \frac{2n_1n_2}{n_1+n_2} + 1 \qquad \ldots(18{\cdot}20)$$

$$\text{Var } r = \sigma_r^2 = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \qquad \ldots(18{\cdot}21)$$

and we reject the null hypothesis if $r < \bar{r} - 1{\cdot}645\sigma_r$. The above approximation can be used if both n_1, n_2 exceed 10.

That run test can also be used for *testing the randomness of a given sample, i.e.,* to test whether the sample observations are independent and identically distributed or in other words to test whether the phenomenon yielding the data is in the *statistical control.* Let x_1, x_2,.....,x_n denote the observations in the order obtained. For each observation we note whether it is above or below the sample median. We write a if it is above and b if it is below the sample median, obtaining a sequence of a's and b's such as, say,

aa bbb ab aa bb... ...(18·22)

If the sample *is actually* a random sample, the number of runs above and below the sample median is a random variable with

$$E(r)=\frac{n+2}{2}, \quad \text{Var}\,(r)=\frac{n}{4}\left(1-\frac{1}{n-1}\right) \qquad \ldots(18{\cdot}23)$$

The formulae are exact only if n is even and population is continuous. For large n, say above 25, the asymptotic normality of the distribution of r can be used.

Another test for randomness can be obtained by writing + or − between each pair of successive observations, depending on whether the second observation of the pair is larger or smaller than the first and counting the number of runs of positive or negative signs. Under the assumption of randomness, the distribution of the number of runs is asymptotically normal with

$$E(r)=\tfrac{1}{3}\,(2n-1) \text{ and } \text{Var}\,(r)=\tfrac{1}{90}\,(16n-29) \qquad \ldots(18{\cdot}24)$$

Ex. 1. For $n_1=6$, $n_2=8$, give the probability distribution of r and find r_0 for 5 per cent level of signification.

Ex. 2. If in a sample of size 50, there are 20 runs above and below the median : is the sample random ?

Ex. 3. Thirty observations given below are obtained :

14, 25, 2, 40, 50, 60, 58, 39, 70, 15, 59, 18, 18, 1, 73, 2.
27, 24, 44, 65, 35, 85, 65, 16, 33, 47, 84, 38, 53, 35, 39, 9.

Test their randomness by considering the sequence of positive and negative signs.

18·4·5. The Median Test. The run test discussed above is sensitive to both differences in shape and location of the two distributions. If we are interested in differences in location only, the median test is to be preferred as it is very little sensitive to differences in shapes.

Let $x_1, x_2, \ldots\ldots, x_{n_1}$ and $y_1, y_2, \ldots\ldots, y_{n_2}$ be two ordered random samples from the two populations $f(x)$ and $f(y)$ respectively and let $z_1, z_2, \ldots\ldots, z_{n_1+n_2}$ be the combined ordered sample z_a as its median. The null hypothesis is that the two populations have the same median. Under this null hypothesis, the probability that m_1 of the x's and m_2 of the y's are less than z_a is given by

$$^{n_1}C_{m_1} \cdot {}^{n_2}C_{m_2} \Big/ {}^{n_1+n_2}C_{m_1+m_2} \qquad \ldots(18{\cdot}25)$$

where we take a to be $(n_1-n_2)/2$ if n_1+n_2 is even and $(n_1+n_2+1)/2$ if n_1+n_2 is odd. If the null hypothesis is true, m_1 has a hypergeometric distribution with mean $n_1/2$ and variance $(n_1n_2)/4\,(n-1)$. For small n_1, n_2, the computation can be done numerically, while for large values of n_1, n_2 we can use the asymptotic normal distribution.

We can also arrange observations in a 2×2 contingency table and use χ^2 test.

Ex. The random samples of size 20 each are drawn and 5 of one and 15 of the other are found to be below the overall median. Prove that the probability of seven or less observations being below the overall median is ·056. Find the probability of six or less and discuss the significance of the observed result.

18·4·6. Wilcoxon-Mann Whitney Test. Let two ordered samples be combined and ordered in order of magnitude as follows, say

$$x_1, y_1, x_2, x_3, y_2, y_3\cdots\cdots \qquad \ldots(18\cdot26)$$

For each y we count the number of '*inversions*' *i.e.*, the number of x's that precede that y in the sequence. Thus for y_1, y_2, y_3 in the above sequence, there are 1, 3, 3 inversions. Alternatively for each pair of observations x_i, y_j, let

$$z_{ij}=\begin{cases} 1 \text{ if } x_i<y_j \\ 0 \text{ if } x_i>y_j \end{cases} \qquad \ldots(18\cdot27)$$

Then the total number of inversions would be

$$u=\sum_{j=1}^{n_2}\sum_{i=1}^{n_1} z_{ij} \qquad \ldots(18\cdot28)$$

Under the null hypothesis z_{ij} is a Bernoullian variate with $p=\frac{1}{2}$, and

$$E(u)=\frac{n_1 n_2}{2}, \quad \text{Var } u=\frac{n_1 n_2}{2}(n_1+n_2+1) \qquad \ldots(18\cdot29)$$

Also u is asymptotically normal with these parameters.

Ex. 1. The *rank* of any y (or x) is its position in the combined ordered sample. Let R_y, R_x; $U_y\ U_x$ be the sum of ranks and inversion respectively of y's and x's. Show that

$$R_y=U_y+\frac{n_1(n_2+1)}{2}, \quad R_x=U_x+\frac{n_1(n_1+1)}{2}$$

$$E(R_x)=\frac{n_1n_2+n_1(n_1+1)}{2}, \quad \text{Var } R_x=\text{Var } U_x.$$

Ex. 2. Two sets of data are given below. Test the hypothesis that there has been no change in the centre of population at 5% level.

(1) $-\cdot1, 3\cdot6, -7, 3, 1, 2, -8, 3, 2, 1, 8, -4, 2, 1, 2, 3, 5, 6, -1$

(2) $3, 1, 6, -1, 1, 2, 3, 4, -4, 1, 0, 6, 2, 0, -5, 1, 1, 2.$

18·4·7. Kolmorgov-Smirnov Test. The test for testing the goodness of fit between observed and expected frequencies uses the statistics

$$D_n=l.u.b.\ |\ F_n(x)-F(x)\ | \qquad \ldots(18\cdot30)$$

where $F(x)$ is the specific population distribution function being tested $F_n(x)$ is the sample distribution function given by

$$F_n(x)=\sum_{x_i\leqslant x}\frac{f_i}{n} \qquad \ldots(18\cdot31)$$

and *l.u.b.* refer to the least upper bound of the absolute difference between $F_n(x)$ and $F(x)$ *for all* x. Both the large sample and small sample distributions of D_n are independent of $F(x)$.

For samples of moderate size, some of the 5% significant values are :

n	1	3	5	7	9	11	13	15	20	50	100
5% value	·915	·708	·563	·486	·432	·391	·361	·338	·204	·19	·14

For large n the asymptotic formulae are :

Significance level	·20	·10	·05	·10
Formulae	$\frac{1\cdot07}{\sqrt{n}}$	$\frac{1\cdot22}{\sqrt{n}}$	$\frac{1\cdot36}{\sqrt{n}}$	$\frac{1\cdot63}{\sqrt{n}}$

For testing the significance of the difference between two samples of sizes n_1, n_2, *Kolmorgov-Smirnov* test uses the statistic

$$D = l.u.b. \mid F_{n1}(x) F_{n2}(x) \mid \qquad \ldots(18{\cdot}32)$$

For small values of n_1, n_2, tables are available. For large values of n_1, n_2, we reject the null hypothesis of the identity of two populations if D exceeds $a \times \sqrt{\frac{n_1+n_2}{n_1 n_2}}$, where a is 1·36 at 5% level and is 1·63 at 1% level.

18·5. Statistical Decision Theory. Statistics may be considered as the science of decision making in the presence of uncertainty. This testing of a hypothesis H_0 against a hypothesis H_1 may by regarded as a decision between two actions (*i*) accepting H_0 or (*ii*) accepting H_1. In estimation, out of a finite or infinite set of possible values of the parameter, we have to choose a subset. This again involves decision in the face of uncertainty. The problems of design of experiments is again one of decision, since here we have to decide, in the face of uncertainty, as to how many and what sort of observations should be made during the course of experimentation.

Abraham Wald was the first to give a comprehensive 'statistical decision theory' which includes estimation, testing of hypothesis and design of experiments as particular cases. The theory has had tremendous impact on the development of statistics as well as on fields like economics, military science, business, industry and operations research, for in all these fields, we are concerned with decisions in the presence of uncertainty.

An Example. Consider the problem of determining the proportions of defectives in a lot. We take a sample and on the basis of the sample we may do the following : (*i*) we may like to

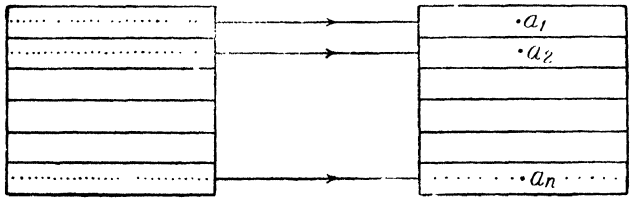
estimate the proportion of defectives in the population. This is the problem of estimation ; (*ii*) we may like to test the hypothesis that the proportion of defectives in the population is p. This is the problem of testing of hypothesis ; (*iii*) we may decide between the following actions : (*a*) rejection of the lot ; (*b*) acceptance of the lot ; (*c*) selling it at reduced price. This will be a problem in the general statistical decision theory. When we have to make a choice out of a finite number of possible decisions, we call it a multiple classification problem.

18·5·1. Basic Concepts. Each possible sample $x=(x_1, x_2,\ldots,x_n)$ may be regarded as a point in a Euclidean space. The set of all such possible points is called the *outcome space* X.

If there are m parameters $\omega_1, \omega_2, \ldots\ldots, \omega_m$. we may regard $\omega=(\omega_1, \omega_2,\ldots\ldots, \omega_m)$ as a point in m dimensional Euclidean space. The set of all such possible points is called *parameter space* Ω.

Corresponding to each ω in Ω, there is a probability distribution P_w. The set of all such probability distributions will be denoted by P.

The set A of the possible actions which the statistician can take in the light of the observations is called the action space A.



Outcoming space X Action space A

The statistician partitions the outcome space X into a number of disjoint subjects. If the observed sample is in the first subset, the statistician takes the first action, if it lies in the second, he chooses the second action and so. The outcome space X may however be divided into a countably infinite or even non-countably infinite number of subsets depending on the number of possible actions available. In every case, however, we define a mapping of every possible outcome uniquely to some action. This mapping or function is called *decision function*. A decision function d is thus a rule which assigns to every possible outcome x a unique action so that

$$a=d(x)$$

Given an outcome x, we choose an action as determined by the decision function. Since there is uncertainty, there can be some loss, positive and negative, which depends on a and the unknown parameter ω. This *loss function* is denoted by

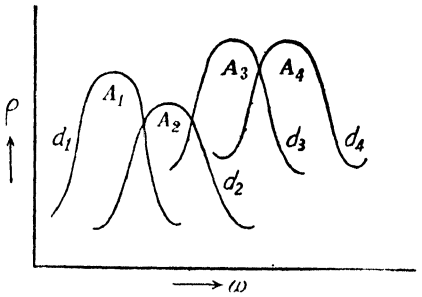
$$L(\omega, a)=L[\omega, d(x)]$$

For given (ω, d) the loss function depends on the outcome x and is thus a random variable. Its expected value, *i.e.*, its average over all possible outcomes is called the *risk function* and is denoted by

$$\rho(\omega, d)=E[L\{\omega, d(x)\}]$$

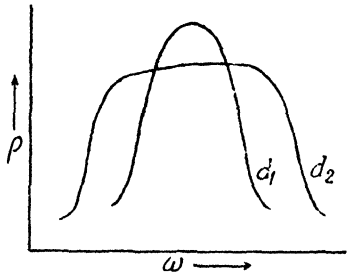
The risk function depends on the unknown ω and the decision function d.

18·5·2. Minimax Decision Rule. Our object in statistical decision theory is to determine the decision function d. For this we want to take such a decision as will minimise risk. But the risk function depends also upon the unknown ω. We now adopt the principle of minimizing the minimum risk. For a fixed d that we may choose, $\rho(\omega, d)$ depends upon ω. We find that value of ω in Ω for which this is maximum. Call this value ω_d. Then $\rho(\omega_d, d)$ is a function of d only and for each d it gives the maximum risk. The minimax principle asserts that we choose such a decision rule d which minimises $\rho(\omega_d, d)$.



In the figure we give some typical risk functions for different decision rules, A_1, A_2, A_3, A_4 give the maximum risks for different decision rules. We choose decision rule d_2 since it gives the minimum of the maximum risks.

18·5·3. Bayes Decision Rule. In the maximum decision rule we give importance to one point *viz.*, the maximum point in the decision curve.



This may not always be desirable *e.g.*, in the above figure, according to minimax rule we will prefer d_2 to d_1, since its maximum is smaller than that of d_1, but it is obvious that except for a narrow range of ω, d_1 always gives a smaller risk and as such should be preferred.

This suggests that instead of the maximum of the risk curve we may try to find some measure of average risk for each d. The average of risk function $\rho(\omega, d)$ can be found only if the *a priorie* distribution of ω is known. For a given *a priorie* distribution, we find the expected value of $p(\omega, d)$ and then we choose the d so as to minimize this average. Such a decision rule is called Bayes decision rule with respect to the given *a priorie* distribution for ω.

In many statistic al situations, the *a priorie* distribution cannot be postulated and even if it can be postulated, it is not known to the experimenter. As such the practical value of Bayes decision rule is limited. It is, however, important in the theory of decision functions. In fact, it appears that a good decision must be a Bayes decision or the limit of a sequence of Bayes decisions.

18·5·4. An Example. We use the above principles in every-day life also. An engineer when he designs a dam has to take decisions in the face of uncertainty represented by weather conditions. For any set of decision rules he makes, he considers the design which covers the maximum risk due to earthquakes and large amounts of rainfall in the catchment area. He then chooses a design out of these which minimizes the cost.

18·5·5. Games against Nature. A decision rule is quite often regarded as a game against nature. There are two players *viz.*, the statistician and Nature. The statistician can choose d and Nature can choose ω. For any rule statistician chooses Nature chooses ω so as to maximize $\rho(w, d)$. The statistician therefore chooses d so as to get the minimum of these maxima.

There is, however, no reason why Nature should be so perverse (some persons tried to prove the perversity of Nature by throwing a buttered piece of bread a large number of times. They found that in a majority of cases, the buttered face came down !) Mathematically however it is easier to determine the consequences of this complete opposition of Nature than those of partial opposition.

18·5·6. Admissible Decision Functions and Complete Classes of Decision Functions. A decision function d is said to be *admissible.* If there exists no other decision function d^+ which is uniformly better than d *i.e.*, if there exists no decision function d^+ satisfying the following two conditions

(*i*) $\rho(\omega, d^+) \leqslant \rho(\omega, d)$ for all ω in Ω

(*ii*) $\rho(\omega, d^+) < \rho(\omega, d)$ for at least one ω in Ω

A class C of decision function d is said to be *complete* if for any d not in C, we can find a d^+ in C such that d^+ is uniformly better than d. A complete class is said to be a minimal *complete class* if no proper sub-class of it is complete. It can be shown that

(*i*) if a minimal complete class exists, it must be equal to the class of all admissible decision functions.

(*ii*) if the class of all admissible decision functions is complete, it must be minimal.

The problem of finding minimal complete classes of decision functions is important in statistics. The principles of sufficiency and invariance have been used for this purpose, but we shall not go into these. It is sufficient to say that there are many unsolved mathematical problems in this area.

EXERCISES 18

1. Write an essay on 'Statistical Inference' explaining clearly the basic concepts of point and interval estimation, testing of hypotheses, sequential distribution free methods and statistical decision theory.

2. A sample of size n is drawn from each of k normal populations with the same variance. Obtain the likelihood ratio criteria for testing the hypothesis

that the means are all zero. Show that the criterion is a function of ratio which has the F distribution.

3. Describe the likelihood ratio test for tesing whether the correlation of a bivariate normal population is zero.

4. Two random samples of same size n are drawn from two bivariate normal populations with means m_1, m_2, standard deviations σ_1, σ_2 and correlation coefficients ρ_1, ρ_2. Show that for testing the hypothesis $\rho_1=\rho_2$, the likelihood ratio is given by

$$\lambda=[\operatorname{sech}\tfrac{1}{2}(z_1-z_2)]^{\frac{1}{2}n}$$

where $r_1=\tanh z_1$, $r_2=\tanh z_2$, are the same correlation coefficients.

5. In testing between the values μ_0 and μ_1, for the mean of normal population, show that the probabilities of both types of errors can be made arbitrarily small by taking a sufficiently large sample.

6. How would you make a two-sided test of $\mu=\mu_0$ for normal population with known variance ? Is this a uniformly most powerful test ?

7. It is desired to test the hypothesis $\lambda t=1$ against $\lambda t=2$ assuming that the population is of Poisson type with parameters λt. Using the mean as the basis o f the test, determine the rejection region and sample size corresponding to $\alpha=\beta=\cdot 01$.

8. Show that the one-side sequential test for the mean of a normal population with known variance σ^2 may be performed by plotting the lines

$$y=a\log A+cn,\quad y=b\log B+cn$$

where
$$a=\frac{\sigma^2\log A}{\mu_1-\mu_0},\quad b=\frac{\sigma^2\log B}{\mu_1-\mu_0},\quad c=\frac{\mu_0+\mu_1}{2}$$

in $x-y$ plane concluding the test when point $(n, \sum_{i=1}^{n} y_i)$ crosses one of the lines.

9. Show that the power function and the average sample size of the above test are given by

$$P(\mu)\cong\frac{1-e^{2(c-\mu)b/\sigma^2}}{e^{2(c-a)/\sigma^2}-e^{2(c-\mu)b/\sigma^2}}$$

$$E(\mu)\cong\frac{b+P(\mu)(b-a)}{\mu-c}.$$

10. (*a*) Develop likelihood ratio for testing binomial probability $p=p_0$ against $p=p_1$. Prove that if T_m represents the number of success in first m trials,

$$\lambda_m=\frac{p_1^{S_m}\,q_1^{m-S_m}}{p_0^{S_m}\,q_0^{m-S_m}}.$$

Discuss the test for $p_0=0\cdot 5$, $p_1=0\ 7$, $\alpha=0\cdot 5$, $\beta=\cdot 10$

(*b*) Prove that h is given by

$$p\left(\frac{p_1}{p_0}\right)^h+(1-p)\left(\frac{1-p_1}{1-p_0}\right)^h=1$$

and the operating characteristic curve is obtained by eliminating h between the above equation and

$$L(p)=\frac{A^h-1}{A^h-B^h}.$$

c) Prove that for the above test

$$E(n) \cong \frac{[1-L(p)] \log A + L(p) \log B}{p \log \frac{p_1}{p} + (1-p) \log [(1-p_1)/(1-p_0)]}$$

11. Show that the expected value of the larger of a sample of two observations from a normal population with mean x and variance σ^2 is $\mu + \sigma/\sqrt{\pi}$.

12. If x, y, is an observation from a bivariate normal population with zero mean, variances unity and correlation ρ, show that the expected value of the large of x and y is $\sqrt{(1-\rho)/\pi}$.

13. Derive the L.R.T. based on n independent observations for the hypothesis that $\theta=1$, where θ is the parameter occurring in the distribution

$$\text{Const } e^{-\theta x} x^2 \, dx,$$

and obtain the distribution of the test criterion. (*M.A. Bombay, 1955*)

14. Draw the sequential OC curve for the binomial distribution to test p_0 against p_1.

$$p_0=0{\cdot}01, \quad p_1=0{\cdot}15$$
$$\alpha=0{\cdot}01, \quad \beta=0{\cdot}02$$

obtain from the curve the probability of accepting a lot which is 20% defective.

15. Random samples of size 10 are taken from $dF=\theta^{-1} e^{-x/\theta} \, dx$, $x \geqslant 0$, $\theta>0$ obtain U.M.P. test for testing $H_0 : \theta=4$ against $H : \theta>4$ taking 0·05 as the load of significance. Draw the power curve of the test.

16. Obtain the SPRT for the hypothesis $m=m_0$ against the alternative hypothesis $m=m_0$ ($>m_0$) where m is the mean of a Poisson population. Derive the OC and ASN functions of the test. What will be the OC and ASN curve of the test if there is a population which is normally distributed with *s. d.* 3 and $H_0 : m=5$ against $H_1 : m=6$ and the probabilities of both types of error are 0·05.

17. A sequential sampling plan is to be adopted for acceptance of lots of very large size. The probability of rejection of a lot with 20% defective is to be kept at 0·05 and the probabilities of rejection of a lot with 8% defectives is to be kept at 0·1. Give the graphical and tabular procedure for currying out the above plan.

18. From each lot of very large size, 100 articles are taken at random by inspector A. He passes the lot unless there are more than 2 defectives amongst the 100 he has chosen. Inspector B follows the same procedure with another sample of 100 articles. If the two inspectors agree the lot is accepted or rejected as the case may be. If they disagree a further sample of size 200 is taken. The lot is finally accepted unless there are more than 4 defectives amongst the 200 chosen. Obtain the OC and ASN curves of the plan.

19. Discuss testing of hypothesis point and interval estimation as a special case of the general decision problem.

20. What is the relation of admissible decision functions to U.M.P. tests ?

21. A decision d_0 is said to be uniformly best if there is decision which is uniformly better than d_0. Do uniformly best decisions exist in the case of one-sided tests hypotheses ?

22. If D is a complete class of decisions and C is any other class, show that DUC is also complete.

23. Show that a complete class of decisions always exists.

19

Elements of Stochastic Processes

19·1. Basic concepts and illustrations. So far in this book, we have been concerned with the 'static' aspects of statistical theory. In the present chapter we shall be concerned with the 'dynamic' aspect or with 'statistics of change'. A stochastic process is a process which has some random or stochastic element involved in its structure.

Mathematically a stochastic process is a set of random variables $\{x_t\}$ or $\{x(t)\}$ depending on some real parameter like the time t. As examples we may consider the following :

(*i*) **A queuing system.** Here the number of persons joining the queue in a time interval t, the number of persons being serviced in the time interval t, the number of the persons in the queue at time t are all random variables depending on the parameter t.

(*ii*) **Turbulent fluid flow.** Here the fluid velocity components u, v, w are random variables depending on the space co-ordinates x, y, z and the time t.

(*iii*) **A random walk model.** A particle moves in a straight line in steps of unit length. At each stage it can move one step to the right with probability p or one step to the left with probability q where $p+q=1$. If it starts from the origin, its position after n moves is a random variable depending on the discrete parameter n.

(*iv*) **An epidemic process.** Here the number of persons injected with a disease, the number of persons susceptible to the disease, the number of persons recovered and immune from the disease are all the discrete random variables depending on the time parameter t.

(*v*) **A communication process.** Here the amplitude of the signals to be transmitted, and amplitude of the noise or disturbance produced in the channel are both random variables depending on time.

We shall come across many more examples of stochastic processes in this chapter.

A stochastic process may be classified according as (*i*) the random variable is discrete or continuous (*ii*) the parameter is discrete or continuous. If the joint distribution of x_{t_1} and x_{t_2} depends on only t_1-t_2, the process is called *stationary*, otherwise it may be termed *evolutionary*. A stochastic process is called *Gaussian* or *normal* if the joint distribution of any number of variables x_{t_1}, x_{t_2}...is a normal distribution.

Ex. 1. Give some more examples of stochastic processes.

Ex. 2. Show that $x_t=A\cos\lambda_t+B\sin\lambda_t$ where A and B are independent normal variates with zero means and unit variances, is a normal stochastic process. Show also it is a stationary process.

Ex. 3. Show that $x_t=R\cos(wt+\theta)$ is a stochastic process.

19·2. Gamblers Ruin Problem. A gambler with capital z plays against an adversary with capital $a-z$. The game is played in stages. At each stage the gambler can win one unit with probability p and lose one unit with probability q where $p+q=1$. The capital of the gambler after n stages is a random variate depending on n. Even without finding the distribution of the variate we can study some important characteristics of this process.

19·2·1. Probability of ruin. Let q_z denote the probability of ultimate ruin of the gambler. After the first trial, the gambler's capital is $z+1$ if he wins (with probability p) or it is $z-1$ if he loses (with probability q). The probability of his ruin after one trial will be q_{z+1} and q_{z-1} respectively, so that

$$q_z=pq_{z+1}+qq_{z-1} \qquad 1<z<a-1 \qquad ...(19·1)$$

It is also easily seen that

$$q_1=pq_2+q, \qquad z=1 \qquad ...(19·2)$$

$$q_{a-1}=qq_{a-2}, \qquad z=a-1 \qquad ...(19·3)$$

All these three equations are included in

$$q_z=pq_{z+1}+qq_{z-1}. \quad 1\leqslant z\leqslant a-1. \qquad ...(19·4)$$

If we use the convention

$$q_0=1, \qquad q_a=0 \qquad ...(19·5)$$

To solve (19·4), we try the solution $q_z=\rho^z$. Substitution in (19·4) gives

$$p\rho^2-\rho+q=0. \qquad ...(19·6)$$

which has roots

$$\rho=1, \quad \rho=\frac{q}{p}.$$

Thus if $p\neq\frac{1}{2}$, (19·4) has two independent solutions 1 and q/p and since (19·4) is a linear difference equation, its general solution is

$$q_z=A+B(q/p)^z. \qquad \ldots(19\cdot7)$$

Using (19·5) to find A and B, we get the final solution

$$q_z=\frac{\left(\frac{q}{p}\right)^a-\left(\frac{q}{p}\right)^z}{\left(\frac{q}{p}\right)^a-1} \qquad \ldots(19\cdot8)$$

The probability p_z of the opponent's ruin can be obtained from this by writing $a-z$ for z and interchanging p and q so that

$$p_z=\frac{\left(\frac{q}{p}\right)^z-1}{\left(\frac{q}{p}\right)^a-1} \qquad \ldots(19\cdot9)$$

From (19·8) and (19·9)

$$p_z+q_z=1, \qquad \ldots(19\cdot10)$$

so that the game has to terminate.

If $p=q=\frac{1}{2}$, (19·4) becomes

$$2q_z=q_{z+1}+q_{z-1}, \qquad \ldots(19\cdot11)$$

of which the solution, subject to (19·5) is easily found to be

$$q_z=1-\frac{z}{a} \qquad \ldots(19\cdot12)$$

In this case writing $a-z$ for z, we get

$$p_z=\frac{z}{a}, \qquad \ldots(19\cdot13)$$

so that (19·10) is still true.

19·2·2. Expected Duration of the game. Let d_z denote the expected duration of the game. If the first trial leads to a win for the gambler, the expected total duration would be $1+d_{z+1}$ and if it leads to a loss, the duration would be $1+d_{z-1}$, so that we get the difference equation

$$\begin{aligned} d_z&=p(1+d_{z+1})+q(1+d_{z-1}) \\ &=1+pd_{z+1}+qd_{z-1}. \end{aligned} \qquad \ldots(19\cdot14)$$

It is easily verified that this equation is true for $z=1$ and $z=a-1$ if also we take

$$d_0=0, \; d_a=0. \qquad \ldots(19\cdot15)$$

The equation (19·14) is non-homogeneous and its homogeneous part is the same as (19·4) when q_z is replaced by d_z. We get general solution by adding to (19·7) a particular solution λz is seen to be a particular solution of (19·4) if

$$\lambda z = 1 + p\lambda(z+1) + q\lambda(z-1) \text{ or } \lambda = 1/(q-p) \quad \ldots(19\cdot16)$$

The general solution of (19·4) is therefore

$$d_z = A + B\left(\frac{q}{p}\right)^z + \frac{1}{q-p}z \quad \ldots(19\cdot17)$$

Using (19·15), we get A and B so that finally

$$d_z = \frac{z}{q-p} - \frac{a}{q-p} \cdot \frac{1-\left(\frac{p}{p}\right)^z}{1-\left(\frac{q}{p}\right)^a} \quad \ldots(19\cdot18)$$

If $p = q = \frac{1}{2}$, we try the particular solution μz, so that

$$\mu z^2 = 1 + \mu\tfrac{1}{2}(z+1)^2 + \mu\tfrac{1}{2}(z-1)^2 \text{ giving } \mu = 1 \quad \ldots(19\cdot19)$$

The general solution is then

$$d_z = A + Bz - z^2. \quad \ldots(19\cdot20)$$

Using (19·15), we get

$$d_z = z(a-z) \quad \ldots(19\cdot21)$$

19·2·3. Gambler's Ruin with Correlation. In this case the probabilities of the gambler's winning or losing at a trial are p_1, q_1 if he wins in the earlier trial and these are q_2, p_2 if he loses in the earlier trial. If x, y denote his respective gains in the two trials and if the score of $+1$ has the same probability as a score of -1 in the first game, then

$x=1,$	$y=1,$	with probability $\frac{1}{2}\,p$
$x=1,$	$y=-1,$	with probability $\frac{1}{2}\,q$
$x=-1,$	$y=-1,$	with probability $\frac{1}{2}\,p$
$x=-1,$	$y=1,$	with probability $\frac{1}{2}\,q$

giving

$$\Sigma px = 0,\ \Sigma py = \tfrac{1}{2}(p_1+q_2-p_2-q_1),\ \Sigma px^2 = 1$$
$$\Sigma py^2 = 1,\ \Sigma pxy = p_1 - q_2$$

so that the correlation coefficient is given by

$$R_{xy} = \frac{p_1-q_2}{\sqrt{\tfrac{1}{4}(p_1+q_2-p_2-q_1)^2 - \tfrac{1}{4}(p_1+q_2-p_2-q_1)^2}}$$
$$= \frac{p_1-q_2}{\sqrt{(p_1+q_2)(p_2+q_1)}} \quad \ldots(19\cdot22)$$

Ex. 5. Find P for the cyclical random walk in which the system can be stated $E_1, E_2 \ldots E_q$, and E_a has neighbours E_{a-1} and E.

19·5. Poisson Process. A Poisson process is characterised by any of the following three equivalent properties :—

(a) The total number of arrivals during any given time interval is independent of the number of arrivals that have already occurred and the probability of an arrival in time dt is $\lambda dt + 0(dt)^2$ where λ is constant. The probability of more than one arrival in time dt is of the order of $(dt)^2$ and can be neglected. Here λ represents the mean arrival rate, *i.e.*, number of units of the incoming traffic per unit time and $1/\lambda$ is the mean length of time interval between consecutive arrivals.

(b) If n is the stochastic variate giving the number of arrivals in some time interval of length T, then n obeys a Poisson distribution with parameter λT *i.e.*

$$P(n \text{ arrivals in time } T) = \frac{e^{-\lambda T}(\lambda T)^n}{n\,!}$$

(c) The distribution of time intervals between consecutive arrivals is exponential, *i.e.*, the density function for the time interval between consecutive arrivals is

$$a(T) = \lambda e^{-\lambda T}$$

To show that these three properties are equivalent, we show that $(a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (a)$.

We have already shown that $(a) \Rightarrow (b)$ in Chapter 6.

To show that $(b) \Rightarrow (c)$, we note that

P[time interval between consecutive arrivals lies between T and $T+dT$] $= P$[no arrival during time interval T]

$\times P$[one arrival during time interval dT]

$= e^{-\lambda T}\lambda dT$

To show that $(c) \Rightarrow (a)$, we note that

P[time interval between consecutive arrivals $\geqslant T$].

$$\int_t^\infty \lambda\ e^{-\lambda T}\ dT = e^{-\lambda T}$$

$\therefore$ P[no arrival taking place in time T] $= 1 - e^{-\lambda T}$

If T is small

P[one or more arrivals in time dT]

$$= 1 - e^{-\lambda dT} = \lambda dT + 0(dT)^2$$

Ex. 1. If the random variates X_t and Y_t represent the independent stoch astic processes parameters λ and μ, prove that the process defined by

$$S_t = X_t + Y_t$$

is also a Poisson process with parameter $\lambda+\mu$.

Ex. 2. Let X_t be a random variate representing a Poisson process with parameter λ, show that

$$E(X_t)=\lambda t, \qquad \text{Var}\,(X_t)=\lambda t, \qquad E[X_t(X_{t+\tau}-X_t)]=0$$

$$E[X_t X_{t+\tau}]=\lambda t(\lambda t+\lambda t+1)$$

Deduce that the correlation coefficient between X_t and $X_{t+\tau}$ is

$$\left[\frac{t}{t+\tau}\right]^{\frac{1}{2}}$$

19·6. The Simple birth process. Consider a population of organisms multiplying in accordance with the following rules :

(*a*) the sub populations generated by two co-existing individuals develop in complete independence of one another.

(*b*) an individual existing at time t has a chance $\lambda dt+0(dt)^2$ of multiplying by binary fission during the following time-interval of length dt.

(*c*) the 'birth-rate' λ is the same for all individuals in the population at time t.

The 'birth' in (*b*) will usually be interpreted in the sense that at each birth just one new member is added to the population, but mathem atically it is not possible to distinguish between this and an alternative interpretation in which a parent dies at birth and is replaced by two new members.

Let N_0 be the number of individuals at time $t=0$. For the present we take $N_0=1$. Also let $P_n(t)$ be the probability that the population size $N(t)$ has the value n at time t, then from the given rules it follows that

$$P_n(t+dt)=P'_n(t)\,[1-n\lambda dt]+P_{n-1}(t)[(n-1)\lambda dt]+o(dt)^2,\ (n>1).$$

Proceeding to the limit as $dt\to o$, we get

$$P'_n(t)=-n\lambda P_n(t)+(n-1)\lambda P_{n-1}(t),\ (n>1) \qquad \ldots(19{\cdot}31)$$

Proceeding in the same way, we get

$$P'_1(t)=-\lambda P_1(t) \qquad \ldots(19{\cdot}32)$$

so that (19·31) also holds for $n=1$.

Consider the probability generating function,

$$\phi(z,\ t)=\sum_{n=1}^{\infty} P_n(t)z^n \qquad \ldots(19{\cdot}33)$$

so that

$$\frac{\partial \phi}{\partial t} = \sum_{n=1}^{\infty} P_n'(t)\, z^n \; ; \frac{\partial \phi}{\partial z} = \sum_{n=1}^{\infty} n P_n(t) z^{n-1} \qquad \ldots(19{\cdot}34)$$

Using (19·31) and (19·32) we get from (19·34)

$$\frac{\partial \phi}{\partial t} = \lambda z(z-1)\ \frac{\partial \phi}{\partial z} \qquad \ldots(19{\cdot}35)$$

Let

$$\frac{dz}{z(z-1)} = du$$

or

$$e^u = A\left(1 - \frac{1}{z}\right) \qquad \ldots(19{\cdot}36)$$

then (19·35) becomes

$$\frac{\partial \phi}{\partial t} - \lambda\ \frac{\partial \phi}{du} = 0, \qquad \ldots(19{\cdot}37)$$

so that

$$\phi(z, t) = f(u + \lambda t) = \psi(e^{u+\lambda t}) = \phi[(1 - 1/z)(e^{\lambda t})] \qquad \ldots(19{\cdot}38)$$

but

$$\phi(z, o) = \sum_{n=1}^{\infty} P_n(o) z^n = z \qquad \ldots(19{\cdot}39)$$

since there is only one individual at time $t=0$.

From (19·38) and (19·39)

$$\phi\left(1 - \frac{1}{z}\right) = z$$

or

$$\phi(z) = \frac{1}{1-z}$$

or

$$\phi\left[\left(1 - \frac{1}{z}\right)e^{-\lambda t}\right] = \frac{1}{1 - \left(1 - \frac{1}{z}\right)e^{\lambda t}}$$

so that

$$\phi(z, t) = z e^{-\lambda t}\,[1 + z(1 - e^{-\lambda t})]^{-1} \qquad \ldots(19{\cdot}40)$$

Expanding in powers of z we get the probability distribution for the population size at any time t, so that

$$P_n(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1}, \ (n \geqslant 1) \qquad \ldots(19{\cdot}41)$$

Remarks. (*i*) The above probability distribution is of the geometrical form and the common ratio tends to unity as t tends to infinity.

(*ii*) The mean or the expected number of individuals at time t is given by

$$E[N(t)]=N(t)=\sum_{1}^{\infty} nP_n(t)=\left.\frac{\partial \phi}{\partial z}\right]_{z=1}=e^{\lambda t} \qquad ...(19\cdot42)$$

(*iii*) The variance of the number of individuals at time t is given by

$$\text{Var } [N(t)]=E[\bar{N}(t)-N(t)]^2=\left.\frac{\partial^2 \phi}{\partial z^2}\right]_{z=1}=\bar{N}(\bar{N}-1)$$

$$=e^{\lambda t}(e^{\lambda t}-1) \qquad ...(19\cdot43)$$

(*iv*) In the more general case when the number of individuals in the population at time $t=0$ is N_0, the probability generating function (*p.g. f.*) in (19·40) will have to be raised to power N_0, since the *p.g. f.* for the sum of n independent stochastic variates is the product of the *n.p.g. f.* Also since the mean and variance of n independent stochastic variables are the sum of the means and variances respectively, the means and variances in (19·42) and (19·43) will have to be multiplied by N_0. Thus if there are N_0 individuals in the population at $t=0$ then

$$\phi(z, t)=\sum_{n=1}^{N_0} e^{-N_0\lambda t}[1+z(1-e^{-\lambda t})]^{-N_0} \qquad ...(19\cdot44)$$

$$P_n(t)=C\, e^{-N_0\lambda t}(1-e^{-\lambda t})^{n-N_0} \text{ for } n \geqslant N_0 \qquad ...(19\cdot45)$$

$$=0 \text{ for } n<N_0 \qquad ...(19\cdot46)$$

$$\bar{N}(t)=N_0e^{\lambda t}$$

$$\text{Var}[N(t)-N_0(e^{\lambda t}-1) \qquad ...(19\cdot47)$$

Alternatively (19·44) could be derived as follows :

In this case, (19·39) would become

$$\phi(z, 0)=\sum_{n=1}^{\infty} z^n P_n(t^0)=z^{No} \qquad ...(19\cdot48)$$

$$\phi\left(1-\frac{1}{z}\right)=z^{N_0} \text{ or } \phi(z)=\left(1-\frac{1}{z}\right)^{N_0} \qquad ...(19\cdot49)$$

$$\phi\left[\left(1-\frac{1}{z}\right)e^{\lambda t}\right]=\frac{N_0}{\left[1-\left(1-\frac{1}{z}\right)e^{\lambda t}\right.}$$

$$\phi(z, t)=z^{N_0}\, e^{-N_0\lambda t}[1+z(1-e^{-\lambda t})]^{-N_0} \qquad ...(19\cdot50)$$

which is same as (19·44).

Ex. 1. Solve (19·32) subject to $P_1(0)=1$. Then solve (19·31) for $n=2, 3, 4...$ in succession and verify (19·41).

Ex. 2. Solve (19·31) and (19·32) subject to the condition $P_n=(0)=1$ if $n=N_0$ and $P_n(0)=0$ if $n \neq N_1$ and verify (19·45) and (19·46).

Ex. 3. Starting with (19·45) and (19·46) verify (19·44), (19·47) and (19·48).

Ex. 4. Show that for large t, variance (NOT) is asymptotically equal to $N_0 e^{2\lambda t}$.

Ex. 5. A population of initial N_0 grows deterministically in such a fashion that at any time its rate of increase is λ times its size. Show that its size at time t is the same as the average size in the pure birth stochastic case.

19·7. The simple birth and death process. To the conditions (*a*), (*b*), (*c*), given in the last section, we add the following :

(*d*) an individual existing at time t has a chance $\mu\, dt+0\,(dt)$ of dying in the following time interval dt.

(*e*) the 'death rate' μ is the same for all individuals at time t.

From conditions (*a*) to (*e*), we get

$$\begin{aligned} P_n(t+dt)=P_n(t)[1-n\lambda dt-n\mu dt] \\ +P_{n-1}(t)[(n-1)\lambda\, dt][1-(n-1)\mu dt] \\ +P_{n+1}(t)(n+1)\mu\, dt[1-(n+1)\lambda\, dt]+0(dt) \end{aligned}$$

Proceeding to the limit as $dt \to 0$, we get

$$P'_n(t)=-nP_n(t)\,(\lambda+\mu)+(n-1)\lambda P_{n-1}(t)+(n+1)\mu P_{n+1}(t),$$

so that

$$\begin{aligned} \Sigma z^n P'_n(t)=-(\lambda+\mu)\,\Sigma n z^n P_n(t)+\lambda\,\Sigma(n-1)z^n P_{n-1}(t) \\ +\mu\Sigma(n+1)\,z^n P_{n+1}(t) \end{aligned} \qquad \ldots(19\cdot51)$$

If

$$\phi(z,\,t)=\Sigma\, z^n P_n(t) \qquad \ldots(19\cdot52)$$

denotes the probability generating function, then (19·54) gives

$$\frac{\partial \phi}{\partial t}=-(\lambda+\mu)\,z\,\frac{\partial \phi}{\partial z}+\lambda z^2\,\frac{\partial \phi}{\partial z}+\mu\,\frac{\partial \phi}{\partial z}$$

or

$$\frac{\partial \phi}{\partial t}=(\lambda z-\mu)(z-1)\,\frac{\partial \phi}{\partial z} \qquad \ldots(19\cdot53)$$

Let

$$\frac{dz}{(\lambda z-\mu)(z-1)}=du$$

or

$$\frac{\lambda z-\mu}{z-1}=e^{(\mu-\lambda)u} \qquad \ldots(19\cdot54)$$

then

$$\phi(z, t)=\phi\left[\frac{\lambda z-\mu}{z-1}\quad e^{(\mu-\lambda)t}\right] \qquad \ldots(19{\cdot}55)$$

but

$$\phi(z, 0)=z$$

so that

$$\phi(z)=\frac{\mu-z}{\lambda-z}$$

$$\phi(z, t)=\frac{\mu-\dfrac{\lambda z-\mu}{z-1}\ e^{(\mu-\lambda)t}}{\lambda-\dfrac{\lambda z-\mu}{z-1}\ e^{(\mu-\lambda)t}}$$

$$=\frac{\mu-\lambda-\lambda/\eta}{\lambda}+\frac{1}{\eta t}\frac{(1-\xi)(1-\eta)}{1-z\eta} \qquad \ldots(19{\cdot}56)$$

where

$$\frac{\xi}{\mu}=\frac{\eta}{\lambda}=\frac{e^{(\lambda-\mu)t}-1}{\lambda e^{(\lambda-\mu)t}-\mu} \qquad \ldots(19{\cdot}57)$$

Expanding (19·56) in powers of t, we get

$$P_n(t)=(1-\xi)(1-\eta)(\eta)^{n-1}\ (\eta\geqslant 1) \qquad \ldots(19{\cdot}58)$$

In the present case, it is possible that at time $t=0$, there is no individual in the population. To find its probability we have

$$P_0(t)=1-\sum_{n=1}^{\infty}P_n(t)=1-\sum_{n=1}^{\infty}(1-\xi)(1-\eta)(\eta)^{n-1}$$

$$=1-(1-\xi)=\xi \qquad \ldots(19{\cdot}59)$$

The same result can be seen from (19·56) and (19·57), since from here

$$\phi(1, t)=\frac{\mu}{\lambda}+1-\frac{1}{\eta}+\frac{1}{\eta}-\frac{\xi}{\eta}=1$$

and the constant term in $\phi(z, t)$

$$=\frac{\mu}{\lambda}+1-\frac{1}{\eta}+(1-\xi)\left(\frac{1}{\eta}-1\right)=\xi \qquad \ldots(19{\cdot}60)$$

This can also be seen from first principles, since

$$P_0(t+dt)=P_0(t)(1-o\lambda\ dt-o\mu\ dt)+P_1(t)\ \mu\ dt+0(dt)$$

so that

$$P'_0(t)=\mu P_1(t)=\mu(1-\xi)(1-\eta) \qquad \ldots(19{\cdot}61)$$

or

$$\frac{dP_0(t)}{dt}=\frac{\mu(\lambda-\mu)^2\ e^{(\mu-\lambda)t}}{[\lambda-\mu e^{(\mu-\lambda)t}]^2}$$

Integrating and remembering that $P_0(0)=0$,

$$P_0(t)=\frac{\mu\, e^{(\lambda-\mu)t}-1}{\lambda\, e^{(\lambda-\mu)t}-\mu}=\xi \qquad \ldots(19{\cdot}62)$$

Remarks : The probabilities at a given time t from a geometric series with modified first term (corresponding to no individual). The common ratio is

$$\eta=\lambda\,\frac{e^{(\lambda-\mu)t}-1}{\lambda e^{(\lambda-\mu)t}-\mu}=\lambda\,\frac{1-e^{(\lambda-\mu)t}}{\lambda-e^{(\lambda-\mu)t}} \qquad \ldots(19{\cdot}63)$$

so the ratio $\to 1$, if $\lambda>\mu$, it approaches λ/μ if $\lambda<\mu$. If $\lambda=\mu$.

$$\xi=\eta=\frac{\lambda t}{1+\lambda t} \qquad \ldots(19{\cdot}64)$$

and again the ratio approaches unity.

(*ii*) $$E[N(t)]=\bar{N}(t)=\left.\frac{\partial\phi}{\partial z}\right|_{z=1}=e^{(\lambda-\mu)t} \qquad \ldots(19{\cdot}65)$$

$$\text{Var}\,[N(t)]=\left.\frac{\partial^2\phi}{\partial z^2}\right|_{z=1}=\bar{N}\,(\bar{N}-1)$$

$$=\frac{\lambda+\mu}{\lambda-\mu}\,e^{(\lambda-\mu)t}\left\{\,e^{(\lambda-\mu)t}-1\,\right\} \qquad \ldots(19{\cdot}66)$$

If there are N_0 individuals in the population at time $t=0$, both mean and variance have to be multiplied by N_0.

When $\lambda=\mu$,

$$\bar{N}=1,\ Var\ N(t)=2\lambda t. \qquad \ldots(19{\cdot}67)$$

Thus if birth and death rates are equal, the average number of individuals at any time t will be the same as that the time $t=0$, but the variations in numbers will increase with both N_0 and t.

(*iii*) We have, for all N_0

$$\lim_{t\to\infty} P_0(t)=\begin{cases}1 & \text{if } \lambda\leqslant\mu\\ \left(\dfrac{\mu}{\lambda}\right)^{N_0} & \text{if } \lambda>\mu\end{cases} \qquad \ldots(19{\cdot}68)$$

so that the probability of extinction is certain when $t\to\infty$ not only when birth rate is less than death rate, but even when birth and death rates are equal. When birth rate is greater than the death rate, there is still a definite probability of extinction which decreases as N_0 increases.

Ex. 1. For the simple death process, prove that

(*i*) $\phi(z, t)=\{-e^{-\mu t}(1-z)\}^{N_0}$

(*ii*) $P_n(t)={}^{N_0}C_n\ e^{-n\mu t}\ (1-e^{-\mu t})^{N_0-n}$ $\qquad 0\leqslant n\leqslant N_0$

(*iii*) $P_0(t)\to 1$ as $t\to\infty$

(*iv*) mean$=N_0\, e^{-\mu t}$

(*v*) Variance$=N_0\, e^{-\mu t}(1-e^{-\mu t})$.

Ex. 2. For the birth and death process, prove that

$$P_n(t)=\sum_{j=0}^{min(N_0,\,n)} {}^{N_0}C_j\ {}^{N_0+n-j-1}C_{N_0-1}\ \xi^{N_0-j}\ \eta^{n-j}\ (1-\xi-\eta)^j$$

$P_0(t)=\xi^{N_0}$

Ex. 3. For the case of $\lambda=\mu$, show that

$$\phi(z, t)=\left\{\frac{1-(\lambda t-1)(z-1)}{1-\lambda t(z-1)}\right\}^{N_0}$$

Ex. 4. For the case $\lambda=\mu$, $N_0=1$, show that

$$P_n(t)=\frac{(\lambda t)^{n-1}}{(1+\lambda t)^{n+1}}, \qquad n\geqslant 1$$

$$P_n(t)=\frac{\lambda t}{1+\lambda t}$$

Find mean and variance of $N(t)$.

Ex. 5. Prove that

$$P_0(t)=\left[\frac{\mu\ e^{(\lambda-\mu)t}-1}{\mu\ e^{(\lambda-\mu)t}-\mu}\right]^{N_0} \qquad \text{when } \lambda\neq\mu$$

$$=\left(\frac{\lambda t}{\lambda t+1}\right)^{N_0} \qquad \text{when } \lambda=1$$

Ex. 6. Show that when $\lambda=\mu$, the mean remains constant, yet the chance of extinction is unity. Explain this.

Ex. 7. Find the differential equation satisfied by the cumulant function $k(z, t)$.

Put $\qquad k(z, t)=k_1 z+k_2\,\frac{z^2}{2!}+\ldots$

and equate the various coefficients of z. Solve the resulting equation for k_1, k_2.

Ex. 8. By substituting $P_n(t)=\phi(t)\ [\psi(t)^{n-1}$, $p_0(t)=1-\sum_{n=1}^{\infty} P_n(t)$ (19·56) shows that $\phi(t)$, $\psi(t)$ can be independent of n. Find $\mu_1(t)$ and $\mu_2(t)$ from these expressions.

Ex. 9. From (19·65) obtain differential equations for $\mu_1(t)$ and $\mu_2(t)$ and solve them.

Ex. 10. Compare the expressions $t\,\lambda_1(t)$ and $\mu_2(t)$ obtained in questions 8 and 9 to get $\phi(t)$ and $\lambda(t)$. Verify the results of question 4.

19·8. Birth, Death and Immigration Processes.

One interesting variant of the simple birth and death process is obtained if we add the condition :

(f) in an infinitesimal time interval dt, there is a chance $k\,dt+0\,(dt)$ that a single member will be added to the population by immigration from the outside world.

The characteristic feature of immigration effect is that it acts at an expected rate which is independent of the population size.

The generating function for (λ, μ, k) process satisfies the differential equation

$$\frac{\partial \phi}{\partial t} = (\lambda z - \mu)(z-1)\frac{\partial \phi}{\partial z} + k(z-1)\phi \qquad \ldots(19{\cdot}69)$$

Also $\phi(z, 0)=1$, initial population is zero $\qquad \ldots(19{\cdot}70)$

On solving (19·69), subject to (19·70), we get

$$\phi(z, t)=\left[\frac{\lambda-\mu}{\lambda e^{(\lambda-\mu)t}-\mu}\right]^{\frac{k}{\lambda}}\left[1-z\,\frac{\lambda\{(\lambda e^{(\lambda-\mu)t}-1\}}{\lambda e^{(\lambda-\mu)t}-\mu}\right]^{\frac{-k}{\lambda}} \text{ when } \lambda \neq \mu$$

$\ldots(19{\cdot}71)$

and $$\phi(z, t)=(1+\lambda t)^{\frac{k}{\lambda}}\left(1-\frac{\lambda t z}{1+\lambda t}\right)^{\frac{-k}{\mu}} \text{ when } \lambda \neq \mu. \qquad \ldots(19{\cdot}72)$$

The mean number of individuals is

$$\overline{n} = \frac{k}{\lambda-\mu}(e^{(\lambda-\mu)t}-1) \qquad \text{when } \lambda \neq \mu \qquad \ldots(19{\cdot}73)$$

$$= kt \qquad \text{when } \lambda = \mu \qquad \ldots(19{\cdot}74)$$

Remarks : (i) when $\lambda<\mu$ and $t\to\infty$, the limiting generating function is $$\phi(z)=\left(1-\frac{\lambda}{\mu}\right)^{\frac{-k}{\lambda}}\left(1-\frac{\lambda z}{\mu}\right)^{\frac{-k}{\lambda}} \qquad \ldots(19{\cdot}75)$$

and in the limit the mean population size is

$$\overline{n} = \frac{k}{\mu-\lambda}. \qquad \ldots(19{\cdot}76)$$

This is related to the stable distribution of population which immigration can just maintain against the excess of μ over λ.

(ii) When $\mu=0$, *i.e.*, when there are only births and immigration,

$$\phi(z, t)=e^{kt}[1-z(1-e^{-\lambda t})]^{\frac{-k}{\lambda}} \qquad \ldots(19{\cdot}77)$$

which is the *p.g.f.* for the negative binomial distribution.

(*iii*) when $\lambda=0$ *i.e.*, when there is immigration and emigration the latter depending on the size of the population

$$\phi(z, t)=\exp\left[\frac{k}{\mu}(1-e^{-\mu t})(z-1)\right] \quad \ldots(19\cdot78)$$

which is the *p.g. f.* of Poisson process.

When $t\to\infty$, it gives

$$\phi(z)=\exp\left[\frac{k}{\mu}(z-1)\right] \quad \ldots(19\cdot79)$$

(*iv*) When $\lambda=0$, $\mu=0$ *i.e.*, when there is only immigration

$$\phi(z, t)=\exp[kt(z-1)] \quad \ldots(19\cdot80)$$

$$=e^{-kt}\sum_{n=0}^{\infty}\frac{(kt)^n}{n!}$$

$$P_n(t)=e^{-kt}\frac{(kt)^n}{n!} \quad \ldots(19\cdot81)$$

which again gives probability distribution for a Poisson distribution with parameter kt.

Ex. 1. Show that for the case when the initial population size is N_0.

$$\phi(z, t)=\frac{(\lambda-\mu)^{k/\lambda}\{\mu e^{(\lambda-\mu)t}-1)-(\mu e^{(\lambda-\mu)t}-1)z\}^{N_0}}{\left\{(\lambda e^{(\lambda-\mu)t}-\mu)-\lambda(e^{(\lambda-\mu)t}-\lambda)z\right\}^{N_0+\frac{k}{\lambda}}}$$

Deduce (19·71).

Ex. 2. Discuss the limiting case when $k\to 0$.

Ex. 3. Obtain $\overline{N}(t)$ from (19·77), (19·78), (19·79) and (19 88).

Ex. 4. Discuss the model when immigration is independent of the population size and emigration is proportional to the rate of population size. Interpret this model to discuss the colloidal particle suspension or the movements in spermatozoa.

Ex. 5. Find $\overline{N}(t)$ from problem 1.

19·9. Generalised Birth and Death Process. In the present section, we postulate that the probability of going from state E_n to state E_{n+1} in time dt is $\lambda_n dt+0(dt)$ and the probability of going from state E_n to E_{n-1} is $\mu_n dt+0(dt)$, then following the method of earlier sections, we get

$$P_n(t+dt)=P_n(t)[1-(\lambda_{n+1}\mu_n)dt]+P_{n+1}(t)\mu_{n+1}dt+P_{n-1}(t)\lambda_{n-1}dt+0(dt)$$

so that

$$P_n'(t)=-(\lambda_n+\mu_n)P_n(t)+\lambda_{n-1}P_{n-1}(t)+\mu_{n+1}P_{n+1}(t),\ n\geqslant 1 \quad \ldots(19\cdot82)$$

$$P_0'(t)=-\lambda_0P_0(t)+\mu_1P_1(t) \quad \ldots(19\cdot83)$$

If at time zero, the system is in a state E_r, the initial conditions are :

$$P_r(0)=1\ ;\ P_n(0)=0 \text{ for } n\neq r \quad \ldots(19\ 84)$$

$$\frac{e^{P(t)}}{z-1}-\int_0^t \lambda(t)\, e^{P(t)}\, dt=\text{const.}$$

where

$$P(t)=\int_0 \{\mu(t)-\lambda(t)\}dt$$

Deduce $\phi(z, t)$ when $\phi(z, 0)=z^4$.

11. Discuss the particular case when

$$\lambda(t)=a+bt,\ \mu(t)=c+dt.$$

12. Suppose a particle performs unrestricted symmetrical random walk in two dimensions starting from the origin. Let r_n be the distance of the moving particle from the origin after the nth step. Prove that

$$E(r_n^2)=n.$$

13. For the queue with discouragement prove that the queue length a Poisson distribution with parameter λ/μ.

14. Show that the following processes are identically distributed

(*i*) $X(t)=A\cot \omega t+B \sin \omega t$,

(*ii*) $X(t)=R\cos(\omega t+\theta)$,

when A and B are independent normally distributed random variables with means 0 and variances σ^2, θ is uniformly distributed on 0 to 2π and R is Rayleigh distribution with probability density function

$$\frac{r}{\sigma^2}\, e^{-\frac{1}{2}\frac{r^2}{\sigma^2}} \qquad r>0$$

15. Let $X_1(t)$ and $X_2(t)$ be the independent Poisson processes with mean rates λ_1, μ_2 respectively. For example $X_1(t)$ may represent the number of waiting taxis and $X_2(t)$ represents the number of waiting passengers at a taxi stand). Show that if

$$X(t)=X_1(t)-X_2(t),$$

then $X(t)$ is not a Poisson process. Also find $P[Y(t)=X(s)=k]$

For $t>s\geqslant 0$ and $k=0, \pm 1\ \pm 2, \ldots$

[Such stochastic processes are useful in studying double ended queuing systems.]

16. For any Poisson process $X(t)$, show that for $s<t$.

$$P\{N(S)=k|N(t)=n\}={}^nc_k\left(\frac{s}{t}\right)^k\left(1-\frac{s}{t}\right)^{n-k}$$

17. For the stochastic process (*i*) $X(t)=A+Bt+Ct^2$, where A, B, C are independent random variables each with mean 1 and variance 1 (*ii*) $X(t)=A\cos wt+B\sin wt$, in which w is positive constant and A and B are uncorrelated random variables with 0 mean and variance σ^2, find

(*a*) $E\{X(t)\}$, (*b*) cov $\{X(S), X(t)\}$.

18. Particles arrive in groups. For $k=1, 2, \ldots$ suppose that the event that k particles arrive simultaneously occurs in accordance with a Poisson process $N_k(t)$ at a mean rate μ_k per unit time. Assume independence of that process. Let $N(t)$ be the total number of particles that arrive in interval $(0, t)$. Show that

(*i*) $N(t)=N_1+2N_2(t)+\ldots+kN_k(t)+\ldots$

(*ii*) $E(N(t)=\mu_1 t$, where $\mu_1=\sum_{k=1}^{\infty} k\lambda_k$

(*iii*) Var $N(t)=\mu_2 t$ where $\mu_2=\sum_{k=1}^{\infty} k^2\lambda_k$.

APPENDIX I
Some Continuous Distributions

Distribution (1)	*Range of values* (2)	*Density function* (3)	*Mean* (4)	*Variance* (5)	*Mode* (6)
Uniform	(a, b)	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	—
Normal	$(-\infty, \infty)$	$\frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left[-\frac{x-a)^2}{2\sigma^2}\right]$	a	σ^2	a
Logarithmic normal	$(0, \infty)$	$\frac{1}{x\sigma(2\pi)^{\frac{1}{2}}} \exp\left[-\frac{(\log x-a)^2}{2\sigma^2}\right]$	$\exp[a+\frac{1}{2}\sigma^2]$	$\exp[2a+\sigma^2]\ (\exp[\sigma^2]-1)$	$\exp[a-\sigma^2]$
Weibull's	$(0, \infty)$	$acx^{a-1} \exp\left[-cx^a\right] \frac{\left(\Gamma(1+\frac{1}{\alpha})\right)}{a^{1/\alpha}}$	$\frac{\Gamma\left(1+\frac{2}{\alpha}\right)-\Gamma^2\left(1+\frac{1}{\alpha}\right)}{c^{2/a}}$		(for $\alpha>\frac{1}{2}$) $\left(\frac{\alpha-1}{c\alpha}\right)$
Gamma	$(0, \infty)$	$\frac{\beta^a}{\Gamma(a)} x^{a-1} c^{\beta x}$	α/β	α/β^2	$[(\alpha-1)/\beta]$, $\sigma<1$
Cauchy	$(-\infty, \infty)$	$\frac{1}{\pi} \cdot \frac{1}{1+(x/a)^2}$	a	—	a

Appendix I—(*Contd.*)

(1)	(2)	(3)	(4)	(5)	(6)
Exponential	$(0, \infty)$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$	—
Chi-square	$(0, \infty)$	$\frac{z^{\frac{1}{2}(k-1)}}{\Gamma(\frac{1}{2}k)2^{\frac{1}{2}k}}\ e^{-\frac{1}{2}\chi^2}$	k	$2k$	$k-2$
Beta	$(0, 1)$	$\frac{1}{B(a, b)}\ x^{a-1}\ (1-x)^{b-1}$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2\ (a+b+1)}$	$\frac{a-1}{a+b-2}$
Student's	$(-\infty, \infty)$	$\left[\ 2^{\frac{1}{2}(n-1)}\ \ \Gamma\left(\frac{1}{2n}\right)(\pi n)^{\frac{1}{2}}\right]^{-1}$ $\times\left(\ 1+\frac{1}{2}x^2\ \right)^{-\frac{1}{2}(n+1)}$	0	$\frac{1}{\frac{1}{2}n-2}$	0
Fisher's	$(0, \infty)$	$\frac{\Gamma[\frac{1}{2}(n_1+n_2)]}{\Gamma(\frac{1}{2}\ n_1)\ \Gamma(\frac{1}{2}\ n_2)} \cdot \left(\ \frac{n_1}{n_2}\ \right)^{\frac{1}{2}n_1}\ \ x^{\frac{1}{2}n_1-1}$ $\times\left(\ 1+\frac{n_1}{n_2}\ \right)^{-\frac{1}{2}(n_1+n_2)}$	$\frac{n_2}{(n_2-2)}$	$\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$	$\frac{(n_1-2)n_2}{2n_1^2-n_2}$

APPENDIX II

Some Discrete Distributions

Distribution	*Possible Values*	*Probability*	*Mean*	*Variance*
Binomial	0, 1, 2, ..., n	$P_n(m)={}^nC_m p^m q^{n-m}$	np	npq
Hypergeometic	0, 1, ..., min (m, n)	$P_m=\frac{{}^mC_x\ {}^nC_{r-x}}{{}^{m+n}C_r}$	$\frac{mr}{m+n}$	$\frac{mnr(m+n-r)}{(m+n)^2(m+n-1)}$
Poisson	0, 1, 2, ...,	$P_m=\frac{\lambda^m}{m!}\cdot e^{-\lambda}$	λ	λ
Geometric	1, 2, ...	$P_m=pq^{m-1}$	$\frac{1}{p}$	$\frac{q}{p^2}$
Negative binomial	r, $r+1$,.....	$P_m={}^{m-1}C_{n-1}p^r q^{m-r}$	$\frac{r}{p}$	$\frac{qr}{p^2}$

APPENDIX III
Bibliography

1. **Bartlett, M.S.** : *An Introduction to Stochastic Processes.* Cambridge (1955).
2. **Bharucha Reid, A.T.** : *Elements of Theory of Markov Processes and their Application,* McGraw Hill (1960).
3. **Birbaum, Z.W.** : *Introduction to Probability and Mathematical Statistics.* Harper (1962).
4. **Brownlee, K.A.** : *Statistical Theory and Methodology in Science and Engineering.* John Wiley (1960)
5. **Brunk, H.** : *An Introduction to Mathematical Statistics,* Grinn and Co. (1962).
6. **Chernoff, H. and Moser, I.F.** : *Elementary Decision Theory.* John Wiley (1959).
7. **Chung, K.L.** : *Markov Chains with Stationary Transition Probabilities.* Springer (1967).
8. **Clark, C.** : *An Introduction to Statistics.* John Wiley (1953).
9. **Cramer, H.** : *The Elements of Probability Theory and some of its Application.* John Wiley (1955).
10. **Cramer, H.** : *Mathematical Methods of Statistics.* Princeton University Press (1946).
11. **Dixon, W.J. and Stuart, A.** : *Introduction Statistical Analysis.* McGraw Hill (1957)
12. **Doob, J.L.** : *Stochastic Processes.* John Wiley (1953).
13. **Feller., W.** : *An Introduction to Probability Theory and its Application.* John Wiley, Vol. 1, 2nd ed. (1967). Vol. II (1965).
14. **Fisher, R.A.** : *Statistical Methods and Scientific Inference.* Hafner Publishing Co. (1956).
15. **Fisher, R.A.** : *Statistical Methods for Research Workers.* Oliver and Boyd (1925).
16. **Fraser D.A.S.** : *Statistics, an Introduction.* John Wiley (1962).
17. **Goldberg, S.** : *Probability Introduction.* Prentice Hall (1960).
18. **Genedenko, B.V.** : *Theory of Probability.* Chalsea (1962).

19. **Genedenko, B.V. Belyayeo, Y.K. Solovojev, A.D.** : *Mathematical Methods for Reliability Theory.* Academic Press (1969).

20. **Ganurman, V.B.** : *Fundamentals of Probability Theory and Mathematical Statistics*, American-Elsevier (1968).

21. **Graybill, F.A.** : *Linear Statistical Models.* McGraw Hill (1961).

22. **Guenther, W.C.** : *Concepts of Statistical Inference.* McGraw Hill (1965).

23. **Hald, A.** : *Statistical Theory with Engineering Applications.* John Wiley (1952).

24. **Hypel, P.G.** : *Elementary Statistics.* John Wiley (1962).

25. **Hodges, J.L. and Lehman, E.L.** : *Basic Concepts of Probability and Statistics.* Holden Day (1964).

26. **Hogg, R. and A Craig** : *Introduction to Mathematical Statistics.* Mcmillian (1959).

27. **Freund, J.E.** : *Mathematical Statistics.* Prentice-Hall (1954).

28. **Jeffreys, H.** : *Theory of Probability.* Clarendon Press (1948).

29. **Kac, M.** : *Probability and Related Topics in Physical Sciences.* Inter-Science (1959).

30. **Kendell, M.G. and Stuart, A** : *Advanced Theory of Statistics, Vol. I, II, III.* Charles Griffen (1961).

31. **Kemeny, J. G. and Snell J.K.** : *Finite Markov Chains.* Van Nostrand (1960).

32. **Kenney J.F. and E.S. Keeping** : *Mathematics of Statistics.* Van Nostrand (1951), (East-West 1964).

33. **Keeping, E.S.** : *Introduction to Statistical Inference.* Van Nostrand, East West (1964).

34. **Kolmogroff, A.N.** : *Foundations of the Theory of Probability.* Chalsea Publishing Co. (1950).

35. **Kricheberg, K.** : *Probability Theory;* Addison Wesley (1965).

36. **Lamperti, J.** : *Probability.* W.A. Benjamin (1965).

37. **Krutchkoff, R.G.** : *Probability and Statistical Inference.* Gordon and Beach (1970).

38. **Levi, H., Roth L.** : *Elements of Probability.* Clarendon Press (1936).

39. **Lindley, D.V.** : *Introduction to Probability and Statistics.* Camb. Uni. Press (1965).

40. **Lehman, E.L.** : *Testing Statistical Hypotheses.* John Wiley (1959).

41. **Miller, I. and Freund, J.E.** : *Probability and Statistics for Engineers.* Prentice Hall (1965).

42. **Mood, A.M. and Graybull, F.A.** : *Introduction to the Theory of Statistics,* McGraw Hill (1963).

43. **Mosteller, F., Rourke, R.E. and Thomas, G B.** : *Probability and Statistics.* Addison Wesley (1964).

44. **Minroe, M.** : *Theory of Probability,* MaGraw Hill (1951)

45. **Neven, J.** : *Mathematical Foundations of Calculus of Probability.* Holden Day (1965).

46. **Neyman, J.** : *First Course in Probability and Statistics.* Hôlt Rinehart and Winston (1950).

47. **Prazen, E.** : *Modern Probability Theory and its Applications.* John Wiley (1966).

48. **Parzen, E.** : *Stochastic Processes.* Holden Day (19 67).

49. **Rao, C R.** : *Advanced Statistical Methods in Biometrical Research.* John Wiley (1952).

50. **Rao, C.R.** : *Linear Statistical Inference and its Application.* John Wiley (1965).

51. **Rousander, A C.** : *Elementary Principles of Statistics.* Von Nostrand East-West (1965).

52. **Savage, L.J.** : *The Foundation of Statistics.* John Wiley (1954).

53. **Snedecor, G.W.** : *Statistical Method.* Iowa State College Press (1956).

54. **Tucker, W.G.** : *Introduction to Probability and Mathematical Statistics.* Academic Press (1962).

55. **Tucker, H.** : *A Graduate Course in Probability.* Academic Press (1967).

56. **Uspensky, J.** : *I ntroduction to Mathematical Probability* McGraw Hill (1957).

57. **Weatherburn, C.F.** : *Mathematical Statistics,* Cambridge University Press.

58. **Walker, H.M. and Leoi** : *Statistical Inference.* Holt Rineholt and Winston (1953).

59. **Wadsworth, G.P. and Bryan, J.G.** : *Introduction to Probability and Random, Variables.* McGraw Hill.

60. **Renyi, A.** : *Foundations of Probability.* Holden Day (1970).

61. **Weiss, A.** : *Statistical Decision Theory.* McGraw Hill.

62. **Wilkes, S.S.** : *Elementary Statistical Analysis.* Princeton University Press (1948).

63. **Wilkes, S.S.** : *Mathematical Statistics.* Princeton University Press (1947).

64. **Wolf** : *Elements of Probability and Statistics.* McGraw Hill.

65. **Yale, G.U. and Kendell, M.G.** : *Introduction to the Theory of Statistics.* Hafner (1954).

Index

A

Absolute moments, 45
Absorbing barrier, 545
Additive property of
  Chi-square, 378
  Cumulants, 109
  Gamma Variate, 245
  Normal Variate, 224
  Poisson Variate, 118
Alternative hypothesis, 517
Analysis of variance, 438
Arithmetic mean, 13
Array, 6
Association of attributes, 474
Average deviation,
Average sample size, 523
Average sampling number, 523

B

Bartlet test, 472
Bayes theorem, 70
Bernoullian theorem, 190
Bertrand paradox, 184
Beta distribution function, 245
Bias in statistical test, 522
Binomial distribution, 127
Binomial frequency distribution, 128
Birth and death process, 534
Bivariate normal distribution, 36
Bool's inequality, 62
Buffon needle problem, 184

C

Cauchy distribution, 247
Causal distribution, 140
Central limit theorem, 341
Central tendency, 12
Characteristic function, 175
Charlier check, 42
Chebyshev inequality, 184
Chi-square distribution, 369
Composite hypothesis, 517
Compound probability, 67
Conditional expectation, 106
Conditional probability, 67
Conditional variance, 106
Confidence coefficient, 343
Confidence interval, 343
Consistent estimator, 482
Contingency table, 464
Continuous probability distribution, 159
Converge stochastically, 159
Correlation coefficient, 101
Correlation of ranks, 293
Correlation ratio, 295
Covariance, 101
Critical region, 517
Cumulative frequency, 10
Cumulants, 109
Cumulative function, 104
Curve fitting, 272
Curvilinear regression, 241

D

Deciles, 17
Degrees of freedom, 370
Density function, 160
Desion function, 534
Discrete probability distribution, 127
Dispersion, 18
Distribution function, 93
Drift,

E

Efficient estimator, 485
Elastic barrier, 543
Error function, 235
Errors of first and second kind, 518
Equally likely, 53
Ergodic chain, 548
Estimation, estimators, 482
Exclusive cases, 153
Exhaustive, 153
Expectation, 97
Exponential distribution, 171, 249
Erlang family, 249
Event, 53

F

Factorial moments, 45
  generating function, 108
Favourable cases, 53
Fiducial limits, 343
Fisher's z-transformation, 420
Fourier inversion theorem, 176
Frequency distribution, 6

Frequency polygon, 9
F-distribution, 332
Function—loss, 535
  decision, 535
  risk, 535

G

Gamma distribution, variates, 244
Geometric distribution, 141
Geometric mean, 16
Geometric probability,
Goodness of fit, test of, 458
Gram Charlier series, 258

H

Harmonic mean, 16
Hermite polynomial, 255
Histogram, 9
Hypergeometric distribution, 144
Hyperphere, 369
Homoscedastic, 303

I

Incomplete beta function, 246
Incomplete gama function, 245
Independent events, 76
Interclass correlation, 31
Interdecile, 32
Interquartile range, 32
Intraclass correlation, 305
Inversion theorem, 176

J

J-shaped curve, 12

K

Kolwogorov smirnov test, 532
Kurtosis, 45

L

Laplace double exponential,
Large sample,
Law of large numbers, 184
Least square, method of, 271
Leptokurtic, 45
Level of significance, 335
Likelihood ratio test, 521
Linear regression, 286
Lines of regression, 286
Location, 12
Logarithmic normal distribution, 250

M

Marginal distribution, 193
Mark off inequality, 186
Markov chain, 546
Mathematical expectation, 97
Mean deviation, 32
Mean difference, 36
Median, 13
Median test, 531
Method of maximum likelihood, 492
Method of minimum variance, 501
Method of movements, 500
Minimum decision rule, 502
Minimum chi-square, 502
Mode, modal value, 13
Moments, 36
Multiple correlation coefficient, 317
Multinomial distribution, 146
Mutually exclusive, 53

N

Negative binomial distribution, 141
Normal distribution, 226
Normal equation, 269
Normal probability curve, 301
Normal correlation surface, 301
Null hypothesis, 332

O

Ogive, 9
Operating characteristic curve, 517
Ordinates of normal curve, 235

P

Pascal distribution, 121
Parabola of degree, 273
Parameter, 13, 340
Partial correlation, 317
Partition values, 8
Pearson β-and γ-coefficients, 44
Pearson distribution, 251
Percentiles, 16
Platykurtic, 45
Poisson distribution, 136
Power function, 514
Powerful test, 519
Primary subscripts, 318
Probability, 53
Probability density function, 166
Probable error, 237
Probability generating function, 112
Probability differential, 138

Q

Quantiles, 17
Quartile, 13
  distribution of, 412
Quartile deviation, 28

R

Random sampling, 280
Random variable, 93
Random walk, 545
Random variate, 92
Range, 32
  distribution of, 352
Reflecting barrier, 547
Rectangular distribution, 141
Reproductive property, 139
Regression, 286

Residual, 317
Risk function, 535
Root mean square deviation, 32
Run, 528
Run test, 527
Ruin problem, 53

S

Sample space, 71
Sampling, 331
  attribute of, 331
Sampling distribution, 287
  of mean, 341
  of variance, 341
  of correlation coefficient, 415
  of F, 388
  of *t*, 382
  of Z, 420
Schwartz inequality, 282
Scatter diagram, 282
Secondary scripts, 317
Sequatal analysis, 523
Set, 71
Simple trunk, 549
Subset, 71
Sheppard, 41
Sign test, 526
Skew curve, 11
Skewness, 39
Standard deviation, 35
Standard error, 293
Statistic, 3, 340
Statistical control, 464
Statistical decision, 533
Stochastic variates, 82
Sufficient estimator, 488

T

Tests of significance, 6
Tippet, 332
Trial, 53
Transition matrix, 545

U

Uniform distribution, 141
Univariate distribution, 201
U-shaped curve, 12

V

Variance, 35

W

Weilbull distribution, 249
Wilcoxion-Man test, 582

Y

Yates correction, 468
Yule's coefficient of associate, 474
Yule's coefficient of colligate, 474