

Statistics for Textile Engineers

Statistics for Textile Engineers

Prof. J. R. Nagla

WOODHEAD PUBLISHING INDIA PVT LTD

New Delhi • Cambridge • Oxford • Philadelphia

Published by Woodhead Publishing India Pvt. Ltd.
Woodhead Publishing India Pvt. Ltd., 303, Vardaan House, 7/28, Ansari Road,
Daryaganj, New Delhi - 110002, India
www.woodheadpublishingindia.com

Woodhead Publishing Limited, 80 High Street, Sawston, Cambridge,
CB22 3HJ UK

Woodhead Publishing USA 1518 Walnut Street, Suite 1100, Philadelphia
www.woodheadpublishing.com

First published 2014, Woodhead Publishing India Pvt. Ltd.
© Woodhead Publishing India Pvt. Ltd., 2014

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission. Reasonable efforts have been made to publish reliable data and information, but the authors and the publishers cannot assume responsibility for the validity of all materials. Neither the authors nor the publishers, nor anyone else associated with this publication, shall be liable for any loss, damage or liability directly or indirectly caused or alleged to be caused by this book.

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming and recording, or by any information storage or retrieval system, without permission in writing from Woodhead Publishing India Pvt. Ltd. The consent of Woodhead Publishing India Pvt. Ltd. does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific permission must be obtained in writing from Woodhead Publishing India Pvt. Ltd. for such copying.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

Woodhead Publishing India Pvt. Ltd. ISBN: 978-93-80308-37-1
Woodhead Publishing Ltd. ISBN: 978-1-78242-067-5
Woodhead Publishing Ltd. e-ISBN: 978-1-78242-071-2

Prepress by DSM Soft, Chennai
Printed and bound by Thomson Press Pvt. Ltd.

Dedicated to
DKTE and My Family

Contents

<i>Preface</i>	<i>xiii</i>
1 Introduction and need of statistics	1
1.1 Introduction	1
1.2 Nature of textile industry today	1
1.3 Need for SQC techniques	2
2 Statistics and basic concepts	5
2.1 Introduction of statistics	5
2.2 Methods of data collection	7
3 Classification and graphical representations	9
3.1 Introduction	9
3.2 Frequency distribution table	10
3.3 Cumulative frequency distribution table	16
3.4 Graphical representations of the frequency distribution	18
3.5 Graphical representations of the cumulative frequency distribution (Ogive Curve)	20
3.6 Exercise	22
4 Measures of central tendency	25
4.1 Introduction of the central tendency	25
4.2 Measure of central tendency	25
4.3 Arithmetic mean	25

4.4	Computation of AM for the frequency distribution	26
4.5	Median	31
4.6	Computation of median	31
4.7	Graphical determination of median	32
4.8	Mode	35
4.9	Determination of mode	35
4.10	Determination of mode graphically	36
4.11	Exercise	38
5	Partition values	40
5.1	Introduction	40
5.2	Quartiles	40
5.3	Deciles	45
5.4	Percentiles	50
5.5	Exercise	54
6	Measures of dispersion	56
6.1	Introduction of the dispersion	56
6.2	Range	56
6.3	Quartile deviation	58
6.4	Mean deviation	61
6.5	Standard deviation	65
6.6	Relative measures of dispersion	72
6.7	Exercise	76
7	Skewness and kurtosis	80
7.1	Introduction	80
7.2	Skewness	80
7.3	Symmetric frequency distribution	80
7.4	Skewed frequency distribution	81
7.5	Measures of skewness	82

Contents	ix	
7.6	Interpretation of the coefficient of skewness	83
7.7	Kurtosis	85
7.8	Measures of kurtosis	86
7.9	Exercise	88
8	Correlation and regression	90
8.1	Introduction	90
8.2	Bivariate data	90
8.3	Correlation analysis	90
8.4	Coefficient of determination	93
8.5	Spearman's rank correlation coefficient	95
8.6	Regression analysis	100
8.7	Exercise	108
9	Multivariate analysis	111
9.1	Introduction	111
9.2	Multivariate data	111
9.3	Multiple correlation analysis	111
9.4	Multiple regression analysis	114
9.5	Exercise	117
10	Probability	119
10.1	Introduction	119
10.2	Basic concepts	119
10.3	Definition of the probability (classical approach)	122
10.4	Laws of the probability	122
10.5	Results of permutation and combination	123
10.6	Exercise	129
11	Probability distributions	131
11.1	Basic concepts	131
11.2	Probability distribution of a random variable	132

11.3	Some properties of the random variable and its probability distribution	136
11.4	Some standard probability distributions	138
11.5	Exercise	138
12	Standard discrete probability distributions	140
12.1	Binomial probability distribution	140
12.2.	Poisson probability distribution	143
12.3	Poisson approximation to the binomial distribution	146
12.4	Fitting of binomial and Poisson probability distributions	147
12.5	Exercise	151
13	Standard continuous probability distributions	154
13.1	Normal problem distribution	154
13.2	Standard normal variable and standard normal probability distribution	157
13.3	Chi-square probability distribution (χ^2 distribution)	162
13.4	Student's <i>t</i> -probability distribution	164
13.5	<i>F</i> -probability distribution	166
13.6	Exercise	168
14	Testing of hypothesis	170
14.1	Introduction	170
14.2	Large sample tests (Z-tests)	172
14.3	Small sample tests	185
14.4	Exercise	216
15	Estimation	220
15.1	Introduction	220
15.2	Point estimation	220
15.3	Interval estimation (confidence interval)	222
15.4	Exercise	229

Contents	xi
16 Analysis of variance	231
16.1 Introduction	231
16.2 One-way analysis of variance	232
16.3 Two-way analysis of variance	238
16.4 Exercise	249
17 Design of experiments	253
17.1 Introduction	253
17.2 Completely randomized design	254
17.3 Randomized block design	255
17.4 Latin square design (LSD)	257
17.5 Factorial experiments	262
17.6 Exercise	274
18 Statistical quality control	278
18.1 Introduction	278
18.2 Process control	279
18.3 Control chart	279
18.4 Interpretation of control chart	280
18.5 Specification limits	281
18.6 \bar{X} chart	281
18.7 R-Chart	282
18.8 np -Chart	284
18.9 p -Chart	285
18.10 C-chart	288
18.11 Lot control	290
18.12 Some basic concepts related to rectifying single sampling plan	291
18.13 Exercise	294

Area under standard normal curve from $Z = 0$ to $Z = z$	296
$t_{n,\infty/2}$ values for t-distribution	298
$\chi^{2n,\infty}$ values for χ^2-distribution	300
$F_{m,n, 0.05}$ values for F-distribution	302
$F_{m,n, 0.01}$ values for F-distribution	304
Statistical constants for control charts	306
Index	308

Preface

The inspiration of writing this book is the result of great need from textile engineering students, textile engineers and textile professionals working in the textile industries for the book on statistics in simple and easy language, which will guide and help them to use different statistical methods in decision making.

I hope this book will meet all the needs of textile professionals and textile industries in their decision making as it discusses everything related to statistics right from basic data collection to up to design of experiments with suitable illustrations. Though this book is written for textile-related people only, it is also useful for engineers or researchers of any discipline.

The statistical tables of the book are created using Microsoft Excel functions.

I am thankful to Shri. K. Venkatrayan (BITRA Mumbai) for his guidance in writing this book. I am very much thankful to Professor S. D. Mahajan (TEI Ichalkaranji) for writing the introductory chapter, also editing and proof reading the book. I am thankful to my student Raghav Bhala for his help in checking calculations. I am very much thankful to our principal Dr. P. V. Kadole and the management of DKTE Society for their support and help in completing this book. Finally, I am thankful to those who have inspired and helped me directly or indirectly in writing this book.

I will be happy if the readers point out mistakes and give feedback about the book.

Introduction and need of statistics

1.1 Introduction

From many angles textile industry holds unique position in the Indian industries and economy. It is next to farming in providing employment (direct and indirect) to the people. Share of textile industry in industrial products is almost fourteen percent. Thirty to thirty-five percent of our foreign exchange earning is thorough textile products. It is the only industry linking all states and union territories. It will be appropriate to say that our country our vast country is tied together by the threads of textile industry.

Thus, well being of Indian economy depends to a large extent upon well being of textile industry. And, if judiciously used statistical quality control (SQC) can be of good help in maintaining well being of Indian Textile Industry.

1.2 Nature of textile industry today

Cloth was, is and will be the primary need of the civilized mankind – for protection from heat, cold, rain etc. and also for decoration of human body. Art of converting natural fibers like cotton, silk, wool, jute and flax into cloth was known to man from centuries. Before industrial revolution in Europe textile industry was more in the form of decentralized cottage industry and ‘Art’ element was playing major role.

But, industrial revolution totally changed the nature of textile industry all over the world. Concept of ‘Mass Production’ took over the charge and to that extent ‘Art’ element receded. First major change was establishment of big capacity mills – change from decentralized sector to centralized, organized sector. Second change was drastic increase in the production rates because of use of high-speed machines. This necessarily put forward the challenge of maintaining reasonable consistency in quality of mass production. In addition with the developments in science and technology many new areas got introduced in textile industry with stricter demands on quality. Manmade textile industry, development of garment sector and fast developing field of

technical textiles are some examples of new areas. All these developments have put enormous pressure on textile industry for achieving quality standards. Today's textile industry has really become divergent and all serious efforts are being made to satisfy divergent quality demands from all these sectors. One thing can be certainly said – nature of today's textile industry was beyond imagination even few decades back.

Today textile industry can be divided in the following well-defined sectors:

1. Sectors processing and preparing natural fibres for further processing [Ginning and pressing of cotton, breeding of cocoons and reeling in case of silk are the examples of this sector.]
2. Units manufacturing manmade fibres like viscose, nylon, polyester, polypropylene, acrylic etc.
3. Spinning mills converting fibres into yarn.
4. Sizing units converting coned yarn into warp beams for weaving.
5. Weaving units manufacturing cloth. In India this sector is further divided in four sub sectors – hand looms, non-automatic power looms, automatic power looms and shuttleless looms.
6. Knitting units manufacturing hosiery products.
7. Latest sector in fabric forming is of non wovens. This is a high tech industry with high growth potential.
8. Process houses for chemical treatments on cloth to make it more attractive and suitable for use. Bleaching, dyeing, printing and finishing are the common chemical treatments carried out.
9. Units manufacturing ready-made garments from the finished cloth.
10. A relatively new sector is of manufacturing textile composites for industrial applications.
11. Vast and fast growing sector of technical textiles.

1.3 Need for SQC techniques

Out of the above sectors age old sectors are ginning/pressing, silk breeding, reeling, spinning, sizing, weaving, knitting and chemical processing. All these sectors are having some common features:

1. Variability in raw materials like fibres, chemicals, water etc. All natural fibres are characterized by inherent variation in their properties. For example in case of cotton the basic properties like

length, strength, fineness, colour varies considerably even for same variety, station, lot, bale. Variations in raw materials become a major source of variation in the finished product at each stage. Even chemical process houses feel the impact of variability in fibres.

2. All these industries are labour intensive with low level of automation. As such finished product quality depends to a large extent on the work practices followed by the workers.
3. Many of these industrial units work on three shifts basis. It is almost impossible to have similar work practices from all workers from all shifts. Thus the system brings along with it the variability from person to person.
4. In majority of cases processing of material is done, not in a continuous fashion, but in a batch wise fashion. One lot is divided in number of batches depending on the capacities of available machines. For example, if lot requirement is of 30,000 metres it will be simultaneously woven on number of looms and even chemical processing like dyeing is done in 30 batches of 1000 metres each. One cannot expect exactly same quality of finished material when processing is done in number of batches. Thus, batch wise processing is one reason for variations.
5. Another major reason for variability in yarn quality is very high number of production centers. Every ring frame spindle is a full-fledged production center having its own check points. And so, yarn produced on one spindle will differ in properties from yarn spun on another spindle. Similarly, same cloth variety woven on two looms will show some variability in properties. In the existing processing conditions such variations are unavoidable.
6. In textiles number of inputs is too large till the fibre gets converted into readymade garments. The big list includes fibers, dyes, chemicals, water, steam, fuels, electricity, machines, workers, maintenance inventory etc. Their role in controlling finished product quality is many times quite complex. Perfect quality control to all these inputs is next to impossible.

Textile engineers are generally interested in studying such variations occurring from material to material, test to test, sample to sample, machine to machine, time to time and place to place. Before any type of study textile engineers generally have following questions.

1. How many tests are to be carried out for getting the desired results?
2. How to analyze the results or the data collected for the purpose of the study?
3. How to interpret the results of analysis?

Answers of all above questions can be obtained with the help of the subject ‘Statistics’. Thus, for studying the variation in the data and interpret them the textile engineer must know the theory and different methods of the subject ‘Statistics’.

All eighteen chapters of this book discuss various statistical methods and techniques which are useful for study and analysis of textile data.

Chapter one is about the subject textile technology and need of statistics in textiles.

Chapter two discusses about the subject Statistics and the basic terminology used in statistics.

Chapter three discusses about classification and graphical representation of the data.

Chapter four discusses about the measures of central tendency.

Chapter five discusses about the partition values and its use.

Chapter six discusses about measures of dispersion and their importance.

Chapter seven discusses about the skewness and kurtosis of the frequency distribution and their interpretations.

Chapter eight discusses about the study of the bivariate data using correlation and regression.

Chapter nine discusses about the study of the multivariate data using multiple and partial correlation and the multiple regression.

Chapter ten discusses about the probability theory.

Chapter eleven discusses about the probability distribution of the random variables.

Chapter twelve discusses about some standard discrete distributions.

Chapter thirteen discusses about some standard continuous distributions.

Chapter fourteen discusses about the testing of hypothesis.

Chapter fifteen discusses about the estimation.

Chapter sixteen discusses about analysis of variance (ANOVA).

Chapter seventeen discusses about design of experiments.

Chapter eighteen discusses about statistical quality control (SQC).

Statistics and basic concepts

2.1 Introduction of statistics

In previous chapter we have discussed the need of statistics in textile technology. Obviously, the first question which arises in mind is what is statistics and how is it defined?

Statistics is the subject, which deals with different methods of collection of the data or the numbers (according to the objective of study), methods of classification and summarization of the data, methods of analyzing the data and finally drawing the conclusions on the basis of analysis of the data.

Population

The data or the numbers in Statistics are generally collected from a well-defined group of individuals; this group of individuals is called as the population. Thus, population is the collection of members or the individuals from which the data or the numbers are collected for the study.

For example, population may be the collection of ring bobbins produced during a shift of production or it may be collection of rolls of fabrics produced in a textile mill or it may be collection of garments produced by a garment factory.

Individual

The single member of the population from which the data is actually collected for the purpose of the study is called as an individual.

For example, an individual may be a single cotton fiber or a single ring bobbin or a single piece of fabric or a single piece of garment.

Characteristic

Any physical property (of interest) possessed by the individuals of the population, about which the numbers are collected is called as the characteristic.

Quantitative type characteristic and Qualitative type characteristic are the two different types of characteristics which are defined as follows:

Quantitative type characteristic

The characteristic, which is measured in any unit of measurement, is called as the quantitative type characteristic.

For example, quantitative type characteristics may be length or count or strength or weight etc.

Qualitative type characteristic

The characteristic, which cannot be measured in any unit of measurement, is called as the qualitative type characteristic.

For example, qualitative type characteristics may be external appearance or the luster of the fabric, or the color or the brightness of the garment.

Attribute

The qualitative type characteristic, which changes from individual to individual is called as an attribute and is denoted by the notations A, B, C, etc.

For example, attribute may be external appearance of the fabric or the choice of color of the garment etc.

Variable

The quantitative type characteristic, whose value changes from individual to individual, is called as the variable. The variables are always denoted by notations X, Y, Z, etc.

For example, the variables may be staple length of the fiber or the count of the yarn or the strength of the fabric etc.

Discrete Variable and Continuous Variable are two different types of variables, which are defined as follows:

Discrete variable

The variable whose possible values are finite or countably infinite that is, the variable which is counted as 0, 1, 2 etc. is called as the discrete variable. For example, the discrete variable may be the number of defective needles in a pack of 10 needles or the number of defective garments in a sample of 10 garments or the number of accidents in a textile mill during a month etc.

Continuous variable

The variable whose possible values are uncountably infinite or the variable, which is measured in any unit of measurement, is called as the continuous variable. For example, the variable may be the staple length of the fiber or the count of the yarn or the strength of the fabric etc.

2.2 Methods of data collection

We have seen that, in Statistics the population under study is the source of the data. This population may be finite (small) or infinite (large). If the size of the population is small then it is easy to collect information from each and every member of the population, but if the size of the population is large then because of the limitations like time, labor, money etc., it is not possible to collect the observations or information from each and every member of the population. In such cases a subgroup of the population is selected and the data is collected from the members of this subgroup. Such subgroup in Statistics is called as the sample. The sample selected by giving equal chance of selection to each and every member of the population is called as the random sample. Such sample is free from biasness and partiality. Such sample is a proper representative of the population. Hence, the accuracy of the conclusions is also more. Census Survey and Sample Survey are two different methods of data collection, which are defined as follows:

Census survey

The procedure of the collection of the data or the observations from each and every member of the population is called as the census survey. For example, government of India conducts census survey after the interval of every 10 years.

Sample survey

The procedure of the collection of the data or the observations from the members of the sample, selected from the population is called as the sample survey.

The data collected by the census or the sample survey is called as the raw statistical data, which can be defined as follows:

Raw statistical data

The collection of the observations recorded in the order, in which they are collected, is called as the raw statistical data. It is always represented in the

mathematical set notation. For example, if the variable X represents the count of the yarn, then raw statistical data is represented by the set

$$\{38.0, 38.5, 37.2, \dots, 37.6\}.$$

There are following three types of the raw statistical data.

Univariate data

The data related to only one variable is called as the univariate data. Thus, if ' X ' is the variable of the data, then in general the univariate data of ' n ' observations is given as follows:

$$\{x_1, x_2, \dots, x_n\}$$

Bivariate data

The data related to the two different variables is called as the bivariate data. Thus, if X and Y are the two variables of the data, then in general the bivariate data of ' n ' observations is given as follows:

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

Multivariate data

The data related to the three or more variables is called as the multivariate data. Thus if X_1, X_2, \dots, X_k are k different variables of the data, then in general the multivariate data of ' n ' observations is given as follows:

$$\{(x_{11}, x_{21}, \dots, x_{k1}), (x_{12}, x_{22}, \dots, x_{k2}), \dots, (x_{1n}, x_{2n}, \dots, x_{kn})\}$$

In this chapter we have seen that, in statistics data collected for the study is always in the raw format. To study some thing from it and to draw some conclusion from it, the data is must be converted into the simple and understandable format. The procedure of how to convert the data in the simple and the understandable format is discussed in the next chapter.

Classification and graphical representations

3.1 Introduction

In previous chapter we have seen, what statistics is, what is importance of statistics in textile industry and how to collect the data (by census or by sample survey) for the study of different properties of fibers, yarns, fabrics and garments statistically.

After collecting the data, the person who has collected the data may have several questions in his mind. For example after collecting hundred results of count tests the textile engineer may be interested in knowing how many count tests show count less than or more than certain value or how many count tests show count in between some values. His questions can be easily answered if he can represents the data in the suitable form, so that it is self explanatory, easy to understand and easy to answer his questions about the data and hence about the population from which the data is collected. Classification, summarization and graphical representation of the data are different methods of representing the data in suitable and self explanatory forms. Thus, classification and summarization is the first step towards studying or understanding the data. It means sorting the data according to the similarity and then representing it in a tabular format. There are following two types of classification.

Classification according to attribute

When the data collected is classified according to the attributes possessed by the individuals and represented in table form then it is called as classification according to the attributes. This procedure of table formation is also called as the ‘Tabulation.’

For example, the data of 1800 workers of a textile mill is classified in the table form according to sex, religion and education and is shown in Table 3.1.

Table 3.1

Sex	Hindu				Muslim				Others			
	Below	10th	12th	Above	Below	10th	12th	Above	Below	10th	12th	Above
Total												
Male	150	142	115	407	200	107	54	361	182	98	62	342
Female	152	135	110	397	46	26	12	84	136	50	23	209
	302	277	225	804	246	133	66	445	318	148	85	551

Similarly, the production data of 2000 cones of yarn are classified according to blend, fiber and count and represented in the form of Table 3.2.

Table 3.2

Count	PV				PC				Others			
	40 × 60	50 × 50	60 × 40	Total	40 × 60	50 × 50	60 × 40	Total	40 × 60	50 × 50	60 × 40	Total
30's	200	140	215	555	207	100	56	363	180	100	65	345
40's	172	160	117	449	40	20	10	70	130	65	23	218
Total	372	300	332	1004	247	120	66	433	433	165	88	563

Classification according to the variables

When the data collected is classified according to the possible values of the variables and represented in table form then it is called as the classification according to the variables. Such table is also called as the frequency distribution table and is discussed below in more details.

3.2 Frequency distribution table

The tabular representation of the classified data according to the values of the variables is called as the frequency distribution table. There are following two types of frequency distribution table which are discussed below.

Ungrouped frequency distribution table

When the variable under study is discrete and has finite possible values then the data of this variable is classified according to individual possible values and is represented in the table from, such representation in statistics is called as the ungrouped frequency distribution table. It is sometimes also called as the discrete frequency distribution.

For example, the data of 105 observations of the textile firm related to ‘Number of workers absent’ recorded for 105 days is classified and represented in Table 3.3.

Table 3.3

Number of workers absent	Number of days
0	10
1	15
2	35
3	30
4	15
Total	105

From Table 3.2 it is clear that for 10 days number of absent workers was ‘0’. Here 10 is called as the frequency of the observation ‘0’. Similarly 15, 35, 30 and 15 are the frequencies of the observations 1, 2, 3 and 4 respectively. Thus number of repetitions of the observation in the data is called its frequency.

In general the ungrouped frequency distribution table can be represented as per Table 3.4.

Table 3.4

X	Frequency
x_1	f_1
x_2	f_2
...	...
x_k	f_k
Total	$N = \sum f_i$

Where, X is the variable and x_1, x_2, \dots, x_k are the possible values of the variable X . Also f_i is the frequency of the observation x_i and N represents the total number of observations.

Construction of ungrouped frequency distribution table

This type of frequency distribution table is constructed, if the possible values of the variable are finite or limited. Following steps are followed for the construction of ungrouped frequency distribution table.

1. Identify the variable and its possible values.
2. Write all the possible values in the first column of the table.
3. Make the tally bars in the second column of the table.
4. Count tally bars and write as the frequency in the third column of the table.

Example 3.1

Following are the observations of ‘number of major weaving defects’ obtained by inspecting 40 pieces of the fabric.

{0,1,5,3,4,3,0,0,1,1,1,3,2,1,2,2,2,3,3,0,0,0,1,1,1,2,2,2,3,4,5,2,5,4,4,1,1,1,2,2}

Construct the frequency distribution table for above data and represents in the table form.

Solution

Suppose that, ‘number of major weaving defects in a piece of fabric’ is represented by the variable X . The ungrouped frequency distribution table for the data of this variable X can be constructed by writing all possible values and making tally bars as shown in Table 3.5.

For marking tally bars start with first observation and continue up to last observation. The resultant Table 3.5 is the ungrouped frequency distribution for the given data.

Table 3.5

X	Tally bars	Number of fabric pieces (frequency)
0	I	6
1	I	11
2		10
3	I	6
4		4
5		3
Total		40

Grouped frequency distribution table

When the variable under study is discrete/continuous with infinite (large) possible values then the data of variable is classified by making groups of the possible values of the variable and is represented in the table from, such representation in statistics is called as the grouped frequency distribution table. It is sometimes also called as the discontinuous/continuous frequency distribution. Here the groups formed are also called as the class intervals which can be inclusive (discontinuous) or exclusive (continuous). The frequency distribution

having classes of inclusive type is called as discontinuous where as frequency distribution having exclusive type classes is called as continuous frequency distribution. The ends of inclusive class intervals are called class limits and are included in that class interval that is observations 10 and 19 are included in the class interval 10–19. Where as, the ends of exclusive class intervals are called class boundaries and upper end is not included in that class interval that is observation 20 is not included in the class interval 10–20. Here number of observations belonging to the class interval is called as the class frequency.

For example Table 3.6 is a grouped frequency distribution with inclusive classes and Table 3.7 is a grouped frequency distribution with exclusive classes.

Table 3.6

Class interval	Frequency (no. of observations)
60.0–62.9	7
63.0–65.9	15
66.0–68.9	6
69.0–71.9	2
Total	30

Table 3.7

Class interval	Frequency (no. of observations)
60.0–63.0	7
63.0–66.0	15
66.0–69.0	6
69.0–72.0	2
Total	30

In general the grouped frequency distribution table with inclusive type of class intervals can be written in the form of Table 3.8 and the grouped frequency distribution table with exclusive type of class intervals can be written in the form of Table 3.9.

Table 3.8

Class interval	Class frequency
L_1-U_1	f_1
L_2-U_2	f_2
....	...
L_k-U_k	f_k
Total	$N = \sum f_i$

Where, L_i – is lower limit, U_i is the upper limit and f_i – is the frequency of the class interval $L_i - U_i$

Table 3.9

Class interval	Class frequency
$L_1 - L_2$	f_1
$L_2 - L_3$	f_2
....	...
$L_k - L_{k+1}$	f_k
Total	$N = \sum f_i$

Where, L_i – is lower boundary and L_{i+1} is the upper boundary the class interval $L_i - L_{i+1}$. Also f_i – is the frequency of the class interval $L_i - L_{i+1}$.

Note that, the inclusive type of class intervals can be converted into exclusive type of class intervals by first calculating CF (Correction Factor), then subtracting it from all lower limits and adding it to all upper limits. Where, CF is defined as follows:

$$CF = \frac{\text{lower limit of second class} - \text{Upper limit of first class}}{2}$$

Actual conversion is shown with illustration in Example 3.2.

Construction of grouped frequency distribution table

This type of frequency distribution table is constructed, if the possible values of the variable are very large. The procedure of construction of the grouped frequency distribution is as follows:

1. Identify the variable and its possible values.
2. Make tally bars in the second column of the table.
3. Count tally bars and write as the frequency in the third column of the table.

The grouped frequency distribution can be constructed with inclusive type class intervals or with exclusive type class intervals. There are no hard and fast rules of formation of the groups but, following rules are generally followed for formation of the groups or the class intervals.

1. The number of groups or class intervals should not be too large or too small.
2. As far as possible the size of each class interval must be same.
3. The groups formed should be non-overlapping.

Example 3.2

Following are the results of strength tests carried out on 40 pieces of a fabric.

60.0, 71.2, 62.4, 65.0, 62.8, 64.2, 64.0, 68.0, 63.5, 64.2, 68.2, 65.2, 61.8, 70.0, 60.8, 61.2, 67.5, 65.2, 65.0, 68.4, 62.8, 64.2, 64.0, 68.0, 63.5, 64.2, 68.2, 65.2, 61.8, 70.0, 67.6, 64.4, 64.0, 64.0, 65.0, 65.0, 64.8, 65.6, 66.0, 62.0

1. Construct the inclusive type of grouped frequency distribution table by taking the class intervals as 60.0–62.9, 63.0–65.9 and so on.
2. Convert the inclusive type classes into exclusive type classes.

Solution:

Here, the variable X represents ‘strength of the fabric piece’ and the grouped frequency distribution constructed for above data is given in Table 3.10.

Table 3.10

Strength (class interval)	Tally bars	Frequency (number of observations)
60.0–62.9	IIII III	9
63.0–65.9	IIII IIID IIID	20
66.0–68.9	IIII III	8
69.0–71.9	III	3
Total		40

For converting into exclusive type class interval,

$$CF = \frac{\text{Lower limit of second class} - \text{Upper limit of first class}}{2}$$

$$CF = \frac{63.0 - 62.9}{2} \\ = 0.05$$

Table 3.11 is the table with exclusive type classes and is obtained by subtracting CF from lower limits and adding to the upper limits.

Table 3.11

Strength (class interval)	Tally bars	Frequency (number of observations)
59.95–62.95	III III	9
62.95–65.95	III III III III	20
65.95–68.95	III III	8
68.95–71.95	III	3
Total		40

3.3 Cumulative frequency distribution table

The frequency distribution table obtained by adding or cumulating the frequencies, is called as the cumulative frequency distribution table. According to the way of addition of the frequencies there are following two types of cumulative frequency distribution.

Less than type cumulative frequency distribution table

The cumulative frequency distribution table of less than type is obtained by adding the frequencies from the top to the bottom. Such frequencies are called as the less than type cumulative frequencies and are denoted by the notation $CF(L)$. How to find the cumulative frequencies of less than type is illustrated in Table 3.12 and Table 3.13. Table 3.14 is another representation of Table 3.13.

Table 3.12

Number of weaving defects X	Number of pieces frequency	$CF(L)$
0	6	6
1	11	$6 + 11 = 17$
2	10	$17 + 10 = 27$
3	6	$27 + 6 = 33$
4	4	$33 + 4 = 37$
5	3	$37 + 3 = 40$
Total	40	

Table 3.13

Strength X	Number of samples frequency	$CF(L)$
60.0–62.5	6	6
62.5–65.0	9	$6 + 9 = 15$
65.0–67.5	8	$15 + 8 = 23$
67.5–70.0	5	$23 + 5 = 28$
70.0–72.5	2	$28 + 2 = 30$
Total	30	

Table 3.14

Strength	Frequency
Less than 62.5	6
Less than 65.0	15
Less than 67.5	23
Less than 70.0	28
Less than 72.5	30

More than type cumulative frequency distribution table

The cumulative frequency distribution table of more than type is obtained by adding the frequencies from the bottom to the top. Such frequencies are called as the more than type cumulative frequencies and are denoted by the notation $CF(M)$. Finding the cumulative frequencies of more than type is illustrated in Table 3.15 and Table 3.16. Table 3.17 is another representation of Table number 3.16.

Table 3.15

Number of weaving defects X	Number of pieces frequency	$CF(M)$
0	6	$34 + 6 = 40$
1	11	$23 + 11 = 34$
2	10	$13 + 10 = 23$
3	6	$7 + 6 = 13$
4	4	$3 + 4 = 7$
5	3	3
Total	40	

Table 3.16

Strength X	Number of samples frequency	$CF(M)$
60.0–62.5	6	30
62.5–65.0	9	$15 + 9 = 24$
65.0–67.5	8	$7 + 8 = 15$
67.5–70.0	5	$2 + 5 = 7$
70.0–72.5	2	2
Total	30	

Table 3.17

Strength	Frequency
More than & equal to 60.0	30
More than & equal to 62.5	24
More than & equal to 65.0	15
More than & equal to 67.5	7
More than & equal to 70.0	2

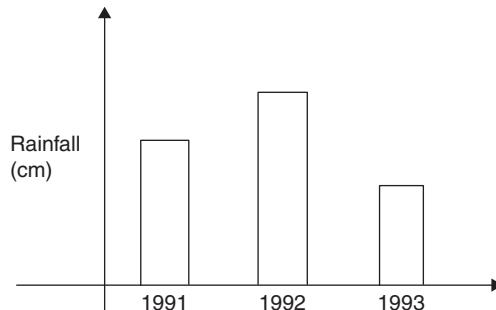
3.4 Graphical representations of the frequency distribution

Some times in statistics the frequency distribution may be represented in the graphical form because representing the frequency distribution in the graphical or the pictorial form makes it easy to understand, easy to interpret. There are following different types of graphical representations of the frequency distribution.

1. Bar diagram
2. Histogram
3. Frequency polygon
4. Frequency curve

Bar diagram

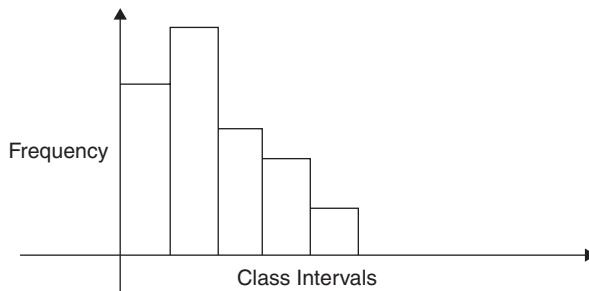
The graphical representation of an ungrouped frequency distribution is called as the bar diagram. Figure 3.1 is the typical example of the bar diagram.



3.1

Histogram

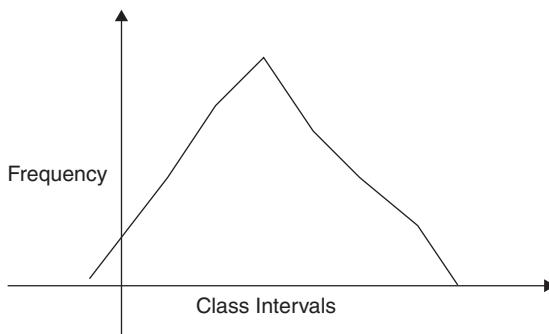
The graphical representation of the grouped frequency distribution with exclusive type of classes (If the classes are inclusive, they are converted into exclusive) is called as the Histogram. It is drawn by erecting vertical bars over the class intervals as the base. The height of the bar is taken according to the frequency of the class interval. Figure 3.2 is the typical example of the histogram.



3.2

Frequency polygon

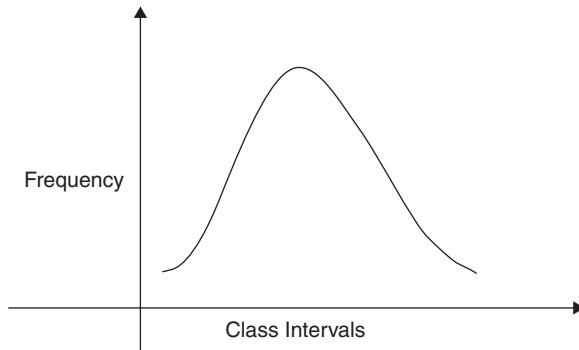
This is a graphical representation of the grouped frequency distribution with exclusive type of classes. This is obtained by plotting the points (x_i, f_i) and then joining them in order by straight lines, where, x_i and f_i are the mid point and the frequency of the i^{th} class interval. Figure 3.3 is the typical example of the frequency polygon.



3.3

Frequency curve

This is also a graphical representation of the grouped frequency distribution with exclusive type of classes. This is obtained by plotting the points (x_i, f_i) and then joining them in order by smooth curves, where, x_i and f_i are the mid point and the frequency of the i^{th} class interval. Figure 3.4 is the typical example of the frequency curve.



3.4

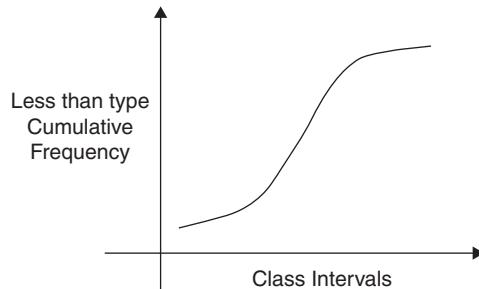
3.5 Graphical representations of the cumulative frequency distribution (Ogive Curve)

Some times in statistics the cumulative frequency distribution is also represented in the graphical form and the graphical representation of the cumulative frequency distribution is called as the ogive curve. There are two

different types of ogive curves according to the type of cumulative frequency distribution.

Less than type ogive curve

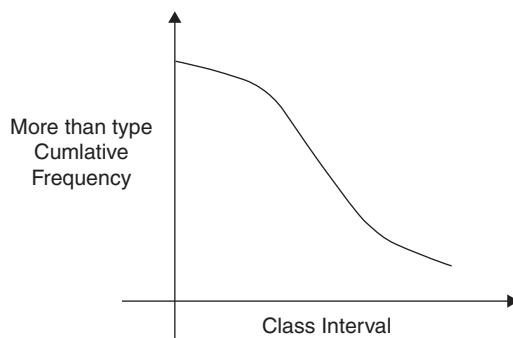
The graphical representation of the less than type of cumulative frequency distribution is called as the less than type ogive curve. This is obtained by plotting the points (*upper boundary, CF(L)*) and then joining them in order by smooth curve. Figure 3.5 is the typical example of the Less than type ogive curve.



3.5

More than type ogive curve

The graphical representation of the more than type of cumulative frequency distribution is called as the more than type ogive curve. This is obtained by plotting the points (*lower boundary, CF(M)*) and then joining them in order by smooth curve. Figure 3.6 is the typical example of the more than type ogive curve.



3.6

In this chapter we have seen, how to classify, summarize and graphically represents the summarized data. Obvious question after this is what next? After classification, next is the study of central tendency of the data which is discussed in the next chapter.

3.6 Exercise

1. Construct the frequency distribution for the following data of number of weaving defects observed in the 30 pieces of fabric.

4, 0, 1, 5, 2, 6, 0, 4, 4, 1, 5, 4, 5, 4, 3

0, 2, 3, 3, 2, 6, 4, 1, 2, 1, 3, 4, 3, 2, 3

2. Construct the ungrouped frequency distribution for the following data of number of end breaks observed on a ring frame for the 30 days.

10, 8, 5, 10, 7, 7, 11, 8, 10, 12, 11, 12, 6, 6, 7

10, 10, 12, 10, 10, 6, 7, 8, 10, 8, 8, 12, 11, 10, 10

3. Following are the figures of number of workers absent in a textile mill every day:

2, 3, 1, 1, 2, 3, 2, 0, 0, 1, 2, 5, 0, 0,

2, 1, 1, 4, 3, 2, 0, 2, 1, 2, 2, 4, 3, 3, 3, 4

Construct the ungrouped frequency distribution for the above data.

4. Following are the results of 'hairiness index' obtained from 20 tests:

6.72, 6.52, 6.00, 6.85, 6.08, 6.64, 6.88, 6.91, 6.35, 6.52

6.26, 7.20, 7.12, 6.36, 6.82, 6.53, 6.78, 6.75, 6.49, 6.62

Construct the grouped frequency distribution by taking classes 6.00–6.20, 6.20–6.40 and so on.

5. Following are the number of accidents per day observed in a textile-mill during a month of 30 days:

4, 3, 3, 3, 4, 2, 2, 1, 2, 0, 2, 3, 4, 1, 1, 2, 0,

0, 5, 2, 1, 0, 0, 2, 3, 2, 1, 1, 3, 2.

Construct an ungrouped frequency distribution.

6. Thirty pieces of fabric were observed for the number of defects and following results were obtained.

1, 0, 0, 2, 3, 2, 1, 1, 3, 2, 4, 3, 4, 1, 1

2, 0, 0, 5, 2, 3, 3, 4, 2, 2, 1, 2, 0, 2, 3

Construct an ungrouped frequency distribution.

7. Construct the frequency distribution for the following results of the daily sale (in 1000 Rs.) in a textile firm recorded for a month.

21, 21, 21, 22, 26, 20, 20, 20, 21, 21, 20, 20, 22, 20, 21

23, 21, 22, 21, 25, 18, 24, 25, 19, 25, 18, 26, 23, 18, 24

8. Thirty linear density tests made on the yarn have shown following results.

14.8, 14.2, 13.8, 13.5, 14.0, 14.2, 14.3, 14.6, 13.9, 14.0,

14.1, 13.2, 13.0, 14.2, 13.5, 13.0, 12.8, 13.9, 14.8, 15.0,

12.8, 13.4, 13.2, 14.0, 13.8, 13.9, 14.0, 14.0, 13.9, 14.8

Construct the grouped frequency distribution by taking class intervals 12.6–13.0; 13.1–13.5 and so on.

9. Following is the distribution of fabric samples according to the strength of fabric:

Strength	150–155	155–160	160–165	165–170	170–175
No. of samples	8	12	17	10	7

Draw the histogram for the above data and interpret it.

10. Draw the more than type ogive curve for the following frequency distribution.

No. of workers absent	0–4	5–9	10–14	15–19	20–24
No. of days	45	55	30	20	15

11. Draw the less than type ogive curve for the following data of CV%.

CV%	3.0–3.4	3.5–3.9	4.0–4.4	4.5–4.9	5.0–5.4
No. of tests	4	10	25	16	10

12. Following data is related to the daily wages of the workers.

Daily Wages	140–150	150–160	160–170	170–180	180–190
No. of Workers	45	55	30	20	15

Draw the histogram for the above data.

13. Draw the less than type ogive curve for the following data of the breaking strength.

Breaking strength	190–195	195–200	200–205	205–210	210–215
No. of samples	10	20	50	30	15

14. Draw the more than type ogive curve for the following data of number of absent workers.

No. of absent workers	0–9	10–19	20–29	30–39	40–49	50–59
No. of days	7	17	22	15	7	2

15. Draw the histogram for the following frequency distribution of count of the yarn.

Yarn Count	29.1–29.5	29.6–30.0	30.1–30.5	30.6–31.0	31.0–31.5
Frequency	5	26	54	15	2

Measures of central tendency

4.1 Introduction of the central tendency

In Chapter 3, we have seen that in order to understand the data collected in statistics, the data can be classified and represented graphically. What next? After classifying the data, their central tendency can be studied. Generally in statistics, the observations of the data collected are concentrated around the central value of the data. This tendency of the observations toward the central value is called the central tendency. If this central tendency of the data is quantified or measured, then it can be treated as the representative of the data and can be used for the comparison of the central tendency of the data. For example, if a textile engineer makes 10 count tests on the same yarn, he will get 10 different values. What conclusion should he come to about the count of the yarn? But if all the 10 results are observed carefully, then they will be concentrated around some value that can be measured and can be representative of the 10 results.

4.2 Measure of central tendency

Any numeric figure or the value, which gives idea regarding the central tendency of the data, is called the measure of central tendency. In practice, it is also called an “average.” There are several measures of central tendency, and each of them has some advantages and disadvantages. Arithmetic mean (AM), median, mode, weighted AM, harmonic mean, and geometric mean are some of the main averages.

4.3 Arithmetic mean

This is the most popular and commonly used measure of central tendency, as it is based on all observations and is simple by its definition. It can also be used for further mathematical calculations. The AM is defined as follows:

$$\text{A.M.} = \frac{\text{Total of all observations}}{\text{Total number of observations}}$$

In particular,

Suppose X is the variable of the data. The AM of the variable X will be denoted by \bar{X} and will be defined as follows:

Case I If the data contain only ' n ' observations x_1, x_2, \dots, x_n of the variable X , then the AM is defined as follows:

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n}$$

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies, then the AM of the data is defined as follows:

$$\bar{X} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \dots + f_k \cdot x_k}{N} = \frac{\sum f_i \cdot x_i}{N}$$

where, $N = \sum f_i$

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid-points of the class intervals and f_1, f_2, \dots, f_k as the class frequencies, then the AM of the data is defined as follows:

$$\bar{X} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \dots + f_k \cdot x_k}{N} = \frac{\sum f_i \cdot x_i}{N}$$

where, $N = \sum f_i$

Properties of the AM

1. If \bar{X}_1 is the AM of first data of n_1 observations and \bar{X}_2 is the AM of second data of n_2 observations, then the mean of the combined data of $n_1 + n_2$ observations can be given as follows:

$$\text{Combined mean} = \bar{X} = \frac{n_1 \cdot \bar{X}_1 + n_2 \cdot \bar{X}_2}{n_1 + n_2}$$

2. AM is affected by change of origin as well as the scale.

4.4 Computation of AM for the frequency distribution

When the data are given in the form of the frequency distribution, then AM can be calculated by the following two methods:

1. Direct method

In this case, AM is calculated directly using the formula and by preparing a table such as Table 4.1.

Table 4.1

X	f_i	$f_i x_i$
x_1	f_1	$f_1 x_1$
x_2	f_2	$f_2 x_2$
...
x_k	f_k	$f_k x_k$
Total	$N = \sum f_i$	$\sum f_i x_i$

$$\bar{X} = \frac{\sum f_i \cdot x_i}{N}$$

2. Indirect method

In this case, AM is calculated indirectly by transforming the variable X into another variable U . There are two ways of transforming the variable X into the variable U .

Change-of-origin method

In the case of change-of-origin method, the variable U is defined as $U = X - A$ and the AM for the variable X is calculated using the relationship as $\bar{X} = A + \bar{U}$ the AM is affected by the change of origin.

Where,

$$\bar{U} = \frac{\sum f_i \cdot u_i}{N} \text{ is the AM of the variable } U.$$

In addition, the AM is calculated by preparing a table (see Table 4.2) and by using the above formulae as follows:

Table 4.2

X	f_i	u_i	$f_i u_i$
x_1	f_1	u_1	$f_1 u_1$
x_2	f_2	u_2	$f_2 u_2$
....
x_k	f_k	u_k	$f_k u_k$
Total	$N = \sum f_i$		$\sum f_i u_i$

$$\bar{U} = \frac{\sum f_i \cdot u_i}{N}$$

$$\bar{X} = A + h\bar{U}$$

Change-of-origin and scale method

In the case of change-of-origin and scale method, the variable U is defined as $U = \frac{X - A}{h}$ and the AM for the variable X is calculated using the relationship

$$\bar{X} = A + h\bar{U} \text{ as the AM is affected by the change of origin and scale.}$$

Where,

$$\bar{U} = \frac{\sum f_i \cdot u_i}{N} \text{ is the AM of the variable } U.$$

Also, the AM is calculated by preparing a table (see Table 4.3) and by using the above formulae as follows:

Table 4.3

X	Frequency	u_i	$f_i u_i$
x_1	f_1	u_1	$f_1 u_1$
x_2	f_2	u_2	$f_2 u_2$
....
x_k	f_k	u_k	$f_k u_k$
Total	$N = \sum f_i$		$\sum f_i u_i$

$$\bar{U} = \frac{\sum f_i \cdot u_i}{N}$$

$$\bar{X} = A + h\bar{U}$$

Example 4.1

Ten-count tests made on certain yarn have shown following results:

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Calculate AM using the above data and comment on it.

Solution

Here, X represents count of the yarn

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n} = \frac{229.9}{10} = 22.99$$

Thus, from the value of the AM, it can be said that the average count of the yarn is 22.99 units.

Example 4.2

The following data are related to the number of defective garments produced by a group of workers in an industry.

No. of defective garments	28	32	40	50	58	62
No. of days	5	10	12	20	10	3

Calculate AM using the above data and comment on it.

Solution

Here, X represents number of defective garments produced by the workers in a day (See Table 4.4.).

Table 4.4

X	f_i	$f_i x_i$
28	5	140
32	10	320
40	12	480
50	20	1000
58	10	580
62	3	186
Total	60	2706

$$\bar{X} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \dots + f_k \cdot x_k}{N} = \frac{\sum f_i \cdot x_i}{N} = \frac{2706}{60} = 45.1$$

Thus, from the value of the AM, it can be concluded that on an average 45 defective garments are produced by the group of workers everyday.

Example 4.3

The following data are related to the linear density of yarn.

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Calculate AM using the above data and comment on it.

Solution

Here, X represents linear density of the yarn

- a. Calculation of AM by direct method

In this method, a table is prepared (see Table 4.5)

Table 4.5

X	f_i	x_i	$f_i x_i$
13.00–13.25	8	13.125	105.0000
13.25–13.50	12	13.375	160.5000
13.50–13.75	20	13.625	272.5000
13.75–14.00	25	13.875	346.8750
14.00–14.25	22	14.125	310.7500
14.25–14.50	10	14.375	143.7500
14.50–14.75	3	14.625	43.8750
Total	100		1383.2500

$$\bar{X} = \frac{\sum f_i \cdot x_i}{N} = \frac{1383.25}{100} = 13.8325$$

Thus, from the value of the AM, it is clear that the average linear density of the yarn is 13.8325 units.

- b. Calculation of AM by indirect method. See Table 4.6.

Table 4.6

X	f_i	x_i	$u_i = x_i - 13.875$	$f_i u_i$
13.00–13.25	8	13.125	-3	-24
13.25–13.50	12	13.375	-2	-24
13.50–13.75	20	13.625	-1	-20
13.75–14.00	25	13.875	0	0
14.00–14.25	22	14.125	1	22
14.25–14.50	10	14.375	2	20
14.50–14.75	3	14.625	3	9
Total	100			-17

$$\bar{U} = \frac{\sum f_i \cdot u_i}{N} = \frac{-17}{100} = -0.17$$

$$\bar{X} = A + h\bar{U} = 13.875 + (.25 \times -0.17) = 13.8325$$

Example 4.4

Ten-count tests carried out on the yarn of a cone have shown AM 25.62 and another 15-count tests carried out on the yarn of same cone have shown AM 26.25. What is the AM of the yarn if all 25 tests taken together?

Solution

Here $\bar{X}_1 = 25.62$ is the AM of first data of $n_1 = 10$ observations and $\bar{X}_2 = 26.25$ is the AM of second data of $n_2 = 15$ observations.

Now, the mean of the combined data of $n_1 + n_2 = 25$ observations can be given as follows:

$$\text{Combined mean} = \bar{X} = \frac{n_1 \cdot \bar{X}_1 + n_2 \cdot \bar{X}_2}{n_1 + n_2} = \frac{10 \times 25.62 + 15 \times 26.25}{25} = 25.998$$

Therefore, the AM of 25 tests combined together is 25.998 units.

4.5 Median

Median is another measure of central tendency and is defined as the value or the observation, which divides the data into two parts of equal size. That is, it is the value below which there are 50% observations and above which there are 50% observations. Thus, each part on both sides of median contain 50% of the observations.

4.6 Computation of median

Case I If the data contain only ‘ n ’ observations x_1, x_2, \dots, x_n of the variable X , then after arranging the observations in the increasing/decreasing order of magnitude

$$\text{Median} = \left[\frac{n+1}{2} \right]^{\text{th}} \text{ value}$$

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies, then

$$\text{Median} = \left[\frac{N+1}{2} \right]^{\text{th}} \text{ value}$$

where, $N = \text{Total number of observations} = \sum f_i$

Note that here median that is $\left[\frac{N+1}{2} \right]^{\text{th}}$ value can be obtained with the help of cumulative frequencies of less than type.

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid-points of the class intervals and f_1, f_2, \dots, f_k as the class frequencies, then

$$\text{Median} = \left[\frac{N}{2} \right]^{\text{th}} \text{ value}$$

In this case, median that is $\left[\frac{N}{2} \right]^{\text{th}}$ value can be obtained using following steps:

Step 1 Convert the classes into exclusive type (continuous), if they are not of exclusive.

Step 2 Determine the class interval containing median (median class) by using the value $\left[\frac{N}{2} \right]$ and the cumulative frequencies of less than type.

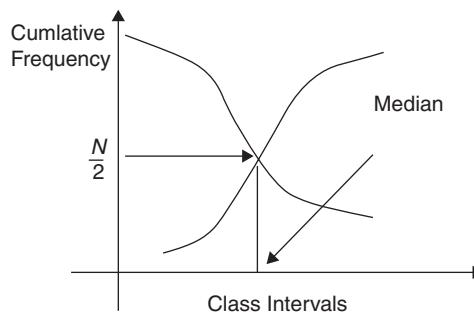
Step 3 Determine median using the formula

$$\text{Median} = L + \frac{h}{f} \left[\frac{N}{2} - \text{CF}(p) \right]$$

where, "L" is the lower boundary of the median class; "h" is the class width of the median class, 'f' is the frequency of the median class and $\text{CF}(p)$ is the cumulative frequency of premedian class.

4.7 Graphical determination of median

Graphically, median can be determined by drawing the less than type and the more than type ogive curve as follows:



4.1

Example 4.5

Ten strength tests made on certain yarn have shown the following results:

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Find median for the above data and comment on it.

Solution

Here variable X represents count of the yarn

The observations in the increasing order of the magnitude are

22.6, 22.8, 22.9, 22.9, 23.0, 23.0, 23.0, 23.1, 23.2, 23.2, 23.

Now,

$$\begin{aligned}\text{Median} &= \left[\frac{n+1}{2} \right]^{\text{th}} \text{ value} = \left[\frac{10+1}{2} \right]^{\text{th}} \text{ value} = 5.5^{\text{th}} \text{ value} \\ &= 5.5^{\text{th}} \text{ value} \\ &= 5^{\text{th}} \text{ value} + 0.5(6^{\text{th}} \text{ value} - 5^{\text{th}} \text{ value}) \\ &= 23 + 0.5(23 - 23) \\ &= 23\end{aligned}$$

Thus, from the value of the median, we can say that the average count of the yarn is 23 units.

Example 4.6

The following data are related to the number of defective garments produced by a group of workers in an industry.

No. of defective garments	28	32	40	50	58	62
No. of days	5	10	12	20	10	3

Determine median for the above data and comment on it.

Solution

Here X represents number of defective garments produced

Table 4.7

x_i	f_i	CF(L)
28	5	5
32	10	15
40	12	27
50	20	47
58	10	57
62	3	60
Total	60	

$$\text{Median} = \left[\frac{N+1}{2} \right]^{\text{th}} \text{ value} = \left[\frac{60+1}{2} \right]^{\text{th}} \text{ value} = 30.5^{\text{th}} \text{ value}$$

$$= 30^{\text{th}} \text{ value} + 0.5 (31^{\text{st}} \text{ value} - 30^{\text{th}} \text{ value})$$

$$= 50 + 0.5 (50 - 50)$$

$$= 50$$

Therefore, from the value of the median, we can say that on an average 50 defective garments are produced by the group of workers everyday.

Example 4.7

The following data are related to the linear density of yarn:

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Calculate median for the above data and comment on it.

Solution

Let, the variable X represents linear density of the yarn

Table 4.8

C.I	f_i	CF(L)
13.00–13.25	8	8
13.25–13.50	12	20
13.50–13.75	20	40
13.75–14.00	25	65
14.00–14.25	22	87
14.25–14.50	10	97
14.50–14.75	3	100
	100	

$$\text{Median} = \left[\frac{N}{2} \right]^{\text{th}} \text{ value} = \left[\frac{100}{2} \right]^{\text{th}} = 50^{\text{th}} \text{ value}$$

Using cumulative frequencies of less than type, the median class is 13.75–14.00
Thus,

$$\begin{aligned}\text{Median} &= L + \frac{h}{f} \left[\frac{N}{2} - \text{CF}(p) \right] \\ &= 13.75 + \frac{.25}{25} [50 - 40] \\ &= 13.75 + 0.1 = 13.85\end{aligned}$$

Thus, from the value of the median, we can say that the average linear density of the yarn is 13.85 units.

4.8 Mode

This is also one of the important measures of central tendency, which is used widely in real life. It is defined as the value or the observation of the data, which occurs maximum number of times, that is, which has maximum frequency, or maximum frequency density around it.

4.9 Determination of mode

Case I If the data contain only ‘ n ’ observations x_1, x_2, \dots, x_n of the variable X , then the mode is the observation, which occurs maximum number of times.

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies, then mode is the observation with maximum frequency.

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid-points of the class intervals and f_1, f_2, \dots, f_k as the class frequencies, then mode is the observation with maximum frequency density. The mode in this case can be determined using the following steps:

- Step 1 Convert the classes into the exclusive type (continuous type), if they are not exclusive.
- Step 2 Determine the class interval containing mode (Modal class) according to the maximum frequency.

Step 3 Determine the mode using the following formula:

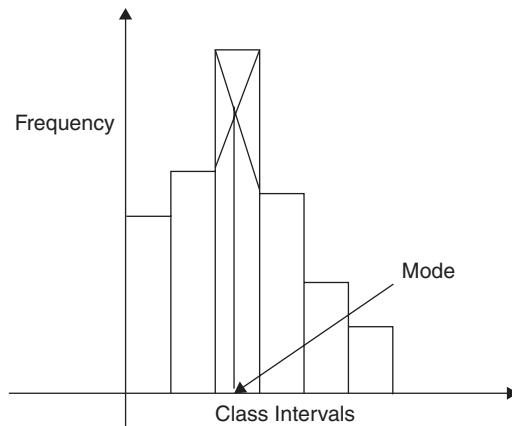
$$\text{Mode} = L + h \left[\frac{f_m - f_0}{2f_m - f_1 - f_0} \right]$$

Where,

L represents lower boundary of the modal class, h represents width of the modal class; f_m represents frequency of modal class, f_0 represents frequency of premodal class, and f_1 represents frequency of postmodal class interval.

4.10 Determination of mode graphically

Mode can also be determined graphically for the data in the form of grouped frequency distribution by drawing the histogram as follows:



4.2

Example 4.8

Ten strength tests made on certain yarn have shown following results:

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Calculate mode for the above data and comment on it.

Solution

Here, variable X represents count of the yarn

Also for the given data, mode = 23.0 as observation 23.0 occurs maximum number of times.

Thus, from the value of the mode, it can be said that on an average the strength of the yarn is 23.0 units.

Example 4.9

The following data are related to the number of defective garments produced by a group of workers in an industry.

No. of defective	28	32	40	50	58	62
No. of days	5	10	12	20	10	3

Find mode using the above data and comment on it.

Solution

Here, the variable X represents number of defective garments produced by the workers per day.

Also for the given data mode = 50 as observation 50 has maximum frequency, Thus, from the value of the mode, it is clear that on an average 50 defective articles are produced by the group of workers everyday.

Example 4.10

The following data are related to the linear density of yarn.

Linear density 13.00–13.25 13.25–13.50 13.50–13.75 13.75–14.00 14.00–14.25 14.25–14.50 14.50–14.75

No. of tests	8	12	20	25	22	10	3
--------------	---	----	----	----	----	----	---

Calculate mode using the above data and write the conclusion.

Solution

Let variable X represent linear density of the yarn.

Here, the class interval 13.75–14.00 has the maximum frequency hence; the modal class interval is 13.75–14.00.

$$\begin{aligned} \text{Mode} &= L + h \left[\frac{f_m - f_0}{2f_m - f_1 - f_0} \right] \\ &= 13.75 + 0.25 \left[\frac{25 - 20}{2 \times 25 - 22 - 20} \right] \\ &= 13.9063 \end{aligned}$$

Thus, from the value of the mode, we can say that on an average linear-density of the yarn is 13.9063 units.

4.11 Exercise

- What is central tendency? What are the different measures of central tendency? Describe any one.
- The following are the results of seven count tests made on a yarn:

30.0	31.5	29.5	30.5	30.0	30.5	31.0
------	------	------	------	------	------	------

Find AM and comment on it.

- Calculate the AM for the following data of breaking strength of yarn and comment on it.

Breaking strength(gm)	190–195	195–200	200–205	205–210	210–215
No. of samples	15	30	50	20	10

- Find AM for the following data of fabric production (in meters) and comment on it.

Production	390–395	395–400	400–405	405–410	410–415	415–420	420–425
No. of days	10	20	35	50	40	30	15

- Determine the mode from the following frequency distribution and comment on it.

Length of fiber(mm)	10–15	15–20	20–25	25–30	30–35	35–40
No. of fibers	12	19	26	21	15	9

- Compute median and mode for the data given below and comment on them.

Daily salary (Rs.)	50–60	60–70	70–80	80–90	90–100
No. of workers	4	6	10	18	12

- Following frequency distribution represents marks obtained by 100 students in an examination. Compute median of the marks and comment on it.

Marks	0–20	20–40	40–60	60–80	80–100
No. of students:	5	12	32	40	11

- Compute median and mode for the following data:

Linear density	13–13.5	13.5–14	14–14.5	14.5–15	15–15.5
No. of observations	5	15	25	20	10

9. The following frequency distribution represents the number of accidents in a month. Compute AM and comment on it.

No. of accidents	0	1	2	3	4	5
No. of days	3	5	7	8	5	2

10. AM of weights of 100 boys is 50 kg and the AM of weights of 50 girls is 45 kg. Calculate AM of weights of combined group of boys and girls.
11. The average and the standard deviation of the salary of 50 nontechnical staff of a textile mill are Rs. 1200 and 200, respectively. The corresponding figures for 75 technical staff are Rs. 1500 and 250, respectively. What are the average and standard deviation of the salary of all employees in the mill?

5.1 Introduction

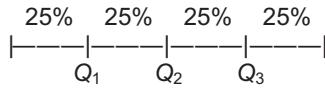
Most of the times after collecting data or from the available data, we are interested in finding some values that may have some percentage of values below/above them. Such values in statistics are also called the partition values. For example, from the data of 50 strength results obtained from a fabric, the manufacturer may be interested in knowing the strength value above which he will get 90% of the test results or 75% of the test results, etc. These are called the partition values because they divide the data under study into number of parts. Thus, partition values are the values that divide the data or the set of observations into number of parts of equal size. There are three different types of partition values according to the number of parts in which the data are divided. These three different partition values are known as the quartiles, deciles, and percentiles.

5.2 Quartiles

Quartiles are the three different values denoted by Q_1 , Q_2 , and Q_3 which that divide the data under study into four parts of equal size. These values are known as the first, second, and third quartiles.

Sometimes, these are also called the lower, middle, and upper quartiles. Here, each of the four parts contains 25% of the observations. Thus, first quartile Q_1 is the value below which there are 25% of the observations and above which there are 75% of the observations or third quartile Q_3 is the value above which there are 25% of observations and below which there are 75% of the observations. Note that, second quartile Q_2 is the median of the data because it will have 50% of the values below and above it.

The meaning of quartiles can be easily understood from the following figure:



Computation of Quartiles

Case I If the data contain only ‘ n ’ observations x_1, x_2, \dots, x_n of the variable X and they are “arranged in the increasing order of the magnitude,” then,

$$j^{\text{th}} \text{ quartile} = Q_j = \left[j \times \frac{n+1}{4} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2, \text{ and } 3$$

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies, then

$$j^{\text{th}} \text{ quartile} = Q_j = \left[j \times \frac{N+1}{4} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2 \text{ and } 3$$

Where, $N = \sum f_i$ = Total frequency

Note that here j^{th} quartile i.e. Q_j i.e. $\left[j \times \frac{N+1}{4} \right]^{\text{th}}$ value is obtained with the help of the CF(L)

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid-points of the class intervals and f_1, f_2, \dots, f_k as the frequencies, then

$$j^{\text{th}} \text{ quartile} = Q_j = \left[j \times \frac{N}{4} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2, \text{ and } 3$$

where, $N = \sum f_i$ = Total frequency

Note that here j^{th} quartile i.e. Q_j i.e. $\left[j \times \frac{N}{4} \right]^{\text{th}}$ value that lies in some class

interval which is again decided with the help of the CF(L). How to decide such value using CF(L) will be discussed with the help of illustration while solving the examples afterward.

In this case, j^{th} quartile, i.e., Q_j , i.e. $\left[j \times \frac{N}{4} \right]$ value can be computed by using following steps:

Step 1 Convert the classes of the frequency distribution into continuous (exclusive) type if they are not continuous.

Step 2 Determine the class interval containing j^{th} quartile, i.e., Q_j according to the value $\left[j \times \frac{N}{4} \right]$ and $CF(L)$. How to decide such class interval using $CF(L)$ will also be discussed with the help of illustration while solving the examples afterward.

Step 3 Find the j^{th} quartile, i.e., Q_j using following formula

$$j^{\text{th}} \text{ quartile} = Q_j = L + \frac{h}{f} \times \left[\left[j \times \frac{N}{4} \right] - CF(p) \right] \quad \text{for } j = 1, 2 \text{ and } 3$$

Where,

L – lower boundary of the class interval containing j^{th} quartile.

h – class width of the class interval containing j^{th} quartile.

f – frequency the class interval containing j^{th} quartile.

$CF(p)$ – cumulative frequency of the previous class of the class interval containing j^{th} quartile.

Example 5.1

Ten-count tests made on certain yarn have shown following results:

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Find first and third quartiles using the above data and comment about each of them.

Solution

Suppose that the variable X represents count of the yarn.

Observations in increasing order are

22.6, 22.8, 22.9, 22.9, 23.0, 23.0, 23.0, 23.1, 23.2, 23.4

Now the first quartile is computed as follows:

$$\begin{aligned} \text{Therefore, } Q_1 &= \left[\frac{n+1}{4} \right]^{\text{th}} \text{ value} = \frac{11}{4}^{\text{th}} \text{ value} = 2.75^{\text{th}} \text{ value} \\ &= 2^{\text{nd}} + 0.75 \times (3^{\text{rd}} - 2^{\text{nd}}) \\ &= 22.8 + 0.75 \times (22.9 - 22.8) = 22.875 \end{aligned}$$

Thus, from the value of first quartile, it can be concluded that in 25% of the cases count of the yarn will be below 22.875 and in 75% of the cases it will be above 22.875.

Similarly, the third quartile can be computed as follows:

$$\begin{aligned}\text{Therefore, } Q_3 &= \left[3 \times \frac{n+1}{4} \right]^{\text{th}} \text{ value} = 3 \times \frac{11}{4}^{\text{th}} \text{ value} = 8.25^{\text{th}} \text{ value} \\ &= 8^{\text{th}} + 0.25 \times (9^{\text{th}} - 8^{\text{th}}) \\ &= 23.1 + 0.75 \times (23.2 - 23.1) = 23.175\end{aligned}$$

In addition, from the value of third quartile, it can be concluded that in 75% of the cases count of the yarn will be below 23.175 and in 25% of the cases it will be above 23.175.

Example 5.2

The following data are related to the number of defective garments produced by a group of workers in an industry.

No. of defective garments	28	32	40	50	58	62	Total
No. of days	5	10	12	20	10	3	60

Find first and third quartiles using the above data and comment about each of them.

Solution

Suppose that variable X represents number of defective garments. The cumulative frequencies of less than type are obtained in Table 5.1.

Table 5.1

x_i	28	32	40	50	58	62	Total
f_j	5	10	12	20	10	3	60
CF(L)	5	15	27	47	57	60	

Now the first quartile is computed as follows:

$$\begin{aligned}\text{Therefore, } Q_1 &= \left[\frac{N+1}{4} \right]^{\text{th}} \text{ value} = \frac{61}{4}^{\text{th}} \text{ value} = 15.25^{\text{th}} \text{ value} \\ &= 15^{\text{th}} + 0.25(16^{\text{th}} - 15^{\text{th}}) \\ &= 32 + 0.25 \times (40 - 32) = 34\end{aligned}$$

Note that here 15^{th} value is obtained by comparing 15 with cumulative frequencies of less than type, i.e., $\text{CF}(L)$. Here, the value x_i having $\text{CF}(L)$ just ≥ 15 is selected as the 15^{th} value.

Thus, from the value of first quartile, it can be concluded that in 25% of the cases number of the defective garments produced will be <40 and in 75% of the cases it will be >40.

Similarly, the third quartile can be computed as follows:

$$\begin{aligned}\text{Therefore, } Q_3 &= \left[3 \times \frac{N+1}{4} \right]^{\text{th}} \text{ value} = 3 \times \frac{61}{4}^{\text{th}} \text{ value} = 45.75^{\text{th}} \text{ value} \\ &= 45^{\text{th}} + 0.75(46^{\text{th}} - 45^{\text{th}}) \\ &= 50 + 0.75 \times (50 - 50) = 50\end{aligned}$$

In addition, from the value of third quartile, it can be concluded that in 25% of the cases number of the defective garments produced will be >50 and in 75% of the cases it will be <50.

Example 5.3

The following data are related to the linear density of yarn:

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Find first and third quartiles using the above data and comment about each of them.

Solution

Suppose that variable X represents linear density of the yarn.

The cumulative frequencies of less than type are obtained in Table 5.2.

Table 5.2

X	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
F	8	12	20	25	22	10	3
CF(L)	8	20	40	65	87	97	100

Now the quartiles are computed as follows:

$$\text{Therefore, } Q_1 = \left[\frac{N}{4} \right]^{\text{th}} \text{ value} = \frac{100}{4}^{\text{th}} \text{ value} = 25^{\text{th}} \text{ value}$$

Comparing 25 with CF(L) the cumulative frequency 40 is just >25, hence class interval containing Q_1 is 13.50–13.75

$$\text{Therefore, } Q_1 = L + \frac{h}{f} \left[\frac{N}{4} - CF(P) \right] = 13.50 + \frac{0.25}{20} [25 - 20] = 13.5625$$

Thus, from the value of first quartile, it can be concluded that in 25% of the cases linear density of the yarn will be <13.5625; and in 75% of the cases, it will be >13.5625.

Similarly, the third quartile can be computed as follows:

$$\text{Therefore, } Q_3 = \left[3 \times \frac{N}{4} \right]^{\text{th}} \text{ value} = 3 \times \frac{100^{\text{th}}}{4} \text{ value} = 75^{\text{th}} \text{ value}$$

Comparing 75 with CF(L) class interval containing Q_3 is 14.00–14.25

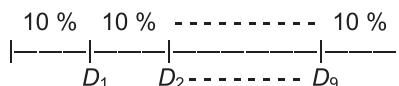
$$\text{Therefore, } Q_3 = L + \frac{h}{f} \left[3 \times \frac{N}{4} - CF(P) \right] = 14.00 + \frac{0.25}{20} [75 - 65] = 14.1136$$

In addition, from the value of third quartile, it can be concluded that in 75% of the cases linear density of the yarn will be <14.1136; and in 25% of the cases, it will be >14.1136.

5.3 Deciles

Deciles are the nine different values denoted by D_1, D_2, \dots, D_9 , which divide the data under study into 10 parts of equal size. These values are known as the first, second, ..., ninth deciles.

Here, each of the 10 parts contains 10% of the observations. Thus, first decile D_1 is the value below which there are 10% of the observations and above which there are 90% of the observations or third decile D_3 is the value above which there are 70% of observations and below which there are 30% of the observations. The meaning of deciles can be easily understood from the following figure.



Computation of Deciles

Case I If the data contain only ‘ n ’ observations x_1, x_2, \dots, x_n of the variable X and they are “arranged in the increasing order of the magnitude,” then

$$j^{\text{th}} \text{decile} = D_j = \left[j \times \frac{n+1}{10} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2, \dots, 9$$

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies, then

$$j^{\text{th}} \text{decile} = D_j = \left[j \times \frac{N+1}{10} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2, \dots, 9$$

Where, $N = \sum f_i$ = Total frequency

Note that here also j^{th} decile, i.e., D_j , i.e., $\left[j \times \frac{N+1}{10} \right]^{\text{th}}$ value is obtained with the help of the $CF(L)$

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid-points of the class intervals and f_1, f_2, \dots, f_k as the frequencies, then

$$j^{\text{th}} \text{decile} = D_j = \left[j \times \frac{N}{10} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2, \dots, 9$$

Where, $N = \sum f_i$ = Total frequency

Note that here j^{th} decile, i.e., D_j , i.e., $\left[j \times \frac{N}{10} \right]^{\text{th}}$ value that lies in some class interval, which is again decided with the help of the $CF(L)$. How to decide such value using $CF(L)$ is already discussed with the help of illustration while solving the examples of quartile.

In this case also j^{th} decile, i.e., D_j , i.e., $\left[j \times \frac{N}{10} \right]^{\text{th}}$ value can be computed by using following steps:

Step 1 Convert the classes of the frequency distribution into continuous (exclusive) type if they are not continuous.

Step 2 Determine the class interval containing j^{th} decile, i.e., D_j according to the value $\left[j \times \frac{N}{10} \right]$ and $CF(L)$. How to decide such class interval using $CF(L)$ is already discussed with the help of illustration while solving the example of quartiles.

Step 3 Find the j^{th} decile, i.e., D_j note the following formula:

$$j^{\text{th}} \text{decile} = D_j = L + \frac{h}{f} \times \left[\left[j \times \frac{N}{10} \right] - CF(p) \right] \quad \text{for } j = 1, 2, \dots, 9$$

Where,

L – lower boundary of the class interval containing j^{th} decile.

h – class width of the class interval containing j^{th} decile.

f – frequency of the class interval containing j^{th} decile.

$\text{CF}(p)$ – cumulative frequency of the previous class of the class interval containing j^{th} decile.

Example 5.4

Ten-count tests made on certain yarn have shown following results:

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Find first and third deciles using the above data and comment about each of them.

Solution

Suppose that the variable X represents count of the yarn.

Observations in increasing order are

22.6, 22.8, 22.9, 22.9, 23.0, 23.0, 23.0, 23.1, 23.2, 23.4

Now the first decile is computed as follows:

$$\begin{aligned}\text{Therefore, } D_1 &= \left[\frac{n+1}{10} \right]^{\text{th}} \text{ value} = \frac{11}{10}^{\text{th}} \text{ value} = 1.1^{\text{th}} \text{ value} \\ &= 1^{\text{st}} + 0.1 \times (2^{\text{nd}} - 1^{\text{st}}) \\ &= 22.6 + 0.1 \times (22.8 - 22.6) = 22.62\end{aligned}$$

Thus, from the value of first decile, it can be concluded that in 10% of the cases, the count of the yarn will be <22.62 ; and in 90% of the cases, it will be >22.62 .

Similarly the third decile can be computed as follows:

$$\begin{aligned}\text{Therefore, } D_3 &= \left[3 \times \frac{n+1}{10} \right]^{\text{th}} \text{ value} = 3 \times \frac{11}{10}^{\text{th}} \text{ value} = 3.3^{\text{rd}} \text{ value} \\ &= 3^{\text{rd}} + 0.3 \times (4^{\text{th}} - 3^{\text{rd}}) \\ &= 22.9 + 0.3 \times (22.9 - 22.9) = 22.9\end{aligned}$$

In addition, from the value of third decile, it can be concluded that in 30% of the cases, the count of the yarn will be <22.9 ; and in 70% of the cases, it will be >22.9 .

Example 5.5

The following data are related to the number of defective garments produced by a group of workers in an industry.

No. of defective garments	28	32	40	50	58	62	Total
No. of days	5	10	12	20	10	3	60

Find fourth and seventh deciles using the above data and comment about each of them.

Solution

Suppose that variable X represents number of defective garments. The cumulative frequencies of less than type are obtained in Table 5.3.

Table 5.3

x_i	28	32	40	50	58	62	Total
f_i	5	10	12	20	10	3	60
CF(L)	5	15	27	47	57	60	

Now the fourth decile is computed as follows:

$$\text{Therefore, } D_4 = \left[4 \times \frac{N+1}{10} \right]^{\text{th}} \text{ value} = 4 \times \frac{61}{10}^{\text{th}} \text{ value} = 24.4^{\text{th}} \text{ value}$$

$$D_4 = 24^{\text{th}} + 0.4(25^{\text{th}} - 24^{\text{th}}) = 40 + 0.4 \times (40 - 40) = 40$$

Note that here also 24^{th} value is obtained by comparing 24 with cumulative frequencies of less than type that is $CF(L)$ as discussed earlier.

Thus, from the value of fourth decile, it can be concluded that in 40% of the cases, the number of the defective garments produced will be <40 ; and in 60% of the cases, it will be >40 .

Similarly, the seventh decile can be computed as follows:

$$\text{Therefore, } D_7 = \left[7 \times \frac{N+1}{10} \right]^{\text{th}} \text{ value} = 7 \times \frac{61}{10}^{\text{th}} \text{ value} = 42.7^{\text{th}} \text{ value}$$

$$= 42^{\text{nd}} + 0.7(43^{\text{rd}} - 42^{\text{nd}})$$

$$D_7 = 50 + 0.7 \times (50 - 50) = 50$$

In addition, from the value of seventh decile, it can be concluded that in 30% of the cases, the number of the defective garments produced will be >50 ; and in 70% of the cases, it will be <50 .

Example 5.6

The following data are related to the linear density of yarn:

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Find first and third deciles using the above data and comment about each of them.

Solution

Suppose that variable X represents linear density of the yarn.

The cumulative frequencies of less than type are obtained in Table 5.4.

Table 5.4

X	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
f	8	12	20	25	22	10	3
CF(L)	8	20	40	65	87	97	100

Now the deciles are computed as follows:

$$\text{Therefore, } D_1 = \left[\frac{N}{10} \right]^{\text{th}} \text{ value} = \frac{100}{10}^{\text{th}} \text{ value} = 10^{\text{th}} \text{ value}$$

Comparing 10 with CF(L) class interval containing D_1 is 13.25–13.50

$$\text{Therefore, } D_1 = L + \frac{h}{f} \left[\frac{N}{10} - CF(P) \right] = 13.25 + \frac{0.25}{12} [10 - 8] = 13.2917$$

Thus, from the value of first decile, it can be concluded that in 10% of the cases, the linear density of the yarn will be <13.2917 ; and in 90% of the cases, it will be >13.2917 .

Similarly, the third decile can be computed as follows:

$$\text{Therefore, } D_3 = \left[3 \times \frac{N}{10} \right]^{\text{th}} \text{ value} = 3 \times \frac{100}{10}^{\text{th}} \text{ value} = 30^{\text{th}} \text{ value}$$

Comparing 30 with CF(L) class interval containing D_3 is 13.50–13.75

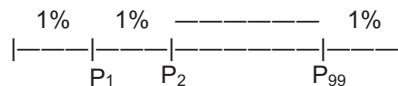
$$\text{Therefore, } D_3 = L + \frac{h}{f} \left[3 \times \frac{N}{10} - CF(P) \right] = 13.50 + \frac{0.25}{20} [30 - 20] = 13.625$$

Also, from the value of third decile, it can be concluded that in 70% of the cases, the linear density of the yarn will be >13.625 ; and in 30% of the cases, it will be <13.625 .

5.4 Percentiles

Percentiles are the 99 different values denoted by P_1, P_2, \dots, P_{99} , which divide the data under study into hundred parts of equal size. These values are known as the first, second, ..., ninety-ninth percentiles.

Here, each of the hundred parts contains 1% of the observations. Thus, first percentile P_1 is the value below which there are 1% of the observations and above which there are 99% of the observations or 30th percentile P_{30} is the value above which there are 70% of observations and below which there are 30% of the observations. The meaning of percentiles can be easily understood from the following figure.



Computation of percentiles

Case I If the data contain only ' n ' observations x_1, x_2, \dots, x_n of the variable X and they are arranged in the increasing order of the magnitude, then,

$$j^{\text{th}} \text{ percentile} = P_j = \left[j \times \frac{n+1}{100} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2, \dots, 99$$

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies, then

$$j^{\text{th}} \text{ percentile} = P_j = \left[j \times \frac{N+1}{100} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2, \dots, 99$$

Where, $N = \sum f_i$ = Total frequency

Note that here also j^{th} percentile, i.e., P_j , i.e., $\left[j \times \frac{N+1}{100} \right]^{\text{th}}$ value is obtained with the help of the CF(L)

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid-points of the class intervals and f_1, f_2, \dots, f_k as the frequencies, then

$$j^{\text{th}} \text{ percentile} = P_j = \left[j \times \frac{N}{100} \right]^{\text{th}} \text{ value} \quad \text{for } j = 1, 2, \dots, 99$$

Where, $N = \sum f_i$ = Total frequency

Note that here j^{th} percentile, i.e., P_j , i.e., $\left[j \times \frac{N}{100} \right]^{\text{th}}$ value that lies

in some class interval, which is again decided with the help of the $CF(L)$. How to decide such value using $CF(L)$ is already discussed with the help of illustration while solving the examples of quartile.

In this case also j^{th} percentile, i.e., P_j , i.e., $\left[j \times \frac{N}{100} \right]^{\text{th}}$ value can be computed by using the following steps.

Step 1 Convert the classes of the frequency distribution into continuous (exclusive) type if they are not continuous.

Step 2 Determine the class interval having j^{th} percentile, i.e., P_j according to the value $\left[j \times \frac{N}{100} \right]$ and $CF(L)$. How to decide such class interval using $CF(L)$ is already discussed with the help of illustration while solving the example of quartiles.

Step 3 Find the j^{th} percentile, i.e., P_j using the following formula:

$$j^{\text{th}} \text{ percentile} = P_j = L + \frac{h}{f} \times \left[\left[j \times \frac{N}{100} \right] - CF(p) \right] \text{ for } j = 1, 2, \dots, 99$$

Where,

L – lower boundary of the class interval containing j^{th} percentile.

h – class width of the class interval containing j^{th} percentile.

f – frequency of the class interval containing j^{th} percentile.

$CF(p)$ – cumulative frequency of the previous class of the class interval containing j^{th} percentile.

Example 5.7

Ten-count tests made on certain yarn have shown following results:

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Find tenth and seventy fifth percentiles using the above data and comment about each of them.

Solution

Suppose that the variable X represents count of the yarn.

Observations in increasing order are

22.6, 22.8, 22.9, 22.9, 23.0, 23.0, 23.0, 23.1, 23.2, 23.4

Now the tenth percentile is computed as follows:

$$\begin{aligned}\text{Therefore, } P_{10} &= \left[10 \times \frac{n+1}{100} \right]^{\text{th}} \text{ value} = 10 \times \frac{11^{\text{th}}}{100} \text{ value} = 1.1^{\text{th}} \text{ value} \\ &= 1^{\text{st}} + 0.1 \times (2^{\text{nd}} - 1^{\text{st}}) \\ &= 22.6 + 0.1 \times (22.8 - 22.6) = 22.62\end{aligned}$$

Thus, from the value of tenth percentile, it can be concluded that in 10% of the cases, the count of the yarn will be <22.62 ; and in 90% of the cases, it will be >22.62 .

Similarly, the seventy-fifth percentile can be computed as follows:

$$\text{Therefore, } P_{75} = \left[75 \times \frac{n+1}{100} \right]^{\text{th}} \text{ value} = 75 \times \frac{11^{\text{th}}}{100} \text{ value} = 8.47^{\text{rd}} \text{ value}$$

$$P_{75} = 8.25^{\text{th}} = 8^{\text{th}} + 0.25(9^{\text{th}} - 8^{\text{th}}) = 23.1 + 0.25 \times (23.2 - 23.1) = 23.125$$

Also, from the value of the seventy-fifth percentile, it can be concluded that in 30% of the cases, the count of the yarn will be <23.125 ; and in 70% of the cases, it will be >23.147 .

Example 5.8

The following data are related to the number of defective garments produced by a group of workers in an industry:

No. of defective garments	28	32	40	50	58	62	Total
No. of days	5	10	12	20	10	3	60

Find 40th and 70th percentiles using the above data and comment about each of them.

Solution

Suppose that variable X represents number of defective garments. The cumulative frequencies of less than type are obtained in the Table 5.5.

Table 5.5

x_i	28	32	40	50	58	62	Total
f_i	5	10	12	20	10	3	60
CF(L)	5	15	27	47	57	60	

Now the 40th percentile is computed as follows:

$$\text{Therefore, } P_{40} = \left[40 \times \frac{N+1}{100} \right]^{\text{th}} \text{ value} = 40 \times \frac{61^{\text{th}}}{100} \text{ value} = 24.4^{\text{th}} \text{ value}$$

$$P_{40} = 24^{\text{th}} + 0.4(25^{\text{th}} - 24^{\text{th}}) = 40 + 0.4 \times (40 - 40) = 40$$

Note that here also 24th value is obtained by comparing 24 with cumulative frequencies of less than type that is CF(L) as discussed earlier.

Thus, from the value of 40th percentile, it can be concluded that in 40% of the cases, the number of the defective garments produced will be <40 ; and in 60% of the cases, it will be >40 .

Similarly, the 70th percentile can be computed as follows:

$$\text{Therefore, } P_{70} = \left[70 \times \frac{N+1}{100} \right]^{\text{th}} \text{ value} = 70 \times \frac{61^{\text{th}}}{100} \text{ value} = 42.7^{\text{th}} \text{ value}$$

$$= 42^{\text{nd}} + 0.7 \times (43^{\text{rd}} - 42^{\text{nd}})$$

$$P_{70} = 50 + 0.7 \times (50 - 50) = 50$$

Also, from the value of 70th percentile, it can be concluded that in 30% of the cases, the number of the defective garments produced will be >50 ; and in 70% of the cases, it will be <50 .

Example 5.9

The following data are related to the linear density of yarn:

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Find 10th and 75th percentiles using the above data and comment about each of them.

Solution

Suppose that variable X represents linear density of the yarn.

The cumulative frequencies of less than type are obtained in Table 5.6.

Table 5.6

X	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
f	8	12	20	25	22	10	3
CF(L)	8	20	40	65	87	97	100

Now the percentiles are computed as follows:

$$\text{Therefore, } P_{10} = \left[10 \times \frac{N}{100} \right]^{\text{th}} \text{ value} = 10 \times \frac{100}{100}^{\text{th}} \text{ value} = 10^{\text{th}} \text{ value}$$

Comparing 10 with $\text{CF}(L)$ class interval containing P_{10} is 13.25–13.50

$$\text{Therefore, } P_{10} = L + \frac{h}{f} \left[10 \times \frac{N}{100} - \text{CF}(P) \right] = 13.25 + \frac{0.25}{12} [10 - 8] = 13.2917.$$

Thus, from the value of 10th percentile, it can be concluded that in 10% of the cases, the linear density of the yarn will be <13.2917; and in 90% of the cases, it will be >13.2917.

Similarly, the 75th percentile can be computed as follows:

$$\text{Therefore, } P_{75} = \left[75 \times \frac{N}{100} \right]^{\text{th}} \text{ value} = 75 \times \frac{100}{100}^{\text{th}} \text{ value} = 75^{\text{th}} \text{ value.}$$

Comparing 75 with $\text{CF}(L)$ class interval containing P_{75} is 14.00–14.25.

$$\text{Therefore, } P_{75} = L + \frac{h}{f} \left[75 \times \frac{N}{100} - \text{CF}(P) \right] = 14.00 + \frac{0.25}{22} [75 - 65] = 14.1136$$

Also, from the value of 75th percentile, it can be concluded that in 75% of the cases, the linear density of the yarn will be <14.1136; and in 25% of the cases, it will be > 14.1136.

5.5 Exercise

- Find the third quartile and 60th percentile for the following data of daily wages of the temporary workers and comment on them.

Daily wages (Rs)	40–49	50–59	60–69	70–79	80–89
No. of workers	15	20	30	45	25

- Calculate the third decile, second percentile, and 70th percentile.

Fabric strength	45–49	50–54	55–59	60–64	65–69	70–74
No. of samples	5	10	15	20	10	5

- Calculate the third decile, first quartile, and 70th percentile for the following data and comment on them.

CV%	3.0–3.5	3.5–4.0	4.0–4.5	4.5–5.0	5.0–5.5
No. of tests	4	10	25	16	10

4. Compute median, seventh decile, and 38th percentile for the following data:

Weight in kg	10–19	20–29	30–39	40–49	50–59	60–69
No. of goods	6	10	16	14	8	4

5. Compute the first quartile, seventh decile, and 38th percentile for following data of U% of certain yarn:

2.25	4.25	2.65	3.85	5.82	3.42	4.44	2.86	2.88	4.46
------	------	------	------	------	------	------	------	------	------

6. For the following data find fourth decile and 70th percentile. Also comment on them.

Time required for replacing bobbin (in s)	15.1–15.2	15.3–15.4	15.5–15.6	15.7–15.8	15.9–16.0	16.1–16.2	16.3–16.4
No. of occasions	1	7	17	22	15	7	2

7. Find the first quartile and sixth decile for the following data of length (in cm) and comment on them.

Length (in cm)	1.0–1.5	1.5–2.0	2.0–2.5	2.5–3.0	3.0–3.5
No. of fibers	5	26	24	15	5

8. Calculate the second decile, 60th percentile, and third quartile for the following frequency distribution of number of absent workers and comment on each of them.

No. of workers absent	0–4	5–9	10–14	15–19	20–24	25–29
No. of days	45	55	30	20	15	5

9. Calculate the second quartile, sixth decile, and 52nd percentile for the following frequency distribution of the hank of the sliver.

Hank	0.1120–0.1160	0.1160–0.1180	0.1180–0.1220	0.1220–0.1260	0.1260–0.1300
No. of tests	5	20	35	28	12

Measures of dispersion

6.1 Introduction of the dispersion

In Chapter 5, we have seen that measures of central tendency help us in locating the value around which observations of the data collected are concentrated, which also acts as the representative of the data and can be used for the comparison of two or more data with each other. But study of central tendency (average) alone is not sufficient. For example, if two different companies produce yarn of same count, then the following question arises: Yarn produced by which company is called the better yarn? Thus, in statistics, it is necessary to study dispersion (variation) too. The smaller the variation in the values of the data, the larger the consistency and better may be the results obtained from the data, which is why the yarn with smaller variation in count is a better yarn. The variation of the data can be calculated using measures of dispersion. Range, quartile deviation, mean deviation, and standard deviation are the four main measures of dispersion, which are discussed below one by one.

6.2 Range

Range is the simplest measure of dispersion, which gives us an idea about the variation present in the data. It is defined as follows:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Thus, if ' X ' is the variable of the data, then

$$\text{Range} = X_{\max} - X_{\min}$$

This is a rough measure of dispersion and does not give exact variation present in the data as it is based only on two extreme values of the data. But this is suitable in the cases where we are interested in studying variation of the data roughly without going for thorough calculations.

Computation of range

Computation of range is very easy in all types of data (only “ n ” observations, ungrouped frequency distribution, or the grouped frequency distribution) because of its simple definition it is explained with suitable examples of each case.

Example 6.1

Ten-count tests made on certain yarn have shown following results:

$$22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0$$

Calculate range using the above data and comment about the variation of the data.

Solution

Suppose that variable X represents count of the yarn. Now maximum count value is $X_{\max} = 23.4$ and minimum count value is $X_{\min} = 22.6$. Therefore, the range of the data is

$$\text{Range} = X_{\max} - X_{\min} = 23.4 - 22.6 = 0.8$$

Thus from the given data of 10 observations it can be said that the variation in the count of the yarn is 0.8 units.

Example 6.2

The following data are related to the number of defective garments produced by a group of workers in an industry.

No. of defective garments	28	32	40	50	58	62	Total
No. of days	5	10	12	20	10	3	60

Find the range for the above data and comment about the variation of the data.

Solution

Suppose that variable X represents number of defective garments. Now maximum number of defective garments is $X_{\max} = 62$ and minimum number of defective garments is $X_{\min} = 28$. Therefore the range of the data is

$$\text{Range} = X_{\max} - X_{\min} = 62 - 28 = 34$$

Thus, from the given data of 60 days, it can be said that the variation in production of number of defective garments is 34 garments.

Example 6.3

The following data are related to the linear density of yarn.

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Calculate range using the above data and comment about the variation of the data.

Solution

Suppose that variable X represents linear density of the yarn. Now the mid-value or the average of first-class interval is maximum linear density; hence, $X_{\max} = 14.625$ and the mid value or the average of the last class interval is minimum linear density, hence $X_{\min} = 13.125$. Therefore, the range of the data is

$$\text{Range} = X_{\max} - X_{\min} = 14.625 - 13.125 = 1.5$$

Thus, from the given data of 100 linear density tests, it can be said that the variation in the linear density of the yarn is 1.5 units.

6.3 Quartile deviation

Quartile deviation is also the simple measure of dispersion, which gives us an idea about the variation present in the data. It is defined as follows

$$QD = \frac{Q_3 - Q_1}{2}$$

Where, Q_3 is the third quartile and Q_1 is the first quartile.

This is also a rough measure of dispersion and does not give exact variation present in the data as it is based on only mid-50% values of the data. But this is better than the range and suitable in the cases where we are interested in studying variation of the data roughly without going for thorough calculations.

Computation of Quartile deviation

Computation of Quartile deviation is also very easy. For computing quartile deviation, we should first find Q_3 (third quartile) and Q_1 (first quartile) by the procedure explained in Chapter 5, substitute them in quartile deviation formula and simplify.

Example 6.4

Ten-count tests made on certain yarn have shown the following results:

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Calculate quartile deviation using the above data and comment about the variation of the data.

Solution

Suppose that the variable X represents count of the yarn.

Observations in increasing order are

22.6, 22.8, 22.9, 22.9, 23.0, 23.0, 23.0, 23.1, 23.2, 23.4

Now the quartiles are computed as follows:

$$\begin{aligned}\text{Therefore, } Q_1 &= \left[\frac{n+1}{4} \right]^{\text{th}} \text{ value} = \frac{11}{4}^{\text{th}} \text{ value} = 2.75^{\text{th}} \text{ value} \\ &= 2^{\text{nd}} + 0.75(3^{\text{rd}} - 2^{\text{nd}}) \\ &= 22.8 + 0.75 \times (22.9 - 22.8) = 22.875\end{aligned}$$

$$\begin{aligned}\text{Therefore, } Q_3 &= \left[3 \times \frac{n+1}{4} \right]^{\text{th}} \text{ value} = 3 \times \frac{11}{4}^{\text{th}} \text{ value} = 8.25^{\text{th}} \text{ value} \\ &= 8^{\text{th}} + 0.25(9^{\text{th}} - 8^{\text{th}})\end{aligned}$$

$$Q_3 = 23.1 + 0.25 \times (23.2 - 23.1) = 23.125$$

$$\text{Therefore, } QD = \frac{Q_3 - Q_1}{2} = \frac{23.125 - 22.875}{2} = 0.15$$

Thus, from the given data of 10 observations and from quartile deviation it can be said that the variation in the count of the yarn is 0.15 units.

Example 6.5

The following data are related to the numbers of defective garments produced by a group of workers in an industry.

No. of defective garments	28	32	40	50	58	62	Total
No. of days	5	10	12	20	10	3	60

Find the quartile deviation for the above data and comment about the variation of the data.

Solution

Suppose that variable X represents number of defective garments. The cumulative frequencies of less than type are obtained in Table 6.1.

Table 6.1

x_i	28	32	40	50	58	62	Total
f_i	5	10	12	20	10	3	60
CF(L)	5	15	27	47	57	60	

Now the quartiles are computed as follows:

$$\begin{aligned} \text{Therefore, } Q_1 &= \left[\frac{N+1}{4} \right]^{\text{th}} \text{ value} = \frac{61}{4}^{\text{th}} \text{ value} = 15.25^{\text{th}} \text{ value} \\ &= 15^{\text{th}} + 0.25(16^{\text{th}} - 15^{\text{th}}) \\ &= 32 + 0.25 \times (40 - 32) = 34 \end{aligned}$$

$$\begin{aligned} \text{Therefore, } Q_3 &= \left[3 \times \frac{N+1}{4} \right]^{\text{th}} \text{ value} = 3 \times \frac{61}{4}^{\text{th}} \text{ value} = 45.75^{\text{th}} \text{ value} \\ &= 45^{\text{th}} + 0.75(46^{\text{th}} - 45^{\text{th}}) \\ &= 50 + 0.75 \times (50 - 50) = 50 \end{aligned}$$

$$\text{Therefore, } QD = \frac{Q_3 - Q_1}{2} = \frac{50 - 34}{2} = 8$$

Thus from the given data of 60 days and from quartile deviation, it can be said that the variation in production of number of defective garments is 8 garments.

Example 6.6

The following data are related to the linear density of yarn.

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Calculate quartile deviation for the above data and comment about the variation of the data.

Solution

Suppose that variable X represents linear density of the yarn.

The cumulative frequencies of less than type are obtained in Table 6.2.

Table 6.2

X	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
F	8	12	20	25	22	10	3
CF(L)	8	20	40	65	87	97	100

Now the quartiles are computed as follows:

$$\text{Therefore, } Q_1 = \left[\frac{N}{4} \right]^{\text{th}} \text{ value} = \frac{100}{4}^{\text{th}} \text{ value} = 25^{\text{th}} \text{ value}$$

Comparing 25 with CF(L) class interval containing Q_1 is 13.50–13.75

$$\text{Therefore, } Q_1 = L + \frac{h}{f} \left[\frac{N}{4} - \text{CF}(P) \right] = 13.50 + \frac{0.25}{20} [25 - 20] = 13.5625$$

$$\text{Also Therefore, } Q_3 = \left[3 \times \frac{N}{4} \right]^{\text{th}} \text{ value} = 3 \times \frac{100}{4}^{\text{th}} \text{ value} = 75^{\text{th}} \text{ value}$$

Comparing 75 with CF(L) class interval containing Q_3 is 14.00–14.25

$$\text{Therefore, } Q_3 = L + \frac{h}{f} \left[3 \times \frac{N}{4} - \text{CF}(P) \right] = 14.00 + \frac{0.25}{20} [75 - 65] = 14.1136$$

$$\text{Therefore, } QD = \frac{Q_3 - Q_1}{2} = \frac{14.1136 - 13.5625}{2} = 0.2756$$

Thus, from the given data of 100 linear density tests and from quartile deviation it can be said that the variation in the linear density of the yarn is 0.2756 units.

6.4 Mean deviation

Mean deviation is one of the best measures of dispersion, which gives us an idea about the variation present in the data. The mean deviation about the constant A is defined as follows:

$$\text{Mean deviation} = \frac{\text{Sum of absolute deviations from } A}{\text{Total number of observations}}$$

Where, A is some central value of the data or assumed mean.

This is one of the best measures of dispersion and gives exact variation present in the data as it is based on all values of the data. But this is not

popularly used in real life because common people do not know meaning of absolute deviation (mathematically modulus of the number).

In particular, If X is the variable of the data, then the mean deviation about constant A be denoted by the notation $MD(A)$ and will be defined as follows:

Case I If the data contain only ' n ' observations x_1, x_2, \dots, x_n of the variable X , then the mean deviation about A is defined as follows:

$$MD(A) = \frac{|x_1 - A| + |x_2 - A| + \dots + |x_n - A|}{n} = \frac{\sum |x_i - A|}{n}$$

Where $|x_i - A|$ is called the modulus of $(x_i - A)$ or the absolute deviation of x_i from A and $N = \sum f_i$.

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies, then the mean deviation about A of the data is defined as follows:

$$MD(A) = \frac{f_1|x_1 - A| + f_2|x_2 - A| + \dots + f_k|x_k - A|}{N} = \frac{\sum f_i|x_i - A|}{N}$$

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid points of the class intervals and f_1, f_2, \dots, f_k as the frequencies, then the mean deviation about A of the data is defined as follows:

$$MD(A) = \frac{f_1|x_1 - A| + f_2|x_2 - A| + \dots + f_k|x_k - A|}{N} = \frac{\sum f_i|x_i - A|}{N}$$

If the constant A is mean (\bar{x}) or median or mode, then mean deviation is called the mean deviation about mean or mean deviation about median or mean deviation about the mode.

Note that in this chapter, we will be discussing how to compute mean deviation about mean only with suitable examples and other mean deviations can be computed similarly.

Example 6.7

Ten-count tests made on certain yarn have shown following results.

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Calculate mean deviation about mean using the above data and comment about the variation of the data.

Solution

Suppose that the variable X represents count of the yarn.

Mean deviation about mean is calculated by preparing Table 6.3 as follows:

Table 6.3

Observation no.	x_i	$ x_i - \bar{x} $
1	22.8	0.19
2	23.2	0.21
3	22.9	0.09
4	22.6	0.39
5	23.4	0.41
6	23.0	0.01
7	23.1	0.11
8	23.0	0.01
9	22.9	0.09
10	23.0	0.01
Total	229.9	1.52

$$\bar{X} = \frac{\sum x_i}{n} = \frac{229.9}{10} = 22.99$$

$$MD(\bar{x}) = \frac{\sum |x_i - \bar{x}|}{n} = \frac{1.52}{10} = 0.152$$

Thus from the given data of 10 observations and from mean deviation about mean it can be said that the variation in the count of the yarn is 0.152 units.

Example 6.8

The following data are related to the numbers of defective garments produced by a group of workers in an industry.

No. of defective garments	28	32	40	50	58	62	Total
No. of days	5	10	12	20	10	3	60

Find the mean deviation about mean for the above data and comment about the variation of the data.

Solution

Here, X represents number of defective garments produced by the workers in a day

Preparing table like Table 6.4

Table 6.4

X	f_i	$f_i x_i$	$ x_i - \bar{x} $	$f_i x_i - \bar{x} $
28	5	140	17.1	85.5
32	10	320	13.1	131
40	12	480	5.1	61.2
50	20	1000	4.9	98
58	10	580	12.9	129
62	3	186	16.9	50.7
Total	60	2706		555.4

$$\bar{X} = \frac{\sum f_i \cdot x_i}{N} = \frac{2706}{60} = 45.1$$

$$MD(\bar{x}) = \frac{\sum f_i |x_i - \bar{x}|}{N} = \frac{555.4}{60}$$

$$= 9.2567$$

Thus from the given data of 60 observations and from mean deviation about mean it can be said that the variation in the number of defective garments is 9.2567 units.

Example 6.9

The following data are related to the linear density of yarn.

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Calculate mean deviation about mean from the above data and comment about the variation of the data.

Solution

Suppose that variable X represents linear density of the yarn.

In this method preparing table like Table 6.5

Table 6.5

X	f_i	x_i	$f_i x_i$	$ x_i - \bar{x} $	$f_i x_i - \bar{x} $
13.00–13.25	8	13.125	105	0.7075	5.66
13.25–13.50	12	13.375	160.5	0.4575	5.49
13.50–13.75	20	13.625	272.5	0.2075	4.15
13.75–14.00	25	13.875	346.875	0.0425	1.0625
14.00–14.25	22	14.125	310.75	0.2925	6.435
14.25–14.50	10	14.375	143.75	0.5425	5.425
14.50–14.75	3	14.625	43.875	0.7925	2.3775
Total	100		1383.25	30.6	

$$\bar{X} = \frac{\sum f_i x_i}{N} = \frac{1383.25}{100} = 13.8325$$

$$MD(\bar{x}) = \frac{\sum f_i |x_i - \bar{x}|}{N} = \frac{30.6}{100} = 0.306$$

Thus from the given data of 100 observations and from mean deviation about mean, it can be said that the variation in the linear density of the yarn is 0.306 units.

6.5 Standard deviation

Standard deviation is most popularly used measure of dispersion as it is based on all observations and is simple by definition. It gives idea about the variation present in the data. The Standard deviation is denoted by the notation σ or σ_x and is defined as follows:

$$\sigma_x = \text{Standard deviation of the variable } X = +\sqrt{\sigma_x^2}$$

Where σ_x^2 is called the variance of the variable X . The variance of the variable X can be treated as the measure of variation; but it will have its unit in square, and therefore the standard deviation is defined as the positive square root of the variance and it is treated as the measure of variation. The variance is also called the mean squared deviation. It is also denoted by $V(X)$ and is defined as follows:

$$\text{Variance of } X = V(X) = \frac{\text{Sum of squared deviations from mean}}{\text{Total number of observations}}$$

In particular,

If X is the variable of the data, then mean squared deviation or the variance of the variable X is defined as follows:

Case I If the data contain only ' n ' observations x_1, x_2, \dots, x_n of the variable X then

$$V(X) = \sigma_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} = \frac{\sum (x_i - \bar{x})^2}{n}$$

Simplifying,

$$V(X) = \sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2 - \bar{x}^2}{n}$$

where $(x_i - \bar{x})^2$ is called the squared deviation x_i from \bar{x} and $N = \sum f_i$.

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies, then

$$V(X) = \sigma_x^2 = \frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \dots + f_k(x_k - \bar{x})^2}{N} = \frac{\sum f_i(x_i - \bar{x})^2}{N}$$

Simplifying,

$$V(X) = \sigma_x^2 = \frac{\sum f_i(x_i - \bar{x})^2}{N} = \frac{\sum f_i x_i^2}{N} - \bar{x}^2$$

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid points of the class intervals and f_1, f_2, \dots, f_k as the frequencies, then

$$V(X) = \sigma_x^2 = \frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \dots + f_k(x_k - \bar{x})^2}{N} = \frac{\sum f_i(x_i - \bar{x})^2}{N}$$

Simplifying,

$$V(X) = \sigma_x^2 = \frac{\sum f_i(x_i - \bar{x})^2}{N} = \frac{\sum f_i x_i^2}{N} - \bar{x}^2$$

Note that,

By definition variance is always positive.

Variance and hence the standard deviation is not affected by the change of origin, but it is affected by the change of scale.

that is if we define $U = X - A$, then $\sigma_x^2 = \sigma_u^2$

if we define $U = \frac{X - A}{h}$, then $\sigma_x^2 = h^2 \sigma_u^2$

If \bar{x}_1 and σ_1^2 are the AM and variance of first data n_1 of observations and \bar{x}_2 and σ_2^2 are the AM and variance of second data of n_2 observations, then the mean of the combined data of $n_1 + n_2$ observations can be given as follows

$$\text{Combined mean} = \bar{X} = \frac{n_1 \cdot \bar{X}_1 + n_2 \cdot \bar{X}_2}{n_1 + n_2}$$

Also combined variance is denoted by σ^2 and can be given as follows:

$$\sigma^2 = \frac{n_1 \times (d_1^2 + \sigma_1^2) + n_2 \times (d_2^2 + \sigma_2^2)}{n_1 + n_2}$$

Where, $d_1 = \bar{x}_1 - \bar{x}$ and $d_2 = \bar{x}_2 - \bar{x}$

Computation of the standard deviation

As discussed in the Chapter 4, standard deviation can also be computed by two different methods.

1. Direct method

Case I If the data contain only ‘ n ’ observations x_1, x_2, \dots, x_n of the variable X , then the standard deviation is calculated directly using the formula and by preparing a table (see Table 6.6):

$$\bar{x} = \frac{\sum x_i}{n}$$

Table 6.6

X	x_i^2	$V(X) = \sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}$
x_1	x_1^2	
x_2	x_2^2	
....	$= \frac{\sum x_i^2}{n} - \bar{x}^2$
x_n	x_n^2	
$\sum x_i$	$\sum x_i^2$	$\sigma_x = +\sqrt{\sigma_x^2}$

Case II If the data are in the ungrouped frequency distribution form with x_1, x_2, \dots, x_k as the possible values and f_1, f_2, \dots, f_k as the frequencies,

then the standard deviation is calculated directly using the formula and by preparing a table (see Table 6.7).

Table 6.7

X	f_i	$f_i x_i$	$f_i x_i^2$	$\bar{X} = \frac{\sum f_i \cdot x_i}{N}$
x_1	f_1	$f_1 x_1$	$f_1 x_1^2$	
x_2	f_2	$f_2 x_2$	$f_2 x_2^2$	
....	$\sigma_x^2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} = \frac{\sum f_i x_i^2}{N} - \bar{x}^2$
x_k	f_k	$f_k x_k$	$f_k x_k^2$	
Total	$N = \sum f_i$	$\sum f_i x_i$	$\sum f_i x_i^2$	$\sigma_x = +\sqrt{\sigma_x^2}$

Case III If the data are in the grouped frequency distribution form with x_1, x_2, \dots, x_k as the mid points of the class intervals and f_1, f_2, \dots, f_k as the frequencies, then also the standard deviation is calculated directly using the formula and by preparing a table (see Table 6.5).

2. Indirect method

This method is generally used when the data under study is in the frequency distribution form. Particularly if the data are in the ungrouped frequency distribution form the standard deviation is calculated indirectly by transforming the variable X into another variable U . In this case the variable U is defined by the change of origin as $U = X - A$ and standard deviation is calculated indirectly using the formula and by preparing a table (see Table 6.8).

Table 6.8

X	f_i	u_i	$f_i u_i$	$f_i u_i^2$	$\bar{u} = \frac{\sum f_i u_i}{N}$
x_1	f_1	u_1	$f_1 u_1$	$f_1 u_1^2$	
x_2	f_2	u_2	$f_2 u_2$	$f_2 u_2^2$	
....	$\sigma_u^2 = \frac{\sum f_i (u_i - \bar{u})^2}{N} = \frac{\sum f_i u_i^2}{N} - \bar{u}^2$
x_k	f_k	u_k	$f_k u_k$	$f_k u_k^2$	
Total	$N = \sum f_i$		$\sum f_i u_i$	$\sum f_i u_i^2$	

Now,

$$\sigma_x^2 = \sigma_u^2$$

Further, if the data are in the grouped frequency distribution form the standard deviation is calculated indirectly by transforming the variable X into another variable U . In this case, the variable U is defined by the change of origin and the scale method as $U = \frac{X - A}{h}$ and standard deviation is calculated indirectly using the formula and by preparing a table (see Table 6.6). But in case of grouped frequency distribution, x_1, x_2, \dots, x_k are the mid-points of the class intervals f_1, f_2, \dots, f_k are the frequencies and h is the class width. Where,

$$\sigma_x^2 = h^2 \sigma_u^2$$

Example 6.10

Ten-count tests made on certain yarn have shown following results.

22.8, 23.2, 22.9, 22.6, 23.4, 23.0, 23.1, 23.0, 22.9, 23.0

Calculate standard deviation using the above data and comment about the variation of the data.

Solution

Suppose that the variable X represents count of the yarn.

Standard deviation is calculated by preparing the Table 6.9 as follows:

Table 6.9

Observation no.	x_i	x_i^2	
1	22.8	519.84	$\bar{X} = \frac{\sum x_i}{n} = \frac{229.9}{10} = 22.99$
2	23.2	538.24	
3	22.9	524.41	$\sigma_x^2 = V(X) = \frac{\sum x_i^2}{n} - \bar{x}^2$
4	22.6	510.76	$= \frac{5285.83}{10} - 22.99^2$
5	23.4	547.56	$= 0.0429$
6	23.0	529	
7	23.1	533.61	$\sigma_x = +\sqrt{\sigma}$
8	23.0	529	
9	22.9	524.41	$\sigma_x = 0.2071$
10	23.0	529.00	
Total	229.9	5285.83	

Thus from the given data of 10 observations and from the standard deviation it can be said that the variation in the count of the yarn is 0.2071 units.

Example 6.11

The following data are related to the numbers of defective garments produced by a group of workers in an industry.

No. of defective garments	28	32	40	50	58	62	Total
No. of days	5	10	12	20	10	3	60

Find the standard deviation for the above data and comment about the variation of the data.

Solution

Here, X represents number of defective garments produced by the workers in a day

1. Direct method

Preparing table like Table 6.10

Table 6.10

X	f_i	$f_i x_i$	$f_i x_i^2$	$\bar{X} = \frac{\sum f_i x_i}{N} = \frac{2706}{60} = 45.1$
28	5	140	3920	$\sigma_x^2 = V(X) = \frac{\sum f_i x_i^2}{N} - \bar{x}^2 = \frac{128532}{60} - 45.1^2$
32	10	320	10240	$= 108.19$
40	12	480	19200	$\sigma_x = +\sqrt{\sigma^2}$
50	20	1000	50000	$\sigma_x = 10.4014$
58	10	580	33640	
62	3	186	11532	
Total	60	2706	128532	

Thus from the given data of 60 observations and from the standard deviation, it can be said that the variation in the number of defective garments is 10.4014 units.

2. Indirect method

Preparing table like Table 6.11

Table 6.11

X	f_i	$u_i = x_i - 50$	$f_i u_i$	$f_i u_i^2$
28	5	-22	-110	2420
32	10	-18	-180	3240
40	12	-10	-120	1200
50	20	0	0	0
58	10	8	80	640
62	3	12	36	432
Total	60		-294	7932

$$\bar{u} = \frac{\sum f_i u_i}{N} = \frac{-294}{60} = -4.9$$

$$\sigma_u^2 = V(U) = \frac{\sum f_i u_i^2}{N} - \bar{u}^2 = \frac{7932}{60} - (-4.9)^2 = 108.19$$

Now,

$$\sigma_x^2 = \sigma_u^2 \quad \text{Therefore, } \sigma_x = +\sqrt{\sigma_x^2} \quad \text{Therefore, } \sigma_x = 10.4014$$

Example 6.12

The following data are related to the linear density of yarn.

Linear density	13.00–13.25	13.25–13.50	13.50–13.75	13.75–14.00	14.00–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Calculate standard deviation for the above data and comment about the variation of the data.

Solution

Suppose that variable X represents linear density of the yarn.

In this method preparing table like Table 6.12

1. Direct method

Table 6.12

X	f_i	x_i	$f_i x_i$	$f_i x_i^2$
13.00–13.25	8	13.125	105.0000	1378.1250
13.25–13.50	12	13.375	160.5000	2146.6880
13.50–13.75	20	13.625	272.5000	3712.8130
13.75–14.00	25	13.875	346.8750	4812.8910
14.00–14.25	22	14.125	310.7500	4389.3440
14.25–14.50	10	14.375	143.7500	2066.4060
14.50–14.75	3	14.625	43.8750	641.6719
Total	100		1383.2500	19147.9389

$$\bar{X} = \frac{\sum f_i \cdot x_i}{N} = \frac{1383.25}{100} = 13.8325$$

$$\begin{aligned}\sigma_x^2 &= V(X) = \frac{\sum f_i x_i^2}{N} - \bar{x}^2 = \frac{19147.9389}{100} - 13.8325^2 \\ &= 0.1413\end{aligned}$$

$$\sigma_x = +\sqrt{\sigma_x^2} \text{ Therefore, } \sigma_x = 0.3759$$

Thus from the given data of 100 observations and from the standard deviation it can be said that the variation in the linear density of the yarn is 0.3759 units.

In this method preparing table like Table 6.13

2. Indirect method

Table 6.13

X	f_i	x_i	$u_i = \frac{x_i - 13.875}{0.25}$	$f_i x_i$	$f_i x_i^2$
13.00–13.25	8	13.125	-3	-24	72
13.25–13.50	12	13.375	-2	-24	48
13.50–13.75	20	13.625	-1	-20	20
13.75–14.00	25	13.875	0	0	0
14.00–14.25	22	14.125	1	22	22
14.25–14.50	10	14.375	2	20	40
14.50–14.75	3	14.625	3	9	27
Total	100			-17	229

$$\bar{u} = \frac{\sum f_i \cdot u_i}{N} = \frac{-17}{100} = -0.17$$

$$\sigma_u^2 = V(U) = \frac{\sum f_i u_i^2}{N} - \bar{u}^2 = \frac{299}{100} - (-0.17)^2 = 2.2611$$

Now,

$$\sigma_x^2 = h^2 \sigma_u^2 = 0.1413 \text{ Therefore, } \sigma_x = +\sqrt{\sigma_x^2} \text{ Therefore, } \sigma_x = 0.3759$$

6.6 Relative measures of dispersion

Up to this stage, we have seen that the range, quartile deviation, mean deviation, and standard deviation are the main measures of variation. These measures are also called the absolute measures of variation as they all depend on the units of measurement. Hence, comparison of the variation of two or more data sets on the basis of above measures of dispersion is

not possible if their units of measurements are different. Therefore, another measures of dispersion are defined, which are called the relative measures of dispersion. Relative measures of dispersion are independent of the units of measurement, and therefore variability of any two data sets can be easily compared with each other. Following are the different relative measures of dispersion:

1. Coefficient of range
2. Coefficient of quartile deviation
3. Coefficient of mean deviation
4. Coefficient of variation

These relative measures of dispersion are defined as follows

$$\text{Coefficient of range} = \frac{X_{\max} - X_{\min}}{X_{\max} + X_{\min}}$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$\text{Coefficient of mean deviation about mean} = \frac{MD(\bar{x})}{\bar{x}}$$

If coefficient of mean deviation is multiplied by 100, then it is expressed in percentage and the corresponding result is sometimes also called the percentage mean deviation (PMD)

$$\text{PMD} = \frac{MD(\bar{x})}{\bar{x}} \times 100$$

The coefficient of variation is also expressed in percentage and is well known as the CV%. This CV% is generally used for comparing the consistency of two or more data.

The data having less CV% is having less variation, and hence it is said to be more consistent among all.

$$\text{CV\%} = \frac{\sigma_x}{\bar{x}} \times 100$$

Example 6.13

Following are the results of the fabric strength (in 10 gm) obtained from the samples of two different fabrics.

Fabric A	22.0	21.5	22.8	21.0	23	20.9	21.6	22.0	22.8	21.2
Fabric B	22.3	21.6	22.0	22.1	22.0	22.3	21.8	21.8	21.6	21.8

Calculate CV% for both fabrics and state which fabric is more consistent in terms of the strength.

Find the mean and standard deviation of the strength if results of the strength of two different fabrics are taken together.

Solution

Here,

Variable X_1 is the strength of the fabric of Type A.

Variable X_2 is the strength of the fabric of Type B.

Preparing the Table 6.14 as follows:

Tables 6.14

x_{1i}	x_{1i}^2	x_{2i}	x_{2i}^2
22.0	484.00	22.3	497.29
21.5	462.25	21.6	466.56
22.8	519.84	22.0	484.00
21.0	441.00	22.1	488.41
23.0	529.00	22.0	484.00
20.9	436.81	22.3	497.29
21.6	466.56	21.8	475.24
22.0	484.00	21.8	475.24
22.8	519.84	21.6	466.56
21.2	449.44	21.8	475.24
218.8	4792.74	219.3	4809.83

For fabric A

$$\bar{X} = \frac{\sum x_i}{n} = \frac{218.8}{10} = 21.88$$

$$\begin{aligned}\sigma_x^2 &= V(X) = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{4792.74}{10} - 21.88^2 \\ &= 0.5396\end{aligned}$$

$$\sigma_x = +\sqrt{\sigma_x^2} \quad \sigma_x = 0.7346$$

$$CV\% = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{0.7346}{21.88} \times 100 = 3.3574\%$$

For fabric B

$$\bar{X} = \frac{\sum x_i}{n} = \frac{219.3}{10} = 21.93$$

$$\begin{aligned}\sigma_x^2 &= V(X) = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{4809.83}{10} - 21.93^2 \\ &= 0.0581\end{aligned}$$

$$\sigma_x = +\sqrt{\sigma_x^2} \quad \sigma_x = 0.2410$$

$$CV\% = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{0.2410}{21.93} \times 100 = 1.0989\%$$

Here, it is clear that $CV\%(A) > CV\%(B)$. That is, variation in the strength is more for fabric A than that of fabric B. Hence, Fabric B is more consistent in terms of the strength.

Further,

Combining results of two fabrics combined average strength is

$$\bar{x} = \frac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2}{n_1 + n_2} = \frac{10 \times 21.88 + 10 \times 21.93}{10 + 10} = 21.905$$

Also, combined variance is denoted by σ^2 and can be given as follows:

$$\sigma^2 = \frac{n_1 \times (d_1^2 + \sigma_1^2) + n_2 \times (d_2^2 + \sigma_2^2)}{n_1 + n_2}$$

$$\text{where, } d_1 = \bar{x}_1 - \bar{x} = 21.88 - 21.905 = -0.025$$

$$d_2 = \bar{x}_2 - \bar{x} = 21.93 - 21.905 = 0.025$$

$$\begin{aligned}\sigma^2 &= \frac{n_1 \times (d_1^2 + \sigma_1^2) + n_2 \times (d_2^2 + \sigma_2^2)}{n_1 + n_2} \\ &= \frac{10 \times (-0.025^2 + 0.5396) + 10 \times (0.025^2 + 0.0581)}{10 + 10} = 0.2995\end{aligned}$$

Therefore,

$$\text{Combined standard deviation} = \sigma = 0.5472$$

6.7 Exercise

- What is dispersion? What are different measures of dispersion? Describe any one.
- What is dispersion? What are different measures of dispersion? What is the difference between absolute and relative measures of dispersion?
- Following is the frequency distribution of number of weaving defects observed in 50 pieces of fabric:

No. of defects	0	1	2	3	4	5
No. of pieces	15	20	9	3	2	1

Find mean deviation about mean and comment on it.

- Five strength tests each carried out on the two fabrics have shown following results:

A	15.2	15.5	15.0	15.4	15.6
B	15.0	14.8	13.6	15.5	15.2

Find CV% for both and decide which fabric is more consistent?

- Find mean, variance, and CV% for the following data of number of thick spots observed in 100 pieces of fabric and comment on them.

No. of thick spots	0	1	2	3	4	5
No. of pieces	30	25	20	10	10	5

- Following are the results of count tests made on the production of two shifts:

Shift I	11.7	12.2	12.0	11.6	10.9	13.0	11.0	12.8	11.5	12.0
Shift II	12.2	11.6	12.0	12.1	12.0	12.3	11.8	11.6	11.8	12.0

Using CV% decide which production is more consistent.

- A spinning master has made 10-count tests each on the yarn spun by two different ring frames. The results are as follows:

Count R/F-A	12.2	11.8	12.3	12.4	11.9	11.9	11.5	12.0	12.1	11.9
Count R/F-B	12.5	11.2	12.8	11.5	12.6	12.4	11.6	11.4	12.7	11.3

Decide which ring frame is more consistent in terms of count.

8. The time taken by an operator to replace bobbin on a spinning frame are as follows:

Time (in s)	15.1–15.2	15.3–15.4	15.5–15.6	15.7–15.8	15.9–16.0	16.1–16.2	16.3–16.4
No. of occasions	1	7	17	22	15	7	2

Calculate quartile deviation for the above data and comment on it.

9. Find AM and Standard deviation for the following data of breaking strength (in some units).

Breaking strength	390–395	395–400	400–405	405–410	410–415	415–420	420–425
No. of observations	10	20	35	50	40	30	15

10. Calculate mean deviation about mean for the following data and comment on it.

CI	10–11	11–12	12–13	13–14	14–15
Frequency	5	10	25	10	5

11. Following is the frequency distribution of number of defects on pieces of fabric.

No. of defects	0	1	2	3	4	5
No. of pieces	2	5	8	7	6	3

Compute quartile deviation.

12. Following are the results of number of workers absent in a textile firm. Compute the mean deviation about mean for the following frequency distribution.

No. of absent workers	0	1	2	3	4	5	6
No. of days	60	80	50	30	20	15	10

13. Following are the results of sale of two types of knitted garments:

Type A	10	12	11	14	10	11	15	12	14	13
Type B	12	10	17	13	12	10	10	18	14	15

Find CV% and decide sale of which type of garment is more consistent.

14. Following frequency distribution represents the number of accidents in a month.

Compute AM and variance of the data.

No. of accidents	0	1	2	3	4	5
No. of days	3	5	7	8	5	2

15. Following data gives number of defective articles produced by a batch of 10 workers during day and night shift of the week:

No. of defectives (day shift)	122	120	125	115	118	120	125
No. of defectives (night shift)	132	110	128	140	115	115	130

- i) Find AM for each shift and comment on them.
 - ii) Which shift is more consistent in terms of number of defectives produced?
16. The average and standard deviation of the salary of 50 nontechnical staff of a mill are Rs. 1200 and 200, respectively. The corresponding figures for 75 technical staff are Rs. 1500 and 250, respectively. What are the average and standard deviation of the salary of all employees in the mill?
17. Following are the results of seven count tests made on a yarn.

30.0	31.5	29.5	30.5	30.0	30.5	31.0
------	------	------	------	------	------	------

Find AM, standard deviation and CV%.

18. Following data represents number of end breaks recorded on two ring frames at ten occasions.

R/F-I	12	10	8	15	9	11	14	15	8	10
R/F-II	10	11	12	9	12	9	11	12	11	10

Find which ring frame is more consistent in terms of end breaks.

19. Calculate mean deviation about mean for the following data:

CI	10–13	13–16	16–19	19–22	22–25
Frequency	5	10	25	10	5

20. Calculate the mean deviation about mean using following data and comment on it.

Diameter (mm)	130	135	140	145	150
No. of screws	5	10	25	10	5

21. Following are the results of single yarn strength (in some units) obtained for two different types of yarns.

Yarn A	12	11.5	12.8	11.0	13.0	10.9	11.6	12	12.2	11.7
Yarn B	12.2	11.6	12.0	12.1	12.0	12.3	11.8	11.6	11.8	12.0

Calculate CV% for both yarns and decide which yarn is more consistent in terms of single yarn strength.

22. Following data represents the strength of the fabric (in some units).

Strength	55–58	58–61	61–64	64–67	67–70
No. of samples	12	17	23	18	11

Find the quartile deviation for the above data and comment on it.

Skewness and kurtosis

7.1 Introduction

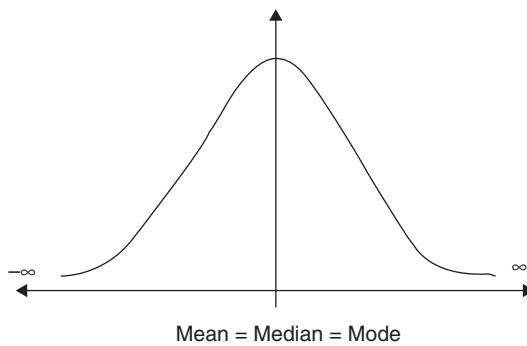
Sometimes in statistics it is necessary to study the shape of the frequency curve of a frequency distribution. Skewness and kurtosis are the two terms that help in understanding this.

7.2 Skewness

Skewness is the property related to the frequency distribution. Using skewness, we can study the distribution of the frequencies around the AM.

7.3 Symmetric frequency distribution

A frequency distribution is called the symmetric frequency distribution, if the frequencies are equally distributed on both sides of mean. The symmetric frequency distributions always satisfy the relationship $AM = \text{median} = \text{mode}$. That is, the mean, mode, and median are the same for a symmetric frequency distribution. The graph of the symmetric frequency distribution is of the form as shown in Fig. 7.1.



7.4 Skewed frequency distribution

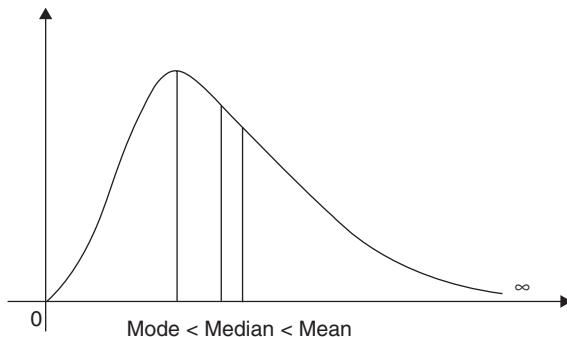
A frequency distribution is called the skewed frequency distribution, if the frequencies are not equally distributed on both sides of the mean. That is, the frequency distribution, which is not symmetric (asymmetric) is called the skewed frequency distribution or it is said to have skewness. The two different types of skewness are positive and negative skewness.

Positive skewness or positively skewed frequency distribution

If the larger part of the frequencies is distributed below mean, then the frequency distribution is said to have positive skewness or it is called the positively skewed frequency distribution.

The graph of the positively skewed frequency distribution is of the following form. Positively skewed frequency distributions always satisfy the relationship

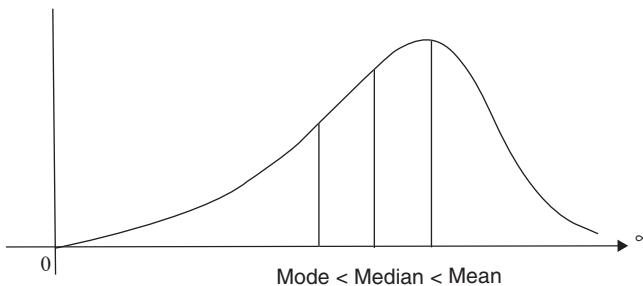
Mode < median < mean. That is, for a positively skewed frequency distribution mean is largest, mode is smallest, and median is in middle. The graph of the positively skewed frequency distribution is of the form as shown in Fig. 7.2.



7.2

Negative skewness or negatively skewed frequency distribution

If the larger part of the frequencies is distributed above mean, then the frequency distribution is said to have negative skewness or it is called the negatively skewed frequency distribution. Negatively skewed frequency distributions always satisfy the relationship mode > median > mean. That is, for a negatively skewed frequency distribution, mean is smallest, mode is largest and median is in middle. The graph of the negatively skewed frequency distribution is of the form as shown in Fig. 7.3.



7.3

7.5 Measures of skewness

Any value or any term, which gives idea regarding the skewness of the frequency distribution, is called the measure of skewness. Karl Pearson's coefficients of skewness and Bowley's coefficient of skewness are the two main measures of skewness.

Karl Pearson's coefficient of skewness

This is the measure of skewness based on the mean, mode, and median; and it is defined as follows:

Karl Pearson's coefficient of skewness

$$\begin{aligned}
 &= \frac{(\text{mean} - \text{Mode})}{\text{Standard deviation}} && \text{if Mode is determinable} \\
 &= \frac{3(\text{mean} - \text{Median})}{\text{Standard deviation}} && \text{if Mode is not determinable}
 \end{aligned}$$

Note that,

If mode is not determinable, then $(\text{mean} - \text{mode})$ is replaced by $3(\text{mean} - \text{median})$ using the relationship “ $\text{mean} - \text{mode} = 3(\text{mean} - \text{median})$,” which is called the empirical relationship for the asymmetric frequency distribution.

Bowley's coefficient of skewness

This is the measure of skewness based on the quartiles, and it is defined as follows:

$$\text{Bowley's coefficient of skewness} = \frac{Q_3 + Q_1 - 2 \times Q_2}{Q_3 - Q_1}$$

This coefficient of skewness is useful when mean is not determinable. That is, when the classes of the frequency distribution are open-ended classes.

7.6 Interpretation of the coefficient of skewness

1. The frequency distribution is called the positively skewed frequency distribution, if the value of the coefficient of skewness is positive.
2. The frequency distribution is called the negatively skewed frequency distribution, if the value of the coefficient of skewness is negative.
3. The frequency distribution is called the symmetric frequency distribution, if the value of the coefficient of skewness is zero.

Example 7.1

The following data are related to the linear density of yarn.

Linear density	13.0–13.25	13.25–13.50	13.50–13.75	13.75–14.0	14.0–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Calculate Karl Pearson and Bowley's coefficient of skewness using the above data and comment on it.

Solution

Here, variable X represents linear density of the yarn.

We have,

$$KP \text{ coefficient of skewness} = \frac{\text{mean} - \text{mode}}{\text{SD}}$$

$$\text{Bowley's coefficient of skewness} = \frac{Q_3 + Q_1 - 2 \times Q_2}{Q_3 - Q_1}$$

Table 7.1

CI	f_i	x_i	$f_i x_i$	$f_i x_i^2$	$CF(L)$
13.00–13.25	8	13.125	105.0000	1378.1250	8
13.25–13.50	12	13.375	160.5000	2146.6880	20
13.50–13.75	20	13.625	272.5000	3712.8130	40
13.75–14.00	25	13.875	346.8750	4812.8910	65
14.00–14.25	22	14.125	310.7500	4389.3440	87
14.25–14.50	10	14.375	143.7500	2066.4060	97
14.50–14.75	3	14.625	43.8750	641.6719	100
Total	100		1383.2500	19147.9389	

$$\bar{X} = \frac{\sum f_i \cdot x_i}{N} = \frac{1383.25}{100} = 13.8325$$

$$\sigma_x^2 = V(X) = \frac{\sum f_i x_i^2}{N} - \bar{x}^2 = \frac{19147.9389}{100} - 13.8325^2$$

$$\sigma_x = +\sqrt{\sigma_x^2} \quad \text{Therefore, } \sigma_x = 0.3759$$

Here the class interval 13.75–14.00 has the maximum frequency; hence, the modal class interval is 13.75–14.00.

$$\begin{aligned}\text{Mode} &= L + h \left[\frac{f_m + f_o}{2f_m - f_l - f_o} \right] \\ &= 13.75 + 0.25 \left[\frac{25 - 20}{2 \times 25 - 22 - 20} \right] \\ &= 13.9063\end{aligned}$$

$$\text{Thus, } K.P \text{ coefficient of skewness} = \frac{\text{mean} - \text{mode}}{\text{std.dev.}} = \frac{13.8322 - 13.9063}{0.3759} = -0.1963$$

Also, the quartiles are computed as follows:

$$\text{Therefore, } Q_1 = \left[\frac{N}{4} \right]^{\text{th}} \text{ value} = \frac{100}{4}^{\text{th}} \text{ value} = 25^{\text{th}} \text{ value}$$

Comparing 25 with CF(L) class interval containing Q_1 is 13.50–13.75

$$\text{Therefore, } Q_1 = L + \frac{h}{f} \left[\frac{N}{4} - \text{CF}(P) \right] = 13.50 + \frac{0.25}{20} [25 - 20] = 13.5625$$

$$\text{Therefore, } Q_2 = \left[\frac{N}{2} \right]^{\text{th}} \text{ value} = \frac{100}{2}^{\text{th}} \text{ value} = 50^{\text{th}} \text{ value}$$

Comparing 50 with CF(L) class interval containing Q_1 is 13.75–14.00

$$\text{Therefore, } Q_2 = L + \frac{h}{f} \left[\frac{N}{2} - \text{CF}(P) \right] = 13.75 + \frac{0.25}{20} [50 - 40] = 13.85$$

$$\text{Therefore, } Q_3 = \left[3 \times \frac{N}{4} \right]^{\text{th}} \text{ value} = 3 \times \frac{100}{4}^{\text{th}} \text{ value} = 75^{\text{th}} \text{ value}$$

Comparing 75 with CF(L) class interval containing Q_3 is 14.00–14.25

$$\text{Therefore, } Q_3 = L + \frac{h}{f} \left[3 \times \frac{N}{4} - \text{CF}(P) \right] = 14.00 + \frac{0.25}{22} [75 - 65] = 14.1136$$

$$\begin{aligned}\text{Bowley's coefficient of Skewness} &= \frac{Q_3 + Q_1 - 2 \times Q_2}{Q_3 - Q_1} = \frac{14.1136 + 13.5625 - 2 \times 13.85}{14.1136 - 13.5625} \\ &= -0.0432\end{aligned}$$

Thus, from the values of the Karl Pearson and Bowley's coefficient of skewness, it can be concluded that the given frequency distribution is negatively skewed.

7.7 Kurtosis

Kurtosis is the property related to the peakness of the frequency curve of the frequency distribution. Leptokurtic, mesokurtic, and platykurtic are the three types of the kurtosis according to the peakness of the frequency curve.

Leptokurtic frequency distribution

If the curve of the frequency distribution is very highly peaked, then the frequency distribution is called the leptokurtic frequency distribution.

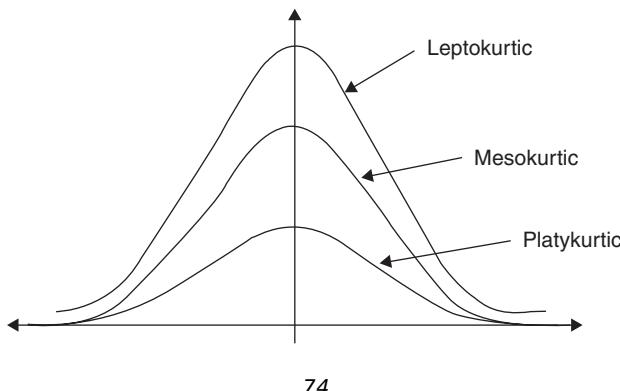
Mesokurtic frequency distribution

If the curve of the frequency distribution is not very highly peaked or not very less peaked, that is it has normal peakness, then the frequency distribution is called the mesokurtic frequency distribution.

Platykurtic frequency distribution

If the frequency curve of the frequency distribution has very low peakness, then the frequency distribution is called the platykurtic frequency distribution.

Above three types of kurtosis can be shown graphically in Fig. 7.4.



7.8 Measures of kurtosis

Any value or any term, which gives idea regarding the peakness (kurtosis) of frequency distribution, is called the measure of kurtosis. Coefficient of kurtosis based on moments is the main measure of kurtosis. It is defined as follows

$$\text{Coefficient of Kurtosis} = \frac{\mu_4}{\mu_2^2} - 3$$

Where, μ_2 is second central moment and μ_4 is fourth central moment. The moment is the generalized concept of mean and variance. Raw moments and the central moments are the two different types of the moments which are defined as follows:

$$\begin{aligned} r^{\text{th}} \text{ raw moment} &= \mu_r' = \frac{\sum_1^n x_i^r}{n} && \text{if } n \text{ observations} \\ &= \frac{\sum_1^k f_i \cdot x_i^r}{N} && \text{if frequency distribution} \end{aligned}$$

$$\begin{aligned} r^{\text{th}} \text{ central moment} &= \mu_r = \frac{\sum_1^k (x_i - \bar{x})^2}{n} && \text{if } n \text{ observations} \\ &= \frac{\sum_1^k f_i \cdot (x_i - \bar{x})^2}{N} && \text{if frequency distribution} \end{aligned}$$

Note that, here $r = 1, 2, \dots$

Also, it can be easily proved that the relationship between central moments and the raw moments are as follows:

$$\begin{aligned}\mu_1 &= 0 \\ \mu_2 &= \mu'_2 - \mu'^2_1 \\ \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1\end{aligned}$$

Properties of coefficient of kurtosis

If the coefficient of kurtosis is positive, then the frequency distribution is leptokurtic.

If the coefficient of kurtosis is negative, then the frequency distribution is platykurtic.

If the coefficient of kurtosis is zero, then the frequency distribution is mesokurtic.

Example 7.2

The following data are related to the linear density of yarn.

Linear density	13.0–13.25	13.25–13.50	13.50–13.75	13.75–14.0	14.0–14.25	14.25–14.50	14.50–14.75
No. of tests	8	12	20	25	22	10	3

Calculate coefficient of kurtosis using the above data and comment on it.

Solution

Here, Variable X represents linear density of the yarn.

Table 7.2

CI	f_i	x_i	$f_i x_i$	$f_i x_i^2$	$f_i x_i^3$	$f_i x_i^4$
13.00–13.25	8	13.125	105	1378.125	18087.89	237403.6
13.25–13.50	12	13.375	160.5	2146.688	28711.95	384022.3
13.50–13.75	20	13.625	272.5	3712.813	50587.07	689248.8
13.75–14.00	25	13.875	346.875	4812.891	66778.86	926556.6
14.00–14.25	22	14.125	310.75	4389.344	61999.48	875742.7
14.25–14.50	10	14.375	143.75	2066.406	29704.59	427003.5
14.50–14.75	3	14.625	43.875	641.6719	9384.451	137247.6
Total	100		1383.25	19147.94	265254.3	3677225.1

Now the raw moments are as follows:

$$\mu'_1 = \frac{1383.25}{100} = 13.8325$$

$$\mu'_2 = \frac{19147.94}{100} = 191.4794$$

$$\mu'_3 = \frac{265254.3}{100} = 2652.543$$

$$\mu'_4 = \frac{3677225.1}{100} = 36772.25$$

$$\text{Therefore, } \mu_2 = \mu'_2 - \mu'^2_1 = 0.1413$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 = 0.0687$$

$$\text{Coefficient of Kurtosis} = \frac{\mu_4}{\mu_2^2} - 3 = \frac{0.0687}{0.1413^2} - 3 = 0.4409$$

Thus, from the value of coefficient of kurtosis, we can say that the given frequency distribution is leptokurtic that is a highly peaked curve.

7.9 Exercise

- What is skewness? State the measure of skewness based on the quartiles. Write its properties.
- What is kurtosis? What are its types? How is it measured?
- Find Bowley's coefficient of skewness for the following data of number of workers absent in a textile mill and comment on it:

No. of absent workers	0	1	2	3	4	5	6
No. of days	60	80	50	30	20	15	10

- Calculate Karl Pearson's coefficient of skewness for the following data and comment on it.

Salary	4000–5000	5000–6000	6000–7000	7000–8000
No. of employees	12	45	94	74
Salary	8000–9000			
No. of employees	32			

5. The distribution of fibers chosen at random according to their length is as follows:

Length (in cm)	8.1–8.2	8.3–8.4	8.5–8.6	8.7–8.8	8.9–9.0	9.1–9.2	9.3–9.4
No. of fibers	3	8	24	32	26	6	1

Find Karl Pearson's coefficient of skewness and write conclusion.

6. Compute Bowley's coefficient of skewness for following data of heights of students and comment on it.

Height (inch)	50–55	55–60	60–65	65–70
No. of students	14	25	40	11

7. Calculate Bowley's coefficient of skewness for the following data and comment on it.

Daily wages	40–50	50–60	60–70	70–80	80–90
No. of workers	4	10	25	18	3

8. Compute Bowley's coefficient of skewness for the following data and comment on it.

No. of printing mistakes	0	1	2	3	4	5	6
No. of Pages	5	21	25	20	15	10	4

9. Calculate coefficient of skewness based on quartiles for the following data of "linear density" and comment on it.

Linear density (in text)	59–61	61–63	63–65	65–67	67–69
No. of observations	4	30	45	15	6

10. Following data represents breaking strength of certain thread

Breaking strength	390–394	395–399	400–404	405–409	410–414
No. of Samples	10	20	35	50	15

Calculate Karl Pearson's coefficient of skewness and comment on it.

11. Find Karl Pearson's coefficient of skewness and comment on it.

Production of yarn	500–550	550–600	600–650	650–700	700–750
No. of days	5	10	15	10	5

Correlation and regression

8.1 Introduction

Sometimes in statistics, the data under study are related to two different variables, that is, information related to two different variables is collected from each and every individual of the population or the sample. From such data, it may be of interest to find out whether there is any mutual relation between the two variables of the data, or is it possible to find value of any one variable on the basis of the data if the value of another variable is known. Correlation and regression analysis are the answers to the above questions.

8.2 Bivariate data

The data related to two different variables or the collection of paired observation is called the bivariate data. If X and Y are the two different variables under study, then the bivariate data corresponding to the variables X and Y is given as follows:

$$\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$$

Thus, there are “ n ” paired observations in the bivariate data. Correlation analysis and regression analyses are the two different ways of studying the bivariate data.

8.3 Correlation analysis

When the bivariate data are studied for the mutual relationship between the two different variables, then it is called the correlation analysis.

Correlation

If the change in the value of one variable causes change in the value of the other variable, then the two variables are said to have correlation or are correlated.

Positive correlation and negative correlation are the two different types of the correlation according to the direction of the change in the values of the two variables.

Positive correlation

If the change in the values of both the variables is in the same direction, then the two variables have positive correlation. That is increase/decrease in the values of one variable causes increase/decrease in the values of another variable.

Negative correlation

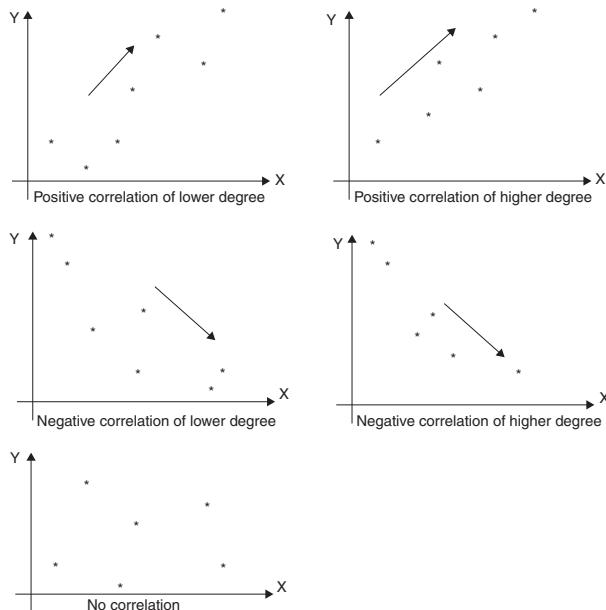
If the change in the values of both the variables is in the opposite direction, then the two variables have negative correlation. That is, increase/decrease in the values of one variable causes decrease/increase in the values of another variable.

Measure of correlation

Scatter diagram and Karl Pearson's coefficient of correlation are the two different ways of studying or measuring the correlation between the two variables.

Scatter diagram

The graphical representation of the bivariate data is called the scatter diagram. It helps in studying the type and the degree of correlation between the two variables. Following are some of the typical scatter diagrams representing the different types and the degree of the correlation as shown in Fig. 8.1.



8.1

Note that, using scatter diagram, we can decide whether the two variables under study are correlated, what is the type of correlation and what is its degree. But, we do not get exact idea about the correlation between any two variables, and hence we cannot use it for comparison. Therefore, it is necessary to define a measure that can measure correlation exactly. Karl Pearson's coefficient of correlation is the main measure of correlation.

Karl Pearson's coefficient of correlation

It is the measure of correlation or the numeric value, which gives idea regarding the correlation between the two variables. It is denoted by the notation “ r ” or “ r_{xy} ” and is defined as follows:

$$r_{xy} = \frac{\text{Cov}(X,Y)}{\sqrt{V(X) \cdot V(Y)}} = \frac{\text{Cov}(X,Y)}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}} = \frac{\text{Cov}(X,Y)}{\sigma_x \cdot \sigma_y}$$

Where, $\text{Cov}(X,Y)$ represents joint variation between X and Y and is defined as follows:

$$\text{Cov}(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x} \bar{y}$$

$$\text{Also, } V(X) = \sigma_x^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 \text{ and } V(Y) = \sigma_y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2.$$

Computation of correlation coefficient

Karl Pearson's coefficient of correlation (r_{xy}) can also be computed by direct as well as indirect method as studied in case of mean computation. In direct method, r_{xy} is calculated directly by using the formula. Where as in case of indirect method, first r_{uv} is calculated by transforming variables X and Y into new variables U and V and, then using r_{uv} the required correlation coefficient r_{xy} is calculated.

Properties of Karl Pearson's correlation coefficient

1. The value of correlation coefficient always lies in between -1 to 1 , that is, $-1 \leq r_{xy} \leq 1$.
2. If the value of r_{xy} is positive, then the two variables X and Y are positively correlated.
3. If the value of r_{xy} is negative, then the two variables X and Y are negatively correlated.
4. If the value of r_{xy} is zero, then the two variables are uncorrelated.

5. If the value of $r_{xy} = +1$, then the two variables have perfect linear relationship and the correlation is of positive type.
6. If the value of $r_{xy} = -1$, then the two variables have perfect linear relationship and the correlation is of negative type.
7. The correlation coefficient is not affected by the change of origin as well as change of scale.

That is,

$$r_{xy} = r_{uv}$$

When,

the new variables U and V are defined using
change of origin that is $\{U = X - A \text{ & } V = Y - B\}$.
and

$$\text{change of origin and scale that is } \left\{ U = \frac{X - A}{h} \text{ and } V = \frac{Y - B}{k} \right\}$$

8.4 Coefficient of determination

If r_{xy} represents correlation coefficient between any two variables X and Y , then r_{XY}^2 represents coefficient of determination. Sometimes, it is multiplied by 100 and is expressed in percentage. The larger the coefficient of determination, the larger the variation explained.

For example, if $r_{XY} = 0.8 \Rightarrow r_{XY}^2 = 0.64$, it means only 64% variation will be explained by the linear relation between X and Y .

Example 8.1

The following data are related to the percentage of humidity and the warp breakage rate recorded for a week in a loom shed.

Percentage humidity	54	85	86	50	42	75	65	56
Warp breakage rate	2.45	1.21	1.20	2.84	3.25	1.86	1.90	2.32

Find Karl Pearson's coefficient of correlation using the above data and comment on it.

Solution

Let us suppose that, $X \Rightarrow$ Percentage humidity
 $Y \Rightarrow$ Warp breakage rate

1. Direct method:

Completing the calculations as shown in Table 8.1

Table 8.1

x	y	x^2	y^2	xy	
54	2.45	2916	6.00	132.30	$\bar{x} = \frac{\sum x_i}{n} = 64.125$
85	1.21	7225	1.46	102.85	$\bar{y} = \frac{\sum y_i}{n} = 2.1287$
86	1.20	7396	1.44	103.20	
50	2.84	2500	8.07	142.00	
42	3.25	1764	10.56	136.50	$\sigma_x^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 = 236.3594$
75	1.86	5625	3.46	139.50	
65	1.90	4225	3.61	123.50	
56	2.32	3136	5.38	129.92	$\sigma_y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2 = 0.4668$
Total	513	17.03	34787	39.99	1009.77

$$\text{Cov}(X, Y) = \frac{\sum x_i y_i - \bar{x} \bar{y}}{n} = -10.2848$$

$$r_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{V(X) \cdot V(Y)}} = -0.9792$$

2. Indirect method

Suppose here change of origin is used. That is, suppose $U = X - 60$ and $V = Y - 2.0$. Completing the calculations as shown in Table 8.2

Table 8.2

x	y	u	v	u^2	v^2	uv	
54	2.45	-6	0.45	36	0.2025	-2.70	$\bar{u} = \frac{\sum u_i}{n} = 4.125$
85	1.21	25	-0.79	625	0.6241	-19.75	$\bar{v} = \frac{\sum v_i}{n} = 0.1287$
86	1.20	26	-0.8	676	0.6400	-20.80	
50	2.84	-10	0.84	100	0.7056	-8.40	$\sigma_u^2 = \frac{\sum u_i^2}{n} - \bar{u}^2 = 236.3594$
42	3.25	-18	1.25	324	1.5625	-22.50	
75	1.86	15	-0.14	225	0.0196	-2.10	$\sigma_v^2 = \frac{\sum v_i^2}{n} - \bar{v}^2 = 0.4661$
65	1.90	5	-0.1	25	0.0100	-0.50	
56	2.32	-4	0.32	16	0.1024	-1.28	
Total	513	17.03	33	2027	3.8667	-78.03	$\text{Cov}(U, V) = \frac{\sum u_i v_i}{n} - \bar{u} \bar{v}$ = -10.2816

$$r_{xy} = r_{uv} = \frac{\text{Cov}(U, V)}{\sqrt{V(U) \cdot V(V)}} = -0.9796$$

Thus, from the value of correlation coefficient, it can be concluded that percentage humidity and warp breakage rates have very high degree negative correlation. That is warp breakage rate is highly negatively affected by the percentage humidity.

Note that, if we find coefficient of determination, $r_{xy}^2 = (-0.9796)^2 = 0.9588 \Rightarrow$ nearly 96% variation of warp breakage rate is due to variation in percentage humidity.

8.5 Spearman's rank correlation coefficient

From the earlier discussion, it is clear that correlation coefficient is useful in studying the degree and the type of correlation between any two variables. But sometimes in statistics, study of correlation or the association between any two attributes may be of interest. For example, 10 pieces of fabric dyed by 10 different students may be judged and ranked by two different judges, and it may be of interest to decide whether the ranks assigned by the judges have correlation. That is, is it possible to conclude that the two judges are having similar or opposite opinion. This can be achieved with the help of rank correlation coefficient. Thus, rank correlation coefficient is the measure of correlation between the ranks of any two attributes. Here, ranks are the positive numbers assigned to the individual or individual observation under study.

For example, the ranks can be assigned to the data of marks as shown in Table 8.3.

Table 8.3

Marks (X)	65	56	85	80	45	70	65	68	72
Ranks (R_x)	6.5	8	1	2	9	4	6.5	5	3

Note that here rank 1 is assigned to highest, rank 2 is assigned to second highest, and so on. Also average of 6 and 7 is taken as 6.5 because observation 65 is repeating twice.

Definition of rank correlation coefficient

The rank correlation coefficient is denoted by the notation “ R ” and is defined as follows:

$$R = 1 - \frac{6 \times \sum d_i^2}{n(n^2 - 1)} \quad \text{if ranks are repeating}$$

$$R = 1 - \frac{6 \times (\sum d_i^2 + CF)}{n(n^2 - 1)} \quad \text{if ranks are repeating}$$

Where,

“ n ” represents number of pairs of ranks.

$d_i = R_{X_i} - R_{Y_i}$ = difference of i^{th} pair of ranks.

$$CF = CF(X) + CF(Y) = \sum \frac{m(m^2 - 1)}{12} + \sum \frac{k(k^2 - 1)}{12}$$

“ m ” represents length of tie (number of repetitions) for a rank in X series and $\sum \frac{m(m^2 - 1)}{12}$ is sum of quantities $\frac{m(m^2 - 1)}{12}$ for all such ties. “ k ” represents length of tie (number of repetitions) for a rank in Y series. and $\sum \frac{k(k^2 - 1)}{12}$ is sum of quantities $\frac{k(k^2 - 1)}{12}$ for all such ties.

Properties of Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient satisfies all the properties which are satisfied by the Karl Pearson's coefficient of correlation.

Example 8.2

Following are the marks obtained by eight different students in the subject of mathematics and the subject of statistics:

Student no.	1	2	3	4	5	6	7	8
Marks in mathematics	48	85	70	72	70	68	65	80
Marks in statistics	42	92	65	80	67	62	60	72

Calculate rank correlation coefficient using the above data and comment on it.

Solution

Suppose,

X represents marks in mathematics and Y represents marks in statistics.

R_x represents ranks for X observations and R_y represents ranks for Y observations.

Assigning the ranks and completing the calculations as shown in Table 8.4.

Table 8.4

X_i	Y_i	R_{X_i}	R_{Y_i}	$d_i = R_{X_i} - R_{Y_i}$	d_i^2
48	42	8	8	0	0
85	92	1	1	0	0
70	65	4.5	5	-0.5	0.25
72	80	3	2	1	1
70	67	4.5	4	0.5	0.25
68	62	6	6	0	0
65	60	7	7	0	0
80	72	2	3	-1	1
Total				2.5	

$$R = 1 - \frac{6 \times (\sum d_i^2 + CF)}{n(n^2 - 1)} \quad \text{as ranks are repeating}$$

Now,

As ranks are repeating only in X series

$$\begin{aligned} CF(X) &= \sum \frac{m(m^2 - 1)}{12} = \frac{2(2^2 - 1)}{12} = 0.5 \text{ and } CF(Y) = 0 \\ R &= 1 - \frac{6 \times (2.5 + 0.5)}{8(8^2 - 1)} = 0.96423 \end{aligned}$$

Thus, from the value of rank correlation coefficient, it can be said that the ranks are highly positively correlated. That is, intelligence in mathematics and in statistics have very high-degree positive correlation.

Example 8.3

Following are the ranks assigned by three different judges to 10 different pieces of fabric dyed in a laboratory.

Piece number		1	2	3	4	5	6	7	8	9	10
Ranks	Judge I	9	7.5	1	5.5	4	7.5	3	5.5	10	2
	Judge II	5	3	7	10	5	9	5	1	8	2
	Judge III	10	8	2	6	3	7	4	5	9	1

Using rank correlation coefficient, decide which pair of judges has nearest approach of giving ranks.

Solution

Suppose,

R_1 represents rank given by Judge I, R_2 represents rank given by Judge II and R_3 represents rank given by Judge III.

Note that, largest positive value of rank correlation coefficient indicates nearest approach.

Therefore, computing pair wise rank correlation the decision regarding nearest approach can be made.

Consider the pair of Judge I and Judge II.

Complete the calculations as shown in Table 8.5.

Table 8.5

R_{1i}	R_{2i}	$d_i = R_{X_i} - R_{Y_i}$	d_i^2
9	5	4.0	16
7.5	3	4.5	20.25
1	7	-6.0	36.00
5.5	10	-4.5	20.25
4	5	-1.0	1.00
7.5	9	-1.5	2.25
3	5	-2.0	4.00
5.5	1	4.5	20.25
10	8	2.0	4.00
2	2	0.0	0.00
Total			124.00

$$R = 1 - \frac{6 \times (\sum d_i^2 + CF)}{n(n^2 - 1)} \quad \text{as ranks are repeating}$$

Now,

As ranks are repeating in both X and Y series

$$CF(X) = \Sigma \frac{m(m^2 - 1)}{12} = \frac{2(2^2 - 1)}{12} + \frac{2(2^2 - 1)}{12} = 1$$

$$CF(Y) = \Sigma \frac{k(k^2 - 1)}{12} = \frac{3(3^2 - 1)}{12} = 2$$

$$R = 1 - \frac{6 \times (124 + 3)}{10(10^2 - 1)} = 0.2303$$

Consider the pair of Judge II and Judge III.

Complete the calculations as shown in Table 8.6.

Table 8.6

R_{2i}	R_{3i}	$d_i = R_{x_i} - R_{y_i}$	d_i^2
5	10	-5	25
3	8	-5	25
7	2	5	25
10	6	4	16
5	3	2	4
9	7	2	4
5	4	1	1
1	5	-4	16
8	9	-1	1
2	1	1	1
Total		118	

$$R = 1 - \frac{6 \times (\sum d_i^2 + CF)}{n(n^2 - 1)} \quad \text{as ranks are repeating}$$

Now,

As ranks are repeating only in X series

$$CF(X) = \Sigma \frac{m(m^2 - 1)}{12} = \frac{3(3^2 - 1)}{12} = 2$$

$$CF(Y) = 0$$

$$R = 1 - \frac{6 \times (118 + 2)}{10(10^2 - 1)} = 0.2727$$

Consider the pair of Judge I and Judge III.

Complete the calculations as shown in Table 8.7.

Table 8.7

R_{1i}	R_{3i}	$d_i = R_{x_i} - R_{y_i}$	d_i^2
9	10	-1	1
7.5	8	-0.5	0.25
1	2	-1	1
5.5	6	-0.5	0.25
4	3	1	1
7.5	7	0.5	0.25
3	4	-1	1
5.5	5	0.5	0.25
10	9	1	1
2	1	1	1
Total		7	

$$R = 1 - \frac{6 \times (\sum d_i^2 + CF)}{n(n^2 - 1)} \quad \text{as ranks are repeating}$$

Now,

As ranks are repeating only in X series

$$CF(X) = \Sigma \frac{m(m^2 - 1)}{12} = \frac{2(2^2 - 1)}{12} + \frac{2(2^2 - 1)}{12} = 1$$

$$CF(Y) = 0$$

$$R = 1 - \frac{6 \times (7+1)}{10(10^2 - 1)} = 0.9515$$

As the value of rank correlation coefficient is highest for the pair of first and third judges, this pair has nearest approach of giving ranks.

8.6 Regression analysis

As discussed earlier with the help of correlation analysis, it is possible to decide whether the two variables of the bivariate data are correlated and what is the type of correlation. But sometimes, it may be of interest to find value of any one variable when the value of another variable is known or given. This can be done with the help of regression analysis. Thus, if we try to find or estimate value of any one variable on the basis of the value of another variable, then it is called the regression analysis.

Such estimation is possible only if we know the functional relation between the two variables of the data. According to the functional relationship, linear regression and nonlinear regression are the two different types of regression. Nonlinear regression is beyond the scope of this book, hence only linear regression is discussed here.

Linear regression analysis

When we try to find or estimate value of any one variable on the basis of the value of another variable using linear relation or equation, then it is called the linear regression analysis. Linear regression of X on Y and linear regression of Y on X are the two different types of linear regression.

Linear regression of X on Y

When we try to find or estimate value of variable X on the basis of the value of variable Y using linear relation or equation, then it is called the linear

regression of X on Y . In this case X is the dependent. As the linear equation of two variables represents a straight line in the plane variable and Y is independent variable, the linear regression equation of X on Y is also called the line of regression of X on Y . The equation of line of regression of X on Y is derived using the mathematical method known as the “Method of least squares.” The method in short is as follows:

Suppose

$$x = a + by \dots \dots (1) \quad \text{is the required equation.}$$

Finding the above equation is equivalent to find the constants “ a ” and “ b ” in such a way that total error sum of squares, that is the function

$$\text{ESS} = \sum (x_i - x)^2 = \sum (x_i - a - by_i)^2 \text{ is minimum.}$$

Differentiating the above function partially with respect to “ a ” and “ b ” and equating these partial derivatives $\frac{\partial \text{ESS}}{\partial a}$ and $\frac{\partial \text{ESS}}{\partial b}$ to zero results in two equations in two unknowns “ a ” and “ b ”. These equations are known as the normal equations and are as follows:

$$\sum x_i = na + b \sum y_i$$

$$\sum x_i y_i = na + b \sum y_i^2$$

Solving these equations for “ a ” and “ b ,” substitution of values of the constants “ a ” and “ b ” in equation 1 and simplification gives the required final form of the equation of line of regression of X on Y , which is as follows:

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

or

$$x = (\bar{x} - b_{xy} \bar{y}) + b_{xy} y$$

Linear regression of Y on X

The equation of line of regression of Y on X can be derived in the same manner as derived in case of regression equation of X on Y , only by interchanging the variables X and Y in the above derivations and the final form of the equation of line of regression of Y on X can be given as follows:

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

or

$$y = (\bar{y} - b_{yx} \bar{x}) + b_{yx} x$$

where,

\bar{x} , \bar{y} are the means of variables X and Y , also b_{xy} and b_{yx} are called the regression coefficients of X on Y and Y on X , respectively. These regression coefficients are defined as follows;

$$b_{xy} = \frac{\text{Cov}(X,Y)}{V(Y)}$$

$$b_{yx} = \frac{\text{Cov}(X,Y)}{V(X)}$$

Note that, definitions of $\text{COV}(X,Y)$, $V(X)$ and $V(Y)$ are already discussed in correlation analysis.

Properties of regression coefficients

1. The regression coefficient b_{yx} represents slope of the regression line of Y on X .
2. Similarly, $1/b_{xy}$ represents slope of the regression line of X on Y .
3. Either both regression coefficients b_{xy} and b_{yx} are positive or both are negative.
4. If both regression coefficients b_{xy} and b_{yx} are positive, then the variables X and Y have positive correlation.
5. If both regression coefficients b_{xy} and b_{yx} are negative, then the variables X and Y have negative correlation.
6. From the definitions of correlation coefficient and regression coefficients, we have:

$$b_{xy} = \frac{\text{Cov}(X,Y)}{V(Y)} = \frac{r \cdot \sigma_x \cdot \sigma_y}{\sigma_y^2} = r \cdot \frac{\sigma_x}{\sigma_y}$$

$$b_{yx} = \frac{\text{Cov}(X,Y)}{V(X)} = \frac{r \cdot \sigma_x \cdot \sigma_y}{\sigma_x^2} = r \cdot \frac{\sigma_y}{\sigma_x}$$

Further,

$$b_{xy} \cdot b_{yx} = r^2$$

$$\text{Therefore, } r = \pm \sqrt{b_{xy} \cdot b_{yx}}$$

Note that, here one should take positive square root if both regression coefficients b_{xy} and b_{yx} are positive and negative square root if both regression coefficients b_{xy} and b_{yx} are negative.

7. The regression coefficients are not affected by the change of origin but are affected by the change of scale. That is,

$$b_{xy} = b_{uv} \quad \text{and} \quad b_{yx} = b_{vu}$$

If, the new variables U and V are defined using change of origin that is

$$\{U = X - A \quad \text{and} \quad V = Y - B\}.$$

But,

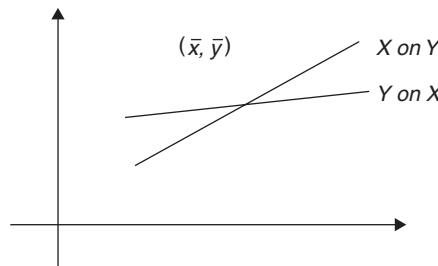
$$b_{xy} = \frac{h}{k} \cdot b_{uv} \quad \text{and} \quad b_{yx} = \frac{k}{h} \cdot b_{vu}$$

If, the new variables U and V are defined using change of origin and scale that is

$$\left\{ U = \frac{X - A}{h} \quad \text{and} \quad V = \frac{Y - B}{k} \right\}$$

Property of regression lines

The regression lines of X on Y and Y on X always intersect at the point (\bar{x}, \bar{y}) as shown if Fig. 8.2.



8.2

Example 8.4

The following data are related to the percentage of humidity and the warp breakage rate recorded for a week in a loom shed.

Percentage humidity	54	85	86	50	42	75	65	56
Warp Breakage rate	2.45	1.21	1.20	2.84	3.25	1.86	1.90	2.32

Using the above data, find two equations of lines of regression. In addition, find warp breakage rate if humidity percentage on a specific day is 60 and find percentage humidity required for the target warp breakage rate of 1.50%.

Solution:

Let us suppose that $X \Rightarrow$ percentage humidity

$Y \Rightarrow$ warp breakage rate

1. Direct method:

Completing the calculations as per Table 8.8.

Table 8.8

x	y	x^2	y^2	xy	
54	2.45	2916	6.00	132.3	$\bar{x} = \frac{\sum x_i}{n} = 64.125$
85	1.21	7225	1.46	102.85	$\bar{y} = \frac{\sum y_i}{n} = 2.1287$
86	1.2	7396	1.44	103.2	
50	2.84	2500	8.07	142	
42	3.25	1764	10.56	136.5	$\sigma_x^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 = 236.3594$
75	1.86	5625	3.46	139.5	
65	1.9	4225	3.61	123.5	
56	2.32	3136	5.38	129.92	$\sigma_y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2 = 0.4661$
Total	513	17.03	34787	39.99	1009.77

$$\text{Cov}(X, Y) = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} = -10.2816$$

$$b_{xy} = \frac{\text{Cov}(X, Y)}{V(Y)} = \frac{-10.2848}{0.4668} = -22.0588$$

$$b_{yx} = \frac{\text{Cov}(X, Y)}{V(X)} = \frac{-10.2848}{236.3595} = -0.0435$$

Now, the equation of line of regression of X on Y is

$$x = (\bar{x} - b_{xy} \bar{y}) + b_{xy} y$$

$$\text{Therefore, } x = (64.125 - (-22.0326 \times 2.1256)) - 22.0326y$$

$$\text{Therefore, } x = 111.0816 - 22.0326y$$

This is, the required equation of line of regression of X on Y.

Further, if y = Warp breakage rate = 1.50

$$\text{Therefore, } x = 111.0816 - 22.0326 \times 1.50 = 77.9221$$

Thus for achieving target warp breakage rate of 1.50 humidity should be nearly 77.92%.

Similarly, the equation of line of regression of Y on X is

$$y = (\bar{y} - b_{yx} \bar{x}) + b_{yx} x$$

$$\text{Therefore, } y = (2.1256 - (-0.0435 \times 64.125)) - 0.0435x$$

$$\text{Therefore, } y = 4.9181 - 0.0435x$$

This is the required equation of line of regression of Y on X.

Further, if x = humidity = 60%

$$\text{Therefore, } y = 4.9181 - 0.0435 \times 60 = 2.305$$

Thus, for humidity level 60% the warp breakage rate will be approximately 2.31.

2. Indirect method

Suppose, here change of origin is used. That is, suppose $U = X - 60$ and $V = Y - 2.0$. Completing the calculations as shown in Table 8.9.

Table 8.9

x	y	u	v	u^2	v^2	uv	
54	2.45	-6	0.45	36	0.2025	-2.7	$\bar{u} = \frac{\sum u_i}{n} = 4.125$
85	1.21	25	-0.79	625	0.6241	-19.75	$\bar{v} = \frac{\sum v_i}{n} = 0.1287$
86	1.2	26	-0.8	676	0.64	-20.8	
50	2.84	-10	0.84	100	0.7056	-8.4	$\sigma_u^2 = \frac{\sum u_i^2}{n} - \bar{u}^2 = 236.3594$
42	3.25	-18	1.25	324	1.5625	-22.5	
75	1.86	15	-0.14	225	0.0196	-2.1	$\sigma_v^2 = \frac{\sum v_i^2}{n} - \bar{v}^2 = 0.4661$
65	1.9	5	-0.1	25	0.01	-0.5	
56	2.32	-4	0.32	16	0.1024	-1.28	
Total	513	17.03	33	2027	3.8667	-78.03	$\text{Cov}(U, V) = \frac{\sum u_i v_i}{n} - \bar{u} \bar{v} = -10.2816$

$$b_{xy} = b_{uv} = \frac{\text{Cov}(U,V)}{V(V)} = \frac{-10.2848}{0.4668} = -22.0588$$

$$b_{yx} = b_{vu} = \frac{\text{Cov}(U,V)}{V(U)} = \frac{-10.2848}{236.3595} = -0.0435$$

Also $\bar{x} = A + \bar{u} = 60 + 4.125 = 64.125$ and $\bar{y} = B + \bar{v} = 2.0 + 0.1256 = 2.1256$

Now, the equation of line of regression of X on Y is

$$x = (\bar{x} - b_{xy} \bar{y}) + b_{xy} \cdot y$$

$$\text{Therefore, } x = (64.125 - (-22.0326 \times 2.1256)) - 22.0326y$$

$$\text{Therefore, } x = 111.0816 - 22.0326y$$

This is the required equation of line of regression of X on Y .

Further, if y = Warp breakage rate = 1.50

$$\text{Therefore, } x = 111.0816 - 22.0326 \times 1.50 = 77.9221$$

Thus for achieving target warp breakage rate of 1.50 humidity should be approximately 77.92%.

Similarly, the equation of line of regression of Y on X is

$$y = (\bar{y} - b_{yx} \bar{x}) + b_{yx} \cdot x$$

$$\text{Therefore, } y = (2.1256 - (-0.0435 \times 64.125)) - 0.0435x$$

$$\text{Therefore, } y = 4.9181 - 0.0435x$$

This is the required equation of line of regression of Y on X .

Further, if x = humidity = 60%

$$\text{Therefore, } y = 4.9181 - 0.0435 \times 60 = 2.305$$

Thus, for humidity level 60% the warp breakage rate will be approximately 2.31.

Example 8.5

Find the means of variables X and Y and also find the coefficient of correlation from the following two equations of regression.

$$3x + 4y = 26$$

$$8x + 2y = 10$$

Solution

We know that the regression lines of X on Y and Y on X always intersect at the point (\bar{x}, \bar{y}) , and the point of intersection of two lines can be obtained by solving the two equations of lines simultaneously.

Suppose,

$$3x + 4y = 26 \quad \rightarrow \quad (1)$$

$$8x + 2y = 10 \quad \rightarrow \quad (2)$$

Now,

$$\text{Equation (1) - equation (2)} \times 2 \text{ gives } x = \bar{x} = \frac{6}{-13} = -0.4615$$

Substituting $\bar{x} = 0.4615$ in equation (1) gives $\bar{y} = 6.8461$

Thus the means of the variables X and Y are -0.4615 and 6.8461 , respectively.

Further for finding the correlation coefficient,

Suppose,

$$3x + 4y = 26 \quad \rightarrow \text{Equation of } X \text{ on } Y$$

$$8x + 2y = 10 \quad \rightarrow \text{Equation of } Y \text{ on } X$$

Now,

Comparing first equation with the general equation of line of regression of X on Y

$$\begin{aligned} x &= (\bar{x} - b_{xy}\bar{y}) + b_{xy} \cdot y \\ \Rightarrow b_{xy} &= -\frac{4}{3} \end{aligned}$$

Also,

Comparing second equation with the general equation of line of regression of Y on X

$$\begin{aligned} y &= (\bar{y} - b_{yx}\bar{x}) + b_{yx} \cdot x \\ \Rightarrow b_{yx} &= -\frac{8}{2} = -4 \end{aligned}$$

$$\text{Therefore, } r = \pm \sqrt{b_{xy} \cdot b_{yx}} = -\sqrt{-\frac{4}{3} \times -4} = -\sqrt{\frac{16}{3}} = -\frac{4}{\sqrt{3}} = -5.333$$

But this is contradiction as the value of correlation coefficient lies in between $[-1, +1]$.

This is because of wrong assumption made initially.

Therefore changing assumption,

Suppose,

$$3x + 4y = 26 \quad \rightarrow \text{Equation of } Y \text{ on } X$$

$$8x + 2y = 10 \quad \rightarrow \text{Equation of } X \text{ on } Y$$

Now,

Comparing first equation with the general equation of line of regression of Y on X

$$\begin{aligned} y &= (\bar{y} - b_{yx}\bar{x}) + b_{yx} \cdot y \\ \Rightarrow b_{yx} &= -\frac{3}{4} \end{aligned}$$

Also,

Comparing second equation with the general equation of line of regression of X on Y

$$x = (x - b_{xy}\bar{y}) + b_{xy}y$$

$$\Rightarrow b_{xy} = -\frac{2}{8}$$

$$\text{Therefore, } r = \mp \sqrt{b_{xy} \cdot b_{yx}} = -\sqrt{-\frac{3}{4} \times -\frac{2}{8}} = -\sqrt{\frac{6}{32}} = -0.433$$

8.7 Exercise

- The following data are related to the monthly income of the worker (in Rs.) and average savings (in Rs.) collected from a textile company.

Monthly income	Average savings
0–1000	150
1000–2000	350
2000–3000	400
3000–4000	450
4000–5000	600
5000–6000	800

Calculate coefficient of correlation for the above data and comment on it.

- Compute Karl Pearson's coefficient of correlation from the following data of price of garment (in Rs.) and Number of garments sold and comment on it.

Price of garment	100	105	110	115	120	125	130
Number of garments sold	85	80	70	75	65	60	60

- The results of percentage humidity and Warp breakages were recorded from a weaving factory as follows:

Percentage humidity	60	62	59	70	56	58	75
Warp breakages	112	110	120	102	122	118	90

Compute Karl Pearson's coefficient of correlation from the above data and comment on it.

4. The following data are related to the monthly income of the worker (in Rs.) and average savings (in Rs.) collected from a textile company.

Monthly income	Average savings
0–1000	150
1000–2000	350
2000–3000	400
3000–4000	450
4000–5000	600

- Find the two equations of lines of regression using the above data.
- Find the average saving of the worker having income of Rs. 4800.
- What is the income of the worker having average saving of Rs. 500
- Find the equation of line of regression to estimate number of garments sold using following data of price of garment (in Rs.) and number of garments sold. Also find number of garments sold if the price of garment is Rs. 112.

Price of garment	100	105	110	115	120	125	130
Number of garments sold	85	80	70	75	65	60	60

- The results of percentage humidity and end breakages recorded from a ring frame of a spinning mill are as follows:

Percentage humidity	60	62	59	70	56	58	75
End breakages	10	12	10	14	6	8	15

- Find the two equations of lines of regression using the above data.
- Find the end breakages if the humidity is 65%.
- What should be the humidity percentage for getting only 5 end breakages?
- Find the two equations of line of regression to estimate production (in dozens) and number of workers present using following data. Also find number of workers required for getting production of 60 dozens.

Number of workers present	100	105	110	115	120	125	130
Production (in dozens)	58	64	65	70	75	80	85

8. If the value of correlation coefficient between the two variables X and Y is 0.83, find the equations of regression of X on Y and Y on X using following data.

		Variables	
		X	Y
Mean		112	65
Standard deviation		2.5	3.2

9. Following are the ranks assigned by three different judges to 10 different participants in a fashion show. Using rank correlation coefficient, decide which two judges have concordance.

Participant no.	1	2	3	4	5	6	7	8	9	10
Judge 1	8	1	6	3	9	3	7	10	3	5
Judge 2	7	1	5.5	2	10	4	5.5	8	3	9
Judge 3	6.5	2	6.5	5	8	1	4	9	3	10

10. Find the Spearman's rank correlation coefficient for the following data of marks in physics and textile testing and comment on it.

Marks in physics	70	60	65	50	58	75	68
Marks in textile testing	74	62	70	40	45	78	56

Multivariate analysis

9.1 Introduction

Sometimes in statistics, the data under study are related to three or more variables, that is, information related to three or more variables is collected from each and every individual of the population or the sample. From such data, it may be of interest to find out whether there is any relation of one variable with all other variables of the data or is it possible to find the value of any one variable on the basis of the data if the values of all other variables are known. Multiple correlation and multiple regression analysis are the answer to the above questions.

9.2 Multivariate data

The data related to the three or more variables is called the multivariate data. In particular, if X , Y , and Z are any three variables, then the data related to these variables is called the trivariate data and the trivariate data of n observations can be given as follows

$$\{(x_1, y_1, z_1); (x_2, y_2, z_2); \dots; (x_n, y_n, z_n)\}$$

Multiple correlation analysis and multiple regression analysis are the two different ways of studying multivariate data.

9.3 Multiple correlation analysis

The study of the relationship between the multiple variables of the data is called the multiple correlation analysis. Study of the multiple correlation of one variable with all others and the partial correlation between any two variables assuming all others are at fixed or the constant levels are the two different types of multiple correlation analysis.

Multiple correlation of one variable with others

In this case, the relationship of any one variable is studied with all other variables. That is, in this case, the value of one variable is influenced by all

other variables. Thus, any one variable of the multivariate data is said to have multiple correlation with all other variables, if the value of this variable is affected by the change in the values of all other variables. Multiple correlation coefficient is the main measure of multiple correlation of one variable with all others.

Multiple correlation coefficients

Any value or any term, which gives idea regarding the multiple correlation of one variable with all others, is called the multiple correlation coefficient. These are denoted by the notations $R_{1,23\dots k}$, $R_{2,13\dots k}$, and so on.

For example,

$R_{1,23\dots k}$ represents multiple correlation coefficient of X_1 with X_2, X_3, \dots, X_k
Similarly,

$R_{2,13\dots k}$ represents multiple correlation coefficient of X_2 with X_1, X_3, \dots, X_k

Partial correlation between any two variables

In this case, the relationship between any two variables is studied assuming that all other variables are at the constant or the fixed levels. That is, in this case, the values of any two variables are influenced by each other, when all other variables are at the constant or the fixed levels. Thus, any two variables of the multivariate data are said to have partial correlation with each other, if they affect each other when all other variables are at the fixed or the constant levels. Partial correlation coefficient is the main measure of the partial correlation between any two variables.

Partial correlation coefficients

Any value or any term that gives idea regarding the partial correlation between any two variables, assuming all other variables are at constant or the fixed levels, is called the partial correlation coefficient. These are denoted by the notations $r_{12,3\dots k}$ or $r_{13,2\dots k}$ etc.

For example, $r_{12,3\dots k}$ represents partial correlation coefficient between X_1 and X_2 assuming X_3, \dots, X_k are at the fixed or the constant levels.

Similarly,

$r_{13,2\dots k}$ represents partial correlation coefficient between X_1 and X_3 assuming X_2, \dots, X_k are at the fixed or the constant levels.

**Definitions of multiple and partial correlation coefficient
(for three variables data)**

Suppose X_1 , X_2 , and X_3 are the three different variables under study and suppose that r_{12} , r_{13} , r_{23} are the correlation coefficients of X_1 and X_2 , X_1 , and X_3 and X_2 and X_3 respectively.

Suppose,

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

That is the matrix R is the matrix of the correlation coefficient which is a square symmetric matrix.

Further Suppose that,

R_{ij} = Minor corresponding to the element r_{ij} of the matrix R

Also,

ω_{ij} = Cofactor corresponding to the element r_{ij} of the matrix R

$$\omega_{ij} = (-1)^{i+j} R_{ij}$$

With all the above notations the multiple correlation coefficient of X_1 with X_2 and X_3 is defined as follows:

$$R^2_{1.23} = 1 - \frac{|R|}{R_{11}} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

Similarly, the multiple correlation coefficient of X_2 with X_1 and X_3 is defined as follows:

$$R^2_{2.13} = 1 - \frac{|R|}{R_{22}} = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}$$

Also, the partial correlation coefficient between X_1 and X_2 assuming X_3 is at the fixed or the constant level is defined as follows:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Similarly, the partial correlation coefficient between X_1 and X_3 assuming X_2 is at the fixed or the constant level is defined as follows:

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

Properties of multiple correlation coefficient

1. The value of the multiple correlation coefficient always lies in between $[0, 1]$. That is,

$$0 \leq R_{i,12 \dots k} \leq 1$$

2. If, $R_{i,12 \dots k} = 0$, then X_i does not have multiple correlation with all other variables.
3. If, $R_{i,12 \dots k} = 1$, then X_i has perfect linear correlation with all other variables. That is, X_i has perfect linear relationship with all other variables.
4. If $R_{i,12 \dots k}$ is nearer to 0, then X_i has lower degree multiple correlation with all other variables.
5. If $R_{i,12 \dots k}$ is nearer to 1, then X_i has higher degree multiple correlation with all other variables.

Properties of partial correlation coefficient

The partial correlation coefficient is the correlation coefficient of two variables assuming all other variables at fixed level. Hence, it satisfies all the properties satisfied by the correlation coefficient of two variables (r_{xy}) discussed in Chapter 8.

9.4 Multiple regression analysis

The procedure of estimating or finding the value of any one variable on the basis of the values of all other variables is called the multiple regression analysis. In this case, to find the value of one variable on the basis of values of all other variables, a linear functional relationship is established which is called the linear regression equation or the equation of the plane of regression.

In particular, the equation of the plane of regression of X_1 with X_2 and X_3 can be obtained by using the method of least squares and the final simplified form of the regression equation can be obtained as follows:

Equation of plane of regression of X_1 with X_2 and X_3

Suppose that the general regression equation of the plane of X_1 with X_2 and X_3 is as follows:

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$

Now, the values of the constants a , $b_{12,3}$, and $b_{13,2}$ are estimated using the method of least square, that is by minimizing error of estimation. Substituting these estimated values into the original equation, the final simplified form of the equation of the plane of regression of X_1 with X_2 and X_3 can be given as follows:

$$\left(\frac{x_1 - \bar{x}_1}{\sigma_1} \right) \omega_{11} + \left(\frac{x_2 - \bar{x}_2}{\sigma_2} \right) \omega_{12} + \left(\frac{x_3 - \bar{x}_3}{\sigma_3} \right) \omega_{13} = 0$$

Where,

σ_1 , σ_2 and σ_3 are the standard deviations of the variables X_1 , X_2 and X_3 respectively. Also ω_{11} , ω_{12} and ω_{13} are the cofactors of the elements r_{11} , r_{12} and r_{13} of the matrix R . Note that the constants $b_{12,3}$ and $b_{13,2}$ are the regression coefficients.

Similarly, the final simplified form of the equation of the plane of regression of X_2 with X_1 and X_3 can be given as follows:

$$\left(\frac{x_1 - \bar{x}_1}{\sigma_1} \right) \omega_{21} + \left(\frac{x_2 - \bar{x}_2}{\sigma_2} \right) \omega_{22} + \left(\frac{x_3 - \bar{x}_3}{\sigma_3} \right) \omega_{23} = 0$$

Also, the final simplified form of the equation of the plane of regression of X_3 with X_1 and X_2 can be given as follows:

$$\left(\frac{x_1 - \bar{x}_1}{\sigma_1} \right) \omega_{31} + \left(\frac{x_2 - \bar{x}_2}{\sigma_2} \right) \omega_{32} + \left(\frac{x_3 - \bar{x}_3}{\sigma_3} \right) \omega_{33} = 0$$

Example 9.1

Following results are obtained from the 10 observations of the variables X_1 (Count), X_2 (Strength) and X_3 (% Elongation).

	X_1	X_2	X_3
Mean	30.12	12.56	5.25
SD	2.12	1.50	0.80

Also, $r_{12} = -0.65$, $r_{13} = -0.58$ and $r_{23} = 0.62$.

Using the above data

- Find the multiple correlation coefficient of X_3 with X_1 and X_2 and comment on the result.

2. Find the partial correlation coefficient between X_2 and X_3 assuming X_1 constant and comment on the result.
3. Find the equation of the plane of regression of X_3 with X_1 and X_2 and estimate the % elongation of the yarn having count 32's and strength 10 gm.

Solution

Here, the variables

X_1 —Count of the yarn, X_2 —Strength of the yarn and X_3 —% elongation of the yarn
Now,

1. Multiple correlation coefficient of X_3 with X_1 and X_2

$$R_{3.12}^2 = 1 - \frac{|R|}{R_{33}} = \frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}$$

Therefore,

$$\text{Therefore, } R_{3.12}^2 = 0.4386$$

$$\text{Therefore, } R_{3.12} = 0.6623.$$

Thus, X_3 (% elongation) does not have high-degree multiple correlation with the count (X_1) and the strength (X_2).

2. Partial correlation coefficient between X_2 and X_3 assuming X_1 constant

$$r_{23.1} = \frac{r_{23} - r_{12}r_{31}}{\sqrt{(1 - r_{12}^2)(1 - r_{31}^2)}}$$

$$r_{13.2} = 0.3925$$

Thus, strength (X_2) and % elongation (X_3) are positively correlated, if count (X_1) is kept constant and the degree of correlation is very low.

3. The equation of the plane of regression of X_3 with X_1 and X_2

$$\left(\frac{x_1 - \bar{x}_1}{\sigma_1} \right) \omega_{31} + \left(\frac{x_2 - \bar{x}_2}{\sigma_2} \right) \omega_{32} + \left(\frac{x_3 - \bar{x}_3}{\sigma_3} \right) \omega_{33} = 0$$

Here,

$$\text{Matrix of correlation } R = \begin{bmatrix} 1 & -0.65 & -0.58 \\ -0.65 & 1 & 0.62 \\ -0.58 & 0.62 & 1 \end{bmatrix}$$

$$\omega_{31} = (-1)^{3+1} R_{31} = \begin{vmatrix} -0.65 & 0.58 \\ 1 & 0.62 \end{vmatrix} = 0.177$$

$$\omega_{32} = (-1)^{3+2} R_{32} = - \begin{vmatrix} 1 & -0.58 \\ -0.65 & 0.62 \end{vmatrix} = -0.243$$

$$\omega_{33} = (-1)^{3+3} R_{33} = \begin{vmatrix} 1 & -0.65 \\ -0.65 & 1 \end{vmatrix} = 0.5775$$

Thus,

$$\left(\frac{x_1 - 30.12}{2.12} \right) \times 0.177 + \left(\frac{x_2 - 12.56}{1.5} \right) \times -0.243 + \left(\frac{x_3 - 5.25}{0.8} \right) \times 0.5775 = 0$$

$$(x_1 - 30.12) \times 0.0835 - (x_2 - 12.56) \times 0.162 + (x_3 - 5.25) \times 0.7219 = 0$$

$$x_3 = -4.2703 - 0.1157x_1 + 0.2244x_2$$

This is the required equation of the plane of regression of X_3 with X_1 and X_2 . Further,

$$X_1 = 32 \text{ and } X_2 = 10 \text{ then,}$$

$$x_3 = -4.2703 - 0.1157 \times 32 + 0.2244 \times 10$$

$$x_3 = 4.451\%$$

Thus, if the yarn count is 32^s and strength is 10 g, the percentage elongation of the yarn will be 4.451%.

9.5 Exercise

- Define multiple and partial correlation. What are the measures of multiple and partial correlation?
- What is multiple regression? State the equation of plane of regression of X_1 on X_2 and X_3 and explain it.
- What is multiple regression? What is plane of regression? Describe by taking suitable example.
- Following are the results of correlation between the variables X_1 , X_2 and X_3 .

$$r_{12} = 0.68, r_{13} = 0.65, r_{23} = 0.78.$$

Calculate multiple correlation coefficient of X_1 with X_2 and X_3 and write conclusion.

- Find the plane of regression of X_3 with X_1 and X_2 , using the following results of X_1 : rainfall (in cm), X_2 : area under cultivation (in acres) and X_3 : production of cotton (in tons). Also find estimate of X_3 , if $X_1 = 30$ and $X_2 = 600$.

	Variables		
	X_1	X_2	X_3
Mean	28.025	491.594	75.45
Standard deviation	4.42	11.0	8.5

$$r_{12} = -0.8, r_{23} = -0.56, \text{ and } r_{13} = 0.4$$

From the following results of trivariate data, find multiple correlation coefficient of X_2 with X_1 and X_3 and partial correlation coefficient of X_1 and X_2 keeping X_3 constant. Also comment on them.

6. Find the equation of plane of regression of X_3 with X_1 and X_2 . Also Find X_3 when $X_1 = 24$ and $X_2 = 20$ and comment on it.

	Variables		
	X_1	X_2	X_3
Mean	22.5	20.25	42.15
Standard deviation	2.5	1.75	2.2

$$\text{Also } r_{12} = -0.65, r_{13} = -0.68 \text{ and } r_{23} = 0.65$$

7. Find multiple correlation coefficient of X_2 with X_1 and X_3 and partial correlation coefficient of X_1 and X_2 assuming X_3 is constant. Also comment on them.

$$R = \begin{vmatrix} 1 & 0.36 & 0.88 \\ 0.36 & 1 & 0.65 \\ 0.88 & 0.65 & 1 \end{vmatrix}$$

8. If $r_{12} = 0.86, r_{13} = 0.62$ and $r_{23} = 0.58$, find $R_{2,13}$ and comment on it.
From the following results of trivariate data, find the equation of the plane of regression of X_2 with X_1 and X_3 and find value of X_2 if $X_1 = 35$ and $X_3 = 8$.

	Variable		
	X_1	X_2	X_3
Mean	30	20	10
SD	4.5	3.5	2.5

$$\text{Also } r_{12} = -0.56, r_{13} = -0.4 \text{ and } r_{23} = 0.8.$$

10.1 Introduction

In real life, most often we come across random phenomena and in such cases one may be interested in finding, what is the chance of happening of certain thing. Sometimes one can take proper decision if exact value of chances of happening is known. For example, a manufacturer of garments may be interested in knowing, what is the chance that the consignment of packs of 10 garments each, dispatched to the customer will be rejected by the customer. The spinning master may want to know, how many count tests made on yarn will show count in between 26–29 or > 30 or < 25 , etc. The chance of happening of certain incidence can be measured and its value always lies in between 0 and 1. This value of measure of chance is called the probability. Thus, probability is a real number belonging to the interval $[0, 1]$, which gives an idea regarding the happening of an incident or an event. It can be used for the comparison of happenings of two or more events with each other.

10.2 Basic concepts

Experiment

Any task or phenomenon, which gives us some outcome or result when it is performed, is called an experiment. There are two types of an experiment as follows:

Sure experiment

The experiment whose outcome is almost sure is called the sure experiment.

For example, throwing a stone up in the sky is a sure experiment.

Random experiment

The experiment whose all possible outcomes are known, but when the experiment is actually performed, which outcome will happen or occur is

not predictable is called the random experiment. That is, the experiment whose outcome depends on the chance is called the random experiment. For example, tossing a coin, throwing a die, count test of the yarn, and strength test of the fabric are random experiments.

Sample space

The set or the collection of all possible outcomes of the random experiment is called the sample space. It is denoted by the notation S or Ω . For example, if the random experiment is tossing a coin, then $\Omega = \{H, T\}$; if the random experiment is throwing a die, then $\Omega = \{1, 2, 3, 4, 5, 6\}$ and if the random experiment is count test, then Ω may be interval (10.5, 13.5) for nominal count of 12^s. There are two types of the sample spaces as follows:

Finite sample space

If the number of elements belonging to the sample space is finite or limited, then it is called the finite sample space. For example, $\Omega = \{1, 2, 3, 4, 5, 6\}$ is a finite sample space.

Infinite sample space

If the number of elements belonging to the sample space is infinite, then it is called an infinite sample space. Infinite sample space can be divided into two categories. For example, $\Omega = \{1, 2, 3, \dots, \infty\}$ is a countable infinite sample space and $\Omega = (0, \infty)$ is an uncountable infinite sample space.

Event

Any subset of the sample space is called an event. The notations, namely, A , B , C , etc., always denote an event. For example, if $\Omega = \{H, T\}$ then, $A = \{H\}$ or $A = \{T\}$ are the events and if $\Omega = \{1, 2, 3, 4, 5, 6\}$ then, $A = \{\text{Even numbers}\}$ or $A = \{\text{Odd numbers}\}$ are the events.

Note that, the null set or the empty set (\emptyset) is an event, as it is the subset of the sample space. Similarly, the sample space Ω is also an event, as it is subset of itself.

There are different types of an event as follows:

Sure event

Any event, which is almost sure to happen, is called the sure event. For example, from a box full of white balls, if any one ball is selected at random, then an event $A = \{\text{White ball in the draw}\}$ is a sure event.

Impossible event

Any event, which is impossible to happen, is called an impossible event. For example, from a box full of white balls, if any one ball is selected at random, then an event

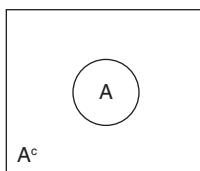
$A = \{\text{Black ball in the draw}\}$ is an impossible event.

Mutually exclusive events

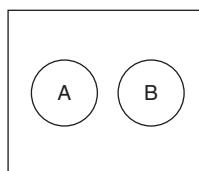
Any two events A and B are called the mutually exclusive events, if they cannot happen at a time or they cannot occur simultaneously. For example, if the experiment is tossing a coin, and the events A and B denote getting head and getting tail, then A and B are the mutually exclusive events. Also, if the experiment is throwing a die, and the event A and B denote getting even number and getting odd number, then A and B are the mutually exclusive events.

Venn diagram

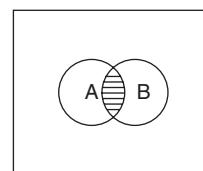
The pictorial representation of a sample space and the events is called the “Venn diagram”. Following are typical examples of Venn diagrams as shown in Fig. 10.1



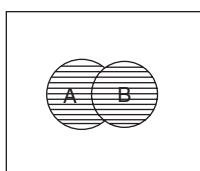
A denote an event A^c denote compliment of event A



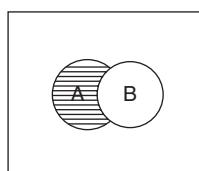
A and B are mutually exclusive events



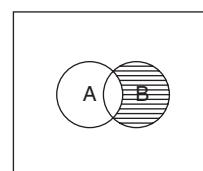
$A \cap B$ means A and B both will happen



$A \cup B$ means A happens or B happens or both happens



$A \cap B^c$ means only A happens



$A^c \cap B$ means only B happens

10.3 Definition of the probability (classical approach)

Let Ω be the sample space of the random experiment containing “ n ” elements and suppose that A is any event belonging to Ω containing “ m ” elements, then probability of event A is denoted by the notation $P(A)$ and is defined as follows:

$$P(A) = \frac{m}{n} = \frac{\text{Number of favorable cases}}{\text{Total number of cases}}$$

10.4 Laws of the probability

$$0 \leq P(A) \leq 1 \\ P(A) + P(A^c) = 1 \Rightarrow P(A^c) = 1 - P(A)$$

For any two events A and B , $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

If events A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$

1. Law of independence

Any two events A and B are independent of each other, if $P(A \cap B) = P(A) \times P(B)$

2. Law of conditional probability

If A and B are any two events, then the probability of event A under the condition that event B has already happened is called the conditional probability of A . It is denoted by the notation $P(A/B)$ and pronounced as probability of A given B . Also, it is defined as follows:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Similarly, the probability of B given A is defined as follows:

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Thus, from above the property, we get:

$$P(A \cap B) = P(A/B) \times P(B) = P(B/A) \times P(A)$$

3. De' Morgan's Law

$$P(A \cup B)^c = P(A^c \cap B^c) \\ P(A \cap B)^c = P(A^c \cup B^c)$$

If A is impossible event, then $P(A) = 0$.

If A is sure event, then $P(A) = 1$.

$$P(\text{only } A \text{ happens}) = P(A \cup B^c) = P(A) - P(A \cap B) \\ P(\text{only } B \text{ happens}) = P(A^c \cup B) = P(B) - P(A \cap B)$$

10.5 Results of permutation and combination

1. A sample of n elements one by one without replacement from the group of N elements can be selected in

$$\binom{N}{n} \text{ or } {}^N C_n \text{ ways.}$$

2. From a population of size N elements, containing N_1 elements of first type, N_2 elements of second type, and so on N_k elements of k^{th} type, a sample of size n containing n_1 elements of first type, n_2 elements of second type, and so on n_k elements of k^{th} type can be selected in following number of ways:

$$\binom{N}{N_1} \times \binom{N}{N_2} \times \cdots \times \binom{N}{N_k}$$

where, $N_1 + N_2 + \cdots + N_k = N$

3. The collection of “ N ” digits or letters, containing N_1 of first type, N_2 of second type, and so on N_k of k^{th} type can be rearranged in following number of ways:

$$\frac{N!}{N_1! \times N_2! \times \cdots \times N_k!}$$

where, $N_1 + N_2 + \cdots + N_k = N$

Further, if all the digits are occurring only once or all digits are distinct, then the number of possible arrangements = $N!$

Example 10.1

In manufacturing certain component, two independent defects are likely to occur with the corresponding probabilities 0.05 and 0.1. What is the probability that a randomly chosen component:

- (a) does not have either kind of defect?
- (b) has only one kind of defect?

Solution

Let, event A = defect of 1st type and event B = defect of 2nd type.

Given, $P(A) = 0.05$ and $P(B) = 0.1$

The defects are independent.

$$\text{Thus, } P(A \cap B) = P(A) \times P(B) = 0.05 \times 0.1 = 0.005$$

Now,

$$\begin{aligned} P[\text{does not contain either kind of defect}] &= 1 - P[\text{contains either kind of defect}] \\ &= 1 - P[A \cup B] = 1 - (P(A) + P(B) - P(A \cap B)) \\ &= 1 - (P(A) + P(B) - P(A) \times P(B)) \\ &= 1 - [0.05 + 0.1 - 0.005] \\ &= 1 - (0.145) = 0.855 \end{aligned}$$

That is, 85.5% of items may be non-defective.

Thus,

$$\begin{aligned} P[\text{item has only one kind of defect}] &= P\{\text{[item contains only 1st kind] or [item contains only 2nd kind]}\} \\ &= P(\text{only } A \text{ happens}) + P(\text{only } B \text{ happens}) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) \\ &= 0.05 - 0.005 + 0.1 - 0.005 \\ &= 0.15 - 0.010 \\ &= 0.14 \end{aligned}$$

That is, there is 14% chance that the product will contain only one kind of defect.

Example 10.2

In the world cup tournament, probability that India will enter semifinal (A) is 0.5, probability that it will win final (B) is 0.6 and probability that it will win final given that it has entered in semifinals is 0.9. Find the probability that at least one of A or B will happen.

Solution

Here, suppose event A = India enters semifinal and event B = India will win final.

$$\text{Given, } P(A) = 0.5 \text{ and } P(B) = 0.6 \text{ and } P(B/A) = 0.9$$

Now,

$$\begin{aligned} P[\text{At least one of } A \text{ & } B \text{ will happen}] &= P[A \cup B] \\ &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(B/A) \times P(A) \\ &= 0.5 + 0.6 - (0.9 \times 0.5) = 1.1 - 0.45 \\ &= 0.65 \end{aligned}$$

Thus, there is 65 % chance that at least one of two will happen.

Example 10.3

A machine is made up of the three components A , B , and C . It works, only if all the three components are working. The probabilities that components A , B , and C will fail are 0.01, 0.1, and 0.02, respectively. What is the probability that the machine will work?

Solution

Let, event A = component A works, event B = component B works and event C = component C works.

Thus,

A^c = component A fails, B^c = component B fails and C^c = component C fails.

Also,

$$P(A^c) = 0.01, P(B^c) = 0.1 \text{ and } P(C^c) = 0.02$$

Now,

$$\begin{aligned} P[\text{that machine works}] &= P[A \text{ works} \& B \text{ works} \& C \text{ works}] = P[A \cap B \cap C] \\ &= P(A) \times P(B) \times P(C) \\ &= [1 - P(A^c)] \times [1 - P(B^c)] \times [1 - P(C^c)] \\ &= [1 - 0.01] \times [1 - 0.1] \times [1 - 0.02] \\ &= 0.99 \times 0.9 \times 0.98 \\ &= 0.87 \end{aligned}$$

That is, there is 87% chance that the machine works.

Example 10.4

A student can win prize X with probability 0.4 and win at least one of the two prizes X and Y with probability 0.7. Find the probability that he will win prize Y , if

- (a) the events are mutually exclusive.
- (b) the events are independent.

Solution

Let, event A = student wins prize X , event B = student wins prize Y .

Given, $P(A) = 0.4$ and $P(A \cup B) = 0.7$.

We have to find $P(B)$,

- (a) If the events are mutually exclusive, that is $P(A \cap B) = 0$

Therefore,

$$\text{Therefore, } P(A \cup B) = P(A) + P(B)$$

$$\begin{aligned} \text{Therefore, } P(B) &= P(A \cup B) - P(A) \\ &= 0.7 - 0.4 \\ &= 0.3 \end{aligned}$$

Thus, there is 30% chance of student winning prize Y , if A, B are mutually exclusive.

(b) if the events are independent that is $P(A \cap B) = P(A) \times P(B)$

Therefore

$$\begin{aligned} \text{Therefore, } P(B) &= (A \cup B) - P(A) + P(A) \times P(B) \\ &= 0.7 - 0.4 + 0.4 \times 0.3 \\ &= 0.42 \end{aligned}$$

Thus, there is 42% chance of student winning prize Y , if A, B are independent.

Example 10.5

A box contains 4 white and 3 black balls and another box contains 3 white and 4 black balls. If one ball is chosen at random from each box, then what is the probability that the balls will be of different colors?

Solution

Given that, Box-1 contains 4 white (W) balls and 3 black (B) balls.

Box-2 contains 3 white (W) balls and 4 black (B) balls.

Now,

$$\begin{aligned} P[\text{balls chosen are of different colors}] &= P[(W \text{ and } B) \text{ or } (B \text{ and } W)] \\ &= P[(W \cap B) \cup (B \cap W)] \\ &= P(W \cap B) + P(B \cap W) \\ \\ &= P[W \text{ from 1}^{\text{st}}] \cdot P[B \text{ from 2}^{\text{nd}}] + P[B \text{ from 1}^{\text{st}}] \cdot P[W \text{ from 2}^{\text{nd}}] \\ &= \frac{4}{7} \times \frac{4}{7} + \frac{3}{7} \times \frac{3}{7} \\ &= \frac{25}{49} \\ &= 0.51 \end{aligned}$$

Thus, there is 51% chance that balls are of different colors.

Example 10.6

A box contains 10 cards that are numbered one to ten. Two cards are drawn one by one without replacement, what is the probability that both cards selected will show odd number.

Solution

Given that, there are 10 cards numbered from 1 to 10.

Two cards are selected one by one without replacement
Now,

Total number of ways of selecting two cards out of 10 = ${}^{10}C_2 = 45$
Total number of odd cards is 5,

Therefore,

Total number of ways of selecting two odd cards = ${}^5C_2 = 10$

$$P[\text{both cards are odd}] = 10/45 = 0.22$$

Example 10.7

Three bags X , Y , and Z are having composition of balls as 4 white and 6 black, and 3 white and 7 black and 6 white and 4 black, respectively. If two balls are selected at random from any one of the bags, what is the probability that both balls are white?

Solution

Suppose, event X = bag X is selected, event Y = bag Y is selected, event Z = bag Z is selected,

Also,

Suppose, event A = two white balls are selected,

Now,

$$P(X) = P(Y) = P(Z) = \frac{1}{3}$$

Therefore,

$$\begin{aligned} P(A) &= P[(A \cap X) \cup (A \cap Y) \cup (A \cap Z)] \\ &= P(A \cap X) + P(A \cap Y) + P(A \cap Z) \\ &= P(A/X)P(X) + P(A/Y)P(Y) + P(A/Z)P(Z) \\ &= \frac{\binom{4}{2}}{\binom{10}{2}} \times \frac{1}{3} + \frac{\binom{3}{2}}{\binom{10}{2}} \times \frac{1}{3} + \frac{\binom{6}{2}}{\binom{10}{2}} \times \frac{1}{3} \\ &= 0.1777. \end{aligned}$$

Example 10.8

If a five-figure number is formed by the digits 1,2,3,4, and 5 without repetition. What is the probability that, the number formed is divisible by two?

Solution

The total number of ways of forming five digit number without repetition = $5! = 120$.
 Also, The number of ways in which number formed divisible by 2 = $4! + 4! + 4! = 72$
 Thus,

$$P[\text{The number formed is divisible by 2}] = 72/120 = 0.6$$

Example 10.9

A group of students appeared for three papers. It was found that 25% have passed paper 1, 18% have passed paper 2, 15% have passed paper 3, 10% have passed paper 1 and 2, 7% have passed 2 and 3, 6% have passed paper 1 and 3, and 4% have passed all 3. If a student is chosen at random, what is the probability that

- (a) he has passed at least one paper?
- (b) he has passed only one paper?

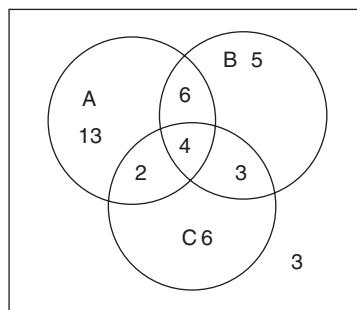
Solution:

Suppose,

Event A = student passed paper 1, event B = student passed paper 2,
 event C = student passed paper 3.

Given: $P(A) = 0.25$, $P(B) = 0.18$, $P(C) = 0.15$, $P(A \cap B) = 0.10$, $P(B \cap C) = 0.07$,
 $P(A \cap C) = 0.06$ and $P(A \cap B \cap C) = 0.04$

Meanings of the above probabilities are interpreted in Fig. 10.2



10.2

$$(a) P[\text{Passed in at least one paper}] = 1 - P[\text{failed in all}] = 1 - \frac{61}{100} = 0.39$$

$$(b) P[\text{passed in only one}] = P[(\text{only } 1^{\text{st}}) \text{ or } (\text{only } 2^{\text{nd}}) \text{ or } (\text{only } 3^{\text{rd}})] \\ = P[(\text{only } 1^{\text{st}}) \cup (\text{only } 2^{\text{nd}}) \cup \text{only } 3^{\text{rd}}]$$

$$\begin{aligned} &= P(\text{only 1}^{\text{st}}) + P(\text{only 2}^{\text{nd}}) + P(\text{only 3}^{\text{rd}}) \\ &= 0.13 + 0.05 + 0.06 \\ &= 0.24 \end{aligned}$$

10.6 Exercise

1. What is probability? What is its use? State the laws of the probability.
2. A machine has two electric fuses. The machine stops working if both fuses burn. The probability that a fuse will burn in a month is 0.25. What is the probability that the machine will not stop working in next month?
3. The probability that a movie will get award for good acting is 0.5, the probability that it will get award for good direction is 0.4 and probability that it will get awards for both is 0.3. Find the probability that the movie will get at least one award.
4. Two boxes of garments contain 10 red, 5 black, and 5 red and 10 black garments. If two garments are chosen from each box, what is the probability that all four garments are of same color?
5. Two boxes contain 5 gold 4 silver and 4 gold 6 silver coins. If one coin is selected at random from each box, what is probability that
 - (i) the coins will be of same metal?
 - (ii) the coins will be of different metals?
6. Two boxes each contain 5 white and 5 black balls. If two balls are chosen at random from each box, what is the probability that all balls are of same color?
7. Out of 20 employees of a textile store five are graduate. Three employees are selected at random, what is the probability that
 - (i) all of them are graduate?
 - (ii) only one is graduate?
8. A bag contains 7 red, 12 white, and 4 green balls. If three balls are selected at random from the bag, what is the probability that
 - (i) all three balls are of different colors?
 - (ii) all three balls are of same color?

9. Two groups of the students contain 4 boys, 6 girls and 5 boys, 5 girls. If one student is selected at random from each group, what is the probability that
 - (i) both will be of boys or girls?
 - (ii) one will be boy and one will be girl?
10. A bag contains 3 white and 7 black balls and another bag contains 7 white and 3 black balls. If one ball is chosen at random from each bag, what is the probability that
 - (i) both balls will be of same color?
 - (ii) both balls will be of different colors?

Probability distributions

11.1 Basic concepts

In Chapter 10, we have seen that a manufacturer of garments may be interested in knowing what is the chance that the consignment of packs of 10 garments each dispatched to the customer will be rejected by the customer. The spinning master may want to know, how many count tests made on a production of yarn will show count in between 26 and 29 or >30 or <25 , etc. All the above questions can be easily answered if the corresponding probability distribution is known. The concept of probability distribution can be easily understood by understanding following basic terms.

Sample space

As already discussed in Chapter 10, the sample space is the collection of all possible outcomes of a random experiment, and it is denoted by Ω or “S.”

For example,

$\Omega = \{H, T\}$ or $\Omega = \{1, 2, 3, 4, 5, 6\}$ are the sample spaces.

Random variable

The variable whose values are associated with the elements of the sample space and hence depend on the chance is called the random variable. The notations X, Y, Z , etc., always denote the variables.

For example, for the random experiment of tossing a coin with the sample space $\Omega = \{H, T\}$, the random variable X can be defined as follows:

$$\begin{aligned} X &= 0 && \text{if } H \text{ occurs} \\ X &= 1 && \text{if } T \text{ occurs.} \end{aligned}$$

Discrete random variable and continuous random variable are the two different types of the random variable.

Discrete random variable

If the possible values of the random variable are finite or countable infinite, then the random variable is called the discrete random variable.

For example:

$$\begin{aligned} X = 0 & \quad \text{if H occurs} \\ X = 1 & \quad \text{if T occurs} \end{aligned}$$

X = number on the upper face of the die,

X = number of defective ring bobbins in a pack of bobbins,

X = number of accidents in a textile mill,

are the discrete random variables.

Continuous random variable

If the possible values of the random variable are unaccountably infinite, that is if the random variable can take all possible values within certain range or interval, then it is called the continuous random variable.

For example,

X = count of the yarn

X = strength of the fabric

X = length of the fiber

are the continuous random variables.

11.2 Probability distribution of a random variable

The distribution of the probabilities to all possible values of the random variable, in such a way that the total of the distributed probabilities is one is called the probability distribution of the random variable.

For example, for the random experiment of tossing a coin with the sample space $\Omega = \{H, T\}$ and the random variable X defined as above, the probability distribution can be given by Table 11.1

Table 11.1

X	Probability $P(X)$
0	0.5
1	0.5
Total	1

Similarly, for the random experiment of throwing a die with the sample space

$\Omega = \{1, 2, 3, 4, 5, 6\}$ and the random variable X representing number on the upper face of the die the probability distribution can be given by Table 11.2 as follows:

Table 11.2

X	1	2	3	4	5	6	Total
Probability	1/6	1/6	1/6	1/6	1/6	1/6	1

Discrete probability distribution and continuous probability distribution are two different types of the probability distribution.

Discrete probability distribution

The probability distribution of the discrete random variable is called the discrete probability distribution. The discrete probability distribution is always represented by the function called the probability mass function (pmf), and it is denoted by the notation $P(x)$ or $P(X = x)$. Every probability mass function satisfies the condition

$$\sum_x P(x) = 1$$

For example,

$$P(x) = \frac{1}{6} \quad x = 1, 2, 3, 4, 5, 6$$

is the pmf of the random variable X , then

$$\sum_x P(x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

Example 11.1

Find value of k for the following probability distribution and find the following probabilities $p(X = 2)$, $p(X > 3)$ and $p(X \leq 4)$.

x	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	k	$2k$	$3k$	$4k$	$5k$	$6k$	$7k$	$8k$	$9k$	$10k$

Solution

The given pmf must satisfy,

$$\sum_x P(x) = 1$$

Here,

$$\sum_{x=0}^{10} P(x) = k + 2k + \dots + 10k = k(1+2+\dots+10) = k \times \frac{10 \times (10+1)}{2} = 55k = 1$$

$$\text{Therefore, } k = \frac{1}{55}$$

Hence, by substituting value of k , the probability distribution of random variable X is given in Table 11.1

Table 11.3

x	1	2	3	4	5	6	7	8	9	10
$P(X=x)$	1/55	2/55	3/55	4/55	5/55	6/55	7/55	8/55	9/55	10/55

Further,

$$P(X=2) = \frac{2}{55}$$

$$\begin{aligned} P(X > 3) &= 1 - p(X \leq 3) = 1 - (P(1) + P(2) + P(3)) \\ &= 1 - \frac{6}{55} = \frac{49}{55} = 0.8909 \end{aligned}$$

$$\begin{aligned} P(X \leq 4) &= p(1) + p(2) + p(3) + p(4) \\ &= \frac{10}{55} = 0.1818 \end{aligned}$$

Continuous probability distribution

The probability distribution of the continuous random variable is called the continuous probability distribution. The continuous probability distribution is always represented by a function called the probability density function (pdf), and it is denoted by the notation $f(x)$. Every pdf satisfies the condition.

$$\int f(x) dx = 1$$

Note that,

1. In case of continuous distribution, equality does not have any meaning. That is,

$$P(a \leq X \leq b) = p(a < x < b)$$

2. Thus, area under probability distribution curve in between a and b

$$= P(a < X < b) = P(a \leq X \leq b) = \int_a^b f(x) dx$$

Example 11.2

Show that following function is a probability density function.

$$f(x) = \frac{1}{(b-a)} \quad a < x < b$$

Solution

Every pdf satisfies the condition.

$$\int f(x)dx = 1$$

Here,

$$\begin{aligned} \int f(x)dx &= \int_a^b \frac{1}{b-a} dx \\ &= \frac{1}{b-a} \int_a^b dx \\ &= \frac{1}{b-a} \times (x)_a^b \\ &= \frac{1}{b-a} \times (b-a) \\ &= 1 \end{aligned}$$

Thus, the given function is a pdf.

Example 11.3

Find the value of “ k ” such that the function $f(x) = k(2 - x^2)$ $1 < x < 3$, is a pdf and hence find the probability $P(2 < X < 3)$.

Solution

For any function $f(x)$ to be pdf, it must satisfy the condition

$$\int f(x)dx = 1$$

Here,

$$\int f(x)dx = \int_1^3 k(2 - x^2)dx = 1$$

$$\text{Therefore, } k \int_1^3 (2 - x^2)dx = 1$$

$$\text{Therefore, } k \left[2x - \frac{x^3}{3} \right]_1^3 = 1$$

Hence $k = -\frac{3}{14}$

Thus,

$$f(x) = \frac{3}{14}(x^2 - 2) \quad 1 < x < 3$$

Now,

$$\begin{aligned} P(2 < X < 3) &= \int_2^3 \frac{3}{14}(x^2 - 2) dx \\ &= \frac{3}{14} \times \left[\frac{x^3}{3} - 2x \right]_2^3 \\ &= \frac{3}{14} \times \frac{13}{3} = \frac{13}{14} = 0.9286 \end{aligned}$$

Thus, there is 92.8 chance that the random variable X lies in between 2 and 3.

11.3 Some properties of the random variable and its probability distribution

There are some important properties that are associated with the random variable and its probability distribution. These properties are useful in studying basic structure and the nature of the probability distribution.

Expectation of a random variable X

If X is a random variable under study then, expectation of the random variable is denoted by the notation $E(X)$ and is defined as follows:

$$\begin{aligned} E(X) &= \sum_x x \cdot p(x) \quad \text{if } X \text{ is discrete variable} \\ &= \int x \cdot f(x) dx \quad \text{if } X \text{ is continuous variable} \end{aligned}$$

This is also called the theoretical mean of the random variable X .

Variance of a random variable X

If X is a random variable under study then, variance of the random variable is denoted by the notation $V(X)$ or σ^2 and is defined as follows:

$$\sigma^2 = V(X) = E(X - E(X))^2 = E(X^2) - E(X)^2$$

Thus, standard deviation of the random variable $X = SD(X) = \sqrt{\sigma^2}$

Moments of a random variable

Moments are the generalized concepts of mean and variance. Raw moments and central moments are the two different types of the moments.

Raw moments

If X is a random variable under study then, r^{th} raw moment of the random variable is denoted by the notation μ'_r and is defined as follows:

$$\begin{aligned}\mu'_r = E(X^r) &= \sum_x x^r \cdot p(x) && \text{if } X \text{ is discrete variable} \\ &= \int x^r \cdot f(x) dx && \text{if } X \text{ is continuous variable}\end{aligned}$$

Note that, if $r = 1$, then $\mu'_1 = E(X)$ = mean of r.v. X

Central moments

If X is a random variable under study then, r^{th} central moment of the random variable is denoted by the notation μ_r and is defined as follows:

$$\mu_r = E(X - E(X))^r$$

Note that,

$$\begin{array}{ll} \text{For } r = 1 & \mu_1 = 0 \\ \text{For } r = 2 & \mu_2 = \sigma^2 = V(X) = E(X - E(X))^2 = E(X^2) - E(X)^2 \end{array}$$

Moment generating function

The function, which provides moments, is called the moment generating function (MGF). It is denoted by the notation $M_x(t)$ and is defined as follows:

$$\begin{aligned}M_x(t) = E(e^{tX}) &= \sum_x e^{tx} \cdot p(x) && \text{if } X \text{ is discrete variable} \\ &= \int e^{tx} \cdot f(x) dx && \text{if } X \text{ is continuous variable}\end{aligned}$$

Further,

$$E(X^r) = \left\{ \frac{d^r}{dt^r} M_x(t) \right\}_{t=0}$$

Thus,

$$\text{Mean} = E(X) = \left\{ \frac{d}{dt} M_x(t) \right\}_{t=0} \quad \text{for } r = 1$$

11.4 Some standard probability distributions

There are some standard probability distributions that have large number of applications in the real life, particularly in the industries. These probability distributions are classified into discrete and continuous types according to their associated random variables. There are so many standard discrete and continuous probability distributions out of which only few probability distributions that have large number of applications in textile engineering are listed below and will be discussed in the subsequent chapters.

Some standard discrete probability distributions

1. Binomial probability distribution
2. Poisson probability distribution

Some standard continuous probability distributions

1. Normal probability distribution
2. Chi-square (χ^2) probability distribution
3. Student's " t " probability distribution
4. " F " probability distribution

11.5 Exercise

1. Explain the following terms:
 - (i) Random variable
 - (ii) Discrete random variable
 - (iii) Continuous random variable
2. What is probability distribution? What are its types? Describe any one.
3. Find the value of k for the following pmf; hence, find $P[1 < X < 4]$ and $P[X > 2]$.

x	0	1	2	3	4	5	6
$P[X = x]$	$2k$	$4k$	$6k$	$8k$	$10k$	$12k$	$14k$

4. Find the value of constant k for the following probability mass function.

$$P(X = x) = \frac{k}{10} \quad x = 0, 1, 2, \dots, 20$$

Also find the probability $P[X \geq 3]$.

5. Find the value of the constant k for the following probability mass function.

$$P(X = x) = \frac{kx}{2} \quad x = 0, 1, 2, \dots, 5$$

Also find $P(x < 3)$.

6. Find the value of constant “ k ” for the following probability density function.

$$f(x) = 3kx^2 \quad 1 \leq x \leq 5$$

Also find the probability $P[2 \leq X \leq 3]$.

7. Find the value of the constant k for the following probability density function. Also find $P(2 < X < 5)$.

$$f(x) = ke^{-x} \quad 0 \leq x \leq \infty$$

8. A continuous random variable X has pdf

$$f(x) = k(1 + x^2) \quad 0 \leq x \leq 1$$

Find the value of k and hence compute $P(X \geq 0.4)$.

Standard discrete probability distributions

12.1 Binomial probability distribution

This is the most popularly used discrete probability distribution that has large number of applications in real life. It is defined as A discrete random variable “ X ” denoting number of successes in “ n ” independent Bernoulli trials is said to follow binomial probability distribution if its pmf is as follows:

$$p(X = x) = p(x) = \binom{n}{x} p^x q^{n-x} \quad x = 1, 2, \dots, n$$

where, the random experiment, which has only two possible outcomes, is called the Bernoulli trial. These two outcomes of the Bernoulli trial are generally called the success and failure.

Also, $p = P[\text{Success in a trial}]$ and $q = P[\text{Failure in a trial}]$

Thus, $p + q = 1$ and $q = 1 - P$

Properties of the binomial probability distribution

1. In case of binomial distribution n and p are known as the parameters of the binomial distribution because it depends on these two values. Hence, the binomial distribution can be represented by the notation:

$$X \sim B(n, p)$$

For example, $X \sim B(5, 0.5) \Rightarrow n = 5$ trials, $p = P(\text{Success}) = 0.5$

X represents number of successes in 5 trials.

$$p(X = x) = p(x) = \binom{5}{x} 0.5^x 0.5^{5-x} \quad x = 1, 2, \dots, 5$$

2. The pmf of Binomial probability distribution satisfies the condition $\sum p(x) = 1$

Proof: Here,

$$\sum_x p(x) = \sum_{x=0}^n \binom{n}{x} p^x q^{n-x} = \binom{n}{0} p^0 q^{n-0} + \binom{n}{1} p^1 q^{n-1} + \dots + \binom{n}{n} p^n q^{n-n} = (p+q)^n = 1$$

3. The mean of the binomial distribution is “ np .” That is $E(X) = np$.
4. The variance of the binomial distribution is “ npq .”
That is $V(X) = npq$ Therefore, $SD(X) = \sqrt{npq}$
5. In case of binomial probability distribution $V(X)$ is always smaller than $E(X)$.
6. The relation

$$p(x+1) = \frac{p}{q} \cdot \frac{n-x}{x+1} \cdot p(x) \quad x = 0, 1, \dots, (n-1)$$

is called the recurrence relation for the binomial probability distribution.

Derivation of mean and variance for the binomial distribution using MGF

Let, $X \sim B(n, p)$

$$\text{Therefore, } p(x) = \binom{n}{x} p^x q^{n-x} \quad x = 1, 2, \dots, n$$

Now, MGF of X is

$$M_X(t) = E(e^{tx}) = \sum_x e^{tx} \cdot p(x) = \sum_{x=0}^n e^{tx} \cdot \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^n \binom{n}{x} (pe^t)^x q^{n-x} = (q + pe^t)^n$$

Now,

$$\begin{aligned} \text{Mean} = E(X) &= \left[\frac{d}{dt} M_X(t) \right]_{t=0} \\ &= \left[\frac{d}{dt} (q + pe^t)^n \right]_{t=0} = \left[n(q + pe^t)^{n-1} \cdot pe^t \right]_{t=0} = np \end{aligned}$$

Also,

$$\begin{aligned} E(X^2) &= \left\{ \frac{d^2}{dt^2} M_X(t) \right\}_{t=0} \\ &= \left[\frac{d^2}{dt^2} (q + pe^t)^n \right]_{t=0} = \left[\frac{d}{dt} n \cdot (q + pe^t)^{n-1} \cdot pe^t \right]_{t=0} \\ &= \left[n \cdot (q + pe^t)^{n-1} \cdot pe^t + pe^t \cdot n \cdot (n-1) \cdot (q + pe^t)^{n-2} \cdot pe^t \right]_{t=0} \\ &= (np + n(n-1)p^2) \end{aligned}$$

$$\begin{aligned}\text{Therefore, } V(X) &= E(X^2) - (E(X))^2 \\ &= (np + n(n-1)p^2) - (np)^2 = np - np^2 = np(1-p) = npq\end{aligned}$$

Example 12.1

1. A company produces knitting needles and supplies them in the packs of 10 needles each. The manufacturer knows from his past experience that 5% of the needles produced by his company may be defective. If the manufacturer selects one pack at random from the production line and checks it, what is the probability that
 - (i) he will find exactly one defective needle?
 - (ii) he will find all defective needles?
 - (iii) he will find at least one defective needle?

Solution

Here, X —Number of defective needles in a pack of 10 needles.

Let, $X \sim B(n = 10, p)$

Given that $p = 5/100 = 0.05$

Therefore,

$$p(X=x) = p(x) = \binom{10}{x} 0.05^x 0.95^{10-x} \quad x = 1, 2, \dots, 10$$

Now,

$$(i) p\{\text{exactly one defective needle}\} = p\{X=1\} = \binom{10}{1} 0.05^1 0.95^{10-1} = 0.3151$$

$$(ii) p\{\text{all defective needles}\} = p\{X=10\} = \binom{10}{10} 0.05^{10} 0.95^{10-10} = 0$$

$$\begin{aligned}(iii) p\{\text{at least one defective needle}\} &= p\{X \geq 1\} = 1 - p\{X=0\} \\ &= 1 - 0.5987 = 0.4013\end{aligned}$$

Example 12.2

A company produces knitted garments and supplies them in the packs of 10 garments each. The manufacturer knows from his past experience that 1% of the garments produced by his company may be defective. The customer rejects the pack if it contains two or more defective garments. If the manufacturer supplies such 500 packs to the customer, how many packs the customer will reject?

Solution

Here, X —Number of defective garments in a pack of 10 garments.

Let, $X \sim B(n = 10, p)$

Given that $p = 1/100 = 0.01$

Therefore,

$$p(X = x) = p(x) = \binom{10}{x} 0.01^x 0.99^{10-x} \quad x = 1, 2, \dots, 10$$

Now,

$$\begin{aligned} P\{\text{Pack is rejected}\} &= p\{\text{at least two defective garments}\} \\ &= p\{X \geq 2\} \\ &= 1 - [p\{X = 0\} + p\{X = 1\}] \\ &= 1 - [0.9044 + 0.0914] \\ &= 0.0042 \end{aligned}$$

Thus, the number of packs rejected $= 500 \times 0.0042 = 2.1$

That is, approximately two packs will be rejected by the customer.

12.2. Poisson probability distribution

This is also most popularly used discrete probability distribution which has large number of applications in real life. It is defined as follows:

Definition

A discrete random variable X , denoting number of occurrences of an event during a fixed time period, interval or area, is said to follow Poisson probability distribution, if its pmf is as follows:

$$p(X = x) = p(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \quad x = 1, 2, \dots, \infty$$

where,

λ denotes average number of occurrences of an event during a fixed time period, interval, or area.

Note that, in this case the events under study are very common events but for the fixed period, area or interval they are rare events. For example, accidents, weaving defects, end breaks, and neps are very common events but for the fixed period, area, or interval are the rare events.

Properties of the Poisson probability distribution

1. The constant λ is called the parameter of the Poisson distribution. Hence, Poisson distribution can be represented by the notation

$$X \sim P(\lambda).$$

2. The pmf of Poisson distribution satisfies the condition $\sum_x p(x) = 1$

Proof: Here,

$$\begin{aligned}\sum_x p(x) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda} \cdot \lambda^x}{x!} = e^{-\lambda} + e^{-\lambda} \cdot \lambda + e^{-\lambda} \frac{\lambda^2}{2!} + \dots \\ &= e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \dots \right) = e^{-\lambda} \cdot e^{\lambda} = 1\end{aligned}$$

3. The mean of the Poisson distribution is m, that is $E(X) = \lambda$.
 4. The variance of the Poisson distribution is also m, that is $V(X) = \lambda$.
 5. Recurrences relationship of the Poisson distribution is

$$p(x+1) = \frac{\lambda}{x+1} \cdot p(x) \quad x = 0, 1, \dots, (n-1)$$

Derivation of mean and variance for Poisson distribution

Let, $X \sim P(\lambda)$

$$\text{Therefore, } p(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \quad x = 1, 2, \dots, \infty$$

Now, MGF of X is

$$M_X(t) = E(e^{tX}) = \sum_x e^{tx} \cdot p(x) = \sum_{x=0}^{\infty} e^{tx} \cdot \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \cdot (\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{-\lambda(1-e^t)}$$

Now,

$$\begin{aligned}\text{Mean} &= E(X) = \left[\frac{d}{dt} M_X(t) \right]_{t=0} \\ &= \left[\frac{d}{dt} e^{-\lambda(1-e^t)} \cdot \lambda e^t \right]_{t=0} = [e^{-\lambda(1-e^t)} \cdot \lambda e^t]_{t=0} = \lambda\end{aligned}$$

Also,

$$\begin{aligned}E(X^2) &= \left\{ \frac{d^2}{dt^2} M_X(t) \right\}_{t=0} \\ &= \left[\frac{d^2}{dt^2} e^{-\lambda(1-e^t)} \cdot \lambda e^t \right]_{t=0} = \left[\frac{d}{dt} e^{-\lambda(1-e^t)} \cdot \lambda e^t \right]_{t=0}\end{aligned}$$

$$= \left[e^{-\lambda(1-e^t)} \cdot \lambda e^t + \lambda e^t \cdot e^{-\lambda(1-e^t)} \cdot \lambda e^t \right]_{t=0} \\ = (\lambda + \lambda^2)$$

$$\text{Therefore, } V(X) = E(X^2) - (E(X))^2 \\ = (\lambda + \lambda^2) - (\lambda)^2 = \lambda$$

Example 12.3

If on an average, 12 accidents occur during 1 year in a textile mill, what is the probability that in the coming year there will be

- (i) no accident?
- (ii) at most two accidents?

Solution

Here, X -number of accidents during 1 year

Let, $X \sim P(\lambda)$

Where,

$\lambda = 12$ = average number of accidents during 1 year.

$$p(X=x) = p(x) = \frac{e^{-12} \cdot 12^x}{x!} \quad x = 1, 2, \dots, \infty$$

$$(i) p\{\text{no accidents during next year}\} = P(X=0) \\ = p(x) = \frac{e^{-12} \cdot 12^0}{0!} = 6.144 \times 10^{-6} \approx 0$$

$$(ii) p\{\text{there are at most two accidents in 1 year}\} = P[x \leq 2] \\ = P(x=0) + P(x=1) + P(x=2) \\ = e^{-12} + e^{-12} \times 12 + e^{-12} \times 144/2 \\ = .000006144 + 0.00007373 + 0.0004424 \\ = 0.0005$$

Example 12.4

A 100 m² roll of the fabric is expected to contain on an average eight weaving defects scattered uniformly over the full fabric. If the fabric is cut into four pieces of equal size, what is the probability that

- (i) a piece will be free from the weaving defects?
- (ii) all the pieces will be free from the weaving defects?

Solution

Here, X —number of weaving defects in a piece of size 25 m^2

Let, $X \sim P(\lambda)$

Where,

$$\Lambda = \text{average number of weaving defects in a piece of size } 25 \text{ m}^2 = \frac{25 \times 8}{100} = 2$$

$$p(X=x) = p(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \quad x = 1, 2, \dots, \infty$$

$$(i) P[\text{a piece will be free from weaving defect}] = P(x=0) = e^{-\lambda} = 0.1353$$

$$(ii) P[\text{all pieces are free from weaving defects}]$$

$$\begin{aligned} &= P([1^{\text{st}} \text{ is free}] \cap [2^{\text{nd}} \text{ is free}] \cap [3^{\text{rd}} \text{ is free}] \cap [4^{\text{th}} \text{ is free}]) \\ &= P(1^{\text{st}} \text{ is free}) \cdot P(2^{\text{nd}} \text{ is free}) \cdot P(3^{\text{rd}} \text{ is free}) \cdot P(4^{\text{th}} \text{ is free}) \\ &= (0.1353)^4 \\ &= 0.0003354 \end{aligned}$$

That is, there is approximately 0% chance that all pieces are free from weaving defects.

12.3 Poisson approximation to the binomial distribution

Let $X \sim B(n, p)$.

Now

If $n \rightarrow \infty$ that is, number of trials are very large

$p \rightarrow 0$ that is, probability of success in a trial is very small

then,

$X \sim P(\lambda = np)$ approximately.

Practically, if $np < 5$, then Poisson approximation can be used.

Example 12.5

A company produces air jet loom valves and supplies them in the pack of 50 valves each. The manufacturer knows from the past experience that 1% of the valves produced by his company may be defective. If the manufacturer selects any one pack at random from the production line and checks it, what is the probability that he will find exactly one defective valve?

Solution

Here, X : Number of defective valves in a pack of 50

Let, $X \sim B(n, p)$

Given that,

$n = 50$ (Very large)

and

$p = p(\text{success}) = P(\text{defective valves}) = 1\% = 0.01$ (Very small)

Hence,

$X \sim P(\lambda = np = 0.5)$ approximately.

Thus,

$$p(x) = \frac{e^{-0.5} \cdot 0.5^x}{x!} \quad x = 1, 2, \dots, 50$$

Now,

$P(\text{the pack contains 1 defective valve}) = P[X = 1] = e^{-0.5} \cdot 0.5 = 0.3033$.

That is, there is approximately 30% chance that the pack will contain only one defective valve.

12.4 Fitting of binomial and Poisson probability distributions

Suppose the data for which binomial distribution is to be fitted is as follows:

X	0	1	2		n	Total
Frequency	f_1	f_2			f_n	N

Now, fitting of the probability distribution means, finding the probabilities and expected frequencies for all the possible values of the random variable X with the help of the given data.

Algorithm for fitting binomial distribution

Step 1 Identify the random variable X and its possible values “ n .”

Step 2 Assume $X \sim B(n, p)$

Step 3 If “ p ” is known, go to step 4

Other wise if p is unknown find it using the formula

$$p = \frac{1}{n} \left(\frac{\sum f_i x_i}{N} \right)$$

Step 4 Using given “ p ” or calculated “ p ” in the step number 3, find all the probabilities with the help of the probability mass function of binomial distribution or the recurrence relationship for binomial distribution.

Step 5 Find expected frequencies by multiplying each probability by N .

Algorithm for fitting Poisson distribution

Step 1 Identify the random variable X and its possible values.

Step 2 Assume $X \sim P(\lambda)$

Step 3 If “ λ ” is known, go to step 4

Otherwise, if λ is unknown, find it using the formula

$$\lambda = \left(\frac{\sum f_i x_i}{N} \right)$$

Step 4 Using given “ λ ” or calculated “ λ ” in the step number 3, find all the probabilities with the help of the probability mass function of Poisson distribution or the recurrence relationship for Poisson distribution.

Step 5 Find expected frequencies by multiplying each probability by N .

Example 12.6

Following is the distribution of 100 samples of five garments each, according to the number of defective garments.

Number of defective garments	0	1	2	3	4	5	Total
Number of samples	52	20	12	8	5	3	100

Fit the Binomial probability distribution to the above data assuming,

- (i) Probability of the defective garments “ p ” is unknown.
- (ii) 15% of the garments are defective.

Solution

Here, X - Number of defective garments in a sample of 5.

Let $X \sim B(n = 5, p)$

- (i) Here, parameter “ p ” is unknown

X	0	1	2	3	4	5	Total
f_i	52	20	12	8	5	3	100
$f_i x_i$	0	20	24	24	20	15	103

$$p = \frac{1}{n} \left(\frac{\sum f_i x_i}{N} \right)$$

$$= 0.206$$

$$q = 0.794.$$

Required fitting of distribution is Table 12.1

Table 12.1

X	Observed frequency	$P(X=x)$	Expected frequency
0	52	0.3156	31.557490
1	20	0.4094	40.937300
2	12	0.2124	21.242030
3	8	0.0551	5.511156
4	5	0.0071	0.714923
5	3	0.0004	0.037097
	100	1	100

(ii) Here, parameter “ p ” is known and $p = 0.15$

Required fitting of distribution is Table 12.2

Table 12.2

X	Observed frequency	$P(X=x)$	Expected frequency
0	52	0.4437	44.370530
1	20	0.3915	39.150470
2	12	0.1382	13.817810
3	8	0.0244	2.438438
4	5	0.0022	0.215156
5	3	0.0001	0.007594
	100	1	100

Note that,

In the above example, the probabilities are calculated as follows:

Case I As $n = 5$ and $p = 0.206$

$$p(0) = \binom{5}{0} 0.206^0 \cdot (1 - 0.206)^{5-0} = 0.794^5 = 0.3156$$

$$p(1) = \binom{5}{1} 0.206^1 \cdot (1 - 0.206)^{5-1} = 0.4094$$

Case II As $n = 5$ and $p = 0.15$

$$p(0) = \binom{5}{0} 0.15^0 \cdot (1 - 0.15)^{5-0} = 0.85^5 = 0.4437$$

$$p(1) = \binom{5}{1} 0.15^1 \cdot (1 - 0.15)^{5-1} = 0.3915$$

Similarly remaining probabilities can be calculated in both cases by changing value of x from 0 to 5.

Example 12.7

Following is the distribution of 100 samples of garments, according to the number of defects observed in the garments.

Number of defects in a garment	0	1	2	3	4	5	Total
Number of samples	52	20	12	8	5	3	100

Fit the Poisson probability distribution to the above data assuming,

- (i) average number of the defects in a garment “ λ ” is unknown.
- (ii) average number of defect per garment as one.

Solution

Here, X -Number of defects in a garment.

Let $X \sim P(\lambda)$

- (i) Here parameter “ λ ” is unknown.

X	0	1	2	3	4	5	Total
f_i	52	20	12	8	5	3	100
$f_i x_i$	0	20	24	24	20	15	103

$$\lambda = \left(\frac{\sum f_i x_i}{N} \right) = 1.03$$

Required fitting of distribution is Table 12.3

Table 12.3

X	Observed frequency	$P(X = x)$	Expected frequency
0	52	0.3570	35.700700
1	20	0.3677	36.771720
2	12	0.1894	18.937430
3	8	0.0650	6.501852
4	5	0.0167	1.674227
5 and more	3	0.0041	0.414073
	100	1	100

- (ii) Here parameter λ is known and $\lambda = 1$

Required fitting of distribution is Table 12.4

Table 12.4

X	Observed frequency	$P(X=x)$	Expected frequency
0	52	0.3679	36.787940
1	20	0.3679	36.787940
2	12	0.1839	18.393970
3	8	0.0613	6.131324
4	5	0.0153	1.532831
5 and more	3	0.0037	0.365985
	100	1	100

Note that,

in the above example the probabilities are calculated as follows:

Case I As $\lambda = 1.03$

$$P(0) = \frac{e^{-1.03} \cdot 1.03^0}{0!} = e^{-1.03} = 0.3570$$

$$P(1) = \frac{e^{-1.03} \cdot 1.03^1}{1!} = e^{-1.03} \cdot 1.03 = 0.3677$$

Case II As $\lambda = 1$

$$P(0) = \frac{e^{-1} \cdot 1^0}{0!} = e^{-1} = 0.3679$$

$$P(1) = \frac{e^{-1} \cdot 1^1}{1!} = e^{-1} = 0.3679$$

Similarly remaining probabilities can be calculated in both cases by changing value of x from 0 to 5.

12.5 Exercise

1. Define “binomial probability distribution” and state its properties.
2. Define “Poisson probability distribution” and state its properties.
3. Derive the MFG of binomial distribution and show that mean of binomial distribution is greater than its variance.
4. Derive the MGF of Poisson probability distribution and find its mean and variance.

5. A company produces needles and supplies them in packs of 10 needles each. If 5% of the needles are defective, what is the probability that a pack will
 - (i) contain at least one defective needle?
 - (ii) all defective needles?
6. The probability that a woman does not know swimming is $2/5$. If seven women in a city are selected at random, find probability that
 - (i) four women know swimming.
 - (ii) at least one woman knows swimming.
7. In a computer center, there are six computers; the chance for their failure is same during the given period and is equal to $1/3$. Use binomial distribution to compute the probability that during given period
 - (i) at least two computers will fail.
 - (ii) exactly four computers will fail.
8. A company produces bulbs and supplies in pack of 50 bulbs each. If 2% of the bulbs are defective and customer rejects the pack if it contains more than two defective bulbs, how many packs out of 100 supplied will be rejected by the customer?
9. A company produces raincoats and 95% of its coats satisfy the waterproof test. If 5 raincoats are tested by the manager, using binomial probability distribution, find the probability that all coats will satisfy the test and probability that all coats will not satisfy the test.
10. A company produces air-jet machine valves and knows from the experience that one percent of the valves are defective. If the manager selects a sample of 10 valves and inspects it, using binomial probability distribution find the probability of getting at least one defective valve in the sample.
11. A company produces knitting needles and supplies them in the pack of 10 needles each. The manager knows from his experience that 5% of the needles are generally defective. If the manager inspects 100 packs, using binomial probability distribution, find the number of packs containing 3 or more defective needles.
12. There are on an average 15 end breaks on a ring-frame per 1000 spindle hours. If the ring-frame is observed for 200 spindle hours, what is the probability that it will show at most 3 end breaks?

13. There are on an average 10 weaving defects in 100 m^2 of fabric. If the fabric is cut into 5 pieces of equal size, what is the probability that a piece of fabric will contain at the most two weaving defects?
14. A manufacturer produces garments and supplies them in the packs of 50 garments each. He knows from his experience that 2% of his garments are defective. If the manufacturer checks any one pack at random, using Poisson approximation, find the probability that it will contain more than two defective garments.
15. In a loom shed of forty looms a jobber has to attend on an average three looms everyday. What is the probability that he has to attend more than four looms on a specific day?
16. Fit the Poisson probability distribution to the following data related to pieces of yarn.

No. of weak spots	0	1	2	3	4	5	6
No. of pieces	65	45	35	25	15	10	5

17. Following are the results of number of faults obtained from 500 pieces of yarn.

No. of faults	0	1	2	3	4	5
No. of pieces	150	150	130	55	40	20

Fit the Poisson distribution; assume average faults in a piece as one and find expected frequencies.

18. For the following data, fit the binomial probability distribution and find expected frequencies:

No. of defective blades	0	1	2	3	4	5
No. of packs	100	80	60	40	20	10

19. Following is the distribution of 300 packs of 5 needles each according to defective needles.

No. of defective needles	0	1	2	3	4	5
No. of packs	80	100	70	20	20	10

Fit the binomial probability distribution to the above data and find expected frequencies.

Standard continuous probability distributions

13.1 Normal probability distribution

This is the most important and the most popularly used continuous probability distribution, which has very large number of applications in real life because most of the variables of interest are measurable and their values are expected to be concentrated around mean. The normal probability distribution is defined as follows:

Definition

A continuous random variable “ X ” is said to follow normal probability distribution if its pdf is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad 0 < x < \infty$$

Properties of the normal probability distribution

1. In case of normal distribution μ and σ are known as the parameters as the normal distribution depends on these two values. Hence, the normal distribution can be represented by the notation

$$X \sim N(\mu, \sigma) \text{ or } X \sim N(\mu, \sigma^2)$$

2. The pdf of normal distribution satisfies the condition $\int f(x)dx = 1$

Proof: Here,

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &\text{put } \left(\frac{x-\mu}{\sigma}\right) = z \end{aligned}$$

$$\text{Therefore, } x = \mu + \sigma z$$

$$dx = \sigma dz$$

$$\begin{aligned} \text{Therefore, } \int_{-\infty}^{\infty} f(x)dx &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z)^2} \sigma dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z)^2} dz \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}(z)^2} dz \\ &\text{put } \frac{z^2}{2} = t \end{aligned}$$

$$\text{Therefore, } z = \sqrt{2t}$$

$$dz = \frac{1}{2\sqrt{2t}} \cdot 2dt$$

$$\begin{aligned} \text{Therefore, } \int_{-\infty}^{\infty} f(x)dx &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-t} \frac{1}{2\sqrt{2t}} \cdot 2dt = \frac{1}{\sqrt{\pi}} \int_0^{\infty} e^{-t} t^{-\frac{1}{2}} \cdot dt = \frac{1}{\sqrt{\pi}} \Gamma \left(\frac{1}{2} \right) \\ &= \frac{1}{\sqrt{\pi}} \cdot \sqrt{\pi} = 1 \end{aligned}$$

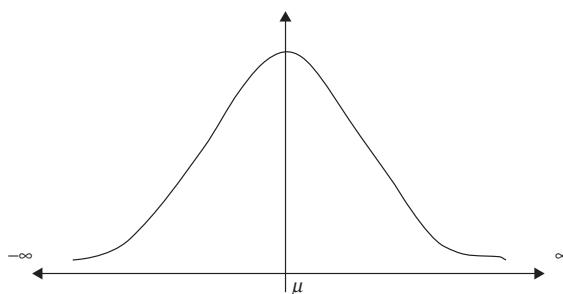
3. The mean of the normal distribution is μ . That is $E(X) = \mu$.

4. The variance of the normal distribution is σ^2 .

That is $V(X) = \sigma^2$ Therefore, $SD(X) = \sqrt{\sigma^2} = \sigma$

5. Graph of the normal distribution

The graph of the normal distribution is bell-shaped and symmetric about mean as shown below in Fig. 13.1.



13.1

Note that,

(a) As the normal distribution is symmetric distribution,

$$\text{Mean} = \text{Mode} = \text{Median} = \mu$$

- (b) Total area under normal curve = $P(-\infty < X < +\infty) = \int_{-\infty}^{\infty} f(x)dx = 1$
- (c) $P(-\infty < X < \mu) = P(\mu < X < +\infty) = 0.5$
- (d) $P(\mu - t < X < \mu) = P(\mu < X < \mu + t)$
- (e) $P(-\infty < X < \mu - t) = P(\mu + t < X < +\infty)$

Derivation of mean and variance for the binomial distribution using MGF

Let, $X \sim N(\mu, \sigma^2)$

$$\text{Therefore, } f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad 0 < x < \infty$$

Now, MGF of X is

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\text{put } \left(\frac{x-\mu}{\sigma}\right) = z$$

$$\text{Therefore, } x = \mu + \sigma z$$

$$dx = \sigma dz$$

$$\begin{aligned} \text{Therefore, } M_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(\mu+\sigma z)} \cdot e^{-\frac{1}{2}(z^2)} dz = \frac{2e^{t\mu}}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}(z^2+2t\sigma z)} dz \\ &= \frac{2e^{t\mu}}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}(z^2+2t\sigma z+t^2\sigma^2-t^2\sigma^2)} dz = \frac{2e^{t\mu+\frac{1}{2}t^2\sigma^2}}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}(z+t\sigma)^2} dz \end{aligned}$$

$$\text{put } z + t\sigma = x$$

$$dz = dx$$

$$\text{Therefore, } M_X(t) = \frac{2e^{t\mu+\frac{1}{2}t^2\sigma^2}}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}x^2} dx$$

$$\text{put } \frac{x^2}{2} = t$$

$$\text{Therefore, } x = \sqrt{2t}$$

$$dx = \frac{1}{2\sqrt{2t}} \cdot 2dt$$

$$\text{Therefore, } M_X(t) = \frac{2e^{t\mu+\frac{1}{2}t^2\sigma^2}}{\sqrt{2\pi}} \int_0^{\infty} e^{-t} \frac{1}{2\sqrt{2t}} \cdot 2dt = \frac{e^{t\mu+\frac{1}{2}t^2\sigma^2}}{\sqrt{\pi}} \int_0^{\infty} e^{-t} t^{\frac{1}{2}} \cdot dt$$

$$\text{Therefore, } M_X(t) = \frac{e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}}}{\sqrt{\pi}} \cdot \frac{1}{2} = \frac{e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}}}{\sqrt{\pi}} \cdot \sqrt{\pi} = e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}}$$

Now,

$$\begin{aligned}\text{Mean} &= E(X) = \left[\frac{d}{dt} M_X(t) \right]_{t=0} \\ &= \left[\frac{d}{dt} e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}} \right]_{t=0} = \left[e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}} \cdot (\mu + t\sigma)^2 \right]_{t=0} = \mu\end{aligned}$$

Also,

$$\begin{aligned}E(X^2) &= \left\{ \frac{d^2}{dt^2} M_X(t) \right\}_{t=0} \\ &= \left[\frac{d^2}{dt^2} e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}} \right]_{t=0} = \left[\frac{d}{dt} e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}} \cdot (\mu + t\sigma)^2 \right]_{t=0} \\ &= \left[e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}} \cdot \sigma^2 + e^{\frac{t\mu + \frac{1}{2}t^2\sigma^2}{}} \cdot (\mu + t\sigma^2)^2 \right]_{t=0} \\ &= (\sigma^2 + \mu^2)\end{aligned}$$

$$\begin{aligned}\text{Therefore, } V(X) &= E(X^2) - (E(X))^2 \\ &= (\sigma^2 + \mu^2) - \mu^2 = \sigma^2\end{aligned}$$

13.2 Standard normal variable and standard normal probability distribution

Standard normal variable (definition)

If, $X \sim N(\mu, \sigma^2)$ then the variable

$$Z = \left(\frac{X - \mu}{\sigma} \right)$$

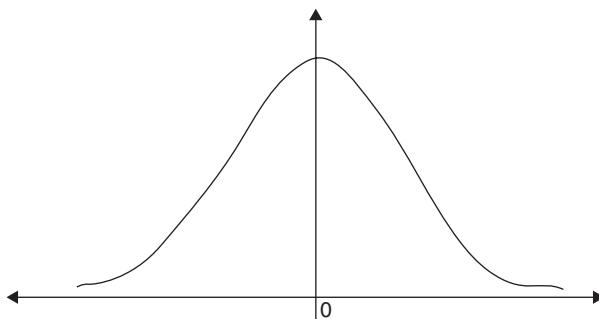
is called the standard normal variable (SNV)

Standard normal probability distribution

The probability distribution followed by the random variable Z is also a normal distribution and is called the standard normal distribution.

Properties of standard normal distribution

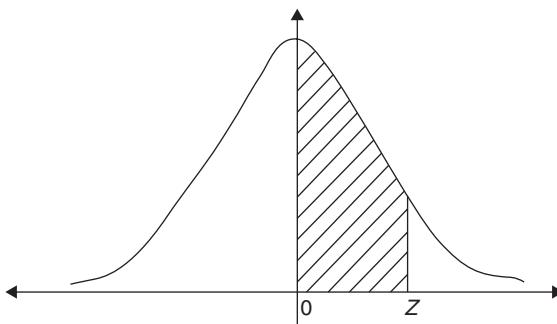
1. The mean of the standard normal distribution is always zero. That is $E(Z) = 0$.
2. The variance of standard normal distribution is always one. That is $V(Z) = 1$.
3. Graph of the standard normal distribution is as follows & shown in Fig. 13.2.



13.2

Note that,

1. From the graph it is clear that
 - (a) Total area under standard normal distribution curve
$$= P(-\infty < Z < +\infty) = \int_{-\infty}^{\infty} f(z) dz = 1$$
 - (b) $P(-\infty < Z < 0) = P(0 < Z < +\infty) = 0.5$
 - (c) $P(-t < Z < 0) = P(0 < Z < t)$
 - (d) $P(-\infty < Z < -t) = P(t < Z < +\infty)$
 2. The statistical table for the standard normal distribution provides the probabilities of the type $P(0 < Z < z)$ where z is any positive number
- That is the statistical table for the standard normal distribution provides the area of the shaded region as shown in the following Fig. 13.3.



13.3

Determination of probabilities in normal distribution

In normal probability distribution, generally interest is to find probabilities means to find area under the standard normal curve that is to solve the integrals of following type.

$$P(X < a) = \int_{-\infty}^a f(x) dx$$

$$P(X > a) = \int_a^{\infty} f(x) dx$$

$$P(a < X < b) = \int_a^b f(x) dx$$

All the above integrals are incomplete Gamma integrals and are not solvable directly. Hence, the above probabilities are obtained by using alternative procedure that is given below.

Procedure of finding probabilities in normal probability distribution

1. Convert the normal random variable (X) into the standard normal random variable (Z) using the definition

$$Z = \left(\frac{X - \mu}{\sigma} \right)$$

2. Use the statistical table for finding area under standard normal distribution curve available in the statistical tables book and the

properties of the standard normal distribution to determine the required probability.

For example,

Let $X \sim N(\mu, \sigma^2)$

Thus,

$$P(X > a) = P\left(\frac{X - \mu}{\sigma} > \frac{a - \mu}{\sigma}\right) = P(Z > z) \text{ assuming } \frac{a - \mu}{\sigma} = z$$

Further,

as the area under standard normal distribution curve from $(0, \infty)$ is 0.5

$$\text{Therefore, } P(X > a) = P\left(\frac{X - \mu}{\sigma} > \frac{a - \mu}{\sigma}\right) = P(Z > z) = 0.5 - P(0 < Z < z)$$

Now the required probability can be calculated by subtracting $P(0 < Z < z)$, obtained from the statistical table according to any real value of "z."

Example 13.1

A ring-frame is expected to spin the yarn of average count 40^s and CV% 1.2. If a ring bobbin is selected and tested for the count, what is the probability that

- (i) the ring bobbin will show count more than 40.5?
- (ii) the ring bobbin will show count in between 39 to 39.5?

Solution

Here, X represents count of the yarn.

Let, $X \sim N(\mu, \sigma^2)$

Given that, mean = $\mu = 40$ and $\text{CV\%} = \frac{\sigma}{\mu} \times 100 = 1.2 \Rightarrow \sigma = 0.48$

Now,

$$\begin{aligned} P(X > 40.5) &= P\left(\frac{X - 40}{0.48} > \frac{40.5 - 40}{0.48}\right) = P(Z > 1.04) \\ &= 0.5 - 0.3508 \\ &= 0.1492 \end{aligned}$$

Thus, there is approximately 14.92% chance that the count of the yarn will be greater than 40.5

Also,

$$P[\text{count is between 39 to 39.5}] = P[39 < X < 39.5]$$

$$\begin{aligned}
 &= P\left(\frac{39-40}{0.48} < \frac{X-40}{0.48} < \frac{39.5-40}{0.48}\right) = P(-2.08 < Z < -1.04) \\
 &= P(1.04 < Z < 2.08) \\
 &= P(0 < Z < 2.08) - P(0 < Z < 1.04) \\
 &= 0.4812 - 0.3508 = 0.1304
 \end{aligned}$$

That is, there is ~13% chance that the count is in between 39 and 39.5.

Example 13.2

Out of hundred pieces of fabric tested for the strength, 15 pieces have shown strength >75 units and 20 pieces have shown strength <60 units. Assuming normal distribution for the strength of the fabric, find the average and the standard deviation of the strength of the fabric. Also find the probability that the strength of such fabric is >70 units.

Solution

Here, X – strength of the fabric.

Let $X \sim N(\mu, \sigma^2)$

Given that,

$$P(X > 75) = 0.15 \text{ and } P(X < 60) = 0.2$$

From first information,

$$P(X > 75) = P\left(\frac{X-\mu}{\sigma} > \frac{75-\mu}{\sigma}\right) = P(Z > z) \quad \text{where } z = \frac{75-\mu}{\sigma}$$

Now from the statistical table for the standard normal distribution, it is clear that $z = 1.04$

$$\Rightarrow 1.04 = \frac{75-\mu}{\sigma} \quad \text{Therefore, } \mu + 1.04\sigma = 75 \rightarrow (1)$$

Similarly from second information,

$$P(X < 60) = P\left(\frac{X-\mu}{\sigma} < \frac{60-\mu}{\sigma}\right) = P(Z < -z) \quad \text{where } -z = \frac{60-\mu}{\sigma}$$

Now from the statistical table for the standard normal distribution, it is clear that $z = 1.04$ that is $-z = -1.04$

$$\Rightarrow -0.84 = \frac{60-\mu}{\sigma} \quad \text{Therefore, } \mu - 0.84\sigma = 60 \rightarrow (2)$$

Thus, solving equations (1) and (2) simultaneously for μ and σ , we get:
 $\sigma = 7.9787$ and $\mu = 66.7$.

Hence, average strength of the fabric is 66.7 units and the standard deviation of the strength of the fabric is 7.9787 units.

Further,

$$\begin{aligned} P(X > 70) &= P\left(\frac{X - 66.7}{7.9787} > \frac{70 - 66.7}{7.9787}\right) = P(Z > 0.41) = 0.5 - P(0 < Z < 0.41) \\ &= 0.5 - 0.1591 \\ &= 0.3409 \end{aligned}$$

That is approximately 34% of the times strength of the fabric pieces will be greater than 70 units.

13.3 Chi-square probability distribution (χ^2 distribution)

This is also an important and most popularly used continuous probability distribution which has very large number of applications in real life. Sometimes, this is also known as small sample or small sampling distribution. The Chi-square probability distribution is defined as follows:

Definition

A continuous random variable “ X ” is said to follow Chi-square probability distribution with “ n ” degrees of freedom if its pdf is as follows:

$$f(x) = \frac{1}{2^{\frac{n}{2}} \cdot \gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} \quad 0 < x < \infty$$

Properties of the Chi-square probability distribution

1. In case of Chi-square distribution the degrees of freedom “ n ” is known as the parameter of the Chi-square distribution.

Hence the Chi-square distribution can be represented by the notation

$$X \sim \chi^2_n$$

2. The pdf of Chi-square distribution satisfies the condition $\int f(x)dx = 1$
3. If $X \sim N(0,1)$ that is, X is a SNV then, X^2 will follow χ^2 distribution with 1 degree of freedom. That is, $X \sim \chi^2_1$

Further, if X_1, X_2, \dots, X_n are “ n ” independent SNVs then

$T = X_1^2 + X_2^2 + \dots + X_n^2 = \sum X_i^2$ will follow χ^2 -distribution with n degrees of freedom.

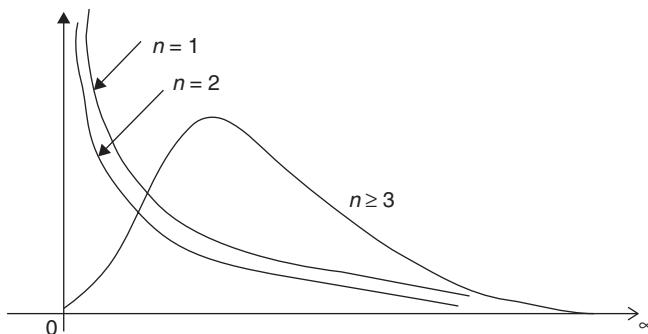
From the above property, note that degrees of freedom can be defined as the number of independent random variables used for defining the new χ^2 random variable.

Further, suppose that the above variable T is defined under the condition $\sum X_i = 0$ then $T \sim \chi^2_{n-1}$

4. If, $X \sim \chi^{2n}$, $Y \sim \chi^{2m}$ and if X and Y are independent then, $X + Y \sim \chi^2_{m+n}$

The mean of the Chi-square distribution is “ n ” and the variance is “ $2n$ ”.

The graph of the χ^2 -distribution is positively skewed, and can be shown as follows (see Fig. 13.4)



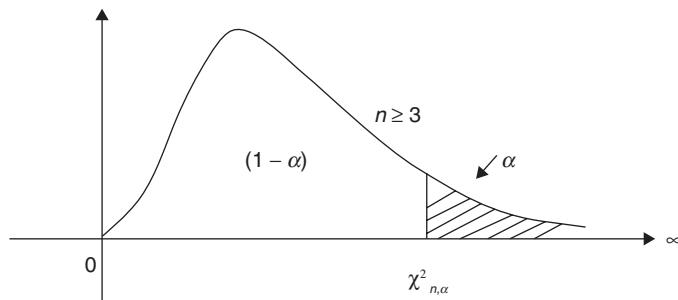
13.4

5. The statistical table for the χ^2 -distribution provides the points $\chi^2_{n,\alpha}$ for different values of n and α . Where the point is $\chi^2_{n,\alpha}$ is such that the area (as shown in the following figure), under the χ^2 -distribution curve above this point is α .

That is,

$$P\{\chi^2_n > \chi^2_{n,\alpha}\} = \alpha$$

Also graphically can be shown in Fig. 13.5.



13.5

For example, $\chi^2_{5,0.05} = 11.07 \Rightarrow$ Only in 5% cases value of the variable χ^2_5 will be > 11.07 .

13.4 Student's *t*-probability distribution

This is also an important and most popularly used continuous probability distribution which has very large number of applications in real life. Sometimes this is also known as small sample or small sampling distribution. The *t*-probability distribution is defined as follows:

Definition

A continuous random variable “ X ” is said to follow *t*-probability distribution with “ n ” degrees of freedom if its pdf is as follows:

$$f(x) = \frac{1}{\sqrt{n} \cdot B\left(\frac{n}{2}, \frac{1}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{\frac{(n+1)}{2}} \quad -\infty < x < \infty$$

Properties of the t-probability distribution

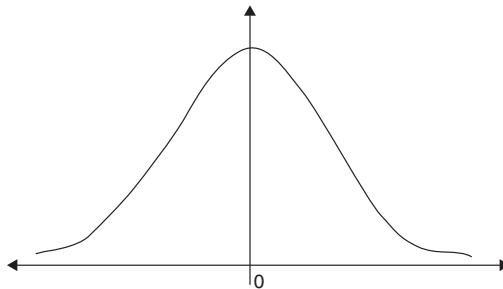
1. In case of *t*-probability distribution also the degrees of freedom “ n ” is known as the parameter of the *t*-distribution. Hence, the *t*-distribution can be represented by the notation

$$X \sim t_n$$

2. The pdf of *t*-distribution satisfies the condition $\int f(x)dx = 1$
3. If $X \sim N(0,1)$ that is, X is a SNV; Y follows χ^2 -distribution with n degrees of freedom that is $Y \sim \chi^2_n$ and if both X and Y are independent, then

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$$

4. The graph of the t -distribution is symmetric and can be shown as follow in Fig. 13.6.

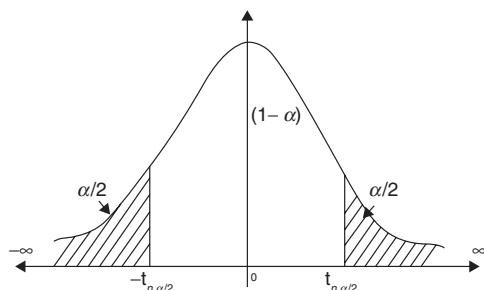


13.6

5. The statistical table for the t -distribution provides the points $t_{n,\alpha/2}$ for different values of n and α . Where the point $t_{n,\alpha/2}$ is such that, the area (as shown in Fig. 13.7) under the t -distribution curve above this point is $\alpha/2$ and the area below the point $-t_{n,\alpha/2}$ is also $\alpha/2$. Thus the total area below and above is α . That is,

$$P(|t_n| > t_{n,\alpha/2}) = \alpha$$

Graphically can be shown in Fig. 13.7,



13.7

For example, if $n = 5$ and $\alpha = 0.05$ then

$$t_{n,\alpha/2} = t_{5,0.025} = 2.571$$

$$\Rightarrow p(|t_5| > t_{5,0.025}) = 0.05$$

Thus, there is 2.5% chance that the variable t_5 has value more than 2.571 and the chance 2.5% that it is less than -2.571.

13.5 F-probability distribution

This is also an important and most popularly used continuous probability distribution which has very large number of applications in real life. Sometimes, this is also known as small sample or small sampling distribution. The F-probability distribution is defined as follows:

Definition

A continuous random variable “X” is said to follow F-probability distribution with “m” and “n” degrees of freedom if its pdf is as follows:

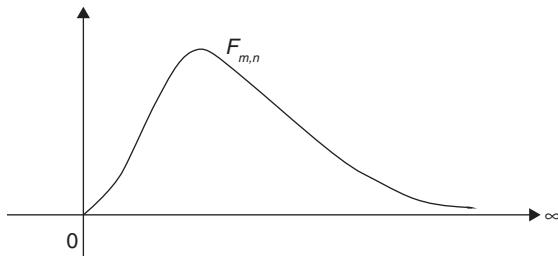
$$f(x) = \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \times \frac{x^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n}x\right)^{\frac{(m+n)}{2}}} \quad 0 < x < \infty$$

Properties of the F-probability distribution

1. In case of F-probability distribution also the degrees of freedom “m” and “n” are known as the parameters of the F-distribution. Hence the F-distribution can be represented by $X \sim F_{m,n}$
2. The pdf of F-distribution satisfies the condition $\int f(x)dx = 1$
3. If X follows χ^2 -distribution with n degrees of freedom that is $X \sim \chi^{2n}$; Y follows χ^2 distribution with n degrees of freedom that is $Y \sim \chi^2_m$ and if both X and Y are independent then

$$F = \frac{X/m}{Y/n} \sim F_{m,n}$$

4. The graph of the F-probability distribution is also positively skewed and can be shown as follow in Fig. 13.8.

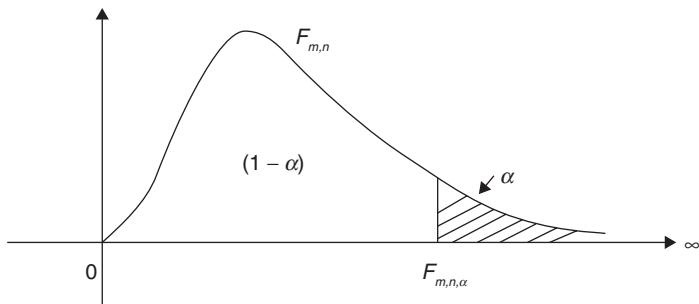


13.8

5. The statistical table for the F -probability distribution provides the points $F_{m,n,\infty}$ for different values of m and n but only for two values of α ($\alpha = 0.05$ that is 5% and $\alpha = 0.01$ that is 1%), where the point $F_{m,n,\infty}$ is such that the area under the probability distribution curve of F -probability distribution, above this point is α . That is,

$$p(F_{m,n} > F_{m,n,\infty}) = \alpha$$

Graphically can be shown in Fig. 13.9,



13.9

For example,

If $m = 5$ – numerator degrees of freedom, $n = 8$ – denominator degrees of freedom and $\alpha = 0.05$, then $F_{m,n,\infty} = F_{5,8,0.05}$

Thus from 5% table, $F_{m,n,\infty} = F_{5,8,0.05} = 3.69$

That is in 5% cases value of the variable $F_{5,8}$ can be greater than 3.69.

Also,

If $m = 5$ – numerator degrees of freedom, $n = 8$ – denominator degrees of freedom and $\alpha = 0.01$ then

$$F_{m,n,\infty} = F_{5,8,0.01}$$

Thus from 1% table, $F_{m,n,\infty} = F_{5,8,0.01} = 6.63$

That is in 1% cases value of the variable $F_{5,8}$ can be greater than 6.63.

6. F -probability distribution satisfies following relationship.

$$F_{m,n,(1-\alpha)} = \frac{1}{F_{n,m,\infty}} \Rightarrow F_{m,n,(1-\alpha)} \cdot F_{n,m,\infty} = 1$$

For example,

If we are interested in finding the value $F_{5,8,0.95}$, then for finding this value we need 95% table and 95% table for F -distribution is not available. In such cases, the above property can be used and required value can be obtained as follows:

$$F_{5,8,0.95} = \frac{1}{F_{8,5,0.05}} = \frac{1}{4.82} = 0.2075$$

13.6 Exercise

1. Define normal probability distribution. State its properties. Give two examples of random variables which follow normal distribution.
2. Derive expression of MGF for the normal probability distribution; hence find its mean and the variance.
3. Define student's "t" distribution. State its properties
4. Define F -probability distribution. State its properties.
5. Define Chi-square probability distribution. State its properties.
6. The single thread strength of a yarn is expected to follow normal probability distribution with mean 50 gms. and std. dev. 5 gm. If 100 strength tests are made on this yarn, how many tests will show strength in between 35 gm to 45 gms?
7. The sliver of draw frame has nominal hank 0.12 and std. dev. 0.0025. If 50 hank tests are made on the production of this draw frame, using normal probability distribution, find how many tests will show hank in between 0.1150 to 0.1250.
8. The linear density of the yarn is normally distributed with mean = 14 units and $CV\% = 1.5$. If a linear density test is made on this yarn,

what is the probability that the test will show linear density more than 13.0 units?

9. The strength of the yarn is expected to follow normal probability distribution with mean 65 gms. and standard deviation 5 gm. If 100 strength tests are made on this yarn, how many tests will show strength in between 50 gm to 60 gm?
10. The weekly production of a mill is normally distributed with mean 125 tons and standard deviation 7.5 tons. What is the probability that in the coming week
 - (i) the production will be >135 tons?
 - (ii) the production will be >120 tons?
11. The weight of garment is normally distributed with mean 250 gms and standard deviation 5 gms. If 100 garments are measured, how many garments will show
 - (i) weight in between 262 and 270 gm
 - (ii) more than 270 gm?
12. A soft drink machine is designed to regulate the average discharge of 7 ounce per cup. If the amount of drink is normally distributed with SD of 5 ounce, find:
 - (i) the percentage of cups containing more than 7.8 ounces,
 - (ii) how many cups out of 1000 are expected to contain 7 to 7.5 ounces?
13. The life times of a certain battery have an average of 300 h with S.D. of 35 h. Assuming that the distribution of lifetime is normal, how many batteries, from a lot of 1000, are expected to have lifetime more than 275 h?

Testing of hypothesis

14.1 Introduction

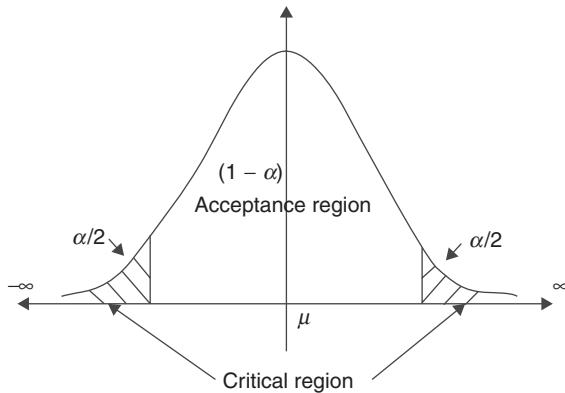
Testing of hypothesis and estimation are the two important parts of statistical inference. Engineers in the industry, scientist of the research institutes, and statisticians are generally interested in studying the unknown parameters of certain population under study. This study of unknown population parameters is carried out on the basis of the sample selected from that population. Testing of hypothesis and estimation are the two different types of such study.

Testing of hypothesis

Testing of hypothesis is one of the two important ways of studying unknown population parameters. In testing of hypothesis, generally some assumption is made regarding unknown population parameter; this assumption is called the hypothesis. Null hypothesis and alternate hypothesis are the two different types of hypothesis. The hypothesis of no difference is always called the null hypothesis and any alternate to the null hypothesis is called the alternate hypothesis. The notation " H_0 " denotes the null hypothesis, and the notation " H_1 " denotes the alternate hypothesis. Thus, in testing of hypothesis the decision regarding acceptance (true) or rejection (false) of null hypothesis is made on the basis of some function or the value, calculated from the sample selected from the population for the purpose of the study, this function of the sample observable variables is called the statistic. The probability distribution followed by the statistic is called the sampling distribution of the statistic and the standard deviation of the sampling distribution of this statistic is called the standard error. The standard error is treated as the measure of accuracy of the statistic; the smaller the value of the standard error, the larger the accuracy of the statistic, and vice versa. The standard error depends on the size of the sample selected for the study; the larger the size of the sample, the smaller the standard error and the better the accuracy of the statistic, and vice versa. The decision regarding the acceptance or

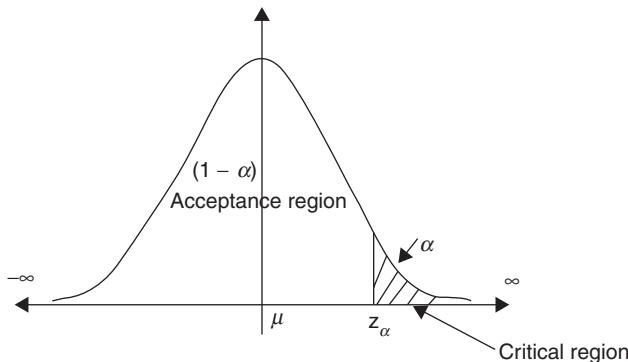
rejection of null hypothesis is made with the help of certain region under the probability density curve of the sampling distribution of the statistic. This region is called the critical region (CR). If the value of statistic, calculated from the sample, falls in this region, then the null hypothesis H_0 is rejected. Two-tailed critical region and one-tailed critical region (left tailed or right tailed) are two different types of critical regions. As an illustration, different types of critical region under the standard normal distribution curve can be shown in Figs. 14.1, 14.2, and 14.3.

1. Two-tailed critical region



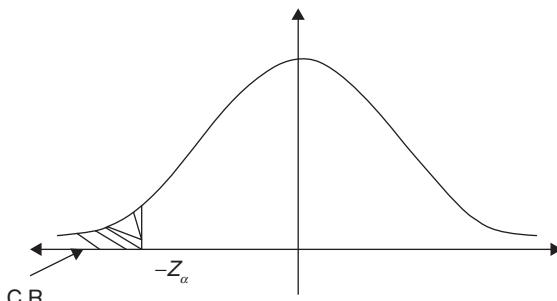
14.1

2. Right-tailed critical region



14.2

3. Left-tailed critical region



14.3

The size of critical region is always measured in terms of probability and its value is always α . Thus the size of acceptance region (AR) is $1 - \alpha$. Here, α is related to level of significance (*los*), which represents maximum probability of making type-one error and type-one error represents the probability of rejecting null hypothesis H_0 when it is true. The level of significance is always represented in percentage and decided in advance before testing of hypothesis.

Depending upon the parameter to be studied and the size of the sample selected for study, there are number of different test procedures, which are classified as the large sample tests (sample size > 30) and small sample tests (sample size ≤ 30).

14.2 Large sample tests (Z-tests)

Large sample test for the population mean (Z-test)

This test is used, if the size of the sample is large and interest is to test the hypothesis of the following type, related to mean of the only one population under study.

Suppose,

X is the variable of the population under study

μ is the mean and σ is standard deviation of the variable X of the population under study.

μ_0 is the expected or the nominal value of the population mean μ .

Thus, in this test, it is interesting to test the hypothesis:

$$\begin{aligned} H_0: \mu = \mu_0 & \quad \text{Vs} \quad H_1: \mu \neq \mu_0 \\ & \quad \text{or} \\ & \quad H_1: \mu < \mu_0 \\ & \quad \text{or} \\ & \quad H_1: \mu > \mu_0 \end{aligned}$$

For testing the above hypothesis H_0 , a large sample of size “ n ” is selected and using the law of large numbers the statistic Z is defined as follows:

$$\begin{aligned} Z &= \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad \text{if } \sigma \text{ is known} \\ &= \frac{\bar{x} - \mu_0}{S / \sqrt{n}} \quad \text{if } \sigma \text{ is unknown} \end{aligned}$$

where,

$$\begin{aligned} \bar{x} &= \text{Sample mean} = \frac{\sum x_i}{n} \\ S^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{1}{n} \left(\sum x_i^2 - n\bar{x}^2 \right) = \text{Sample Variance} \\ S &= \sqrt{S^2} = \text{Sample standard deviation} \end{aligned}$$

Here, the statistic Z is expected to follow the standard normal distribution that is $Z \sim N(0,1)$.

Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

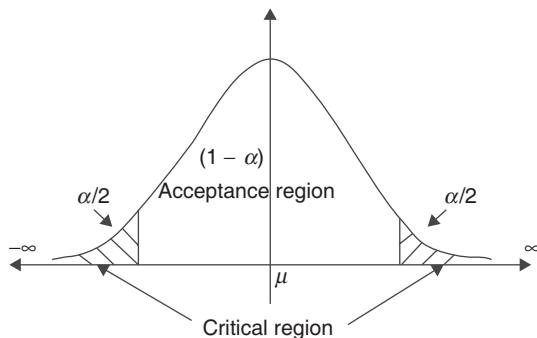
Case I Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu \neq \mu_0$

if, $|Z_{cal}| > Z_{\alpha/2}$

Where, the point $Z_{\alpha/2}$ is such that

$$P(|Z| > Z_{\alpha/2}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.4

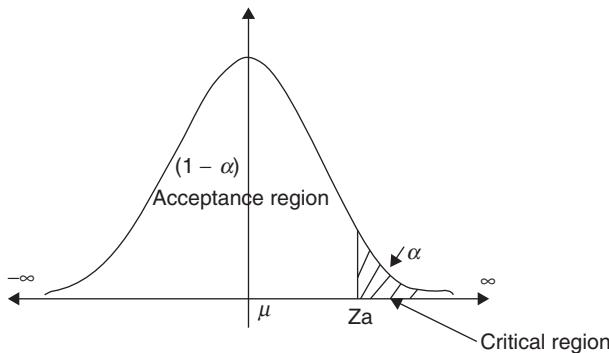


14.4

Case II Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu > \mu_0$ if, $Z_{cal} > Z_\alpha$
Where, the point Z_α is such that,

$$P(Z > Z_\alpha) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.5



14.5

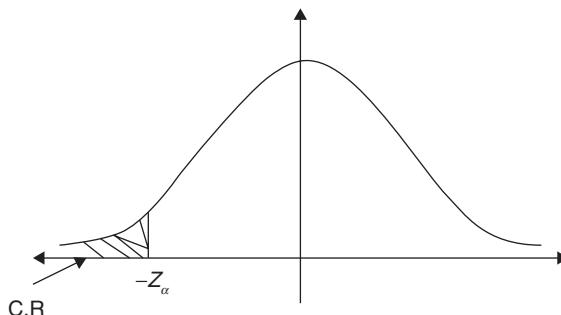
Case III Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu < \mu_0$

if, $Z_{cal} < -Z_\alpha$

Where, the point $-Z_\alpha$ is such that,

$$P(Z < -Z_\alpha) = \alpha \text{ i.e. } P(Z > Z_\alpha) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.6



14.6

Example 14.1

The nominal linear density of the yarn spun during a shift is 14 tex. But the sample of 45 leas tested has shown average linear density 14.8 tex and the CV% 2.5 tex. From the sample results, can we say that the production of the shift is of the required linear density?

Solution

Here,

$\text{Population} \Rightarrow \text{Production of the yarn during the shift}$
 $X \Rightarrow \text{Linear density of the yarn.}$

Suppose,

μ is the population mean of the variable X and σ is the population standard deviation of variable X

Thus, interest is to test the hypothesis,

$$H_0: \mu = 14 \quad vs \quad H_1: \mu \neq 14$$

For testing the above hypothesis, the large sample of size $n = 45$ is selected
Hence, the statistic Z is calculated as follows:

$$Z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}} \quad \text{as } \sigma \text{ is unknown}$$

Given that,

Sample mean $= \bar{x} = 14.8$

$$\text{Coefficient of Variation} = \text{CV\%} = \frac{s}{\bar{x}} \times 100 = 2.5$$

$$\text{Therefore, } S = \frac{2.5 \times 14.8}{100} = 0.37$$

$$\text{Therefore, } Z = \frac{14.8 - 14}{\frac{0.37}{\sqrt{45}}} = 14.51$$

Now, at 5% *los*, that is for $\alpha = 0.05$

$$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$$

Here,

$$|Z_{\text{cal}}| = 14.51 > Z_{\frac{\alpha}{2}} = 1.96$$

\Rightarrow Reject $H_0 \Rightarrow$ accept $H_1 \Rightarrow \mu \neq 14$

\Rightarrow Average linear density of the yarn produced during the shift is not 14, which is not as per requirement.

Example 14.2

A sample of 35 leas has shown average lea weight 14.5 units. Can we say that this sample is selected from the population having mean lea weight 15 units and the standard deviation of lea weight as 1.00?

Solution

Here,

Population \Rightarrow Collection of leas under study

$X \Rightarrow$ Lea weight.

Suppose,

μ is the population mean of the variable X and σ is the population standard deviation of variable X , respectively

Thus, interest is to test the hypothesis,

$$H_0: \mu = 15 \quad \text{Vs} \quad H_1: \mu \neq 15$$

For testing the above hypothesis, the large sample of size $n = 35$ is selected

Hence, the statistic Z is calculated as follows:

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad \sigma \text{ is known}$$

Given that,

Sample mean $= \bar{x} = 14.5$

Population standard deviation $= \sigma = 1.00$

$$\text{Therefore, } Z = \frac{14.5 - 15}{\frac{1.00}{\sqrt{35}}} = -2.95$$

Now, at 5% los, that is for $\alpha = 0.05$

$$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$$

Here,

$$|Z_{\text{cal}}| = 2.95 > Z_{\frac{\alpha}{2}} = 1.96$$

\Rightarrow Reject $H_0 \Rightarrow$ accept $H_1 \Rightarrow \mu \neq 15$

\Rightarrow Average linear density of the yarn produced during the shift is not 15, which is not as per requirement.

Large sample test for equality of population means (Z-test)

This test is used, if the sizes of the samples are large and interest is to test the hypothesis of the following type, related to the means of two different populations under study.

$$H_0: \mu_1 = \mu_2 \quad \text{Vs} \quad H_1: \mu_1 \neq \mu_2$$

or

$$H_1: \mu_1 > \mu_2$$

or

$$H_1: \mu_1 < \mu_2$$

where,

μ_1 & μ_2 are the population means of variables X_1 & X_2 of two different populations under study.

For testing the hypothesis H_0 , two large sample of sizes n_1 & n_2 are selected and using the law of large numbers, the statistic Z is defined as follows

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)}} \quad \text{if } \sigma_1 \text{ & } \sigma_2 \text{ are known and } \sigma_1 \neq \sigma_2$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}} \quad \text{if } \sigma_1 \text{ & } \sigma_2 \text{ are known and } \sigma_1 \neq \sigma_2$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{S \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{if } \sigma_1 \text{ & } \sigma_2 \text{ are known and } \sigma_1 = \sigma_2$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{if } \sigma_1 \text{ & } \sigma_2 \text{ are known and } \sigma_1 = \sigma_2$$

where,

$$\bar{x}_1 = \text{Sample mean for first sample} = \frac{\sum x_{1i}}{n_1} \text{ &}$$

$$\bar{x}_2 = \text{Sample mean for second sample} = \frac{\sum x_{2i}}{n_2}$$

$$S_1^2 = \frac{\sum x_{1i}^2}{n_1} - \bar{x}_1^2 = \frac{1}{n_1} \left(\sum x_{1i}^2 - n_1 \bar{x}_1^2 \right) = \text{Sample variance for first sample}$$

$$S_2^2 = \frac{\sum x_{2i}^2}{n_2} - \bar{x}_2^2 = \frac{1}{n_2} \left(\sum x_{2i}^2 - n_2 \bar{x}_2^2 \right) = \text{Sample variance for second sample}$$

$$S_p^2 = \text{Pooled sample variance for both samples together} = \frac{n_1 \times S_1^2 + n_2 \times S_2^2}{n_1 + n_2}$$

$$S_p = \sqrt{S_p^2} = \text{Pooled sample standard deviation}$$

σ_1 & σ_2 are the population standard deviations of variables X_1 & X_2 of two different populations

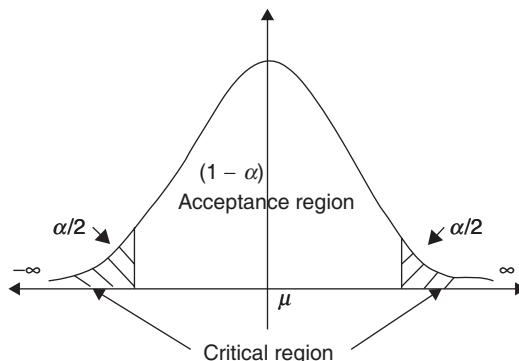
$$\sigma_1 = \sigma_2 = \sigma = \text{population standard deviation}$$

Here, the statistic Z is expected to follow the standard normal distribution that is $Z \sim N(0,1)$.

Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

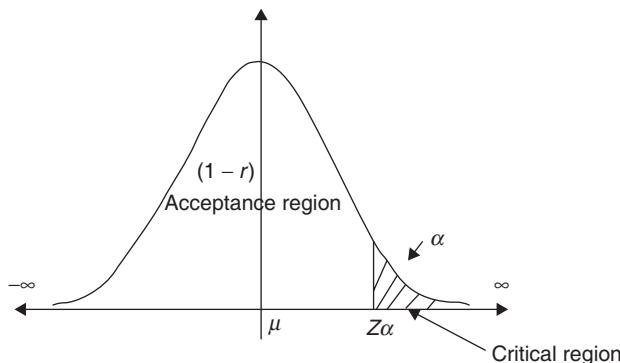
Case I Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu_1 \neq \mu_2$ if,
 $|Z_{cal}| > Z_{\alpha/2}$

Where, the point $Z_{\alpha/2}$ is similar as discussed in first case of previous test and the graph of the critical region can be shown in Fig. 14.7



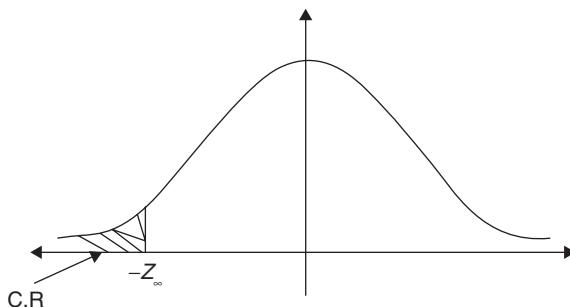
14.7

Case II Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu_1 > \mu_2$ if, $Z_{cal} > Z_\alpha$
 Where, the point Z_α is similar as discussed in second case of previous test and the graph of the critical region can be shown a in Fig. 14.8



14.8

Case III Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu_1 < \mu_2$ if, $Z_{cal} < -Z_\alpha$
 Where, the point $-Z_\alpha$ is similar as discussed in third case of previous test and the graph of the critical region can be shown in Fig. 14.9



14.9

Example 14.3

Two ring frames are expected to spin the yarn of the same strength. The samples of 35 and 40 ring bobbins selected from these ring frames have shown following results.

	R/F-1	R/F-2
Sample size	35	40
Mean strength	60 units	56 units
Std. deviation of strength	1.25 units	1.50 units

From the sample results, is there any evidence that the yarn of first ring frame is having strength more than the yarn of ring frame 2? Use 1% *los*.

Solution

Here,

Population 1 \Rightarrow Yarn spun by R/F-1 and Population 2 \Rightarrow Yarn spun by R/F-2

X_1 \Rightarrow Strength of yarn spun by R/F-1 and X_2 \Rightarrow Strength of yarn spun by R/F-2

Suppose, μ_1 and σ_1 are mean and standard deviation of variable X_1 and μ_2 and σ_2 are mean and standard deviation of variable X_2

Thus, interest is to test the hypothesis,

$$H_0: \mu_1 = \mu_2 \quad \text{Vs} \quad H_1: \mu_1 > \mu_2$$

For testing the hypothesis, large samples of sizes, $n_1 = 35$ and $n_2 = 40$ are selected from the two populations under study. Also σ_1 & σ_2 are unknown and $\sigma_1 \neq \sigma_2$ hence, the statistic Z is calculated as follows:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{60 - 56}{\sqrt{\left(\frac{1.25^2}{35} + \frac{1.5^2}{40}\right)}} = 12.6$$

Now, at 1% los, that is for $\alpha = 0.01$, $Z_\alpha = Z_{0.01} = 2.33$

Here, $Z_{\text{cal}} = 12.6 > Z_\alpha = 2.33$

Hence, Reject $H_0 \Rightarrow$ accepts $H_1 \Rightarrow \mu_1 > \mu_2 \Rightarrow$ average strength of the R/F-1 yarn is more than that of the yarn of R/F-2.

Large sample test for the population proportion (Z-test)

This test is used, if the size of the sample is large and interest is to test the hypothesis of the following type, related to proportion of some specific type of articles in the population under study with the assumption that, the only one population under study is made up of two types of articles which are generally regarded as the defective and non-defectives.

$$\begin{array}{ll} H_0: P = P_0 & \text{Vs} \\ & H_1: P \neq P_0 \\ & \quad \text{or} \\ & H_1: P < P_0 \\ & \quad \text{or} \\ & H_1: P > P_0 \end{array}$$

where,

P is the proportion of specific type of articles and $Q = 1 - P$ is the proportion of other type of articles in the population under study.

P_0 is the expected or the nominal value of the population proportion P .

For testing the hypothesis H_0 , a large sample of size “ n ” is selected and using the law of large numbers, the statistic Z is defined as follows:

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}$$

where,

p is sample proportion of corresponding articles

$$p = \frac{\text{No.of articles of corresponding type}}{\text{sample size}} = \frac{d}{n}$$

Here, the statistic Z is expected to follow the standard normal distribution that is $Z \sim N(0,1)$.

Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

Case I Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: P = P_0$

if, $|Z_{\text{cal}}| > Z_{\alpha/2}$

Where, the point Z_α
and the graph of the critical region are similar as discussed in first case of previous tests.

Case II Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: P > P_0$
if, $Z_{cal} > Z_\alpha$ Where, the point Z_α
and the graph of the critical region are similar as discussed in second case of previous tests.

Case III Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: P < P_0$
if, $Z_{cal} < -Z_\alpha$ Where, the point $-Z_\alpha$
and the graph of the critical region are similar as discussed in third case of previous tests.

Example 14.4

A batch of fibers is expected to contain the mix of cotton and viscose in the ratio 1:3. A textile engineer has selected the sample of 300 fibers from this batch and found 110 cotton fibers. From the sample result, what conclusion should the textile engineer make out?

Solution

Here, Population is the batch of fibers under study.

Let P is the proportion of cotton fibers in the batch under study.

Expected proportion of cotton fibers as per the ratio 1:3 $\Rightarrow P_0 = 1/3$

Thus the interest is to test,

$$H_0: P = \frac{1}{3} \quad Vs \quad H_1: P \neq \frac{1}{3}$$

For testing the hypothesis H_0 , a large sample of size $n = 300$ is selected and using the law of large numbers the statistic Z is calculated follows:

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = \frac{\frac{110}{300} - \frac{1}{3}}{\sqrt{\frac{\frac{1}{3} \times \frac{2}{3}}{300}}} = 1.22497$$

Here, the statistic Z is expected to follow the standard normal distribution.

Now, at $100\alpha\%$ los, that is for $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$

Here, $|Z_{cal}| = 1.2247 < Z_{\alpha/2} = 1.96$

Thus, accept $H_0 \Rightarrow P = 1/3 \Rightarrow$ the batch contains the mix of cotton and viscose in the ratio 1:3

Large sample test for the equality of the proportions of two different populations (Z-test)

This test is used, if the sizes of the samples are large and interest is to test the hypothesis of the following type, related to proportions of some specific type

of articles of two different populations under study with the assumption that, both the populations under study are made up of two types of articles which are generally regarded as the defective and non-defective.

$$\begin{aligned} H_0: P_1 &= P_2 && \text{Vs} && H_1: P_1 \neq P_2 \\ &&&&& \text{or} \\ &&&&& H_1: P_1 > P_2 \\ &&&&& \text{or} \\ &&&&& H_1: P_1 < P_2 \end{aligned}$$

where,

P_1 & P_2 are the population proportions of some specific type of articles of two different populations under study.

For testing the hypothesis H_0 , two different large samples of sizes n_1 & n_2 are selected and using the law of large numbers, the statistic Z is defined as follows:

$$Z = \frac{p_1 - p_2}{\sqrt{\bar{P}\bar{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where,

$$p_1 = \text{Proportion of above assumed articles in first sample} = \frac{d_1}{n_1}$$

$$p_2 = \text{Proportion of above assumed articles in second sample} = \frac{d_2}{n_2}$$

$$\bar{P} = \frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2} \text{ and } \bar{Q} = 1 - \bar{P}$$

Here, the statistic Z is expected to follow the standard normal distribution that is $Z \sim N(0,1)$.

Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

Case I Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: P_1 \neq P_2$

if, $|Z_{\text{cal}}| > Z_{\alpha/2}$ where, the point $Z_{\alpha/2}$

and the graph of the critical region are similar as discussed in first case of previous tests.

Case II Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: P_1 > P_2$

if, $Z_{\text{cal}} > Z_\alpha$ where, the point Z_α

and the graph of the critical region are similar as discussed in second case of previous tests.

Case III Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: P_1 < P_2$, if, $Z_{cal} < -Z_\alpha$ where, the point $-Z_\alpha$ and the graph of the critical region are similar as discussed in third case of previous tests.

Example 14.5

Two batches of fibers are expected to contain same mix of cotton and viscose fibers. To investigate the expectation two different samples of 200 and 250 fibers were selected from the two batches respectively and it was found that the samples contain 55 and 82 viscose fibers respectively. From the sample results is it reasonable to say that the expectation is true?

Solution

Here, Population 1 is the batch 1 and Population 2 is the batch 2

Let, P_1 and P_2 are the proportions of viscose fibers in batch 1 and batch 2, respectively. Thus, the interest is to test the hypothesis,

$$H_0: P_1 = P_2 \quad \text{Vs} \quad H_1: P_1 \neq P_2$$

For testing the above hypothesis, two large samples of sizes 200 and 250 are selected. Hence, the statistic Z is calculated as follows:

$$Z = \frac{P_1 - P_2}{\sqrt{\bar{P}\bar{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Given that

$$p_1 = \text{Proportion of viscose fibers in first sample} = \frac{d_1}{n_1} = \frac{55}{200}$$

$$p_2 = \text{Proportion of viscose fibers in second sample} = \frac{d_2}{n_2} = \frac{82}{250}$$

$$\bar{P} = \frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2} = \frac{200 \times \frac{55}{200} + 250 \times \frac{82}{250}}{200 + 250} = 0.304$$

$$\bar{Q} = 1 - \bar{P} = 1 - 0.304 = 0.696$$

$$Z = \frac{\frac{55}{200} - \frac{82}{250}}{\sqrt{0.304 \times 0.696 \times \left(\frac{1}{200} + \frac{1}{250}\right)}} = -1.2145$$

Here, the statistic Z is expected to follow the standard normal distribution.

Now, at $100\alpha\%$ los, that is for $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$

Here, $|Z_{\text{cal}}| = 1.2145 < Z_{\alpha/2} = 1.96$

Thus, accept $H_0 \geq P_1 = P_2 \geq$ the batches contain same mix of cotton and viscose fibers.

14.3 Small sample tests

Small sample test for the population mean (t-test)

This test is used, if the size of the sample is small and interest is to test the hypothesis of the following type, related to mean of the only one population under study.

Suppose,

X is the variable of the population under study

μ is the mean and σ is standard deviation of the variable X of the population under study.

μ_0 is the expected or the nominal value of the population mean μ .

Thus, in this test interest is to test the hypothesis,

$$\begin{aligned} H_0: \mu = \mu_0 \quad &\text{Vs} \quad H_1: \mu \neq \mu_0 \\ &\text{or} \\ &H_1: \mu < \mu_0 \\ &\text{or} \\ &H_1: \mu > \mu_0 \end{aligned}$$

For testing the hypothesis H_0 , a small sample of size “ n ” is selected.

As the size of the sample is small in this case a statistic “ t ” is defined as follows:

$$t = \frac{\bar{x} - \mu_0}{\hat{s} / \sqrt{n}}$$

where,

$$\bar{x} = \text{Sample mean} = \frac{\sum x_i}{n}$$

$$\hat{s}^2 = \frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2) = \text{Estimate of population variance } \sigma^2$$

$$\hat{s} = \sqrt{\hat{s}^2} = \text{Estimate of population standard deviation } \sigma$$

$$n \cdot S^2 = (n-1) \cdot \hat{s}^2$$

Note that the above definition of statistic “ t ” is valid under the assumptions:

1. The population under study is normal population. That is $X \sim N(\mu, \sigma^2)$
2. Population standard deviation σ is unknown.

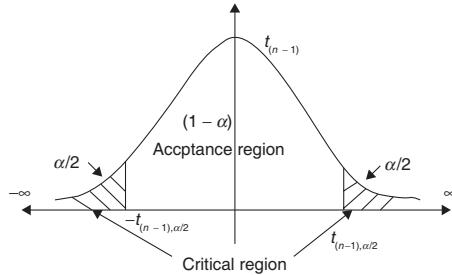
Here, the statistic “ t ” is expected to follow “ t ” probability distribution with $(n - 1)$ degrees of freedom that is $t \sim t_{n-1}$

Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

Case I Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu \neq \mu_0$ if, $|t_{cal}| > t_{n-1,\alpha/2}$
Where, the point $t_{n-1,\alpha/2}$ is such that,

$$P(|t_{n-1}| > t_{n-1,\alpha/2}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.10



14.10

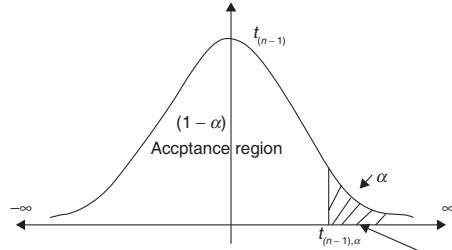
Case II Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu > \mu_0$.

if, $t_{cal} > t_{n-1,\alpha}$

where, the point $t_{n-1,\alpha}$ is such that,

$$P(|t_{n-1}| > t_{n-1,\alpha}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.11

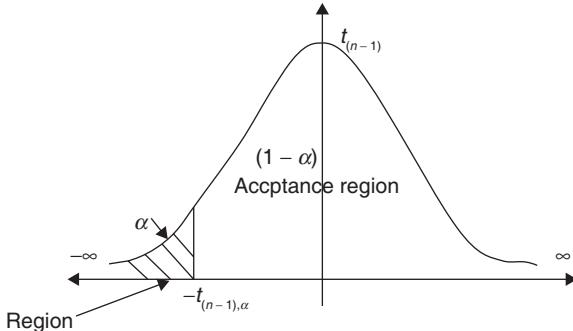


14.11

Case III: Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu < \mu_0$ if, $t_{cal} < -t_{n-1,\alpha}$
 Where, the point $-t_{n-1,\alpha}$ is such that,

$$P(t_{n-1} < -t_{n-1,\alpha}) = \alpha \quad \text{i.e. } P(t_{n-1} > -t_{n-1,\alpha}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.12



14.12

Example 14.6

The nominal linear density of the yarn spun during a shift is 14 tex. But the sample of 15 leas tested has shown average linear density 14.8 tex and the CV% 2.5 tex. From the sample results, can we say that the production of the shift is of the required linear density?

Solution

Here,

Population \Rightarrow Production of the yarn during the shift

$X \Rightarrow$ Linear density of the yarn

Suppose,

μ is the population mean of the variable X and σ is the population standard deviation of variable X

Thus, interest is to test the hypothesis,

$$H_0: \mu = 14 \quad \text{Vs} \quad H_1: \mu \neq 14$$

For testing the above hypothesis, the small sample of size $n = 15$ is selected

Hence, the statistic t is calculated as follows:

$$t = \frac{\bar{x} - \mu_0}{\hat{s} / \sqrt{n}}$$

Given that,

$$\text{Sample mean} = \bar{x} = 14.8$$

$$\text{Coefficient of Variation} = \text{CV\%} = \frac{s}{\bar{x}} \times 100 = 2.5$$

$$\begin{aligned}\text{Therefore, } S &= \frac{2.5 \times 14.8}{100} = 0.37 \Rightarrow \hat{s}^2 = \frac{n}{n-1} S^2 = \frac{15}{14} \times 0.37^2 \\ &= 0.1467 \Rightarrow \hat{s} = \sqrt{0.1467} \\ &= 0.383\end{aligned}$$

$$\text{Therefore, } t = \frac{14.8 - 14}{0.383 / \sqrt{15}} = 8.0895$$

Now, at 5% los, that is for $\alpha = 0.05$

$$t_{n-1, \alpha/2} = t_{14, 0.025} = 2.145$$

Here,

$$|t_{\text{cal}}| = 8.0895 > t_{n-1, \alpha/2} = 2.145$$

$$\Rightarrow \text{Reject } H_0 \Rightarrow \text{accept } H_1 \Rightarrow \mu \neq 14$$

\Rightarrow Average linear density of the yarn produced during the shift is not 14, which is not as per requirement.

Example 14.7

A sample of 10 leas has shown lea weights as follows:

$$14.2, 15.0, 14.9, 15.7, 15.4, 14.6, 14.5, 15.0, 15.4, 15.2$$

On the basis of the above results can we say that this sample is selected from the population having mean lea weight 15 units?

Solution

Here,

Population \Rightarrow Collection of leas under study

$X \Rightarrow$ Lea weight

Suppose,

μ is the population mean of the variable X and σ is the population standard deviation of variable X

Thus, interest is to test the hypothesis,

$$H_0: \mu = 15 \quad \text{Vs} \quad H_1: \mu \neq 15$$

For testing the above hypothesis, the small sample of size $n = 10$ is selected

Hence, the statistic t is calculated as follows:

$$t = \frac{\bar{x} - \mu_0}{\hat{s} / \sqrt{n}}$$

Now,

$$\bar{x} = \text{Sample mean} = \frac{\sum x_i}{n} = \frac{149.9}{10} = 14.99$$

$$\hat{s}^2 = \frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2) = \frac{1}{14} (2248.91 - 10 \times 14.99^2) = 0.1364$$

$$\hat{s} = \sqrt{\hat{s}^2} = 0.3693$$

$$\text{Therefore, } t = \frac{14.99 - 15}{0.3693 / \sqrt{10}} = -0.0856$$

Now, at 5% los, that is for $\alpha = 0.05$

$$t_{n-1, \alpha/2} = t_{9, 0.025} = 2.262$$

Here,

$$|t_{\text{cal}}| = 0.0856 < t_{n-1, \alpha/2} = 2.262$$

$$\Rightarrow \text{Accept } H_0 \Rightarrow \mu = 15$$

\Rightarrow Average lea weight of the production is 15 units.

Small sample test for equality of population means (t-test)

This test is used, if the sizes of the samples are small and interest is to test the hypothesis of the following type, related to the means of two different populations under study.

$$\begin{aligned} H_0: \mu = \mu_0 \quad &\text{Vs} \quad H_1: \mu_1 \neq \mu_2 \\ &\text{or} \\ &H_1: \mu_1 > \mu_2 \\ &\text{or} \\ &H_1: \mu_1 < \mu_2 \end{aligned}$$

where,

μ_1 & μ_2 are the population means of variables X_1 & X_2 of two different populations under study.

For testing the hypothesis H_0 , two small samples of sizes n_1 & n_2 are selected. As the sizes of the samples are small the statistic t is defined under the assumptions

1. Both populations under study are the independent normal populations. that is, $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, and both are independent.
2. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and σ^2 is unknown as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where,

\bar{x}_1 & \bar{x}_2 are sample means of two samples defined in same way as discussed earlier

$$S_1^2 = \frac{\sum x_{1i}^2}{n_1} - \bar{x}_1^2 = \frac{1}{n_1} \left(\sum x_{1i}^2 - n_1 \bar{x}_1^2 \right) = \text{Sample variance for first sample}$$

$$S_2^2 = \frac{\sum x_{2i}^2}{n_2} - \bar{x}_2^2 = \frac{1}{n_2} \left(\sum x_{2i}^2 - n_2 \bar{x}_2^2 \right) = \text{Sample variance for second sample}$$

$$\hat{s}_p^2 = \text{Pooled estimate of unknown variance } \sigma^2 = \frac{n_1 \times S_1^2 + n_2 \times S_2^2}{n_1 + n_2 - 2}$$

Here, the statistic “ t ” is expected to follow “ t ” probability distribution $n_1 + n_2 - 2$ with degrees of freedom that is $t \sim t_{n_1+n_2-2}$

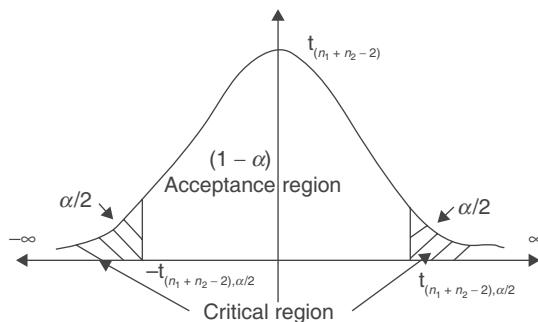
Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

Case I Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu_1 \neq \mu_2$ if, $|t_{cal}| > t_{n_1+n_2-2, \alpha/2}$

where, the point $t_{n_1+n_2-2, \alpha/2}$ is such that,

$$P(|t_{n_1+n_2-2}| > t_{n_1+n_2-2, \alpha/2}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.13.



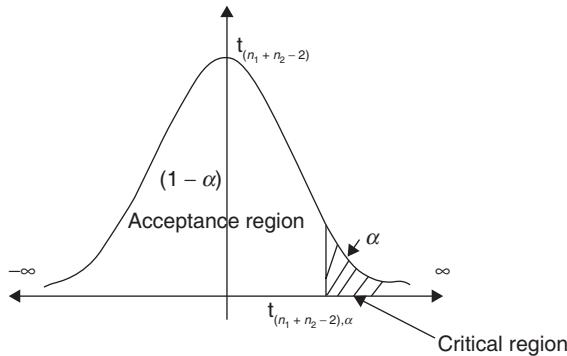
14.13

Case II Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu_1 > \mu_2$ if, $t_{cal} > t_{n_1+n_2-2, \alpha}$

Where, the point $t_{n_1+n_2-2,\alpha}$ is such that,

$$P(t_{n_1+n_2-2} > t_{n_1+n_2-2,\alpha}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.14.



14.14

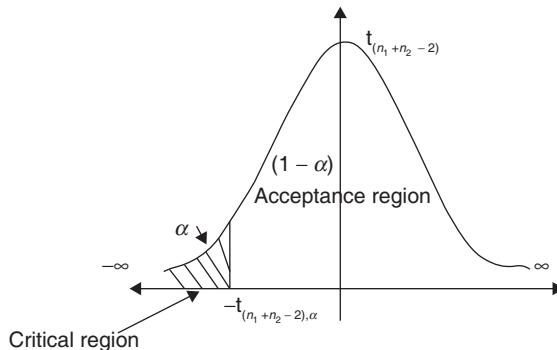
Case III Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \mu_1 < \mu_2$

if, $t_{cal} < -t_{n_1+n_2-2,\alpha}$

where, the point $-t_{n_1+n_2-2,\alpha}$ is such that,

$$P(t_{n_1+n_2-2} < -t_{n_1+n_2-2,\alpha}) = \alpha \text{ i.e. } P(t_{n_1+n_2-2} > t_{n_1+n_2-2,\alpha}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.15.



14.15

Example 14.8

Ten ring-bobbins selected from the production of the day shift and fifteen ring bobbins selected from the production of the night shift have shown following results

	Day shift	Night shift
No of tests	10	15
Average count	40.2	39.3
Std. dev. of count	2.5	3.8

From these sample results, is there any evidence that the yarn spun during night shift is coarser than the day shift? Use 10% *los*.

Solution

Here,

Population 1 \Rightarrow Yarn spun during day shift and Population 2 \Rightarrow Yarn spun during night shift.

X_1 \Rightarrow Count of yarn spun during day shift and X_2 \Rightarrow Count of yarn spun during night shift.

Suppose,

μ_1 and σ_1 are mean and standard deviation of variable X_1 and μ_2 and σ_2 are mean and standard deviation of variable X_2

Thus, interest is to test the hypothesis,

$$H_0: \mu_1 = \mu_2 \quad \text{Vs} \quad H_1: \mu_1 > \mu_2$$

For testing the hypothesis, small samples of sizes, $n_1 = 10$ and $n_2 = 15$ are selected from the two populations under study. Also $\sigma_1 = \sigma_2 = \sigma$ is unknown; hence, the statistic t is calculated as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Given,

$$\bar{x}_1 = 40.2 \text{ & } \bar{x}_2 = 39.3$$

$$S_1 = 2.5 \text{ & } S_2 = 3.8$$

$$\hat{s}_p^2 = \frac{n_1 \times S_1^2 + n_2 \times S_2^2}{n_1 + n_2 - 2} = \frac{10 \times 2.5^2 + 15 \times 3.8^2}{10 + 15 - 2} = 12.1348$$

$$\hat{s}_p = \sqrt{12.1348} = 3.4835$$

$$\text{Therefore, } t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{40.2 - 39.3}{3.4835 \times \sqrt{\frac{1}{10} + \frac{1}{15}}} = 0.6328$$

Here the statistic “ t ” is expected to follow “ t ” probability distribution with $n_1 + n_2 - 2 = 23$ degrees of freedom, that is $t \sim t_{23}$

Therefore at 10% los that is for $\alpha = 0.1$

$$t_{n_1+n_2-2,\alpha} = t_{23,0.1} = 1.319$$

Now,

$$tcal = 0.1055 < t_{n_1+n_2-2,\alpha} = 1.319$$

\Rightarrow Accept $H_0 \Rightarrow \mu_1 = \mu_2 \Rightarrow$ Average count of the yarn spun during two shifts is same.

Therefore we cannot say that yarn spun during night shift is coarser than the day shift.

Example 14.9

Five strength tests each carried out on fabric woven with same raw material on two different looms have shown following results of strength.

Fabric of loom-I	123	122	130	125	128
Fabric of loom-II	125	127	132	130	132

From the above results is there any evidence that the strength of fabric woven on second loom is more than that of first?

Solution

Here,

Population 1 \Rightarrow Fabric woven on first loom and Population 2 \Rightarrow Fabric woven on second loom.

$X_1 \Rightarrow$ Strength of fabric woven on first loom and $X_2 \Rightarrow$ Strength of fabric woven on second loom.

Suppose,

μ_1 and σ_1 are mean and standard deviation of variable X_1 and μ_2 and σ_2 are mean and standard deviation of variable X_2

Thus, interest is to test the hypothesis,

$$H_0: \mu_1 = \mu_2 \quad Vs \quad H_1: \mu_1 < \mu_2$$

For testing the hypothesis, small samples of sizes, $n_1 = 5$ and $n_2 = 5$ are selected from the two populations under study. Also $\sigma_1 = \sigma_2 = \sigma$ is unknown hence, the statistic t is calculated as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

By calculation,

$$\bar{x}_1 = 125.6 \text{ & } \bar{x}_2 = 129.2$$

$$S_1 = 3.0067 \text{ & } S_2 = 2.7857$$

$$\hat{s}_p^2 = \frac{n_1 \times S_1^2 + n_2 \times S_2^2}{n_1 + n_2 - 2} = \frac{5 \times 3.0067^2 + 5 \times 2.7857^2}{5 + 5 - 2} = 10.5$$

$$\hat{s}_p = \sqrt{13.1249} = 3.2404$$

$$\text{Therefore, } t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{125.6 - 130.25}{3.2404 \times \sqrt{\frac{1}{5} + \frac{1}{5}}} = -2.26895$$

Here the statistic “ t ” is expected to follow “ t ” probability distribution with $n_1 + n_2 - 2 = 8$ degrees of freedom, that is $t \sim t_8$

Therefore at 5% *los* that is for $\alpha = 0.05$

$$t_{n_1+n_2-2,\alpha} = t_{8,0.05} = 1.860$$

Now,

$$t_{\text{cal}} = -2.26895 < -t_{n_1+n_2-2,\alpha} = -1.860$$

\Rightarrow Reject $H_0 \Rightarrow \mu_1 < \mu_2 \Rightarrow$ Average strengths of the fabric woven on second looms is more than that of first loom.

Therefore we cannot say that strength of the fabric woven on first loom is lesser than the second loom fabric.

Small sample test for the population variance (χ^2 -test)

This test is used, if the size of the sample is small and interest is to test the hypothesis of the following type, related to variance of the only one population under study.

Suppose,

X is the variable of the population under study.

μ is the mean and σ is standard deviation of the variable X of the population under study.

σ_0 is the expected or the nominal value of the population standard deviation σ . Thus in this test interest is to test the hypothesis,

$$\begin{aligned} H_0: \sigma^2 = \sigma_0^2 &\quad \text{Vs} \quad H_1: \sigma^2 \neq \sigma_0^2 \\ &\quad \text{or} \\ &H_1: \sigma^2 > \sigma_0^2 \\ &\quad \text{or} \\ &H_1: \sigma^2 = \sigma_0^2 \end{aligned}$$

For testing the hypothesis H_0 , a small sample of size “ n ” is selected.

As the size of the sample is small in this case a statistic χ^2 is defined under the assumption that the population under study is normal population that is $X \sim N(\mu, \sigma^2)$ as follows:

$$\begin{aligned} \chi^2 &= \frac{\sum (x_i - \mu)^2}{\sigma_0^2} \quad \text{if } \mu \text{ is known} \\ &= \frac{n \cdot S^2}{\sigma_0^2} \quad \text{if } \mu \text{ is unknown} \end{aligned}$$

Note that in the above formula,

$n \cdot S^2$ can be replaced by $(n-1) \cdot \hat{s}^2$

where,

$$S^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{1}{n} \left(\sum x_i^2 - n\bar{x}^2 \right) = \text{Sample Variance}$$

Here, the statistic χ^2 is expected to follow χ^2 probability distribution with n degrees of freedom if μ is known and it follows χ^2 probability distribution with $(n-1)$ degrees of freedom if μ is unknown.

That is,

$$\chi^2 \sim \chi^2_n \text{ if } \mu \text{ is known and } \chi^2 \sim \chi^2_{n-1} \text{ if } \mu \text{ is unknown.}$$

Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

Case I Population mean μ is known.

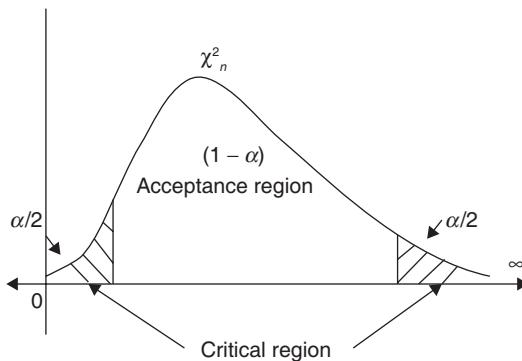
Subcase 1: Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \sigma^2 \neq \sigma_0^2$ if,

$$\chi^2_{cal} > \chi^2_{n, \frac{\alpha}{2}} \quad \text{or} \quad \chi^2_{cal} < \chi^2_{n, 1 - \frac{\alpha}{2}}$$

where, the point $\chi^2_{n,\frac{\alpha}{2}}$ and $\chi^2_{n,1-\frac{\alpha}{2}}$ are such that,

$$P\left(\chi^2 > \chi^2_{n,\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \text{ or } P\left(\chi^2 < \chi^2_{n,1-\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

That is, graphically the critical region can be shown in Fig. 14.16.

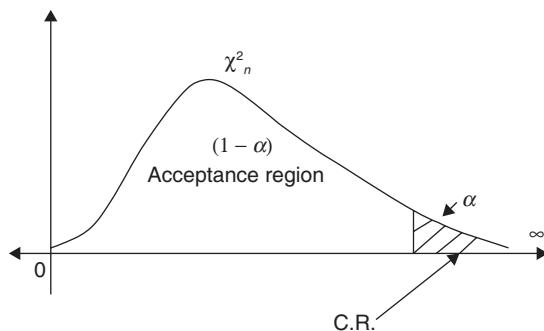


14.16

Subcase 2: Reject null hypothesis H_0 at $100\alpha\%$ los against $H_0: \sigma^2 > \sigma_0^2$ if, $\chi^2_{\text{cal}} > \chi^2_{n,\alpha}$
Where, the point $\chi^2_{n,\alpha}$ is such that,

$$P(\chi^2 > \chi^2_{n,\alpha}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.17.



14.17

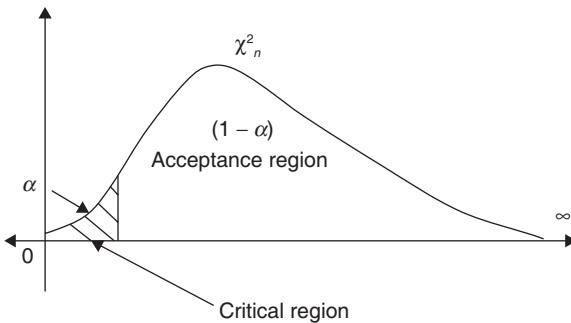
Subcase 3: Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \sigma^2 < \sigma_0^2$ if,

$$\chi_{\text{cal}}^2 < \chi_{n, 1-\alpha}^2$$

Where, the point $\chi_{n, 1-\alpha}^2$ is such that,

$$P(\chi^2 < \chi_{n, 1-\alpha}^2) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.18.



14.18

Case II Population mean μ is unknown.

In this case the critical region for rejection of null hypothesis H_0 is same as that of Case I and can be obtained simply by replacing degrees of freedom “ n ” by “ $n - 1$ ” in all the above subcases of *Case I*.

Example 14.10

A sample of 10 leas has shown lea weights as follows:

14.2, 15.0, 14.9, 15.7, 15.4, 14.6, 14.5, 15.0, 15.4, 15.2

On the basis of the above results can we say that this sample is selected from the population having mean lea weight 15 units and variance of lea weight 1 unit?

Solution

Here,

Population \Rightarrow Collection of leas under study

$X \Rightarrow$ Lea weight.

Suppose,

μ is the population mean of the variable X and σ is the population standard deviation of variable X

Thus, interest is to test the hypothesis,

$$H_0: \sigma^2 = 1 \quad \text{Vs} \quad H_1: \sigma^2 \neq 1$$

As here mean $\mu = 15$ is given and size of the sample selected for study $n = 10$ is small, the statistic χ^2 is calculated as follows:

$$\chi^2 = \frac{\sum(x_i - \mu)^2}{\sigma_0^2} = \frac{1.91}{1} = 1.91$$

Here it is expected that the statistic χ^2 follows χ^2 probability distribution with $n = 10$ degrees of freedom.

Therefore at 5% los that is for $\alpha = 0.05$

$$\chi^2_{n, \frac{\alpha}{2}} = \chi^2_{10, 0.025} \text{ and } \chi^2_{n, 1 - \frac{\alpha}{2}} = \chi^2_{10, 0.975} \text{ are obtained as follows:}$$

$$\chi^2_{10, 0.01} = 23.209 \text{ and } \chi^2_{10, 0.05} = 18.307 \Rightarrow \chi^2_{10, 0.025} = 21.3708$$

$$\chi^2_{10, 0.95} = 3.94 \text{ and } \chi^2_{10, 0.99} = 2.558 \Rightarrow \chi^2_{10, 0.975} = 3.0762$$

Here,

$$\chi^2_{cal} = 1.91 < \chi^2_{n, 1 - \frac{\alpha}{2}} = 3.0762$$

Hence the null hypothesis H_0 is rejected $\sigma^2 \neq 1$. That is variation of lea weight is not one unit.

Example 14.11

The nominal linear density of the yarn spun during a shift is 14 tex with the standard deviation of 0.75 tex. But the sample of 15 leas tested has shown average linear density 14.8 tex. and the CV% 0.25 tex. From the sample results can we say that the production of the shift is having more variation in linear density than the expected?

Solution

Here,

Population \Rightarrow Production of the yarn during the shift

$X \Rightarrow$ Linear density of the yarn.

Suppose,

μ is the population mean of the variable X and σ is the population standard deviation of variable X .

Thus, interest is to test the hypothesis,

$$H_0: \sigma^2 = 0.75^2 \quad \text{Vs} \quad H_1: \sigma^2 > 0.75^2$$

For testing the above hypothesis, the size of sample selected is $n = 15$ which is small.

Hence the statistic χ^2 can be calculated as follows:

$$\chi^2 = \frac{n \cdot S^2}{\sigma_0^2}$$

$$\text{Here Coefficient of Variation} = \text{CV\%} = \frac{s}{\bar{x}} \times 100 = 0.25$$

$$\text{Therefore, } S = \frac{0.25 \times 14.8}{100} = 0.037$$

$$\text{Therefore, } \chi^2 = \frac{n \cdot S^2}{\sigma_0^2} = \frac{15 \times 0.037^2}{0.75^2} = 0.0365$$

Here it is expected that the statistic χ^2 follows χ^2 probability distribution with $n - 1 = 14$ degrees of freedom.

Therefore at 5% los that is for $\alpha = 0.05$

$$\chi^2_{n,\alpha} = \chi^2_{14,0.05} = 23.685$$

$$\text{As } \chi^2_{\text{cal}} = 0.0365 < \chi^2_{14,0.05} = 23.685 \Rightarrow \text{Accept } H_0 \Rightarrow \sigma^2 = 0.75^2$$

Thus from the sample results we say that the production of the shift is not having more variation in linear density than the expected.

Small sample test for equality of population variances (F-test)

This test is used, if the sizes of the samples are small and interest is to test the hypothesis of the following type, related to the variances of two different populations under study.

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{Vs} \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

or

$$H_1: \sigma_1^2 > \sigma_2^2$$

or

$$H_1: \sigma_1^2 < \sigma_2^2$$

where,

μ_1 & μ_2 are the population means of variables X_1 & X_2 and σ_1^2 & σ_2^2 are the variances of two different populations under study.

For testing the hypothesis H_0 , two small samples of sizes n_1 & n_2 are selected. As the sizes of the samples are small, under the assumption of independent normal populations that is $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ and both are independent, the statistic F is defined as follows:

$$F = \begin{cases} \frac{\hat{s}_1^2}{\hat{s}_2^2} & \text{if } \hat{s}_1^2 > \hat{s}_2^2 \\ \frac{\hat{s}_2^2}{\hat{s}_1^2} & \text{if } \hat{s}_1^2 < \hat{s}_2^2 \end{cases}$$

where,

$$\hat{s}_1^2 = \frac{1}{n_1 - 1} \left(\sum x_{1i}^2 - n_1 \bar{x}_1^2 \right) \text{ and } \hat{s}_2^2 = \frac{1}{n_2 - 1} \left(\sum x_{2i}^2 - n_2 \bar{x}_2^2 \right)$$

Here it is expected that this statistic F will follow F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom if $\hat{s}_1^2 > \hat{s}_2^2$ and it is expected to follow F distribution with $n_2 - 1$ and $n_1 - 1$ degrees of freedom if $\hat{s}_1^2 < \hat{s}_2^2$.

Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

Case I $\hat{s}_1^2 > \hat{s}_2^2$

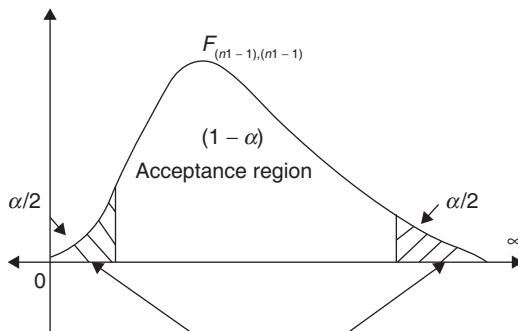
Subcase 1: Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \sigma_1^2 \neq \sigma_2^2$

$$\text{if, } F_{\text{cal}} > F_{n_1-1, n_2-1, \frac{\alpha}{2}} \text{ or } F_{\text{cal}} < F_{n_1-1, n_2-1, 1-\frac{\alpha}{2}}$$

where, the point $P\left(F_{n_1-1, n_2-1} > F_{n_1-1, n_2-1, \frac{\alpha}{2}}\right) = \frac{\alpha}{2}$ or $-P\left(F_{n_1-1, n_2-1} < F_{n_1-1, n_2-1, 1-\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$

are such that,

That is, graphically the critical region can be shown in Fig. 14.19.



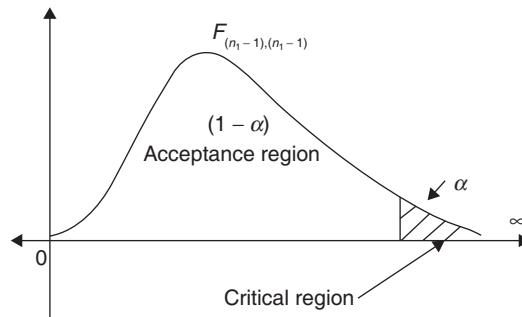
14.19

Subcase 2: Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \sigma_1^2 > \sigma_2^2$ if, $F_{\text{cal}} > F_{n_1-1, n_2-1, \alpha}$

Where, the point $F_{n_1-1, n_2-1, \alpha}$ is such that,

$$P(F_{n_1-1, n_2-1} > F_{n_1-1, n_2-1, \alpha}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.20.

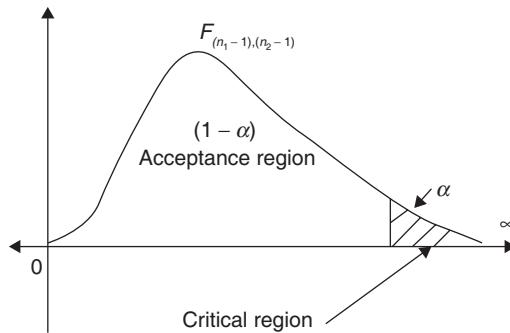


14.20

Subcase 3: Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \sigma_1^2 > \sigma_2^2$ if, $F_{\text{cal}} > F_{n_1-1, n_2-1, \alpha}$
where, the point $F_{n_1-1, n_2-1, \alpha}$ is such that,

$$P(F_{n_1-1, n_2-1} > F_{n_1-1, n_2-1, \alpha}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.21.



14.21

Case II $\hat{s}_1^2 < \hat{s}_2^2$

In this case the criteria for rejection of null hypothesis H_0 is same as the above and can be obtained by interchanging $n_1 - 1$ and $n_2 - 1$ with each other in all the above cases.

Example 14.12

Two ring frames are expected to spin the yarn of the same strength. The samples of 15 and 10 ring bobbins selected from these ring frames have shown following results.

	R/F-1	R/F-2
Sample size	15	10
Mean strength	60 units	56 units
Std. deviation of strength	1.25 units	1.50 units

From the samples results, can it be said that the variation in yarn strength is same for both ring frames? Use 10 % los.

Solution

Here,

Population 1 \Rightarrow Yarn spun by R/F-1 and Population 2 \Rightarrow Yarn spun by R/F-2.

X_1 \Rightarrow Strength of yarn spun by R/F-1 and X_2 \Rightarrow Strength of yarn spun by R/F-2.

Suppose, μ_1 and σ_1 are mean and standard deviation of variable X_1 and μ_2 and σ_2 are mean and standard deviation of variable X_2 .

Thus, interest is to test the hypothesis,

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{Vs} \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

Given that,

$$S_1 = 1.25 \quad \text{and} \quad S_2 = 1.50$$

Now,

$$\hat{s}_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{15}{14} \times 1.25^2 = 1.6741$$

$$\hat{s}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{10}{9} \times 1.5^2 = 2.5$$

Here $\hat{s}_1^2 < \hat{s}_2^2$

$$\text{Therefore, } F = \frac{\hat{s}_2^2}{\hat{s}_1^2} = \frac{2.5}{1.6741} = 1.4933$$

Here it is expected that this statistic F will follow F distribution with $n_1 - 1 = 14$ and $n_2 - 1 = 9$ degrees of freedom. That is $F \sim F_{14,9}$

Therefore at 10% los that is for $\alpha/2 = 0.05$ $F_{n_1-1, n_2-1, \alpha/2} = F_{14,9,0.05}$

$$F_{12,9,0.05} = 3.07 \text{ and } F_{24,9,0.05} = 2.90 \Rightarrow F_{14,9,0.05} = 3.0471$$

As $F_{\text{cal}} = 1.4933 < F_{14,9,0.05} = 3.0471 \Rightarrow \text{Accept } H_0 \Rightarrow \sigma_1^2 = \sigma_2^2 \Rightarrow \text{variation in yarn strength is same for both ring frames}$

Example 14.13

Five strength tests each carried out on fabric woven with same raw material on two different looms have shown following results of strength.

Fabric of loom – I	123	122	130	125	128
Fabric of loom – II	125	127	132	130	132

From the above results, is there any evidence that the strength of fabric woven on second loom is more than that of first?

Solution

Here,

Population 1 \Rightarrow Fabric woven on first loom and Population 2 \Rightarrow Fabric woven on second loom.

X_1 \Rightarrow Strength of fabric woven on first loom and X_2 \Rightarrow Strength of fabric woven on second loom.

Suppose,

μ_1 and σ_1 are mean and standard deviation of variable X_1 and μ_2 and σ_2 are mean and standard deviation of variable X_2

Thus, interest is to test the hypothesis,

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{Vs} \quad H_1: \sigma_1^2 < \sigma_2^2$$

By calculation,

$$\bar{x}_1 = 125.6 \text{ & } \bar{x}_2 = 129.2$$

$$S_1 = 3.0067 \text{ & } S_2 = 2.7857$$

Now,

$$\hat{s}_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{5}{4} \times 3.0067^2 = 11.3$$

$$\hat{s}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{5}{4} \times 2.7857^2 = 9.7$$

Here $\hat{s}_1^2 > \hat{s}_2^2$

$$\text{Therefore, } F = \frac{\hat{s}_1^2}{\hat{s}_2^2} = \frac{11.3}{9.7} = 1.1649$$

Here it is expected that this statistic F will follow F distribution with $n_1 - 1 = 4$ and $n_2 - 1 = 4$ degrees of freedom. That is $F \sim F_{4,4}$

Therefore at 10% los that is for $\alpha/2 = 0.05$ $F_{n_1-1, n_2-1, \alpha/2} = F_{4,4,0.05} = 6.39$

As $F_{\text{cal}} = 1.1649 < F_{14,9,0.05} = 6.39 \Rightarrow \text{Accept } H_0 \Rightarrow \sigma_1^2 = \sigma_2^2 \Rightarrow$ variation in fabric strength is same for both fabrics. That is variation in strength of first loom fabric is not larger than the second loom.

Test for significance of the population correlation coefficient (*t*-test)

This test is used, if the interest is to test the hypothesis related to the correlation coefficient “ ρ ” of the two different variables X and Y of the bivariate population under study which is given as follows:

$$H_0: \rho = 0 \quad \text{Vs} \quad H_1: \rho \neq 0$$

For testing the above hypothesis H_0 , a small sample of size “ n ” is selected and using the property of the *t*-distribution the statistic t is defined as follows:

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

where, $r = r_{xy}$ = Sample correlation coefficient between the variables X and Y which is already discussed in previous chapter.

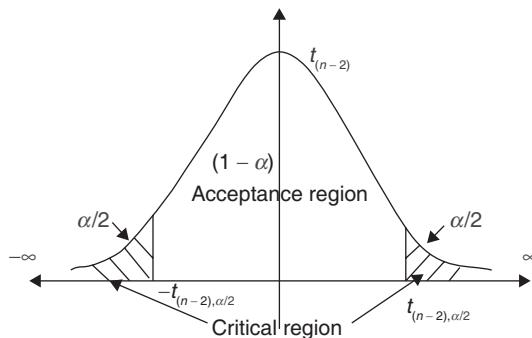
Here, the statistic t is expected to follow the *t* probability distribution with $(n - 2)$ degrees of freedom. Hence, the rejection criteria for the null hypothesis H_0 (Critical Region) can be given as follows:

Reject null hypothesis H_0 at $100\alpha\%$ los against $H_1: \rho \neq 0$ if, $|t_{\text{cal}}| > t_{n-2, \alpha/2}$

Where, the point $t_{n-2, \alpha/2}$ is such that,

$$(|t_{n-2}| > t_{n-2, \alpha/2}) = \alpha$$

That is, graphically the critical region can be shown in Fig. 14.22.



14.22

Example 14.14

10 sample regions selected from India have shown correlation coefficient 0.5214 between price and demand of the cotton fabric. From this result can we say that there is significant correlation between price and demand of the cotton fabric?

Solution

Suppose, ρ is the correlation coefficient between two different variables X (price) and Y (demand of cotton fabric).

Thus interest is to test,

$$H_0: \rho = 0 \quad \text{Vs} \quad H_1: \rho \neq 0$$

For testing the hypothesis H_0 , a sample of size “ $n = 10$ ” is selected and using the property of the t -distribution the statistic t can be calculated as follows:

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2} = \frac{0.5214}{\sqrt{1-0.5214^2}} \times \sqrt{10-2} = 1.7283$$

Here, the statistic t is expected to follow the t probability distribution with $(n-2)=8$ degrees of freedom.

At 5% los, that is for $\alpha = 0.05$, $t_{(n-2), \alpha/2} = t_{8, 0.025} = 2.306$

Now, $|t_{\text{cal}}| = 1.7283 < t_{8, 0.025} = 2.306$

Hence, the null hypothesis H_0 is accepted $\Rightarrow \rho = 0 \Rightarrow$ price and demand of the cotton fabric are not significantly correlated.

Test for goodness of fit (χ^2 -test)

Sometimes in statistics, the data under study is expected to follow a specific probability distribution or the probability law. To verify, whether the data actually follow this particular probability distribution or the probability law, χ^2 -test for the goodness of fit is used.

In this case, first the expected probability distribution or the probability law is fitted to the given data and the expected frequencies are obtained. After fitting, the observed frequencies are compared with the expected frequencies to decide whether the data follows the expected probability distribution or the probability law.

Thus, the hypothesis to be tested is,

H_0 : The observed frequencies (O_i) and the expected frequencies (E_i) do not differ significantly, that is, the fitting is good.

Vs

H_1 : The observed frequencies (O_i) and the expected frequencies (E_i) differ significantly, that is, the fitting is not good.

Where,

O_1, O_2, \dots, O_k and E_1, E_2, \dots, E_k are the observed frequencies and the expected frequencies corresponding to the k different classes of the given data.

For testing the above hypothesis, the statistic χ^2 is calculated from the given data, using the formula,

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Here, it is expected that the statistic χ^2 will follow the χ^2 -probability distribution with the $(k - r - 1)$ degrees of freedom.

Where,

k represents effective number of classes

r represents number of parameters calculated (estimated) for the fitting of the probability distribution or the probability law.

Criterion for the rejection of null hypothesis H_0

Reject null hypothesis H_0 at $100\alpha\%$ los, if $\chi^2_{\text{cal}} > \chi^2_{(k-r-1),\alpha}$

Where,

The point $\chi^2_{(k-r-1),\alpha}$ is such that,

$$P(\chi^2 > \chi^2_{(k-r-1),\alpha}) = \alpha$$

Note that,

1. Sometimes, after fitting the probability distribution or the probability law, the expected frequencies of the extreme classes may become less than five by calculation, in such cases the continuity correction must be applied before calculating Chi-square statistic. The continuity correction can be applied by adding the expected frequencies until the total becomes greater than or equal to five.
2. While fitting the Binomial distribution the parameter “ p ” is calculated or estimated from the data if it is unknown by using the formula,

$$P = \frac{\bar{x}}{n} = \frac{1}{n} \left(\frac{\sum f_i \cdot x_i}{N} \right)$$

Also the required probabilities can be obtained by using the pmf of the Binomial distribution or the recurrence relation for the Binomial distribution given as follows:

$$P(X = x + 1) = \frac{p}{q} \cdot \frac{(n - x)}{(x + 1)} \cdot P(X = x)$$

3. While fitting the Poisson distribution the parameter “ λ ” is calculated or estimated from the data if it is unknown by using the formula

$$\lambda = \bar{x} = \left(\frac{\sum f_i \cdot x_i}{N} \right)$$

Also the required probabilities can be obtained by using the pmf of the Poisson distribution or the recurrence relation for the Poisson distribution given as follows:

$$P(X = x + 1) = \frac{\lambda}{(x + 1)} \cdot P(X = x)$$

Example 14.15

Following is the distribution of 200 samples of five garments each, according to the number of defective garments.

Number of defective garments	0	1	2	3	4	5	Total
Number of samples	50	80	40	15	10	5	200

Fit the Binomial probability distribution to the above data and test the goodness of fit assuming,

- (i) Probability of the defective garments “ p ” is unknown.
- (ii) 20% of the garments are defective.

Solution

Here, X —Number of defective garments in a sample of 5.

Let $X \sim B(n = 5, P)$

Case I Here, parameter “ p ” is unknown therefore it is calculated by preparing Table 14.1

Table 14.1

X	0	1	2	3	4	5	Total
f_i	50	80	40	15	10	5	200
$f_i \cdot x_i$	0	80	80	45	40	25	270

$$P = \frac{\bar{x}}{n} = \frac{1}{n} = \left(\frac{\sum f_i \cdot x_i}{N} \right) = \frac{1}{5} \left(\frac{270}{200} \right) = 0.27 \Rightarrow q = 0.73$$

Now,

By using pmf of Binomial distribution the calculations for fitting are shown in Table 14.2.

Table 14.2

X	Observed frequency O_i	$p(x) = \binom{n}{x} p^x q^{n-x}$	Excepted frequency $E_i = N \times p(x)$
0	50	0.2073	41.46
1	80	0.3834	76.68
2	40	0.2836	56.72
3	15	0.1049	20.98
4	10	0.0194	3.88
5	5	0.0014	0.28
Total	200	1	200

Or by using recurrence relation of Binomial distribution the calculations for fitting are shown in Table 14.3.

Table 14.3

X	Observed frequency O_i	$\frac{p}{q} \cdot \frac{(n-x)}{(x+1)}$	$p(x+1) = \frac{p}{q} \cdot \frac{(n-x)}{(x+1)} \cdot P(x)$	Excepted frequency $E_i = N \times p(x)$
0	50	1.849315	0.2073	41.46
1	80	0.739726	$1.8493 \times 0.2073 = 0.3834$	76.68
2	40	0.369863	0.2836	56.72
3	15	0.184932	0.1049	20.98
4	10	0.073973	0.0194	3.88
5	5	0	0.0014	0.28
Total	200		1	200

Now the interest is to test the hypothesis,

H_0 : The observed frequencies (O_i) and the expected frequencies (E_i) do not differ significantly that is, the fitting is good.

Vs

H_1 : The observed frequencies (O_i) and the expected frequencies (E_i) differ significantly that is, the fitting is not good.

For testing the above hypothesis, χ^2 the statistic is calculated from the given data, using the following formula and the calculations are shown in Table 14.4,

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Table 14.4

X	Observed frequency O_i	Expected frequency E_i	$\frac{(O_i - E_i)^2}{E_i}$
0	50	41.46	1.7591
1	80	76.68	0.1437
2	40	56.72	4.9287
3–5	30	25.14	0.9395
Total			7.771

$$\text{Therefore, } \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 7.771$$

Here, it is expected that the statistic χ^2 will follow the χ^2 -probability distribution with the $(k - r - 1) = 4 - 1 - 1 = 2$ degrees of freedom.

Thus, at 5% los, that is for $\alpha = 0.05$, $\chi^2_{(k-r-1), \alpha} = \chi^2_{2, 0.05} = 5.991$

As, $\chi^2_{\text{cal}} = 7.771 > \chi^2_{(k-r-1), \alpha} = 5.991$

The null hypothesis H_0 is rejected $\Rightarrow H_1$ accepted \Rightarrow the observed frequencies (O_i) and the expected frequencies (E_i) differ significantly, that is, the fitting is not good.

Case II Here, it is given that the probability of the defective garments is 0.20, that is the parameter $p = 0.20$ and $q = 0.80$.

Now, by using pmf of Binomial distribution the calculations for fitting are shown in Table 14.5.

Table 14.5

X	Observed frequency O_i	$p(x) = \binom{n}{x} p^x q^{n-x}$	Expected frequency $E_i = N \times p(x)$
0	50	0.3277	65.54
1	80	0.4096	81.92
2	40	0.2048	40.96
3	15	0.0512	10.24
4	10	0.0064	1.28
5	5	0.0003	0.06
Total	200	1	200

Again, for testing the above hypothesis, the statistic χ^2 is calculated as above using the given data and the formula,

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 33.0521$$

Here, it is expected that the statistic χ^2 will follow the χ^2 -probability distribution with the $(k - r - 1) = 4 - 0 - 1 = 3$ degrees of freedom.

Thus, at 5% los, that is for $\alpha = 0.05$, $\chi^2_{(k-r-1), \alpha} = \chi^2_{3, 0.05} = 7.815$

As, $\chi^2_{\text{cal}} = 33.0521 > \chi^2_{(k-r-1), \alpha} = 7.815$

The null hypothesis H_0 is rejected $\Rightarrow H_1$ accepted \Rightarrow the observed frequencies (O_i) and the expected frequencies (E_i) differ significantly, that is, the fitting is not good. Thus the above data do not follow binomial distribution with $n = 5$ and $p = 0.20$.

Example 14.16

Following is the distribution of 200 garments, according to the number of defects in a garment.

Number of defects in garment	0	1	2	3	4	5	Total
Number of garments	50	80	40	15	10	5	200

Fit the Poisson probability distribution to the above data and test the goodness of fit assuming,

- (i) Average number of defects in a garment “ λ ” is unknown.
- (ii) On an average one defect per garment.

Solution

Here, X – Number of defects in a garment.

Let $X \sim P(\lambda)$

Case I Here, parameter “ λ ” is unknown preparing Table 14.6

Table 14.6

X	0	1	2	3	4	5	Total
f_i	50	80	40	15	10	5	200
$f_i \cdot x_i$	0	80	80	45	40	25	270

$$\lambda = \bar{x} = \left(\frac{\sum f_i \cdot x_i}{N} \right) = \left(\frac{270}{200} \right) = 1.35$$

Now,

By using pmf of Poisson distribution the calculations for fitting are shown in Table 14.7.

Table 14.7

X	Observed frequency O_i	$p(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$	Excepted frequency $E_i = N \times p(x)$
0	50	0.2592	51.84805
1	80	0.3500	69.99487
2	40	0.2362	47.24654
3	15	0.1063	21.26094
4	10	0.0359	7.175568
5	5	0.0124	2.47403
Total	200	1	200

By using recurrence relation of Poisson distribution the calculations for fitting are shown in Table 14.8.

Table 14.8

X	Observed frequency O_i	$\frac{\lambda}{(x+1)}$	$p(x+1) = \frac{\lambda}{(x+1)} \cdot P(x)$	Excepted frequency $E_i = N \times p(x)$
0	50	1.35	0.2592	51.84805
1	80	0.675	$1.35 \times 0.2592 = 0.3500$	69.99487
2	40	0.45	0.2362	47.24654
3	15	0.3375	0.1063	21.26094
4	10	0.27	0.0359	7.175568
5	5	0.225	0.0124	2.47403
Total	200		1	200

Now the interest is to test the hypothesis,

H_0 : The observed frequencies (O_i) and the expected frequencies (E_i) do not differ significantly, that is, the fitting is good.

Vs

H_1 : The observed frequencies (O_i) and the expected frequencies (E_i) differ significantly, that is, the fitting is not good.

For testing the above hypothesis, the statistic χ^2 is calculated from the given data, using the formula and the calculations are shown in Table 14.9,

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Table 14.9

X	Observed frequency O_i	Expected frequency E_i	$\frac{(O_i - E_i)^2}{E_i}$
0	50	51.84805	0.065871
1	80	69.99487	1.430142
2	40	47.24654	1.111453
3	15	21.26094	1.843728
4 and more	15	9.649598	2.966631
Total	200	200	7.4178

$$\text{Therefore, } \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 7.4178$$

Here, it is expected that the statistic χ^2 will follow the χ^2 -probability distribution with the $(k - r - 1) = 4 - 1 - 1 = 2$ degrees of freedom.

Thus, at 5% los, that is for $\alpha = 0.05$, $\chi^2_{(k-r-1), \alpha} = \chi^2_{2, 0.05} = 5.991$

$$\text{As, } \chi^2_{\text{cal}} = 7.4178 > \chi^2_{(k-r-1), \alpha} = 5.991$$

The null hypothesis H_0 is rejected $\Rightarrow H_1$ accepted \Rightarrow the observed frequencies (O_i) and the expected frequencies (E_i) differ significantly, that is, the fitting is not good.

Case II Here, it is given that average number of the defects in a garment is 1, that is the parameter $\lambda = 1$. Now, by using pmf of Poisson distribution the calculations for fitting are shown in Table 14.10.

Table 14.10

X	Observed frequency O_i	$p(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$	Expected frequency $E_i = N \times p(x)$
0	50	0.3679	73.57589
1	80	0.3679	73.57589
2	40	0.1839	36.78794
3	15	0.0613	12.26265
4	10	0.0153	3.065662
5	5	0.0031	0.613132
Total	200	1	200

Again, for testing the above hypothesis, the statistic χ^2 is calculated as above using the given data and the formula,

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 20.7939$$

Here, it is expected that the statistic χ^2 will follow the χ^2 -probability distribution with the $(k - r - 1) = 4 - 0 - 1 = 3$ degrees of freedom.

Thus, at 5% los, that is for $\alpha = 0.05$, $\chi^2_{(k-r-1), \alpha} = \chi^2_{3,0.05} = 7.815$

As, $\chi^2_{\text{cal}} = 20.7939 > \chi^2_{(k-r-1), \alpha} = 7.815$

The null hypothesis H_0 is rejected $\Rightarrow H_1$ accepted \Rightarrow the observed frequencies (O_i) and the expected frequencies (E_i) differ significantly, that is, the fitting is not good.

Thus the above data do not follow Poisson distribution with $\lambda=1$.

χ^2 test for the independence of the attributes

Sometimes in statistics one is interested in checking dependence or association of two different attributes on the basis of the data related to these two different attributes under study. This can be easily done with the help of the χ^2 test for the independence of the attributes. For example, sometimes we may be interested in knowing dependence of eye color of son with the eye color of father or a textile garment manufacturer may be interested in knowing whether the choice of color of garments depends upon the age of the person. The test procedure is as follows:

Suppose A and B are the two different attributes whose association or dependence is to be tested. Further suppose that, attribute A has m levels or classes denoted by A_1, A_2, \dots, A_m and attribute B has n levels or classes denoted by B_1, B_2, \dots, B_n .

Thus, interest is to test the hypothesis,

H_0 : Attribute A does not depend on attribute B (A and B are independent).

Vs

H_1 : Attribute A depends on attribute B .

For testing the above hypothesis, suppose a sample of size N is selected and is classified in the form of contingency table and shown in Table 14.11.

Table 14.11

$A B$	B_1	B_2	...	B_n	Total
A_1	O_{11}	O_{12}		O_{1n}	(A_1)
A_2	O_{21}	O_{22}		O_{2n}	(A_2)
...					
A_m	O_{m1}	O_{m2}		O_{mn}	(A_m)
Total	(B_1)	(B_2)		(B_n)	N

where,

O_{ij} is observed frequency corresponding to the levels A_i and B_j

E_{ij} is expected frequency corresponding to the levels A_i and B_j

$$E_{ij} = \frac{(Ai) \cdot (Bj)}{N}$$

(A_i) is the total of the frequencies corresponding to the level A_i .

(B_j) is the total of the frequencies corresponding to the level B_j .

Now, for testing the above null hypothesis H_0 , the statistic χ^2 is calculated as follows:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Here it is expected that, the statistic will χ^2 follow χ^2 -probability distribution with $(m - 1) \cdot (n - 1)$ degrees of freedom.

Criterion for rejection of null hypothesis H_0

Reject null hypothesis H_0 at $100\alpha\%$ los, if $\chi^2_{\text{cal}} > \chi^2_{(m-1),(n-1),\alpha}$

Where, the point $\chi^2_{(m-1),(n-1),\alpha}$ is such that,

$$P(\chi^2 > \chi^2_{(m-1),(n-1),\alpha}) = \alpha$$

Graphically the critical region can be shown in the same way as above by changing degrees of freedom as $(m - 1) \cdot (n - 1)$.

Example 14.17

From the following data collected by the readymade garment manufacturer, is it possible to say that, the choice of the color depends on the age of the customer?

		Color of the jeans			Total
		Light-blue	Blue	Black	
Age (years)	3–10	50	60	75	185
	11–15	40	80	100	220
	16–30	30	90	25	145
	Total	120	230	200	550

Solution

Suppose, Attribute A denotes age of the customer and attribute B denotes choice of color of the jeans.

Thus, the interest is to test the hypothesis,

H_0 : Choice of the color of the jeans does not depend on the age of the customer.
Vs

H_1 : Choice of the color of the jeans depends on the age of the customer.

For testing the above hypothesis, the sample of size $N = 550$ is selected and the statistic χ^2 can be calculated and are shown in Table 14.12.

Table 14.12

		Choice of color of the jeans			Total
		Light-blue	Blue	Black	
Age (years)	3–10	50/40.36	60/77.36	75/67.27	185
	11–15	40/48	80/92	100/80	220
	16–30	30/31.64	90/60.64	25/52.73	145
	Total	120	230	200	550

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 43.9317$$

Now, it is expected that the statistic χ^2 will follow χ^2 -probability distribution with $(m - 1) \cdot (n - 1) = (3 - 1) \cdot (3 - 1) = 4$ degrees of freedom.

Thus, at 5% los, that is for $\alpha = 0.05$, $\chi^2_{(m-1) \cdot (n-1), \alpha} = \chi^2_{4, 0.05} = 9.488$

Here, $\chi^2_{\text{cal}} = 43.9317 > \chi^2_{(m-1) \cdot (n-1), \alpha} = \chi^2_{4, 0.05} = 9.488$

Hence null hypothesis H_0 is rejected \Rightarrow Accept $H_1 \Rightarrow$ Choice of the color of the jeans depends on the age of the customer.

14.4 Exercise

- What is the testing of hypothesis? Explain the test for the equality of population proportion of two different populations.
- The average weight of the knitted garment is expected to be 200 g with std. dev. 10 g. From the production line the manager had selected 35 garments and he finds that average weight of the 35 garments is 212 g. Using Z test, can we say that the average weight of the garment is more than the expected?
- A ring bobbin manufacturing company claims that 5% of its ring bobbins are defective. A customer had selected 75 ring bobbins and found that 10 ring bobbins were defective. From the sample results, can you say that the proportion of defective ring bobbins is more than the expected?
- Describe the *F*-test used for testing equality of the variation of two different populations. If ten ring-bobbins each selected from the day shift and night shift productions have shown standard deviations of the count as 1.2 and 2.2 units respectively, can you say that variation in count is more for the night shift production than that of day shift production?
- Forty hank tests made on the sliver of a draw frame have shown average hank as 0.1245 and standard deviation as 0.0025. Using the *Z*-test, can you say that the nominal hank of the sliver is 0.12 units?
- Ten tests each, made on two different yarns have shown standard deviations of count and strength product [CSP] as 8.5 and 6.2 respectively. Using the *F*-test can we say that the variation in count and strength product [CSP] is more for the first yarn than that of second yarn?
- Applying chi-square test to the following data, can we say that the choice of color depends on the age of the person?

	Color		
	Light	Medium	Dark
Below 20	15	28	65
20–30	52	40	20
Age 30–40	10	55	25
40 & above	20	40	05

8. What is the t -test for the significance of population correlation coefficient? If the sample of size 10 selected from a population has shown correlation coefficient 0.712, can we say that the population correlation coefficient is significant?
9. A spinning master had tested 50 ring bobbins each, from two ring-frames and had found following results.

	R/F - A	R/F - B
Average strength	14.5	15.8
Standard deviation	0.50	0.65

Using Z -test, can we say that the average strength of the yarn spun by ring-frame B is more than that of ring-frame A?

10. A company had supplied knitting needles to the customer and claimed that 5% of its needles may be defective. The customer had selected 100 needles from the consignment and found that 9 needles are defective. Using Z -test can we say that the claim of the company is not true?
11. Five ring bobbins each selected from two ring frames were tested for the count and the results are as follows:

R/F I	35.5	36.2	37.0	35.8	36.0
R/F II	37.2	36.8	37.5	36.9	35.9

Using F -test, can we say that variation in count is same for both ring frames? Take 10% α .

12. Five hank tests made on a sliver have shown following results:
0.1195, 0.1205, 0.12, 0.1210, 0.1190.

Using the above sample results, can the manufacturer claim that the average and standard deviation of the hank of the sliver are 0.12 and 0.0025 units?

13. The results of the strength tests made on the two types of fabrics are as follows:

	Fabric A	Fabric B
No. of tests	10	10
Average strength	6.0 kg	6.5 kg
Std. dev.	1.5 kg	1.2 kg

Using t -test, can we say that the strength of the Fabric A is less than the Fabric B?

14. The average weight of the knitted garment is expected to be 200 g with std. dev. 10 gms. From the production line the manager had selected 35 garments and he found that average weight of the 35 garments is 212 g. Using Z test, can we say that the average weight of the garment is more than the expected?
15. Five count tests each, made on day shift and night shift production have shown following results:

Day	37.2	36.8	37.5	36.9	35.9
Night	33.5	36.2	37.0	35.8	36.2

Using *F*-test, can we say that variation in the count is more for the night shift production?

16. Following data represents number of weaving defects observed in 200 pieces of cloth:

No. of weaving defects	0	1	2	3	4	5	6
No. of pieces	70	60	25	20	10	8	7

Fit the Poisson probability distribution to the above data and test the goodness of fit.

17. Ten ring bobbins selected from each ring frame have shown following results.

	R/F-I	R/F-II
Average strength	16.12	16.20
Standard deviation	1.5	2.6

Using *F*-test, can we say that the variation in strength is more for the second ring frame than the first?

18. Twenty leas tested for the count have shown average and c.v.% of the count as 40.8 and 1.2 units resp. Using *t*-test, can we say that the average count of the lea is 40 units?
19. A spinning master had selected five ring bobbins each from two ring-frames and found following results:

	R/F-I	R/F-II
Average linear density	14.08	14.80
Standard deviation	0.0136	0.048

From the above results and using F -test, can we say that the yarn spun by ring-frame II has more variation in linear density than that of first?

20. A yarn was tested for the single thread strength and following results were obtained:

15.5, 14.5, 12.5, 16.5, 16.4, 15.6, 11.8, 12.4

Using the t -test, can we say that these results support the hypothesis Average strength of the yarn is 15 units?

21. Following is the distribution of 300 packs of 5 needles each according to defective needles.

No. of defective needles	0	1	2	3	4	5
No. of packs	80	100	70	20	20	10

Fit the binomial probability distribution to the above data to find expected frequencies and test the goodness of fit.

15.1 Introduction

Estimation means finding the values of the unknown population parameters on the basis of sample selected from that population. Point estimation and Interval estimation are the two different types of the estimation. For example when we carry out some sample count tests on the yarn and on the basis of those sample results when we find mean such as “39.5,” then it is point estimation, but when we find the interval of type (39.2,40.8), then it is interval estimation.

15.2 Point estimation

When the unknown population parameter is estimated in the form of the single value, on the basis of sample selected from that population, then it is called the point estimation. Unbiased estimation, consistent estimation, sufficient estimation etc. are the different types of estimation.

Unbiased estimator

A statistic “ T ” is called the unbiased estimator of the unknown population parameter θ , if $E(T) = \theta$

Results

1. The sample mean \bar{X} is the unbiased estimator of the population mean μ . That is,

$$E(\bar{X}) = \mu$$

Proof

Suppose, X_1, X_2, \dots, X_n is the sample of size “ n ” selected from a population with mean μ and variance σ^2 . That is these are the “ n ” independent random variables with mean

$$E(X_i) = \mu \text{ and variance } V(X_i) = \sigma^2.$$

Now,

$$E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}(E(\sum X_i)) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\sum \mu = \frac{1}{n}n\mu = \mu$$

2. The sample variance of mean is $V(\bar{X}) = \sigma^2/n$

Proof

$$\begin{aligned} V(\bar{X}) &= V\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2}(V(\sum X_i)) = \frac{1}{n^2}\sum V(X_i) = \frac{1}{n^2}\sum \sigma^2 \\ &= \frac{1}{n^2}n\sigma^2 = \sigma^2/n \end{aligned}$$

3. The sample mean square $\hat{S}^2 = \frac{1}{n-1}(\sum x_i^2 - n\bar{x}^2)$ is unbiased estimator of population variance σ^2 . That is $E(\hat{S}^2) = \sigma^2$

Proof

Consider,

$$\begin{aligned} \sum(X_i - \mu)^2 &= \sum(X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum(X_i - \bar{X})^2 - 2(\bar{X} - \mu)\sum(X_i - \bar{X}) + \sum(\bar{X} - \mu)^2 \\ &= \sum(X_i - \bar{X})^2 + n \cdot (\bar{X} - \mu)^2 \end{aligned}$$

$$\text{Therefore, } \sum(X_i - \bar{X})^2 = \sum(X_i - \mu)^2 - n \cdot (\bar{X} - \mu)^2$$

Dividing both sides by $(n - 1)$,

$$\text{Therefore, } \hat{S}^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{1}{n-1}\sum(X_i - \mu)^2 - \frac{n}{n-1}(\bar{X} - \mu)^2$$

$$\text{Therefore, } E(\hat{S}^2) = E\left(\frac{1}{n-1}\sum(X_i - \mu)^2\right) - \frac{n}{n-1}E(\bar{X} - \mu)^2$$

$$\text{Therefore, } E(\hat{S}^2) = \left(\frac{1}{n-1}\sum E(X_i - \mu)^2\right) - \frac{n}{n-1}E(\bar{X} - \mu)^2$$

$$\text{Therefore, } E(\hat{S}^2) = \left(\frac{1}{n-1}\sum \sigma^2\right) - \frac{n}{n-1}\frac{\sigma^2}{n} = \left(\frac{1}{n-1}n \cdot \sigma^2\right) - \frac{\sigma^2}{n-1}$$

$$\text{Therefore, } E(\hat{S}^2) = \sigma^2$$

15.3 Interval estimation (confidence interval)

The procedure of estimating or finding the value of the unknown population parameter θ , on the basis of the sample selected from that population, in the form of the interval (L, U) is called the interval estimation. The interval estimate is always obtained with some guarantee or the confidence; hence it is also called the confidence interval. The confidence of the confidence interval is always given in percentage and can be represented by $(1 - \alpha)100\%$. Where α is related to the los.

For example, if los = 5%, then $\alpha = 0.05$

$(1 - \alpha)100\% = 95\%$ is the confidence of the confidence interval (CI) $\Rightarrow p\{L < \theta < U\} = 1 - \alpha$

Results

1. Confidence interval for the population mean “ μ ” based on the large sample

If the size of the sample “ n ” selected for the study is large, then 100 $(1 - \alpha)\%$ confidence Interval for the population mean “ μ ” can be given as follows:

$$\left(\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \quad \text{If } \sigma \text{ is known}$$

$$\left(\bar{x} - Z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{S}{\sqrt{n}} \right) \quad \text{If } \sigma \text{ is unknown}$$

Where, S is sample standard deviation.

Proof

From the property of the standard normal distribution, we have

$$P[|Z| > Z_{\alpha/2}] = \alpha$$

$$\text{Therefore, } P[|Z| \leq Z_{\alpha/2}] = 1 - \alpha$$

$$\text{Therefore, } P[-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}] = 1 - \alpha$$

$$\text{Therefore, } P\left[-Z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\right] = 1 - \alpha$$

$$\text{Therefore, } P\left[-Z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq \bar{x} - \mu \leq Z_{\alpha/2} \cdot \sigma/\sqrt{n}\right] = 1 - \alpha$$

$$\text{Therefore, } P\left[-\bar{x} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq -\mu \leq -\bar{x} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}\right] = 1 - \alpha$$

$$\text{Therefore, } P\left[\bar{x} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \bar{x} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}\right] = 1 - \alpha$$

Thus, $100(1 - \alpha)\%$ confidence Interval for the population mean “ μ ” can be given as follows:

$$\left(\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

Further, if los is 5%, that is the confidence is 95%, then the 95% confidence interval for the population mean μ based on the large sample is,

$$\left(\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \quad \text{If } \sigma \text{ is known}$$

$$\left(\bar{x} - 1.96 \times \frac{S}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{S}{\sqrt{n}}\right) \quad \text{If } \sigma \text{ is unknown}$$

Also, if los is 1%, that is the confidence is 99%, then the 99% confidence interval for the population mean μ based on the large sample is,

$$\left(\bar{x} - 2.575 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.576 \times \frac{\sigma}{\sqrt{n}}\right) \quad \text{If } \sigma \text{ is known}$$

$$\left(\bar{x} - 2.575 \times \frac{S}{\sqrt{n}}, \bar{x} + 2.575 \times \frac{S}{\sqrt{n}}\right) \quad \text{If } \sigma \text{ is unknown}$$

Example 15.1

Thirty-five hank tests made on a draw frame sliver have shown the mean and the standard deviation of the hank as the 0.12 and 0.025 respectively. Find 95 and 99% confidence interval for the mean hank of the sliver and comment on each of them.

Solution

Here, Population is the production of the draw frame sliver.

Suppose, X denotes hank of the sliver, μ is the mean and σ is the standard deviation of the hank of the sliver.

Now, as the sample size $n = 35$ is large and σ is unknown, the $100(1 - \alpha)\% = 95\%$ confidence interval for the average hank is

$$\left(\bar{x} - 1.96 \times \frac{S}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{S}{\sqrt{n}} \right)$$

also, given that $\bar{x} = 0.12$ and $S = 0.025$.

Thus, substituting values of \bar{x} and S , the 95% confidence interval for the population mean μ is $(0.1117, 0.1283)$.

That is, with 95% guarantee the value of the population mean μ (average hank of the sliver) lies in between 11.9917 to 12.0083 units.

2. Confidence interval for the population mean “ μ ” based on the small sample

If the size of the sample “ n ” selected for the study is small, then $100(1 - \alpha)\%$ confidence Interval for the population mean “ μ ” can be given as follows:

$$\left(\bar{x} - t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}, \bar{x} + t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}} \right)$$

Where, \hat{s} is estimate of population standard deviation.

Proof

From the property of t-probability distribution, we have

$$P\left[|t| > t_{(n-1), \alpha/2}\right] = \alpha$$

$$\text{Therefore, } P\left[|t| \leq t_{(n-1), \alpha/2}\right] = 1 - \alpha$$

$$\text{Therefore, } P\left[-t_{(n-1), \alpha/2} \leq t \leq t_{(n-1), \alpha/2}\right] = 1 - \alpha$$

$$\text{Therefore, } P\left[-t_{(n-1), \alpha/2} \leq \frac{\bar{x} - \mu}{\hat{S}/\sqrt{n}} \leq t_{(n-1), \alpha/2}\right] = 1 - \alpha$$

Therefore, $P\left[-t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}} \leq \bar{x} - \mu \leq t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}\right] = 1 - \alpha$

Therefore, $P\left[-\bar{x} - t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}} \leq -\mu \leq -\bar{x} + t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}\right] = 1 - \alpha$

Therefore, $P\left[\bar{x} - t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}\right] = 1 - \alpha$

Thus, $100(1 - \alpha)\%$ confidence Interval for the population mean “ μ ” can be given as follows:

$$\left(\bar{x} - t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}, \bar{x} + t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}\right)$$

3. Confidence interval for the population variance “ σ^2 ” based on the small sample

If the size of the sample “ n ” selected for the study is small, then $100(1 - \alpha)\%$ Confidence Interval for the population variance “ σ^2 ” can be given as follows:

$$\left(\frac{\sum(x_i - \mu)^2}{\chi^2_{n, \alpha/2}}, \frac{\sum(x_i - \mu)^2}{\chi^2_{n, (1-\frac{\alpha}{2})}}\right) \quad \text{If } \mu \text{ is known}$$

$$\left(\frac{nS^2}{\chi^2_{n-1, \alpha/2}}, \frac{nS^2}{\chi^2_{n-1, (1-\frac{\alpha}{2})}}\right) \text{ or } \left(\frac{(n-1)\hat{S}^2}{\chi^2_{n-1, \alpha/2}}, \frac{(n-1)\hat{S}^2}{\chi^2_{n-1, (1-\frac{\alpha}{2})}}\right) \quad \text{If } \mu \text{ is unknown}$$

Example 15.2

Ten leas were weighed and following results were obtained.

1.55, 1.50, 1.62, 1.58, 1.68, 1.50, 1.55, 1.65, 1.59, 1.58

From the above results find the unbiased estimates of the average and the variance of lea weight. Also find the 95% confidence interval for the average and variance of lea weight each and comment on them.

Solution

Here, Population is the collection of leas or the production of the yarn.

Suppose, X denotes lea weight, μ is the mean and σ is the standard deviation of the lea weight.

Now,

Unbiased estimator of population mean μ is sample mean $\bar{x} = \frac{\sum x_i}{n} = 1.58$

Unbiased estimator of population Variance σ^2 is sample mean square \hat{S}^2

$$\hat{S}^2 = \frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2) = \frac{1}{9} (24.9952 - 10 \times 1.58^2) = 0.0035$$

$$\text{Therefore, } \hat{S} = \sqrt{\hat{S}^2} = 0.0588$$

Thus, the average lea weight is 1.58 units and the variance of lea weight is 0.0312 units.

Now,

the sample size $n = 10$ is small therefore the 95% confidence interval for the average lea weight is

$$\text{Therefore, } \left(\bar{x} - t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}, \bar{x} + t_{(n-1), \alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}} \right)$$

Further

$$95\% \text{ confidence} \Rightarrow \alpha = 0.05 \text{ Therefore, } t_{(n-1), \alpha/2} = t_{9, 0.025} = 2.262$$

Substituting values we get required 95% confidence interval for mean μ as (1.5379, 1.6222).

That is, with 95% guarantee the value of the population mean μ (average lea weight) lies in between 1.5379 to 1.6222 units.

Also,

the sample size $n = 10$ is small and population mean μ is unknown therefore the 95% confidence interval for the variance of lea weight is

$$\left(\frac{nS^2}{\chi^2_{n-1, \alpha/2}}, \frac{nS^2}{\chi^2_{n-1, \left(1-\frac{\alpha}{2}\right)}} \right)$$

$$95\% \text{ confidence} = \alpha = 0.05$$

$$\text{Therefore, } \chi^2_{9, 0.05} = 16.919 \text{ and } \chi^2_{9, 0.025} = 19.679$$

$$\Rightarrow \chi^2_{n-1, \alpha/2} = \chi^2_{9, 0.025} = 18.644$$

$$\text{Therefore, } \chi^2_{9, 0.95} = 3.325 \text{ and } \chi^2_{9, 0.98} = 2.532$$

$$\Rightarrow \chi^2_{n-1, \left(1-\frac{\alpha}{2}\right)} = \chi^2_{9, 0.0975} = 2.8294$$

$$S^2 = \frac{n-1}{n} \cdot \hat{S}^2 = .0031$$

Substituting values we get required 95% confidence interval for variance σ^2 as (0.0017, 0.0109).

Determination sample size for the given level of confidence and the level of accuracy

The formula or the determination of the sample such that, we can get $(1 - \alpha)$ 100% confidence as well as $\{\text{Mean} \pm k\% \text{ of mean}\}$ as the accuracy can be obtained as follows:

Case I

- (a) Equating the confidence interval for the population mean μ based on the large sample

$\left(\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$ and the interval (Mean $- k\%$ of mean, Mean $+ k\%$ of mean)

We get,

$$Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = k\% \text{ of mean} = \mu \times \frac{k}{100}$$

$$\Rightarrow \sqrt{n} = \frac{Z_{\alpha/2} \cdot \sigma}{\mu \times \frac{k}{100}} = \frac{Z_{\alpha/2} \cdot \frac{\sigma}{\mu} \times 100}{k} = \frac{Z_{\alpha/2} \times \text{CV}\%}{k}$$

$$\text{Therefore, } n = \frac{Z_{\alpha/2}^2 \times \text{CV}^2}{k^2}$$

Further if level of confidence is 95% that is $\alpha = 0.05 \Rightarrow Z_{\alpha/2} = 1.96$ then,

$$n = \frac{1.96^2 \times \text{CV}^2}{k^2}$$

- (b) Equating the confidence interval for the population mean μ based on the large sample

$\left(\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$ and the interval (Mean $- k$, Mean $+ k$)

we get,

$$\begin{aligned} Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= k \\ \Rightarrow \sqrt{n} &= \frac{Z_{\alpha/2} \cdot \sigma}{k} \\ \text{Therefore, } n &= \frac{Z_{\alpha/2}^2 \times \sigma^2}{k^2} \end{aligned}$$

Further if level of confidence is 95% that is $\alpha = 0.05 \Rightarrow Z_{\alpha/2} = 1.96$ then,

$$n = \frac{1.96^2 \times \sigma^2}{k^2}$$

Example 15.3

If the expected average linear density is 14.0 Tex and coefficient of variation of linear density is 1.2% Tex find the sample size or the number of tests to be made, so that confidence of 95% can be achieved with the accuracy of mean $\pm 0.5\%$ of mean.

Solution

Here confidence level is 95% and CV% = 1.2.

Also level of accuracy is mean $\pm 0.5\%$ of mean $\Rightarrow k = 0.5$

Therefore, Using the formula given as the above

$$n = \frac{1.96^2 \times CV^2}{k^2} = \frac{1.96^2 \times 1.2^2}{0.5^2} = 22.13$$

Thus for achieving 95% confidence and mean $\pm 0.5\%$ of mean accuracy no. of tests to be carried out are approximately 22.

Case II Equating the confidence interval for the population mean μ based on the small sample

$\left(\bar{x} - t_{(n-1), \alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}, \bar{x} + t_{(n-1), \alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}} \right)$ and the interval (Mean $- k\%$ of mean,

Mean $+ k\%$ of mean)

We get,

$$t_{(n-1), \alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}} = k\% \text{ of mean} = \mu \times \frac{k}{100}$$

Thus in this case number of tests can be determined by starting with $n = 2$ and continuing until LHS and RHS of the above equation are

nearly same. The value of “ n ,” for which LHS and RHS are approximately same is the required sample size.

For example if mean $\mu = 20$, standard deviation estimate $\hat{s} = 0.75$ confidence level is 95% and required accuracy is mean $\pm 0.5\%$ of mean, then preparing Table 15.1

Table 15.1

n	$t_{(n-1), \alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}$
2	6.738483
3	1.863103
4	1.193417
5	0.931248
6	0.787077
7	0.693634
8	0.627016

Where mean $\pm 0.5\%$ of mean \Rightarrow

$$\mu \times \frac{k}{100} = 20 \times \frac{0.5}{100} = 0.1$$

Continuing this way we find that for $n = 130$, $t_{(n-1), \alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}$ and $\mu \times \frac{k}{100}$ are same. Hence, the required sample size is 130.

15.4 Exercise

1. What is estimation? What are its types? Describe any one.
2. Define unbiased estimator. Prove that the sample mean is an unbiased estimator of the population mean.
3. Show that $\hat{S}^2 = \frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2)$ is an unbiased estimator of population variance σ^2 .
4. What is confidence interval? Derive $100(1 - \alpha)\%$ confidence interval for the population mean based on large sample.
5. What is the confidence interval? Derive $100(1 - \alpha)\%$ confidence interval for population mean based on small sample.
6. What is the confidence interval? Find the 95% confidence interval for the population mean, if mean and standard deviation of the sample of size 40 are 78.25 and 3.2 respectively. Also comment on it.

7. What is confidence interval? State the confidence interval for the population mean based on the large sample. Find 99% confidence interval using the sample results of 40 leas having average and standard deviation of the mass as 75.6 and 4.2 grams respectively and comment on your answer.
8. State the $100(1 - \alpha)\%$ confidence interval based on small sample for the population mean. Also, find 95% confidence interval, if mean and standard deviation of a sample of size 10 are 50 and 2.5 respectively and comment on your answer.
9. Show that sample mean \bar{X} is unbiased estimator of population mean μ .

16.1 Introduction

Professor R.A. Fisher first invented the technique “analysis of variance” (ANOVA) in 1920s, when he was dealing with the agricultural data. This is a powerful tool for testing the hypothesis, $H_0: \mu_1 = \mu_2 = \dots = \mu_k$. That is, it is a powerful tool for comparing means of three or more populations with each other. These means of three or more populations are compared using the variation present in the observations. Hence, it is called the analysis of variance technique. The variation present in the observations occurs due to number of causes, which are classified as assignable causes and chance causes.

Assignable causes

The causes of variation, which can be identified or detected, are called the assignable causes. The human beings can control these causes. The variation in the observations, occurring due to these causes is called the assignable cause variation.

Chance causes

The causes of variation, which cannot be identified or detected, are called the chance causes. The human beings cannot control these causes hence they cannot be totally eliminated. But by properly designing the experiment (discussed in next chapter) one can try to minimize them. The variation in the observations, occurring due to these is called the error variation. Smaller the error variation better is the experiment.

Thus, in analysis of variance technique, the total variation present in the observations is divided into number of components, where each of the components represents the variation due to some assignable causes and chance causes. After division, the variations of assignable causes are compared with the variation due to chance causes and decision regarding the acceptance or the rejection of hypothesis is made. One-way analysis of variance and two-way analysis of variance are the two different types of the analysis of variance technique.

16.2 One-way analysis of variance

When there is only one assignable cause or factor, apart from the chance cause, which is responsible for the variation in the observations of the data, then it is called the one-way analysis of variance. Suppose this only one factor is denoted by A and suppose that it has “ k ” different levels denoted by A_1, A_2, \dots, A_k . Further suppose that, $\mu_1, \mu_2, \dots, \mu_k$ denote mean effects of the levels A_1, A_2, \dots, A_k . Thus, in case of one-way analysis of variance interest is to compare the mean effects of the levels A_1, A_2, \dots, A_k with each other. That is to test the hypothesis,

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \quad \text{Vs} \quad H_1: \mu_i \neq \mu_j \text{ for at least one pair } (i, j).$$

For testing this hypothesis, suppose total “ n ” observations are collected such that for each level A_i there are n_i observations and $\sum n_i = n$ where, $i = 1, 2, \dots, k$.

These n observations are shown in Table 16.1.

Table 16.1

A_1	A_2	A_k
x_{11}	x_{21}	x_{k1}
x_{21}	x_{22}	x_{k2}
...
...
x_{1n_1}	x_{2n_2}	x_{nk}

Where, x_{ij} is the j^{th} observation corresponding to the i^{th} level of factor A

$$i = 1, 2, \dots, k \text{ and } j = 1, 2, \dots, n_i.$$

Mathematical analysis in case of one-way ANOVA

Suppose the mathematical model for the observation or the response x_{ij} is as follows:

$$x_{ij} = \mu_i + \epsilon_{ij}$$

Where,

μ_i = Mean effect of the level A_i of factor A

ϵ_{ij} = Random error component.

$\epsilon_{ij} \sim N(0, \sigma^2)$ and are all independent for all (i, j) .

Further,

Suppose that the above model is rewritten in the following form,

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Where,

μ = General mean effect or overall average effect.

α_i = fixed effect of the level A_i of the factor A .

Thus, from this new model, testing the above hypothesis is equivalent to test the following hypothesis.

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k \quad Vs \quad H_1: \alpha_i \neq 0 \text{ for at least one } i.$$

For testing this hypothesis the total variation in the observations is divided into two components as follows.

Total sum of squares = Sum of squares due to factor A + Sum of squares due to error

$$TSS = SSA + SSE$$

Here, with the above assumptions from the Cochran's theorem, it is expected that,

$$SSA \sim \sigma^2 \chi^2_{k-1} \text{ i.e. } \frac{SSA}{\sigma^2} \sim \chi^2_{k-1}$$

$$SSE \sim \sigma^2 \chi^2_{n-k} \text{ i.e. } \frac{SSE}{\sigma^2} \sim \chi^2_{n-k}$$

and both components are independent of each other.

Hence, from the property of the F probability distribution,

$$F = \frac{\frac{SSA}{\sigma^2} / k - 1}{\frac{SSE}{\sigma^2} / n - k} = \frac{MSSA}{MSSE} = \frac{\text{Mean sum of squares due to factor } A}{\text{Mean sum of squares due to error}}$$

All the above results of mathematical analysis are written in a tabular form called the analysis of variance table (ANOVA table) is shown in Table 16.2.

Table 16.2

Source	<i>df</i>	<i>SS</i>	MSS	<i>F</i> ratio
Factor A	$k - 1$	SSA	$\text{MSSA} = \text{SSA}/k - 1$	$F = \text{MSSA}/\text{MSSE}$
Error	$n - k$	SSE	$\text{MSSE} = \text{SSE}/n - k$	
	$n - 1$	TSS		

Here, it is expected that the statistic

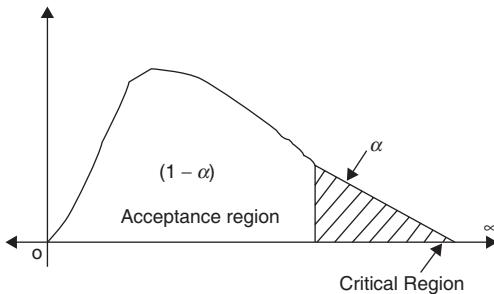
$$F \sim F_{(k-1), (n-k)}$$

Thus,

the null hypothesis H_0 is rejected at $100\alpha\%$ los against H_1 , if $F_{\text{cal}} > F_{(k-1), (n-k), \alpha}$

Where, the value $F_{(k-1), (n-k), \alpha}$ is such that, area above this value is under F distribution curve is α .

Hence, the critical region can be shown in Fig. 16.1.



16.1

Here, if H_0 is accepted, then it implies that $\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ that is the fixed effects of different levels of factor A are not significant. Thus $\mu_1 = \mu_2 = \dots = \mu_k = 0$ that is the mean effects of different levels of A are same.

Hence on an average the values of the characteristic X are same for all levels of A .

Remarks

1. Analysis of variance is not affected by change of origin that is if a constant value is subtracted from all the observations, then the analysis of variance answers are not affected as variance is not affected by change of origin.
2. Sometimes by calculations if MSSA is less than MSSE, then also the F calculated should be calculated as usual as $F = \text{MSSA}/\text{MSSE}$.

Procedure and formulae for computation

1. Identify factor A and its levels “ k ”
2. Find totals for each level of factor A and denote them by A_1, A_2, \dots, A_k
3. Find Grand Total $G = \sum \sum x_{ij} = \sum A_i$
4. Find $CF = \frac{G^2}{n}$
5. Find Raw SS = $\sum \sum x_{ij}^2$ that is find sum of squares of all observations.
6. Find TSS = Raw SS – CF.
7. Find SSA = $\sum \frac{A_i^2}{n_i} - CF$
8. Find SSE = TSS – SSA
9. Prepare ANOVA table and write conclusions.

Example 16.1

An experiment was carried out to study the effect of the speed of the ring frame on the number of end breakages and following results were obtained by taking five readings at each speed.

		Speed (rpm)			
		15000	160000	17000	18000
End breakages	18	22	23	30	
	20	20	23	32	
	18	21	25	28	
	15	18	24	28	
	17	23	23	35	

Carry out analysis of the above data and write the conclusion.

Solution

Here, factor “ A ” is the speed of the ring frame and its levels are,

$$A_1 \sim 15000, A_2 \sim 16000, A_3 \sim 17000 \text{ and } A_4 \sim 18000$$

The variable under study is

$$X \sim \text{Number of end breakages.}$$

Suppose, $\mu_1, \mu_2, \dots, \mu_4$ denote mean effects of the levels A_1, A_2, \dots, A_4 .
Also $\alpha_1, \alpha_2, \dots, \alpha_4$ denote fixed effects of the levels A_1, A_2, \dots, A_4 .

Thus, in case interest is to test the hypothesis,

$$\begin{array}{ll} H_0: \mu_1 = \mu_2 = \dots = \mu_4 & \text{Vs} \quad H_1: \mu_i \neq \mu_j \text{ for at least one pair } (i, j). \\ & \text{or} \\ H_0: \alpha_1 = \alpha_2 = \dots = \alpha_4 & \text{Vs} \quad H_1: \alpha_i \neq 0 \text{ for at least one } i. \end{array}$$

For testing the above hypothesis, the analysis calculations are carried out as follows:

	15000	16000	17000	18000
18	22	23	30	
20	20	23	32	
18	21	25	28	
15	18	24	28	
17	23	23	35	
$A_1 = 88$	$A_2 = 104$	$A_3 = 118$	$A_4 = 153$	

$$\text{Grand Total} = \sum \sum x_{ij} = \sum A_i = 88 + 104 + 118 + 153 = 463.$$

$$CF = \frac{G^2}{n} = \frac{463^2}{20} = 10718.45$$

$$\text{Raw SS} = \sum \sum x_{ij}^2 = 11245$$

$$\text{TSS} = \text{Raw SS} - CF = 11245 - 10718.45 = 526.55$$

$$\text{SSA} = \sum \frac{A_i^2}{n_i} - CF = 11178.6 - 10718.45 = 460.15$$

$$\text{SSE} = \text{TSS} - \text{SSA} = 526.55 - 460.15 = 66.4$$

Thus, the ANOVA table is shown in Table 16.3.

Table 16.3

Source	df	SS	MSS	F ratio
Factor A (Speed)	3	460.15	153.38	36.96
Error	16	66.4	4.15	
	19	526.55		

Here, it is expected that the statistic "F" will follow F distribution with $(k - 1) = 3$ and $(n - k) = 16$ df. Thus, at 5% los that is for $\alpha = 0.05$

$$F_{(k-1), (n-k), \alpha} = F_{(3), (16), 0.05} = 3.24$$

$$\text{As, } F_{\text{cal}} = 36.96 > F_{(3), (16), 0.05} \text{, } 3.24$$

Reject H_0 at 5% los. $\Rightarrow \alpha_i \neq 0$ for at least one "i" \Rightarrow fixed effect of at least one speed is significant: $\mu_i \neq \mu_j$ for at least one pair of speeds \Rightarrow the mean effects of the at least one pair of speeds are not same.

Hence, Average number of end breaks is not same for different speeds.

Example 16.2

An experiment was conducted to study the effect of a dye produced by four different companies (A, B, C, and D) on the strength of the fabric and following results were obtained.

		Dye			
		A	B	C	D
Fabric strength	250	225	250	300	
	275	250	275	250	
	275	225	250	275	
	300	300	250	300	

Carry out the analysis of the above data and write the conclusion.

Solution

Here, factor “A” is the Dye and its levels are,

$$A_1 \Rightarrow A, A_2 \Rightarrow B, A_3 \Rightarrow C \text{ and } A_4 \Rightarrow D$$

The variable under study is

$X \Rightarrow$ Strength (in gms.) of the fabric.

Suppose, $\mu_1, \mu_2, \dots, \mu_4$ denote mean effects of the levels A_1, A_2, \dots, A_4 .

Also $\alpha_1, \alpha_2, \dots, \alpha_4$ denote fixed effects of the levels A_1, A_2, \dots, A_4 .

Thus, in case interest is to test the hypothesis,

$$H_0: \mu_1 = \mu_2 = \dots = \mu_4 \quad \text{Vs} \quad H_1: \mu_i \neq \mu_j \text{ for at least one pair } (i, j).$$

or

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_4 \quad \text{Vs} \quad H_1: \alpha_i \neq 0 \text{ for at least one } i.$$

For testing the above hypothesis, the analysis is as follows:

As ANOVA is not affected by change of origin and scale hence we use change of origin and scale and change all values using relation $(X - 250)/25$

		Dye			
	A	B	C	D	
0	0	-1	0	2	
1	1	0	1	0	
1	1	-1	0	1	
2	2	2	0	2	
	$A_1 = 4$	$A_2 = 0$	$A_3 = 1$	$A_4 = 5$	

$$\text{Grand Total} = \sum \sum x_{ij} = \sum A_i = 4 + 0 + 1 + 5 = 10.$$

$$CF = \frac{G^2}{n} = \frac{10^2}{16} = 6.25$$

$$\text{Raw } SS = \sum \sum x_{ij}^2 = 22$$

$$\text{TSS} = \text{Raw } SS - CF = 22 - 6.25 = 15.75$$

$$\text{SSA} = \sum \frac{A_i^2}{n_i} - CF = 10.5 - 6.25 = 4.25$$

$$\text{SSE} = \text{TSS} - \text{SSA} = 15.75 - 4.25 = 11.5$$

Thus, the ANOVA table is shown in Table 16.4.

Table 16.4

Source	df	SS	MSS	F ratio
Factor A (Dye)	3	4.25	1.4167	1.4783
Error	12	11.50	0.9583	
	15	15.75		

Here, it is expected that the statistic “F” will follow F distribution with $(k - 1) = 3$ and $(n - k) = 12$ df. Thus, at 5% los that is for $\alpha = 0.05$

$$F_{(k-1), (n-k), \alpha} = F_{(3), (12), 0.05} = 3.49$$

$$\text{As, } F_{\text{cal}} = 1.3734 < F_{(3), (12), 0.05} = 3.49$$

Accept H_0 at 5% los. $\Rightarrow \alpha_i = 0$ for each “ i ” \Rightarrow fixed effect of every dye is not significant.

$\therefore \mu_1 = \mu_2 = \dots = \mu_4$ the mean effects of all dyes are same.

On an average strength of fabric is same for all dyes.

16.3 Two-way analysis of variance

Two-way ANOVA with one observation per cell (without repetition) and two-way ANOVA with more than one observation per cell (with repetition) are the two types of two-way ANOVA.

Two-way ANOVA with one observation per cell (without repetition)

When two different assignable causes or factors, apart from the chance cause, are responsible for the variation in the observations, then it is called the two-way analysis of variance. Suppose, these two different factors are denoted by A and B. Suppose that factor A has “ p ” different levels denoted by A_1, A_2, \dots, A_p and factor B has “ q ” different levels denoted by B_1, B_2, \dots, B_q .

Further suppose that, $\mu_1, \mu_2, \dots, \mu_p$ denote mean effects of the levels A_1, A_2, \dots, A_p and $\mu_{.1}, \mu_{.2}, \dots, \mu_{.q}$ denote mean effects of the levels B_1, B_2, \dots, B_q .

Thus, in case of two-way analysis of variance interest is to compare the mean effects of the level A_1, A_2, \dots, A_p with each other as well as to compare mean effects of the levels B_1, B_2, \dots, B_q with each other.

That is to test the hypotheses,

$$H_0^A: \mu_1 = \mu_2 = \dots = \mu_p \quad \text{Vs} \quad H_1^A: \mu_i \neq \mu_j \text{ for at least one pair } (i, j).$$

$$H_0^B: \mu_{.1} = \mu_{.2} = \dots = \mu_{.q} \quad \text{Vs} \quad H_1^B: \mu_{.i} \neq \mu_{.j} \text{ for at least one pair } (i, j).$$

For testing these hypotheses, suppose one observation is collected corresponding to each pair of levels A_i and B_j , thus total “ pq ” observations are collected such that for each level A_p , there are q observations and for each level B_j there are p observations. These “ pq ” observations are shown in Table 16.5.

Table 16.5

Factor		B				
		B_1	B_2	B_q
A	A_1	x_{11}	x_{21}	x_{q1}
	A_2	x_{12}	x_{22}	x_{q2}

	A_p	x_{p1}	x_{p2}	x_{pq}

where,

x_{ij} = Observation corresponding to the i^{th} level of factor A and j^{th} level of factor B

$$i = 1, 2, \dots, p \quad \text{and} \quad j = 1, 2, \dots, q$$

Mathematical analysis in case of two-way ANOVA (without repetition)

Suppose the mathematical model for x_{ij} is as follows:

$$x_{ij} = \mu_{ij} + \epsilon_{ij}$$

Where, μ_{ij} is the combined mean effect of i^{th} level of factor A and j^{th} level of factor B.

ϵ_{ij} is the random error component, $\epsilon_{ij} \sim N(0, \sigma^2)$ and all ϵ_{ij} are independent. The above model is rewritten as follows:

$$x_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$$

Where, μ is general mean effect.

α_i is the fixed effect of i^{th} level of factor A.

β_j is the fixed effect of j^{th} level of factor B.

γ_{ij} is the fixed interaction effect of i^{th} level of factor A and j^{th} level of factor B.

In this case the parameter γ_{ij} cannot be studied or estimated as there is only one observation per cell (I, j) . Hence, it is assumed that $\gamma_{ij} = 0$

Thus, in this case interest is to test following hypotheses,

$$H_0^A: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0 \quad \text{Vs} \quad H_1^A: \alpha_i \neq 0 \text{ for at least one } i$$

$$H_0^B: \beta_1 = \beta_2 = \dots = \beta_n = 0 \quad \text{Vs} \quad H_1^B: \beta_j \neq 0 \text{ for at least one } j$$

For testing the above hypotheses in this case also total variation is divided into three components as follows:

$$\text{TSS} = \text{SSA} + \text{SSB} + \text{SSE}$$

Thus from the Cochran's theorem, it is expected that

$$\text{SSA} \sim \sigma^2 \chi_{p-1}^2 \text{ i.e. } \frac{\text{SSA}}{\sigma^2} \sim \chi_{p-1}^2$$

$$\text{SSB} \sim \sigma^2 \chi_{q-1}^2 \text{ i.e. } \frac{\text{SSB}}{\sigma^2} \sim \chi_{q-1}^2$$

$$\text{SSE} \sim \sigma^2 \chi^2 \text{ i.e. } \frac{\text{SSE}}{\sigma^2} \sim \chi^2$$

and all three components are independent of each other.

Hence, from the property of the F probability distribution,

$$F_A = \frac{\frac{\text{SSA}}{\sigma^2} / (p-1)}{\frac{\text{SSE}}{\sigma^2} / ((p-1)(q-1))} = \frac{\text{SSA}/(p-1)}{\text{SSE}/((p-1)(q-1))} = \frac{\text{MSSA}}{\text{MSSE}}$$

Similarly,

$$F_B = \frac{\text{MSSB}}{\text{MSSE}}$$

Where,

MSSA represents mean sum of squares due to factor A, MSSB represents mean sum of squares due to factor B and MSSE represents mean sum of squares due to error.

All the above results of mathematical analysis are written in the form of analysis of variance table (ANOVA table) is shown in Table 16.6.

Table 16.6

Source	df	SS	MSS	F ratio
Factor A	$(p - 1)$	SSA	MSSA	$F_A = \frac{\text{MSSA}}{\text{MSSE}}$
Factor B	$(q - 1)$	SSB	MSSB	$F_B = \frac{\text{MSSB}}{\text{MSSE}}$
Error	$(p - 1)(q - 1)$	SSE	MSSE	
Total	$(pq - 1)$	TSS		

Here it is expected that, the statistics F_A will follow “F” probability distribution with $(p - 1)$ and $(p - 1) . (q - 1)$ degrees of freedom and F_B will follow “F” probability distribution with $(q - 1)$ and $(p - 1) . (q - 1)$ degrees of freedom.

Criteria for Rejection of H_0

1. Reject null hypothesis H_0^A at $100\alpha\%$ los, if

$$F_{\text{cal}}^A > F_{(p-1), (p-1). (q-1), \alpha}$$

2. Reject null hypothesis H_0^B at $100\alpha\%$ los, if

$$F_{\text{cal}}^B > F_{(q-1), (p-1). (q-1), \alpha}$$

Procedure and formulae for computation

1. Identify factor A, and factor B using the given data. Also decide their levels (p and q).
2. Find totals of observations for each level of factor A and each level of factor B and denote them by A_1, A_2, \dots, A_p and B_1, B_2, \dots, B_q .
3. Find Grand total $G = \sum \sum x_{ij} = \sum A_i = \sum B_j$
4. Find CF using $CF = \frac{G^2}{pq}$
5. Find Raw sum of squares using Raw S.S = $\sum \sum x_{ij}^2$ = that is find sum of squares of each and every observation.
6. Find TSS using $TSS = \text{Raw SS} - CF$
7. Find SSA using $SSA = \sum \frac{A_i^2}{qm} - CF$

8. Find SSB using $SSB = \sum \frac{B_j^2}{pm} - CF$
9. Find SSE using $SSE = TSS - SSA - SSB$
10. Prepare analysis of variance table and write conclusions by comparing calculated F values with corresponding table values of F .

Example 16.3

An experiment was conducted to study the effect of the speed of the ring frame on the count of the yarn. The yarn was spun with four different speeds on three different ring frames and the results yarn counts are as follows:

		Speed (rpm)			
		15000	16000	17000	18000
R/F	I	85	88	85	90
	II	70	85	90	95
	III	80	82	88	92

Carry out analysis of the above data and write the conclusions.

Solution

Here, Factor A: Ring frame and with three levels $A_1 \Rightarrow R/F-I, A_2 \Rightarrow R/F-II, A_3 \Rightarrow R/F-III$

Factor B: Speed of the ring frame with four levels $B_1 \Rightarrow 15000, B_2 \Rightarrow 16000, B_3 \Rightarrow 17000, B_4 \Rightarrow 18000$.

Thus, interest is to test,

$$\begin{array}{lll} H_0^A: \mu_{1.} = \mu_{2.} = \dots = \mu_{3.} & \text{Vs} & H_1^A: \mu_i \neq \mu_j \text{ for at least one pair } (i, j). \\ H_0^B: \mu_{1.} = \mu_{2.} = \dots = \mu_{4.} & \text{Vs} & H_1^B: \mu_i \neq \mu_j \text{ for at least one pair } (i, j). \end{array}$$

That is interest is to test,

$$\begin{array}{lll} H_0^A: \alpha_1 = \alpha_2 = \dots = \alpha_3 = 0 & \text{Vs} & H_1^A: \alpha_i \neq 0 \text{ for at least one } i \\ H_0^B: \beta_1 = \beta_2 = \dots = \beta_4 = 0 & \text{Vs} & H_1^B: \beta_j \neq 0 \text{ for at least one } j \end{array}$$

For testing these hypotheses, the analysis is as follows:

$$\begin{aligned} A_1 &= 85 + 88 + 85 + 90 = 348, A_2 = 70 + 82 + 90 + 95 = 340, A_3 = 80 + 82 + 88 + 92 = 342 \\ B_1 &= 85 + 70 + 90 = 235, B_2 = 88 + 85 + 82 = 255, B_3 = 85 + 90 + 88 = 263, \\ B_4 &= 90 + 95 + 92 = 277 \end{aligned}$$

Thus, Grand total = $G = \sum A_i = \sum B_j = 1030$

$$CF = \frac{G^2}{pq} = \frac{1030^2}{12} = 88408.3333$$

$$\text{Raw SS} = \sum \sum x_{ij}^2 = 88876$$

$$\text{TSS} = \text{Raw SS} - CF = 88876 - 88408.3333 = 467.6667$$

$$\text{SSA} = \sum \frac{A_i^2}{q} - CF = \frac{353668}{4} - 88404.3333 = 8.6667$$

$$\text{SSB} = \sum \frac{B_j^2}{p} - CF = \frac{266148}{3} - 88408.3333 = 307.6667$$

$$\text{SSE} = \text{TSS} - (\text{SSA} + \text{SSB}) = 467.6667 - (8.6667 + 307.6667) = 151.3333$$

Thus, ANOVA table is shown in Table 16.7.

Table 16.7

Source	df	SS	MSS	F ratio
Count	2	8.6667	4.3334	$F_A = 0.1718$
Speed	3	307.6667	102.5556	$F_B = 4.06608$
Error	6	151.3333	25.2222	
Total	11	467.6667		

Here it is expected that the statistic F_A and F_B will follow F probability distribution as follows:

$$F_A \sim F_{(p-1),(p-1)(q-1)} \text{ that is, } F_A \sim F_{2,6} \text{ and } F_B \sim F_{(q-1),(p-1)(q-1)} \text{ that is, } F_B \sim F_{3,6}$$

Thus, at 5% los \Rightarrow for $\alpha = 0.05$

$$F_{(p-1),(p-1)(q-1)} = F_{2,6,0.05} = 5.14 \text{ and } F_{(q-1),(p-1)(q-1)} = F_{3,6,0.05} = 4.76$$

Here, $F_{\text{Acal}} = 0.1718 < F_{2,6,0.05} = 5.14$

Thus, we accept $H_0^A \Rightarrow$ fixed effects of different Ring frames are not significant \Rightarrow mean effects of the ring frames are same and hence average yarn count is same for different ring frames.

Similarly, $F_{\text{Bcal}} = 4.0661 < F_{3,6,0.05} = 4.76$

Thus, we accept $H_0^B \Rightarrow$ fixed effect of different speeds are not significant \Rightarrow mean effects of the speeds are same and hence average yarn count is same for different speeds.

Two-way ANOVA with "m" observation per cell (with repetition)

We have seen that in case of two-way analysis of variance interest is to compare the mean effects of the level A_1, A_2, \dots, A_p with each other as well as to compare mean effects of the levels B_1, B_2, \dots, B_q with each other. That is to test the hypotheses,

$$\begin{array}{ll} H_0^A: \mu_{1.} = \mu_{2.} = \dots = \mu_{p.} & \text{Vs} \\ H_1^A: \mu_{i.} \neq \mu_{j.} \text{ for at least one pair } (i, j). & \\ H_0^A: \mu_{.1} = \mu_{.2} = \dots = \mu_{.q} & \text{Vs} \\ H_1^A: \mu_{.i} \neq \mu_{.j} \text{ for at least one pair } (i, j). & \end{array}$$

For testing these hypotheses, suppose "m" observations are collected corresponding to each pair of levels A_i and B_j , thus total " pqm " observations are collected such that for each level A_i , there are qm observations and for each level B_j there are pm observations. These " pqm " observations are shown in Table 16.8

Table 16.8

Factor		B					
		B_1		B_1		B_q	
A	A_1	$x_{111}, x_{112}, \dots, x_{11m}$	$x_{121}, x_{122}, \dots, x_{12m}$	$x_{1q1}, x_{1q2}, \dots, x_{1qm}$	
	A_2	$x_{211}, x_{212}, \dots, x_{21m}$		$x_{2q1}, x_{2q2}, \dots, x_{2qm}$	

	A_p	$x_{p11}, x_{p12}, \dots, x_{p1m}$				$x_{pq1}, x_{pq2}, \dots, x_{pqm}$	

Where,

x_{ijk} = k^{th} observation corresponding to the i^{th} level of factor A and j^{th} level of factor B.

$$i = 1, 2, \dots, p; j = 1, 2, \dots, q \text{ and } k = 1, 2, \dots, m$$

Mathematical analysis in case of two-way ANOVA (with repetition)

Suppose the mathematical model for x_{ijk} is as follows:

$$x_{ijk} = \mu_{ijk} + \epsilon_{ijk}$$

Where, μ_{ijk} is the combined mean effect of i^{th} level of factor A and j^{th} level of factor B.

ϵ_{ijk} is the random error component, $\epsilon_{ijk} \sim N(0, \sigma^2)$ and all ϵ_{ij} are independent. The above model is rewritten as follows:

$$x_{ijk} = \mu_i + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where, μ is general mean effect.

α_i is the fixed effect of i^{th} level of factor A.

β_j is the fixed effect of j^{th} level of factor B.

γ_{ij} is the fixed interaction effect of i^{th} level of factor A and j^{th} level of factor B.

In this case the parameter γ_{ij} can also be studied or estimated.

Thus in this case interest is to test following hypotheses,

$$\begin{array}{lll} H_0^A: \alpha_1 = \alpha_2 = \dots = \alpha_p = 0 & \text{Vs} & H_1^A: \alpha_i \neq 0 \text{ for at least one } i \\ H_0^B: \beta_1 = \beta_2 = \dots = \beta_q = 0 & \text{Vs} & H_1^B: \beta_j \neq 0 \text{ for at least one } j \\ H_0^{AB}: \gamma_{11} = \gamma_{12} = \dots = \gamma_{pq} = 0 & \text{Vs} & H_1^{AB}: \gamma_{ij} \neq 0 \text{ for at least } (i, j) \end{array}$$

For testing the above hypotheses in this case also total variation is divided into four components as follows:

$$\text{TSS} = \text{SSA} + \text{SSB} + \text{SSE}$$

Thus from the Cochran's theorem, it is expected that,

$$\text{SSA} \sim \sigma^2 \chi_{p-1}^2 \text{ i.e. } \frac{\text{SSA}}{\sigma^2} \sim \chi_{p-1}^2$$

$$\text{SSB} \sim \sigma^2 \chi_{q-1}^2 \text{ i.e. } \frac{\text{SSB}}{\sigma^2} \sim \chi_{q-1}^2$$

$$\text{SSAB} \sim \chi_{(p-1)(q-1)}^2 \text{ i.e. } \frac{\text{SSAB}}{\sigma^2} \sim \chi_{(p-1)(q-1)}^2$$

$$\text{SSE} \sim \chi_{(pq)(m-1)}^2 \text{ i.e. } \frac{\text{SSE}}{\sigma^2} \sim \chi_{(pq)(m-1)}^2$$

and all four components are independent of each other.

Hence, from the property of the F probability distribution,

$$F_A = \frac{\frac{\text{SSA}}{\sigma^2} / p - 1}{\frac{\text{SSE}}{\sigma^2} / ((p-1)(q-1))} = \frac{\text{SSA}/p - 1}{\text{SSE}/((p-1)(q-1))} = \frac{\text{MSSA}}{\text{MSSE}}$$

Similarly,

$$F_B = \frac{\text{MSSB}}{\text{MSSE}} \quad \text{and} \quad F_{AB} = \frac{\text{MSSAB}}{\text{MSSE}}$$

Where,

MSSA represents mean sum of squares due to factor A, MSSB represents mean sum of squares due to factor B, MSSAB represents mean sum of squares due to factor A and B and MSSE represents mean sum of squares due to error.

All the above results of mathematical analysis are written in the form of analysis of variance table (ANOVA table) is shown in Table 16.9.

Table 16.9

Source	df	SS	MSS	F ratio
Factor A	$(p - 1)$	SSA	MSSA	$F_A = \frac{\text{MSSA}}{\text{MSSE}}$
Factor B	$(q - 1)$	SSB	MSSB	$F_B = \frac{\text{MSSB}}{\text{MSSE}}$
Interaction AB	$(p - 1)(q - 1)$	SSAB	MSSAB	$F_{AB} = \frac{\text{MSSAB}}{\text{MSSE}}$
Error	$pq(m - 1)$	SSE	MSSE	
Total	$(pqm - 1)$	TSS		

Here it is expected that, the statistics F_A will follow “F” probability distribution with $(p - 1)$ and $pq.(m - 1)$ degrees of freedom, F_B will follow “F” probability distribution with $(q - 1)$ and $pq.(m - 1)$ degrees of freedom. and F_{AB} will follow “F” probability distribution with $(p - 1).(q - 1)$ and $pq.(m - 1)$ degrees of freedom.

Criteria for Rejection of H_0

1. Reject null hypothesis H_0^A at $100\alpha\%$ los, if

$$F_{Cal}^A > F_{(p-1), pq(m-1), \alpha}$$

2. Reject null hypothesis H_0^B at $100\alpha\%$ los, if

$$F_{Cal}^B > F_{(p-1), pq(m-1), \alpha}$$

3. Reject null hypothesis H_0^{AB} at $100\alpha\%$ los, if

$$F_{Cal}^{AB} > F_{(p-1).(q-1), pq(m-1), \alpha}$$

Procedure and formulae for computation

1. Identify factor A, and factor B using the given data. Also decide their levels (p and q) and number of observations per cell (m).
2. Find totals of observations for each level of factor A and each level of factor B and denote them by A_1, A_2, \dots, A_p and B_1, B_2, \dots, B_q . Also find totals of observations for each cell and denote them by $T_{11}, T_{12}, \dots, T_{pq}$.
3. Find Grand total $G = \sum \sum \sum x_{ijk} = \sum A_i = \sum B_j = \sum \sum T_{ij}$
4. Find CF using $CF = \frac{G^2}{pqm}$
5. Find Raw sum of squares using Raw SS = $\sum \sum \sum x_{ijk}^2$ that is find sum of squares of each and every observation.
6. Find TSS using $TSS = \text{Raw SS} - CF$
7. Find SSA using $SSA = \sum \frac{A_i^2}{qm} - CF$
8. Find SSB using $SSB = \sum \frac{B_j^2}{pm} - CF$
9. Find SSAB using $SSAB = \left(\frac{\sum \sum T_{ij}^2}{m} - CF \right) - SSA - SSB$
10. Find SSE using $SSE = TSS - SSA - SSB - SSAB$
11. Prepare analysis of variance table and write conclusions by comparing calculated F values with corresponding table values of F .

Example 16.4

An experiment was conducted to study the effect of the speed of the ring frame on the count of the yarn. The yarn was spun with four different speeds on three different ring frame and the results yarn counts are as follows:

		Speed (rpm)			
		15000	16000	17000	18000
R/F	I	85,80,82	88,85,86	85,82,84	80,82,84
	II	79,82,80	85,80,78	80,86,83	85,88,86
	III	80,81,80	82,85,84	88,85,86	84,85,86

Carry out analysis of the above data and write the conclusions.

Solution

Here, Factor A: Ring frame and with three levels $A_1 \Rightarrow R/F - I$, $A_2 \Rightarrow R/F-II$, $A_3 \Rightarrow R/F-III$

Factor B: Speed of the ring frame with four levels $B_1 \Rightarrow 15000$, $B_2 \Rightarrow 16000$, $B_3 \Rightarrow 17000$, $B_4 \Rightarrow 18000$.

Thus interest is to test,

$$\begin{array}{lll} H_0^A: \alpha_1 = \alpha_2 = \dots = \alpha_3 = 0 & \text{Vs} & H_1^A: \alpha_i \neq 0 \text{ for at least one } i \\ H_0^B: \beta_1 = \beta_2 = \dots = \beta_4 = 0 & \text{Vs} & H_1^B: \beta_j \neq 0 \text{ for at least one } j \\ H_0^{AB}: \gamma_{11} = \gamma_{12} = \dots = \gamma_{34} = 0 & \text{Vs} & H_1^{AB}: \gamma_{ij} \neq 0 \text{ for at least one } (i, j) \end{array}$$

For testing these hypotheses, the analysis is as follows:

$$A_1 = 1003, A_2 = 992, A_3 = 1006$$

$$B_1 = 729, B_2 = 753, B_3 = 759, B_4 = 760.$$

$$\begin{aligned} T_{11} &= 247, T_{12} = 259, T_{13} = 251, T_{14} = 246, T_{21} = 241, T_{22} = 243, T_{23} = 249, T_{24} = 259, \\ T_{31} &= 241, T_{32} = 251, T_{33} = 259, T_{34} = 255 \end{aligned}$$

Thus, Grand total = $G = \sum A_i = \sum B_j = \sum \sum T_{ij} = 3001$

$$CF = \frac{G^2}{pqm} = \frac{3001^2}{36} = 250166.7$$

$$\text{Raw SS} = \sum \sum x_{ij}^2 = 250431$$

$$\text{TSS} = \text{Raw SS} - CF = 250431 - 250166.7 = 264.3056$$

$$\text{SSA} = \sum \frac{A_i^2}{qm} - CF = 9.0556$$

$$\text{SSB} = \sum \frac{B_j^2}{pm} - CF = 70.0833$$

$$\text{SSAB} = \left(\frac{\sum \sum T_{ij}^2}{m} - CF \right) - \text{SSA} - \text{SSB} = 89.8333$$

$$\text{SSE} = \text{TSS} - (\text{SSA} + \text{SSB} + \text{SSAB}) = 95.3333$$

Thus, ANOVA table is shown in Table 16.10.

Table 16.10

Source	df	SS	MSS	F ratio
Count	2	9.0556	4.5278	$F_A = 1.1399$
Speed	3	70.0833	23.3611	$F_B = 5.8811$
Interaction	6	89.8333	14.9722	$F_{AB} = 3.7692$
Error	24	95.3333	3.9722	
Total	35	264.3056		

Here it is expected that the statistic F_A , F_B and F_{AB} will follow F probability distribution as follows:

$F_A \sim F_{(p-1), pq(m-1)}$ that is, $F_A \sim F_{2,24}$; $F_B \sim F_{(q-1), pq(m-1)}$ that is, $F_B \sim F_{3,24}$ and
 $F_{AB} \sim F_{(p-1)(q-1), pq(m-1)}$ that is, $F_{AB} \sim F_{6,24}$

Thus, at 5% los \Rightarrow for $\alpha = 0.05$

$$F_{(p-1), pq(m-1)} = F_{2,24,0.05} = 3.40; F_{(p-1), pq(m-1)} = F_{3,24,0.05} = 3.01 \text{ and}$$

$$F_{(p-1)(q-1), pq(m-1), \infty} = F_{6,24,0.05} = 2.51$$

Here,

$$F_{\text{Acal}} = 1.1399 < F_{2,24,0.05} = 3.40$$

Thus, we accept $H_0^A \Rightarrow$ fixed effects of different ring frames are not significant \Rightarrow mean effects of the ring frames are same and hence average yarn count is same for different ring frames.

Similarly,

$$F_{\text{Bcal}} = 5.8811 > F_{3,24,0.05} = 3.01$$

Thus, we reject $H_0^B \Rightarrow$ fixed effect of different speed is significant for at least one speed \Rightarrow mean effects of the speeds are not same and hence average yarn count is same for different speeds.

Also here,

$$F_{\text{ABcal}} = 3.7692 > F_{6,24,0.05} = 2.51$$

Thus, we Reject $H_0^{AB} \Rightarrow$ fixed interaction effect is significant for at least one pair of count and speed.

16.4 Exercise

- What are assignable and chance causes of variation? How these causes are used for testing of hypothesis in ANOVA technique?
- Describe analysis of variance technique.
- What is analysis of variance? What are its types? Describe any one.
- Using mathematical model describe how the F -test is suitable for testing hypothesis in one way ANOVA.
- Following results show fluidity values obtained under different temperatures in a mercerization experiment.

Fluidity Values					
Temp	25°	4.2	3.9	2.9	4.1
	40°	3.5	3.8	2.6	4.4
	60°	3.6	3.2	2.8	4.0
	75°	3.5	4.1	3.0	3.8

Using one way ANOVA can it be said that average fluidity values are same under different temperatures?

6. Following are the results of fabric production (in meters) collected from three different looms.

Loom	I	525	500	510	510
	II	500	495	505	515
	II	520	540	525	530

using one way ANOVA test whether the average production is same for all looms?

7. Following are the results of fabric strength (in units) woven from three different looms.

Loom	I	52.0	54.0	52.5	53.0
	II	52.5	51.5	51.0	52.0
	II	50.0	50.5	49.5	51.5

Analyze the above data using one-way ANOVA and write conclusions.

8. An experiment was conducted to study effect of four different dyes on the strength of the fabric and following results of fabric strength are obtained.

Dye	A	8.67	8.68	8.66	8.65	
	B	7.68	7.58	8.67	8.65	8.62
	C	8.69	8.67	8.92	7.7	
	D	7.7	7.90	8.65	8.20	8.60

Carry out analysis of the above data and write conclusions.

9. An experiment was conducted to study performance of four detergents at three different temperatures on the whiteness of fabric and following data were obtained.

		Detergent			
		A	B	C	D
Temperature	Cold	57	55	67	60
	Warm	49	52	68	62
	Hot	54	46	58	65

Carry out analysis of the above data and write conclusions.

10. Following table shows fluidity values obtained under different temperatures with different levels of caustic soda.

		% Caustic soda			
		18	22	24	30
Temp	25°	4.2	3.9	2.9	4.1
	40°	3.5	3.8	2.6	4.4
	60°	3.6	3.2	2.8	4.0
	75°	3.5	4.1	3.0	3.8

Carry out analysis of the above data using a two-way ANOVA and write conclusions.

11. A textile mill has large number of output giving looms. It is expected that each loom gives same output and to verify this, an engineer collected following data of output (in meters) for three shifts from three different looms with four different operators.

		Operator			
		A	B	C	D
Loom	I	57,65,59	55,54,60	67,60,58	60,66,54
	II	49,55,60	52,48,50	68,65,62	62,60,66
	III	54,64,58	46,50,48	58,60,62	65,60,62

Analyze the above data and write conclusions.

12. Following is the ANOVA table for one-way. Complete the table and write conclusions.

Source	df	SS	MSS	F ratio
Factor A	4	-	-	-
Error	-	42.50	-	-
Total	14	140.65		

13. Following is the ANOVA table for two-way. Complete the table and write conclusions.

Source	df	SS	MSS	F ratio
Factor A	4	-	12.56	-
Factor B	3	-	-	-
Error	12	42.50	-	-
Total	19	440.65		

14. Following is the ANOVA table for two-way with 3 results per cell.
Complete the table and write conclusions.

Source	df	SS	MSS	F ratio
Factor A	4	—	12.56	—
Factor B	—	252.35	—	—
Interaction AB	12	—	—	—
Error	—	142.50	—	—
Total	59	1440.65		

Design of experiments

17.1 Introduction

It is the branch of statistics which deals with designing or planning the experiment according to the objective of the study, conducting the experiment according to the planning, collecting the results, analyzing the data using analysis of variance technique and finally writing the conclusions on the basis of analysis.

Basic concepts and terminology

Treatment

Any factor or any parameter whose effect is to be studied on the variable of the interest is called the treatment.

Experimental material

It is the actual raw material on which the treatment is applied for the purpose of the study.

Experimental unit (plot)

It is the smallest or singular unit of the experimental material to which the treatment is actually applied for the purpose of the study.

Yield/response

It is the singular numerical observation or result, which is collected from the experimental unit after application of the treatment.

Block

A sub group of the experimental material is called the block. Each block is homogeneous within itself but there is variation from the block to the block.

Basic principles of design of experiment

Principle of randomization, principle of replication, and principle of local control are the three basic principles of the design of experiments.

Principle of randomization

Randomization simply means applying the treatments at random to the experimental units. Randomization removes personal bias from the results, and therefore, it increases efficiency of the design. It also assures independence of the observations.

Principle of replication

Replication means repetitions of the treatments while applying to the experimental units. Larger the number of replications, larger is the accuracy of the results obtained from the statistical analysis. The number of replications depends on several factors such as type of trial, time, money, labor, etc.

Principle of local control

Sometimes the experimental material used for the purpose of the study is not homogeneous according to one or more properties. In such cases, the experimental material is divided into number of subgroups to reduce the error of experiment. That is, to increase the efficiency of the experiment. This procedure of dividing experimental material into number of sub groups is called the principle of local control. Such subgroups are also called the blocks.

The three principles of the design of experiments help in reducing the error of the experiment and hence to increase the efficiency of the experiment. Completely randomized design (CRD), randomized block design (RBD) and latin square design (LSD) are the three basic designs based on three principles of the design of experiments.

17.2. Completely randomized design

This is the simplest standard design, which is based on the principle of randomization and replication only. In case of completely randomized design (CRD), it is assumed that, there are “ k ” different treatments denoted by T_1, T_2, \dots, T_k whose mean effects are to be compared with each other. For the purpose of comparison, the “ k ” different treatments are applied to the homogeneous experimental material, made up of “ n ” experimental units, at random. That is the treatments are applied using the principle of randomization. Further it is assumed that, there are n_1 repetitions (replications)

of treatment T_1 , n_2 replications of treatment T_2, \dots, n_k replications of treatment T_k , such that $n_1 + n_2 + \dots + n_k = n$. Thus, after conducting the CRD experiment, total “ n ” observations obtained, which are shown as follows:

Treatment				
T_1	T_2	...	T_k	
x_{11}	x_{21}	...	x_{k1}	
x_{12}	x_{22}	...	x_{k2}	
:	:		:	
x_{1n1}	x_{2n2}	...	x_{knk}	

Where,

x_{ij} denote j^{th} observation corresponding to the i^{th} treatment.

The results of the CRD can be analyzed using the technique of one-way analysis of variance and the analysis of variance table for the CRD can be given in Table 17.1.

Table 17.1

Source	Degrees of freedom	S.S.	M.S.S.	F ratio
Treatment	$k - 1$	SST	MSST	$F_T = \frac{\text{MSST}}{\text{MSSE}}$
Error	$n - k$	SSE	MSSE	
Total	$n - 1$	TSS		

Note

Examples of CRD are same as that of one-way ANOVA. That is examples of CRD are solved using one-way ANOVA method.

17.3. Randomized block design

Randomized block design (RBD) is an improvement over the completely randomized design.

This is more efficient design than CRD as it is based on all the three principles of design of experiments. In case of RBD, also the main assumption is that, there are “ v ” treatments denoted by T_1, T_2, \dots, T_v whose mean effects are to be compared with each other and the experimental material used for

the purpose study is not homogeneous according to one property or in one direction. For this purpose, the experimental material is divided into the “ b ” blocks, such that each block is homogeneous within itself and there is variation from block to block. Now to these “ b ” blocks “ v ” treatments are applied at random in such a way that each treatment will occur once and only once in each block. Thus the size of the block is “ v ” and each treatment will have “ b ” replications. After conducting the randomized block design total “ vb ” observations are obtained which can be shown as follows:

		Blocks			
Treatment		B ₁	B ₂	...	B _b
T ₁	x ₁₁	x ₁₂	...	x _{1b}	
T ₂	x ₂₁	x ₂₂	...	x _{2b}	
:	:	:			:
T _v	x _{v1}	x _{v2}	...	x _{vb}	

Where,

x_{ij} denotes observation corresponding to the i^{th} treatment (T_j) and j^{th} block (B_j).

The results of randomized block design can be analyzed using the technique of two-way analysis of variance. Thus the analysis of variance table for the randomized block design is shown in Table 17.2.

Table 17.2

Source	d.f.	S.S.	M.S.S.	F-ratio
Treatment	$v - 1$	SST	MSST	$F_T = \frac{\text{MSST}}{\text{MSSE}}$
Block	$b - 1$	SSB	MSSB	$F_B = \frac{\text{MSSB}}{\text{MSSE}}$
Error	$(v - 1)(b - 1)$	SSE	MSSE	
Total	$(vb - 1)$	TSS		

Note

Examples of RBD are same as that of two-way ANOVA. That is examples of RBD are solved using two-way ANOVA method.

17.4. Latin square design (LSD)

Latin square design is the further improvement over the randomized block design. Like RBD it is also based on all the three principles of design of experiments. In case of Latin square design also, the main assumption is that, there are “ m ” treatments denoted by T_1, T_2, \dots, T_m whose mean effects are to be compared with each other and the experimental material under study is not homogeneous according to the two different properties or in two different directions. For this purpose the experimental material is divided into “ m ” rows and “ m ” columns in such a way that each row/column is homogeneous within itself and there is variation from row to row and column to column. Hence the design is also called the $m \times m$ LSD. Now, to these “ m ” rows and “ m ” columns, the “ m ” treatments are applied at random in such a way that, each treatment will occur once and only once in each row and in each column. Thus by conducting LSD experiment, total “ m^2 ” observations are obtained which are denoted by the notation x_{ijk} where, x_{ijk} represents the observation corresponding to i^{th} row, j^{th} column and k^{th} treatment. For example a 4×4 layout corresponding to four treatments A, B, C and D can be given as follows:

		Column			
		C ₁	C ₂	C ₃	C ₄
Row	R ₁	D	A	B	C
	R ₂	C	D	A	B
	R ₃	A	B	C	D
	R ₄	B	C	D	A

For the analysis of the results of $m \times m$ LSD a mathematical model is assumed for x_{ijk} which is as follows:

$$x_{ijk} = \mu_{ijk} + \epsilon_{ijk}$$

where, μ_{ijk} is the combined mean effect of i^{th} row, j^{th} column and k^{th} treatment.

ϵ_{ijk} is the random error component, $\epsilon_{ijl} \sim N(0, \sigma^2)$ and all ϵ_{ijk} are independent. The above model is rewritten as follows:

$$x_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \epsilon_{ijk}$$

where. μ is general mean effect,

α_i is the fixed effect of i^{th} row,

β_j is the fixed effect of j^{th} column,

τ_k is the fixed effect of k^{th} treatment.

Thus in this case interest is to test following hypotheses,

$$H_0^R: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0 \text{ Vs } H_1^R: \alpha_i \neq 0 \text{ for at least one } i$$

$$H_0^C: \beta_1 = \beta_2 = \dots = \beta_m = 0 \text{ Vs } H_1^C: \beta_j \neq 0 \text{ for at least one } j$$

$$H_0^T: \tau_1 = \tau_2 = \dots = \tau_m = 0 \text{ Vs } H_1^T: \tau_k \neq 0 \text{ for at least one } i$$

For testing the above hypotheses in this case also total variation is divided into four components as follows:

$$\text{TSS} = \text{SSR} + \text{SSC} + \text{SST} + \text{SSE}$$

Thus from the Cochran's theorem, it is expected that,

$$\text{SSR} \sim \sigma^2 \chi_{m-1}^2 \text{ i.e. } \frac{\text{SSR}}{\sigma^2} \sim \chi_{m-1}^2$$

$$\text{SSC} \sim \sigma^2 \chi_{m-1}^2 \text{ i.e. } \frac{\text{SSC}}{\sigma^2} \sim \chi_{m-1}^2$$

$$\text{SST} \sim \sigma^2 \chi_{m-1}^2 \text{ i.e. } \frac{\text{SST}}{\sigma^2} \sim \chi_{m-1}^2$$

$$\text{SSE} \sim \sigma^2 \chi_{(m-1)(m-2)}^2 \text{ i.e. } \frac{\text{SSE}}{\sigma^2} \sim \chi_{(m-1)(m-2)}^2$$

and all four components are independent of each other.

Hence, from the property of the F probability distribution,

$$F_R = \frac{\frac{\text{SSR}}{\sigma^2} / (m-1)}{\frac{\text{SSE}}{\sigma^2} / ((m-1)(m-2))} = \frac{\text{SSR}/m-1}{\text{SSE}/(m-1)(m-2)} = \frac{\text{MSSR}}{\text{MSSE}}$$

Similarly,

$$F_C = \frac{\text{MSSC}}{\text{MSSE}}$$

$$F_T = \frac{\text{MSST}}{\text{MSSE}}$$

where,

MSST represents mean sum of squares due to treatment, MSSR represents mean sum of squares due to rows, MSSC represents Mean sum of squares due to columns and MSSE represents Mean sum of squares due to error.

All the above results of mathematical analysis are written in the form of analysis of variance table (ANOVA table) is given in Table 17.4.

Table 17.4

Source	d.f	S.S	MSS	F ratio
Treatment	$(m - 1)$	SST	MSST	$F_T = \frac{\text{MSST}}{\text{MSSE}}$
Rows	$(m - 1)$	SSR	MSSR	$F_R = \frac{\text{MSSR}}{\text{MSSE}}$
Column	$(m - 1)$	SSC	MSSC	$F_B = \frac{\text{MSSB}}{\text{MSSE}}$
Error	$(m - 1)(m - 2)$	SSE	MSSE	
Total	$(m^2 - 1)$	TSS		

Here it is expected that, the statistics F_T , F_R and F_B will follow "F" probability distribution with $(m - 1)$ and $(m - 1) \cdot (m - 2)$ degrees of freedom.

Criteria for rejection of H_0

1. Reject null hypothesis H_0^R at $100\alpha\%$ los, if

$$F_{Cal}^R > F_{(m-1), (m-1)(m-2), \alpha}$$

2. Reject null hypothesis H_0^C at $100\alpha\%$ los, if

$$F_{Cal}^C > F_{(m-1), (m-1).(m-2), \alpha}$$

3. Reject null hypothesis H_0^T at $100\alpha\%$ los, if

$$F_{Cal}^T > F_{(m-1), (m-1).(m-2), \alpha}$$

Procedure and formulae for computation

1. Identify rows, columns, and treatments using the given data. Also decide their levels (m).
2. Find totals of observations for each row and denote them by R_1, R_2, \dots, R_m .
3. Find totals of observations for each column and denote them by C_1, C_2, \dots, C_m .
4. Find totals of observations for each treatment and denote them by T_1, T_2, \dots, T_m .

5. Find Grand total $G = \sum \sum \sum x_{ijk} = \sum R_i = \sum C_j = T_k$
6. Find CF using $CF = \frac{G^2}{n}$
7. Find Raw sum of squares using Raw S. S. = $\sum \sum \sum x_{ijk}^2$ that is find sum of squares of each and every observation.
8. Find TSS using $TSS = \text{Raw SS} - CF$
9. Find SSR using $SSR = \sum \frac{R_i^2}{m} - CF$
10. Find SSC using $SSC = \sum \frac{C_j^2}{m} - CF$
11. Find SST using $SST = \sum \frac{T_k^2}{m} - CF$
12. Find SSE using $SSE = TSS - SSR - SSC - SST$
13. Prepare analysis of variance table and write conclusions by comparing calculated F values with corresponding table values of F .

Example 17.1

A 4×4 LSD was conducted to study the effects of four different dyes A, B, C and D on the strength of the fabric. To remove the variation of the laboratory and the operators four different operators conducted the experiment in four different laboratories and the results obtained are as follows:

Lab	I	II	III	IV
I	66(B)	74(D)	70(A)	72(C)
II	75(D)	68(A)	68(C)	65(B)
III	69(A)	72(C)	63(B)	75(D)
IV	70(C)	65(B)	74(D)	70(A)

Carry out analysis of the above data and write the conclusions.

Solution

Here, Rows \Rightarrow Laboratory, Columns \Rightarrow Operators and Treatments \Rightarrow Dye

4×4 LSD \Rightarrow there are four levels of each which are as follows:

$R_1 \Rightarrow$ Lab-1, $R_2 \Rightarrow$ Lab-2, $R_3 \Rightarrow$ Lab-3 and $R_4 \Rightarrow$ Lab-4

$C_1 \Rightarrow$ Operator-1, $C_2 \Rightarrow$ Operator-2, $C_3 \Rightarrow$ Operator-3, and $C_4 \Rightarrow$ Operator-4

$T_1 \Rightarrow 5\%$, $T_2 \Rightarrow 7\%$, $T_3 \Rightarrow 10\%$ and $T_4 \Rightarrow 15\%$

Thus, the interest is to test the hypotheses

$$H_0^R: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0 \text{ Vs } H_1^R: \alpha_i \neq 0 \text{ for at least one } i$$

$$H_0^C: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \text{ Vs } H_1^C: \beta_j \neq 0 \text{ for at least one } j$$

$$H_0^T: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0 \text{ Vs } H_1^T: \tau_k \neq 0 \text{ for at least one } k$$

For testing these hypotheses the analysis is carried out as follows:

$$R_1 = 282, R_2 = 276, R_3 = 279 \text{ and } R_4 = 279$$

$$C_1 = 280, C_2 = 279, C_3 = 275 \text{ and } C_4 = 282$$

$$T_1 = 277, T_2 = 259, T_3 = 282 \text{ and } T_4 = 298$$

$$\text{Grand Total } G = \sum \sum \sum x_{ijk} = \sum R_i = \sum C_j = \sum T_k = 1116$$

$$CF = \frac{G^2}{n} = 1116^2 / 16 = 77841$$

$$\text{Raw S.S} = \sum \sum \sum x_{ijk}^2 = 78054$$

$$\text{TSS} = \text{Raw SS} - CF = 78054 - 77841 = 213$$

$$SSR = \sum \frac{R_i^2}{m} - CF = 4.5$$

$$SSC = \sum \frac{C_j^2}{m} - CF = 6.5$$

$$SST = \sum \frac{T_k^2}{m} - CF = 193.5$$

$$SSE = TSS - (SSR + SSC + SST) = 8.5$$

Thus the analysis of variance table can be given in Table 17.5.

Table 17.5

Source	df	SS	MSS	F ratio
Dye	3	193.5	64.5	$F_T = 45.5283$
Labs	3	4.5	1.5	$F_R = 1.0588$
Operator	3	6.5	2.1667	$F_C = 1.5294$
Error	6	8.5	1.4167	
Total	15	213		

Here it is expected that the statistics F_T , F_R and F_C will follow F distribution with 3 and 6 degrees of freedom.

Hence at 5% los, that is for $\alpha = 0.05$, $F_{3,6,0.05} = 4.76$

$$\text{As, } F_{T\text{Cal}} = 45.5283 > F_{3,6,0.05} = 4.76$$

The treatment hypothesis is rejected \Rightarrow the fixed effect of the treatment is significant for at least one treatment \Rightarrow average fabric strength is not same for different dyes.

Also,

$$F_{real} = 1.0588 < F_{3,6,0.05} = 4.76$$

The row hypothesis is accepted \Rightarrow laboratory effects are not significant \Rightarrow average fabric strength is same for all labs.

$$F_{real} = 1.5294 < F_{3,6,0.05} = 4.76$$

The column hypothesis is accepted \Rightarrow operator effects are not significant \Rightarrow average fabric strength is same for all operators.

17.5 Factorial experiments

When two or more factors, each with two or more levels, are studied in an experiment, then it is called the factorial experiment. Thus, in factorial experiments several factors are studied at a time in terms of their effects. These factors can be studied by taking the experimental material, which is homogeneous (CRD), the experimental material is non-homogeneous according to only one property (RBD) and the experimental material is non-homogeneous according to two different properties (LSD). Generally, the factorial experiments are conducted with RBD taking experiment material, which is non-homogeneous according to only one property. Symmetric factorial experiments (all factors with same levels), asymmetric factorial experiments (all factors with different levels) and mixed factorial experiments are the three different types of the factorial experiments.

Symmetric factorial experiments

The factorial experiment in which all factors are taken at the same number of levels is called the symmetric factorial experiment. 2^n factorial experiments, 3^n factorial experiments etc. are the main types of the symmetric factorial experiments.

2^n factorial experiments

When there are “ n ” factors, each with 2 levels under study, then it is called the 2^n factorial experiment. The lower level and the higher level generally denote the two levels of the experiment. If $n = 2$, that is there are only two factors, each with two levels, then the experiment is called the 2^2 factorial experiment and if $n = 3$, that is there are three different factors, each with two levels, then the experiment is called the 2^3 factorial experiment and so on.

2² factorial experiments

If only two factors, each with two levels are studied in an experiment, then it is called a 2² factorial experiment. These two factors of the 2² factorial experiment are generally denoted by A and B. The two levels of the factor A are denoted by a_0 (lower) and a_1 (higher). Also the two levels of the factor B are denoted by b_0 (lower) and b_1 (higher). Thus in 2² factorial experiment, a_0b_0 , a_1b_0 , a_0b_1 and a_1b_1 are the four (2²) different combinations of the factors A and B. These four different combinations can be represented by the simple notations called the Yates's notations as follows:

Combination	Yates notations
a_0b_0	1
a_1b_0	a
a_0b_1	b
a_1b_1	ab

Now, in 2² factorial experiments above four combinations are treated as the treatments and are generally studied with "r" blocks or replicates. Hence, "2².r" total observations can be obtained by conducting a 2² factorial experiment. These observations can be easily analyzed by using the Yates method which is as follows:

Yate's method

In this method the factorial effects are divided into the two main effects (A and B) and the interaction effect AB , which are represented by the contrasts and are calculated by preparing Yates table. The last column of the Yates table gives the factorial effect totals for the main and interaction effects and can be obtained as follows:

Treatment combination	"r"		Factorial effect total
	replication totals		
1	(I)	$(1) + (a)$	$(1) + (a) + (b) + (ab) = G$
a	(a)	$(b) + (ab)$	$(a) - (1) + (ab) - (b) = [A]$
b	(b)	$(a) - (1)$	$(b) + (ab) - (1) - (b) = [B]$
ab	(ab)	$(ab) - (b)$	$(ab) - (b) - (a) + (1) = [AB]$

where,

G represents grand total,

$[A]$ represents factorial effect total for main effect A ,

$[B]$ represents factorial effect total for main effect B ,

$[AB]$ represents factorial effect total for interaction effect AB .

Now, here the total variation is divided into the five components as follows:

$$\text{TSS} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SS Blocks} + \text{SSE}$$

Also,

The estimate of the main effect $A = \frac{[A]}{2^2}$ and $\text{SSA} = \frac{[A]^2}{2^2.r}$

The estimate of the main effect $B = \frac{[B]}{2^2}$ and $\text{SSB} = \frac{[B]^2}{2^2.r}$

The estimate of the interaction effect $AB = \frac{[AB]}{2^2}$ and $\text{SSAB} = \frac{[AB]^2}{2^2.r}$

Similarly,

The Sum of squares of block is $\text{SSBlock} = \sum \frac{B_j^2}{2^2} - \text{CF}$

where, B_1, B_2, \dots, B_r are the totals of “ r ” blocks.

Thus the analysis of variance table for a 2^2 factorial experiment with “ r ” replicates can be given in Table 17.6.

Table 17.6

Source	df	SS	MSS	F ratio
Main effects				
A	1	SSA	MSSA	F_A
B	1	SSB	MSSB	F_B
Interaction effect				
AB	1	SSAB	MSSAB	F_{AB}
Block	$(r - 1)$	SS Blocks	MSSBlock	
Error	$(2^2 - 1)(r - 1)$	SSE	MSSE	
Total	$2^2.r - 1$	TSS		

Example 17.2

A 2^2 factorial experiment was conducted to study the effect of the speed of the ring frame (A) and the traveler weight (B) on the hairiness index of the yarn. The experiment was carried out by spinning the yarns of three different counts on the same R/F. The results of hairiness indices obtained from the samples are as follows:

20 ^s	40 ^s	60 ^s
(b) 7.8	(ab) 10.0	(1) 6.5
(1) 6.2	(a) 8.5	(ab) 9.5
(ab) 10.2	(b) 7.5	(b) 7.0
(a) 8.2	(I) 6.8	(a) 8.2

Carry out the analysis of the above data and write the conclusions.

Solution

Here the two factors are $A \Rightarrow$ speed of the ring frame

$B \Rightarrow$ traveler weight

Also,

the two levels of the factor A are denoted by a_0 (lower) and a_1 (higher) and the two levels of the factor B are denoted by b_0 (lower) and b_1 (higher).

The 2^2 factorial is conducted in three replicates(counts), that is $r = 3$.

Thus here the interest is to test the hypotheses,

H_0^A : Main effect of the speed (A) is not significant
Vs

H_1^A : Main effect of the speed (A) is significant

H_0^B : Main effect of the traveler weight (B) is not significant
Vs

H_1^B : Main effect of the traveler weight (B) is significant

H_0^{AB} : Interaction effect of the speed and the traveler weight (AB) is not significant
Vs

H_1^{AB} : Interaction effect of the speed and the traveler weight (AB) is significant

H_0^{Blocks} : Block (Count) effects are not significant
Vs

H_1^{Blocks} : At least one block (Count) effect is significant

Now, for testing the above hypotheses the analysis is carried out using Yates method by preparing Yate's table such as Table 17.7.

Table 17.7

(1)	(2)	(3)	(4)
Combination	Total		
I	(1) = 19.5	44.4	96.4 = G
a	(a) = 24.9	52	12.8 = [A]
b	(b) = 22.3	5.4	7.6 = [B]
ab	(ab) = 29.7	7.4	2.0 = [AB]

Now,

$$CF = \frac{G^2}{n} = 96.4^2 / 12 = 774.4133$$

Raw S.S = Sum of squares of all observations = 794.04

Therefore, TSS = Raw S.S - C.F = 19.5967

$$SSA = \frac{[A]^2}{2^2.r} = \frac{12.6^2}{12} = 13.65333$$

$$SSB = \frac{[B]^2}{2^2.r} = \frac{7.8^2}{12} = 4.813333$$

$$SSAB = \frac{[AB]^2}{2^2.r} = \frac{2.2^2}{12} = 0.333333$$

$$SSBlock = \frac{\sum B_j^2}{r} - CF = 0.3467$$

Therefore, SSE = TSS - (SSA +SSB + SSAB+ SSBlock) = 0.48

Thus the ANOVA table for 2^2 factorial experiment is shown in Table 17.8.

Table 17.8

Source	df	SS	M.S.S	F ratio
Speed (A)	1	13.65333	13.65333	170.6667
Traveler weight (B)	1	4.813333	4.813333	60.16667
Interaction AB	1	0.333333	0.333333	4.166667
Count (Blocks)	2	0.346667	0.173333	2.166667
Error	6	0.48	0.08	
Total	11	19.62667		

Here,

it is expected that F_A , F_B and F_{AB} will follow F distribution with 1 and 6 degrees of freedom.

Also F_{Block} will follow F distribution with 2 and 6 degrees of freedom.
Now,

at 5% los, that is for $\alpha = 0.05$, $F_{1,6,0.05} = 5.99$ and $F_{2,6,0.05} = 5.14$

$$F_{A\text{cal}} = 170.1667 > F_{1,6,0.05} = 5.99$$

⇒ The main effect of the speed of the ring frame is significant on the hairiness index of the yarn. That is, the hairiness is affected by the speed of the ring frame.
Also,

$$F_{B\text{cal}} = 60.1667 > F_{1,6,0.05} = 5.99$$

⇒ The main effect of the traveler weight of the ring frame is significant on the hairiness index of the yarn. That is, the hairiness is affected by the traveler weight of the ring frame.

Further,

$$F_{AB\text{cal}} = 4.1667 < F_{1,6,0.05} = 5.99$$

⇒ The interaction effect of the speed and the traveler weight of the ring frame is not significant on the hairiness index of the yarn. That is, the hairiness is not affected by the interaction of the speed and the traveler weight of the ring frame.

Also,

$$F_{\text{Block}} = 2.1667 < F_{2,6,0.05} = 5.14$$

⇒ Effect of the count of the yarn is not significant on the hairiness.

Thus, the hairiness is affected by the speed and the traveler weight but the interaction of speed and traveler weight does not affect the hairiness. Also, count variation does not affect hairiness index.

2^3 factorial experiments

We have seen that the experiment in which only two factors, each with two levels are studied is called the 2^2 factorial experiment. Similarly the experiment in which three different factors, each with two levels are studied is called the 2^3 factorial experiment. These three factors of the 2^3 factorial experiment are generally denoted by A, B and C. The two levels of the factor A are denoted by a_0 (lower) and a_1 (higher); the two levels of the factor B are denoted by b_0 (lower) and b_1 (higher) and the two levels of the factor C are denoted by c_0 (lower) and c_1 (higher). Thus in 2^3 factorial experiment, $a_0 b_0 c_0$, $a_1 b_0 c_0$, $a_0 b_1 c_0$ and so on $a_1 b_1 c_1$ are the eight (2^3) different

combinations of the factors A, B and C. Here also these eight different combinations can be represented by the simple notations called the Yates's notations as follows:

Combination	Yates notations
$a_0 b_0 c_0$	1
$a_1 b_0 c_0$	a
$a_0 b_1 c_0$	b
$a_1 b_1 c_0$	ab
$a_0 b_0 c_1$	c
$a_1 b_0 c_1$	ac
$a_0 b_1 c_1$	bc
$a_1 b_1 c_1$	abc

Thus, in 2^3 factorial experiments above eight combinations are treated as the treatments and are generally studied with "r" blocks or replicates. Hence, " $2^3 \cdot r$ " total observations can be obtained by conducting a 2^3 factorial experiment. These observations can be easily analyzed by using the Yates's method which is as follows:

Yate's method

In this method the factorial effects are divided into the three main effects (A, B and C), the two factor interaction effects AB, AC and BC and three factor interaction effect ABC and these are represented by the contrasts and are calculated by preparing Yates table. The last column of the Yates table gives the factorial effect totals for the main and interaction effects and can be obtained such as Table 17.9.

Table 17.9

Treatment combination	"r"				Factorial effect total
	replication totals	(1)	$(1) + (a)$	$(1) + (a) + (b) + (ab)$	$(1) + (a) + (b) + (ab) + (c) + (ac)$ + $(bc) + (abc) = G$
a		(a)	$(b) + (ab)$	$(c) + (ac) + (bc) +$ (abc)	$-(1) + (a) - (b) + (ab) - (c)$ + $(ac) - (bc) + (abc) = [A]$
b		(b)	$(c) + (ac)$	$(a) - (1) + (ab) - (b)$	$-(1) - (a) + (b) + (ab) - (c)$ - $(ac) + (bc) + (abc) = [B]$
ab		(ab)	$(bc) + (abc)$	$(ac) - (c) + (abc) - (bc)$	$(1) - (a) - (b) + (ab) + (c) - (ac)$ - $(bc) + (abc) = [AB]$

Treatment combination	“r” replication totals	Factorial effect total				
		c	(c)	(a)-(1)	(b)+(ab)-(1)-(a)	- (1) - (a) - (b) - (ab) + (c) + (ac) + (bc) + (abc) = [C]
ac	(ac)			(ab)-(b)	(bc)+(abc)-(c)-(ac)	(1) - (a) + (b) - (ab) - (c) + (ac) - (bc) + (abc) = [AC]
bc	(bc)			(ac)-(c)	(ab)-(b)-(a)+(1)	(1) + (a) - (b) - (ab) - (c) - (ac) + (bc) + (abc) = [BC]
abc	(abc)			(abc)-(bc)	(abc)-(bc)-(ac)+(c)	- (1) + (a) + (b) - (ab) + (c) - (ac) - (bc) + (abc) = [ABC]

where,

G represents grand total, $[A]$ represents factorial effect total for main effect A , $[B]$ represents factorial effect total for main effect B , $[AB]$ represents factorial effect total for interaction effect AB , $[C]$ represents factorial effect total for main effect C , $[AC]$ represents factorial effect total for interaction effect AC , $[BC]$ represents factorial effect total for interaction effect BC and $[ABC]$ represents factorial effect total for interaction effect ABC .

Now, here the total variation is divided into the five components as follows:

$$\begin{aligned} \text{TSS} &= \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSC} + \text{SSAC} \\ &\quad + \text{SSBC} + \text{SSABC} + \text{SS Blocks} + \text{SSE} \end{aligned}$$

Also,

The estimate of the main effect

$$A = \frac{[A]}{2^3} \text{ and } \text{SSA} = \frac{[A]^2}{2^3.r}$$

The estimate of the main effect

$$B = \frac{[B]}{2^3} \text{ and } \text{SSB} = \frac{[B]^2}{2^3.r}$$

The estimate of the interaction effect

$$AB = \frac{[AB]}{2^3} \text{ and } \text{SSAB} = \frac{[AB]^2}{2^3.r}$$

The estimate of the main effect

$$C = \frac{[C]}{2^3} \text{ and } \text{SSC} = \frac{[C]^2}{2^3.r}$$

The estimate of the main effect AC

$$= \frac{[AC]}{2^3} \text{ and } SSAC = \frac{[AC]^2}{2^3.r}$$

The estimate of the interaction effect

$$BC = \frac{[BC]}{2^3} \text{ and } SSAC = \frac{[BC]^2}{2^3.r}$$

The estimate of the interaction effect

$$ABC = \frac{[ABC]}{2^3} \text{ and } SSABC = \frac{[ABC]^2}{2^3.r}$$

Similarly,

The Sum of squares of block is

$$SSBlock = \sum \frac{B_j^2}{2^3} - CF$$

where, B_1, B_2, \dots, B_r are the totals of “ r ” blocks.

Thus the analysis of variance table for 2^3 factorial experiments with “ r ” replicates is shown in Table 17.10.

Table 17.10

Source	df	SS	MSS	F ratio
Main effects				
A	1	SSA	MSSA	F_A
B	1	SSB	MSSB	F_B
C	1	SSC	MSSC	F_C
Interaction effects				
AB	1	SSAB	MSSAB	F_{AB}
AC	1	SSAC	MSSAC	F_{AC}
BC	1	SSBC	MSSBC	F_{BC}
ABC	1	SSABC	MSSABC	F_{ABC}
Block	$(r - 1)$	SSBlocks	MSSBlocks	F_{Blocks}
Error	$(2^3 - 1)(r - 1)$	SSE	MSSE	
Total	$(2^3.r - 1)$	TSS		

Example 17.3

A 2^3 factorial experiment was conducted to study the effect of the speed of the ring frame (A), the traveler weight (B) and the ring diameter (C) on the hairiness index of the yarn. The experiment was carried out by spinning the yarns of three different counts on the same R/F. The results of hairiness indices obtained from the samples are as follows:

20 ^s	40 ^s	60 ^s
(b) 7.8	(ab) 10.0	(c) 6.5
(1) 6.2	(c) 8.5	(abc) 9.5
(ab) 10.2	(b) 7.5	(b) 7.0
(a) 8.2	(1) 6.8	(a) 8.2
(c) 7.8	(abc) 10.0	(1) 6.5
(ac) 6.2	(a) 8.5	(ac) 9.5
(abc) 10.2	(bc) 7.5	(bc) 7.0
(bc) 8.2	(ac) 6.8	(ab) 8.2

Carry out the analysis of the above data and write the conclusions.

Solution

Here the three factors are $A \Rightarrow$ speed of the ring frame

$B \Rightarrow$ traveler weight

$C \Rightarrow$ ring diameter

Also,

the two levels of the factor A are denoted by a_0 (lower) and a_1 (higher), the two levels of the factor B are denoted by b_0 (lower) and b_1 (higher) and the two levels of the factor C are denoted by c_0 (lower) and c_1 (higher).

The 2^3 factorial is conducted in three replicates (counts), that is $r = 3$.

Thus here the interest is to test the hypotheses,

H_0^A : Main effect of the speed (A) is not significant
Vs

H_1^A : Main effect of the speed (A) is significant

Similarly the other hypotheses H_0^B , H_0^C , H_0^{AB} , H_0^{AC} , H_0^{BC} and H_0^{ABC} can be written.
Further,

H_0^{Blocks} : Block (Count) effects are not significant
Vs

H_1^{Blocks} : At least one block (Count) effect is significant

Now, for testing the above hypotheses the analysis is carried out using Yates method by preparing Yate's table such as Table 17.11.

Table 17.11

Treat comb	Total	Factorial total	
1	(1) = 19.5	44.4	95.1
<i>a</i>	(<i>a</i>) = 24.9	50.7	97.7
<i>b</i>	(<i>b</i>) = 22.3	45.3	11.5
<i>ab</i>	(<i>ab</i>) = 28.4	52.4	6.7
<i>c</i>	(<i>c</i>) = 22.8	5.4	6.3
<i>ac</i>	(<i>ac</i>) = 22.5	6.1	7.1
<i>bc</i>	(<i>bc</i>) = 22.7	-0.3	0.7
<i>abc</i>	(<i>abc</i>) = 29.7	7	7.3
			[<i>ABC</i>] = 6.6

Now,

$$CF = \frac{G^2}{n} = 96.2^2 / 24 = 1548.8267$$

Raw S. S. = Sum of squares of all observations = 1588.08

Therefore, TSS = Raw S.S - C.F = 39.2533

$$SSA = \frac{[A]^2}{2^3.r} = \frac{12.6^2}{24} = 13.8017$$

$$SSB = \frac{[B]^2}{2^3.r} = \frac{7.8^2}{24} = 7.4817$$

$$SSAB = \frac{[AB]^2}{2^3.r} = \frac{2.2^2}{24} = 2.6667$$

$$SSC = \frac{[C]^2}{2^3.r} = 0.2817$$

$$SSAC = \frac{[AC]^2}{2^3.r} = 0.96$$

$$SSBC = \frac{[BC]^2}{2^3.r} = 0.0267$$

$$SSABC = \frac{[ABC]^2}{2^3.r} = 1.815$$

$$SSBlock = \sum \frac{B_j^2}{2^3} - CF = 0.6933$$

Therefore, SSE = TSS - (SSA +SSB + SSAB+ SSBlock) = 11.5915

Thus the ANOVA table for 2^3 factorial experiment is shown in Table 17.12.

Table 17.12

Source	df	SS	MSS	F ratio
Main effects				
A (Speed)	1	13.8017	13.8017	16.7632
B (Tr. Weight)	1	7.4817	7.4817	9.0870
C (Ring Dia.)	1	2.6667	2.6667	3.2389
Interaction effects				
AB	1	0.2817	0.2817	0.3421
AC	1	0.9600	0.9600	1.1660
BC	1	0.0267	0.0267	0.0324
ABC	1	1.8150	1.8150	2.2045
Block	$(r - 1) = 2$	0.6933	0.3467	0.4211
Error	$(2^3 - 1)(r - 1) = 14$	11.5267	0.8233	
Total	$(2^3 \cdot r - 1) = 23$	39.2533		

Here,

it is expected that F_A , F_B , F_C , F_{AB} , F_{AC} , F_{BC} and F_{ABC} will follow F distribution with 1 and 14 degrees of freedom.

Also F_{Block} will follow F distribution with 2 and 14 degrees of freedom.

Now,

at 5% los, that is for $\alpha = 0.05$, $F_{1,14,0.05} = 4.60$ and $F_{2,14,0.05} = 3.74$

Here it is clear that only $F_{A\text{cal}}$ and $F_{B\text{cal}}$ are greater than the table value $F_{1,14,0.05} = 4.60$

⇒ The main effects of the speed of the ring frame and the traveler weight are significant on the hairiness index of the yarn. That is, the hairiness is affected by the speed of the ring frame and the traveler weight.

But here $F_{C\text{cal}}$ and $F_{AB\text{cal}}$, $F_{AC\text{cal}}$, $F_{BC\text{cal}}$, $F_{ABC\text{cal}}$ are less than the table value $F_{1,14,0.05} = 4.60$

⇒ The main effect of the ring diameter and all interaction effects are not significant on the hairiness index of the yarn. That is, the hairiness is not affected by the ring diameter as well as all interactions.

Also,

$$F_{\text{Block}} = 0.4211 < F_{2,6,0.05} = 3.74$$

⇒ Effect of the count of the yarn is not significant on the hairiness.

Thus, the hairiness is affected by the speed and the traveler weight but it is not affected by the ring diameter, all interactions as well as count.

17.6 Exercise

- What is design of experiments? What are its basic principles?
- Describe basic principles of design of experiments.
- Describe completely randomized design.
- Describe randomized block design.
- Describe Latin square design.
- What is factorial experiment? What is its use?
- Describe factorial experiment with two factors.
- Describe 2^3 factorial experiments.
- Following is the ANOVA table for CRD. Complete the table and write conclusions.

Source	d.f.	S.S	M.S.S.	F ratio
Treatment	4	—	—	—
Error	—	42.50	—	
Total	14	140.65		

- Following is the ANOVA table for RBD. Complete the table and write conclusions.

Source	d.f.	S.S	M.S.S.	F ratio
Treatment	4	—	12.56	—
Blocks	3	—	—	—
Error	12	42.50	—	
Total	19	440.65		

- Following is the ANOVA table for 5×5 LSD. Complete the table and write conclusions.

Source	d.f.	S.S	M.S.S.	F ratio
Treatment	—	—	12.56	—
Rows	—	252.35	—	—
Columns	—	—	—	—
Error	—	142.50	—	
Total	—	1440.65		

12. Following is the ANOVA table for 2^2 factorial experiment conducted with three replicates. Complete the table and write conclusions.

Source	d.f.	S.S	M.S.S.	F ratio
Main effects	—	25.65	—	—
A	—	—	12.56	—
B	—	—	—	—
Interaction	—	—	—	—
AB	—	—	—	—
Block	—	5.48	—	—
Error	—	32.50	—	—
Total	—	540.65		

13. A completely randomized experiment was conducted to study effect of four different chemicals on the strength of the fabric and following results of fabric strength are obtained.

chemical	A	8.67	8.68	8.66	8.65	
	B	7.68	7.58	8.67	8.65	8.62
	C	8.69	8.67	8.92	7.7	
	D	7.7	7.90	8.65	8.20	8.60

Carry out analysis of the above data and write conclusions.

14. Following are the results of CCM values obtained from dyeing experiment under different temperatures in different laboratories.

	Laboratories				
		A	B	C	D
Temp	25°	4.2	3.9	2.9	4.1
	40°	3.5	3.8	2.6	4.4
	60°	3.6	3.2	2.8	4.0
	75°	3.5	4.1	3.0	3.8

Analyze the above data using ANOVA technique assuming laboratories as the blocks and write conclusions.

15. A 4×4 LSD was conducted to study effect of four different dyes A, B, C and D on the percentage elongation of the yarn. The experiment

was conducted in 4 different laboratories by four different operators and results of hairiness index obtained are as follows:

		Laboratories			
		I	II	III	IV
Operator	1	4.2 C	3.9 B	2.9 A	4.1 D
	2	3.5 B	3.8 A	2.6 D	4.4 C
	3	3.6 D	3.2 C	2.8 B	4.0 A
	4	3.5 A	4.1 D	3.0 C	3.8 B

Analyze the above data using ANOVA technique and write conclusions.

16. A 2^2 factorial experiment was conducted to study the effect of the ring diameter (A) and the traveler weight (B) on the hairiness index of the yarn. The experiment was carried out by spinning the yarns on three different ring frames. The results of hairiness indices obtained from the samples are as follows:

R/F-I	R/F-II	R/F-III
(b) 4.8	(ab)8.0	(1) 6.5
(1) 6.2	(a) 5.5	(ab) 6.5
(ab) 7.2	(b) 7.5	(b) 7.0
(a) 5.2	(I) 6.8	(a) 5.2

Carry out the analysis of the above data and write the conclusions.

17. A 2^3 factorial experiment was conducted to study the effect of the speed of the ring frame (A), the traveler weight (B) and the ring diameter (C) on the $U\%$ of the yarn. The experiment was carried out

by spinning the yarns of three different counts on the same *R/F*. The results of *U%* obtained from the samples are as follows:

20 ^s	40 ^s	60 ^s
(<i>b</i>) 12.8	(<i>ab</i>) 11.0	(<i>c</i>) 16.5
(<i>1</i>) 14.2	(<i>c</i>) 11.5	(<i>abc</i>) 15.5
(<i>ab</i>) 13.2	(<i>b</i>) 9.5	(<i>b</i>) 12.0
(<i>a</i>) 12.2	(<i>1</i>) 12.8	(<i>a</i>) 11.2
(<i>c</i>) 14.5	(<i>abc</i>) 9.0	(<i>1</i>) 15.5
(<i>ac</i>) 12.2	(<i>a</i>) 11.5	(<i>ac</i>) 12.4
(<i>abc</i>) 13.4	(<i>bc</i>) 12.5	(<i>bc</i>) 10.0
(<i>bc</i>) 14.2	(<i>ac</i>) 9.8	(<i>ab</i>) 15.2

Carry out the analysis of the above data and write the conclusions.

18.1 Introduction

Statistical quality control (SQC) is very important branch of Statistics, as in the today's age of globalization quality of the product has become important aspect of the life. In this competitive age, as the customer has lot of choice, the companies whose products are of good quality will only survive in the markets. Thus, SQC is the branch of statistics, which deals with the statistical tools or the methods, which are used for controlling the quality of the product.

Quality

Different people have defined the term “Quality” in different ways. Dr. Juran, in 1964, defined the quality of the product as the “fitness for the use” by the customer. Philip Crosby, the promoter of the “zero-defects” concept, has defined the quality as the “Conformance to the requirement.” Dr. Deming, in 1986, defined the quality as the “Quality should be aimed at the needs of the customer.” American Society for Quality Control has given the definition of the quality as “Totality of features and characteristics of the product or some service that bear on its ability to satisfy given needs.” The common thread from all the above definitions is that, the quality is the measure of the extent to which customer requirements and expectations are satisfied. The quality is not static. It can change from place to place and from customer to customer.

Production process of an industry

The production process of every industry can be divided in following three stages



where,

Input: It is the stage where the raw material is supplied to the machines or the production processes.

Conversion: It is the stage where the raw material supplied as the input is processed and converted into final product.

Output: It is the stage where the final product produced by the conversion stage is packed and dispatched to the customers.

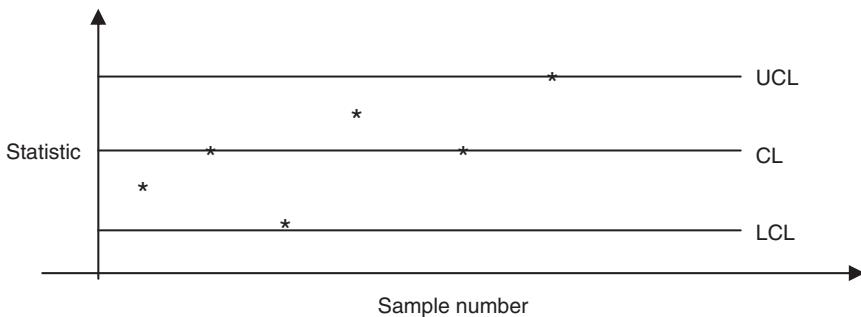
The SQC techniques are applied at each stage of the production process because of the variation occurring in the input/output material. Larger variation is the indication of bad quality and the SQC techniques help in deciding whether the variation occurring is tolerable (non significant). According to the stage of application, Process control and Lot control are the two different types of the SQC.

18.2 Process control

The variation in the quality of the products can occur at the conversion/processing stage of the production process and the techniques applied at the conversion stage of the production are called process control techniques. These techniques help to decide whether the continuous production process is under statistical control. Production process under control means quality of the product produced by the process is good or is as per the requirement. The statistical tool called the “control chart” achieves the process control.

18.3 Control chart

The control chart is the graphical tool used for deciding, whether the continuous production process is under statistical control or not. It is based on the central idea that, if the production process is under control, then the values of the statistic will lie in between “ $\text{mean} \pm 3\sigma$.” Thus, according to “ $\text{mean} \pm 3\sigma$ ” the control chart is made up of three different lines called the Central Line (CL), Lower Control Limit (LCL) and Upper Control limit (UCL). Along with these lines, on the control chart there are some points, which are related to the statistic calculated from the samples selected from the production process at the regular intervals. According to the scattered pattern of these points, the decision regarding the process, whether under control or not is made. The distance between the LCL ($\text{mean} - 3\sigma$) and the CL (mean) and between CL (mean) and UCL ($\text{mean} + 3\sigma$) is 3σ , hence the LCL and UCL are called 3σ control limits. Also, the distance between the LCL ($\text{mean} - 3\sigma$) and UCL ($\text{mean} + 3\sigma$) is 6σ , which is called the natural tolerance or variation. Typically, a control chart can be shown in Fig. 18.1



18.1

According to the statistic calculated and plotted on the control chart, the following are the main types of the control charts:

1. \bar{X} chart (control chart for mean).
2. R chart (control chart for range).
3. np chart (control chart for number of defectives).
4. p chart (control chart for proportion of defectives).
5. C chart (control chart for number of defects).

18.4 Interpretation of control chart

On the basis of control chart, the production process is said to be under control if all the points of the control chart are within the control limits and are at random. The process is said to be not under control if at least one point goes outside the control limits. Sometimes, even though all the points of control chart are within the control limits, the process is said to be not in statistical control as the points are not at random. Such cases of control charts are as follows:

Case I All the points of the control chart show upward trend.

Case II All the points of the control chart show downward trend.

Case III All the points of the control chart are within the upper half (CL to UCL)

Case IV All the points of the control chart are within the lower half (CL to LCL)

Case V All the points of the control chart are very much close to CL or are almost on CL.

18.5 Specification limits

Sometimes, the customer provides the control limits for the production process, such control limits are called the upper specification limit (USL) and lower specification limit (LSL) and the decision regarding the control of the production process is taken by taking care of these specification limits. The difference of the USL and LSL is called the specified tolerance.

18.6 \bar{X} chart

The \bar{X} chart is used, if the variable of the process is continuous. This chart helps in deciding whether the mean or the average of the process is under statistical control or not. To draw \bar{X} chart, k different samples, each of size n are selected and the points \bar{x}_i (means of the i^{th} sample) are plotted and according to these points the decision is made. The control limits for the \bar{X} chart can be obtained as follows:

Suppose,

X is the variable of the production process; μ is the mean and σ is the standard deviation of the process or the parameters of the process.

Case I Suppose the parameters μ and σ are known.

In this case the control limits are,

$$\text{CL} = \mu$$

$$\text{LCL} = \mu - 3\left(\sigma/\sqrt{n}\right) = \mu - A\sigma$$

$$\text{UCL} = \mu + 3\left(\sigma/\sqrt{n}\right) = \mu + A\sigma$$

where,

$A = 3/\sqrt{n}$ and it is tabulated in the table for the different values of n .

Case II Suppose, the parameters μ and σ are unknown.

In this case μ is replaced by \bar{x} and σ is replaced by \bar{R}/d_2 hence, the control limits are,

$$\text{CL} = \bar{\bar{x}}$$

$$\text{LCL} = \bar{\bar{x}} - 3\frac{\bar{R}/d_2}{\sqrt{n}} = \bar{\bar{x}} - A_2 \bar{R}$$

$$\text{UCL} = \bar{\bar{x}} + 3\frac{\bar{R}/d_2}{\sqrt{n}} = \bar{\bar{x}} + A_2 \bar{R}$$

where

$$\bar{\bar{x}} = \frac{\sum \bar{x}_i}{k} \text{ and } \bar{R} = \frac{\sum R_i}{k}$$

\bar{x}_i is the mean of i^{th} sample, R_i is the range of the i^{th} sample and k denotes total samples of size n selected for study is already given.

18.7 R-Chart

The R-chart is used, if the variable of the process is continuous. This chart helps in deciding whether the variation of the process is under statistical control or not. To draw R-chart, k different samples, each of size n are selected and the points R_i (range of the i^{th} sample) are plotted and according to these points the decision is made. Thus, the control limits for the R-chart can be obtained as follows:

Case I Suppose the parameters μ and σ are known.

In this case the control limits are,

$$\begin{aligned} \text{CL} &= d_2\sigma \\ \text{LCL} &= d_2\sigma - 3D\sigma = D_1\sigma \\ \text{UCL} &= d_2\sigma + 3D\sigma = D_2\sigma \end{aligned}$$

where,

d_2 , D_1 and D_2 are tabulated in the statistical table for the different values of n .

Case II Suppose, the parameters μ and σ are unknown.

In this case σ is replaced by \bar{R}/d_2 and the control limits are,

$$\begin{aligned} \text{CL} &= \bar{R} \\ \text{LCL} &= \bar{R} - 3D\bar{R}/d_2 = D_3\bar{R} \\ \text{UCL} &= \bar{R} + 3D\bar{R}/d_2 = D_4\bar{R} \end{aligned}$$

where, D_3 and D_4 are also tabulated in the statistical table.

Typically, R-chart can be shown as follows:

Example 18.1

Following data represents average and range of linear density of the yarn obtained from eight different samples each of size five, selected during a spinning process.

Sample no	1	2	3	4	5	6	7	8
Avg. linear density	19.6	20.1	20.5	19.4	22.3	21.7	20.3	19.9
Range	1.2	2.1	1.6	1.8	2.0	1.7	2.0	1.8

Draw the mean and range charts for the above data and comment on them.

Solution

Here the process is the spinning process and the variable X is the linear density of the yarn.

Suppose μ is the mean and σ is the standard deviation of X . Here μ and σ are unknown.

1. \bar{X} -chart (Mean Chart)

Given that sample size for each sample is $n = 5 \Rightarrow A_2 = 0.577, D_3 = 0$ and $D_4 = 2.115$

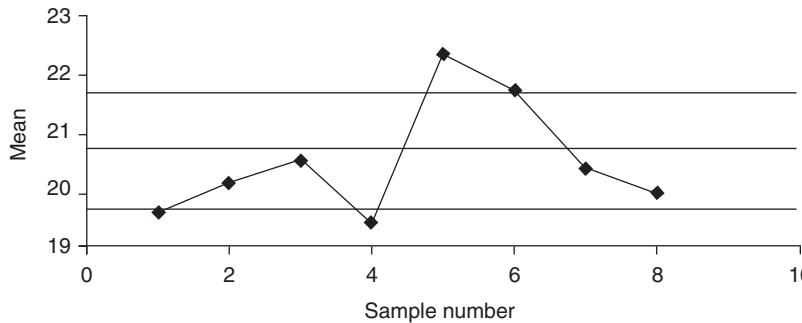
$$\text{Therefore, } \bar{\bar{x}} = \frac{\sum \bar{x}_i}{k} = \frac{163.8}{8} = 20.475 \text{ and } \bar{R} = \frac{\sum R_i}{k} = \frac{14.2}{8} = 1.775$$

$$\text{CL} = \bar{\bar{x}} = 20.475$$

$$\text{LCL} = \bar{\bar{x}} - A_2 \bar{R} = 19.4508$$

$$\text{UCL} = \bar{\bar{x}} + A_2 \bar{R} = 21.4992$$

Thus, the mean chart is drawn like Fig. 18.2.



18.2

As some of the points of the \bar{X} -chart are outside the control limits the spinning process is not under statistical control for the average linear density of the yarn.

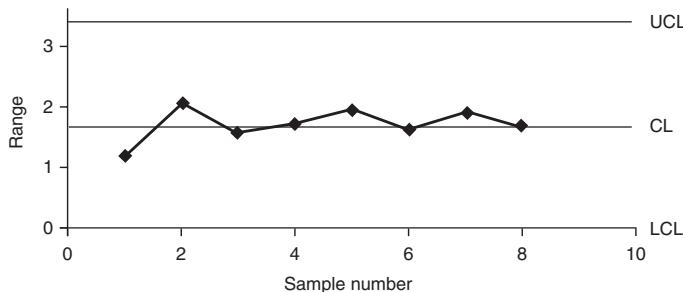
2. R-chart (range chart)

$$\text{CL} = \bar{R} = 1.775$$

$$\text{LCL} = D_3 \bar{R} = 0$$

$$\text{UCL} = D_4 \bar{R} = 3.7541$$

Thus, the Range chart is drawn like Fig. 18.3.



18.3

As all the points of the R-chart are inside the control limits the spinning process is under control for the variation in the linear density of the yarn.

Conclusion: Thus from the above two charts, it can be concluded that the average of the spinning process is not in statistical control but the variation of the process is under control.

18.8 *np*-Chart

The *np*-chart is used, if the variable of the process is discrete. This chart helps in deciding whether the number of defective articles/products produced by the process is under statistical control or not. To draw *np*-chart, k different samples, each of size n are selected from the production process and are inspected. Let d_i be the number of defective articles/products observed in the i^{th} sample. These points $d_i = np_i$ are plotted on the *np* chart and according to these plotted points the decision is made. Thus, the control limits for the *np*-chart can be obtained as follows:

Suppose, variable $X = d$ represents number of defective articles/products observed in the sample of size n . Here X will be binomially distributed random variable with parameters “ n ” and “ p ,” where p is the probability of getting defective article produced by the process.

Case I Suppose the parameter “ p ” is known.

In this case the control limits are as follows:

$$\text{CL} = np$$

$$\text{LCL} = np - 3 \times \sqrt{npq}$$

$$\text{UCL} = np + 3 \times \sqrt{npq}$$

Where, $q = 1 - p$

Case II Suppose, the parameter p is unknown.

In this case p is replaced by \bar{P} and the control limits are as follows:

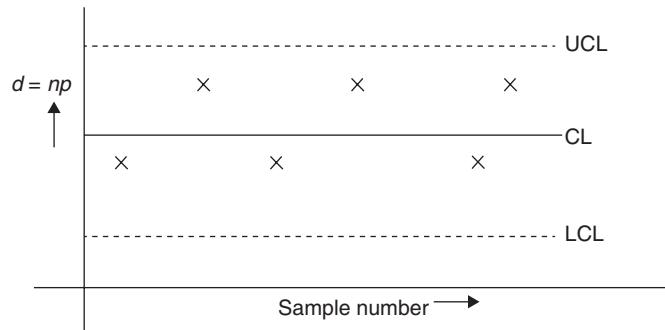
$$CL = n\bar{P}$$

$$LCL = n\bar{P} - 3 \times \sqrt{n\bar{P}\bar{Q}}$$

$$UCL = n\bar{P} + 3 \times \sqrt{n\bar{P}\bar{Q}}$$

where, $\bar{P} = \sum p_i / k = \sum d_i / nk$ and $\bar{Q} = 1 - \bar{P}$

Typically, np -chart can be drawn like Fig. 18.4.



18.4

18.9 p -Chart

The p -chart is also used, if the variable of the process is discrete. This chart helps in deciding whether the proportion (percentage) of defective articles/products produced by the process is under statistical control or not. To draw p -chart also, k different samples, each of size n are selected from the production process and are inspected. Let d_i be the number of defective articles/products observed in the i^{th} sample and $p_i = d_i/n$ be proportion of defective articles/products produced by the process. These points p_i are plotted on the p chart and according to these plotted points the decision is made. Thus, the control limits for the p -chart can be obtained as follows:

Suppose, variable $X = d$ represents number of defective articles/products observed in the sample of size n . Here X will be binomially distributed random variable with parameters “ n ” and “ p ,” where p is the probability of getting defective article produced by the process.

Case I Suppose the parameter “ p ” is known.

In this case the control limits are as follows:

$$CL = p$$

$$LCL = p - 3 \times \sqrt{\frac{pq}{n}}$$

$$UCL = p + 3 \times \sqrt{\frac{pq}{n}}$$

where,

$$q = 1 - p$$

Case II Suppose, the parameter p is unknown.

In this case p is replaced by \bar{P} and the control limits are as follows:

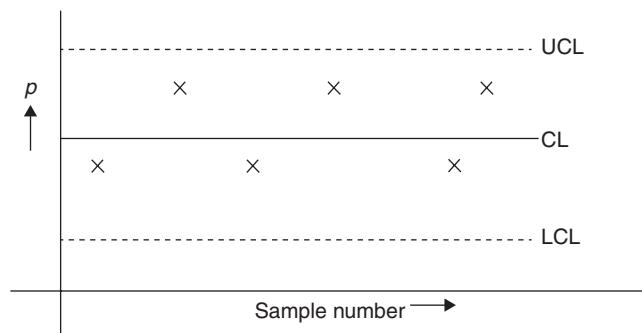
$$CL = \bar{P}$$

$$LCL = \bar{P} - 3 \times \sqrt{\frac{\bar{P}\bar{Q}}{n}}$$

$$UCL = \bar{P} + 3 \times \sqrt{\frac{\bar{P}\bar{Q}}{n}}$$

$$\text{Where, } \bar{P} = \sum \frac{p_i}{k} = \sum \frac{d_i}{n_k} \text{ and } \bar{Q} = 1 - \bar{P}$$

Typically, p -chart can be shown like Fig. 18.5.



Example 18.2

Five knitted garments each were selected at eight different times during the production and following results of number of defective garments were obtained.

Sample number	1	2	3	4	5	6	7	8
No. of defective garments	0	2	1	1	2	0	0	0

Draw the np -chart and p -chart for the above data. Also comment on each of them.

Solution

Here the process is the knitting process and the variable “ d ” denotes number of defective garments in a sample of size $n = 5$.

Suppose “ p ” is the proportion of defective garments produced by the process and is unknown.

1. np -chart

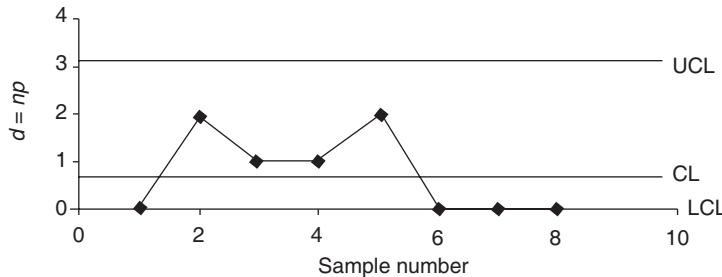
$$\bar{P} = \frac{\sum p_i}{k} = \frac{\sum d_i}{nk} = \frac{6}{5 \times 8} = 0.15$$

$$CL = n\bar{P} = 0.75$$

$$LCL = n\bar{P} - 3\sqrt{n\bar{P}\bar{Q}} = 0.75 - 3\sqrt{0.75 \cdot 0.25} = -1.6453 \approx 0$$

$$UCL = n\bar{P} + 3\sqrt{n\bar{P}\bar{Q}} = 0.75 + 3\sqrt{0.75 \cdot 0.25} = 3.1453 \approx 3.145$$

Thus the np -chart can be drawn like Fig. 18.6.



18.6

Conclusion: As all the points of the np -chart are inside the control limits the garment production process is under control for the number of defective garments.

2. *p*-chart

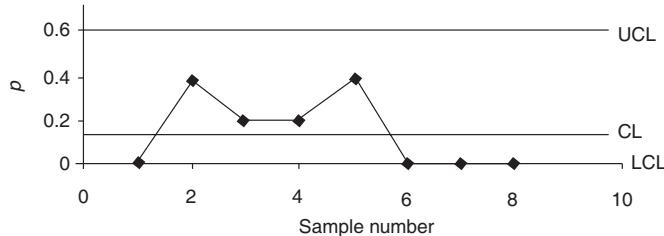
As p is unknown, the control limits are as follows:

$$CL = \bar{P} = 0.15$$

$$LCL = \bar{P} - 3 \times \sqrt{\frac{\bar{P}\bar{Q}}{n}} = 0.15 - 3 \times \sqrt{\frac{0.15 \times 0.85}{10}} = -0.3291 \approx 0$$

$$UCL = \bar{P} + 3 \times \sqrt{\frac{\bar{P}\bar{Q}}{n}} = 0.15 + 3 \times \sqrt{\frac{0.15 \times 0.85}{10}} = 0.6291$$

Thus the *p*-chart can be drawn like Fig. 18.7.



18.7

Conclusion: As all the points of the *p*-chart are inside the control limits the garment production process is under control for the proportion of defective garments.

18.10 C-chart

The *c*-chart is also used if the variable of the process is discrete. This chart helps in deciding whether the number of defects per article/product produced by the process is under statistical control or not. To draw *c*-chart, k different sample units are selected from the production process and are inspected. Let c_i be the number of defects observed in the i^{th} sample unit produced by the process. These points c_i are plotted on the *c*-chart and according to these plotted points the decision is made. Thus, the control limits for the *c*-chart can be obtained as follows.

Suppose, variable $X = C$ represents number of defects observed per article/product. Here X will follow Poisson probability distribution with parameter " λ " where λ is the average number of defects per article produced by the process.

Case I Suppose the parameter " λ " is known.

In this case the control limits are as follows:

$$CL = \lambda$$

$$LCL = \lambda - 3 \times \sqrt{\lambda}$$

$$UCL = \lambda + 3 \times \sqrt{\lambda}$$

Case II Suppose, the parameter λ is unknown.

In this case λ is replaced by \bar{C} and the control limits are as follows":

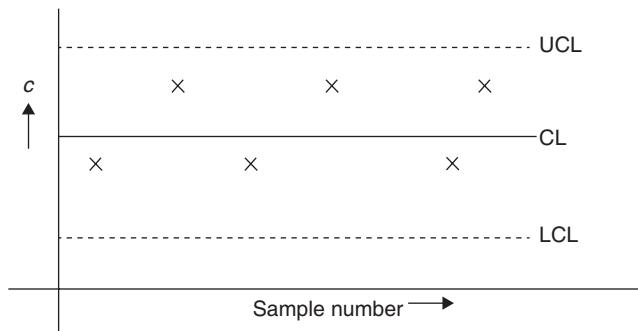
$$CL = \bar{C}$$

$$LCL = \bar{C} - 3 \times \sqrt{\bar{C}}$$

$$UCL = \bar{C} + 3 \times \sqrt{\bar{C}}$$

Where, $\bar{C} = \sum c_i / k$

Typically, c -chart can be shown like Fig. 18.8.



18.8

Example 18.3

The following data are related to number of defects observed in each of eight garments selected from the production and following results of number of defects per garment were obtained.

Garment number	1	2	3	4	5	6	7	8
No. of defects in garment	0	2	1	1	2	0	0	0

Draw the C-chart and comment on it.

Solution

Here production process is garment production process.

Suppose $C \Rightarrow$ Number of defects in a garment

$\lambda \Rightarrow$ average number of defects per garment (Unknown)

Therefore,

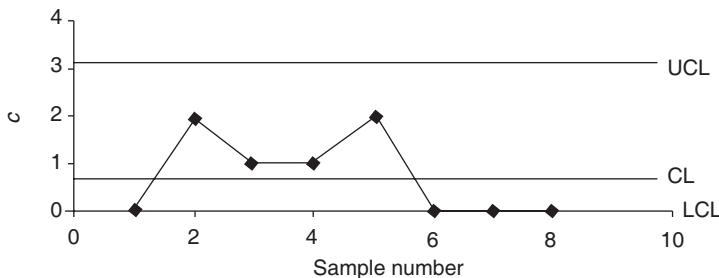
$$\bar{C} = \sum c_i / k = \frac{6}{8} = 0.75$$

$$CL = \bar{C} = 0.75$$

$$LCL = \bar{C} - 3 \times \sqrt{\bar{C}} = -1.8408 \approx 0$$

$$UCL = \bar{C} + 3 \times \sqrt{\bar{C}} = 3.3481$$

Thus the C-chart can be drawn like Fig. 18.9.



18.9

Conclusion: As all the points of the C-chart are inside the control limits the garment production process is under control for the number of defects per garment.

18.11 Lot control

The input/output stage of the production process is generally in the form of batches or lots. The proportion or the percentage defective items in the lot " P " is the indication of the quality of the lot. Smaller the value of " p " better is the quality of the lot that is the lot is under statistical control. The statistical techniques applied to study percentage of defectives in a lot at input or output stage are called lot control techniques. The lot control can be achieved by acceptance sampling methods that are also known as the rectifying acceptance sampling plans. The rectifying sampling plans help in deciding whether the lot under study is to be accepted/rejected on the basis of the sample selected from that lot. Single- and double-sampling plans are the two main types of rectifying sampling plans.

Single-sampling plan

Suppose $N \Rightarrow$ Size of the lot under study

$n \Rightarrow$ Size of the sample selected

$c \Rightarrow$ Acceptance number (Number of defective items allowed)

$d \Rightarrow$ Number of defective items found in the sample.

With all the above assumptions, the single sampling plan is as follows:

- Step 1* Select the sample of size “ n ” from the lot and check it. Suppose the sample contains “ d ” defective items.
- Step 2* If $d \leq c$, then accept the lot by replacing all defectives (d) by good ones.
- Step 3* If $d > c$, then reject the lot and go for 100% inspection, replace all defectives by good ones and then accept it.

Double-sampling plan

Suppose $N \Rightarrow$ Size of the lot under study

$n_1 \Rightarrow$ Size of the first sample selected

$n_2 \Rightarrow$ Size of the second sample selected

c_1 and $c_2 \Rightarrow$ Acceptance numbers (Number of defective items allowed)

d_1 and $d_2 \Rightarrow$ Number of defective items found in the first and second samples respectively.

With these assumptions, the double sampling plan is as follows:

- Step 1* Select the first sample of “ n_1 ” from the lot and check it. Suppose the sample contain “ d_1 ” defective items.
- Step 2* If $d_1 \leq c_1$, then accept the lot by replacing all defectives (d_1) by good ones.
- Step 3* If $d_1 > c_2$, then reject the lot and go for 100% inspection, replace all defectives by good ones and then accept it.
- Step 4* If $c_1 < d_1 \leq c_2$, then select second sample of size n_2 from the lot and inspect it.
Suppose the sample contain d_2 defective articles.
- Step 5* If $d_1 + d_2 \leq c_2$, then accept the lot by replacing all defectives (d_2) by good ones.
- Step 6* If $d_1 + d_2 > c_2$, then reject the lot and go for 100% inspection, replace all defectives by good ones and then accept it.

18.12 Some basic concepts related to rectifying single sampling plan

In the introduction of lot control, it has already been discussed that the lot quality is represented by “ p .” Rectifying sampling plans always affect this “ p ” quality of the lot, which reaches the customer. The size of the sample (n)

and acceptance number (C) in case of single sampling plan should be known in advance by using Dodge and Romig tables (not discussed in this book) which require certain concepts, which are as follows:

Acceptance quality level (AQL)

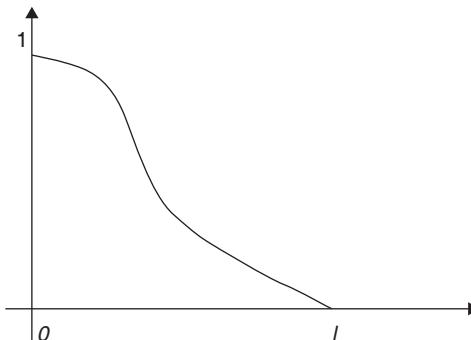
This is the level of quality “ p ” which is acceptable to the customer. It is denoted by notation p_1 .

Lot tolerance proportion defective (LTPD)

This is also known as lot tolerance fraction defective (LTFD). This is the level of quality “ p ” which is not acceptable to the customer. It is denoted by notation p_2 .

Operating characteristic curve (OC curve)

If “ p ” is the quality of the lot, then $P_a(p)$ represents probability of acceptance of the lot of quality “ p .” This $P_a(p)$ is also known as the operating characteristic. In addition, it is clear that $P_a(p)$ changes as the lot quality “ p ” changes. Hence, the graph of “ p ” versus $P_a(p)$ is a curve and this curve is known as the operating characteristic curve. Typically, an OC curve can be shown like Fig. 18.10.



18.10

Producer's risk

The probability of rejection of the lot of quality AQL (p_1) is known as the producer's risk. It is denoted by the notation α .
Thus,

$$\alpha = P(\text{Rejection of lot of quality } p_1) = 1 - P_a(p_1)$$

Customer's risk

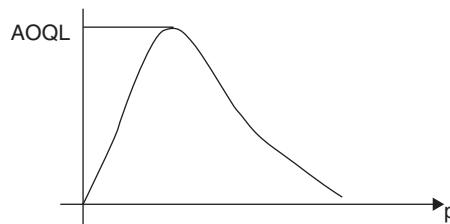
The probability of acceptance of the lot of quality LTPD (p_1) is known as the customer's risk. It is denoted by the notation β .

Thus,

$$\beta = P(\text{Rejection of lot of quality } p_2) = p_a(p_2)$$

Average outgoing quality (AOQ) and AOQL

We have seen that in rectifying sampling plan the lot is accepted after replacing defective items by good ones. This improves the quality of the lot, but some proportion of defective articles remains in the lot after application of sampling plan. This proportion of defective articles remaining in the lot after application of rectifying sampling plan is known as the AOQ. The AOQ also depends on the lot quality "p" and changes as "p" changes. The graph of "p" versus AOQ can be shown in Fig. 18.11.



18.11

From the graph, it is clear that AOQ increases up to certain level for the increase in the value of p after that it decreases. This maximum value of AOQ (as shown in graph), that is maximum proportion of defective items remaining in the lot after application of rectifying sampling plan is called AOQL.

With all the above notations, the expression for AOQ in case of single sampling plan can be easily determined as follows:

If $Y \Rightarrow$ Number of defective items remaining in the lot after application of rectifying sampling plan, then the probability distribution of Y is like Table 18.1.

Table 18.1

Y	Probability
0	$1 - p_a(p_1)$
$(N - n)p$	$p_a(p)$
Total	1

Now,

Expected number of defective items remaining

$$E(Y) = \sum y \cdot p(y) = 0 \times (1 - p_a(p)) + (N - n) \cdot p \times p_a(p) = (N - n) \cdot p \cdot p_a(p)$$

Thus,

$$\text{AOQ} = \frac{(N - n) \cdot p \cdot p_a(p)}{N}$$

18.13 Exercise

1. What is control chart? What are its types? Describe any one.
2. Write short note on “OC curve.”
3. Describe single sampling plan.
4. Describe double sampling plan.
5. Define following terms:
 1. AQL 2. LTPD 3. Producer’s risk 4. Customer’s risk
6. Samples of five ring bobbins each selected from a ring frame for eight shifts have shown following results of count CV%.

Sample no.	1	2	3	4	5	6	7	8
Average	4.16	4.22	4.08	3.96	3.99	4.03	4.02	3.91
Range	0.32	0.38	0.11	0.19	0.12	0.13	0.22	0.11

Draw \bar{X} and R chart for the above data and write conclusion about the state of the process.

7. Following data represents number of defective needles observed in the eight different samples of size 25 each selected from a production line.

Sample no.	1	2	3	4	5	6	7	8
Number of defective	1	0	0	2	4	3	1	1

Draw “ np ” and “ p ” chart for the above data and write conclusions about the state of the process.

8. Following data represents number of end breaks (per 100 spindle hours) observed on a ring frame on eight different days.

Day no.	1	2	3	4	5	6	7	8
Number of end breaks	2	8	3	2	4	6	5	7

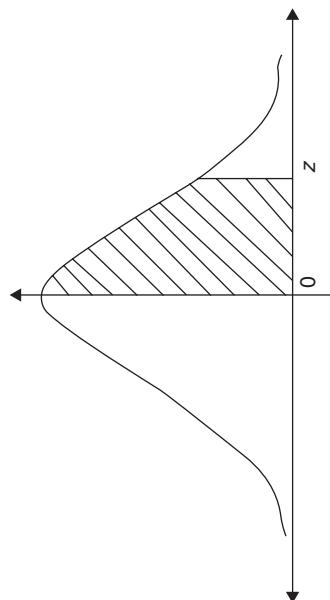
Draw “C” chart for the above data and write conclusion about the state of the process.

9. Samples of five ring bobbins each selected from a ring frame for eight shifts have shown following results of count of yarn.

Sample no.	1	2	3	4	5	6	7	8
Count of yarn	27.5	27.4	25.4	28.5	28.5	28.9	28.0	28.4
	28.5	26.9	26.9	28.0	29.0	29.5	28.5	28.5
	28	26.0	28.0	29.2	28.5	30.0	27.8	28.4
	26.9	28.7	26.7	29.0	28.5	29.4	28.0	28.0
	28.6	29.0	28.2	28.7	28.0	28.9	28.1	28.7

Draw \bar{X} and R chart for the above data and write conclusion about the state of the process.

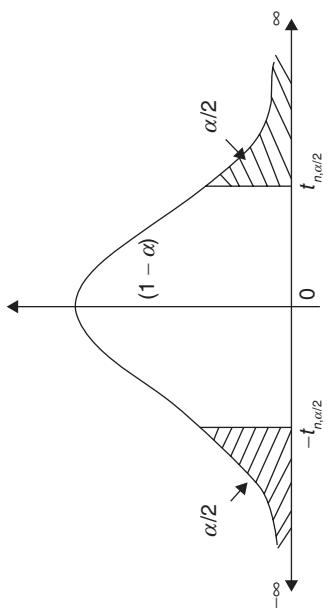
Area under standard normal curve from $Z = 0$ to $Z = z$



Z	0	1	2	3	4	5	6	7	8	9
0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214

1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983

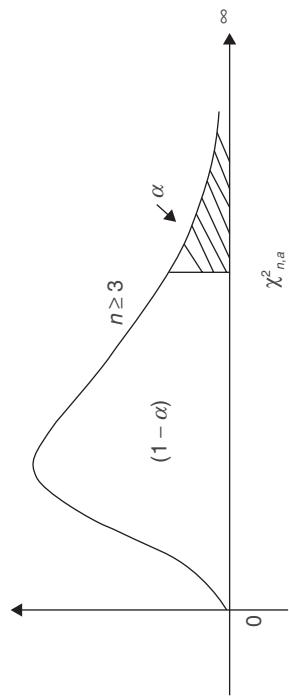
$t_{n,\infty/2}$ values for t-distribution



$df(n)$	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.025	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	25.452	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.205	9.925	31.599
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.177	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.495	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.163	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	2.969	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.841	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.752	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.685	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.634	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.593	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.560	3.055	4.318

13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.533	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.510	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.490	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.473	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.458	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.445	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.433	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.423	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.414	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.405	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.398	2.807	3.768
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.391	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.385	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.379	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.373	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.368	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.364	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.360	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.329	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.045	1.296	1.671	2.000	2.299	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.270	2.617	3.373

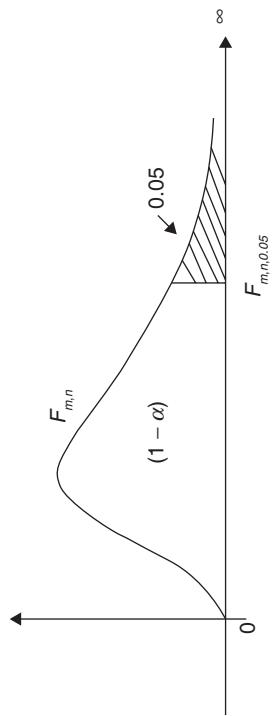
$\chi^2_{n,\infty}$ values for χ^2 -distribution



$df(n)$	0.99	0.975	0.95	0.9	0.8	0.7	0.5	0.2	0.1	0.05	0.025	0.01	0.001
1	0.000	0.001	0.004	0.016	0.064	0.148	0.455	1.642	2.706	3.841	5.024	6.635	10.828
2	0.020	0.051	0.103	0.211	0.446	0.713	1.386	3.219	4.605	5.991	7.378	9.210	13.816
3	0.115	0.216	0.352	0.584	1.005	1.424	2.366	4.642	6.251	7.815	9.348	11.345	16.266
4	0.297	0.484	0.711	1.064	1.649	2.195	3.357	5.989	7.779	9.488	11.143	13.277	18.467
5	0.554	0.831	1.145	1.610	2.343	3.000	4.351	7.289	9.236	11.070	12.833	15.086	20.515
6	0.872	1.237	1.635	2.204	3.070	3.828	5.348	8.558	10.645	12.592	14.449	16.812	22.458
7	1.239	1.690	2.167	2.833	3.822	4.671	6.346	9.803	12.017	14.067	16.013	18.475	24.322
8	1.646	2.180	2.733	3.490	4.594	5.527	7.344	11.030	13.362	15.507	17.535	20.090	26.124
9	2.088	2.700	3.325	4.168	5.380	6.393	8.343	12.242	14.684	16.919	19.023	21.666	27.877
10	2.558	3.247	3.940	4.865	6.179	7.267	9.342	13.442	15.987	18.307	20.483	23.209	29.588
11	3.053	3.816	4.575	5.578	6.989	8.148	10.341	14.631	17.275	19.675	21.920	24.725	31.264
12	3.571	4.404	5.226	6.304	7.807	9.034	11.340	15.812	18.549	21.026	23.337	26.217	32.909
13	4.107	5.009	5.892	7.042	8.634	9.926	12.340	16.985	19.812	22.362	24.736	27.688	34.528

14	4.6660	5.629	6.571	7.790	9.467	10.821	13.339	18.151	21.064	23.685	26.119	29.141	36.123
15	5.229	6.262	7.261	8.547	10.307	11.721	14.339	19.311	22.307	24.996	27.488	30.578	37.697
16	5.812	6.908	7.962	9.312	11.152	12.624	15.338	20.465	23.542	26.296	28.845	32.000	39.252
17	6.408	7.564	8.672	10.085	12.002	13.531	16.338	21.615	24.769	27.587	30.191	33.409	40.790
18	7.015	8.231	9.390	10.865	12.857	14.440	17.338	22.760	25.989	28.869	31.526	34.805	42.312
19	7.633	8.907	10.117	11.651	13.716	15.352	18.338	23.900	27.204	30.144	32.852	36.191	43.820
20	8.260	9.591	10.851	12.443	14.578	16.266	19.337	25.038	28.412	31.410	34.170	37.566	45.315
21	8.897	10.283	11.591	13.240	15.445	17.182	20.337	26.171	29.615	32.671	35.479	38.932	46.797
22	9.542	10.982	12.338	14.041	16.314	18.101	21.337	27.301	30.813	33.924	36.781	40.289	48.268
23	10.196	11.689	13.091	14.848	17.187	19.021	22.337	28.429	32.007	35.172	38.076	41.638	49.728
24	10.856	12.401	13.848	15.659	18.062	19.943	23.337	29.553	33.196	36.415	39.364	42.980	51.179
25	11.524	13.120	14.611	16.473	18.940	20.867	24.337	30.675	34.382	37.652	40.646	44.314	52.620
26	12.198	13.844	15.379	17.292	19.820	21.792	25.336	31.795	35.563	38.885	41.923	45.642	54.052
27	12.879	14.573	16.151	18.114	20.703	22.719	26.336	32.912	36.741	40.113	43.195	46.963	55.476
28	13.565	15.308	16.928	18.939	21.588	23.647	27.336	34.027	37.916	41.337	44.461	48.278	56.892
29	14.256	16.047	17.708	19.768	22.475	24.577	28.336	35.139	39.087	42.557	45.722	49.588	58.301
30	14.953	16.791	18.493	20.599	23.364	25.508	29.336	36.250	40.256	43.773	46.979	50.892	59.703
31	15.655	17.539	19.281	21.434	24.255	26.440	30.336	37.359	41.422	44.985	48.232	52.191	61.098
32	16.362	18.291	20.072	22.271	25.148	27.373	31.336	38.466	42.585	46.194	49.480	53.486	62.487
33	17.074	19.047	20.867	23.110	26.042	28.307	32.336	39.572	43.745	47.400	50.725	54.776	63.870
34	17.789	19.806	21.664	23.952	26.938	29.242	33.336	40.676	44.903	48.602	51.966	56.061	65.247
35	18.509	20.569	22.465	24.797	27.836	30.178	34.336	41.778	46.059	49.802	53.203	57.342	66.619
36	19.233	21.336	23.269	25.643	28.735	31.115	35.336	42.879	47.212	50.998	54.437	58.619	67.985
37	19.960	22.106	24.075	26.492	29.635	32.053	36.336	43.978	48.363	52.192	55.668	59.893	69.346

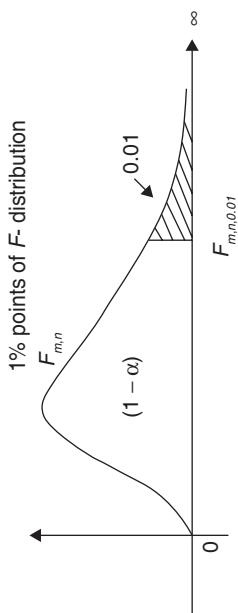
$F_{m,n,0.05}$ values for F -distribution



$n \backslash m$	1	2	3	4	5	6	8	12	24	30
1	161.45	199.50	215.71	224.58	230.16	233.99	238.88	243.91	249.05	250.10
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.46
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.60	2.42	2.38

14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.31
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.25
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.19
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	2.15
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	2.11
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	2.04
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	2.01
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.98
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.20	2.01	1.96
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.94
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.92
26	4.23	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.90
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.13	1.93	1.88
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.12	1.91	1.87
29	4.18	3.33	2.93	2.70	2.55	2.43	2.28	2.10	1.90	1.85
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.84
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.74
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.65
120	3.92	3.07	2.68	2.45	2.29	2.18	2.02	1.83	1.61	1.55

$F_{m,n,0.01}$ values for F -distribution



$n \backslash m$	1	2	3	4	5	6	8	12	24	30
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5981.07	6106.32	6234.63	6260.65
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.47
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.50
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.84
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.38
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	7.23
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.99
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	5.20
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.65
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	4.25
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.94
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.70
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	3.96	3.59	3.51
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.80	3.43	3.35

15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	3.21
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	3.10
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.46	3.08	3.00
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.92
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.84
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.78
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.72
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.67
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.62
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.58
25	7.77	5.57	4.68	4.18	3.85	3.63	3.32	2.99	2.62	2.54
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.50
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.47
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.44
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.41
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.39
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	2.20
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	2.03
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.86

Statistical constants for control charts

Number of observations in Sample, n	Chart for averages			Chart for standard deviations				Factors for control limits				Chart for ranges			
	A	A_1	c_1	Factors for central line	B_1	B_2	B_3	B_4	d_1	d_2	D_1	D_2	D_3	D_4	
2	2.121	3.760	1.880	0.5642	1.7725	0	1.843	0	3.267	1.128	0.8865	0.853	0	3.686	0
3	1.732	2.394	1.023	0.7236	1.3820	0	1.858	0	2.568	1.693	0.5907	0.888	0	4.358	0
4	1.500	1.880	0.729	0.7979	1.2533	0	1.808	0	2.266	2.059	0.4857	0.880	0	4.698	0
5	1.342	1.596	0.577	0.8407	1.1894	0	1.756	0	2.089	2.326	0.4299	0.864	0	4.918	0
6	1.225	1.410	0.483	0.8686	1.1512	0.026	1.711	0.030	1.970	2.534	0.3946	0.848	0	5.078	0
7	1.134	1.277	0.419	0.8882	1.1259	0.105	1.672	0.118	1.882	2.704	0.3698	0.833	0.205	5.203	0.076
8	1.061	1.175	0.373	0.9027	1.1078	0.167	1.638	0.185	1.815	2.847	0.3512	0.820	0.387	5.307	0.136
9	1.000	1.094	0.337	0.9139	1.0942	0.219	1.609	0.239	1.761	2.970	0.3367	0.808	0.546	5.394	0.184
10	0.949	1.028	0.308	0.9227	1.0837	0.262	1.584	0.284	1.716	3.078	0.3249	0.797	0.687	5.469	0.223
11	0.905	0.973	0.285	0.9300	1.0753	0.299	1.561	0.321	1.679	3.173	0.3152	0.787	0.812	5.534	0.256

12	0.866	0.925	0.266	0.9359	1.0684	0.331	1.541	0.354	1.646	3.258	0.3069	0.778	0.924	5.592	0.284	1.716
13	0.832	0.884	0.249	0.9410	1.0627	0.359	1.523	0.382	1.618	3.336	0.2998	0.770	1.026	5.646	0.308	1.692
14	0.802	0.848	0.235	0.9453	0.0579	0.384	1.507	0.406	1.594	3.407	0.2935	0.762	1.121	5.693	0.329	1.671
15	0.775	0.816	0.223	0.9490	1.0537	0.406	1.492	0.428	1.572	3.472	0.2880	0.755	1.207	5.737	0.348	1.652
16	0.750	0.788	0.212	0.9523	1.0501	0.427	1.478	0.448	1.552	3.532	0.2831	0.749	1.285	5.779	0.364	1.636
17	0.728	0.762	0.203	0.9551	1.0470	0.445	1.465	0.466	1.534	3.588	0.2787	0.743	1.359	5.817	0.379	1.621
18	0.707	0.738	0.194	0.9576	1.0442	0.461	1.454	0.482	1.518	3.640	0.2747	0.738	1.426	5.854	0.392	1.608
19	0.688	0.717	0.187	0.9599	1.0418	0.477	1.443	0.497	1.503	3.689	0.2711	0.733	1.490	5.888	0.404	1.596
20	0.671	0.697	0.180	0.9619	1.0376	0.491	1.433	0.510	1.490	3.735	0.2677	0.729	1.548	5.922	0.414	1.586
21	0.655	0.679	0.173	0.9638	1.0396	0.504	1.424	0.523	1.477	3.778	0.2647	0.724	1.606	5.950	0.425	1.575
22	0.640	0.662	0.167	0.9655	1.0358	0.516	1.415	0.534	1.466	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.647	0.162	0.9670	1.0342	0.527	1.407	0.545	1.455	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.632	0.157	0.9684	1.0327	0.538	1.399	0.555	1.445	3.895	0.2567	0.712	1.759	6.031	0.452	1.548
25	0.600	0.619	0.153	0.9696	1.0313	0.548	1.392	0.565	1.435	3.931	0.2544	0.709	1.804	6.058	0.459	1.541

Over 25...

Index

A

absolute measures of variation. *See* measures of variation
acceptance quality level (AQL), 292
algorithm
 for fitting binomial distribution, 147
 for fitting Poisson distribution, 148
alternate hypothesis, 170
AM. *See* Arithmetic mean (AM)
American Society for Quality Control, 278
analysis of variance (ANOVA)
 assignable causes of, 231
 chance causes of, 231
 introduction, 231
 one-way, 232–238
 two-way, 238–249
ANOVA. *See* analysis of variance (ANOVA)
AOQ. *See* average outgoing quality (AOQ)
AOQL, 293–294
AQL. *See* acceptance quality level (AQL)
arithmetic mean (AM)
 computation, for frequency distribution, 26–31
 definition, 25–26
assignable causes of ANOVA, 231
attribute(s)
 and data classification, 9–10
 defined, 6
 Ξ^2 –test for the independence of, 213–215
average. *See also* measure of central tendency
 AM, 25–31
 defined, 25
 median, 31–35
 mode, 35–37
average outgoing quality (AOQ), 293–294

B

bar diagram, 18–19
Bernoulli trial, 140
binomial probability distribution, 140–143
 definition, 140
 derivation of mean and variance for, 141–142
 fitting of, 147–151
 Poisson approximation to, 146–147
 properties of, 140–141
bivariate data, 8, 90
block, defined, 253
Bowley's coefficient of skewness, 82

C

causes of ANOVA, 231
C-chart, 288–290
census survey, 7
central line (CL), 279
central moments, of random variable, 137
central tendency. *See also* average; measure of central tendency
 introduction, 25
 measure of, 25
CF. *See* correction factor (CF)
chance causes of ANOVA, 231
change-of-origin and scale method, 28–31
change-of-origin method, 27–28
characteristic, defined, 5–6
Chi-square probability distribution
 (Ξ^2 distribution), 162–164
CL. *See* central line (CL)
class boundaries, 13
class frequency, 13
classical approach, probability, 122
classification, data. *See* data classification

- class limits, 13
 coefficient of determination, 93–95
 coefficient of range, 73–75
 coefficient of skewness
 Bowley's, 82
 interpretation of, 83–85
 Karl Pearson's, 82
 collection, data. *See* data collection
 completely randomized design (CRD),
 254–255
 conditional probability, law of, 122
 confidence interval. *See* interval estimation
 continuous/discontinuous frequency distribution. *See* grouped frequency distribution table
 continuous probability distribution, 134–136
 continuous random variable, 132
 continuous variable, defined, 7
 control chart, 279–280
 interpretation of, 280
 control limits, 32, 279
 correction factor (CF), 14
 correlation analysis, 90–93
 multiple, 111–114
 correlation coefficients, partial. *See* partial correlation coefficients
 CR. *See* critical region (CR)
 CRD. *See* completely randomized design (CRD)
 critical region (CR), 171–172
 Crosby, Philip, 278
 cumulative frequency distribution table
 graphical representation of, 20–22
 (*See also* ogive curve)
 less than type, 16–17
 more than type, 17–18
 curve, frequency, 20
 customer's risk, 293
- D
- data classification
 according to attribute, 9–10
 according to variables, 10
 data collection
 bivariate data, 8
 census survey, 7
 methods for, 7–8
 multivariate data, 8
 raw statistical data, 7–8
 sample survey, 7
 univariate data, 8
 data variation. *See* dispersion (variation)
 deciles, 45–49
 De' Morgan's law, 122
 derivation
 of mean and variance for Poisson distribution, 144–145
 of mean and variance for the binomial distribution using MGF, 141–142
 design of experiments
 basic concepts and terminology, 253
 basic principles of, 254
 CRD, 254–255
 factorial experiments, 262–273
 introduction, 253
 LSD, 257–262
 RBD, 255–256
 determination
 coefficient of, 93–95
 of normal probability distribution, 159
 deviation
 mean, 61–65
 mean squared, 65
 quartile, 58–61
 standard, 65–72
 diagram
 bar, 18–19
 scatter, 91–92
 Venn, 121
 direct method
 AM computation, 27–31
 and coefficient of determination, 93–94
 standard deviation computation, 67–72
 discontinuous/continuous frequency distribution. *See* grouped frequency distribution table
 discrete frequency distribution. *See* ungrouped frequency distribution table
 discrete probability distribution, 133–134
 discrete random variable, 131–132
 discrete variable, defined, 6
 dispersion (variation)
 error, 231

- introduction, 56
mean deviation, 61–65
quartile deviation, 58–61
range, 56–58
relative measures of, 72–75
standard deviation, 65–72
distribution, probability. *See probability distribution*
distribution table, frequency. *See frequency distribution*
double-sampling plan, 291
- E
- \bar{X} control limits, 279
error
 standard, 170
 variation, 231
estimation
 interval, 222–229
 introduction, 220
 point, 220–221
event, defined, 120
expectation of random variable X, 136
experiment(s)
 defined, 119
 design of (*See design of experiments*)
 factorial (*See factorial experiments*)
experimental material, defined, 253
experimental unit (plot), defined, 253
- F
- factorial experiments, 262–273
 2^2 , 263
 2^3 , 267–268
 2^n , 262
symmetric, 262
Yate's method, 263–267, 268–273
failure and success. *See Bernoulli trial*
finite sample space, defined, 120
Fisher, R. A., 231
fitting of binomial and Poisson probability distributions, 147–151
F-probability distribution, 166–168
frequency curve, 20
frequency distribution, 10–16
- computing AM for, 26–31
cumulative, 16–18
graphical representations of, 18–20
grouped, 12–16
leptokurtic, 85
mesokurtic, 85
negatively skewed, 81–82
platykurtic, 85
positively skewed, 81
skewed, 81–82
symmetric, 80
ungrouped, 10–12
frequency polygon, 19–20
F-test, small sample tests for equality of population variances, 199–203
- G
- graphical determination
 of median, 32–35
 of mode, 36–37
graphical representations of frequency distribution
 bar diagram, 18–19
 cumulative, 20–22 (*See also ogive curve*)
 frequency curve, 20
 frequency polygon, 19–20
 histogram, 19
grouped frequency distribution table, 12–14
 construction of, 14–16
- H
- histogram, 19
hypothesis, testing of. *See testing of hypothesis*
- I
- impossible event, defined, 121
indirect method
 AM computation, 27–31
 and coefficient of determination, 94–95
 standard deviation computation, 68–72
individual, defined, 5
infinite sample space, defined, 120
interpretation, coefficient of skewness, 83–85

interval, confidence. *See* interval estimation
 interval estimation, 222–229

K

Karl Pearson's coefficient
 of correlation, 92–93
 of skewness, 82

kurtosis

description, 85–86
 measures of, 86–88

L

large sample tests (Z-tests), 172–185
 for equality of population means, 177–181
 for the equality of the proportions of two
 different populations, 182–185
 for the population mean, 172–177
 for the population proportion, 181–182

Latin square design (LSD), 257–262

description, 257–259
 procedure and formulae for computation,
 259–262

law of conditional probability, 122

law of independence, 122

LCL. *See* lower control limit (LCL)

left-tailed critical region, 171, 172

leptokurtic frequency distribution, 85

less than type cumulative frequency
 distribution table, 16–17

less than type ogive curve, 21

limits, SQC specification, 281

linear regression analysis, 100–108

of X on Y, 100–101

of Y on X, 101–102

local control, principle of, 254

lot control, 290–291

double-sampling plan, 291

single-sampling plan, 290–291

lot tolerance proportion defective (LTPD), 292

lower control limit (LCL), 279

lower specification limit (LSL), 281

LSD. *See* Latin square design (LSD)LSL. *See* lower specification limit (LSL)

LTPD. *See* lot tolerance proportion
 defective (LTPD)

M

mass production, 1

material, experimental, 253

mathematical analysis

in case of one-way ANOVA, 232–234

in case of two-way ANOVA, 239–241,
 244–246

mean

arithmetic (*See* arithmetic mean (AM))

derivation for binomial distribution,
 141–142

derivation for Poisson distribution,
 144–145

deviation, 61–65

mean squared deviation, 65

measure of central tendency, 25. *See also*
 average

AM (*See* Arithmetic mean (AM))

median, 31–35

mode, 35–37

measure of correlation, 91

measures of kurtosis, 86–88

measures of skewness, 82

measures of variation. *See also* dispersion
 (variation)

mean deviation, 61–65

quartile deviation, 58–61

range, 56–58

relative, 72–75

standard deviation, 65–72

median

computation of, 31–32

definition, 31

graphical determination of, 32–35

mesokurtic frequency distribution, 85

MGF. *See* moment generating function (MGF)

mode

definition, 35

determination of, 35–36

graphical determination of, 36–37

moment generating function (MGF), 137

derivation of mean and variance for the

binomial distribution using, 141–142

moments, of random variable, 137

more than type cumulative frequency

distribution table, 17–18

more than type ogive curve, 21–22
multiple correlation analysis, 111–114
 coefficients, 112
 of one variable with others, 111–112
 partial correlation between any two
 variables, 112
 partial correlation coefficients, 112
multiple correlation coefficients
 definition, 112
 properties of, 114
 for three variables data, 113–114
multiple regression analysis, 114–118
multivariate data
 description, 8, 111
 multiple correlation analysis, 111–114
 multiple regression analysis, 114–118
mutually exclusive events, defined, 121

N

natural tolerance, 279
negative correlation, 91
negatively skewed frequency
 distribution, 81–82
negative skewness, 81–82
normal probability distribution
 definition, 154
 determination of, 159
 procedure of finding probabilities
 in, 159–160
 properties of, 154–157
 standard, 157
normal variable, standard, 157
np-chart, 284–285
null hypothesis, 170

O

OC curve. *See* operating characteristic curve
ogive curve, 20–22
 introduction, 20–21
 less than type, 21
 more than type, 21–22
one-tailed critical region, 171
one-way analysis of variance, 232–238
 description, 232
 mathematical analysis in case of, 232–234

procedure and formulae for computation,
 235–238
operating characteristic curve (OC curve),
 292

P

partial correlation coefficients
 definition, 112
 properties of, 114
 for three variables data, 113–114
partition values
 deciles, 45–49
 introduction, 40
 percentiles, 50–54
 quartiles, 40–45
p-chart, 285–288
pdf. *See* probability density function (pdf)
percentage mean deviation (PMD), 73
percentiles, 50–54
permutation and combination, results
 of, 123–129
platykurtic frequency distribution, 85
plot (experimental unit), defined, 253
PMD. *See* percentage mean deviation (PMD)
pmf. *See* probability mass function (pmf)
point estimation, 220–221
Poisson approximation, to binomial
 probability distribution, 146–147
Poisson probability distribution, 143–146
 definition, 143
 derivation of mean and variance for,
 144–145
 fitting of, 147–151
 properties of, 143–144
polygon, frequency, 19–20
population mean (t-test)
 small sample test for equality of,
 189–194
 small sample tests for, 185–189
population mean (Z-tests), large sample tests
 for, 172–177
population proportion (Z-test)
 large sample tests for, 181–182
 large sample tests for equality of, 182–185
population statistics, 5

- population variance
 F-test, small sample tests for equality of, 199–203
 Ξ^2 -test, small sample tests for, 194–199
- positive correlation, 91
- positively skewed frequency distribution, 81
- positive skewness, 81
- principle of local control, 254
- principle of randomization, 254
- principle of replication, 254
- probability density function (pdf), 134
- probability distribution
 basic concepts, 131–132
 binomial, 140–143
 Chi-square, 162–164
 continuous, 134–136
 discrete, 133–134
 F -, 166–168
 normal, 154–157
 Poisson, 143–146
 properties of, 136–137
 of a random variable, 132–136
 standard, 138
 student's t -, 164–166
- probability mass function (pmf), 133
- probability/probabilities
 basic concepts, 119–121
 classical approach, 122
 finding in normal probability distribution, 159–160
 introduction, 119
 laws of, 122
- process control, 279
- producer's risk, 292
- production
 mass, 1
 process of, 278–279
- properties of regression coefficients, 102–103
- property of regression lines, 103–108
- Q**
- qualitative type characteristic, defined, 6
- quality, definitions of, 278
- quality control, statistical. *See* statistical quality control (SQC)
- quantitative type characteristic, defined, 6
- quartile deviation, 58–61
 coefficient of, 73–75
- quartiles, 40–45
- R**
- random experiment, defined, 119–120
- randomization, principle of, 254
- randomized block design (RBD), 255–256
- random sample, 7
- random variable
 continuous, 132
 definition, 131
 discrete, 131–132
 probability distribution of, 132–136
 properties of, 136–137
- range, 56–58
 coefficient of, 73–75
- rank correlation coefficient, Spearman's, 95–100
- raw moments, of random variable, 137
- raw statistical data, 7–8
- RBD. *See* randomized block design (RBD)
- R-chart, 282–284
- rectifying single-sampling plan, 291–294
- regression analysis, 100–108
 multiple, 114–118
- regression coefficients, properties of, 102–103
- regression lines, property of, 103–108
- relative measures of dispersion, 72–75
- replication, principle of, 254
- response/yield, defined, 253
- right-tailed critical region, 171
- risk
 customer's, 293
 producer's, 292
- S**
- sample, defined, 7
- sample space, defined, 120, 131
- sample survey, 7
- scale method and change-of-origin, 28–31
- scatter diagram, 91–92
- sectors, of textile industry, 2

- single-sampling plan, 290–291
 concepts related to rectifying, 291–294
skewed frequency distribution, 81–82
skewness, 80
 measures of, 82
 negative, 81–82
 positive, 81
small sample tests, 185–215
 for equality of population mean
 (t-test), 189–194
 for equality of population variances
 (F-test), 199–203
population correlation coefficient (t-test),
 test for significance of, 204–205
for the population mean (t-test), 185–189
for the population variance (Σ^2 -test),
 194–199
test for goodness of fit (Σ^2 -test), 205–213
SNV. *See* standard normal variable (SNV)
space, sample, 120, 131
Spearman's rank correlation coefficient,
 95–100
specification limits, SQC, 281
specified tolerance, 281
SQC. *See* statistical quality control (SQC)
squared deviation, 66
 mean, 65
standard deviation, 65–72
 coefficient of, 73–75
 description, 65–67
 direct method computation, 67–72
 indirect method computation, 68–72
standard error, 170
standard normal probability distribution
 definition, 157
 properties of, 158–159
standard normal variable (SNV), 157
standard normal variable, definition, 157
standard probability distribution, 138
statistical data, raw, 7–8
statistical quality control (SQC), 1
 C-chart, 288–290
 control chart, 279–280
 introduction, 278–279
 lot control, 290–291
 np-chart, 284–285
 p-chart, 285–288
process control, 279
purpose of, 2–3
R-chart, 282–284
specification limits, 281
 X chart, 281–282
statistics
 attribute, 6
 characteristic, 5–6
 continuous variable, 7
 discrete variable, 6
 individual, 5
 introduction, 5–7
 population, 5
 qualitative type characteristic, 6
 quantitative type characteristic, 6
 variable, 6
student's t -probability distribution, 164–166
success and failure. *See* Bernoulli trial
sure event, defined, 120
sure experiment, defined, 119
survey
 census, 7
 sample, 7
symmetric factorial experiments, 262
symmetric frequency distribution, 80

T

- tabulation, 9. *See also* data classification
test for goodness of fit (Σ^2 -test), 205–213
test for significance of population correlation
 coefficient (t-test), 204–205
testing of hypothesis
 introduction, 170–172
 large sample tests (Z-tests), 172–185
 small sample tests, 185–215
textile industry
 nature of, 1–2
 sectors of, 2
theoretical mean of random variable X , 136
tolerance
 natural, 279
 specified, 281
 t -probability distribution, student's, 164–166
treatment, defined, 253
trial, Bernoulli, 140
trivariate data, 111. *See also* multivariate data

- t-test
 - for significance of population correlation coefficient, 204–205
 - small sample tests for equality of population mean, 189–194
- 2^2 factorial experiments, 263
- 2^3 factorial experiments, 267–268
- 2^n factorial experiments, 262
- two-tailed critical region, 171
- two-way analysis of variance, 238–249
 - mathematical analysis in case of (with repetition), 244–246
 - mathematical analysis in case of (without repetition), 239–241
 - with “m” observation per cell (with repetition), 244
 - with one observation per cell (without repetition), 238–239
- procedure and formulae for computation, 241–243, 247–249

- U
- UCL. *See* upper control limit (UCL)
- ungrouped frequency distribution table, 10–11
 - construction of, 11–12
- univariate data, 8
- upper control limit (UCL), 279

- V
- values, partition. *See* partition values
- variable
 - and data classification, 10

- defined, 6
- random (*See* random variable)
- standard normal, 157
- variance
 - analysis of (*See* analysis of variance (ANOVA))
 - derivation for binomial distribution, 141–142
 - derivation for Poisson distribution, 144–145
 - of a random variable X, 136
- variation. *See* dispersion (variation)
- Venn diagram, 121

- X
- X chart, 281–282
- Ξ^2 distribution (Chi-square probability distribution), 162–164
- Ξ^2 -test
 - for the independence of the attributes, 213–215
 - small sample tests for the population variance, 194–199

- Y
- Yate’s method, 263–267, 268–273
- yield/response, defined, 253

- Z
- “zero-defects” concept, 278
- Z-tests. *See* large sample tests (Z-tests)