CONTENTS

SYLLABUS

BUSINESS STATISTICS

SECTION-A

1. **Introduction to Statistics :** Frequency Distribution, Graphical Representations.
2. **Measures of Location :** Definition of Central Tendency, Arithmetic Mean (A.M.) Geometric Mean (G.M.) Harmonic Mean (H.M.) Median, Mode, Quartiles, Deciles and Percentiles.
3. **Measures of Dispersion :** Definition, Standard Deviation (S.D.), Mean Deviation (M.D.) Quartile Deviation (Q.D.) Range (R).
4. **Measures of Skewness and Kurtosis :** Moments, Skewness, Kurtosis.

SECTION-B

5. **Correlation and Regressions :** Bivariate Distribution, Correlation—Coefficient, Rank, Multiple and Curvilinear Regressions.
6. **Probability :** Sample Space and Events—Probability—The Axioms of Probability—Some Elementary Theorerms—Conditional Probability—Baye's Theorem.
7. **Probability Distribution :** Random Variables—Discrete and Continuous—Distribution—Distribution Function.
8. **Some Probability Distribution :** Distribution—Binomial—Poisson and Normal Distribution—Related Properties.

SECTION-C

9. **Sampling Theory :** Population and Samples—Sampling Distribution of Mean (Known and Unknown) Proportions, Sums and Differences.
10. **Estimation :** Point Estimation—Interval Estimation—Bayesian Estimation.

SECTION-D

11. **Test of Hypothesis :** Means and Proportions—Hypothesis Concerning One and Two Means—Type I and Type II Errors, One-Tail, Two-Tail Tests, Test of Significance—Student's T-Test, X^2 —Estimation of Proportions.
12. **Analysis of Variances :** ANOVA Table, Randomised Block Design.
13. **Stochastic Process :** Definition, Markov Process, and Markov Chain, Chapman-Kolmogorov Equations, Steady-State and First Entrance Probabilities.

SECTION-A

CHAPTER 1 INTRODUCTION TO STATISTICS

NOTES

★ STRUCTURE ★

- Introduction
- Frequency Distribution
- Graphical Representations
- Summary
- Problems

INTRODUCTION

Statistics is a branch of scientific method comprising of collection, presentation, analysis and interpretation of data which are obtained by measuring some characteristics. However, the word **statistics** is used in both singular and plural forms. For example, statistics is now taught in various disciplines—this is singular sense, whereas the statistics of industrial production of India for the last five years—this is plural sense.

Numerical figures which are the effect of a large number of causes only comprise statistical data. A single train accident is not a statistical data, but the total number of train accidents during a year constitutes the statistical data. A table of values of a mathematical function *viz.*, $\cos x$, $\log x$ etc. will never be called statistical data. Statistics deals with quantitative data only. However, methods have been devised to transfer qualitative data to quantitative.

Statistics is a wide subject and find a very suitable place in various aspects of life. Statistical tools are used in agriculture, biology, behavioural science, geology, physics, psychology, medicines, engineering etc. In business and commerce the statistical tools *viz.*, demand analysis, forecasting, inventory control, network scheduling etc. are needed for proper organisation. For manufacturing industry statistical quality control and sampling theory are two important statistical tools.

Success in Operations Research in military operation and in other phases is because of statistics. The following steps are carried out for any statistical experiment.

(a) Collection of data. The problem which has been formulated requires data for investigation which are collected by any physical methods and techniques.

NOTES

(b) Tabulation. The data which we have collected can be considered as raw data and we do not get any insight of the problem unless we go for tabulation, *i.e.*, represent the data in simple tabular form by diagrams, bar charts, pie charts etc. Construct the frequency distribution.

(c) Statistical inference. Apply the statistical methods on the tabulated data and draw conclusions about the unknown properties of the population from which the data have been drawn.

FREQUENCY DISTRIBUTION

The frequency distribution is a tabulation of data which are obtained from measurement or observation or experiment, arranged in ascending or descending order.

Let us consider the resistance of 50 units of certain electrical product:

3.0	3.4	4.1	4.1	4.3	2.7	3.5	3.7	3.4	3.4
3.8	4.2	3.1	3.9	3.1	4.1	2.8	3.7	4.4	3.5
3.5	3.4	3.7	3.7	2.8	4.3	3.8	3.4	4.1	3.0
4.4	4.1	4.1	3.6	3.4	2.7	3.6	3.0	3.4	4.3
3.8	3.2	4.2	3.9	4.2	3.4	2.9	4.4	3.5	3.9

The following table shows the simple frequency distribution of these data with all data and their frequencies of occurrences.

Resistance	*Tabulation*	*Frequency*
2.7	||	2
2.8	||	2
2.9	|	1
3.0	|||	3
3.1	||	2
3.2	|	1
3.4	𝍸 |||	8
3.5	||||	4
3.6	||	2
3.7	||||	4
3.8	|||	3
3.9	|||	3
4.1	𝍸 |	6
4.2	|||	3
4.3	|||	3
4.4	|||	3

NOTES

When there is a large amount of highly variable data, the above frequency distribution can become large. The data may be grouped into classes to provide a better presentation. But there is no rule about the number of classes to be taken for the given data. In the above, the lowest data is 2.7 and the highest data is 4.4. Let us consider six classes of equal width and the following table is called grouped frequency distribution.

Class	*Frequency*
2.7 — 2.9	5
3.0 — 3.2	6
3.3 — 3.5	12
3.6 — 3.8	9
3.9 — 4.1	9
4.2 — 4.4	9

In the above table, the left side value of each class, *i.e.*, 2.7, 3.0, ..., 4.2 is called lower class limit and the right side of each class, *i.e.*, 2.9, 3.2, ..., 4.4 is called upper class limit. The width of each class is 0.2.

In this example the classes are not continuous. To make it continuous we add 0.05 to the upper limits and subtract 0.05 from the lower limits where 0.05 + 0.05 = 0.1 is the difference between the previous upper limit and the next lower limit between any two consecutive classes. In this case, the class limits are called class boundaries. The middle value of any class is called class mark. So we obtain the following table:

Class boundaries	*Class mark*	*Frequency*
2.65 — 2.95	2.8	5
2.95 — 3.25	3.1	6
3.25 — 3.55	3.4	12
3.55 — 3.85	3.7	9
3.85 — 4.15	4.0	9
4.15 — 4.45	4.3	9

GRAPHICAL REPRESENTATIONS

There are mainly four graphical representation of frequency distribution, *viz*, (*a*) Histogram, (*b*) Frequency polygon, (*c*) Bar chart, (*d*) Ogive.

(*a*) Histogram. In this graph the sides of the column represent the upper and lower class boundaries and their heights are proportional to the respective frequencies. Consider the following grouped frequency distribution.

NOTES

Weight (lbs.)	No. of persons
100 — 110	5
110 — 120	8
120 — 130	15
130 — 140	7
140 — 150	3
150 — 160	2

The histogram is drawn as follows (Fig. 1.1)

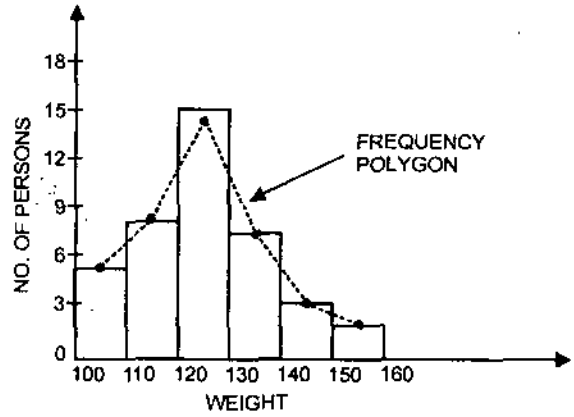


Fig. 1.1

Note. If the width of the class intervals are not same then calculate 'Relative frequency density' (*rfd*) for all classes as follows:

$$rfd = \frac{\text{Frequency}}{\text{Total frequency} \times \text{class width}}$$

Then take *rfd* on *y*-axis and class intervals on *x*-axis to draw histogram.

(*b*) Frequency polygon. Consider the mid-points with a height proportional to class frequency in the histogram. If these points are joined by straight lines then the resultant graph is called frequency polygon.

(*c*) Bar chart. A bar chart is a graphical representation of the frequency distribution in which the bars are centered at the mid-points of the cells. The heights of the bars are proportional to the respective class frequencies.

If a single attribute is presented then it is called simple bar chart (Fig. 1.2). When more than one attribute is presented then it is called multiple bar chart.

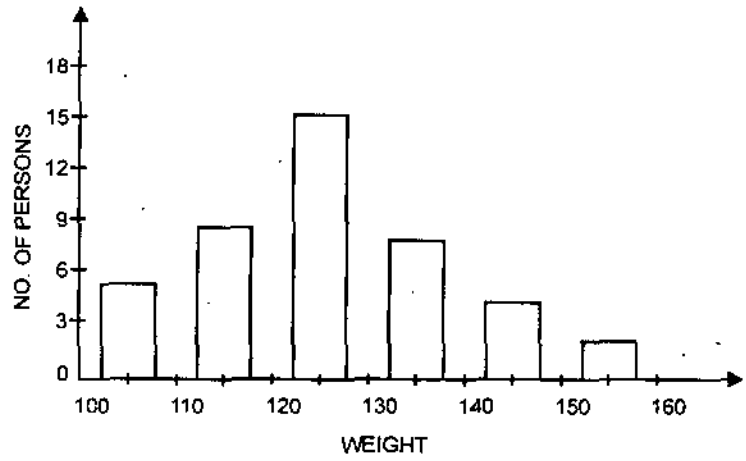


Fig. 1.2

(d) Ogive. There are two types of cumulative frequency distribution—less than type and more than type which are illustrated in the following table.

Daily Wages (in Rs.)	*No. of Workers (Frequency)*	*Cumulative frequency*	
		Less than	*More than*
22	6	6	120
27	12	18	114
32	14	32	102
37	16	48	88
42	19	67	72
47	22	89	53
52	31	120	31

Such a cumulative frequency distributions may be represented graphically and the graph is known as ogive because of its similarity to the ogee curve of the architect and the dam designer. The intersection point of the two curves give the median of the distribution.

For grouped frequency distribution, the 'less than' ogive must be plotted against the upper class boundary and not against the class mark, whereas for 'more than' ogive the cumulative frequency must be plotted against lower class boundary.

In this book if the type of the cumulative distribution is not mentioned it is to be understood that it is less than type.

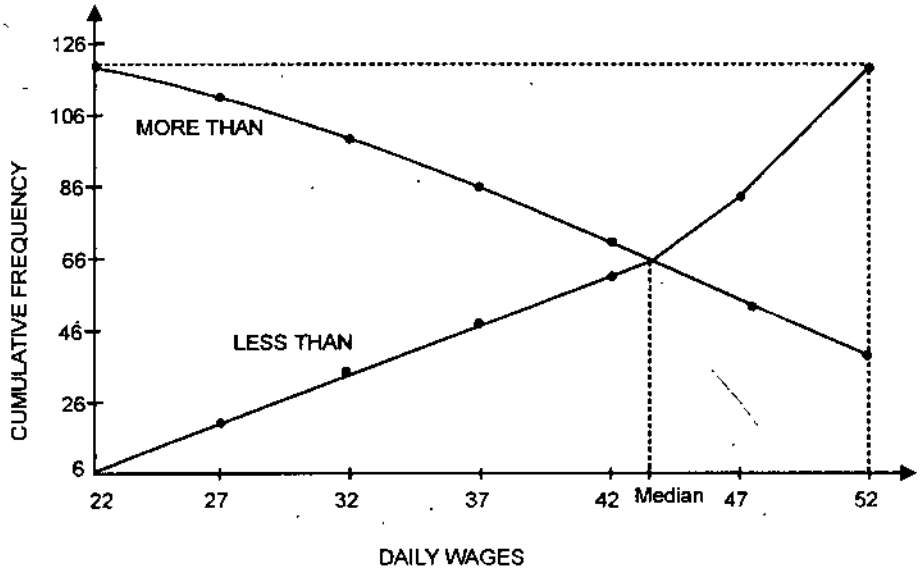


Fig. 1.3

SUMMARY

- Statistics is a branch of scientific method comprising of collection, presentation, analysis and interpretation of data which are obtained by measuring some characteristics.

NOTES

- Statistics is a wide subject and find a very suitable place in various aspects of life. Statistical tools are used in agriculture, biology, behavioural science, geology, physics, psychology, medicines, engineering etc.
- The frequency distribution is a tabulation of data which are obtained from measurement or observation or experiment, arranged in ascending or descending order.

PROBLEMS

1. From the following prepare a frequency distribution table having class intervals of 5:

78	78	93	82	84	92	97	85
84	82	97	78	75	87	84	89
90	91	94	95	93	99	88	82
82	78	96	75	91	93	93	92
88	90	91	78	88	78	91	91

   Also draw the histogram.

2. A machine shop produces steel pins. The width of 30 pins (in mm) was checked after machining and data was recorded as follows:

9.61	9.54	9.51	9.58	9.54	9.52
9.51	9.55	9.57	9.60	9.61	9.58
9.51	9.54	9.54	9.52	9.57	9.58
9.57	9.53	9.55	9.52	9.61	9.50
9.61	9.56	9.61	9.54	9.51	9.55

   Construct the grouped frequency distribution by taking six classes.

3. For a machine making resistors the successive 40 items were checked and the following resistances in ohms were noted:

152	151	150	154	151	156	153	152
150	157	157	154	152	155	155	151
156	157	151	155	155	151	155	156
155	155	150	153	154	157	152	155
157	151	153	151	155	156	154	156

   Prepare the simple frequency distribution table. Draw the ogives.

4. Exhibit the absolute and cumulative frequency in respect of the formations given below:

Height (in cm)	*No. of boys*
Less than 150	2
Less than 155	5
Less than 160	11
Less than 165	16
Less than 170	19
Less than 175	8

5. Exhibit the absolute and cumulative frequency in respect of the formations given below:

Weight (in kg)	*No. of girls*
More than 44	2
More than 46	3
More than 48	6
More than 50	18
More than 52	12
More than 54	5
More than 56	6

6. Consider the following distribution:

Class	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency	8	20	40	22	6	4

Draw the (*a*) Histogram, (*b*) frequency polygon, (*c*) ogives.

7. The hourly wage of employees (Rs. '00) in an organisation is given below:

Hourly wage	*No. of employees*
4.00 — 5.75	31
5.75 — 6.85	20
6.85 — 8.11	15
8.11 — 10.00	14
10.00 — 13.25	9

Draw the Histogram.

NOTES

ANSWERS

4.

Class	*Frequency*	*Cumulative frequency (less than)*
0 — 150	2	2
150 — 155	5	7
155 — 160	11	18
160 — 165	16	34
165 — 170	19	53
170 — 175	8	61

5.

Class	*Frequency*	*Cumulative frequency (less than)*
44 — 46	2	2
46 — 48	3	5
48 — 50	6	11
50 — 52	18	29
52 — 54	12	41
54 — 56	5	46
56 —	6	52

CHAPTER 2 MEASURES OF LOCATION

NOTES

★ STRUCTURE ★

- ❑ Definition of Central Tendency/Location
- ❑ Arithmetic Mean (A.M.)
- ❑ Geometric Mean (G.M.)
- ❑ Harmonic Mean (H.M.)
- ❑ Median
- ❑ Mode
- ❑ Quartiles, Deciles and Percentiles
- ❑ Summary
- ❑ Problems

DEFINITION OF CENTRAL TENDENCY/LOCATION

For quantitative data it is observed that there is a tendency of the data to be distributed about a central value which is a typical value and is called a measure of central tendency. It is also called a measure of location because it gives the position of the distribution on the axis of the variable.

There are three commonly used measures of central tendency, *viz.* Mean, Median and Mode. The mean again may be of three types, *viz.* Arithmetic Mean (A.M.), Geometric Mean (G.M.) and Harmonic Mean (H.M.). Below we shall discuss these different measures.

ARITHMETIC MEAN (A.M.)

The arithmetic mean is simply called 'Average'. For the observations x_1, x_2, ..., x_n the A.M. is defined as

$$\bar{x} = \text{A.M.} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For simple frequency distribution,

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

NOTES

For grouped frequency distribution, x_i is taken as class mark.

The A.M. is sometimes called as 'Average' or 'Sample Mean'.

Example 1. *Find the mean of the following data.*

No. of Matches 2	*3*	*4*	*6*	*7*	*8*		
No. of goals	*0*	*5*	*2*	*4*	*6*	*10*	

Solution. Here N = Total no. of matches = 30

Also $\Sigma x f = 0 + 15 + 8 + 24 + 42 + 80 = 169$

$$\text{Hence mean} = \frac{169}{30} = 5.633.$$

Note. Let $\bar{x}_1$ be the mean of n_1 observations and $\bar{x}_2$ be the mean of n_2 observations then the combined mean $\bar{x}$ is computed as follows:

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

GEOMETRIC MEAN (G.M.)

The geometric mean of the observations $x_1, x_2, \ldots\ldots, x_n$ is defined as

$$\text{G.M.} = (x_1 . x_2 \ldots\ldots\ldots x_n)^{1/n}$$

For simple frequency distribution,

$$\text{G.M.} = \left(x_1^{f_1} . x_2^{f_2} \ldots\ldots\ldots x_n^{f_n}\right)^{1/N}, \quad N = \sum_{i=1}^{n} f_i$$

For grouped frequency distribution, x_i is taken as class mark.

Note. 1. The logarithm of the G.M. of a variate is the A.M. of its logarithm.

2. G.M. = 0 iff a single variate value is zero.

3. G.M. is not used if any variate value is negative.

Example 2. *Find the G.M. of the following distribution :*

Humidity reading	*No. of days*
60	3
62	2
64	4
68	2
70	4

NOTES

Solution. Here N = No. of days = 15.

x	f	$\log x$	$f \log x$
60	3	1.77815	5.33445
62	2	1.79239	3.58478
64	4	1.80618	7.22472
68	2	1.83251	3.66502
70	4	1.84510	7.38040
Σ	15	–	27.18937

Then, $\log \text{ G.M.} = \frac{1}{N} \Sigma f_i \log x_i = \frac{1}{15}(27.18937) = 1.81262$

$\Rightarrow$ G.M. $= 64.9561 \approx 64.96$

HARMONIC MEAN (H.M.)

The reciprocal of the H.M. of a variate is the A.M. of its reciprocal.

For the observations $x_1, x_2, \ldots\ldots, x_n$

$$\text{H.M.} = \frac{n}{\Sigma(1/x_i)}$$

For simple frequency distribution,

$$\text{H.M.} = \frac{N}{\Sigma(f_i/x_i)}, \quad N = \Sigma f_i$$

For grouped frequency distribution x_i is taken as class mark.

Note. A.M. $\geq$ G.M. $\geq$ H.M.

Example 3. *Suppose a train moves 100 km with a speed of 40 km/hr, then 150 km with a speed of 50 km/hr and next 135 km with a speed of 45 km/hr. Calculate the average speed.*

Solution. To get average speed we require harmonic mean of 40,50 and 45 with 100, 150 and 135 as the respective frequency or weights.

$$\text{H.M.} = \frac{100 + 150 + 135}{100 \times \frac{1}{40} + 150 \times \frac{1}{50} + 135 \times \frac{1}{45}}$$

$$= \frac{385}{8.5} = 45.29$$

Hence the average speed per hour is 45.29 km.

Note. In the case of grouped frequency distributions with open end class at one extremity or at both the extremities, the A.M., G.M. and H.M. cannot be computed unless we make some plausible assumptions.

NOTES

MEDIAN

(*a*) For the observations $x_1, x_2, \ldots\ldots, x_n$ the median is the middle value if the number of observations is odd and have been arranged in ascending or descending order of magnitude. For even number of observations the median is taken as the average of two middle values after they are arranged in ascending or descending order of magnitude.

e.g., For the data, 10, 17, 15, 25, 18, let us arrange them in ascending order as 10, 15, 17, 18 and 25. Here the middle value is 17. Hence the median is 17.

Consider another sets of data, 21, 40, 19, 28, 33 30. Let us arrange them in ascending order as 19, 21, 28, 30, 33 and 40. There are two middle values 28 and 30. So the median is (28 + 30)/2 *i.e.*, 29.

(*b*) For simple frequency distribution the median is detained by using less than cumulative frequency distribution. Here the median is that value of the variable for which the cumulative frequency is just greater than $\frac{1}{2}.N$ where N = total frequency.

e.g. consider

x	*f*	*Cumulative frequency*
10	2	2
15	5	7
20	11	18
25	7	25
30	5	30

Here N = 30, $\frac{N}{2} = 15$. So the cumulative frequency just greater than 15 is 18 and the corresponding variable value is 20. Then the median is 20.

(*c*) For grouped frequency distribution, the median is obtained by the following:

$$\text{Median} = L + \frac{h\left(\frac{N}{2} - C\right)}{f}$$

where, L = Lower limit or boundary of the median class.

h = Width of the class interval

f = Frequency of the median class.

N = Total frequency

C = Cumulative frequency of the class preceding the median class.

Example 4. *Find the median of the following data :*

Marks	Less than 40	41–50	51–60	61–70	71–80	81 and above
No. of students	10	20	15	25	10	20

Solution.

Marks	*No. of students (f)*	*Cumulative frequency*
Less than 40	10	10
41 — 50	20	30
51 — 60	15	45
61 — 70	25	70
71 — 80	10	80
81 and above	20	100

Here N = 100, $\frac{N}{2} = 50$, C = 45, the median class is 61 – 70.

L = 61, h = 9, f = 25

$$\therefore \quad \text{Median} = 61 + \frac{9(50-45)}{25} = 62.8.$$

Note. In case of unequal class-intervals median is sometimes preferred to A.M.

MODE

Mode is the value of a variable which occurs most frequently in a set of observations.

For simple frequency distribution the mode is the value of a variable corresponding to the maximum frequency. For grouped frequency distribution the mode is obtained as follows.

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}.h$$

where,

L = Lower limit or boundary of the modal class

h = Width of the modal class

f_1 = Frequency of the modal class

f_0 = Frequency of the class preceeding the modal class

f_2 = Frequency of the class succeeding the modal class.

Note. 1. If $2f_1 - f_0 - f_2 = 0$, then the mode is obtained as follows :

$$\text{Mode} = L + \frac{f_1 - f_0}{|f_1 - f_0| + |f_1 - f_2|} \times h.$$

2. If the maximum frequency is repeated then the above technique is not practicable.

3. For symmetrical distribution,

$$\text{Mean} = \text{Median} = \text{Mode.} \qquad (\text{Mean} = \bar{x}).$$

However, for moderately skewed distribution there is an empirical relationship due to Karl Pearson *i.e.*,

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median}).$$

NOTES

Example 5 : *Find the mode of the problem as given in example 4.*

Solution. Here maximum frequency is 25, so the modal class is 61–70.

$$L = 61,\ h = 9,\ f_1 = 25,\ f_0 = 15,\ f_2 = 10$$

$$\therefore \quad \text{Mode} = 61 + \frac{25-15}{2 \times 25 - 15 - 10} \times 9 = 64.6.$$

QUARTILES, DECILES AND PERCENTILES

As the median divides an array into two parts, the quartiles divide the array into four parts, the deciles divide it into ten parts and the percentiles divide it into one hundred parts.

The first quartile/the lower quartile denoted by Q_1 is computed as follows.

$$Q_1 = L + \frac{\frac{N}{4} - C}{f} \times h$$

where, L = Lower limit of the class containing Q_1

f = Frequency of the class containing the Q_1

h = Width of the class containing the Q_1

C = Cumulative frequency of the class preceding the calss containing Q_1.

Here the cumulative frequency just greater than $\frac{N}{4}$ is the class containing Q_1.

The second quartile is the median.

The third quartile/the upper quartile denoted by Q_3 is computed as follows:

$$Q_3 = L + \frac{\frac{3N}{4} - C}{f} \times h$$

where, L = Lower limit of the class containing Q_3

f = Frequency of the class containing Q_3

h = Width of the class containing Q_3

C = Cumulative frequency of the class preceding the class containing Q_3.

Here the cumulative frequency just greater than $\frac{3N}{4}$ is the class containing Q_3.

The k-th decile denoted by D_k is computed as follows:

C = Cumulative frequency

$$D_k = L + \frac{\frac{k \times N}{10} - C}{f} \times h, \qquad (k = 1, 2, \ldots\ldots, 9)$$

NOTES

where, L = Lower limit of the class containing D_k

f = Frequency of the class containing D_k

h = Width of the class containing D_k

C = Cumulative frequency of the class preceding the class containing D_k,

Here the cumulative frequency just greater than $\frac{k \times N}{10}$ is the class containing D_k (k = 1, 2, ...,9).

The k-th percentile denoted by P_k is computed as follows:

$$P_k = L + \frac{\frac{k \times N}{100} - C}{f} \times h, \qquad (k = 1,2,.....,99)$$

where, L = Lower limit of the class containing P_k

f = Frequency of the class containing P_k

h = Width of the class containing P_k

C = Cumulative frequency of the class preceding the class containing P_k.

Here the cumulative frequency just greater than $\frac{k \times N}{100}$ is the class containing P_k (k = 1,2, ...,99).

Example 6. *Determine (a) Q_1, (b) Q_3, (c) D_5 (d) P_{80} from the following distribution:*

Class	10–15	15–20	20–25	25–30	30–35	35–40
Frequency	20	15	31	22	10	2

Solution.

Class	*Frequency*	*Cumulative frequency*
10–15	20	20
15–20	15	35
20–25	31	66
25–30	22	88
30–35	10	98
35–40	2	100

(*a*) N = 100, $\frac{N}{4}$ = 25. The cumulative frequency just greater than 25 is 35. So the class 15–20 contains Q_1. L = 15, f = 15, h = 5, C = 20. Therefore,

$$Q_1 = 15 + \frac{(25 - 20)}{15} \times 5 = 16.67.$$

(*b*) Here $\frac{3N}{4}$ = 75. The cumulative frequency just greater than 75 is 88. So the class 25 – 30 contains Q_3.

L = 25, f = 22, h = 5, C = 66. Therefore,

NOTES

$$Q_3 = 25 + \frac{(75-66)}{22} \times 5 = 27.05.$$

(*c*) Here $\frac{5N}{10} = 50$. The cumulative frequency just greater than 50 is 66. So the class 20 – 25 contains D_5.

L = 20, $f = 31$, $h = 5$, C = 35. Therefore,

$$D_5 = 20 + \frac{(50-35)}{31} \times 5 = 22.42.$$

(*d*) Here $\frac{80N}{100} = 80$. The cumulative frequency just greater than 80 is 88. So the class 25 – 30 contains P_{80}.

L = 25, $f = 22$, $h = 5$, C = 66. Therefore,

$$P_{80} = 25 + \frac{(80-66)}{22} \times 5 = 28.18.$$

Example 7. *A given machine is assumed to depreciate 30% in value in the first year, 35% in the second year and 80% per annum for the next three years, each percentage being calculated on the diminishing value. Calculate the average annual rate of depreciation.*

Solution. The proportional rates of depreciation for the 5 years are 0.30, 0.35, 0.80, 0.80 and 0.80 respectively.

Let r be the average proportional rate of depreciation.

Then $1 - r$ is the G.M. of $(1 - 0.30)$, $(1 - 0.35)$, $(1 - 0.80)$, $(1 - 0.80)$ and $(1 - 0.80)$ *i.e.*, 0.70, 0.65, 0.20, 0.20 and 0.20.

$$\therefore \quad 1-r = (0.70 \times 0.65 \times 0.20 \times 0.20 \times 0.20)^{1/5}$$

$$\Rightarrow \quad \log(1-r) = \frac{1}{5}\log(0.00364)$$

$$\Rightarrow \quad 1-r = 0.3253$$

$$\Rightarrow \quad r = 0.6747$$

Hence the average annual rate of depreciation is 67.47%.

SUMMARY

- For quantitative data it is observed that there is a tendency of the data to be distributed about a central value which is a typical value and is called a measure of central tendency. It is also called a measure of location because it gives the position of the distribution on the axis of the variable.
- The arithmetic mean is simply called 'Average'. For the observations $x_1, x_2, \ldots\ldots, x_n$ the A.M. is defined as

$$\bar{x} \approx \text{A.M.} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- The geometric mean of the observations $x_1, x_2, \ldots\ldots, x_n$ is defined as

$$\text{G.M.} = (x_1 . x_2 \ldots\ldots\ldots x_n)^{1/n}$$

- The reciprocal of the H.M. of a variate is the A.M. of its reciprocal. For the observations $x_1, x_2, \ldots\ldots, x_n$

$$\text{H.M.} = \frac{n}{\Sigma(1/x_i)}$$

- For the observations $x_1, x_2, \ldots\ldots x_n$ the median is the middle value if the number of observations is odd and have been arranged in ascending or descending order of magnitude. For even number of observations the median is taken as the average of two middle values after they are arranged in ascending or descending order of magnitude.
- Mode is the value of a variable which occurs most frequently in a set of observations.
- As the median divides an array into two parts, the quartiles divide the array into four parts, the deciles divide it into ten parts and the percentiles divide it into one hundred parts.

NOTES

PROBLEMS

1. Suppose a train moves 5 hrs at a speed of 40 km/hr, then 3 hrs at a speed of 45 km/hr and next 5 hrs at a speed of 60 km/hr. Calculate the average speed.
2. A factory has 4 sections, the no. of workers in the different sections being 50, 100, 60 and 150. The average wages per worker are Rs. 100, Rs. 120, Rs. 130 and Rs. 110 respectively, calculate the average wages for all the workers together.
3. Compute the A.M., G.M. and H.M. from the following:

Class	15–19	20–24	25–29	30–34	34–39	40–44
Frequency	13	32	4	42	58	51

4. The mean of 15 items is 34. It was found out, later on, that the two items 48 and 32 were wrongly copied as 84 and 23 respectively. Find out the correct mean.
5. The number of telephone calls received daily in a marketing department of a company for 200 days are given below :

No. of Calls	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45	45–50
Frequency	7	15	24	31	42	30	26	15	10

Calculate the mean, median and mode of the telephone calls.

6. Consider the following distribution of humidity readings in a certain place for 60 days.

Humidity	10–19	20–29	30–39	40–49	50–59	60–69
Frequency	3	15	15	20	5	2

Calculate (*i*) Q_1, (*ii*) Q_3, (*iii*) D_5, (*iv*) P_{90}.

NOTES

7. From the following distribution of marks, calculate

   (*i*) mean, (*ii*) median, (*iii*) mode, (*iv*) D_5 (*v*) P_{85}

Marks (more than)	0	10	20	30	40	50	60	70	80
No. of students	150	140	100	80	80	70	30	14	0

8. Find the missing frequencies in the following distribution when it is known that A.M. = 59.5

Daily wages (Rs.)	10–20	20–35	35–60	60–90	90–105	105–120	120–150
Percentage of wage earners	10	f_1	30	15	f_2	5	5

Where total of percentage of wage earners is 100.

9. An analysis of production rejects resulted in the following figures:

No. of rejects per operator	*No. of operators*	*No. of rejects per operator*	*No. of operators*
21–25	5	41–45	15
26–30	15	46–50	12
31–35	28	51–55	3
36–40	42		

Compute mean, median and mode of the production rejects.

10. An aeroplane travels distances of d_1, d_2 and d_3 km at speeds V_1, V_2 and V_3 km per hour respectively. Show that the average speed is given by v, where

$$\frac{d_1 + d_2 + d_3}{v} = \frac{d_1}{v_1} + \frac{d_2}{v_2} + \frac{d_3}{v_3}.$$

ANSWERS

1. 48.85 km/hr.
2. Rs. 102.22
3. A.M. = 33.33, GM = 32.22, H.M. = 30.95.
4. Corrected mean = 32.2
5. Mean = 27.875, Median = 27.74, Mode = 27.39
6. $Q_1 = 27.5$, $Q_3 = 37.5$, $D_{10} = 45.5$, $P_{90} = 51.5$
7. Mean = 39.27, Median = 45, Mode = 56.46 (using empirical relation)
   $D_5 = 45$, $P_{85} = 64.69$
8. $f_1 = 20$, $f_2 = 15$.
9. Mean = 36.96, Median = 36.64, Mode = 37.207.

CHAPTER 3 MEASURES OF DISPERSION

NOTES

★ STRUCTURE ★

- Definition of Dispersion
- Standard Deviation (S.D.)
- Mean Deviation (M.D.)
- Quartile Deviation (Q.D.)
- Range (R)
- Summary
- Problems

DEFINITION OF DISPERSION

After getting the idea of central value of the quantitative data as discussed in the previous chapter, it is observed that in some cases the values are very close around the central value and in other cases the values are scattered a little wide around the central value. The measure which gives the idea of the amount of scattering of the data around the central value is called the measure of dispersion.

There are four commonly used measures of dispersion *viz.* Standard Deviation (S.D.), Mean Deviation (M.D.), Quartile Deviation (Q.D.) and Range (R). Below we shall discuss these different measures.

STANDARD DEVIATION (S.D.)

Standard deviation (and variance) is a relative measure of the dispersion of a set of data–the larger the standard deviation, the more spread out the data.

If $x_1, x_2, \ldots\ldots, x_n$ be a set of n observations forming a population, then its standard deviation is given by

$$\text{S.D.} = \sigma = \left[\frac{1}{n}.\,\Sigma(x_i - \bar{x})^2\right]^{1/2}, \quad \text{where } \bar{x} = \frac{1}{n}\Sigma x_i$$

$$= \left[\frac{1}{n}\Sigma x^2 - (\bar{x})^2\right]^{1/2}$$

For simple frequency distribution,

$$\text{S.D.} = \sigma = \left[\frac{1}{N}\Sigma f_i (x_i - \bar{x})^2\right]^{1/2}, \text{ where } N = \Sigma f_i$$

For grouped frequency distribution, x_i is taken as class mark.

NOTES

Note. 1. The square of S.D. is known as variance.

2. When $\bar{x}$ is a real number, then the following alternative formula can be used to calculate S.D.

$$\text{S.D.} = \left[\frac{1}{N}\Sigma f x_i^2 - (\bar{x})^2\right]^{1/2}, \quad N = \Sigma f_i$$

3. For comparing the variability of two distributions the coefficient of variation (C.V.) is calculated as follows :

$$\text{C.V.} = \frac{\text{S.D.}}{\bar{x}} \times 100$$

The distribution with less C.V. is said to be more uniform or consistent or less variable or more homogeneous.

4. If the values of x and f are large then for simplicity step-deviation method can be employed in which the deviations of the given values of x from any arbitrary point A is taken.

Then $$\sigma^2 = \frac{1}{N}\Sigma f d^2 - \left(\frac{1}{N}\Sigma f d\right)^2 \text{ where, } d = x - A$$

which shows that the variance and S.D. of a distribution is independent of change of origin.

5. In case of grouped frequency distribution, if h be the width of the class-interval then we can use change of scale also *i.e.*, by taking $u = \frac{x - A}{h}$, we obtain

$$\sigma^2 = \left[\frac{1}{N}\Sigma f u^2 - \left(\frac{1}{N}\Sigma f u\right)^2\right] \times h^2$$

6. **(Combined Variance).** If $\bar{x}_1$ and σ_1 be the mean and S.D. of a group of observations of size n_1 whereas $\bar{x}_2$ and σ_2 be the mean and S.D. of another group of observations of size n_2, then the variance of the combined groups is given by :

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

where $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$ and $\bar{x}$ = Mean of the combined groups.

7. S.D. is independent of origin but not of scale.

Example 1. *A group of 100 items have a mean of 60 and a S.D. of 7. If the mean and S.D. of 60 of these items be 51 and 5.2 respectively, find the S.D. of the other 40 items.*

Solution. Consider the following:

		$\bar{x}$	S.D.
n	100	60	12
n_1	60	51	5.2
n_2	40	m	s

From the combined mean we obtain,

$$60 = \frac{60 \times 51 + 40 \times m}{100}$$

$$\Rightarrow \quad 6000 - 3060 = 40m$$

$$\Rightarrow \quad m = 294/4 = 73.5$$

From the combined variance we obtain,

$$144 = \frac{40 \times s^2 + 60 \times (5.2)^2 + 40 \times (73.5 - 60)^2 + 60 \times (60 - 51)^2}{100}$$

$$\Rightarrow \quad 14400 = 40s^2 + 13772.4$$

$$\Rightarrow \quad s^2 = 627.6/40 = 15.69$$

$$\Rightarrow \quad s = 3.961.$$

NOTES

MEAN DEVIATION (M.D.)

If $x_1, x_2, \ldots\ldots, x_n$ be the n observations then M.D. is defined as

$$\text{M.D.} = \frac{1}{n} \Sigma|x - \text{A}|$$

where A may be mean $(\bar{x})$, or median or mode.

For simple frequency distribution, M.D. is defined as

$$\text{M.D.} = \frac{1}{\text{N}} \Sigma f\,|x - \text{A}|, \quad \text{N} = \Sigma f$$

where A may be mean $(\bar{x})$ or median or mode.

For grouped frequency distribution, x will be taken as class-mark.

Sometimes M.D. about mean $(\bar{x})$ is called simply M.D.

Note. 1. M.D. taken about median is the least, compared to M.D. taken about mean $(\bar{x})$ or mode.

2. It is not a very accurate measure of dispersion particularly when it is calculated from mode.

3. S.D. $\geq$ M.D.

QUARTILE DEVIATION (Q.D.)

It is a location based measure of dispersion and is defined as follows:

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

where Q_1 is lower quartile and Q_3 is upper quartile. This Q.D. is also known as semi inter-quartile range, whereas the quantity $Q_3 - Q_1$ is called inter-quartile range.

Note. 1. For comparative studies of variability of two distributions a relative measure is given as follows :

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

NOTES

2. If there is two much variability between two distributions then the following measure is calculated :

$$\text{Coefficient of variation} = \frac{\text{Q.D.}}{\text{Median}} \times 100.$$

3. In the case of unequal class-intervals Q.D. is sometimes preferred to S.D.

RANGE (R)

It is the simplest measure for dispersion. For the observations $x_1, x_2, \ldots\ldots, x_n$ it is defined as

$$R = (\text{largest value}) - (\text{smallest value})$$

For frequency distribution also, the same definition applies and frequencies do not taken into account.

Note. Of all the measures of dispersion, the S.D. is generally the one with the least sampling fluctuation. However, when the data contain a few extreme values widely different from the majority of the values, S.D. should not be used – Q.D. is the appropriate measure.

Example 2. *Find the mean deviation about the mean and variance for the following :*

x	*Frequency*
47.5	7
48.1	17
45.9	46
44.0	44
40.7	54

Solution. The calculations are shown in the following table :

x	f	xf	$f\|x-\bar{x}\|$	$f(x-\bar{x})^2$
47.5	7	332.5	24.36	84.77
48.1	17	817.7	69.36	282.99
45.9	46	2111.4	86.48	162.58
44.0	44	1936.0	0.88	0.02
40.7	54	2197.8	179.28	595.21
Σ	168	7395.4	360.36	1125.57

$$\text{Mean} = \frac{\Sigma xf}{\Sigma f} = \frac{7395.4}{168} = 44.02$$

$$\text{Mean deviation about mean} = \frac{\Sigma f|x-\bar{x}|}{\Sigma f}$$

$$= \frac{360.36}{168}$$

$$= 2.145$$

$$\text{Variance} = \frac{\Sigma f(x-\bar{x})^2}{\Sigma f}$$

$$= \frac{1125.57}{168} = 6.6998$$

NOTES

Example 3. *Calculate S.D., Q.D., and M.D. of the following data.*

Data class	0 – 4	5 – 9	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34
Frequency	4	5	10	8	7	9	7

Solution.

Class	*Class boundaries*	*Class mark (x)*	*Frequency (f)*	*fx*	fx^2	$f\|x-\bar{x}\|$	*Cumulative frequency*
0 – 4	(–0.5) – 4.5	2	4	8	16	65.6	4
5 – 9	4.5 – 9.5	7	5	35	245	57.0	9
10 – 14	9.5 – 14.5	12	10	120	1440	64.0	19
15 – 19	14.5 – 19.5	17	8	136	2312	11.2	27
20 – 24	19.5 – 24.5	22	7	154	3388	25.2	34
25 – 29	24.5 – 29.5	27	9	243	6561	77.4	43
30 – 34	29.5 – 34.5	32	7	224	7168	95.2	50
		Σ	50	920	21130	395.6	

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{920}{50} = 18.4$$

$$\text{S.D} = \left[\frac{1}{N}\Sigma fx^2 - (\bar{x})^2\right]^{1/2}$$

$$= \left[\frac{1}{50}(21130) - (18.4)^2\right]^{1/2}$$

$$= [84.04]^{1/2} = 9.167$$

Here $\frac{N}{4} = 12.5$, then the cumulative frequency just greater than 12.5 is 19, so 9.5 – 14.5 is the class containing Q_1. L = 9.5, h = 4, f = 10, C = 9.

$$\therefore \quad Q_1 = 9.5 + \frac{(12.5-9)}{10} \times 4 = 10.9$$

Again $\frac{3N}{4} = 37.5$, then the cumulative frequency just greater than 37.5 is 43, so 24.5 – 29.5 is the class containing Q_3. L = 24.5, h = 4, f = 9, C = 34

$$\therefore \quad Q_3 = 24.5 + \frac{(37.5-34)}{9} \times 4 = 26.06$$

Then

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{26.06 - 10.9}{2} = 7.58$$

NOTES

$$\text{M.D.} = \frac{1}{N}\Sigma f|x-\bar{x}| = \frac{395.6}{50} = 7.912.$$

Example 4. *Find out the S.D. of height (in cm) of 10 persons given below : 175, 167, 165, 171, 162, 165, 168, 170, 169 and 166*

Solution. Calculations using step deviation method given below :

Here $n = 10$

x	175	167	165	171	162	165	168	170	169	166	Total
$d = x-165$	10	2	0	6	–3	0	3	5	4	1	28
d^2	100	4	0	36	9	0	9	25	16	1	200

$$\text{S.D.} = \left[\frac{1}{n}\Sigma d^2 - \left(\frac{1}{n}\Sigma d\right)^2\right]^{1/2}$$

$$= \left[\frac{200}{10} - \left(\frac{28}{10}\right)^2\right]^{1/2} = [12.16]^{1/2} = 3.487.$$

Example 5. *For a set of 10 observations the A.M. and S.D. were calculated and were found to be 16 and 3.5 respectively. It was later found on serutiny that the last observation of the set should be 25 instead of 15. Calculate the correct A.M. and S.D.*

Solution. Let the observations be $x_1, x_2, \ldots\ldots, x_{10}$

Corrected $\Sigma x = 10 \times 16 - 15 + 25 = 170$

Hence the corrected A.M. $= \frac{\Sigma x}{10} = \frac{170}{10} = 17$

Now corrected $\Sigma x^2 = 10\left[(16)^2 + (3.5)^2\right] - (15)^2 + (25)^2$

$= 3082.5$

Hence corrected S.D. $= \left[\frac{1}{n}\Sigma x^2 - (\bar{x})^2\right]^{1/2}$

$$= \left[\frac{3082.5}{10} - (17)^2\right]^{1/2}$$

$$= [19.25]^{1/2} = 4.388$$

Example 6. *Calculate mean and standard deviation from the following data:*

Marks	*Frequency*	*Marks*	*Frequency*
More than 80	14	More than 40	280
More than 70	44	More than 30	420
More than 60	86	More than 20	480
More than 50	155	More than 10	550

NOTES

Solution.

Value (more than)	*Class*	*Cum. freq.*	*Frequency (f)*	*Class mark (x)*	$u=\frac{x-45}{10}$	*fu*	fu^2
80	80 – 90	14	14	85	4	56	224
70	70 – 80	44	30	75	3	90	270
60	60 – 70	86	42	65	2	84	168
50	50 – 60	155	69	55	1	69	69
40	40 – 50	280	125	45	0	0	0
30	30 – 40	420	140	35	–1	–140	140
20	20 – 30	480	60	25	–2	– 120	240
10	10 – 20	550	70	15	–3	– 210	630
			Σ:N=550			– 171	1741

$$\text{Mean } (\bar{x}) = 45-\frac{171}{550}.10 = 41.89$$

$$\text{S.D. } (\sigma) = 10\left\{\frac{1741}{550}-\left(\frac{-171}{550}\right)^2\right\}^{1/2}$$

$$= 17.52.$$

Example 7. *A factory produces two types of electric bulbs A and B. In an experiment relating to their life, the following results were obtained.*

Length of life (in hours)	*No. of bulbs*	
	A	*B*
500 – 700	5	4
700 – 900	11	30
900 – 1100	26	12
1100 – 1300	10	8
1300 – 1500	8	6

Find which type of bulb is less variable in length of life.

Solution.

Class	*Class marks(m)*	$u=\frac{m-1000}{100}$	*Bulb A*			*Bulb B*		
			f_1	f_1u	f_1u^2	f_2	f_2u	f_2u^2
500–700	600	–4	5	–20	80	4	–16	64
700–900	800	–2	11	–22	44	30	–60	120
900–1100	1000	0	26	0	0	12	0	0
1100–1300	1200	2	10	20	40	8	16	32
1300–1500	1400	4	8	32	128	6	24	96
		Total	60	10	292	60	–36	312

NOTES

Bulb A

$$\bar{x}_A = A + \frac{h\,\Sigma f_1 u}{N_1} = 1000 + \frac{(100)(10)}{60} = 1016.67$$

$$\sigma_A = h\left[\frac{\Sigma f_1 u^2}{N_1} - \left(\frac{\Sigma f_1 u}{N_1}\right)^2\right]^{1/2}$$

$$= 100\left[\frac{292}{60} - \left(\frac{10}{60}\right)^2\right]^{1/2} = 219.97$$

$$C.V.(A) = \frac{\sigma_A}{\bar{x}_A} \times 100 = \frac{219.97}{1016.67} \times 100 = 21.64$$

Bulb B

$$\bar{x}_B = A + \frac{h\,\Sigma f_2 u}{N_2} = 1000 + \frac{(100)(-36)}{60} = 940$$

$$\sigma_B = h\left[\frac{\Sigma f_2 u^2}{N_2} - \left(\frac{\Sigma f_2 u}{N_2}\right)^2\right]^{1/2}$$

$$= 100\left[\frac{312}{60} - \left(-\frac{36}{60}\right)^2\right]^{1/2} = 220$$

$$C.V.(B) = \frac{\sigma_B}{\bar{x}_B} \times 100 = \frac{220}{940} \times 100 = 23.40$$

Since C.V. (A) < C.V. (B)

⇒ Bulb A is less variable in length of life.

SUMMARY

- The measure which gives the idea of the amount of scattering of the data around the central value is called the measure of dispersion.
- Standard deviation (and variance) is a relative measure of the dispersion of a set of data—the larger the standard deviation, the more spread out the data.
- If $x_1, x_2, \ldots\ldots, x_n$ be the n observations then M.D. is defined as

$$M.D. = \frac{1}{n}\Sigma|x - A|$$

- It is a location based measure of dispersion and is defined as follows:

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

- It is the simplest measure for dispersion. For the observations $x_1, x_2, \ldots\ldots, x_n$ it is defined as

$$R = (\text{largest value}) - (\text{smallest value})$$

NOTES

PROBLEMS

1. Find the M.D. and variance of the following distribution.

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

2. Find the M.D. and S.D. for the following data:

Score	4–5	6–7	8–9	10–11	12–13	14–15
Frequency	4	10	20	15	8	3

3. With reference to the following data verify that the M.D. is the least when deviations are measured about the median instead of mean.

Class	5–9	10–14	15–19	20–24	25–29	30–34
Frequency	21	19	13	8	16	23

4. Calculate the mean age and the variance of the ages of the members of a society from the following table:

Age at nearest birthday	30–34	35–39	40–44	45–49	50–54	55–59	60–64
No. of members	3	9	18	25	26	15	4

5. Calculate the S.D. and M.D. about both mean and median for the first n integers.
6. In a series of 5 observations the values of mean and variance are 4.4 and 8.24. If three observations are 1, 2 and 6, find the other two.
7. Two batsman A and B made the following scores in a series of cricket matches:

A	14	13	26	53	17	29	79	36	84	49
B	37	22	56	52	14	10	37	48	20	4

Who is more consistent player?

8. From the data given below, find which series is more uniform.

Class	*Sereis A*	*Series B*
0 – 10	7	5
10 – 20	6	8
20 – 30	15	12
30 – 40	12	15
40 – 50	10	10

9. Calculate (*i*) Q.D., and (*ii*) S.D. of wages from the following data:

Weekly wages (Rs.)	35–36	36–37	37–38	38–39	39–40	40–41	41–42
No. of persons	14	20	42	54	45	18	7

NOTES

10. Measurements of the lengths in metres of 50 iron rods are distributed as follows:

Class	*Frequency*	*Class*	*Frequency*
2.35 – 2.45	1	2.75 – 2.85	11
2.45 – 2.55	4	2.85 – 2.95	10
2.55 – 2.65	7	2.95 – 3.05	2
2.65 – 2.75	15		

Calculate (*i*) Q.D. and (*ii*) M.D. of the above data (*iii*) Coefficient of Q.D.

ANSWERS

1. M.D. = 1.773, Variance = 4.432

2. M.D. = 2.03, S.D. = 2.4756

4. Mean = 48.15, Variance = 49.43

5. S.D. $= \sqrt{(n^2-1)/12}$, M.D. about mean = M.D. about median

$$= \frac{n^2-1}{4n}, \text{ when } n = \text{odd integer}$$

$$= \frac{n}{4}, \text{ when } n = \text{even integer}$$

6. 4 and 9

7. C.V. (A) = 61.1% C.V. (B) = 58.47%, Batsman B is more consistent.

8. Series B. C.V. (A) = 47.05, C.V. (B) = 36.06

9. (*i*) Q.D. = 1.03, (*ii*) S.D. = 1.46

10. (*i*) Q.D. = 0.096, (*ii*) M.D. = 0.113, (*iii*) 0.035.

CHAPTER 4 MEASURES OF SKEWNESS AND KURTOSIS

NOTES

★ STRUCTURE ★

- Moments
- Skewness
- Kurtosis
- Summary
- Problems

MOMENTS

For n observations $x_1, x_2, \ldots\ldots, x_n$ and an arbitrary constant A, the rth moment about A is defined as,

$$\mu_r' = \frac{1}{n}\sum_i (x_i - A)^r, \quad r = 0, 1, 2, \ldots\ldots$$

For simple frequency distribution, the r th moment about A is defined as

$$\mu_r' = \frac{1}{N}\sum_i f_i (x_i - A)^r, \quad r = 0, 1, 2, \ldots\ldots, N = \Sigma f_i$$

For grouped frequency distribution, x_i will be taken as class mark.

These moments, μ_r', are also known as 'raw moments'.

When $A = \bar{x}$, then these moments are called 'central moments' and we denote as follows:

$$\mu_r = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^r, \quad r = 0, 1, 2, \ldots\ldots, N = \Sigma f_i$$

For $r = 0$, $\mu_0 = \frac{1}{N}\Sigma f = 1$

For $r = 1$, $\mu_1 = \frac{1}{N}\Sigma f(x - \bar{x})$

$$= \frac{1}{N}\Sigma fx - \frac{1}{N}\Sigma f\bar{x}$$

$$= \bar{x} - \bar{x}\frac{1}{N}\Sigma f$$

$$= \bar{x} - \bar{x} = 0$$

NOTES

For $r = 2$, $\mu_2 = \frac{1}{N}\Sigma f (x - \bar{x})^2$ which is called variance.

The third and fourth central moments *i.e.*, μ_3 and μ_4 are used to measure skewness and kurtosis which have been given in the following sections.

The important relations between the central and raw moments are as follows:

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_2\,\mu'_1 + 2(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\mu'_3\,\mu'_1 + 6\,\mu'_2\,(\mu'_1)^2 - (\mu'_1)^4$$

$$\bar{x} = \mu'_1 + A.$$

Note. If the above moments are called as sample moments, then we have to use the notation m'_r instead of μ'_r. See the method of moments in Chapter 10A.

SKEWNESS

Skewness is a measure of symmetry of the shape of frequency distribution, *i.e.*, it reveals the dispersal of value on either side of an average is symmetrical or not.

There are four mathematical measures of relative skewness.

(*a*) Karl Pearson's Coefficient of Skewness

$$SK = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$$

If mode is ill-defined, then we take

$$SK = \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$$

(*b*) Bowley's Coefficient of Skewness

This is based on quartiles and median and is defined as

$$SK = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

This formula is useful when the mode is ill-defined or the distribution has open end classes or unequal class-intervals.

(*c*) Coefficient of skewness based on central moments

Using second and third central moments, the coefficient of skewness is defined as (due to Karl Pearson)

$$\beta_1 = SK = \frac{\mu_3^2}{\mu_2^3}$$

Also $\gamma_1 = \sqrt{\beta_1}$ which is due to R.A. Fishar.

Note. 1. If SK is positive then the frequency distribution is called positively skewed.
If SK is negative then the frequency distribution is called negatively skewed.
If SK is zero, then the frequency distribution is symmetric.

2. There is no theoretical limit to this measure.

KURTOSIS

This is a measure of peakness of a distribution and is defined in terms of second and fourth central moments as follows (due to Karl Pearson) :

$$\beta_2 = \frac{\mu_4}{\mu_2^2},$$

When $\beta_2 = 3$, the distribution is called **Mesokurtic,**

$\beta_2 < 3$, the distribution is called **Platykurtic,**

and $\beta_2 > 3$, the distribution is called **Leptokurtic.**

Another notation due to R.A. Fishar is $\gamma_2 = \beta_2 - 3$ which is also called excess of kurtosis.

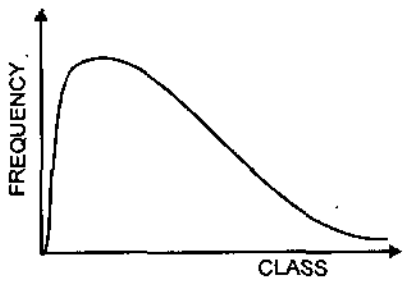


Fig. 4.1 Positively skewed

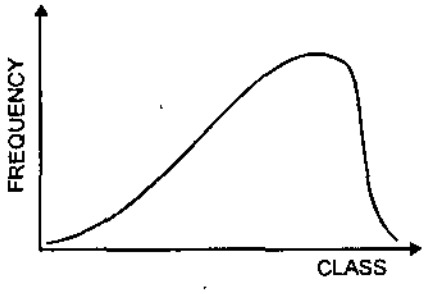


Fig. 4.2 Negatively skewed

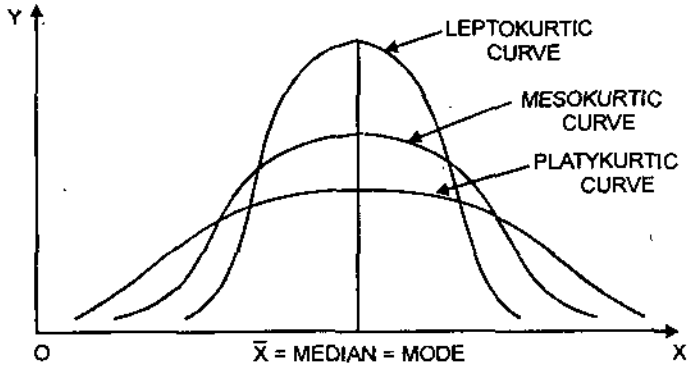


Fig. 4.3 Nature of kurtosis

Example 1. *Calculate the Karl Pearson's coefficient of skewness of the following frequency distribution.*

Class	383–387	388–392	393–397	398–402	403–407
Frequency	8	10	15	17	8

NOTES

Solution.

Class mark (x)	*Frequency* (f)	$d = x - 395$	fd	fd^2
385	8	– 10	– 80	800
390	10	– 5	– 50	250
395	15	0	0	0
400	17	5	85	425
405	8	10	80	800
Σ	58	0	35	2275

$$\therefore \quad \bar{x} = 395 + \frac{\Sigma fd}{\Sigma f} = 395 + \frac{35}{58} = 395.603$$

$$\text{S.D.} = \left[\frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2\right]^{1/2}$$

$$= \left[\frac{2275}{58} - \left(\frac{35}{58}\right)^2\right]^{1/2}$$

$$= [38.864]^{1/2} = 6.234$$

To calculate mode we require class boundaries. Since the maximum frequency is 17, this implies that 397.5 – 402.5 is the modal class.

$$\therefore \quad \text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

$$= 397.5 + \frac{(17-15)}{34-15-8} \times 4$$

$$= 398.227$$

$$\therefore \quad \text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$$

$$= \frac{395.603 - 398.227}{6.234}$$

$$= -0.421$$

which indicate the given distribution is negatively skewed.

Example 2. *The first four moments of a distribution about the value 3 of a variable are 1, 5, 14 and 46.*

Find the mean, variance and comment on the nature of the distribution.

Solution. Given A = 3, $\mu'_1 = 1$, $\mu'_2 = 5$, $\mu'_3 = 14$, $\mu'_4 = 46$.

$$\bar{x} = \mu'_1 + A = 1 + 3 = 4$$

$$\mu_2 = \text{Variance} = \mu'_2 - (\mu'_1)^2 = 5 - 1 = 4$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$$

$$= 14 - 3(5)(1) + 2(1)^3 = 1$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4$$

$$= 46 - 4(14) + 6(5) - 3 = 17$$

Now $$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{1}{64}, \quad \gamma_1 = \sqrt{\beta_1} = \frac{1}{8}$$

Since $\beta_1 > 0 \Rightarrow$ The distribution is positively skewed.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{17}{16} = 1.0625, \qquad \gamma_2 = \beta_2 - 3 = -1.9375$$

Since $\beta_2 < 3 \Rightarrow$ The distribution is platykurtic

Example 3. *For a distribution, Bowley's coefficient of skewness is – 0.65, Q_1 = 15.28 and median = 25.2. What is its coefficient of quartile deviation ?*

Solution. Bowley's coefficient of skewness is

$$SK = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

$$\Rightarrow \quad -0.65 = \frac{Q_3 + 15.28 - 2(25.2)}{Q_3 - 15.28}$$

$$\Rightarrow \quad Q_3 = 27.30$$

Then, coefficient of quartile deviation $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

$$= \frac{27.30 - 15.28}{27.30 + 15.28}$$

$$= 0.28.$$

SUMMARY

- For n observations $x_1, x_2, \ldots\ldots, x_n$ and an arbitrary constant A, the rth moment about A is defined as,

$$\mu_r' = \frac{1}{n}\sum_i (x_i - A)^r, \quad r = 0, 1, 2, \ldots\ldots$$

- Skewness is a measure of symmetry of the shape of frequency distribution, *i.e.*, it reveals the dispersal of value on either side of an average is symmetrical or not.
- Kurtosis is a measure of peakness of a distribution and is defined in terms of second and fourth central moments.

PROBLEMS

1. The mean, median and the coefficient of variation of 100 observations are found to be 60, 56 and 30 respectively. Find the coefficient of skewness of the above system of 100 observations.

NOTES

2. Compute the Karl Pearson's coefficient of skewness of the following distribution:

Class	*Frequency*	*Class*	*Frequency*
2.7 – 2.9	2	4.2 – 4.4	113
3.0 – 3.2	16	4.5 – 4.7	71
3.3 – 3.5	46	4.8 – 5.0	22
3.6 – 3.8	88	5.1 – 5.3	4
3.9 – 4.1	138		

3. Compute mean, variance, β_1 and β_2 if the first four moments about a value 5 of a variable are given as 2, 20, 38 and 52.

4. The first three moments of a distribution about the value 3 are –1, 10, –28. Find the values of mean, standard deviation and the moment measure of skewness.

5. For a distribution the mean is 9, standard deviation is 5, $\sqrt{\beta_1} = 1$ and $\beta_2 = 3$. Obtain the first four moments about the origin *i.e.*, zero.

6. Find the appropriate measure of skewness from the following data :

Value	Less than 10	10–20	20–30	30–40	40–50	50 and above
Frequency	5	9	16	7	6	7

7. Find the skewness and kurtosis of the following distribution by central moments and comment on the type.

Class	0–10	10–20	20–30	30–40	40–50
Frequency	10	20	40	20	10

8. Find the C.V. of a frequency distribution given that its mean is 100, mode = 120 and Karl Pearson's coefficient of skewness = – 0.2.

9. Compare the skewness of two frequency distribution whose moments about the origin are as follows :

Distribution	μ'_1	μ'_2	μ'_3
A	2	5	14
B	2	5	1014

10. Calculate the coefficient of skewness and kurtosis of the following data :

Class	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20
Frequency	4	10	6	12	8

ANSWERS

1. 0.67
2. 0.176
3. $\bar{x} = 7$, $\mu_2 = 16$, $\beta_1 = 1.06$, $\beta_2 = 0.70$
4. Mean = 2, S.D. = 3, $\beta_1 = 0$
5. $\mu'_1 = 9$, $\mu'_2 = 106$, $\mu'_3 = 1529$, $\mu'_4 = 25086$
6. Bowley's coefficient of skewness is suitable and = 0.24
7. $\beta_1 = 0$, $\beta_2 = 2.5$, symmetrical and platykurtic
8. 100
9. A is symmetric, B is positively skew
10. $\beta_1 = 0.038$, $\beta_2 = 1.806$.

CHAPTER 5 CORRELATION AND REGRESSION

NOTES

★ STRUCTURE ★

- Bivariate Distribution
- Coefficient of Correlation
- *Regression Equations*
- Rank Correlation
- Correlation of Bivariate Frequency Distribution
- Multiple Regression
- Curvilinear Regression
- Summary
- Problems

BIVARIATE DISTRIBUTION

When a distribution has two variables then it is called bivariate. For example, if we measure the income and expenditure of a certain group of persons-one variable will measure income and the other variable will measure expenditure and the values will form the bivariate distribution.

There may be any correlation between the variables, *i.e.*, the change in one variable gives a specific change in the other variable. For example, if the increase (or decrease) of one variable results the increase (or decrease) of the other variable, then the correlation is said to be positive. If the increase (or decrease) leads to decrease (or increase) then the correlation is said to be negative.

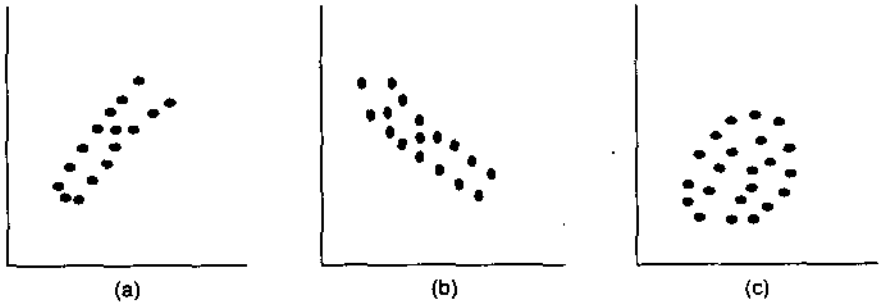


Fig 5.1 Scatter diagram

NOTES

The simplest way to represent the bivariate data in a diagram known as scatter diagram. For the bivariate distribution (x, y), the values (x_i, y_i), $i = 1, 2, \ldots\ldots, n$ of the variables are plotted in the xy-plane which is known as scatter diagram. This gives an idea about the correlation of the two variables.

COEFFICIENT OF CORRELATION

2A. Karl Pearson has given a coefficient to measure the degree of linear relationship between two variables which is known as coefficient of correlation (or, correlation coefficient).

For a bivariate distribution (x, y) the coefficient of correlation denoted by r_{xy} and is defined as

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x . \sigma_y}$$

where cov (x, y) = Covariance between x and y

σ_x = S.D. of x

σ_y = S.D. of y

For the values (x_i, y_i), $i = 1, 2, \ldots\ldots, n$ of a bivariate distribution,

$$r_{xy} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\Sigma\ (y_i - \bar{y})^2}}$$

$$= \frac{\Sigma(x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\ \sqrt{\Sigma(y_i - \bar{y})^2}}$$

2B. Limitations of r_{xy}

1. The coefficient of correlation can be used as a measure of linear relationship between two variables. In case of non-linear or any other relationship the coefficient of correlation does not provide any measure at all. So the inspection of scatter diagram is essential.
2. Correlation must be used to the data drawn from the same source. If distinct sources are used then the two variables may show correlation but in each source they may be uncorrelated.
3. For two variables with a positive or negative correlation it does not necessarily mean that there exists causal relationship. There may be the effect of some other variables in both of them. On elimination of this effect it may be found that the net correlation is nil.

2C. Properties

1. The coefficient of correlation is independent of the origin and scale of reference.
2. $-1 \le r_{xy} \le 1$

Proof. Let $u_i = \frac{x_i - \bar{x}}{\sigma_x}$ and $v_i = \frac{y_i - \bar{y}}{\sigma_y}$

Then $\frac{1}{n}\Sigma u_i^2 = 1,\ \frac{1}{n}\Sigma v_i^2 = 1,\ \frac{1}{n}\Sigma u_i\ v_i = r_{xy}$

Now $$\frac{1}{n}\Sigma(u_i - v_i)^2 \geq 0$$

$$\Rightarrow \quad \frac{1}{n}\Sigma u_i^2 + \frac{1}{n}\Sigma v_i^2 - \frac{2}{n}\Sigma u_i\, v_i \geq 0$$

$$\Rightarrow \quad 2(1 - r_{xy}) \geq 0$$

$$\Rightarrow \quad r_{xy} \leq 1$$

Also, $$\frac{1}{n}\Sigma(u_i + v_i)^2 \geq 0$$

$$\Rightarrow \quad \frac{1}{n}\Sigma u_i^2 + \frac{1}{n}\Sigma v_i^2 + \frac{2}{n}.\,\Sigma u_i\, v_i \geq 0$$

$$\Rightarrow \quad 2(1 + r_{xy}) \geq 0$$

$$\Rightarrow \quad r_{xy} \geq -1$$

Combining we obtain $-1 \leq r_{xy} \leq 1$.

3. Two independent variables are uncorrelated (*i.e.* $r_{xy} = 0$) but the converse is not always true.
4. If r_{xy} is the correlation coefficient in a sample of n pairs of observations, then the standard error of r_{xy} is defined by

$$\text{S.E.}(r_{xy}) = \frac{1 - r_{xy}^2}{\sqrt{n}}$$

5. **Probable error** of the correlation coefficient is defined by

$$\text{P.E.}(r_{xy}) = 0.6745\,\frac{(1 - r_{xy}^2)}{\sqrt{n}}$$

2D. By Step Deviation Method

Let $d_x = x - A$, $d_y = y - B$ which are the deviations and A, B are assumed values, then

$$r_{xy} = \frac{\Sigma d_x\, d_y - \dfrac{\Sigma d_x .\, \Sigma d_y}{n}}{\sqrt{\left[\Sigma d_x^2 - \dfrac{(\Sigma d_x)^2}{n}\right]} \cdot \sqrt{\left[\Sigma d_y^2 - \dfrac{(\Sigma d_y)^2}{n}\right]}}$$

where n = number of observations.

For grouped data,

$$r_{xy} = \frac{\Sigma f\, .d_x\, d_y - \dfrac{\Sigma f d_x .\, \Sigma f d_y}{N}}{\sqrt{\Sigma f d_x^2 - \dfrac{(\Sigma f d_x)^2}{N}} \cdot \sqrt{\Sigma f d_y^2 - \dfrac{(\Sigma f d_y)^2}{N}}}$$

where $N = \Sigma f$

If $X = x - \bar{x}$ and $Y = y - \bar{y}$, then a short-cut formula is

$$r_{xy} = \frac{\Sigma XY}{\sqrt{\Sigma X^2} \cdot \sqrt{\Sigma Y^2}}$$

NOTES

NOTES

Example 1. *Calculate the Karl Pearson's coefficient of correlation of the following data :*

x	25	27	30	35	33	28	36
y	19	22	27	28	30	23	28

Solution. Here $\bar{x} = 31$. $\bar{y} = 25$

x	y	$X = x-31$	$Y = y-25$	X^2	Y^2	XY
25	19	–6	–6	36	36	36
29	20	–2	–5	4	25	10
30	27	–1	2	1	4	–2
35	28	4	3	16	9	12
33	30	2	5	4	25	10
29	23	–2	–2	4	4	4
36	28	5	3	25	9	15
Σ: 217	175	0	0	90	112	85

The Karl Pearson's coefficient of correlation is given by

$$r_{xy} = \frac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}} = \frac{85}{\sqrt{90 \times 112}} = 0.85.$$

Example 2. *Compute Karl Pearson's coefficient of correlation in the following series relating to price and supply of commodity ;*

Price (Rs.)	60	65	70	75	80	85	90	95	100
Demand (Qts)	35	30	25	25	23	21	20	20	18

Solution. Computation table :

Price (x)	$d_x = x - 80$ $<d_x = x-A>$	d_x^2	*Demand (y)*	$d_y = y - 25$ $<d_y = y - B>$	d_y^2	$d_x.d_y$
60	–20	400	35	10	100	–200
65	–15	225	30	5	25	–75
70	–10	100	25	0	0	0
75	–5	25	25	0	0	0
80	0	0	23	–2	4	0
85	5	25	21	–4	16	–20
90	10	100	20	–5	25	–50
95	15	225	20	–5	25	–75
100	20	400	18	–7	49	–140
Σ	0	1500	–	–8	244	–560

$$r_{xy} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x . \Sigma d_y}{9}}{\sqrt{\left[\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{9}\right]} \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{9}}}$$

$$= \frac{-560}{\sqrt{1500} . \sqrt{244 - \frac{64}{9}}} \approx -0.94.$$

Example 3. *For calculation of the correlation coefficient between the variables X and Y, the following informations are obtained : n = 40, ΣX = 120, ΣX² = 600, ΣY = 90, ΣY² = 250, ΣXY = 356. It was, however, later discovered at the time of checking that it had copied down two pairs of observations as*

X	Y
9	11
12	8

while the correct values were

X	Y
8	12
11	9

Obtain the correct value of the correlation coefficient between X and Y.

Solution. We have

Corrected $\Sigma X = 120 - 9 - 12 + 8 + 11 = 118$

Corrected $\Sigma X^2 = 600 - 81 - 144 + 64 + 121 = 560$

Corrected $\Sigma Y = 90 - 11 - 8 + 12 + 9 = 92$

Corrected $\Sigma Y^2 = 250 - 121 - 64 + 144 + 81 = 290$

Corrected $\Sigma XY = 356 - 99 - 96 + 96 + 99 = 356$

∴ Correct value of correlation coefficient is given by

$$r_{xy} = \frac{\Sigma XY - \frac{\Sigma X . \Sigma Y}{n}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} . \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}}$$

$$= \frac{356 - \frac{118 \times 92}{40}}{\sqrt{560 - \frac{(118)^2}{40}} \sqrt{290 - \frac{(92)^2}{40}}} = 0.66.$$

REGRESSION EQUATIONS

In regression analysis we can predict or estimate the value of one variable with the help of the value of other variable of the distribution after fitting to an equation. Hence there are two regression equations. The regression equation of Y on X is used to predict the value of Y with the value of X, whereas the regression equation of X on Y is used to predict the value of X with the value of Y. Here the independent variable is called predictor or explanator or regressor and the dependent variable is called explained or regressed variable.

NOTES

When these regression equations are straight lines then they are called *regression lines.*

(a) Regression Line of Y on X. $Y = a + bX$

Let (x_1, y_1), (x_2, y_2),.....,(x_n, y_n) be the given observations using the method of least square, we can estimate the values of a and b and the resultant equation takes the form

$$Y - \overline{Y} = r.\frac{\sigma_Y}{\sigma_X}(X - \overline{X})$$

or,
$$y - \bar{y} = b_{yx}(x - \bar{x}) \text{ where } b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

= regression coefficient of y on x.

(b) Regression Line of X on Y: $X = c + dY$

Similarly we obtain,

$$X - \overline{X} = r.\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

or,
$$x - \bar{x} = b_{xy}(y - \bar{y}),$$

where $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ = regression coefficient of x on y.

(c) Properties

1. Both the regression lines pass through the mean values $(\bar{x}, \bar{y})$.
2. $b_{xy}.b_{yx} = r^2$

   $\Rightarrow r = \pm\sqrt{b_{xy}.b_{yx}}$ and the sign of r is the same as of regression coefficients.
3. The two regression equations are different, unless $r = \pm 1$, in which case the two equations are identical.
4. The angle between the regression lines is given by, $\tan\theta = \frac{1-r^2}{r}.\frac{\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2}$.

(d) Formulas for Regression Coefficients.

1. For, $X = x - \bar{x}$, $Y = y - \bar{y}$,

$$b_{xy} = \frac{\Sigma XY}{\Sigma Y^2}, \quad b_{yx} = \frac{\Sigma XY}{\Sigma X^2}$$

2. Generally,

$$b_{xy} = \frac{\Sigma xy - \frac{\Sigma x.\Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}, \quad b_{yx} = \frac{\Sigma xy - \frac{\Sigma x.\Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

   where n = no. of observations.

3. For $d_x = x - A$, $d_y = y - B$ where A and B are assumed values,

$$b_{xy} = \frac{\Sigma d_x.d_y - \frac{\Sigma d_x.\Sigma d_y}{n}}{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{n}}$$

$$b_{yx} = \frac{\Sigma d_x . d_y - \frac{\Sigma d_x . \Sigma d_y}{n}}{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{n}}$$

NOTES

4. For grouped data

$$b_{xy} = \frac{\Sigma f d_x d_y - \frac{\Sigma f d_x . \Sigma f d_y}{N}}{\Sigma f d_y^2 - \frac{(\Sigma f d_y)^2}{N}}$$

$$b_{yx} = \frac{\Sigma f . d_x d_y - \frac{\Sigma f d_x . \Sigma f d_y}{N}}{\Sigma f d_x^2 - \frac{(\Sigma f d_x)^2}{N}}$$

where N = Σf.

(e) Standard Error of Estimates.

Consider the regression equation of X on Y. Then the root mean square deviation of the points from the regression line of X on Y is called the standard error of estimate of X which is given by

$$S_x = \sigma_x \sqrt{1 - r^2}$$

Similarly, the standard error of estimate of Y from the regression equation Y on X is

$$S_y = \sigma_y \sqrt{1 - r^2}$$

The standard error of estimate serves a standard deviation of the size of the error of the predicted values of Y (from the equation Y on X) and of X (from the equation X on Y). The size of the standard error also helps up to assess the quality of our regression model.

(f) Coefficient of Determination.

This gives the percentage variation in the dependent variable that is accounted for or explained by the independent variable is given by

$$\text{Coefficient of determination, } R^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

Let Y be the dependent variable and X be the independent variable. If R^2 = 0.85 then we shall be able to reduce or explain 85% of the variation in Y with a knowledge of X.

If $\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i$ $(i = 1, 2, ...n,)$ be the fitted values to the observation (x_i, y_i) $i = 1, 2 .....n$, then

$$R^2 = \frac{n \Sigma y_i^2 - (\Sigma y_i)^2}{n \Sigma (\hat{y}_i)^2 - (\Sigma y_i)^2}, \quad 0 \le R^2 \le 1$$

$R^2 = 1 \Rightarrow$ all n observations lie on the fitted regression line.

Example 4. *From the following data obtain the two regression lines and the correlation coefficient :*

Sales (*x*)	100	98	78	85	110	93	80
Purchase (*y*)	85	90	70	72	95	81	74

Find the value of y when x = 82.

NOTES

Solution. Here $\Sigma x = 644$, $\Sigma y = 567$, $n = 7$

$$\therefore \quad \bar{x} = \frac{\Sigma x}{n} = 92, \qquad \bar{y} = \frac{\Sigma y}{n} = 81$$

x	$X = x - \bar{x}$	X^2	y	$Y = y - \bar{y}$	Y^2	XY
100	8	64	85	4	16	32
98	6	36	90	9	81	54
78	–14	196	70	–11	121	154
85	–7	49	72	–9	81	63
110	18	324	95	14	196	252
95	3	9	81	0	0	0
80	–12	144	70	–11	121	132
	Σ	822			616	687

The regression coefficients are

$$b_{yx} = \frac{\Sigma XY}{\Sigma X^2} = 0.84$$

$$b_{xy} = \frac{\Sigma XY}{\Sigma Y^2} = 1.12$$

Regression equation of y on x :

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\Rightarrow \quad y - 81 = 0.84(x - 92)$$

$$\Rightarrow \quad y = 0.84x + 3.72$$

Regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\Rightarrow \quad x - 92 = 1.12(y - 81)$$

$$\Rightarrow \quad x = 1.12y + 1.28$$

The correlation coefficient = $\sqrt{b_{xy} \cdot b_{yx}}$ (since both coefficients are positive)

$$= \sqrt{(0.84).(1.12)} = 0.97$$

For $x = 82$, the value of y to be obtained from the regression equation of y on x. Hence $y = (0.84)(82) + 3.72 = 72.6$.

Example 5. *Consider the two regression lines : 3X + 2Y = 26 and 6X + Y = 31, (a) Find the mean value and correlation coefficient between X and Y. (b) If the variance of Y is 4, find the S.D. of X.*

NOTES

Solution. (*a*) Intersection of two regression lines gives the mean value *i.e.*, $(\bar{X}, \bar{Y})$.

Solving the two equations, we obtain $\bar{X} = 4$ and $\bar{Y} = 7$.

Let $3X + 2Y = 26$ be the regression line of X on Y and the other line as Y on X.

Then
$$X = -\frac{2}{3}Y + \frac{26}{3} \text{ (X on Y)} \Rightarrow b_{XY} = -\frac{2}{3}$$
$$Y = -6X + 31 \text{ (Y on X)} \Rightarrow b_{YX} = -6$$

but $r^2 = b_{xy}.\, b_{yx} = 4$ which cannot be true.

So we change our assumptions *i.e.*, the line $3X + 2Y = 26$ represents Y on X and the other line as X on Y.

Then
$$Y = -\frac{3X}{2} + 13 \text{ (Y on X)} \Rightarrow b_{YX} = -\frac{3}{2}$$
$$X = -\frac{1}{6}Y + \frac{31}{6} \text{ (X on Y)} \Rightarrow b_{XY} = -\frac{1}{6}$$

$\therefore$
$$r = -\sqrt{b_{YX}.b_{XY}} \text{ (Since both the coefficients are negative)}$$
$$= -\sqrt{\frac{3}{2} \times \frac{1}{6}} = -\frac{1}{2}$$

(*b*) Given $\sigma_y^2 = 4 \Rightarrow \sigma_y = 2$

We have
$$b_{XY} = r.\frac{\sigma_X}{\sigma_Y}$$

$\Rightarrow$
$$\sigma_X = \frac{\sigma_Y.\, b_{XY}}{r}$$
$$= \frac{2 \times (-1/6)}{(-1/2)} = \frac{2}{3}.$$

RANK CORRELATION

Let us suppose that a group of *n* individuals is given grades or ranks with respect to two characteristics. Then the correlation obtained between these ranks assigned on two characteristics is called rank correlation.

Let (x_i, y_i), $i = 1, 2, \ldots\ldots, n$ be the ranks of the i-th individual in two characteristics. Then Spearman's Rank correlation coefficient is given as

$$r = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)}$$

where $d_i = x_i - y_i$.

It is also to be noted that $-1 \le r \le 1$ and the above formula is used when ranks are not repeated.

NOTES

For repeated ranks, a correction factor is required in the formula. If m is the number of times an item is repeated then the factor $\frac{m(m^2-1)}{12}$ is to be added to Σd^2. For each repeated value, this correction factor is to be added.

Example 6. *The ranks of some 10 students in two subjects A and B are given below :*

Ranks in A	5	2	9	8	1	10	3	4	6	7
Ranks in B	10	5	1	3	8	6	2	7	9	4

Calculate the rank correlation coefficient.

Solution. Here $n = 10$.

Ranks in A	5	2	9	8	1	10	3	4	6	7	
Ranks in B	10	5	1	3	8	6	2	7	9	4	
d = A – B	–5	–3	8	5	–7	4	1	–3	–3	3	Σ
d^2	25	9	64	25	49	16	1	9	9	9	216

Rank correlation coefficient is given by

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 216}{10 \times 99} = -0.31.$$

Example 7. *Obtain the rank correlation coefficient for the following data :*

X	85	74	85	50	65	78	74	60	74	90
Y	78	91	78	58	60	72	80	55	68	/70

Solution. Here $n = 10$

X	*Y*	*Rank X* (x)	*Rank Y* (y)	$d = x - y$	d^2
85	78	2.5	3.5	–1	1
74	91	6	1	5	25
85	78	2.5	3.5	–1	1
50	58	10	9	1	1
65	60	8	8	0	0
78	72	4	5	–1	1
74	80	6	2	4	16
60	55	9	10	–1	1
74	68	6	7	–1	1
90	70	1	6	–5	25
			Σ	0	72

In the X series 85 has repeated twice and given ranks 2.5 instead of 2 and 3. For this the correction factor is $\frac{2(4-1)}{12} = \frac{1}{2}$.

Also, 74 has repeated thrice in X series and given ranks 6 instead of 5, 6, 7. For this the correction factor is $\frac{3(9-1)}{12} = 2$.

In the Y series 78 has repeated twice and given ranks 3.5 instead of 3 and 4. For this the correction factor is $\frac{2(4-1)}{12} = \frac{1}{2}$.

So the total correction factors = $\frac{1}{2} + 2 + \frac{1}{2} = 3$

Then the rank correlation coefficient is given by

$$r = 1 - \frac{6(72+3)}{10(100-1)} = 0.545.$$

CORRELATION OF BIVARIATE FREQUENCY DISTRIBUTION

Consider the two way frequency table with marginal and joint distributions of the two variables X and Y.

X / Y	Classes: Mid Points x_1 x_2 ..... x_n	
Classes: Mid-Points y_1	$f(x, y)$	$g(y_1) = \sum_x f(x, y_1)$
y_2	$f(x, y)$	$g(y_2) = \sum_x f(x, y_2)$
y_m	$f(x, y)$	$g(y_m) = \sum_x f(x, y_m)$
	$h(x_1)$ ..... $h(x_n)$ $= \sum_y f(x_1, y)$ $= \sum_y f(x_n, y)$	$N = \sum_x \sum_y f(x, y)$

Here $\bar{x} = \frac{1}{N} \Sigma x_i \, h(x_i), \quad \bar{y} = \frac{1}{N} \Sigma y_j \, g(y_j)$

or simply $= \frac{1}{N} . \Sigma x h(x) \qquad = \frac{1}{N} \Sigma y \, g(y)$

$$\sigma^2_x = \frac{1}{N} \Sigma x^2 h(x) - (\bar{x})^2, \quad \sigma^2_y = \frac{1}{N} \Sigma y^2 \, g(y) - (\bar{y})^2$$

$$Cov(x,y) = \frac{1}{N} \sum_x \sum_y xy \, f(x,y) - \bar{x}.\bar{y}$$

Hence, $$r_{xy} = \frac{Cov.(x,y)}{\sigma_x \cdot \sigma_y}$$

Note. For large data, the two way frequency table is advantageous.

Example 8. *Calculate the correlation coefficient from the following table.*

x / y	0 – 8	8 – 16	16 – '24
1 – 5	2	0	4
5 – 9	3	2	2
9 – 13	2	5	1

Solution. Consider the following table :

	x / y	Mid-values 4	12	20	$g(y)$
Mid-Values	3	2	0	4	6
	7	3	2	2	7
	11	2	5	1	8
$h(x)$		7	7	7	21

$$\bar{x} = \frac{1}{21}[(4)(7) + (12)(7) + (20)(7)] = 12$$

$$\bar{y} = \frac{1}{21}[(3)(6) + (7)(7) + (11)(8)] = 7.38$$

$$\sigma_x^2 = \frac{1}{21}[(16)(7) + (144)(7) + (400)(7)] - (12)^2 = 42.67$$

$$\sigma_y^2 = \frac{1}{21}[(9)(6) + (49)(7) + (121)(8)] - (7.38)^2 = 10.54$$

$$Cov(x,y) = \frac{1}{21}\sum_x\sum_y xy\, f(x,y) - \bar{x}.\bar{y}$$

$$= \frac{1764}{21} - (12)(7.38) = -4.56$$

The correlation coefficient is

$$r_{xy} = \frac{-4.56}{\sqrt{42.67}.\sqrt{10.54}} = -0.22.$$

MULTIPLE REGRESSION

If more than one predictor variable is present in the regression equation then it is called multiple regression. Let us consider two predictor variables and one regressed variable and the multiple linear regression model can be stated as follows:

$$y = a_0 + a_1 x_1 + a_2 x_2$$

To estimate a_0, a_1 and a_2 we take (x_{1i}, x_{2i}, y_i), $i = 1,2,\ldots\ldots, n$ as observed data where the x's are assumed to be known without error while the y values are random variables.

Let $S = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i})]^2$ be the sum of squares of errors. Then to minimize S, we take $\frac{\partial S}{\partial a_0} = 0$, $\frac{\partial S}{\partial a_1} = 0$ and $\frac{\partial S}{\partial a_2} = 0$. From which we obtain three normal equations

$$\Sigma y_i = n a_0 + a_1 \Sigma x_{1i} + a_2 \Sigma x_{2i}$$

$$\Sigma x_i y_i = a_0 \Sigma x_{1i} + a_1 \Sigma x_{1i}^2 + a_2 \Sigma x_{1i} x_{2i}$$

$$\Sigma x_{2i} y_i = a_0 \Sigma x_{2i} + a_1 \Sigma x_{1i} x_{2i} + a_2 \Sigma x_{2i}^2$$

By solving these equations we obtain the least square estimates of a_0, a_1 and a_2.

Note. 1. In the above model if a_2 vanishes then it is the regression line of y on x.

2. The above model represents a regression plane.

Example 9. *Consider the following data:*

x_1	2	4	5	6	3	1
x_2	1	2	1	3	5	2
y	14	16	17	20	18	12

Fit a least squares regression plane.

Solution. Here $n = 6$. Let the regression plane be $y = a_0 + a_1 x_1 + a_2 x_2$.

	x_1	x_2	x_1^2	x_2^2	$x_1 x_2$	$x_1 y$	$x_2 y$	y
	2	1	4	1	2	28	14	14
	4	2	16	4	8	64	32	16
	5	1	25	1	5	85	17	17
	6	3	36	9	18	120	60	20
	3	5	9	25	15	54	90	18
	1	2	1	4	2	12	24	12
Σ	21	14	91	44	50	363	237	97

The three normal equations can be written as follows :

$$97 = 6a_0 + 21 a_1 + 14 a_2$$

$$363 = 21 a_0 + 91 a_1 + 50 a_2$$

$$237 = 14 a_0 + 50 a_1 + 44 a_2.$$

By solving we obtain,

$$a_0 = 9.7, \quad a_1 = 1.3, \quad a_2 = 0.83.$$

CURVILINEAR REGRESSION

Consider one predictor variable and one regressed variable. Then there may be curvilinear (or non-linear) relationship between these variables. We consider some important relationships. To estimate the unknowns, the method of least square is to be applied. Briefly, we illustrate.

(a) $y = a + bx + cx^2$ (Parabolic equation)

Consider S = Sum of squares of errors

$$= \Sigma[y-(a+bx+cx^2)]^2$$

To minimize S, we take $\frac{\partial S}{\partial a}=0, \frac{\partial S}{\partial b}=0$ and $\frac{\partial S}{\partial c}=0$ and obtain the normal equations as follows:

NOTES

$$\Sigma y_i = na + b\,\Sigma x_i + c\,\Sigma x_i^2$$

$$\Sigma x_i y_i = a\,\Sigma x_i + b\,\Sigma x_i^2 + c\,\Sigma x_i^3$$

$$\Sigma x_i^2 y_i = a\,\Sigma x_i^2 + b\,\Sigma x_i^3 + c\,\Sigma x_i^4$$

Solving these equations we obtain the estimates of a, b and c. For $y = a + bx$ (linear fit), there will be the first-two normal equations with $c = 0$.

(b) $y = ae^{bx}$ (Exponential equation)

To estimate a and b, take first log (base 10) on both sides

$$\log y = \log a + bx \log e$$

$$\Rightarrow \quad Y = A + Bx \text{ which is linear regression.}$$

where $Y = \log y,\ A = \log a,\ B = b \log e$

The normal equations are

$$\Sigma Y_i = nA + B\Sigma x_i$$

$$\Sigma x_i Y_i = A\,\Sigma x_i + B\,\Sigma x_i^2$$

The estimation of A and B will give the estimation of a and b.

(c) $y = ax^b$ (Geometric curve / Power function)

To estimate a and b, take first log (base 10) on both sides

$$\log y = \log a + b \log x$$

$$\Rightarrow \quad Y = A + bX$$

where $Y = \log y$, $A = \log a$, $X = \log x$.

The normal equations are

$$\Sigma Y_i = nA + b\,\Sigma X_i$$

$$\Sigma X_i Y_i = A\,\Sigma X_i + b\Sigma X_i^2$$

Here the estimation of A will give the estimation of a, while the estimation of b is obtained directly from normal equations.

(d) $xy^a = b$ (Gas equation)

To estimate a and b, take first log (base 10) on both sides

$$\log x + a \log y = \log b.$$

$$\Rightarrow \quad \log y = \frac{1}{a}\log b - \frac{1}{a}\log x$$

$$\Rightarrow \quad Y = A + BX$$

where $Y = \log y,\ A = \frac{1}{a}\log b,\ B = -1/a,\ X = \log x$

The normal equations are

$$\Sigma Y_i = nA + B\Sigma X_i$$

$$\Sigma X_i Y_i = A\Sigma X_i + B\Sigma X_i^2$$

The estimation of A and B will give the estimation of a and b.

(e) $y = ab^x$ (Growth of bacteria)

To estimate a and b, take first log (base 10) on both sides

$$\log y = \log a + x \log b$$

$$\Rightarrow \quad Y = A + xB.$$

The normal equations are

$$\Sigma Y_i = nA + B\Sigma x_i$$

$$\Sigma x_i Y_i = A\Sigma x_i + B\Sigma x_i^2$$

The estimation of A and B will give the estimation of a and b.

Example 10. *Find the best fitting regression equation of type $y = a + bx + cx^2$ to the following data :*

x	3	2	1	0	–1	–2	–3
y	10	8	3	1	2	6	8

Solution. Here $n = 7$.

	x	y	x^2	x^3	x^4	xy	x^2y
	3	10	9	27	81	30	90
	2	8	4	8	16	16	32
	1	3	1	1	1	3	3
	0	1	0	0	0	0	0
	–1	2	1	–1	1	–2	2
	–2	6	4	–8	16	–12	24
	–3	8	9	–27	81	–24	72
Σ	0	38	28	0	196	11	223

The normal equations are

$$7a + 28c = 38$$

$$28b = 11$$

$$28a + 196c = 223.$$

By solving these equations we obtain,

$a = 2.048$ $\qquad$ $b = 0.393$ $\qquad$ $c = 0.845$

Example 11. *Find the best fitting regression equation of type $y = ax^b$ to the following data :*

x	1	2	3	4	5	6
y	2	16	54	128	250	432

Solution. $\qquad y = ax^b$

Taking log on both sides we obtain

$$\log y = \log a + b \log x$$

$$\Rightarrow \quad Y = A + bX.$$

where, $\quad Y = \log y,\ A = \log a,\ X = \log x.$

Here $\quad n = 6$

x	X	y	Y	XY	X^2
1	0	2	0.3010	0	0
2	0.3010	16	1.2041	0.3624	0.0906
3	0.4771	54	1.7324	0.8265	0.2276
4	0.6021	128	2.1072	1.2687	0.3625
5	0.6990	250	2.3979	1.6761	0.4886
6	0.7782	432	2.6355	2.0509	0.6056
Σ	2.8574	–	10.3781	6.1846	1.7749

The normal equations are

$$6A + 2.8574 = 10.3781$$

$$2.8574A + 1.7749\,b = 6.1846$$

By solving these equations we obtain,

$$A = 0.3011 \Rightarrow a = 2.0004$$

and $$b = 2.9996$$

Hence the best fitting regression equation is

$$y = 2.0004\, x^{2.9996}$$

Example 12. *Fit the curve* $y = ae^{bx}$ *for the following data:*

x	1	5	7	9	12
y	10	15	12	15	21

Solution. $y = ae^{bx}$

Taking log on both sides we obtain,

$$\log y = \log a + bx \log e$$

$$\Rightarrow \quad Y = A + Bx$$

where, $Y = \log y$, $A = \log a$, $B = b \log e$.

Here $n = 5$

x	y	Y	x^2	xY
1	10	1	1	1
5	15	1.1761	25	5.8805
7	12	1.0792	49	7.5544
9	15	1.1761	81	10.5849
12	21	1.3222	144	15.8664
Σ	34	–	5.7536	300	40.8862

The normal equations are

$$5A + 34\,B = 5.7536$$

$$34\,A + 300\,B = 40.8862$$

By solving these equations we obtain,

$$A = 0.9766 \Rightarrow a = 9.48$$

$$B = 0.0256 \Rightarrow b = 0.06$$

Hence the best fitting regression equation is

$$y = 9.48\ e^{0.06x}.$$

NOTES

SUMMARY

- When a distribution has two variables then it is called bivariate.
- Coefficient of Determination gives the percentage variation in the dependent variable that is accounted for or explained by the independent variable is given by

$$\text{Coefficient of determination, } R^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

PROBLEMS

1. *Calculate the (i) two regression coefficients, (ii)* coefficient of correlation and *(iii)* the two regression equations from the following information :

   $n = 10$, $\Sigma X = 350$, $\Sigma Y = 310$, $\Sigma(X - 35)^2 = 162$, $\Sigma (Y - 31)^2 = 222$; *and*

   $\Sigma (X - 35) (Y - 31) = 92$.

2. Calculate the correlation coefficient of the following data :

x	45	46	46	47	48	49	50
y	44	48	45	48	52	51	49

3. *In order to find the correlation* coefficient between two variables X and Y from 20 pairs of observations, the following calculations were made :

   $\Sigma x = 120$, $\Sigma y = 70$, $\Sigma x^2 = 780$, $\Sigma y^2 = 450$ and $\Sigma xy = 500$.

   On subsequent verification it was discovered that the two pairs ($x = 9$, $y = 11$) and ($x = 7$, y = 6) were copied wrongly, the correct values being ($x = 9$, $y = 12$) and ($x = 6$, $y = 7$). Obtain *(i)* the correct value of correlation coefficient, *(ii)* the two lines of regression, and *(iii)* angle between them.

4. *The following data refers to* the percentage of pig iron (x) and the line consumption in cwt, (y) per cast for 50 casts of steel,

   $\Sigma x = 1885$, $\Sigma y = 9291$, $\Sigma x^2 = 72917$, $\Sigma y^2 = 1801653$, $\Sigma xy = 355654$.

   Obtain the lines of regression. Estimate the line consumption for a cast with 45% pig iron.

5. *If the tangent* of the angle between the lines of regression of Y on X and X on Y is 0.6, and the S.D. of Y is twice that of X, find the correlation coefficient between X and Y.

6. Calculate the correlation coefficient and the lines of regression from the following data:

x	57	58	59	59	60	61	62	64
y	77	78	75	78	82	82	79	81

7. (a) In a correlation analysis, the value of the Karl Pearson's coefficient of correlation and its probable error were found to be 0.90 and 0.40 respectively. Find the value of n.

NOTES

(*b*) Given that $x = 4y + 5$ and $y = Kx + 4$ are the regression lines of x on y and y on x, respectively, show that $0 \leq K \leq 25$. If $K = 0.10$ actually, find the means of the variables x and y and also their coefficient of correlation.

8. Three judges give the following ranks of ten different models of a car having different attributes.

Models		1	2	3	4	5	6	7	8	9	10
Judge	A	3	9	6	8	5	4	2	1	10	7
	B	10	5	2	1	7	4	9	3	6	8
	C	7	8	1	5	3	10	6	2	4	9

Discuss which pair of judges have the nearest approach to common tastes of the car models.

9. The coefficient of rank correlation of the marks obtained by 10 students in two subjects was found to be 0.62. It was later discovered that the difference in ranks for one student was taken wrongly as 5 instead of 7. Find the correct coefficient of rank correlation.

10. The following are the data on 5 players

Outdoor exercise x_1 *(in hrs)*	*Indoor exercise* x_2 *(in hrs)*	*Scores* (y)
15	20	240
12	18	380
15	14	250
20	14	450
15	15	310

Fit an equation of the form $y = a_0 + a_1 x_1 + a_2 x_2$ to the given data.

11. Obtain the multiple regression plane $y = a_0 + a_1 x_1 + a_2 x_2$ from the following data :

x_1	1	2	3	1	0	−1
x_2	2	−1	1	3	2	3
y	5	4	6	8	5	6

12. Obtain a second degree regressed polynomial from the following data :

x	0	1	2	3	4
y	1	1.5	2.6	4.2	6.8

13. Consider the production (in '000 units) of an item of a manufacturing company from the years 1990 to 1996 :

Year (x)	1990	1991	1992	1993	1994	1995	1996
Production (y)	6	7	5	4	6	7	5

Fit the trend of the type $y = a + bx + cx^2$ to the above data (Take 1993 as the year of origin).

14. Fit a least square parabola $y = a + bx + cx^2$ to the following data:

x	0	1	2	3	4	5	6
y	2.4	2.1	3.2	5.6	9.3	14.6	21.9

15. Fit a power function to the following price-demand data in six different market areas for a product.

Price *(Rs)* (x)	10	11	12	13	14	15
Demand ('00) (y)	80	70	58	55	32	20

16. Fit a curve $xy^a = b$ to the following data:

x	2	4	6	8	10
y	3	7	11	12	9

17. If the growth of a certain kind of bacteria follows the law $y = ab^x$, then find the best fitting values of a and b using the following data :

x	1	2	3	4	5
y	233.2	253.4	282.3	302.4	332.8

18. For the data given below, find the equation to the best fitting exponential curve of the form $y = ae^{bx}$

x	2	4	6	8	10	12
y	120	102	85	74	62	50

19. An employment bureau asked applicants their weekly wages on jobs last held. The actual wages were obtained for 54 of them and are recorded in the table below: x represents reported wage, y actual wage and the entry in the table represents frequency. Find the correlation coefficient.

y / x	15	20	25	30	35	40
40						2
35				3	5	
30			4	15		
25			20			
20		3	1			
15	1					

20. Calculate the Karl Pearson's coefficient of correlation between X and Y from the bivariate sample of 140 pairs of X and Y as distributed below :

X / Y	10 – 20	20 – 30	30 – 40	40 – 50
10 – 20	20	26	–	–
20 – 30	8	14	37	–
30 – 40	–	4	18	3
40 – 50	–	–	4	6

ANSWERS

1. (*i*) $b_{xy} = 0.41$, $b_{yx} = 0.57$ (*ii*) 0.48, (*iii*) $y = 0.57x + 11.05$, $x = 0.41y + 22.29$

2. 0.754

3. (*i*) 0.754, (*ii*) $y = 3.918x - 20.3$, $x = 0.435y + 4.57$, (*iii*) 13°54'

4. $y = 2.91x + 76.11$, $x = 0.07y + 24.69$, 207.06

5. $r = ½$

6. $r = 0.6$, $y = 0.67x + 39$, $x = 0.55y + 16.9$

7. (*a*) $n = 10$, (*b*) $\bar{x} = 26.67$, $\bar{y} = 7.5$, $r = 0.63$

8. B and C

9. 0.47

10. $y = 216.79 + 11.65x_1 - 4.34x_2$

11. $a_0 = 3.32$, $a_1 = 0.6217$, $a_2 = 1.031$

12. $y = 1.036 + 0.087x + 0.336x^2$

13. $y = 5.429 - 0.071X + 0.071X^2$, where $X = x - 1993$

14. $y = 2.51 - 1.2x + 0.73x^2$

15. $y = 154881.66x^{-3.21}$

16. $a = -1.27$, $b = 0.4$

17. $a = 213.45$, $b = 1.09$

18. $a = 143.64$, $b = -0.09$

19. 0.93

20. – 0.705.

CHAPTER 6 PROBABILITY

NOTES

★ STRUCTURE ★

- Definitions
- Axioms of Mathematical Probability
- Complementation Rule
- Theorem of Total Probability/Addition Theorem
- Theorem of Compound Probability/Multiplication Theorem
- Independent Events
- Subjective Probability
- Baye's Theorem
- Summary
- Problems

DEFINITIONS

Random experiments are those experiments whose results depend on chance. For example, tossing a coin, where head or tail can turn up in a single toss. When a space shuttle takes off, then returning to the ground depends on several chance factors. To satisfy the demand of an item we deal with random experiment.

The single performance of a random experiment will be called an **outcome.** For example, head or tail is the outcome of tossing a coin.

A set of all possible outcomes of an experiment is called a **sample space.** For example, the sample space of tossing two coins is {(H, H), (H, T), (T, H), (T, T)} where H = head and T = tail. Consider two cities A and B connected by two roads R_1 and R_2. Another city C connected to B by a single road R_3.

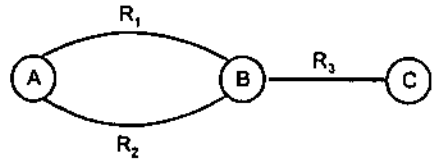


Fig. 6.1

The possible travel times in the roads are as follows:

$R_1 = 3hr$, 3.5 hr.

$R_2 = 3hr$, 3.75 hr.

$R_3 = 2hr$, 1.75 hr.

Then the sample space of possible travel times for this network from A to C is

NOTES

{3 + 2, 3 + 1.75, 3.5 + 2, 3.5 + 1.75, 3 + 2, 3 + 1.75, 3.75 + 2, 3.75 + 1.75}

Any subset of a sample space is called an **event.** Events may be elementary or composite. In case of elementary event it cannot be decomposed into simpler events whereas the composite event is an aggregate of several elementary events.

A sample space is said to be **discrete** if it contains finite or countably infinite elements. For continuum elements, the sample space is said to be **continuous.**

If all the possible events in a random experiment are considered then this set is called **exhaustive.**

In tossing a coin, either head or tail will turn up. Both cannot turn up. These type of events are called **mutually exclusive.** In the above travel example, reaching to B from A can be made either by road R_1 or by road R_2 which is mutually exclusive. If none of the events in a sample space can be expected in preference to another then these events are said to be **equally likely.**

Consider a random experiment with possible results as cases. If there are n exhaustive, mutually exclusive and equally likely cases and of them m are favourable to an event A, then the **probability** of A is defined by

$$p = P(A) = \frac{m}{n},$$

The above definition is known as **classical definition** of probability.

Example 1. *What is the probability of getting 3 tails in tossing 3 coins ?*

Solution. The sample space = {TTT, TTH, THT, THH, HTT, HTH, HHT, HHH}

Since three tails have occurred only once.

$$\therefore \quad P(3T) = \frac{1}{8}$$

The probability of a composite event is the sum of the probabilities of the simple events of which it is composed.

Example 2. *Find the probability of exactly two tails in tossing 3 coins ?*

Solution. From the sample space (given in example 1) exactly two tails occur as TTH, THT and HTT.

Hence P(exactly two tails) $= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$.

AXIOMS OF MATHEMATICAL PROBABILITY

Let A be an event in a sample space S, then

(*i*) $0 \leq P(A) \leq 1$

NOTES

[Here P(A) = 0 means that the event will not occur and P(A) = 1 means that the event is certain]

(*ii*) P(S) = 1

(*iii*) If A and B are two mutually exclusive events then

$$P(A+B)=P(A)+P(B)$$

COMPLEMENTATION RULE

Let A be an event and $\overline{A}$ be its complement, then

$$P(A)+P(\overline{A})=1$$

For example, if the probability of hitting a target by a missile is $\frac{1}{6}$ then its complement of not hitting the target will be given by $1-\frac{1}{6}=\frac{5}{6}$.

Therefore if P(A) denotes the probability of occurrence of event A then $P(\overline{A})$ denotes the probability that event A fails to occur.

Example 3. *Two dice are thrown. What is the probability that the sum of two faces is multiple of 3 ?*

Solution. Each dice has the numbers 1, 2, 3, 4, 5, 6. Then the total number of cases *i.e.,* the sample space will consist of 6^2 = 36 elements.

The favourable cases i.e., the sum of two faces is multiple of 3 are given by

(1,2), (2,1), (5,1), (1,5), (4,2), (2,4), (3,3), (3,6), (6,3), (5,4), (4,5), (6,6)

i.e., 12 cases.

Hence P (sum of two faces is multiple of 3) $=\frac{12}{36}=\frac{1}{3}$.

Example 4. *A batch contains 10 articles of which 3 are defective. If 4 articles are chosen at random, what is the probability that none of them is defective ?*

Solution. Total number of ways of selecting 4 articles out of 10 is $\binom{10}{4}=210$.

If none of the selected articles is defective, which must come from 7 non-defective articles. So the number of favourable cases is $\binom{7}{4}=35$.

Hence the required probability $=\frac{35}{210}=\frac{1}{6}$.

Example 5. *A room has three lamp sockets. From a collection of 10 light bulbs of which only 6 are good, a person selects 3 at random and puts them in the sockets. What is the probability that the room will have light?*

Solution. From the given 10 bulbs, 6 are good and 4 are damaged or bad bulbs. If the person selects at least one good bulb, then the room will have light.

P(room will not have light) = P(the person selects 3 bad bulbs)

$$= \frac{\binom{4}{3} \times \binom{6}{0}}{\binom{10}{3}} = \frac{1}{30}.$$

NOTES

Using complementation rule; we have

P(the room will have light) =1 – P(Room will not have light) = $1 - \frac{1}{30} = \frac{29}{30}$.

THEOREM OF TOTAL PROBABILITY/ADDITION THEOREM

I. If two events A and B are mutually exclusive, then the occurrence of either A or B is given by

$$P(A + B) = P(A) + P(B)$$

[Also we can write $P(A + B) = P(A \cup B)$].

Proof. Let n possible outcomes from a random experiment which are mutually exclusive, exhaustive and equally likely. If n_1 of these outcomes are favourable to the event A, and n_2 outcomes are favourable to the event B, then

$$P(A) = \frac{n_1}{n}, \qquad P(B) = \frac{n_2}{n}$$

Since the events A and B are mutually exclusive *i.e.,* the n_1 outcomes are completely distinct from n_2 outcomes, then the number of outcomes favourable to either A or B is $n_1 + n_2$.

$$\therefore \qquad P(A + B) = \frac{n_1 + n_2}{n} = \frac{n_1}{n} + \frac{n_2}{n} = P(A) + P(B).$$

II. When the two events A and B are not mutually exclusive, then the probability of occurrence of at least one of the 2 events is given by

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

Proof. Here the event A + B means the occurrence of one of the following mutually exclusive events : AB, $A\overline{B}$ and $\overline{A}B$. Therefore

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

Again, we have $\qquad P(A) = P(AB) + P(A\overline{B})$

$\Rightarrow \qquad P(A\overline{B}) = P(A) - P(AB)$

and $\qquad P(B) = P(AB) + P(\overline{A}B)$

$\Rightarrow \qquad P(\overline{A}B) = P(B) - P(AB)$

Hence,

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

Note. 1. For the events A, B and C which may not be mutually exclusive,

$$P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

2. **Boole's inequality.**

NOTES

$$P(A+B) \leq P(A)+P(B).$$

Here equality sign holds when P(AB) = 0, *i.e.*, A and B are mutually exclusive.

3. **Bonferroni's inequality.**

$$P(AB) \geq P(A)+P(B)-1.$$

THEOREM OF COMPOUND PROBABILITY/ MULTIPLICATION THEOREM

Let A and B be two events in a sample space S and $P(A) \neq 0$, $P(B) \neq 0$, then the probability of happening of both the events are given by

(*i*) P(AB) = P(A).P(B/A)

(*ii*) P(AB) = P(B).P(A/B).

[Here P(B/A) denote the conditional probability which means the probability of event B such that the event A has already occurred otherwise event B will not occur. Similarly P(A/B).]

Proof. Let n possible outcomes from a random experiment which are mutually exclusive, exhaustive and equally likely. If n_1 of these outcomes are favourable to the event A, the unconditional probability of A is

$$P(A)=\frac{n_1}{n}$$

Out of these n_1 outcomes, let n_2 outcomes be favourable to another event B *i.e.*, the number of outcomes favourable to A as well as B is n_2. Hence,

$$P(AB)=\frac{n_2}{n}$$

Then the conditional probability of B assuming that A has already occurred is

$$P(B/A)=\frac{n_2}{n_1}$$

Therefore, $$\frac{n_2}{n}=\frac{n_1}{n}.\frac{n_2}{n_1}$$

$\Rightarrow$ P(AB) = P(A). P(B/A) which is (*i*).

Similarly, we can prove (*ii*).

Note. If the occurrence of the event A as well as B as well as C is given by

P(ABC) = P(A). P(B/A) P(C/AB).

INDEPENDENT EVENTS

Two events A and B in a sample space S are said to be independent if P(AB) = P(A).P(B).

For n independent events $A_1, A_2, ..., A_n$

$$P(A_1.A_2...A_n) = P(A_1)\ P(A_2)...P(A_n).$$

If A and B are two independent events, then

$$P(A) = P(A / B) = P(A / \overline{B})$$
$$P(B) = P(B / A) = P(B / \overline{A}).$$

NOTES

SUBJECTIVE PROBABILITY

This is another way to interpret probabilities using personal evaluation of an event.

If the odds in favour of A are $a : b$ then the subjective probability is taken as $P(A) = \frac{a}{a+b}$.

If the odds against A are $a : b$ then the subjective probability is taken as $P(A) = \frac{b}{a+b}$.

These subjective probabilities may or may not satisfy the third axiom of probability.

Example 6. *Given* $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{4}$ *and* $P(AB) = \frac{1}{6}$.

Find the values of $P(A / B)$, $P(\overline{A}B)$, $P(\overline{A}\,\overline{B})$ *and* $P(\overline{A} + B)$.

Solution. $$P(A / B) = \frac{P(AB)}{P(B)} = \frac{1/6}{1/4} = \frac{2}{3}$$

$$P(\overline{A}B) = P(B) - P(AB) = \frac{1}{4} - \frac{1}{6} = \frac{1}{12}$$

$$P(\overline{A}\,\overline{B}) = 1 - P(A + B)$$
$$= 1 - [P(A) + P(B) - P(AB)]$$
$$= 1 - \left[\frac{1}{3} + \frac{1}{4} - \frac{1}{6}\right] = \frac{7}{12}.$$

$$P(\overline{A} + B) = P(\overline{A}) + P(B) - P(\overline{A}B)$$
$$= \left(1 - \frac{1}{3}\right) + \frac{1}{4} - \frac{1}{6} = \frac{3}{4}.$$

Example 7. *An article manufactured by a company consists of two parts I and II. In the process of manufacture of part I, 9 out of 100 are likely to be defective. Similarly, 5 out of 100 are likely to be defective in the manufacture of part II. Calculate the probability that the assembled article will not be defective.*

Solution. Here the assembled article will not be defective means both the parts I and II will not be defective.

NOTES

$$\text{P(defective part I)} = \frac{9}{100}$$

$$\Rightarrow \quad \text{P (non-defective part I)} = 1 - \frac{9}{100} = \frac{91}{100} = 0.91$$

$$\text{P (defective part II)} = \frac{5}{100}$$

$$\Rightarrow \quad \text{P (non-defective part II)} = 1 - \frac{5}{100} = \frac{95}{100} = 0.95$$

Since the manufacturing of part I and part II are independent then P (assembled article will not be defective)

= P (non-defective part I). P (non-defective part II)

= (0.91) (0.95) = 0.8645.

Example 8. *Six men in a company of 15 are engineers. If 3 men are picked out of the 15 at random, what is the probability of at least one engineer ?*

Solution. The event 'at least one engineer' can be split up into three mutually exclusive events:

(*i*) exactly 1 engineer and 2 non-engineers

(*ii*) exactly 2 engineers and 1 non-engineer

(*iii*) exactly 3 engineers and 0 non-engineer.

The probabilities of these cases are respectively,

$$\frac{\binom{6}{1}\binom{9}{2}}{\binom{15}{3}} = \frac{216}{455}, \qquad \frac{\binom{6}{2}\binom{9}{1}}{\binom{15}{3}} = \frac{135}{455}, \qquad \frac{\binom{6}{3}\binom{9}{0}}{\binom{15}{3}} = \frac{20}{455}$$

By the theorem of total probability we obtain,

$$\text{P (at least one engineer)} = \frac{216}{455} + \frac{135}{455} + \frac{20}{455} = \frac{371}{455}.$$

Example 9. *The probability that a student Mr. X passed Mathematics is* $\frac{2}{3}$, *the probability that he passes statistics is* $\frac{4}{9}$. *If the probability of passing at least one subject is* $\frac{4}{5}$, *what is the probability that Mr. X will pass both the subjects ?*

Solution. Let E = Event that Mr. X passed in Mathematics

F = Event that Mr. X passed in Statistics

$$\therefore \quad \text{P (E)} = \frac{2}{3}, \qquad \text{P (F)} = \frac{4}{9}$$

and P (E+F) = Probability that student passes at least one subject

NOTES

$$= \frac{4}{5}.$$

Then, $P(E+F) = P(E) + P(F) - P(EF)$

$\Rightarrow$ $\frac{4}{5} = \frac{2}{3}+\frac{4}{9} - P(EF)$

$\Rightarrow$ $P(EF) = \frac{2}{3}+\frac{4}{9}-\frac{4}{5} = \frac{14}{45}$

$\Rightarrow$ The probability that the student will pass both the subjects is $\frac{14}{45}$.

Example 10. *An investment consultant predicts that the odds against the price of a certain stock will go up during the next week are 2 : 1 and the odds in favour of the price remaining the same are 1 : 3. What is the probability that the price of the stock will go down during the next week ?*

Solution. Let E = Event that stock price will go up

F = Event that stock price will remain same.

Then $P(E) = \frac{1}{3}$ and $P(F) = \frac{1}{4}$

$P(E \cup F)$ = P (stock price will either go up or remains same)

$= P(E) + P(F) = \frac{1}{3}+\frac{1}{4} = \frac{7}{12}.$

So, P(stock price will go down)

$= P(\overline{E}.\overline{F})$

$= 1 - P(E+F) = 1 - \frac{7}{12} = \frac{5}{12}.$

Example 11. *A and B throw alternately with a pair of dice. One who first throws a total of 9 wins. Show that the chances of their winning are 9 : 8.*

Solution. Let E_1 = Event of A throwing a total of 9 with a pair of dice,

E_2 = Event of B throwing a total of 9 with a pair of dice,

and these are independent events.

$\therefore$ $P(E_1) = P(E_2) = \frac{4}{36} = \frac{1}{9}$ and $P(\overline{E}_1) = P(\overline{E}_2) = \frac{8}{9}.$

Assume that A starts the game. Winning of A can be given by the following mutually exclusive cases :

(*i*) E_1, (*ii*) $\overline{E}_1\overline{E}_2E_1$, (*iii*) $\overline{E}_1\overline{E}_2\overline{E}_1\overline{E}_2E_1$ and so on.

Using the theorem of addition,

P [winning of A] $= P(E_1) + P(\overline{E}_1\overline{E}_2E_1) + P(\overline{E}_1\overline{E}_2\overline{E}_1\overline{E}_2E_1) + \ldots$

$= P(E_1) + P(\overline{E}_1)P(\overline{E}_2)P(E_1) + P(\overline{E}_1)P(\overline{E}_2)P(\overline{E}_1)P(\overline{E}_2)P(E_1) + \ldots$

$= \frac{1}{9}+\frac{8}{9}.\frac{8}{9}.\frac{1}{9}+\frac{8}{9}.\frac{8}{9}.\frac{8}{9}.\frac{8}{9}.\frac{1}{9}+\ldots$

NOTES

$$= \frac{1}{9}\left[1+\left(\frac{8}{9}\right)^2+\left(\frac{8}{9}\right)^4+....\right]$$

$$= \frac{\frac{1}{9}}{1-\left(\frac{8}{9}\right)^2} = \frac{9}{17}$$

$\therefore$ P [winning of B] $= 1-\frac{9}{17}=\frac{8}{17}$

Hence the chances of their winning are 9 : 8.

BAYE'S THEOREM

Let us consider B_1, B_2, ..., B_n are mutually exclusive and exhaustive events such that $B_1 + B_2 + ... + B_n = S$.

Let the event A occur in conjunction with only one of the events B_1, B_2,..., B_n. If the probabilities $P(B_1)$, $P(B_2)$, ..., $P(B_n)$ and $P(A/B_1)$, $P(A/B_2)$, ..., $P(A/B_n)$ are known, then

$$P[B_j / A] = \frac{P(B_j)P(A / B_j)}{\sum_{i=1}^{n} P(B_i).P(A / B_i)}, \quad j = 1,2,...,n.$$

Proof. We have

$$P(AB_j) = P(B_jA) = P(B_j)\ P(A/B_j)$$

$$A = AB_1 + AB_2 + ... + AB_n$$ ($\because$ A occurs in conjunction with only one of the events B_1, B_2, ..., B_n)

$\Rightarrow$ $$P(A) = P(AB_1) + P(AB_2) + ... + P(AB_n)$$ ($\because$ all AB_j are mutually exclusive)

$$= P(B_1).P(A / B_1) + P(B_2)P(A / B_2) + ... + P(B_n).P(A / B_n)$$

Then $$P[B_j/A] = \frac{P(B_j.A)}{P(A)} = \frac{P(AB_j)}{P(A)}$$

$$= \frac{P(B_j)P(A / B_j)}{\sum_{i=1}^{n} P(B_i).P(A / B_i)}, \quad j = 1,2,...,n$$

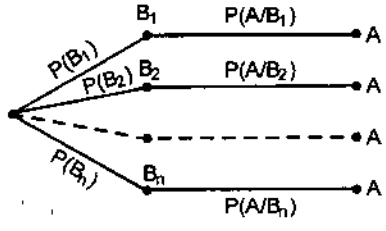


Fig. 6.2 Tree Presentation.

Note. 1. The expression $P(A) = \sum_{i=1}^{n} P(B_i)P(A/B_i)$ is called the rule of elimination or the rule of total probability.

2. The probabilities $P(B_i)$ are called 'a priori' probabilities.

NOTES

Example 12. *Four boxes Q_1, Q_2, Q_3 and Q_4 contain some gold and copper coins. The percentage of the total number of coins in these boxes are respectively 10, 20, 30 and 40. The fractions of gold coins in the boxes are respectively 0.2, 0.3, 0.1 and 0.5.*

(i) If a coin is taken out at random, what is the probability that it is a gold coin?

(ii) If a coin is taken out at random is found to be golden, what is the probability that it is taken from box Q_2?

Solution. Let $P(G)$ = Probability that the coin is a gold coin

$P(Q_i)$ = Probability that the coin come from box Q_i.

Given $P(Q_1) = 0.1$, $P(Q_2) = 0.2$, $P(Q_3) = 0.3$, $P(Q_4) = 0.4$

$P(G/Q_1) = 0.2$, $P(G/Q_2) = 0.3$, $P(G/Q_3) = 0.1$, $P(G/Q_4) = 0.5$

(*i*) $$P(G) = \sum_{i=1}^{4} P(Q_i).P(G/Q_i)$$

$$= (0.1)\,(0.2) + (0.2)\,(0.3) + (0.3)\,(0.1) + (0.4)\,(0.5) = 0.31.$$

(*ii*) $$P(Q_2/G) = \frac{P(Q_2).P(G/Q_2)}{P(G)} \quad \text{(by Baye's theorem)}$$

$$= \frac{(0.2)\,(0.3)}{0.31} = 0.194.$$

Example 13. *In a factory manufacturing bulbs, machines numbered 1, 2, 3 manufacture respectively 20, 45 and 35 percent of the total output. Of their outputs 3, 5 and 4 percent respectively are defective. A bulb is drawn at random from the total output and is found to be defective. Find the probability that it was manufactured by the machine numbered.*

Solution. Let M_i = Event that the bulb is produced by machine i. $(i = 1,2,3)$

$\therefore$ $P(M_1) = 0.2$, $P(M_2) = 0.45$, $P(M_3) = 0.35$

Let A = Event that the bulb is defective.

$\therefore$ $P(A/M_1) = 0.03$, $P(A/M_2) = 0.05$, $P(A/M_3) = 0.04$.

Then using Baye's theorem we calculate the following probabilities.

P(machine 1 producing defective bulb)

$$= P(M_1/A)$$

$$= \frac{P(M_1).P(A/M_1)}{P(M_1).P(A/M_1) + P(M_2).P(A/M_2) + P(M_3).P(A/M_3)}$$

$$= \frac{0.2 \times 0.03}{0.2 \times 0.03 + 0.45 \times 0.05 + 0.35 \times 0.04}$$

$$= 0.14$$

Similarly, $P(M_2/A) = 0.53$ and $P(M_3/A) = 0.33$.

Example 14. *Two persons A and B fire at a target independently and have a probability 0.65 and 0.72 respectively of hitting the target. Find the probability that the target is destroyed.*

NOTES

Solution. Let E_1 = Event that target is hit by A

E_2 = Event that target is hit by B.

Given $P(E_1) = 0.65$, $P(E_2) = 0.72$

$P(\bar{E}_1)$ = Probability of failure to hit by A

$= 1 - P(E_1) = 0.35$

and $P(\bar{E}_2) = 1 - P(E_2) = 0.28$

P(Both of them fails) $= P(\bar{E}_1).P(\bar{E}_2) = 0.35 \times 0.28 = 0.098$

Therefore, P(Target is destroyed)

= Probability that at least one of them hit

= 1 – P(Both of them fails)

= 1 – 0.098 = 0.902.

Alternative Method.

P(Target is destroyed) $= P(E_1+E_2)$

$= P(E_1) + P(E_2) - P(E_1E_2)$

$= P(E_1) + P(E_2) - P(E_1)\,P(E_2)$

$= 0.65 + 0.72 - (0.65 \times 0.72) = 0.902.$

SUMMARY

- The single performance of a random experiment will be called an outcome.
- Axioms of Mathematical Probability

  Let A be an event in a sample space S, then:

  (*i*) $0 \leq P(A) \leq 1$

  [Here P(A) = 0 means that the event will not occur and P(A) = 1 means that the event is certain]

  (*ii*) $P(S) = 1$

  (*iii*) If A and B are two mutually exclusive events then

  $P(A + B) = P(A) + P(B)$
- If two events A and B are mutually exclusive, then the occurrence of either A or B is given by

  $$P(A + B) = P(A) + P(B)$$
- When the two events A and B are not mutually exclusive, then the probability of occurrence of at least one of the 2 events is given by

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

- Let A and B be two events in a sample space S and $P(A) \neq 0$, $P(B) \neq 0$, then the probability of happening of both the events are given by

  (*i*) $P(AB) = P(A).P(B/A)$

  (*ii*) $P(AB) = P(B).P(A/B)$.

- Two events A and B in a sample space S are said to be independent if $P(AB) = P(A).P(B)$.

- If A and B are two independent events, then

$$P(A) = P(A / B) = P(A / \overline{B})$$
$$P(B) = P(B / A) = P(B / \overline{A}).$$

PROBLEMS

1. Let A, B and C be three mutually and exhaustive events. Find P(B), if $\frac{1}{3}.P(C) = \frac{1}{2} P(A) = P(B)$.

2. If $P(A) = 0.50$, $P(B) = 0.40$ and $P(A \cup B) = 0.70$, find P(A/B) and $P(\overline{A} \cup B)$. State whether A and B are independent.

3. Given $P(A) = \frac{1}{4}$, $P(A / B) = \frac{1}{4}$ and $P(B / A) = \frac{1}{2}$, find if

   (*i*) A and B are mutually exclusive, (*ii*) A and B are independent.

4. Out of the numbers 1 to 100, one is selected at random. What is the probability that it is divisible by 7 or 8 ?

5. A committee of 4 people is to be appointed from 3 offices of the production department, 4 officers of the purchase department, 2 officers of the sales department and 1 chartered accountant. Find the probability of forming the committee in the following manner.

   (*i*) There must be one from each category.

   (*ii*) The committee should have at least one from the purchase department.

   (*iii*) The chartered accountant must be in the committee.

6. What is the probability that a leap year selected at random will contain either 53 Thursdays or 53 Fridays ?

7. A candidate is selected for interview for three posts. For the first post there are 6 candidates, for the second there are 9 candidates and for the third there are 5 candidates. What are the chances for his getting at least one post ?

8. The odds that person A speaks the truth are 3 : 2 and the odds that person B speaks the truth are 5 : 3. In what percentage of cases are they likely to contradict each other in stating the same fact ?

9. In an examination, 30% of the students have failed in Engineering Mechanics, 20% of the students have failed in Mathematics and 10% have failed in both the subjects. A student is selected at random.

   (*i*) What is the probability that the student has failed in Engineering Mechanics if it is known that he has failed in Mathematics ?

   (*ii*) What is the probability that the student has failed either in Engineering mechanics or in Mathematics ?

NOTES

10. X can solve 80% of the problems while Y can solve 90% of the problems given in a Statistic book. A problem is selected at random. What is the probability that at least one of them will solve the same ?

11. A box P has 1000 items of which 100 are defective. Another box Q has 500 items of which 20 are defective. The items of both the boxes are mixed and one item is randomly taken out. It is found to be defective. What is the probability that the item belongs to box P ?

12. Villages A, B, C and D are connected by overhead telephone lines joining AB, AC, BC, BD and CD. As a result of severe gales, there is a probability p (the same for each link) that any particular link is broken. Find the probability of making a telephone call from A to B.

13. A man seeks advise regarding one of two possible courses of action from three advisers, who arrive at their recommendations independently. He follows the recommendation of the majority. The probabilities that the individual advisers are wrong are 0.1, 0.05 and 0.05 respectively. What is the probability that the man takes incorrect advise ?

14. An unbiased coin is tossed four times in succession and a man scores 2 or 1 according to the coin as shows head or tail in each throw. E_1, E_2 are the following events : E_1 : total score is even, E_2 : total score is divisible by 3. Determine whether E_1 and E_2 are independent or not.

15. There are two identical boxes containing respectively 4 white and 3 red balls and 3 white and 7 red balls. A box is chosen at random and a ball is drawn from it. Find the probability that the ball is white. If the ball is white, what is the probability that it is from the first box ?

16. A speaks truth 4 out of 5 times. A die is tossed. He reports that there is a six. What is the chance that actually there was six ?

17. If a machine is set correctly it produces 10% defective items. If it is set incorrectly then it produces 10% good items. Chances for a setting to be correct and incorrect are in the ratio 7 : 3. After a setting is made, the first two items produced are found to be good items. What is the chance that the setting was correct ?

18. An insurance company insured 2000 scooter drivers, 4000 car drivers and 6000 truck drivers. The probability of accident is 0.01, 0.03 and 0.15 respectively. One of the insured persons meets an accident, what is the probability that he is a car driver ?

19. A bag has 5 red and 4 green balls, a second bag has 4 red and 6 green balls. One ball is drawn from the first and two from the second. Find the probability that out of three balls

    (*i*) all three balls are red,

    (*ii*) all three balls are green,

    (*iii*) two are red and one is green.

20. Suppose that if a person travels to India, the probability that he will see Delhi is 0.70, the probability that he will see Kolkata is 0.64, the probability that he will see Hyderabad is 0.58, the probabilities that he will see Delhi and Kolkata is 0.42, Delhi and Hyderabad is 0.51, Kolkata and Hyderabad is 0.40 and the probability that he will see all the three cities is 0.21. What is the probability that a person travelling to India will see at least one of these three cities ?

21. Urn I contains 2 white and 4 black balls and urn II contains 4 white and 4 black balls. If a ball is drawn at random from one of the two urns, what is the probability that it is a white ball ?

22. A study of daily rainfall at a place, has shown that in July, the probability of a rainy day following a rainy day is 0.4, a dry day following a dry day is 0.7, a rainy day following a dry day is 0.3 and a dry day following a rainy day is 0.6. If it is observed that a certain July day is rainy, what is the probability that the next two days will also be rainy ?

23. The inflow in a cylindrical water tank in a house is equally likely, to fill 6, 7 or 8 ft of the tank. According to the demand the outflow is 5, 6 or 7 ft. Suppose the water level at the tank is 6 ft at the start of the day.

    (*i*) What are the possible water level in the tank at the end of the day ?

    (*ii*) What is the probability that there will be at least 8 ft of water remaining in the tank at the end of the day ?

NOTES

ANSWERS

1. $P(B) = \frac{2}{7}$
2. $P(A/B) = \frac{1}{2}$, $P(\overline{A} \cup B) = 0.7$, Independent
3. (*i*) Not mutually exclusive, (*ii*) Independent
4. $\frac{1}{4}$
5. (*i*) $\frac{8}{70}$, (*ii*) $\frac{13}{14}$, (*iii*) $\frac{2}{5}$
6. $\frac{3}{7}$
7. $\frac{11}{27}$
8. 47.5% of cases
9. (*i*) $\frac{1}{2}$, (*ii*) 0.40
10. $\frac{49}{50}$
11. $\frac{5}{6}$
12. $1 - 2p^2 + p^3$
13. 0.01175
14. Not independent
15. $\frac{61}{140}$, $\frac{40}{61}$ (Use Baye's theorem)
16. $\frac{4}{9}$ (Use Baye's theorem)
17. 0.99 (Use Baye's theorem)
18. 0.12 (Use Bayes' theorem)
19. (*i*) $\frac{2}{27}$, (*ii*) $\frac{4}{27}$, (*iii*) $\frac{16}{45}$
20. 0.80
21. 0.4166.
22. 0.16
23. (*i*) 5, 6, 7, 8, 9 ft, (*ii*) 1/3.

NOTES

CHAPTER 7 PROBABILITY DISTRIBUTIONS

★ STRUCTURE ★

- Random Variable
- Characteristics of Probability Distributions
- Summary
- Problems

RANDOM VARIABLE

A random variable X is a function whose domain is the sample space S and taking a value in the range set which is the real line with chance.

If the sample space consists of discrete elements then the r.v. is called **discrete r.v.**

If the sample space consists of continuous elements then the r.v. is called **continuous r.v.**

The distribution given by the random variable is called probability distribution. Again on the type of the r.v., the probability distribution is called discrete distribution or continuous distribution.

Any discrete distribution is represented by probability mass function (*pmf*). For example,

x	−1	0	1
$p(x)$	0.2	0.4	0.4

is a discrete distribution. Here the random variable X only takes the values −1, 0 and 1 with probability 0.2, 0.4 and 0.4 respectively.

The characteristic of *pmf* is

(*i*) $p(x) \geq 0$ for all x

(*ii*) $\sum_x p(x) = 1$

Any continuous distribution is represented by probability density function (*pdf*). For example,

$$f(x) = 1, \qquad 0 < x < 1$$
$$= 0, \qquad \text{elsewhere}$$

is a continuous distribution. Here the random variable can take any value between 0 and 1 with probability = 1, for any other value the probability = 0.

The characteristic of *pdf* is

(*i*) $f(x) \geq 0$ for all x

(*ii*) $\int_{-\infty}^{\infty} f(x)dx = 1.$

NOTES

CHARACTERISTICS OF PROBABILITY DISTRIBUTIONS

(*a*) **Distribution function.** For discrete case, the distribution function denoted by $F(x)$ is defined as

$$F(x) = P[X \leq x]$$

For the above example, the distribution function is given by

x	−1	0	1
$F(x)$	0.2	0.6	1

For continuous case, the distribution function is defined as

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(x)\,dx$$

For the above example,

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{x} f(x)\,dx$$

$$= 0 + \int_{0}^{x} 1.\,dx = x.$$

Properties:

(*i*) Discrete case

$$P(a < X < b) = F(b) - F(a)$$

$$P(a \leq X < b) = P[X = a] + [F(b) - F(a)]$$

$$P(a < X \leq b) = [F(b) - F(a)] - P[X = b]$$

$$P(a \leq X \leq b) = F(b) - F(a) + P[X = a] - P[X = b]$$

Continuous case

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = F(b) - F(a).$$

(*ii*) $$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

$$F(\infty) = \lim_{x \to \infty} F(x) = 1$$

(*iii*) $F(x) \leq F(y)$ whenever $x < y$.

(*iv*) $F(a) - F(a - 0) = P[x = a]$ and $F(a + 0) = F(a)$

(*v*) For continuous case, $F'(x) = f(x) \geq 0 \Rightarrow F(x)$ is nondecreasing function.

NOTES

(b) Mean/expectation. Let X be the random variable. Then mean/expectation is defined as

$$\mu = E[X] = \sum_x x.p(x) \qquad \text{(Discrete case)}$$

$$= \int_{-\infty}^{\infty} x f(x)dx \qquad \text{(Continuous case)}$$

This expectation is sometimes called as 'Population Mean'.

Properties :

(*i*) $E[X + Y] = E[X] + E[Y]$

(*ii*) $E[cX] = c\ E[X]$

(*iii*) $E[c] = c$ and $E[X + c] = E[X] + c$

(*iv*) If X and Y are independent then $E(XY) = E(X).\ E(Y)$

(*v*) Physically, expectation represents the centre of mass of the probability distribution.

(c) Variance. Variance of a probability distribution is given by

$$\sigma^2 = V[X] = \sum_x (x-\mu)^2\ p(x) \qquad \text{(Discrete case)}$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2\ f(x)\ dx \qquad \text{(Continuous case)}$$

Alt. $\sigma^2 = E[(X - \mu)^2] = E[X^2] - \{E[X]\}^2$

$$= \sum_x x^2.p(x) - \mu^2 \qquad \text{(Discrete case)}$$

$$= \int_{-\infty}^{\infty} x^2.f(x)dx - \mu^2 \qquad \text{(Continuous case)}$$

Properties :

(*i*) $V[aX + b] = a^2\ V[X]$

(*ii*) Physically variance represents the moment of inertia of the probability mass distribution about a line through the mean perpendicular to the line of the distribution.

Example 1. *For the following distribution*

x	1	2	3	4	5
$p(x)$	0.1	k	.2	$3k$	.3

(*i*) *Find the value of k.*

(*ii*) *Find the mean and variance.*

(*iii*) *Find the distribution function.*

Solution. (*i*) Since this is a *pmf,* we have

$$\Sigma\, p(x) = 1$$

$\Rightarrow \quad 0.1 + k + 0.2 + 3k + 0.3 = 1$

$\Rightarrow \quad 4k + 0.6 = 1$

$\Rightarrow \quad 4k = 0.4$

$\Rightarrow \quad k = 0.1.$

NOTES

(*ii*) μ = Mean = $\Sigma x.p(x) = 1(0.1) + 2(0.1) + 3(0.2) + 4(0.3) + 5(0.3) = 3.6$

σ^2 = Variance = $\Sigma(x-\mu)^2 p(x)$

$= (1 - 3.6)^2\ (0.1) + (2 - 3.6)^2\ (0.1) +$

$(3 - 3.6)^2\ (0.2) + (4 - 3.6)^2\ (0.3)$

$+ (5 - 3.6)^2\ (0.3)$

$= 1.64.$

(*iii*) Distribution function is given as follows :

x	1	2	3	4	5
$p(x)$	0.1	0.2	0.4	0.7	1

Example 2. *Find the mean, variance and distribution function of the pdf*

$f(x) = ax^2, 0 \le x \le 1$

$= 0, \quad elsewhere.$

Solution. Since this is a *pdf*, then $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$\Rightarrow \quad \int_0^1 ax^2\,dx = 1$

$\Rightarrow \quad a\left[\frac{x^3}{3}\right]_0^1 = 1$

$\Rightarrow \quad a = 3$

So the *pdf* can be taken as

$f(x) = 3x^2, 0 \le x \le 1$

$= 0, \quad$ elsewhere

$$\mu = \text{Mean} = \int_0^1 x.f(x)\,dx = 3\int_0^1 x^3\,dx = 3\left[\frac{x^4}{4}\right]_0^1 = \frac{3}{4}.$$

$$\sigma^2 = \text{Variance} = \int_0^1 \left(x - \frac{3}{4}\right)^2 .3x^2\,dx$$

$$= \frac{3}{16}\int_0^1 [16x^4 - 24x^3 + 9x^2]\,dx$$

$$= \frac{3}{16}\left[\frac{16}{5} - \frac{24}{4} + \frac{9}{3}\right] = \frac{3}{80}$$

NOTES

$$\text{Distribution function} = 3\int_0^x x^2 dx, \quad 0 \le x \le 1$$

$$= \begin{cases} x^3 , & 0 \le x \le 1 \\ 0 , & \text{elsewhere.} \end{cases}$$

(*d*) **Moments and moment generating function.** *r*th moment about the mean is defined as

$$\mu_r = E[(X - \mu)^r] = \sum_x (x - \mu)^r . P[X = x] \qquad \text{(Discrete case)}$$

$$= \int_{-\infty}^{\infty} (x - \mu)^r . f(x)\, dx \qquad \text{(Continuous case)}$$

Here $\mu_0 = 1$ and $\mu_1 = 0$ for all random variables.

These are called central moments. If we take moments about any point '*a*', then the moments are called raw moments and is denoted by μ'_r and $\mu'_0 = 1$, $\mu'_1 = \mu$. The moment generating function (*m.g.f.*) about a point '*a*' is defined as

$$M_X(t) = \sum_x e^{t(x-a)} . p(x) \qquad \text{(Discrete case)}$$

$$= \int_{-\infty}^{\infty} e^{t(x-a)} . f(x)\, dx \qquad \text{(Continuous case)}$$

Now consider the discrete case,

$$M_X(t) = \Sigma p_i\, e^{t(x_i - a)}, \text{ denoting } p(x) \text{ by } p_i$$

$$= \Sigma p_i \left[1 + t(x_i - a) + \frac{t^2}{2!}.(x_i - a)^2 + \ldots.\right]$$

$$= \Sigma p_i + t \Sigma p_i (x_i - a) + \frac{t^2}{2!} \Sigma p_i (x_i - a)^2 + \ldots.$$

$$= 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \ldots.. + \frac{t^r}{r!}.\mu'_r + \ldots$$

Therefore, μ'_r = Coefficient of $\frac{t^r}{r!}$ in the expansion of $M_X(t)$.

Alternatively, we have

$$\mu'_r = \left[\frac{d^r}{dt^r}.M_X(t)\right]_{t=0}$$

The relations between the central moments and raw moments are as follows:

$$\mu_2 = \mu'_2 - \mu^2 = \text{Variance}$$
$$\mu_3 = \mu'_3 - 3\mu'_2.\mu + 2\mu^3$$
$$\mu_4 = \mu'_4 - 4\mu'_3.\mu + 6\mu'_2\,\mu^2 - 3\mu^4$$

(The moments obtained from a distribution (discrete/continuous) is called "Population Moments").

Note. 1.For *m.g.f.* if the point a is not given, it is taken as zero.

2. A r.v. X may have no moments although its *m.g.f.* exists.

NOTES

e.g., $f(x) = \dfrac{1}{(x+1)(x+2)}, x = 0, 1, 2, \ldots$ (the reader can verify).

3. A r.v. X may have moments although its *m.g.f.* fail to generate the moments.

Example 3. *The pdf of Rayleigh distribution is given by*

$$f(x) = \begin{cases} \dfrac{x}{a^2} \cdot \exp\left(-\dfrac{x^2}{2a^2}\right), & x \geq 0 \\ 0 & , x < 0 \end{cases}$$

Find the distribution function, mean and variance.

Solution. For $x \geq 0$, $\int_0^x \frac{x}{a^2} \cdot \exp\left(-\frac{x^2}{2a^2}\right) dx = \int_0^x \exp\left(-\frac{x^2}{2a^2}\right) . d\left(\frac{x^2}{2a^2}\right)$

$$= 1 - \exp\left(-\frac{x^2}{2a^2}\right).$$

$\therefore$ Distribution function $= \begin{cases} 1 - \exp\left(-\dfrac{x^2}{2a^2}\right), & x \geq 0 \\ 0 & , x < 0 \end{cases}$

$$\mu = \int_0^\infty \frac{x^2}{a^2} \cdot \exp\left(-\frac{x^2}{2a^2}\right) dx$$

$$\left[\text{Put } \frac{x^2}{2a^2} = t \text{ so that } xdx = a^2 dt\right]$$

$$= a\sqrt{2} \int_0^\infty e^{-t} \cdot t^{1/2} dt$$

$$= a\sqrt{2} \int_0^\infty e^{-t} \cdot t^{3/2 - 1} dt$$

$$= a\sqrt{2} \cdot \boxed{3/2}$$

$$= a\sqrt{2} \cdot \frac{1}{2} \cdot \sqrt{\pi} = \sqrt{\frac{a^2\pi}{2}}$$

Now, $\mu'_2 = \int_0^\infty x^2 \cdot f(x) dx$ (taking the any point $a = 0$ in raw moment)

NOTES

$$= \int_0^\infty \frac{x^3}{a^2} \cdot \exp\left(-\frac{x^2}{2a^2}\right) dx$$

$$\left(\text{Put } \frac{x^2}{2a^2} = t \text{ so that } xdx = a^2 dt\right)$$

$$= 2a^2 \int_0^\infty te^{-t} dt = 2a^2 \left[-te^{-t} - e^{-t}\right]_0^\infty = 2a^2.$$

$$\therefore \quad \text{Variance} = \mu'_2 - \mu^2$$

$$= 2a^2 - \frac{a^2\pi}{2}$$

$$= \left(2 - \frac{\pi}{2}\right)a^2.$$

(e) Skewness and Kurtosis.

$$\text{Skewness} = \beta_1 = \frac{\mu_3^2}{\mu_2^3}, \qquad \gamma_1 = \sqrt{\beta_1}$$

For symmetric distribution, $\beta_1 = 0$. If $\beta_1 > 0$ then the distribution is called positively skewed. If $\beta_1 < 0$, then the distribution is called negatively skewed.

$$\text{Kurtosis} = \beta_2 = \frac{\mu_4}{\mu_2^2}, \quad \gamma_2 = \beta_2 - 3$$

For $\beta_2 < 3$, the distribution is called platykurtic.

$\beta_2 = 3$, the distribution is called mesokurtic.

$\beta_2 > 3$, the distribution is called leptokurtic.

Example 4. *A continuous random variable X has a pdf*

$$f(x) = 3x^2, 0 \le x \le 1$$

Obtain the first four central moments and hence calculate β_1 and β_2.

Solution.

$$\mu'_1 = \int_0^1 x.f(x)\,dx = 3\int_0^1 x^3\,dx = \frac{3}{4}$$

$$\mu'_2 = \int_0^1 x^2.f(x)\,dx = 3\int_0^1 x^4\,dx = \frac{3}{5}$$

$$\mu'_3 = \int_0^1 x^3.f(x)\,dx = 3\int_0^1 x^5\,dx = \frac{1}{2}$$

$$\mu'_4 = \int_0^1 x^4.f(x)\,dx = 3\int_0^1 x^6\,dx = \frac{3}{7}.$$

Now,

$$\mu_1 = \mu'_1 = \frac{3}{4} = 0.75$$

$$\mu_2 = \mu'_2 - (\mu_1)^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80} = 0.0375$$

$$\mu_3 = \mu'_3 - 3\mu'_2.\mu_1 + 2\mu_1^3$$

$$= \frac{1}{2} - 3.\frac{3}{5}.\frac{3}{4} + 2.\left(\frac{3}{4}\right)^3 = -\frac{1}{160} = 0.00625$$

$$\mu_4 = \mu'_4 - 4\mu'_3.\mu_1 + 6\mu'_2.\mu_1^2 - 3\mu_1^4$$

$$= \frac{3}{7} - 4.\frac{1}{2}.\frac{3}{4} + 6.\frac{3}{5}.\frac{9}{16} - 3.\frac{81}{256}$$

$$= \frac{3}{7} - \frac{3}{2} + \frac{81}{40} - \frac{243}{256} = 0.00435$$

Therefore, $$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.00625)^2}{(0.0375)^3} = 0.74$$

Since $\beta_1 > 0$, the distribution is positively skewed.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{0.00435}{(0.0375)^2} = 3.09$$

Since $\beta_2 > 3$, the distribution is leptokurtic.

Example 5. *Find the moment generating function of the following distribution.*

$$P[X = x] = q^x p, x = 0,1,2,\ldots..,\ 0 < p \leq 1,\ q = 1 - p$$
$$= 0, \qquad \text{otherwise.}$$

Hence find the mean and variance.

Solution. The given distribution is a discrete distribution.

$$M_X(t) = E[e^{tx}]$$

$$= \sum_{x=0}^{\infty} e^{tx}.q^x.p = p \sum_{x=0}^{\infty} \left(e^t.q\right)^x$$

$$= p\ (1 - qe^t)^{-1}$$

$$= \frac{p}{1 - qe^t} \text{ which is the } m.g.f.$$

$$\mu'_1 = \left[\frac{d}{dt}.M_X(t)\right]_{t=0}$$

$$= \left[\frac{d}{dt}.p(1 - qe^t)^{-1}\right]_{t=0} = \left[pqe^t(1 - qe^t)^{-2}\right]_{t=0}$$

$$= pq(1 - q)^{-2} = \frac{pq}{p^2} = \frac{q}{p} \qquad (\because\ p = 1 - q)$$

NOTES

$$\mu'_2 = \left[\frac{d^2}{dt^2}.M_X(t)\right]_{t=0}$$

$$= pq\left[\frac{d}{dt}.\{e^t(1-qe^t)^{-2}\}\right]_{t=0} \quad \text{(from the previous)}$$

$$= pq\left[e^t(1-qe^t)^{-2} - 2e^t(1-qe^t)^{-3}.(-qe^t)\right]_{t=0}$$

$$= pq\left[(1-q)^{-2} + 2q(1-q)^{-3}\right] = \frac{q}{p} + \frac{2q^2}{p^2}$$

$$\therefore \quad \text{Mean} = \mu'_1 = \frac{q}{p}$$

$$\text{Variance} = \mu_2 = \mu'_2 - (\mu'_1)^2 = \frac{q}{p} + \frac{2q^2}{p^2} - \frac{q^2}{p^2} = \frac{q}{p} + \frac{q^2}{p^2} = \frac{pq+q^2}{p^2} = \frac{q}{p^2}.$$

(*f*) Mean Deviation (M.D.).

$$\text{M.D.} = \int_{-\infty}^{\infty} |x-\mu|.f(x)\,dx \quad \text{(Continuous case)}$$

$$= \sum_x |x-\mu|.P[X=x] \quad \text{(Discrete case)}$$

(*g*) Median. By definition, the median is the point which divides the entire distribution into two equal parts. For *pdf*, median divides the total area into two equal parts. Then

$$\int_{-\infty}^{M} f(x)\,dx = \frac{1}{2} \quad or, \quad \int_{M}^{\infty} f(x)\,dx = \frac{1}{2}$$

By solving we obtain the value of median.

(*h*) Mode. Mode is the value of x, for which $f(x)$ is maximum and it is given by

$f'(x) = 0$, $f''(x) < 0$ and it should lie in the interval of the distribution.

(*i*) Quartiles. For *pdf*, the quartiles Q_1 and Q_3 are given by

$$\int_{-\infty}^{Q_1} f(x)\,dx = \frac{1}{4} \text{ and } \int_{Q_3}^{\infty} f(x)\,dx = \frac{1}{4}.$$

Note : Let us define two special probability.

(*i*) The probability that a specified magnitude (K) will be exceeded *i.e.*, $P[X > K] = p_0$ is called 'Exceedance Probability'.

(*ii*) Median Exceedance Probability : In a sample of estimates of exceedance probability of a specified magnitude, this is the value that is exceeded by 50 percent of the estimates.

Example 6. *For the following distribution, find the median*

$$f(x) = 3x^2, 0 \le x \le 1.$$

Solution. Let M be the median, then

NOTES

$$\int_{-\infty}^{M} f(x)\,dx = \frac{1}{2}$$

$$\Rightarrow \quad 3\int_{0}^{M} x^2\,dx = \frac{1}{2}$$

$$\Rightarrow \quad \left[x^3\right]_0^M = \frac{1}{2}$$

$$\Rightarrow \quad M^3 = \frac{1}{2} \Rightarrow M = \left(\frac{1}{2}\right)^{1/3} = 0.79.$$

Example 7. *Given the probability distribution, calculate the mean deviation.*

x	0	1	2	3	4
$p(x)$	0.1	0.3	0.4	0.1	0.1

Solution. Here $\mu = \Sigma x\,p(x)$

$= 0 + 1(0.3) + 2(0.4) + 3(0.1) + 4(0.1)$

$= 0.3 + 0.8 + 0.3 + 0.4 = 1.8$

Therefore, $\text{M.D.} = \sum_x |x - \mu|.p(x)$

$= |0 - 1.8|.0 + |1 - 1.8|\ (0.3) + |2 - 1.8|$ $(0.4) + |3 - 1.8|\ (0.1)$

$+ |4 - 1.8|\ (0.1) = 0.66.$

Example 8. *Find the quartiles of the following distribution :*

$$f(x) = 3x^2, 0 \le x \le 1.$$

Solution. For lower quartile, $\int_0^{Q_1} f(x)\,dx = \frac{1}{4}$

$$\Rightarrow \quad 3\int_0^{Q_1} x^2\,dx = \frac{1}{4}$$

$$\Rightarrow \quad \left[x^3\right]_0^{Q_1} = \frac{1}{4}$$

$$\Rightarrow \quad Q_1^3 = \frac{1}{4} \Rightarrow Q_1 = \left(\frac{1}{4}\right)^{1/3}$$

For upper quartile, $3\int_{Q_3}^{1} x^2\,dx = \frac{1}{4}$

$$\Rightarrow \quad \left[x^3\right]_{Q_3}^{1} = \frac{1}{4}$$

$$\Rightarrow \quad 1 - Q_3^3 = \frac{1}{4} \Rightarrow Q_3^3 = \frac{3}{4} \Rightarrow Q_3 = \left(\frac{3}{4}\right)^{1/3}.$$

NOTES

SUMMARY

- A random variable X is a function whose domain is the sample space S and taking a value in the range set which is the real line with chance.
- The probability that a specified magnitude (K) will be exceeded *i.e.*, P[X > K] = p_0 is called 'Exceedance Probability'.
- Median Exceedance Probability : In a sample of estimates of exceedance probability of a specified magnitude, this is the value that is exceeded by 50 percent of the estimates.

PROBLEMS

1. A random variable X has the following probability distribution :

x	–1	0	1	2	3
$p(x)$	0.2	0.1	k	$2k$	0.1

(*a*) Find the value of k.

(*b*) Calculate the mean and variance.

(*c*) Calculate the mean deviation.

(*d*) Find P($1 \leq x \leq 3$), P($x > 1$).

2. Given the probability distribution

x	1	2	3	4	5	6
$p(x)$	a	$2a$	a	a	$3a$	$2a$

(*i*) Find the value of a.

(*ii*) Find the distribution function.

(*iii*) Find the mean and variance.

(*iv*) Find P(X < 4), P(2 < X < 5).

3. A random variable X has the probability distribution :

x	–2	–1	0	1	2
$p(x)$	0.2	0.1	0.1	0.3	0.3

Calculate the first four central moments.

4. Consider the following probability distribution :

$f(x) = kx, \qquad 0 < x < 1$

(*a*) Find the value of k.

(*b*) Find the mean and variance.

(*c*) Find the median, Q_1 and Q_3.

5. The Maxwell-Boltzmann distribution is given by

$$f(x) = 4a.\sqrt{\frac{a}{\pi}}.x^2 e^{-ax^2}, \qquad 0 \leq x < \infty,\ a > 0$$

where, $a = \frac{m}{2kT}$, m = Mass, T = Temperature (K), k = Boltzmann constant and x = Speed of a gas molecule.

(*a*) Verify that this is a *pdf*.

(*b*) Calculate the mean and variance.

NOTES

6. The spectrum of the random variable X consists of the points 1, 2, ..., n and $P[X = i]$ is proportional to $\frac{1}{i(i+1)}$. Determine the distribution function of x. Compute $P[3 < X \leq n]$ and $P[X > 5]$.

7. Three balls are drawn without replacement from an urn containing 4 red and 6 white balls. If X is a random variable which denotes the total number of red balls drawn, construct a table showing the probability distribution of X. Also find the expectation.

8. Find the expected value and variance of the number of heads appearing when two fair coins are tossed.

9. Consider the probability density function

$$f(x) = 0.5\,(x + 1), \quad -1 \leq x \leq 1$$
$$= 0, \quad \text{elsewhere}$$

Calculate (*i*) Mean, (*ii*) Median, (*iii*) Q_1, (*iv*) Q_3, (*v*) β_1, (*vi*) β_2, (*vii*) Distribution function.

10. Find the *m.g.f.* of the following distributions :

(*i*) $f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$
$= 0, \quad$ elsewhere.

(*ii*) $p(x) = \frac{e^{-\lambda}.\lambda^x}{x!}, \quad x = 0,1,2,....$

Hence find the mean and variance in (*ii*).

11. In a shipment of 7 items there are three defective items. At random 4 items are selected. What is the expected number of defective items ?

12. The diameter of an electric cable is assumed to be a continuous random variable X with *pdf* $f(x) = 6x(1 - x)$, $0 \leq x \leq 1$. Determine b such that $P[X < b] = P[X > b]$.

ANSWERS

1. (*a*) 0.2 (*b*) 1.1, 1.69 (*c*) 1.1 (*d*) 0.7, 0.5.

2. (*i*) $a = 0.1$ (*ii*)

x	1	2	3	4	5	6
Cum. $P(x)$	.1	.3	.4	.5	.8	1

(*iii*) 3.9, 2.89 (*iv*) 0.4, 0.2

3. $\mu_1 = 0$, $\mu_2 = 2.24$, $\mu_3 = -1.75$, $\mu_4 = 9.03$

4. (*a*) 2, (*b*) $\frac{2}{3}, \frac{1}{18}$, (*c*) $\frac{1}{\sqrt{2}}, \frac{1}{2}, \frac{\sqrt{3}}{2}$

5. (*b*) Mean $= \frac{2}{\sqrt{a\pi}}$, Variance $= \frac{3}{2a} - \frac{4}{a\pi}$

NOTES

6. $F(x) = \frac{x}{x+1}$, $P[3 \le X \le n] = \frac{1}{4} - \frac{1}{n+1}$, $P[X > 5] = \frac{1}{6}$

7.

X	0	1	2	3
P(X)	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

, E[X] = 1.2

8. Expected value = 1, Variance = $\frac{1}{2}$

9. (*i*) 0.33, (*ii*) 0.414, (*iii*) 0; (*iv*) 0.732, (*v*) 0.3172, (*vi*) 2.4,

(*vii*) $F(x) = \frac{1}{4}(x+1)^2, -1 \le x \le 1$

$= 0$, elsewhere

10. (*i*) $\frac{e^{bt} - e^{at}}{t(b-a)}$, (*ii*) $e^{-\lambda} \cdot e^{(\lambda e^t)}$ **11.** 12/7 **12.** $b = \frac{1}{2}$.

CHAPTER 8 SOME PROBABILITY DISTRIBUTIONS

NOTES

★ STRUCTURE ★

- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Other Distributions
- Problems

BINOMIAL DISTRIBUTION

I. If a random variable X takes two values 1 and 0 with probability p and q respectively and $q = 1 - p$ then this is called Bernoulli distribution. Here p is called the probability of success and q is called the probability of failure.

For n trials, the probability of x successes ($x \leq n$) is given by Binomial distribution. The probability mass function is defined as follows :

$$P[X = x] = \binom{n}{x} p^x . q^{n-x}, \quad x = 0,1,.....,n$$

where,

n = No. of independent trials

x = No. of successes

p = Probability of success on any given trial

$q = 1 - p$

$$\binom{n}{x} = {}^nC_x$$

Generally, it is denoted as B(n, p).

II. Properties

(*i*) $$\sum_{x=0}^{n} P[X = x] = 1.$$

(*ii*) Distribution function

NOTES

$$F(x) = P[X \leq x] = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}.$$

(iii) First two moments about origin.

$$\mu_1 = \sum_{x=0}^{n} x \binom{n}{x} p^x . q^{n-x}$$

$$= np. \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} . q^{n-x}$$

$$= np.(p+q)^{n-1} = np.$$

$$\mu_2 = \sum_{x=0}^{n} x^2 . \binom{n}{x} p^x . q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1) + x] \binom{n}{x} p^x . q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1) . \binom{n}{x} p^x . q^{n-x} + np$$

$$= n(n-1) p^2 \sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} . q^{n-x} + np$$

$$= n(n-1) p^2 . (q+p)^{n-2} + np$$

$$= n(n-1) p^2 + np.$$

Now, $\mu_1 = \mu'_1 = np$, which is the mean.

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$= n(n-1)p^2 + np - n^2p^2$$

$$= np - np^2$$

$$= np\ (1-p)$$

$= npq$, which is the variance.

Similarly we obtain μ_3 and μ_4.

(iv) Skewness : $\beta_1 = \dfrac{(1-2p)^2}{npq}$, $\gamma_1 = \dfrac{1-2p}{\sqrt{npq}}$

(v) Kurtosis : $\beta_2 = 3 + \dfrac{1-6pq}{npq}$, $\gamma_2 = \dfrac{1-6pq}{npq}$

(vi) Mode is the value of x for which $P[X = n]$ is maximum.

When $(n+1)\,p$ is not an integer,

Mode = Integral part of $(n+1)p$.

When $(n+1)\,p$ is an integer, we obtain two modes

i.e., Mode = $(n+1)\,p$ and $(n+1)\,p - 1$.

Example 1. *Determine the binomial distribution for which the mean is 8 and variance 4 and find its mode.*

NOTES

Solution. Given that $np = 8$ and $npq = 4$

On division, we get $q = \frac{1}{2} \Rightarrow p = 1 - q = \frac{1}{2}$

Also $n = \frac{8}{p} = 16$

Thus the given binomial distribution is B(16, ½).

Now, $(n + 1)\, p = (16 + 1)\, \frac{1}{2} = \frac{17}{2} = 8 + \frac{1}{2}$

which implies that mode = 8 (integral part only).

Example 2. *The mean and variance of a binomial distribution are 5 and 2 respectively. Find P[X ≤ 1].*

Solution. Given that $np = 5, \quad npq = \frac{5}{2}$

On division we get $q = \frac{1}{2}, \quad p = \frac{1}{2}$

Also, $n = \frac{5}{p} = 10$

$$\therefore \quad P[X \leq 1] = P[X = 0] + P[X = 1]$$

$$= \binom{10}{0} p^0 \cdot q^{10} + \binom{10}{1} p^1 \cdot q^9$$

$$= \left(\frac{1}{2}\right)^{10} + 10.\left(\frac{1}{2}\right).\left(\frac{1}{2}\right)^9 = \frac{11}{1024} = 0.01.$$

Example 3. *In a shooting competition, the probability of a man hitting a target is 2/5. If he fires 5 times, what is the probability of hitting the target (i) at least twice (ii) at most twice.*

Solution. Let p = hitting a target $= \frac{2}{5}, \quad q = 1 - p = \frac{3}{5}, \; n = 5$

(*i*) P[at least twice hitting] = 1 − [P(no hitting) + P(one hitting)]

$$= 1 - \left[\binom{5}{0} p^0 . q^5 + \binom{5}{1} p^1 . q^4\right]$$

$$= 1 - \left[\left(\frac{3}{5}\right)^5 + 5 \cdot \left(\frac{2}{5}\right) \cdot \left(\frac{3}{5}\right)^4\right]$$

$$= 1 - 0.337 = 0.66$$

(*ii*) P[at most twice hitting) = P(no hitting) + P(one hitting) + P(two hitting)

$$= \binom{5}{0} p^0 . q^5 + \binom{5}{1} p . q^4 + \binom{5}{2} p^2 . q^3$$

$$= \left(\frac{3}{5}\right)^5 + 5 \cdot \left(\frac{2}{5}\right) \cdot \left(\frac{3}{5}\right)^4 + 10 \cdot \left(\frac{2}{5}\right)^2 \cdot \left(\frac{3}{5}\right)^3$$

$$= 0.68.$$

NOTES

Example 4. *If 4 of 12 scooterists do not carry driving licence, what is the probability that a traffic inspector who randomly selects 4 scooterists, will catch*

(i) 1 for not carrying driving licence.

(ii) at least 2 for not carrying driving licence.

Solution. p = Probability that a scooterist does not carry driving licence

$$= \frac{4}{12} = \frac{1}{3}$$

$$q = 1 - p = \frac{2}{3}, \; n = 4$$

(*i*) P (catching one scooterist having no driving licence)

$$= \binom{4}{1} p^1 . q^3$$

$$= 4 \cdot \frac{1}{3} \cdot \left(\frac{2}{3}\right)^3 = \frac{32}{81}.$$

(*ii*) P (catching at least two scooterists having no driving licence)

= 1 – [P(all having licence) + P(1 having no licence)]

$$= 1 - \left[\binom{4}{0} p^0 . q^4 + \binom{4}{1} p . q^3\right]$$

$$= 1 - \left[\left(\frac{2}{3}\right)^4 + 4 \cdot \frac{1}{3} \cdot \left(\frac{2}{3}\right)^3\right]$$

$$= 1 - \frac{48}{81} = \frac{33}{81} = \frac{11}{27}.$$

Example 5. *Fit a binomial distribution to the following distribution:*

x	0	1	2	3	4	5
f	27	14	6	3	0	0

Solution. $\text{Mean} = \frac{\Sigma xf}{\Sigma f} = \frac{35}{50}, \; n = 5$

Therefore, $np = \frac{35}{50} \Rightarrow p = \frac{35}{250} = 0.14$ and $q = 0.86$

The expected frequencies of the fitted binomial distribution can be calculated from

$$50\ (0.86 + 0.14)^5$$

Hence we obtain

x	0	1	2	3	4	5
Expected F	24	19	6	1	0	0

NOTES

PROBLEMS

1. Five prizes are to be distributed among 20 students. Find the probability that a particular student will receive three prizes.
2. Consider an intersection approach in which studies have shown 25% right turns and no left turns. Find the probability of one out of the next four vehicles turning right.
3. The mean and variance of a binomial distribution are 6 and 2 respectively. Find P[X > 1], P[X = 2].
4. In a binomial distribution consisting of 5 independent trials, probabilities of 1 and 2 successes are 0.4096 and 0.2048. Find the parameter p of the distribution.
5. An experiment succeeds twice as often as it fails. What is the probability that in next five trials there will be (*i*) three successes, (*ii*) at least three successes ?
6. A quality control engineer inspects a random sample of 3 calculators from each lot of 20 calculators. If such a lot contains 4 slight defective calculators. What are the probabilities that the inspector's sample will contain
   (*i*) no slight defective calculators,
   (*ii*) one slight defective calculators,
   (*iii*) at least two slight defective calculators.
7. Fit a binomial distribution to the following distribution :

x	0	1	2	3	4	5
f	3	12	21	30	25	9

8. Fit a binomial distribution to the following data:

x	0	1	2	3	4
f	15	12	10	8	5

9. How many tosses of a coin are needed so that the probability of getting at least one head is 87.5% ?
10. A machine produces an average of 20% defective bolts. A batch is accepted if a sample of 5 bolts taken from that batch contains no defective and rejected if the sample contains 3 or more defectives. In other cases, a second sample is taken. What is the probability that the second sample is required ?
11. If the probability of a defective bolt is 0.1, find (*i*) mean, (*ii*) variance, (*iii*) moment coefficient of skewness and, (*iv*) Kurtosis for the distribution of defective bolts in a total of 400.
12. If on an average 1 vessel in every 10 is wrecked, find the probability that out of 5 vessels expected to arrive, at least 4 will arrive safely.

ANSWERS

1. 0.088
2. 0.56
3. 0.999, 0.007
4. $p = \frac{1}{5}$
5. (*i*) $\frac{80}{243}$, (*ii*) $\frac{192}{243}$

NOTES

6. (*i*) $\frac{64}{125}$, (*ii*) $\frac{48}{125}$, (*iii*) $\frac{13}{125}$

7.

x	0	1	2	3	4	5
EF	1	9	25	34	24	7

8.

x	0	1	2	3	4
EF	7	18	17	7	1

9. 3 tosses are required

10. 0.6144

11. (*i*) 40, (*ii*) 36, (*iii*) $\frac{2}{15}$, (*iv*) $\frac{23}{1800}$ 12. $\frac{45927}{50000}$

POISSON DISTRIBUTION

I. When (*i*) the number of trials is indefinitely large *i.e.*, $n \to \infty$,

(*ii*) constant probability of success for each trial is very small *i.e.*, $p \to 0$,

and (*iii*) $np = \lambda$ a finite value.

Poisson distribution is obtained as a limiting case of binomial distribution. The probability mass function is

$$P[X = x] = \frac{e^{-\lambda}.\lambda^x}{x!}, \; x = 0,1,2,.....$$

where λ is called the parameter of this distribution.

Proof. Let $np = \lambda \quad \Rightarrow p = \frac{\lambda}{n}, \; q = 1 - \frac{\lambda}{n}.$

$$\therefore \quad \binom{n}{x} p^x . q^{n-x} \approx \frac{n!}{n-x! \, . \, x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}.$$

when $n \to \infty$

$$\frac{n!}{(n-x)! \, . \, x!} = \frac{(n-k+1)\,(n-k+2)...n}{n^k} \to 1$$

Also, $$ln\left(1 - \frac{\lambda}{n}\right)^{n-x} \simeq (n-x)\left(-\frac{\lambda}{n}\right) \to -\lambda$$

Therefore, $$B\,(n,\, p) \to \frac{e^{-\lambda}.\lambda^x}{x!}$$

II. Properties

(*i*) $$\sum_{x=0}^{\infty} P[X = x] = 1$$

(*ii*) Distribution function

$$F(x) = P[X \le x] = e^{-\lambda} \cdot \sum_{k=0}^{x} \frac{\lambda^k}{k!}, \quad x = 0,1,2,....$$

(*iii*) First two moments about origin.

$$\mu'_1 = \sum_{x=0}^{\infty} x \,.\, P[X=x] = \sum_{x=0}^{\infty} x \,.\, \frac{e^{-\lambda}.\lambda^x}{x!} = \lambda e^{-\lambda}.\sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{x-1!}$$

$$= \lambda \,.\, e^{-\lambda}.\, e^{\lambda} = \lambda.$$

$$\mu'_2 = \sum_{x=0}^{\infty} x^2.\, P[X = x]$$

$$= \sum_{x=0}^{\infty} [x(x-1) + x] \,.\, \frac{e^{-\lambda}.\lambda^x}{x!}$$

$$= \left[\sum_{x=0}^{\infty} x(x-1) \,.\, \frac{e^{-\lambda}.\lambda^x}{x!}\right] + \lambda$$

$$= e^{-\lambda}.\lambda^2 \,.\sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{x-2!} + \lambda$$

$$= e^{-\lambda}.\lambda^2.\, e^{\lambda} + \lambda = \lambda^2 + \lambda$$

$\therefore$ $\mu_1 = \mu'_1 = \lambda$ which is mean.

$\mu_2 = \mu'_2 - (\mu'_1)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$, which is variance.

Similarly we obtain μ_3 and μ_4.

(*iv*) Skewness : $\beta_1 = \frac{1}{\lambda}$, $\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{\lambda}}$.

(*v*) Kurtosis : $\beta_2 = 3 + \frac{1}{\lambda}$, $\gamma_2 = \beta_2 - 3 = \frac{1}{\lambda}$.

(*vi*) Mode :

When λ is not an integer,

Mode = integral part of λ.

When λ is an integer,

Mode = $\lambda - 1$ and λ.

(*vii*) In queueing theory (to be discussed in part B), it can be shown under certain assumptions that the probability of arriving x customers in time t is

$$P_x(t) = \frac{e^{-\lambda t}.\,(\lambda t)^x}{x!}, \quad x = 0,1,2,....$$

which is a Poisson distribution with parameter λt. This is also called Poisson process *i.e.,* the number of customers generated (arriving) until any specific time has a Poisson distribution.

Example 1. *There are 150 misprints in a book of 520 pages. What is the probability that a given page will contain at most 2 misprints ?*

Solution. Here $\lambda = \frac{150}{520}$ and let the misprints follow Poisson distribution.

NOTES

Required probability $= P[X \leq 2]$

$= P[X = 0] + P[X = 1] + P[X = 2]$

$= e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2}{2} \cdot e^{-\lambda}$

$= e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right) = e^{-15/52}\left(1 + \frac{15}{52} + \frac{1}{2}\left(\frac{15}{52}\right)^2\right)$

$= 0.9968.$

Example 2. *Let the probability that an individual suffers a bad reaction from an injection is 0.001. What is the probability that out of 3000 individuals (a) exactly 3, (b) more than 2 individuals will suffer a bad reaction ?*

Solution. Here $\lambda = 3000 \times 0.001 = 3$

(*a*) Required probability $= P[X = 3]$

$= \frac{27 \cdot e^{-3}}{6} = 0.22.$

(*b*) Required probability $= P[X > 2] = 1 - P[X \leq 2]$

$= 1 - (P[X = 0] + P[X = 1] + P[X = 2])$

$= 1 - \frac{17e^{-3}}{2} = 0.58.$

Example 3. *A controlled manufacturing process is 0.2% defective. What is the probability of taking 2 or more defectives from a lot of 100 pieces ? (a) By using binomial distribution. (b) By using Poisson approximation.*

Solution. (*a*) p = Probability of defective = 0.002,

$q = 1 - p = 0.998$

$n = 100$

∴ Probability of finding 2 or more defective

= 1 – [Probability of zero and one defective]

$= 1 - [P[X = 0] + P[X = 1]]$

$= 1 - \left[(0.998)^{100} + 100\,(0.002)\,(0.998)^{99}\right]$

$= 1 - 0.983 = 0.017.$

(*b*) Here $\lambda = np = 100 \times 0.002 = 0.2$

∴ Probability of finding 2 or more defective

$= 1 - [P[X = 0] + P[X = 1]]$

$= 1 - \left[e^{-0.2} + (0.2)\,e^{-0.2}\right]$

$= 1 - 0.982 = 0.018.$

Example 4. *Fit a Poisson distribution to the set of observations :*

x	*0*	*1*	*2*	*3*	*4*
f	*57*	*41*	*28*	*8*	*1*

Solution. $\text{Mean} = \frac{\Sigma xf}{\Sigma f} = \frac{125}{135} = 0.926$

NOTES

$\therefore$ The mean of the poisson distribution is $\lambda = 0.926$

Hence the expected frequencies are given by

$$135 \cdot \frac{e^{-0.926} \cdot (0.926)^x}{x!}, \quad x = 0,1,2,3,4$$

Therefore the fitted distribution is given by

x	0	1	2	3	4
Expected f	53	50	23	7	2

PROBLEMS

1. If X be a Poisson distributed random variable and P[X = 1] = 3P[X = 2], then find P[X > 2].
2. A medicine was supplied in 100 batches (each batch containing a fairly large number of items). A total of 50 items in all the batches were found to be defective. Find the probability that (*a*) a batch has no defective item, (*b*) a batch has at least three defective items.
3. If 2 per cent of electric bulbs manufactured by a certain company are defective, find the probability that in a sample of 200 bulbs (*a*) less than 2 bulbs are defective, (*b*) more than 3 bulbs are defective.
4. Let X follows Poisson distribution, find the value of the mean of the distribution of P[X = 1] = 3P[X = 2].
5. In a certain factory, blades are manufactured in packets of 10. There is a 0.1% probability for any blade to be defective. Using Poisson distribution calculate approximately the number of packets containing two defective blades in a consignment of 10000 packets.
6. Fit a Poisson distribution to the following:

x	0	1	2	3	4	5
f	20	16	11	7	4	2

7. Fit a Poisson distribution to the following:

x	0	1	2	3	4	5
f	120	82	52	22	4	0

8. Records show that the probability is 0.00002 that a car will have a flat tyre while driving over a certain bridge. Use the Poisson distribution to determine the probability that among 20000 cars driven over this bridge, not more than one will have a flat tyre.
9. A typist kept a record of mistakes made per day during 300 working days of a year :

Mistakes per day	0	1	2	3	4	5	6
No. of days	143	90	42	12	9	3	1

Fit an appropriate Poisson distribution to the data.

10. The probability that a Poisson variate X takes a positive value is $\left(1 - e^{-1.5}\right)$. Find the variance and also the probability that X lies between –1.5 and 1.5.
11. A manufacturer, who produces medicine bottles, finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of bottles. Using Poisson distribution, find how many boxes will contain (*i*) no defectives, (*ii*) at least two defectives.
12. What is the probability that in a company of 1000 people only one person will have birth day on new year's day? (Assume that a year has 365 days).

NOTES

ANSWERS

1. 0.0302 2. (*a*) 0.61 (*b*) 0.01

3. (*a*) 0.09 (*b*) 0.58 4. $\frac{2}{3}$ 5. Two packets

6.

x	0	1	2	3	4	5
EF	15	20	15	7	2	1

(The EF corresponding to $x = 1$, has been adjusted to make Σ EF = 60)

7.

x	0	1	2	3	4	5
EF	108	102	49	16	4	1

8. 0.9380

9.

x	0	1	2	3	4	5	6
EF	123	110	49	14	3	1	0

10. Variance = 1.5, $P(-1.5 < X < 1.5) = 2.5\ e^{-1.5}$

11. (*i*) 61, (*ii*) 9 12. 0.18.

NORMAL DISTRIBUTION

1. In the binomial distribution if (*i*) the number of trials is indefinitely large, *i.e.*, $n \to \infty$ and (*ii*) neither p nor q is very small, then the limiting form of the binomial distribution is called 'normal distribution'. It is a continuous distribution and probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-(x-\mu)^2/2\sigma^2}$$

or,
$$= \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp.\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

$-\infty < x < \infty, -\infty < \mu < \infty.$

It is denoted by N (μ, σ^2) where μ = Mean and σ^2 = Variance.

Since this distribution was developed by Carl Friedrich Gauss, sometimes it is also referred as 'Gaussian Distribution'.

II. Properties

(*i*) The normal curve is bell shaped and symmetrical about the line $x = \mu$.

(*ii*) Median : Let M be the median, then

$$\int_{-\infty}^{M} f(x)\,dx = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{M} \exp.\left[-(x-\mu)^2/2\sigma^2\right] dx = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{\mu} \exp.\left[-(x-\mu)^2/2\sigma^2\right] dx + \frac{1}{\sigma\sqrt{2\pi}} \int_{-\mu}^{M} \exp.\left[-(x-\mu)^2/2\sigma^2\right] dx = \frac{1}{2}$$

In the first integral let $z = \frac{x-\mu}{\sigma}$, then we get

$$\frac{1}{\sqrt{2\pi}} \int_0^{\infty} \exp.\left[-z^2/2\right] dz = \frac{1}{2}$$

Therefore, $\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp.\left[-(x-\mu)^2/2\sigma^2\right] dx = \frac{1}{2}$

$$\Rightarrow \quad \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp.\left[-(x-\mu)^2/2\sigma^2\right] dx = 0$$

$$\Rightarrow \quad \mu = M$$

(*iii*) Mode: It is the value of x for which $f(x)$ is maximum *i.e.*, $f'(x) = 0$ and $f''(x) < 0$.

Here we obtain, mode = μ.

Note. Mean, Median and Mode coincides.

(*iv*) Mean deviation (M.D.)

$$\text{M.D.} = \int_{-\infty}^{\infty} |x-\mu| \, . f(x)\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |x-\mu| \, . \exp.\left[-(x-\mu)^2/2\sigma^2\right] dx$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| \exp.\left[-z^2/2\right] dz, \quad \text{taking } z = \frac{x-\mu}{\sigma}$$

$$= \sqrt{\frac{2}{\pi}} \, . \, \sigma \, . \int_0^{\infty} z \, . \exp.\left[-z^2/2\right] dz, \quad (\because \text{ the integrand is even})$$

$$= \sqrt{\frac{2}{\pi}} \, . \, \sigma \, . \int_0^{\infty} e^{-t}\, dt, \qquad \text{taking } \frac{z^2}{2} = t$$

$$= \sqrt{\frac{2}{\pi}} \, . \, \sigma \, . \, 1 = \sqrt{\frac{2}{\pi}} \, . \sigma \approx \frac{4}{5}\, \sigma.$$

(*v*) Q.D : M.D : S.D. = $\frac{2}{3} : \frac{4}{5} : 1$

i.e., 10 : 12 : 15.

(*vi*) Points of inflexion of the normal curve is given at $x = \mu \pm \sigma$.

(*vii*) Central moments : $\mu_{2n+1} = 0$ (odd-order)

and $\mu_{2n} = 1, 3, 5, \ldots, (2n-1)\, \sigma^{2n}$. (even order).

(*viii*) Skewness : $\beta_1 = 0$ ($\because \mu_3 = 0$); $\gamma_1 = 0$.

(*ix*) Kurtosis : $\beta_2 = 3$, $\gamma_2 = \beta_2 - 3 = 0$.

NOTES

(*x*) Distribution function $F(x) = \int_{-\infty}^{x} f(x)\,dx$.

(*xi*) It is a *pdf*, $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

(*xii*) Standard Normal Distribution

If $X \sim N(\mu, \sigma^2)$, then $z = \frac{X - \mu}{\sigma}$ is called standard normal variate with E[*z*] = 0 and var [*z*] = 1 and $z \sim N(0,1)$.

$$\phi(z) = \text{the } pdf = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad -\infty < z < \infty.$$

(*xiii*) $$P[x_1 < X < x_2] = \int_{x_1}^{x_2} f(x)\,dx$$

$$= \int_{0}^{x_2} f(x)\,dx - \int_{0}^{x_1} f(x)\,dx$$

$$= \int_{0}^{z_2} \phi(z)\,dz - \int_{0}^{z_1} \phi(z)\,dz \quad \text{taking } z = \frac{x - \mu}{\sigma}$$

$$= \Phi(z_2) - \Phi(z_1)$$

These values are obtained from the standard normal table.

(*xiv*) Area under the normal curve :

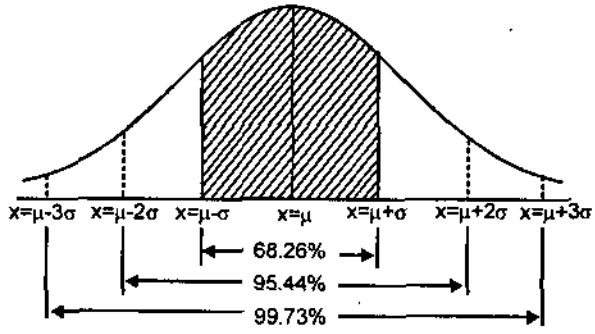


Fig. 8.1

III. Probable Error

Any manufactured items or measurement of any physical quantity shows slight error. All the errors in manufacturing or measurement are random in nature and follow a normal distribution. We define the probable error λ is such that the probability of an error falling within the limits $\mu - \lambda$ and $\mu + \lambda$ is exactly equal to the chance of an error falling outside these limits which implies that the chance of an error lying within $\mu - \lambda$ and $\mu + \lambda$ is ½.

NOTES

$$\Rightarrow \quad \int_{\mu-\lambda}^{\mu+\lambda} f(x)\,dx = \frac{1}{2}, \qquad f(x) = \text{normal } pdf.$$

$$\Rightarrow \quad \int_{0}^{\lambda/\sigma} \phi(z)\,dz = \frac{1}{4} \qquad \left(\text{taking } z = \frac{x-\mu}{\sigma}\right)$$

$$\Rightarrow \quad \frac{\lambda}{\sigma} = 0.6745 \qquad \text{(from normal table)}$$

$$\Rightarrow \quad \lambda = 0.6745\ \sigma \sim \frac{2}{3}\sigma.$$

IV. Normal Approximation To Binomial Distribution

If the number of trials is sufficiently large (*i.e.*, $n \geq 30$) the binomial distribution B(n, p) is approximated by the normal distribution N(μ, σ^2) with $\mu = np$ and $\sigma^2 = npq$. But a continuity correction is required. The discrete integer x in B(n, p) becomes the interval $[x - 0.5, x + 0.5]$ in the N(μ, σ^2). Thus

$$P[X = x] \approx \frac{1}{\sigma\sqrt{2\pi}} \int_{x-0.5}^{x+0.5} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du$$

and

$$P[x_1 < x < x_2] \approx \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1-0.5}^{x_2+0.5} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du$$

(See example 3)

Example 1. *Given a random variable having the normal distribution with $\mu = 18.2$ and $\sigma = 1.25$, find the probabilities that it will take on a value*

(a) less than 16.5,

(b) greater than 18.8,

(c) between 16.5 and 18.8,

(d) between 19.2 and 20.1.

Solution. Let $z = \dfrac{x - 18.2}{1.25}$

(*a*) (Fig. 8.2) $z = \dfrac{16.5 - 18.2}{1.25} = -1.36$

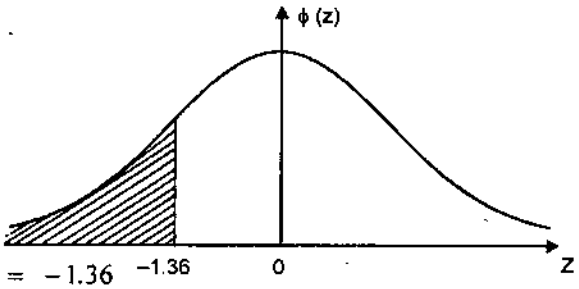


Fig. 8.2

$$P(z < -1.36) = 0.5 - \Phi(1.36)$$
$$= 0.5 - 0.4131$$
$$= 0.0869.$$

(*b*) (Fig. 8.3) $z = \dfrac{18.8 - 18.2}{1.25} = 0.48$

$$P(z > 0.48) = 0.5 - \Phi(0.48)$$

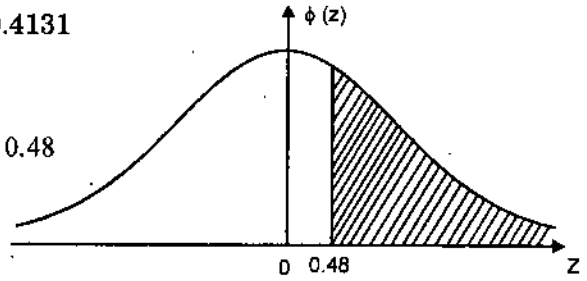


Fig. 8.3

NOTES

$= 0.5 - 0.1844$

$= 0.3156.$

(*c*) (Fig. 8.4) $P(-1.36 < z < 0.48) = \Phi(1.36) + \Phi(0.48)$

$= 0.4131 + 0.1844$

$= 0.5975.$

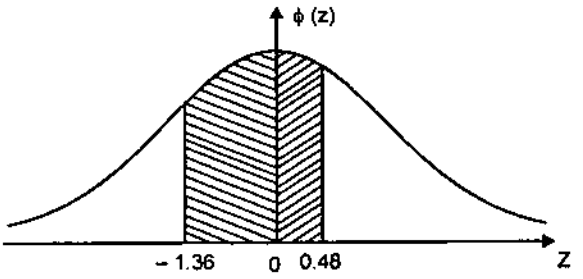


Fig. 8.4

(*d*) (Fig.8.5) When $x = 19.2$, $z_1 = 0.8$

When $x = 20.1$, $z_2 = 1.52$

Here $P(0.8 < z < 1.52) = \Phi(1.52) - \Phi(0.8)$

$= 0.4357 - 0.2881 = 0.1476.$

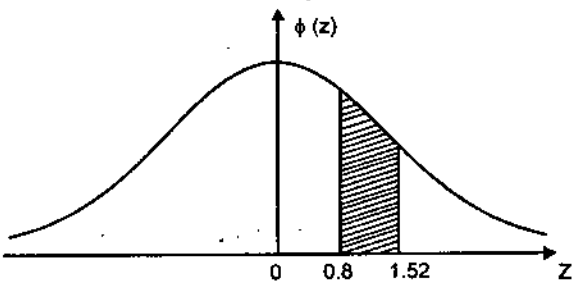


Fig. 8.5

Example 2. *In a large institution 2.28% of employees receive income below Rs. 4500 and 15.87% of employees receive income above Rs. 7500 p.m.. Assuming the income follows normal distribution. Find the mean and S.D. of the distribution.*

Solution. Let μ and σ be the mean and S.D. of the normal distribution.

Given 2.28% of employees receive income below Rs. 4500.

$$\Rightarrow \quad \int_{-\infty}^{z_1} \phi(z)\, dz = 0.0228$$

$$\Rightarrow \quad 0.5 - \Phi(z_1) = 0.0228$$

$$\Rightarrow \quad \Phi(z_1) = 0.4772$$

$$\Rightarrow \quad z_1 = -2 \text{ (from normal table).}$$

Again, 15.87% of employees receive income above Rs. 7500.

$$\Rightarrow \quad \int_{z_2}^{\infty} \phi(z)\, dz = 0.1587$$

$$\Rightarrow \quad 0.5 - \Phi(z_2) = 0.1587$$

$$\Rightarrow \quad \Phi(z_2) = 0.3413$$

$$\Rightarrow \quad z_2 = 1 \text{ (from normal table).}$$

$\therefore$ We obtain two equations as follows :

$$\frac{4500 - \mu}{\sigma} = -2 \qquad \Rightarrow \mu - 2\sigma = 4500$$

and

$$\frac{7500 - \mu}{\sigma} = 1 \qquad \Rightarrow \mu + \sigma = 7500$$

Then the solution gives $\sigma = 1000$ and $\mu = 6500$.

Example 3. *Samples of 40 are taken from a lot, which is on the average 20 percent defective. (a) What is the probability that a sample of 40 will contain exactly 11 defectives ? (b) What is the probability that it will contain 11 or more defectives ?*

Solution. This problem can be solved using normal distribution as an approximation to the binomial.

Here $n = 40, \qquad p = 0.2, \qquad q = 0.8.$

Let $\mu = np = 8$ and $\sigma = \sqrt{npq} = \sqrt{40 \times 0.2 \times 0.8} = 2.53.$

(*a*) Since the normal distribution is continuous, the probability of exactly 11 defectives should be interpreted as meaning the probability of defectives from 10.5 to 11.5.

$\therefore$

$$z_1 = \frac{10.5 - 8}{2.53} = 0.99$$

$$z_2 = \frac{11.5 - 8}{2.53} = 1.38$$

The required probability is

$$\begin{aligned} P(z_1 < z < z_2) &= P(0.99 < z < 1.38) \\ &= \Phi(1.38) - \Phi(0.99) \\ &= 0.4162 - 0.3389 \\ &= 0.0773. \end{aligned}$$

(*b*) The probability of 11 or more defectives should be interpreted to mean the probability of 10.5 or more defectives.

Therefore,

$$\begin{aligned} P(z > 0.99) &= 0.5 - \Phi(0.99) \\ &= 0.5 - 0.3389 \\ &= 0.1611. \end{aligned}$$

Example 4. *Fit a normal curve to the following distribution :*

x	0	1	2	3	4	5
f	5	13	26	32	18	6

NOTES

Solution. $\Sigma f = 100$

$$\text{Mean} = \frac{\Sigma fx}{\Sigma f} = \frac{263}{100} = 2.63$$

$$\text{S.D.} = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - (\text{Mean})^2} = \sqrt{\frac{843}{100} - (2.63)^2} = 1.23$$

Taking $\mu = 2.63$ and $\sigma = 1.23$, the equation of the normal curve is

$$f(x) = \frac{1}{1.23 \cdot \sqrt{2\pi}} . e^{-(x - 2.63)^2/3.026}$$

Let $z = \dfrac{x - 2.63}{1.23}$, then the calculations are presented in the following table:

Mid x	(x_1, x_2)	(z_1, z_2)	*Area between* (z_1, z_2)
0	(–0.5, 0.5)	(–2.54, –1.73)	0.4945 – 0.4582 = 0.0363
1	(0.5, 1.5)	(–1.73, –0.92)	0.4582 – 0.3212 = 0.1370
2	(1.5, 2.5)	(–0.92, –0.11)	0.3212 – 0.0438 = 0.2774
3	(2.5, 3.5)	(–0.11, 0.71)	0.0438 + 0.2611 = 0.3049
4	(3.5, 4.5)	(0.71, 1.52)	0.4357 – 0.2611 = 0.1746
5	(4.5, 5.5)	(1.52, 2.33)	0.4901 – 0.4357 = 0.0544

Expected frequencies (EF) = 100. (Area between (z_1, z_2)).

Therefore, we obtain,

x	0	1	2	3	4	5
EF	4	14	28	31	18	5

Example 5. *Small stones are collected and weights are assumed to be normal. It is found that 5% of the stones are under 30 gm and 80% are under 50 gm. What are the mean and standard deviation of the distribution ?*

Solution. Let weight be represented by a random variable X such that X ~ N (μ, σ^2).

Given that, $P(X < 30) = 0.05$ and

$P(X < 50) = 0.8$ *i.e.*, $P(X > 50) = 0.2$

We obtain the Fig. 8.6.

While converting 30 and 50 into z_1 and z_2 using $z = \dfrac{x - \mu}{\sigma}$ the value z_1 will be in the negative side.

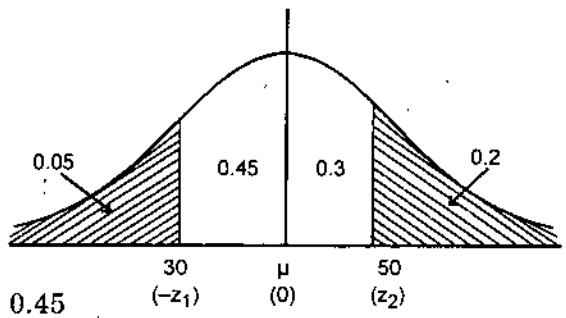


Fig. 8.6

Therefore $\Phi(-z_1) = 0.45$

$\Rightarrow$ $\dfrac{30 - \mu}{\sigma} = -1.65$

$$\Rightarrow \quad \mu - 1.65\sigma = 30 \qquad \ldots(i)$$

and $\Phi(z_2) = 0.3$

$$\Rightarrow \quad \frac{50 - \mu}{\sigma} = 0.85$$

$$\Rightarrow \quad \mu + 0.85\sigma = 50 \qquad \ldots(ii)$$

Solving (i) and (ii) we obtain $\mu = 43.2$ and $\sigma = 8$.

PROBLEMS

1. The diameters of shafts manufactured by a machine are normal random variables with mean 10 cm and standard deviation 1 cm. Show that 81.85 percent of the shafts are expected to these diameters between 9 cm and 12 cm.
2. Tests have indicated that the tensile strengths of certain alloys averages 1885 kg/cm^2 with a standard deviation of 225 kg/cm^2. If the distribution is normal, what percentage of the casting will have tensile strength (*i*) less than 1500 kg/cm^2, (*ii*) more than 1600 kg/cm^2, (*iii*) between 2000 and 2100 kg/cm^2.
3. Assuming that the life in hours of an electric bulb is a random variable following normal distribution with mean of 2000 hours and standard deviation of 500 hours. Find the expected number of bulbs from a sample of 2000 bulbs having life (*i*) more than 2500 hours, (*ii*) between 2600 and 3000 hours.
4. The mean value of the modulus of rupture of a large number of test specimen has been found to be 400 kg/cm^2. If the standard deviation is 70 kg/cm^2 and the distribution is approximately normal, for what percentage of the specimens the modulus of rupture will fall (*i*) between 350 and 450 ? (*ii*) above 300 ?
5. A manufacturer makes chips for use in a mobile phone of which 10% are defective. For a random sample of 200 chips, find the approximate probability that more than 15 are defective.
6. E-mail messages are received by a general manager of a company at an average rate of 1 per hour. Find the probability that in a day the manager receives 24 messages or more.
7. A normal curve has an average of 150.2 and a standard deviation of 3.81. What percentage of the area under the curve will fall between limits of 137.5 and 153.4 ?
8. Suppose that life of a gas cylinder is normally distributed with a mean of 40 days and a S.D. of 5 days. If, at a time, 10,000 cylinders are issued to customers, how many will need replacement after 35 days ?
9. Fit a normal curve to the following :

x	0	1	2	3	4
f	9	21	42	20	8

10. Fit a normal curve to the following :

x	–2	–1	0	1	2	3
f	6	9	24	15	11	5

11. In a distribution exactly normal, 10.03% of the items are under 25 kilogram weight and 97% of the items are under 70 kilogram weight. What are the mean and standard deviation of the distribution ?
12. If the probability of committing an error follows normal distribution, compute the probable error from the following data :

    4.8, 4.2, 5.1, 3.8, 4.4, 4.7, 4.1 and 415.
13. The width of a slot on a forging is normally distributed with mean 0.9″ and standard deviation 0.900″ $\pm$ 0.005″. What percentage of forgings will be defective ?

NOTES

$$\text{Variance} = \exp.\,[2a + b^2].\,[\exp.[b^2] - 1]$$

(g) Hypergeometric distribution. Consider a sample of n units to be drawn from a lot containing N units, of which d are defective.

The x successes (defectives) and $n - x$ failures can be chosen in $\binom{d}{x}\binom{N-d}{n-x}$ ways. Also n items can be chosen from a set of N objects in $\binom{N}{n}$ ways. For sampling without replacement the probability of getting "x successes in n trials" is

$$p\,(x,\ n,\ d,\ N) = \frac{\binom{d}{x}\binom{N-d}{n-x}}{\binom{N}{n}} \quad \text{for } x = 0,\ 1,\ \ldots,\ n$$

$$\text{Mean} = n.\,\frac{d}{N},\ \text{Variance} = \frac{n\,.\,d\,.\,(N-d)\,.\,(N-n)}{N^2\,.\,(N-1)}$$

This distribution approaches to binomial with $p = \frac{d}{N}$ when $N \to \infty$.

(h) Weibull distribution. This is an important distribution used in reliability and life testing of an item. The probability density function is given by

$$f(x) = \begin{cases} \alpha\beta\, x^{\beta-1}.\, e^{-\alpha x^{\beta}}, & \text{for } x > 0,\ \alpha > 0,\ \beta > 0 \\ 0, & \text{elsewhere} \end{cases}$$

where, α and β are two parameters of the distribution.

$$\mu = \int_0^{\infty} x\,.\,\alpha\beta\, x^{\beta-1}.\, e^{-\alpha x^{\beta}}\, dx$$

$$= \alpha^{-1/\beta}\,.\int_0^{\infty} u^{1/\beta}\,.\, e^{-u}\, du \qquad (\text{taking } v = \alpha x^{\beta})$$

$$= \alpha^{-1/\beta}\,.\,\Gamma(1 + 1/\beta)$$

Now $$\mu'_2 = \int_0^{\infty} x^2\,.\,\alpha\beta\, x^{\beta-1}.\, e^{-\alpha x^{\beta}}\, dx$$

$$= \alpha^{-2/\beta}\,.\int_0^{\infty} u^{2/\beta}\,.\, e^{-u}\, du \qquad (\text{taking } \alpha x^{\beta} = u)$$

Using integration by parts,

$$= \alpha^{-2/\beta}\left[\left(\frac{u^{2/\beta}\,.\, e^{-u}}{-1}\right)_0^{\infty} + \frac{2}{\beta}\int_0^{\infty} u^{2/\beta - 1}\,.\, e^{-u}\, du\right]$$

$$= \alpha^{-2/\beta}\left[0 + \frac{2}{\beta} \cdot \Gamma(2/\beta)\right] = \alpha^{-2/\beta} \cdot \Gamma(1 + 2/\beta)$$

Therefore,

NOTES

$$\text{Variance} = \mu_2 = \mu'_2 - (\text{Mean})^2$$

$$= \alpha^{-2/\beta} \cdot \Gamma(1 + 2/\beta) - \alpha^{-2/\beta} \cdot \left[\Gamma(1 + 1/\beta)\right]^2$$

$$= \alpha^{-2/\beta} \cdot \left\{\Gamma(1 + 2/\beta) - \left[\Gamma(1 + 1/\beta)\right]^2\right\}$$

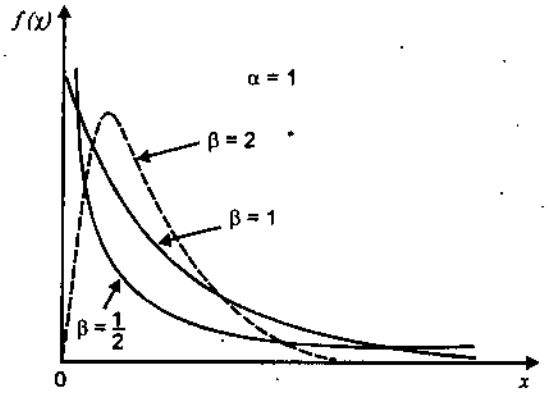


Fig. 8.8 Weibull density

(*i*) Chi-Square (χ^2) distribution. A random variable X is said to follow chi-square distribution if its *pdf* is of the form

$$f(x) = \frac{1}{-(n/2)} \cdot e^{-x/2} \cdot x^{(n/2)-1}, \qquad 0 < x < \infty$$

The parameter n (positive integer) is called the number of degrees of freedom. A variable which follows chi-square distribution is called a chi-square variate.

Properties:

(*i*) The chi-square curve is positively skew.

(*ii*) Mean = n, $\sigma^2 = 2n$, (n = *d.f.*)

(*iii*) If χ^2 is a chi-square variate with n *d.f.*, then $\chi^2/2$ is a Gamma variate with parameter $(n/2)$.

(*iv*) If X_i (i = 1, 2, ..., n) are n independent normal variates with mean μ_i and variance σ_i^2 ($i = 1,2,...,n$) then

$$\chi^2 = \sum_i \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \text{ is a chi-square variate with } n \text{ d.f.}$$

(*v*) $\text{M.G.F.} = (1 - 2t)^{-n/2}, \qquad |2t| < 1$

(*vi*) In practice, for $n \geq 30$, the χ^2 distribution is approximated by normal distribution.

NOTES

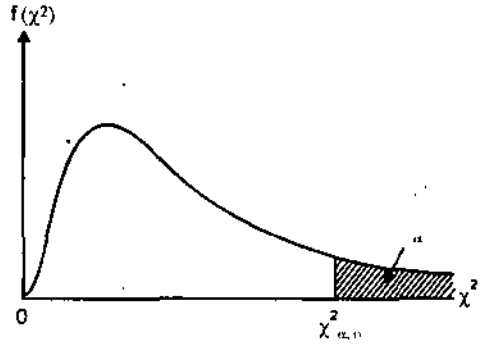


Fig. 8.9 Tabulated values of chi-square distribution

(*j*) Student's t-distribution. A random variable is said to follow student's *t*-distribution or simply *t*-distribution if its *pdf* is given by

$$f(t) = y_0 \cdot \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \qquad -\infty < t < \infty$$

where, y_0 is a constant such that the area under the curve is unity and n is called degrees of freedom.

Properties:

(*i*) The *t*-curve is symmetrical about 0, and leptokurtic *i.e.*, $\beta_1 = 0$, $\beta_2 > 3$.

(*ii*) Mean = 0, Variance $= \frac{n}{n-2}$, $(n > 2)$

(*iii*) For large *d.f.*, the *t*-distribution can be approximated by the standard normal distribution.

(*iv*) Let x_i $(i = 1,2,..., n)$ be the random samples from a normal population having mean μ and variance σ^2, then the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

where, $\bar{x}$ = Sample mean= $\frac{\Sigma x}{n}$ and $S^2 = \frac{1}{n-1}\Sigma\left(x_i - \bar{x}\right)^2$ follows *t*-distribution with $(n - 1)$ *d.f.*

(*k*) F-distribution. A random variable is said to be F-distribution with degrees of freedom (v_1, v_2) if its *pdf* is of the form

$$f(F) = y_0 F^{(v_1/2)-1}(v_2 + v_1 F)^{-(v_1+v_2/2)}, \qquad 0 < F < \infty$$

where, y_0 is a constant such that the area under the curve is unity.

Properties:

(*i*) The F-curve is positively skew.

(*ii*) Mean $= \frac{v_2}{v_2 - 2}$ Variance$= \frac{2(v_1 + v_2 - 2)}{v_1(v_2 - 4)}\left(\frac{v_2}{v_2 - 2}\right)^2$

(*iii*) If y_1 and y_2 are independent chi-square variates with n_1 and n_2 degrees of freedom respectively, then

$$F = \frac{y_1/n_1}{y_2/n_2}$$

follows F-distribution with (n_1, n_2) *d.f.*

SECTION-C

CHAPTER 9 SAMPLING THEORY

NOTES

★ STRUCTURE ★

- Sampling, Population and Samples
- Types of Sampling
- Use of Random Numbers
- Parameter and Statistic
- Sampling Distribution of Mean
- Sampling Distribution of Sample Variance
- Sampling Distribution of Sample Proportion
- Summary
- Problems

SAMPLING, POPULATION AND SAMPLES

Sampling means the selection of a part of the aggregate with a view to draw some statistical informations about the whole. This aggregate of the investigation is called population and the selected part is called sample. A population is finite or infinite according to its size *i.e.*, number of members.

The main objective of the sampling is to obtain the maximum information of the population. The analysis of the sample is done to obtain an idea of the probability distribution of the variable in the population.

Though by applying proper process of sampling we may not be able to represent the characteristics of the population correctly. This discrepancy is called sampling error.

TYPES OF SAMPLING

There are different sampling methods. We describe below some important types of sampling.

(*a*) **Simple random sampling.** In this type of sampling every unit of the population has an equal chance of being selected in a sample. There are two ways of drawing a simple random sample–With Replacement (WR) and Without Replacement (WOR).

NOTES

In WR type, the drawn unit of the population is again returned to the population so that the size of the population remains same before each drawing. In WOR type, the drawn unit of the population is not returned to the population. For finite population the size diminishes as the sampling process continues.

(*b*) **Systematic sampling.** In systematic sampling one unit is chosen at random from the population and the items are selected regularly at predetermined intervals. This method is quite good over the simple random sampling provided there is no deliberate attempt to change the sequence of the units in the population.

(*c*) **Cluster sampling.** When the population consists of certain group of clusters of units, it may be advantageous and economical to select a few clusters of units and then examine all the units in the selected clusters. For example of certain goods which are packed in cartons and repacking is costly it is advisable to select only few cartons and inspect all the inside goods.

(*d*) **Two-stage sampling.** When the population consists of larger number of groups each consisting of a number of items, it may not be economical to select few groups and inspect all the items in the groups. In this case, the sample is selected in two stages. In the first stage, a desired number of groups (primary units) are selected at random and in the second stage, the required number of items are chosen at random from the selected primary units.

(*e*) **Stratified sampling.** Here the population is subdivided into several parts, called strata showing the heterogenity of the items is not so prominent and then a sub sample is selected from each of the strata. All the sub-samples combined together give the stratified sample. This sampling is useful when the population is heterogeneous.

USE OF RANDOM NUMBERS

The random numbers represent a sequence of digits where they appear in a perfectly random order. Selection of a random number from a table of random numbers has the same probability of selection. There are various methods to generate random numbers. Also there are tables of random numbers. Briefly we illustrate the use of random numbers. Let us consider the following two digits random numbers :

23, 04, 82, 07, 14, 66, 54, 10, 72 and 32.

Suppose we have marks of a subject of 100 students and we want to draw a sample of marks of size 10. To draw this, number the students from 00 to 99 and using the above random numbers select the marks of a student whose number is 23 since the first random is 23. Next select a student whose number is 04 since the next random number is 04. Repeating this process we obtain a sample of marks of size 10.

By considering another set of 10 random numbers, we can construct another sample of marks of size 10 and so on.

PARAMETER AND STATISTIC

Any statistical measure relating to the population which is based on all units of the population is called **parameter**, *e.g.*, population mean (μ), population S.D. (σ), moments μ_r, μ'_r etc.

Any statistical measure relating to the sample which is based on all units of the sample is called **statistic,** *e.g.*, sample mean ($\bar{x}$), sample variance, moments m_r, m'_r etc. Hence the value of a statistic varies from sample to sample. This variation is called **'sampling fluctuation'.** The parameter has no fluctuation and it is constant. The probability distribution of a statistic is called 'sampling distribution'. The standard deviation (S.D.) in the sampling distribution is called **'standard error'** of the statistic.

Example 1. *For a population of five units, the values of a characteristic x are given below :*

8, 2, 6, 4 and 10.

Consider all possible samples of size 2 from the above population and show that the mean of the sample means is exactly equal to the population mean.

Solution. The population mean, $\mu = \frac{30}{5} = 6$

Random samples of size two (Without Replacement)

Serial no.	*Sample values*	*Sample mean*	*Serial no.*	*Sample values*	*Sample mean*
1	8, 2	5	6	2, 4	3
2	8, 6	7	7	2, 10	6
3	8, 4	6	8	6, 4	5
4	8, 10	9	9	6, 10	8
5	2, 6	4	10	4, 10	7
	Total	31		Total	29

$\therefore$ Mean of sample means $= \frac{31 + 29}{10} = \frac{60}{10} = 6$ which is equal to the population mean.

SAMPLING DISTRIBUTION OF MEAN

Case I : σ Known

Consider a population having mean μ and variance σ^2. If a random sample of size n is taken from this population then the sample mean $\bar{X}$ is a random variable whose distribution has the mean μ.

If the population is infinite, then the variance of this distribution is $\frac{\sigma^2}{n}$ and the standard error is defined as S.E. $= \frac{\sigma}{\sqrt{n}}$.

NOTES

If the population is finite of size N then the variance of this distribution is $\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}$ and the standard error is defined as

$$S.E. = \frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$

provided the sample is drawn without replacement.

The factor $\frac{N-n}{N-1}$ is called finite population correction factor.

Let us consider the standardized sample mean

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

Then we have the central limit theorem as follows :

If $\bar{X}$ is the mean of a sample of size n taken from a population whose mean is μ and variance is σ^2, then

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0, 1) \text{ as } n \to \infty.$$

If the samples come from a normal population then the sampling distribution of the mean is normal regardless of the size of the sample.

If the population is not normal then the sampling distribution of the mean is approximately normal for small size (n = 25) of the sample.

Example 2. *A random sample of size 100 is taken from an infinite population having the mean μ = 66 and the variance σ^2 = 225. What is the probability of getting an $\bar{x}$ between 64 and 68 ?*

Solution. Let $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, n = 100, μ = 66, σ = 15

Required probability = $P[64 < \bar{x} < 68]$

$= P[-1.33 < z < 1.33]$

$= 2\Phi(1.33) = 2(0.4082)$

$= 0.8164.$

Example 3. *A random sample is of size 5 is drawn without replacement from a finite population consisting of 35 units. If the population standard deviation is 2.25. What is the standard error of sample mean ?*

Solution. Here, n = 5, N = 35, σ = 2.25

$$\text{S.E. of sample mean} = \frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$

$$= \frac{2.25}{\sqrt{5}}\cdot\sqrt{\frac{30}{34}} = 0.95.$$

Case II : σ Unknown

For small sample, the assumption of normal population gives fairly the sampling distribution of $\overline{X}$. However the σ is replaced by sample standard deviation S. Then we have

NOTES

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \quad \text{where,} \quad S^2 = \frac{1}{n-1} \cdot \Sigma (x_i - \bar{x})^2$$

is a random variable having the t distribution with the degrees of freedom $v = n - 1$.

SAMPLING DISTRIBUTION OF SAMPLE VARIANCE

Like sample mean, if we calculate the sample variance for each samples drawn from a population then it shows also a random variable. We have the following result: If a random sample of size n with sample variance S^2 is taken from a normal population having the variance σ^2, then

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad \text{where, } S^2 = \frac{1}{n-1} \cdot \Sigma (x_i - \bar{x})^2$$

is a random variable having the chi-square distribution with the degrees of freedom $v = n - 1$.

(In chi-square distribution table χ^2_α represents the area under the chi-square distribution to its right is equal to α).

If S_1^2 and S_2^2 are the variances of independent random sample of size n_1 and n_2 respectively, taken from two normal populations having the same variance, then

$$F = \frac{S_1^2}{S_2^2}$$

is a random variable having the F distribution with the degrees of freedoms $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$.

Example 4. *If two independent random samples of size $n_1 = 9$ and $n_2 = 16$ are taken from the normal population, what is the probability that the variance of the first sample will be at least four times as large as that of the second sample ?*

Solution. Here $v_1 = 9 - 1 = 8, \quad v_2 = 16 - 1 = 15, \quad S^2_1 = 4S^2_2$

From F distribution table we find that

$$F_{0.01} = 4.00 \text{ for } v_1 = 8 \text{ and } v_2 = 15.$$

Thus, the desired probability is 0.01.

SAMPLING DISTRIBUTION OF SAMPLE PROPORTION

Consider a lot with proportion of defectives P. If a random sample of size n with proportion of defectives p is drawn from this population then the sampling distribution of p is approximately normal distribution with mean = P and S.D.

NOTES

= S.E. of sample proportion $= \sqrt{\frac{PQ}{n}}$ where, Q = 1 – P and the sample size n is sufficiently large. If the random sample is drawn from a finite population without replacement then we have to multiply a correction factor $\sqrt{\frac{N-n}{N-1}}$ to the S.D. formula.

If p_1 and p_2 denote the proportions from independent samples of sizes n_1 and n_2 drawn from two populations with proportions P_1 and P_2 respectively then

$$\text{S.E. of } (p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

where, $P_1 + Q_1 = 1$ and $P_2 + Q_2 = 1$.

Example 5. *It has been found that 3% of the tools produced by a certain machine are defective. What is the probability that in a shipment of 450 such tools, 2% or more will be defective ?*

Solution. Since the sample size $n = 450$ is large, the sample proportion (p) is approximately normally distributed with mean = P = 3% = 0.03.

$$\text{S.D.} = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{(0.03)\ (0.97)}{450}} = 0.008$$

$\therefore$ Required probability $= P[p > 0.02]$

$= P[z > -1.25] = 0.5 + \Phi\ (1.25)$

$= 0.5 + 0.3944 = 0.8944.$

SUMMARY

- Sampling means the selection of a part of the aggregate with a view to draw some statistical informations about the whole. This aggregate of the investigation is called population and the selected part is called sample.
- Any statistical measure relating to the population which is based on all units of the population is called parameter.
- Any statistical measure relating to the sample which is based on all units of the sample is called statistic.

PROBLEMS

1. A population consists of 5 numbers (2, 3, 6, 8, 11). Consider all possible samples of size two which can be drawn with replacement from this population. Calculate the S.E. of sample means.
2. When we sample from an infinite population, what happens to the standard error of the mean if the sample size is (*a*) increased from 30 to 270, (*b*) decreased from 256 to 16?

NOTES

3. A random sample of size 400 is taken from an infinite population having the mean $\mu = 86$ and the variance of $\sigma^2 = 625$. What is the probability that $\overline{X}$ will be greater than 90?
4. The number of letters that a department receives each day can be modeled by a distribution having mean 25 and standard deviation 4. For a random sample of 30 days, what will be the probability that the sample mean will be less than 26 ?
5. A random sample of 400 mangoes was taken from a large consignment and 30 were found to be bad. Find the S.E. of the population of bad ones in a sample of this size.
6. From a population of large number of men with a S.D. 5, a sample is drawn and the standard error is found to be 0.5, what is the sample size ?
7. A population consists of 20 elements, has mean 9 and S.D. 3 and a sample of 5 elements is taken without replacement. Find the mean and S.D. of the sampling distribution of the mean. What will be the S.D. for samples of size 10 ?
8. A machine produces a component for a transistor set of the total produce, 6 percent are defective. A random sample of 5 components is taken for examination from (*i*) a very large lot of produce, (*ii*) a box of 10 components. Find the mean and S.D. of the average number of defectives found among the 5 components taken for examination.
9. A population consists of five numbers 2, 3, 6, 8, 11. Consider all possible samples of size two which can be drawn without replacement from the population. Find
   (*a*) The mean of the population
   (*b*) Standard deviation of the population
   (*c*) The mean of the sampling distribution of means
   (*d*) The standard deviation of the sampling distribution of means.

ANSWERS

1. 2.32 **2.** (*a*) It is divided by 3 (*b*) It is multiplied by 4

3. 0.0007 **4.** 0.9147 **5.** 0.013

6. 100

7. For sample of 5 elements, sampling mean = 8, S.D. = $\sqrt{\frac{27}{19}}$

For sample of 10 elements, sampling mean = 8, S.D. = $\frac{3}{\sqrt{19}}$

8. Mean = 0.06, S.D. = 0.106

9. (*a*) 6, (*b*) 3.29, (*c*) 6, (*d*) 2.12.

CHAPTER 10 ESTIMATION OF PARAMETERS

NOTES

★ STRUCTURE ★

- Estimation
- Point Estimation
- Interval Estimation
- Bayesian Estimation
- Summary
- Problems

ESTIMATION

When we deal with a population, most of the time the parameters are unknown. So we cannot draw any conclusion about the population. To know the unknown parameters the technique is to draw a sample from the population and try to gather information about the parameter through a function which is reasonably close. Thus the obtained value is called an estimated value of the parameter, the process is called estimation and the estimating function is called estimator.

A good estimator should satisfy the four properties which we briefly explain below :

(*a*) Unbiasedness. A statistic t is said to be an unbiased estimator of a parameter θ if, $E[t] = \theta$.

Otherwise it is said to be 'biased'.

Theorem 1. Prove that the sample mean $\bar{x}$ is an unbiased estimator of the population mean μ.

Proof. Let $x_1, x_2, ..., x_n$ be a simple random sample with replacement from a finite population of size N, say, $X_1, X_2, ..., X_N$

Here, $\bar{x} = (x_1 + x_2 + ...+ x_n)/n$

$\mu = (X_1 + X_2 + ...+ X_N)/N$

To prove that $E(\bar{x}) = \mu$

While drawing x_i, it can be one of the population members *i.e.*, the probability distribution of x_i can be taken as follows:

x_i	X_1	X_2	...X_N
Probability	1/N	1/N	1/N

for $i = 1, 2, ..., n$

NOTES

Therefore,

$$E(x_i) = X_1 \cdot \frac{1}{N} + X_2 \cdot \frac{1}{N} + ... + X_N \cdot \frac{1}{N}$$
$$= (X_1 + X_2 + ... + X_N)/N$$
$$= \mu, \quad i = 1, 2, ..., n.$$

and
$$E(\bar{x}) = E[(x_1 + x_2 + ... + x_n)/n]$$
$$= [E(x_1) + E(x_2) + ... + E(x_n)]/n$$
$$= [\mu + \mu + ... + \mu]/n$$
$$= n\mu/n = \mu.$$

The same result is also true for infinite population and the sampling without replacement.

Theorem 2. The sample variance

$$S^2 = \frac{1}{n}.\Sigma (x_i - \bar{x})^2$$

is a biased estimator of the population variance σ^2.

Proof. Let $x_1, x_2, ..., x_n$ be a random sample from an infinite population with mean μ and variance σ^2.

Then
$$E(x_i) = \mu, \text{ Var } (x_i) = E(x_i - \mu)^2 = \sigma^2,$$
for $i = 1, 2, ..., n$.

$$s^2 = \frac{1}{n}.\Sigma (x_i - \bar{x})^2$$

$$= \frac{1}{n}.\Sigma x_i^2 - (\bar{x})^2$$

$$= \frac{1}{n}.\Sigma y_i^2 - (\bar{y})^2, \text{ where, } y_i = x_i - \mu \text{ and S.D is unaffected by change of origin.}$$

$$= \frac{1}{n}.\Sigma (x_i - \mu)^2 - (\bar{x} - \mu)^2$$

$\therefore$
$$E(s^2) = \frac{1}{n}.\Sigma E(x_i - \mu)^2 - E(\bar{x} - \mu)^2$$

$$= \frac{1}{n}.\Sigma \sigma^2 - \text{Var } (\bar{x}) = \sigma^2 - \frac{\sigma^2}{n}$$

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

$\Rightarrow$ s^2 is a biased estimator of σ^2

Note. Let $S^2 = \frac{1}{(n-1)}.\Sigma (x_i - \bar{x})^2$, then

NOTES

$$E(S^2) = \frac{n}{n-1}.E(s^2)$$

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

$$= \sigma^2$$

Thus S^2 is an unbiased estimator of σ^2.

Example 1. *A population consists of 4 values 3, 7, 11, 15. Draw all possible sample of size two with replacement. Verify that the sample mean is an unbiased estimator of the population mean.*

Solution. No. of samples = $4^2 = 16$, which are listed below :

(3, 3),	(7, 3),	(11, 3),	(15, 3)
(3, 7),	(7, 7),	(11, 7),	(15, 7)
(3, 11),	(7, 11),	(11, 11),	(15, 11)
(3, 15),	(7, 15),	(11, 15),	(15, 15)

Population mean, $\mu = \frac{3 + 7 + 11 + 15}{4} = \frac{36}{4} = 9$

Sampling distribution of sample mean

Sample mean ($\bar{x}$)	*Frequency* $f(\bar{x})$	$\bar{x} \cdot f(\bar{x})$
3	1	3
5	2	10
7	3	21
9	4	36
11	3	33
13	2	26
15	1	15
Total	16	144

Mean of sample mean $= \frac{144}{16} = 9$

Since, $E(\bar{x}) = \mu$,

$\Rightarrow$ Sample mean is an unbiased estimator of the population mean.

(*b*) Consistency. A statistic t_n obtained from a random sample of size n is said to be a consistent estimator of a parameter if it converges in probability to θ as n tends to infinity.

Alt, If $E[T_n] \to \theta$ and $Var[T_n] \to 0$ as $n \to \infty$, then the statistic t_n is said to be consistent estimator of θ.

For example, in sampling from a Normal Population $N(\mu, \sigma^2)$,

$$E[\bar{x}] = \mu \text{ and } Var[\bar{x}] = \frac{\sigma^2}{n} \to 0 \quad \text{as } n \to \infty.$$

Hence the sample mean is a consistent estimator of population mean.

(c) Efficiency. There may exist more than one consistent estimator of a parameter. Let T_1 and T_2 be two consistent estimators of a parameter θ. If $\text{Var}(T_1) < \text{Var}(T_2)$ for all n

then T_1 is said to be more efficient than T_2 for all sample size.

If a consistent estimator has least variance than any other consistent estimators of a parameter, then it is called the most efficient estimator.

Let T be the most efficient estimator and T_1 *be any other consistent estimator of a parameter. Then, we define*

$$\text{Efficiency} = \frac{\text{Var}(T)}{\text{Var}(T_1)}$$

which is less than equal to one.

A statistic which is unbiased and also the most efficient, is said to be the Minimum Variance Unbiased Estimator (MVUE).

Note. If T_1 and T_2 are two MVU Estimators of a parameter then $T_1 = T_2$

For example, the sample mean $\bar{x}$ obtained from a normal population is the MVUE for the parameter μ.

Let $x_1, x_2, \ldots, x_n$ be a random sample and

$$T = a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

where $a_1, a_2, \ldots, a_n$ are constants. If T is an MVUE, then T is also called Best Linear Unbiased Estimator (BLUE).

Example 2. *A random sample* $(X_1, X_2, X_3, X_4, X_5, X_6)$ *of size 6 is drawn from a normal population with unknown mean* μ. *Consider the following estimators to estimate* μ.

(i) $$T_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6}$$

(ii) $$T_2 = \frac{X_1 + X_2 + X_3}{2} + \frac{X_4 + X_5 + X_6}{3}$$

(iii) $$T_3 = \frac{1}{2}(X_1 + X_2) + X_3 + X_4 + \frac{1}{3}(X_5 + X_6)$$

Are these estimators unbiased ? Find the estimator which is best among T_1, T_2 *and* T_3.

Solution. Here $E(X_i) = \mu$, $\text{Var}(X_i = \sigma^2$ (say), $\text{Cov}(X_i, X_j) = 0$, $i \neq j$

$$E(T_1) = \frac{1}{6}[E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) + E(X_6)]$$

$$= \frac{1}{6}[\mu + \mu + \mu + \mu + \mu + \mu) = \frac{1}{6}.6\,\mu = \mu.$$

$$E(T_2) = \frac{1}{2}[E(X_1) + E(X_2) + E(X_3)] + \frac{1}{3}[E(X_4) + E(X_5) + E(X_6)]$$

$$= \frac{1}{2}[\mu + \mu + \mu] + \frac{1}{3}[\mu + \mu + \mu] = \frac{3\mu}{2} + \mu = \frac{5\mu}{2}.$$

$$E(T_3) = \frac{1}{2}[E(X_1) + E(X_2)] + E(X_3) + E(X_4) + \frac{1}{3}[E(X_5) + E(X_6)]$$

NOTES

$$= \frac{1}{2}[\mu + \mu] + \mu + \mu + \frac{1}{3}[\mu + \mu]$$

$$= \mu + 2\mu + \frac{2\mu}{3} = \frac{11\mu}{3}.$$

Since $E(T_1) = \mu \Rightarrow T_1$ is unbiased. T_2 and T_3 are biased estimators.

$$\text{Var}(T_1) = \frac{1}{36}[\text{Var}(X_1) + \text{Var}(X_2) + ... + \text{Var}(X_6)]$$

$$= \frac{1}{36}[\sigma^2 + \sigma^2 + ... + \sigma^2] = \frac{1}{36}(6\sigma^2) = \frac{\sigma^2}{6}.$$

$$\text{Var}(T_2) = \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)]$$

$$+ \frac{1}{9}[\text{Var}(X_4) + \text{Var}(X_5) + \frac{1}{9}\text{Var}(X_6)]$$

$$= \frac{1}{4}[\sigma^2 + \sigma^2 + \sigma^2] + [\sigma^2 + \sigma^2 + \sigma^2]$$

$$= \frac{3}{4}\sigma^2 + \frac{3\sigma^2}{9} = \frac{13}{12}\sigma^2.$$

$$\text{Var}(T_3) = \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2)] + \text{Var}(X_3) + \text{Var}(X_4)$$

$$+ \frac{1}{9}[\text{Var}(X_5) + \text{Var}(X_6)]$$

$$= \frac{1}{4}[\sigma^2 + \sigma^2] + \sigma^2 + \sigma^2 + \frac{1}{9}[\sigma^2 + \sigma^2]$$

$$= \frac{\sigma^2}{2} + 2\sigma^2 + \frac{2\sigma^2}{9} = \frac{49}{18}\sigma^2.$$

Since $\text{Var}(T_1)$ is smallest $\Rightarrow T_1$ is best estimator.

$$\text{Efficiency of } T_1 \text{ over } T_2 = \frac{\sigma^2/6}{13\ \sigma^2/12} = \frac{2}{13} = 0.15$$

$$\text{Efficiency of } T_1 \text{ over } T_3 = \frac{\sigma^2/6}{49\ \sigma^2/18} = \frac{3}{49} = 0.06.$$

Example 3. *A random sample (X_1, X_2, X_3, X_4) of size 4 is drawn from a normal population with unknown mean. If*

$$T = 2X_1 + \frac{\lambda}{2}X_2 + 3X_3 - 4X_4$$

be an unbiased estimator of μ, find λ.

Solution. Let $E(X_i) = \mu,\ i = 1, 2, 3, 4.$

For unbiasedness, $E(T) = \mu$

$$\Rightarrow \quad 2E(X_1) + \frac{\lambda}{2} E(X_2) + 3\,E(X_3) - 4\,E(X_4) = \mu$$

$$\Rightarrow \quad 2\mu + \frac{\lambda}{2}\mu + 3\mu - 4\mu = \mu$$

$$\Rightarrow \quad \mu + \frac{\lambda}{2}\mu = \mu$$

$$\Rightarrow \quad \frac{\lambda}{2} = 0$$

$$\Rightarrow \quad \lambda = 0.$$

NOTES

(*d*) **Sufficiency.** Let $x_1, x_2, ..., x_n$ be a random sample from a population whose *p.m.f.* or *pdf* is $f(x, \theta)$. Then T is said to be a sufficient estimator of θ if we can express the following :

$$f(x_1, \theta) \cdot f(x_2, \theta) \ldots f(x_n, \theta) = g_1(T, \theta) \cdot g_2(x_1, x_2, ..., x_n)$$

where $g_1(T, \theta)$ is the sampling distribution of T and contains θ and $g_2(x_1, x_2, ..., x_n)$ is independent of θ.

Sufficient estimators exist only in few cases. However in random sampling from a normal population, the sampling mean $\bar{x}$ is a sufficient estimator of μ.

POINT ESTIMATION

Using sampling if a single value is estimated for the unknown parameter of the population, then this process of estimation is called point estimation. We shall discuss two methods of point estimation below:

I. Method of Maximum Likelihood

Let $x_1, x_2, ..., x_n$ be a random sample from a population whose *p.m.f.* (discrete case) or *p.d.f.* (continuous case) is $f(x, \theta)$ where θ is the parameter. Then construct the likelihood function as follows :

$$L = f(x_1, \theta).\ f(x_2, \theta) \ldots f(x_n, \theta).$$

Since log L is maximum when L is maximum. Therefore to obtain the estimate of θ, we maximize L as follows,

$$\frac{\partial}{\partial \theta}(\log L) = 0 \Rightarrow \theta = \hat{\theta}$$

and

$$\frac{\partial^2}{\partial \theta^2}(\log L) < 0 \text{ at } \theta = \hat{\theta}$$

Here $\hat{\theta}$ is called Maximum Likelihood Estimator (MLE).

Properties of MLE

(*i*) MLE is not necessarily unbiased.

(*ii*) MLE is consistent, most efficient and also sufficient, provided a sufficient estimator exists.

(*iii*) MLE tends to be distributed normally for large samples.

(*iv*) If $g(\theta)$ is a function of θ and $\hat{\theta}$ is an MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

NOTES

Example 4. *A discrete random variable X can take up all non-negative integers and*

$$P(X = r) = p(1-p)^r \ (r = 0, 1, 2, \ldots)$$

where, p $(0 < p < 1)$ is the parameter of the distribution. Find the MLE of p for a sample of size n : $x_1, x_2, \ldots, x_n$ from the population of X.

Solution. Consider the following likelihood function :

$$\begin{aligned} L &= P(X = x_1) \cdot P(X = x_2) \ldots P(X = x_n) \\ &= p(1-p)^{x_1} \cdot p(1-p)^{x_2} \ldots p(1-p)^{x_n} \\ &= p^n (1-p)^{x_1 + x_2 + \ldots + x_n} \\ &= p^n (1-p)^{\Sigma x_i} \end{aligned}$$

Taking log on both sides we obtain

$$\ln L = n \ln p + (\Sigma x_i) \ln (1-p)$$

Now $$\frac{d \ln L}{dp} = 0$$

$$\Rightarrow \quad \frac{n}{p} - \frac{\Sigma x_i}{1-p} = 0$$

$$\Rightarrow \quad \frac{n}{p} = \frac{\Sigma x_i}{1-p}$$

$$\Rightarrow \quad \frac{1-p}{p} = \frac{\Sigma x_i}{n}$$

$$\Rightarrow \quad \frac{1}{p} - 1 = \bar{x}$$

$$\Rightarrow \quad \hat{p} = \frac{1}{1+\bar{x}}$$

Also,

$$\begin{aligned} \frac{d^2 \ln L}{dp^2} &= -\frac{n}{p^2} - \frac{\Sigma x_i}{(1-p)^2} \\ &= -n\left(\frac{1}{p^2} + \frac{\bar{x}}{(1-p)^2}\right) \\ &= -n\left((1+\bar{x})^2 + \frac{\bar{x}(1+\bar{x})^2}{(\bar{x})^2}\right) \end{aligned}$$

at $$\hat{p} = \frac{1}{1+\bar{x}}$$

$$= -n(1+\bar{x})^2 \left(1 + \frac{1}{x}\right) < 0$$

Hence the MLE of p is $\frac{1}{1+\bar{x}}$.

Example 5. *A random variable X has a distribution with density function :*

$$f(x) = \lambda x^{\lambda - 1} \ (0 < x < 1)$$

where λ is the parameter. Find the MLE of λ for a sample of size $n : x_1, x_2, \ldots, x_n$ from the population of X.

Solution. Consider the following likelihood function :

NOTES

$$L = f(x_1).\ f(x_2) \ \ldots \ f(x_n)$$

$$= \lambda x_1^{\lambda - 1} . \lambda x_2^{\lambda - 1} \ldots \lambda x_n^{\lambda - 1}$$

$$= \lambda^n (x_1 . x_2 \ldots x_n)^{\lambda - 1}$$

Taking log on both sides we obtain

$$\ln L = n \ln \lambda + (\lambda - 1) \ln (x_1 . x_2 \ldots . x_n)$$

Now, $$\frac{d \ln L}{d\lambda} = 0 \quad \Rightarrow \quad \frac{n}{\lambda} + \ln (x_1 x_2 \ldots x_n) = 0$$

$$\Rightarrow \quad \frac{n}{\lambda} = - \ln (x_1 x_2 \ldots x_n)$$

$$\Rightarrow \quad \hat{\lambda} = \frac{- n}{\ln (x_1 x_2 \ldots x_n)}$$

Also, $$\frac{d^2 \ln L}{d\lambda^2} = - \frac{n}{\lambda^2} < 0$$

Hence the MLE of λ is $\dfrac{- n}{\ln (x_1 x_2 \ldots x_n)}$

Example 6. *X tossed a biased coin 40 times and got head 15 times, while Y tossed it 50 times and got head 30 times. Find the MLE of the probability of getting head when the coin is tossed.*

Solution. Let P be the unknown probability of getting a head.

Using binomial distribution,

Probability of getting 15 heads in 40 tosses $= \binom{40}{15} P^{15} (1 - P)^{25}$

Probability of getting 30 heads in 50 tosses $= \binom{50}{30} P^{30} (1 - P)^{20}$

The likelihood function is taken by multiplying these probabilities.

$$L = \binom{40}{15} \cdot \binom{50}{30} P^{45} (1 - P)^{45}$$

$$\therefore \quad \log L = \log \left[\binom{40}{15} \cdot \binom{50}{30} \right] + 45 \log P + 45 \log (1 - P)$$

NOTES

Hence $\frac{\partial \log L}{\partial P} = 0 \Rightarrow \frac{45}{P} - \frac{45}{1-P} = 0 \Rightarrow P = 1/2$, which is the MLE.

II. Method of Moments

In this method, the first few moments of the population is equated with the corresponding moments of the sample.

Then $\mu'_r = m'_r$

where $\mu'_r = E(x^r)$ and $m'_r = \Sigma x_i^r / n$

The solution for the parameters gives the estimates. But this method is applicable only when the population moments exist.

Example 7. *Estimate the parameter p of the binomial distribution by the method of moments (when n is known).*

Solution. Here, $\mu'_1 = E(x) = np$ and $m'_1 = \bar{x}$

Taking $\mu'_1 = m'_1$, we have

$$np = \bar{x}$$

$$\Rightarrow \quad p = \frac{\bar{x}}{n}$$

which is the estimated value.

INTERVAL ESTIMATION

In interval estimation we find an interval which is expected to include the unknown parameter with a specified probability, *i.e.*,

$$P(t_1 \leq \theta \leq t_2) = k$$

where, $[t_1, t_2]$ is called confidence interval (C.I.),

t_1, t_2 are called confidence limits,

k is called confidence coefficient of the interval.

(a) C.I. for mean with known S.D. Let us consider a random sample of size n from a Normal Population $N(\mu, \sigma^2)$ **in which σ^2 is known.** To find C.I. for mean μ.

We know that $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ follows standard normal distribution and 95% of the area under the standard normal curve lies between $z = 1.96$ and $z = -1.96$, Then,

$$P[-1.96 \leq z \leq 1.96] = 0.95$$

$$\Rightarrow \quad P\left[-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right] = 0.95$$

i.e., in 95% cases we have

$$-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96$$

$$\Rightarrow \qquad \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

The interval $\left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right]$ is known as 95% confidence interval for μ.

NOTES

Similarly, $\left[\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}}\right]$ is known as 99% C.I. for μ,

$\left[\bar{x} - 3. \frac{\sigma}{\sqrt{n}}, \bar{x} + \frac{3\sigma}{\sqrt{n}}\right]$ is known as 99.73% C.I. for μ.

(b) C.I. for mean with unknown S.D. σ.

In this case, the sampling from a normal population N (μ, σ^2), the statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}, \quad \text{where } s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

follows t distribution with $(n - 1)$ degree of freedom.

Then for 95% confidence interval for mean μ we have

$$-t_{0.025} \leq \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \leq t_{0.025}$$

$$\Rightarrow \qquad \bar{x} - t_{0.025} \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{0.025} \frac{s}{\sqrt{n-1}}$$

Thus $\left[\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n-1}}, \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n-1}}\right]$ *is called 95% C.I. for* μ.

Similarly, $\left[\bar{x} - t_{0.005} \cdot \frac{s}{\sqrt{n-1}}, \bar{x} + t_{0.005} \cdot \frac{s}{\sqrt{n-1}}\right]$ is called 99% C.I. for μ.

(c) C.I. for variance σ^2 with known mean. We know that $(x_i - \mu)^2/\sigma^2$ follows chi-square distribution with n degrees of freedom.

For probability 95% we have

$$\chi^2_{0.975} \leq \Sigma (x_i - \mu)^2 / \sigma^2 \leq \chi^2_{0.025}$$

$$\Rightarrow \qquad \Sigma (x_i - \mu)^2 / \chi^2_{0.025} \leq \sigma^2 \leq \Sigma (x_i - \mu)^2 / \chi^2_{0.975}$$

which is 95% confidence interval for σ^2.

Similarly,

$$\Sigma (x_i - \mu)^2 / \chi^2_{0.005} \leq \sigma^2 \leq \Sigma (x_i - \mu)^2 / \chi^2_{0.995}$$

is the 99% confidence interval for σ.

NOTES

(d) C.I. for variance σ^2 with unknown mean. In this case $ns^2 / \sigma^2 = \Sigma (x_i - \bar{x})^2 / \sigma^2$ follows chi-square distribution with $(n - 1)$ degrees of freedom.

For probability 95% we have

$$\chi^2_{0.975} \leq ns^2 / \sigma^2 \leq \chi^2_{0.025}$$

$$\Rightarrow \quad ns^2/\chi^2_{0.025} \leq \sigma^2 \leq ns^2/\chi^2_{0.975}$$

which is 95% C.I. for σ^2.

Similarly, $ns^2 / \chi^2_{0.005} \leq \sigma^2 \leq ns^2 / \chi^2_{0.995}$ is the 99% C.I. for σ^2.

Some of the Confidence Limits are given below :

(with Normal Population N (μ, σ^2))

Difference of Means ($\mu_1 - \mu_2$) : (S.Ds known).

95% Confidence limits = $(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

99% Confidence limits = $(\bar{x}_1 - \bar{x}_2) \pm 2.58 \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

Difference of Means ($\mu_1 - \mu_2$) : (Common S.D. unknown)

95% Confidence limits = $(\bar{x}_1 - \bar{x}_2) \pm t_{0.025} \cdot s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

99% Confidence limits = $(\bar{x}_1 - \bar{x}_2) \pm t_{0.005} \cdot s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

For Proportion P :

95% Confidence limits = $p \pm 1.96$ (S.E. of p)

99% Confidence limits = $p \pm 2.58$ (S.E. of p)

where, S.E. of $p = \sqrt{\frac{PQ}{n}} \approx \sqrt{\frac{pq}{n}}$

For Difference of Proportions $P_1 - P_2$:

95% Confidence limits = $(p_1 - p_2) \pm 1.96$ [S.E. of $(p_1 - p_2)$]

99% Confidence limits = $(p_1 - p_2) \pm 2.58$ [S.E. of $(p_1 - p_2)$]

where S.E of $(p_1 - p_2) = \sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}} \approx \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$.

Example 8. *A random sample of size 10 was drawn from a normal population with an unknown mean and a variance of 35.4 (cm)². If the observations are (in cms) : 55, 75, 71, 66, 73, 77, 63, 67, 60 and 76, obtain 99% confidence interval for the population mean.*

Solution. Given $n = 10$, $\Sigma x_i = 683$, Then $\bar{x} = \frac{\Sigma x}{n} = 68.3$

Since the population S.D. σ is known, then 99% C.I. for μ is given by

$$\left[\bar{x} - 2.58\,\frac{\sigma}{\sqrt{n}},\ \bar{x} + 2.58\,\frac{\sigma}{\sqrt{n}}\right]$$

i.e.,
$$\left[68.3 - \frac{2.58\,.\,\sqrt{35.4}}{\sqrt{10}},\ 68.3 + \frac{2.58\,.\,\sqrt{35.4}}{\sqrt{10}}\right]$$

i.e., [63.45, 73.15].

NOTES

Example 9. *A random sample of size 10 was drawn from a normal population which are given by 48, 56, 50, 55, 49, 45, 55, 54, 47, 43. Find 95% confidence interval for mean μ of the population.*

Solution. From the given data, $\Sigma x_i = 502$, so $\bar{x} = 50.2$, $n = 10$

Let $d = x - 50$, then the samples are changed to

$-2, 6, 0, 5, -1, -5, 5, 4, -3, -7.$

$\Sigma d = 2, \Sigma d^2 = 190$

$$\therefore \quad s^2 = \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2 = \frac{190}{10} - \left(\frac{2}{10}\right)^2 = 18.96$$

$$s = 4.35$$

Since, the population S.D. σ is unknown, the 95% C.I. for mean μ is

$$\left[\bar{x} - 2.262\,.\frac{s}{\sqrt{n}},\ \bar{x} + 2.262\,.\frac{s}{\sqrt{n}}\right]$$

i.e.,
$$\left[50.2 - (2.262)\,\frac{(4.35)}{\sqrt{10}},\ 50.2 + (2.262)\,\frac{(4.35)}{\sqrt{10}}\right]$$

i.e., [47.09, 53.31].

Example 10. *The standard deviation of a random sample of size 15 drawn from a normal population is 3.2. Calculate the 95% confidence interval for the standard deviation (σ) in the population.*

Solution. Here $n = 15$, sample s.d. $(s) = 3.2$

95% Confidence interval for σ^2 is

$$\frac{n s^2}{\chi^2_{0.025}} \le \sigma^2 \le \frac{n s^2}{\chi^2_{0.975}}$$

From chi-square table with 14 degrees of freedom,

$$\chi^2_{0.025} = 26.12, \quad \chi^2_{0.975} = 5.63$$

Therefore the C.I. is

$$\frac{15.(3.2)^2}{26.12} \le \sigma^2 \le \frac{15.(3.2)^2}{5.63}$$

i.e., $5.88 \le \sigma^2 \le 27.28$

i.e., $2.42 \le \sigma \le 5.22.$

NOTES

Example 11. *A sample of 500 springs produced in a factory is taken from a large consignment and 65 are found to be defective. Estimate the assign limits in which the percentage of defectives lies.*

Solution. There are 65 defective springs in a sample of size $n = 500$.

$\therefore$ The sample proportion of defective is

$$p = \frac{65}{500} = 0.13$$

The limits to the percentage of defectives refer to the C.I., which can be taken as

$$[p - 3 \text{ (S.E. of } p), \; p + 3 \text{ (S.E. of } p)]$$

Here $$\text{S.E. of } p = \sqrt{\frac{PQ}{n}}$$

$$\approx \sqrt{\frac{pq}{n}} = \sqrt{\frac{65}{500}\left(1 - \frac{65}{500}\right)\frac{1}{500}} = 0.02$$

Thus the limits are $[0.13 - 3\,(0.02), 0.13 + 3\,(0.02)]$

i.e., $[0.07, 0.19]$.

BAYESIAN ESTIMATION

Bayesian estimation uses subjective judgement in an engineering design.

For discrete case, let the parameter θ takes the values θ_i, $i = 1, 2, ..., n$ with the probabilities $p_i = P[\theta = \theta_i]$. Let θ_0 be the observed outcome of the experiment. Then by Bayes' theorem we obtain,

$$P[\theta = \theta_i|\theta_0] = \frac{P[\theta_0|\theta = \theta_i] \cdot P[\theta = \theta_i]}{\sum_{j=1}^{n} P[\theta_0|\theta = \theta_j] \cdot P[\theta = \theta_j]} \quad i = 1, 2, ..., n$$

Then the expected value of θ is called Bayesian estimator of the parameter, *i.e.*,

$$\hat{\theta} = E[\theta = \theta_i|\theta_0]$$

$$= \sum_{i=1}^{n} \theta_i \cdot P[\theta = \theta_i|\theta_0]$$

Using this we can calculate

$$P[X \le a] = \sum_{i=1}^{n} P[X \le a|\theta = \theta_i] \cdot P[\theta = \theta_i|\theta_0]$$

For continuous case, let θ be a random variable of the parameter of the distribution given by the density function $f'(\theta)$. Then

$$P[\theta_i < \theta < \theta_i + \Delta\theta] = f'(\theta_i) \cdot \Delta\theta, \; i = 1, 2, ..., n$$

If θ_0 is an observed experimental outcome, then

$$f''(\theta_i)\,\Delta\theta = \frac{P\left[\theta_0|\theta_i\right] . f'(\theta_i)\Delta\theta}{\sum_{j=1}^{n} P\left[\theta_0|\theta_j\right] f'(\theta_j)\Delta\theta}, \quad i = 1, 2, \ldots, n$$

NOTES

In the limit we obtain, $f''(\theta) = \dfrac{P\left[\theta_0|\theta\right] f'(\theta)}{\int_{-\infty}^{\infty} P\left[\theta_0|\theta\right] f'(\theta)\, d\theta}$

Then the Bayesian estimator is

$$\hat{\theta} = E\left[\theta|\theta_0\right] = \int_{-\infty}^{\infty} \theta f''(\theta) d\theta$$

Using this we can calculate

$$P\left[X \leq a\right] = \int_{-\infty}^{\infty} P\left[X \leq a|\theta\right] f''(\theta) d\theta.$$

SUMMARY

- When we deal with a population, most of the time the parameters are unknown. So we cannot draw any conclusion about the population. To know the unknown parameters the technique is to draw a sample from the population and try to gather information about the parameter through a function which is reasonably close. Thus the obtained value is called an estimated value of the parameter, the process is called estimation and the estimating function is called estimator.
- If a consistent estimator has least variance than any other consistent estimators of a parameter, then it is called the most efficient estimator.
- Using sampling if a single value is estimated for the unknown parameter of the population, then this process of estimation is called point estimation.

PROBLEMS

1. A random variable X has a distribution with density function :

$$f(x) = (\alpha + 1)\, x^{\alpha}, \qquad 0 \leq x \leq 1,\ \alpha > -1$$
$$= 0, \qquad \text{otherwise}$$

and a random sample of size 8 produces the data : 0.2, 0.4, 0.8, 0.5, 0.7, 0.9, 0.8 and 0.9. Find the MLE of the unknown parameter α.

2. A random variable X has a distribution with density function :

$$f(x) = \frac{(a + 1)\, x^{a}}{2^{a+1}}, \qquad 0 \leq x \leq 2$$
$$= 0, \qquad \text{otherwise}$$

Find the MLE of the parameter a (> 0).

NOTES

3. Consider a random sample of size n from a population following Poisson distribution. Obtain the MLE of the parameter of this distribution.
4. Consider a random sample $x_1, x_2, ..., x_n$ from a normal population having mean zero. Obtain the MLE of the variance and show that it is unbiased.
5. Consider a random sample $x_1, x_2, ..., x_n$ from a population following binomial distribution having parameters n and p. Find the MLE of p and show that it is unbiased.
6. Find the estimates of μ and σ in the normal populations N (μ, σ^2) by the method of moments.
7. Show that the estimates of the parameter of the Poisson distribution obtained by the method of maximum likelihood and the method of moments are identical.
8. Find a 95% C.I. for the mean of a normal population with $\sigma = 3$, given the sample 2.3, – 0.2, 0.4 and – 0.9.
9. In a sample of size 10, the sample mean is 3.22 and the sample variance 1.21. Find the 95% C.I. for the population mean.
10. A sample of size 10 from a normal population produces the data 2.03, 2.02, 2.01, 2.00, 1.99, 1.98, 1.97, 1.99, 1.96 and 1.95. From the sample find the 95% C.I. for the population mean.
11. A random sample of size 10 from a N (μ, σ^2) yields sample mean 4.8 and sample variance 8.64. Find 95% and 99% confidence intervals for the population mean.
12. The following random sample was obtained from a normal population : 12, 9, 10, 14, 11, 8. Find the 95% C.I. for the population S.D. when the population mean is (*i*) known to be 13, (*ii*) unknown.
13. The marks obtained by 15 students in an examination have a mean 60 and variance 30. Find 99% confidence interval for the mean of the population of marks, assuming it to be normal.
14. 228 out of 400 voters picked at random from a large electorate said that they were going to vote for a particular candidate. Find 95% C.I. for the proportion of voters of the electorate who would in favour of the candidate.
15. In a random sample of 300 road accidents, it was found that 114 were due to bad weather. Construct a 99% confidence interval for the corresponding true proportions.
16. A study shows that 102 of 190 persons who saw an advertisement on a product on T.V. during a sports program and 75 of 190 other persons who saw it advertised on a variety show purchased the product. Construct a 99% confidence interval for the difference of sample proportions.

ANSWERS

1. $\hat{\alpha} = 0.890091$
2. $\hat{a} = \dfrac{n}{\ln\left(2^n / \prod_{i=1}^{n} x_i\right)} - 1$
3. $\hat{\lambda} = \bar{x}$
4. $\sigma^2 = \Sigma x_i^2 / n$
5. $\hat{p} = \bar{x} / n$.
6. $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = s^2$
8. [–2.54, 3.34]
9. [2.39, 4.05]
10. [1.972, 2.008]
11. 95% C.I. [2.233, 7.367], 99% C.I. [1.616, 7.984]
12. (*i*) [1.97, 6.72], (*ii*) [1.35, 5.30]
13. [55.64, 64.36]
14. [0.52, 0.62]
15. [0.31, 0.45]
16. [0.02, 0.28]

SECTION-D

CHAPTER 11 TESTING OF HYPOTHESIS

NOTES

★ STRUCTURE ★

- Statistical Hypothesis and Related terms
- Tests for Large Samples ($n > 3$)
- Tests of Small Samples
- Testing of Proportion
- Testing of a Single Variance
- Testing of Equality of two Variances
- Goodness of Fit Test
- Chi-square Test of Independence
- A 2 × 2 Table (Simplified Form)
- Yate's Correction
- Summary
- Problems

STATISTICAL HYPOTHESIS AND RELATED TERMS

There are many problems in which we have to make decisions about a statistical population on the basis of sample observations. To reach a decision, we make an assumption or statement about the population which is known as a *statistical hypothesis.*

For example, (*i*) the average marks of students in a university is 77% (*ii*) the average lifetime of a certain tires is at least 25,000 miles, (*iii*) difference of resistance between two types of electric wires is 0.025 ohm, etc.

The hypothesis which we are going to test for possible rejection under the assumption is called *'Null Hypothesis'* which is usually denoted by H_0. For example,

$$H_0 : \mu = \mu_0, \quad \text{or,} \quad H_0 : \mu_1 - \mu_2 = K, \; H_0 : \sigma^2 = \sigma_0 \text{ etc.}$$

Any hypothesis which is taken as complementary to the null hypothesis is called an *'Alternative Hypothesis'* and is usually denoted by H_1. For example,

$$H_1 : \mu > \mu_0, \quad \text{or,} \quad H_1 : \mu < \mu_0 \quad \text{or} \quad H_1 : \mu \neq \mu_0 \text{ etc.}$$

NOTES

The next step is to set up a statistic. While accepting or rejecting a hypothesis we commit two types of errors :

Type I Error : Reject H_0 when it is true.

Type II Error : Accept H_0 when it is false.

If we consider

$$P [\text{type I error}] = \alpha,$$

$$P [\text{type II error}] = \beta$$

then α and β are referred to as producer's risk and consumer's risk respectively.

Critical region. A region corresponding to a statistic which amounts to rejection of H_0 is termed as critical region or region of rejection.

Level of significance (α). This is a probability that a random value of the statistic belongs to the critical region. In other words, it is the size of the critical region. Usually, the level of significance is taken as 5 % or 1%. So α = P [Type I error].

Critical value. The value which separates the critical region and the acceptance region is called critical value which is set by seeing the alternative hypothesis.

Type of tests. It is determined based on the alternative hypothesis. For example,

If $H_1 : \mu > \mu_0$, then it is called *right tailed test.*

If $H_1 : \mu < \mu_0$, then it is called left tailed test.

If $H_1 : \mu \neq \mu_0$, then it is called both tailed test (/two tailed test).

Simple and composite hypothesis. If all the parameters are completely specified, the hypothesis is called simple, *e.g.*, $\mu = \mu_0$, $\sigma = \sigma_0$. Otherwise it is called composite hypothesis. *e.g.*, $\mu \leq \mu_0$, $\sigma > \sigma_0$ etc.

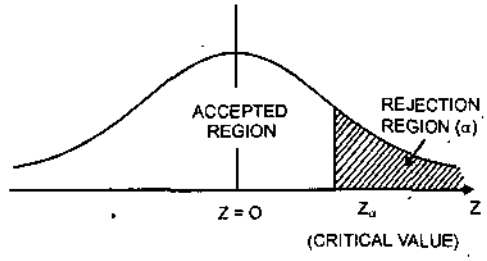


Fig. 11.1 Right tailed test ($\mu > \mu_0$).

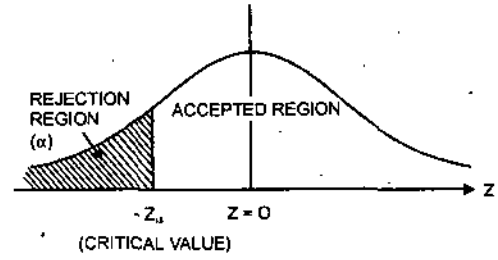


Fig. 11.2 Left tailed test ($\mu < \mu_0$).

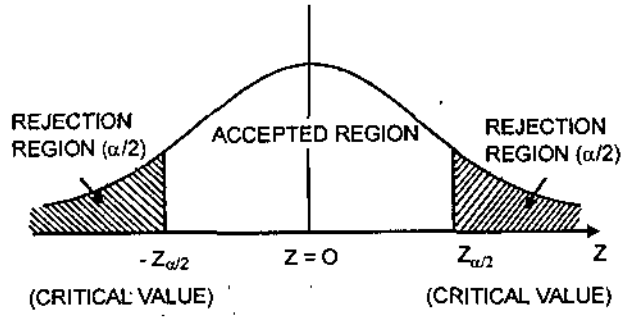


Fig. 11.3 Two tailed test ($\mu \neq \mu_0$).

NOTES

Steps in testing hypothesis:

1. Set up H_0.
2. Set up H_1.
3. Set up test statistic.
4. Set up the level of significance and critical value (s) using statistical table.
5. Compute the value of statistic using sample drawn from population.
6. Take decision. If the calculated value of test statistic lies in the critical region, reject H_0 *i.e.*, the assumption under the null hypothesis cannot be accepted. If the calculated value of test statistic lies in the accepted region *i.e.*, outside the critical region, accept H_0, *i.e.*, the assumption under H_0 can be taken as true value of the parameter.

Power function of the test. Let β = P [Type II error]. Then $1 - \beta$ is called the power function of testing H_0 against H_1.

Example 1. *The fraction of defective items in a large lot is P. To test H_0 : P = 0.1, one considers the number of defectives in a sample of 8 items and accept H_0 if the number of defectives is less than 6. Otherwise he rejects the hypothesis. What is the probability of type I error of this test ? What is the probability of type II error ?*

Solution. Given H_0 : P = 0.1.

Accept H_0 : If the no. of defectives found in the sample is less than equal to 6.

Reject H_0 : If the no. of defectives found in the sample is 7 or 8.

Here, the no. of defectives in a lot follows binomial distribution with $n = 8$ and P = 0.1, Q = 1 – P = 0.9.

P [Type I error] = Probability of rejecting H_0 when H_0 is true

= Probability of 7 or 8 defectives

$= \binom{8}{7} P^7 Q + \binom{8}{8} P^8 Q^0$

$= 8\ (0.1)^7\ (0.9) + (0.1)^8$

= 0.00000073.

P [Type II error] = Probability of accepting H_0 when H_0 is false

= Probability of defectives which are less than equal to 6

= 1 – [Probability of 7 or 8 defectives]

= 1 – *0.00000073*

= 0.99999927

TESTS FOR LARGE SAMPLES ($n > 30$)

(*a*) Testing of A Single Mean. (Inferences about a single mean)

1. Set up H_0 : $\mu = \mu_0$

NOTES

2. Set up $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$
3. Set up the test statistic

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$ which follows standard normal distribution.

4. Set up the level of significance α and the critical value as Z_{tab} from the normal table.
5. Compute the statistic, say Z_{cal}
6. Decisions :

H_1	*Reject* H_0 *if*
$\mu < \mu_0$	$Z_{cal} < -Z_{tab}$
$\mu > \mu_0$	$Z_{cal} > Z_{tab}$
$\mu \neq \mu_0$	$Z_{cal} < -Z_{tab}$, *i.e.*, $-Z_{\alpha/2}$
	$Z_{cal} > Z_{tab}$ *i.e.*, $Z_{\alpha/2}$

Note. If the population S.D. (σ) is not known for large sample we can take the sample S.D. (S) in the test statistic.

Example 2. *The mean lifetime of 100 picture tubes produced by a manufacturing company is estimated to be 5795 hours with a standard deviation of 150 hours. If μ be the mean lifetime of all the picture tubes produced by the company, test the hypothesis $\mu = 6000$ hours against $\mu \neq 6000$ hours at 5 % level of significance.*

Solution.

1. $H_0 : \mu = 6000$ hours
2. $H_1 : \mu \neq 6000$ hours
3. Test statistic : Here $n = 100$ *i.e.*, large sample. Population S.D. is not given but the sample S.D. is given as 150 hours.

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

4. $\alpha = 0.05$, $\alpha/2 = 0.025$

   H_1 shows, this is two tailed test.

   $Z_{\alpha/2} = 1.96$ so the critical values are 1.96 and –1.96.
5. Computation :

$$Z_{cal} = \frac{5795 - 6000}{150 / \sqrt{100}} = -13.67$$

6. Decision :

   Since $Z_{cal} < -1.96$ *i.e.*, it lies in the rejection region

   $\Rightarrow$ H_0 is rejected

   $\Rightarrow$ The claim produced by the company is not true.

Example 3. *A tire company claims that the lives of the tyres have mean of 42000 kilometres with standard deviation of 4000 kilometers. A change in the production process is believed to result in a better product. A test sample of 81 new tyres has a mean life of 42500 kilometers. Test at 5% level of significance that the new product is significantly better than the current one?*

Solution. 1. $H_0 : \mu = 42000$ km.

2. $H_1 : \mu > 42000$ km.

3. Here $n = 81$, Test statistic : $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

4. Alt. hypothesis shows, this is right tailed test.
   $\alpha = 0.05$, $Z_\alpha = 1.64$ which is the critical value.

5. Computation :

$$Z_{cal} = \frac{42500 - 42000}{4000 / \sqrt{81}} = 1.125$$

6. Decision :
   Since $Z_{cal} < Z_\alpha \Rightarrow Z_{cal}$ lies in the acceptance region
   $\Rightarrow$ H_0 is accepted
   $\Rightarrow$ H_1 is rejected
   $\Rightarrow$ New product is not significantly better than the current one.

NOTES

(*b*) Testing of Difference of Two Means. Consider two populations having the means μ_1 and μ_2 and the variance σ_1^2 and σ_2^2 respectively.

1. Set up $H_0 : \mu_1 - \mu_2 = k$

2. Set up $H_1 : \mu_1 - \mu_2 > k$ or $\mu_1 - \mu_2 < k$ or $\mu_1 - \mu_2 \neq k$ where k is a specified constant.

3. Set up the test statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - k}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

where, $\bar{x}_1$ = Means of sample size n_1 from the first population.

$\bar{x}_2$ = Means of sample size n_2 from the second population.

4. Set up the level of significance α and the critical value as Z_{tab} from the normal table.

5. Compute the statistic, says Z_{cal}.

6. Decisions :

H_1	*Reject H_0 if*
$\mu_1 - \mu_2 < k$	$Z_{cal} < -Z_{tab}$ *i.e.,* $(-Z_\alpha)$
$\mu_1 - \mu_2 > k$	$Z_{cal} > Z_{tab}$ *i.e.,* (Z_α)
$\mu_1 - \mu_2 \neq k$	$Z_{cal} < -Z_{tab}$, *i.e.,* $(-Z_{\alpha/2})$
	or, $Z_{cal} > Z_{tab}$ *i.e.,* $(Z_{\alpha/2})$

Note. When both n_1 and n_2 are greater than or equal to 30, the population S.D. can be estimated by sample S.D. in the statistic.

NOTES

Example 4. *A random sample of 100 villages was taken from a district A and the average height of the population per village was found to be 170 cm with a standard deviation of 10 cm. Another random sample of 120 villages was taken from another district B and the average height of the population per village was found to be 176 cm with a standard deviation of 12 cm. Is the difference between averages of the two populations statistically significant?*

Solution. 1. $H_0 : \mu_1 - \mu_2 = 0$, *i.e.*, there is no significant difference between the means of two populations.

2. $H_1 : \mu_1 - \mu_2 \neq 0$

3. Test statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

4. Let the level of significance $\alpha = 0.05$, $\alpha_{/2} = 0.025$, critical values are –1.96 and 1.96.

5. Computation :

Given $\bar{x}_1 = 170, s_1 = 10, n_1 = 100$

$\bar{x}_2 = 176, s_2 = 12, n_2 = 120$

$$Z_{cal} = \frac{170 - 176}{\sqrt{\frac{100}{100} + \frac{144}{120}}} = -4.05$$

6. Conclusion :

$Z_{cal} < -1.96$ *i.e.*, Z_{cal} lies inside the critical region

$\Rightarrow$ Reject H_0

$\Rightarrow$ There is significant difference between the means of two populations.

TESTS OF SMALL SAMPLES

(*a*) Testing of A Single Mean. Here sample is small ($n < 30$) and σ is unknown.

1. Set up $H_0 : \mu = \mu_0$
2. Set up $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$
3. Test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t - \text{distribution with } (n-1) \text{ degrees of freedom.}$$

4. Set up level of significance α and the critical value as t_{tab} (from table of t – distribution).
5. Compute the statistic say t_{cal}

6. Decisions :

H_1	*Reject* H_0, *if*
$\mu > \mu_0$	$t_{cal} > t_{tab}$ *i.e,* t_α
$\mu < \mu_0$	$t_{cal} < -t_{tab}$ *i.e,* $-t_\alpha$
$\mu \neq \mu_0$	$t_{cal} > t_{tab,}$ *i.e,* $t_{\alpha/2}$
	or, $t_{cal} < -t_{tab}$ *i.e,* $-t_{\alpha/2}$

All t_{tab} based on $(n - 1)$ degrees of freedom.

Example 5. *The mean breaking strength of a certain kind of metallic rope is 160 pounds. If six pieces of ropes (randomly selected from different rolls) have a mean breaking strength of 154.3 pounds with a standard deviation of 6.4 pounds, test the null hypothesis $\mu = 160$ pounds against the alternative hypothesis $\mu < 160$ pounds at 1% level of significance. Assume that the population follows normal distribution.*

Solution. 1. $H_0 : \mu = 160$ pounds

2. $H_1 : \mu < 160$ pounds

3. Since $n = 6$, the test statistic is taken as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

4. $\alpha = 0.01$, H_1 indicates it is left tailed test. Critical value at $6 - 1$ *i.e.*, 5 degrees of freedom is -3.365

5. Computation,

$$t_{cal} = \frac{154.3 - 160}{6.4/\sqrt{6}} = -2.18$$

6. Decision

Since $t_{cal} > -3.365$

$\Rightarrow$ it lies in the acceptance region

$\Rightarrow$ H_0 is accepted

$\Rightarrow$ Mean breaking strength of the metallic rope can be taken as 160 pounds.

(*b*) Testing of Difference of Two Means. Here n_1, n_2 or both are small (< 30) and the population variances are unknown but equal and two populations follow normal distribution.

1. Set up $H_0 : \mu_1 - \mu_2 = k$

2. Set up $H_1 : \mu_1 - \mu_2 > k$, or $\mu_1 - \mu_2 < k$ or $\mu_1 - \mu_2 \neq k$

3. Test statistic :

$$t = \frac{\bar{x}_1 - \bar{x}_2 - k}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

NOTES

Here, the statistic follows t – distribution with $n_1 + n_2 - 2$ degrees of freedom.

4. Set up the level of significance, α and the critical value say t_{tab} at $n_1 + n_2 - 2$ degrees of freedom.
5. Compute the test statistic as t_{cal}.
6. Decision

H_1	*Reject* H_0 *if*
$\mu_1 - \mu_2 > k$	$t_{cal} > t_{tab}$ *i.e,* t_α
$\mu_1 - \mu_2 < k$	$t_{cal} < -t_{tab}$ *i.e,* $-t_\alpha$
$\mu_1 - \mu_2 \neq k$	$t_{cal} < -t_{tab}$, *i.e,* $-t_{\alpha/2}$
	or, $t_{cal} > t_{tab}$ *i.e,* $t_{\alpha/2}$

Example 6. *The following are the number of sales with a sample of 6 sales people of gas lighter in a city A and a sample of 8 sales people of gas lighter in another city B made over a certain fixed period of time :*

City A : 63, 48, 54, 44, 59, 52

City B : 41, 52, 38 50, 66, 54, 44, 61

Assuming that the populations sampled can be approximated closely with normal distributions having the same variance, test $H_0 : \mu_1 = \mu_2$ *against* H_1: $\mu_1 \neq \mu_2$ *at the 5% level of significance.*

Solution. 1. $H_0 : \mu_1 = \mu_2$

2. $H_1 : \mu_1 \neq \mu_2$

3. Test statistic : $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

where, $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

4. $\alpha = 0.05$.

The alternate hypothesis shows that it is both tailed test.

$\therefore \quad \alpha/2 = 0.025$

Also $n_1 + n_2 - 2 = 6 + 8 - 2 = 12$

So the critical values are –2.179 and 2.179.

5. Computation :

$$\bar{x}_1 = 53.33, \; s_1^2 = \frac{243.33}{5}$$

$$\bar{x}_2 = 50.75, \; s_2^2 = \frac{653.5}{7}$$

Then, $s_p^2 = \frac{243.33 + 653.5}{12} = 74.74$

$\Rightarrow \quad s_p = 8.65$

$\therefore \quad t_{cal} = \frac{53.33 - 50.75}{8.65\sqrt{\frac{1}{6} + \frac{1}{8}}} = 0.55$

NOTES

6. Decision : Since $t_{cal} < 2.179$

$\Rightarrow$ It lies in the acceptance region

$\Rightarrow$ H_0 is accepted.

Example 7. *Measuring specimens of nylon yarn taken from two machines, it was found that 8 specimens from 1st machine had a mean denier of 9.67 with a standard deviation of 1.81 while 10 specimens from a 2nd machine had a mean denier of 7.43 with a standard deviation 1.48. Assuming the population are normal, test the hypothesis* $H_0 : \mu_1 - \mu_2 = 1.5$ *against*

$H_1 : \mu_1 - \mu_2 > 1.5$ *at 0.05 level of significance.*

Solution. 1. $H_0 : \mu_1 - \mu_2 = 1.5$

2. $H_1 : \mu_1 - \mu_2 > 1.5$

3. Test statistic : $t = \frac{\bar{x}_1 - \bar{x}_2 - k}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

4. $\alpha = 0.05$, the alternative hypothesis shows that it is right tailed test.

Also $n_1 + n_2 - 2 = 8 + 10 - 2 = 16$

So the critical value is $t_{0.05,\ 16} = 1.746$.

5. Computation :

$$S_p^2 = \frac{7 \times (1.81)^2 + 9 \times (1.48)^2}{16} = 2.665 \Rightarrow S_p = 1.632$$

$$t_{cal} = \frac{9.67 - 7.43 - 1.5}{1.632 \cdot \sqrt{\frac{1}{8} + \frac{1}{10}}} = 0.956$$

6. Decision : Since $t_{cal} < 1.746$

$\Rightarrow$ It lies in the accepted region.

$\Rightarrow$ H_0 is accepted.

(c) Paired *t*-Test.

1. Let (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) be n pairs of observations from a bivariate normal population, and the values of x and y are not independent, *i.e.*, they are correlated.

2. $H_0 : \mu_x = \mu_y$

   $H_1 : \mu_x \neq \mu_y$ or $\mu_x > \mu_y$ or $\mu_x < \mu_y$

NOTES

3. Statistic : Let $d_i = x_i - y_i$ for the i-th pair ($i = 1, 2, ..., n$)

and $\bar{d} = \frac{\Sigma d_i}{n}$, $s^2 = \Sigma(d_i - \bar{d})^2 / n$

Then $t = \frac{\bar{d}}{s/\sqrt{n-1}}$ follows t-distribution with $(n - 1)$ degrees of freedom.

4. Set up the level of significance and the critical value (t_{tab}).

5. Decision :

H_1	*Reject* H_0, *if*
$\mu_x > \mu_y$	$t_{cal} > t_{tab}$ *i.e.*, t_α, $n - 1$
$\mu_x < \mu_y$	$t_{cal} < -t_{tab}$ *i.e.*, $-t_\alpha$, $n - 1$
$\mu_x \neq \mu_y$	$t_{cal} > t_{tab}$ *i.e.*, $t_{\alpha/2}$, $n - 1$
	or, $t_{cal} < -t_{tab}$ *i.e.*, $-t_{\alpha/2}$, $n - 1$

Example 8. *An I.Q. test was administered to 5 persons before and after they were trained. The results are given below :*

	I	*II*	*III*	*IV*	*V*
I. Q. before training :	*110*	*120*	*123*	*132*	*125*
I. Q. after training :	*120*	*118*	*125*	*136*	*121*

Test whether there is any change in I.Q. after the training programme [use 1% level of significance].

Solution. 1. $H_0 : \mu_x = \mu_y$

i.e., there is no change in the mean score before and after training.

$H_1 : \mu_x \neq \mu_y$

2. Computation :

x	110	120	123	132	125	Total
y	120	118	125	136	121	
$d = x - y$	–10	2	–2	–4	4	–10

$$\bar{d} = \frac{-10}{5} = -2$$

$$s^2 = \frac{1}{5}[(-10 + 2)^2 + (2 + 2)^2 + (-2 + 2)^2 + (-4 + 2)^2 + (4 + 2)^2]$$

$$= 24$$

Statistic : $t_{cal} = \frac{\bar{d}}{s/\sqrt{n-1}} = \frac{-2}{\sqrt{24}/2} = -0.82.$

3. Critical value : $\alpha = 0.01$, $\alpha/2 = 0.005$, H_1 indicates two tailed test. Then $t_{0.005,4} = 4.604$, *i.e.*, critical values are – 4.604 and 4.604.

4. Decision:

Since $t_{cal} > -4.604$

$\Rightarrow$ H_0 is accepted

$\Rightarrow$ There is no significant change in the mean I.Q. after the training programme.

NOTES

TESTING OF PROPORTION

(*a*) Single Proportion

1. Set up $H_0 : P = p_0$.
2. Set up $H_1 : P > p_0$ or $P < p_0$ or $P \neq p_0$.
3. Set up the test statistics

$$Z = \frac{X - np_0}{\sqrt{np_0 (1 - p_0)}}$$

which approximately follows the standard normal distribution.

4. Set up level of significance α and the critical value say Z_{tab}.
5. Compute the statistic.
6. Decisions :

Here the conclusions are the similar as given in section 2(*a*) for testing *of a single mean of a large sample.*

Example 9. *A die was thrown 500 times and six resulted 100 times. Do the data justify the hypothesis of an unbiased die ?*

Solution. Let us assume that the die is unbiased and the probability of obtaining a six with the die is 1/6.

1. $H_0 : P = 1/6$
2. $H_1 : P \neq 1/6$
3. Test statistic :

$$Z = \frac{X - p_0}{\sqrt{p_0 (1 - p_0)/n}}$$

4. Let $\alpha = 0.05$. Alternative hypothesis suggests for two tailed test.

$\therefore$ $\alpha/2 = 0.025$, critical values are -1.96 and 1.96.

5. Computation : Here out of 500 times throw, six resulted 100 times. So the observed value of proportion (X) of six is

$$X = \frac{100}{500} = 0.2, \; p_0 = 1/6 = 0.167$$

$$\therefore \quad Z_{cal} = \frac{0.2 - 0.167}{\sqrt{\frac{1}{6} \times \frac{5}{6} \times \frac{1}{500}}} = 1.98$$

6. Decision : Since $Z_{cal} > 1.96$

$\Rightarrow$ H_0 is rejected

NOTES

$\Rightarrow$ The given data do not justify the hypothesis of an unbiased die.

Example 10. *A manufacturer claimed that at least 95% of the components of an electronic circuit board which he supplied, conformed to specifications. A random sample of 220 components showed that only 185 were upto the standard. Test his claim at 1% level of significance.*

Solution.

1. $H_0 : P = 0.95$ (components conforming specifications)
2. $H_1 : P < 0.95$
3. Test statistic :

$$Z = \frac{X - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

4. $\alpha = 0.01$, alternative hypothesis shows left tail test. Critical value is $Z_{tab} = -2.33$.
5. Computation :

$$\text{Observed proportions} = X = \frac{185}{220} = 0.84$$

$$n = 220$$

$$Z_{cal} = \frac{0.84 - 0.95}{\sqrt{(0.95)\,(0.05)/220}} = -7.49.$$

6. Decision : Since $Z_{cal} < -2.33$

$\Rightarrow$ Z_{cal} is inside the critical region.

$\Rightarrow$ H_0 is rejected.

$\Rightarrow$ Manufacturer's claim cannot be accepted.

(b) Difference Two Proportions. Let p_1 and p_2 be the proportions in two large samples of sizes n_1 and n_2 drawn respectively from two populations. To test whether the differences $p_1 - p_2$ as observed in the samples has arises only due to fluctuation of sampling.

1. Set up $H_0 : P_1 = P_2$
2. Set up $H_1 : P_1 \neq P_2$
3. Test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where, $$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}, \quad q = 1 - p$$

Here, Z approximately follows standard normal distribution.

4. Set the level of significance and the critical value say, Z_{tab} using normal table.
5. Compute the statistic as Z_{cal}.
6. Decision:

Here, the decisions are the similar as given in testing of difference of two means for large samples in section 2 (*b*).

Example 11. *A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved ?*

NOTES

Solution.

1. Set up $H_0 : P_1 = P_2$ (*i.e.*, proportions of defectives before and after overhauling are equal).
2. $H_1 : P_1 > P_2$
3. Test statistic :

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

where, $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$, $q = 1 - p$.

4. Let $\alpha = 0.05$. Alternative hypothesis shows it is a right tailed test. So the critical value is $Z_{tab} = 1.645$.
5. Here $n_1 = 400,\ n_2 = 300$

$$p_1 = \frac{20}{400} = \frac{1}{20}, \quad p_2 = \frac{10}{300} = \frac{1}{30}$$

$$p = \frac{20 + 10}{400 + 300} = \frac{3}{70}, \quad q = 1 - p = \frac{67}{70}$$

$$Z_{cal} = \frac{\frac{1}{20} - \frac{1}{30}}{\sqrt{\frac{3}{70} \cdot \frac{67}{70}\left(\frac{1}{400} + \frac{1}{300}\right)}} = 1.08$$

6. Decision :

Since $Z_{cal} < Z_{tab}$

$\Rightarrow$ It lies in the acceptance region.

$\Rightarrow$ H_0 is accepted.

$\Rightarrow$ Two population proportions before and after overhauling are equal.

$\Rightarrow$ Machine has not improved.

TESTING OF A SINGLE VARIANCE

Consider a normal population $N(\mu, \sigma^2)$.

1. Set up $H_0 : \sigma^2 = \sigma_0^2$
2. Set up $H_1 : \sigma^2 < \sigma_0^2$ or $\sigma^2 > \sigma_0^2$, or $\sigma^2 \neq \sigma_0^2$
3. Test statistic:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}, \quad S^2 = \text{Sample variance (unbiased)}$$

which follows chi-square distribution with $(n - 1)$ degrees of freedom.

4. Set up the level of significance α and the critical point as χ^2_{tab} using chi-square table with $(n - 1)$ degrees of freedom.

NOTES

5. Compute the statistic, say χ^2_{cal}.

6. Decision :

H_1	*Reject* H_0, *if*
$\sigma^2 < \sigma_0^2$	$\chi^2_{cal} < \chi^2_{tab}$ *i.e.,* $\chi^2_{1-\alpha, n-1}$
$\sigma^2 > \sigma_0^2$	$\chi^2_{cal} > \chi^2_{tab}$ *i.e.,* $\chi^2_{\alpha, n-1}$
$\sigma^2 \neq \sigma_0^2$	$\chi^2_{cal} < \chi^2_{tab}$ *i.e.,* $\chi^2_{1-\alpha/2, n-1}$ or $\chi^2_{cal} > \chi^2_{tab}$ *i.e.,* $\chi^2_{\alpha/2, n-1}$

Note. The above test is also used for testing standard deviation.

Example 12. *Use the 0.05 level of significance to test the null hypothesis that $\sigma = 0.022$ inch for the diameters of certain wire rope against the alternative hypothesis that $\sigma \neq 0.022$ inch, given that a random sample of size 18 yielded $S^2 = 0.000324$.*

Solution. 1. $H_0 : \sigma = 0.022$ inch.

2. $H_1 : \sigma \neq 0.022$ inch.

3. Test statistic:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

which follows chi-square distribution with $(n - 1)$ degrees of freedom.

4. Here $\alpha = 0.05$. Alternative hypothesis shows that it is right tailed test. $\alpha/2 = 0.025$.

Critical values are $\chi^2_{0.975, 17} = 7.564$ and

$\chi^2_{0.025, 17} = 30.191$.

5. Computation :

$$\chi^2_{cal} = \frac{17(0.000324)}{(0.022)^2} = 11.38.$$

6. Decision :

Since $\chi^2_{cal} < 30.191$ and $\chi^2_{cal} > 7.564$

it lies in the acceptance region.

$\Rightarrow H_0$ is accepted.

$\Rightarrow$ True diameter of the metallic rope can be 0.022 inch.

TESTING OF EQUALITY OF TWO VARIANCES

Consider two independent random samples from two normal populations.

1. Set up $H_0 : \sigma_1^2 = \sigma_2^2$.

2. Set up $H_1 : \sigma_1^2 < \sigma_2^2$ or $\sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$.

3. Set up the level of significance α.

In this case the test statistic follows F-distribution and depending on H_1, the critical value is obtained. We list in the following :

H_1	*Test Statistic*	*Reject H_0, if*
$\sigma_1^2 < \sigma_2^2$	$F = \frac{S_2^2}{S_1^2}$	$F_{cal} > F_{\alpha, n_2 - 1, n_1 - 1}$
$\sigma_1^2 > \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$F_{cal} > F_{\alpha, n_1 - 1, n_2 - 1}$
$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{S_L^2}{S_l^2}$	$F_{cal} > F_{\alpha/2, n_L - 1, n_l - 1}$

where S_L^2 represent the larger of the two sample variances and S_l^2 the smaller one corresponding to the sample sizes n_L and n_l respectively.

Example 13. *It is desirable to determine whether there is less variability in the intensity of light by two bulbs made by company A and company B respectively in certain locations. If the independent random samples of size 16 of the two bulbs yield $S_1 = 1.5$ foot-candles and $S_2 = 1.75$ foot-candles, test the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ at the 0.01 level of significance.*

Solution. 1. $H_0 : \sigma_1^2 = \sigma_2^2$.

2. $H_1 : \sigma_1^2 < \sigma_2^2$.

3. *Test statistic:*

$$F = \frac{S_2^2}{S_1^2}$$

which follows F-distribution with degrees of freedom 15 and 15.

4. $\alpha = 0.01$. The alternative hypothesis shows that it is left tailed test.

Critical value $F_{0.01, 15, 15} = 3.52$.

5. Given $S_1 = 1.5$ and $S_2 = 1.75$

$$\therefore \quad F_{cal} = \frac{(1.75)^2}{(1.5)^2} = 1.36.$$

6. Decision : Since $F_{cals} < 3.52$, H_0 is accepted.

$\Rightarrow$ There is no variability in the intensity of light by two bulbs.

GOODNESS OF FIT TEST

This is called distribution free test, *i.e.*, the population may not be normal. Here the null hypothesis is taken as

H_0 : observations are in good agreement with a hypothetical distribution/ population.

NOTES

NOTES

If O_i be the observed frequencies and e_i be the expected frequencies (i = 1, 2, ..., n) then the statistic

$$\chi^2 = \sum_i \frac{(O_i - e_i)^2}{e_i}$$

follows chi-square distribution with (n–1) degrees of freedom. If the calculated value of the statistic is greater than the tabulated value of χ^2 at a level of significance α, then the null hypothesis is rejected.

CHI-SQUARE TEST OF INDEPENDENCE

Consider a $r \times c$ table in which date can be explained in two ways having r rows and c columns. Also this is called *contingency table.* A 3 × 3 table can be taken as

	c_1	c_2	c_3	*Total*
r_1	O_{11}	O_{12}	O_{13}	rt_1
r_2	O_{21}	O_{22}	O_{23}	rt_2
r_3	O_{31}	O_{32}	O_{33}	rt_3
Total	ct_1	ct_2	ct_3	*Grand total*

where r's and c's can be taken as attributes and O_{ij} = observed frequencies.

In the above, the expected cell frequencies can be calculated as

$$e_{ij} = \frac{(i\text{-th row total})\ (j\text{-th column total})}{\text{grand total}}$$

For example, $e_{11} = \dfrac{rt_1 \cdot ct_1}{\text{grand total}}$ etc.

Then $$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

follows chi-square distribution with (r – 1) (c – 1) degrees of freedom.

We set the null hypothesis as the attributes are independent

i.e., $H_0 : p_{i1} = p_{i2} = ... = p_{ic}$, i = 1, 2, ..., r

where p_{ij} = Probability of obtaining an observation belonging to the i-th row and the j-th column and $\sum_{i=1}^{r} p_{ij} = 1$ for each column.

We shall obtain the tabulated chi-square value at a level of significance and (r – 1) (c – 1) degrees of freedom.

If $\chi^2_{cal} > \chi^2_{tab}$, then H_0 is rejected.

If $\chi^2_{cal} < \chi^2_{tab}$, then H_0 is accepted.

A 2 × 2 TABLE (SIMPLIFIED FORM)

	Attribute A		*Total*
Attribute B	a	b	$R_1 = a + b$
	c	d	$R_2 = c + d$
Total	$c_1 = a + c$	$c_2 = b + d$	Grand total $= N = R_1 + R_2 = c_1 + c_2$.

In this case, the statistic can be calculated as

$$\chi^2 = \frac{N(ad - bc)^2}{R_1 R_2 c_1 c_2}$$

with $(2 - 1)(2 - 1) = 1$ degree of freedom.

YATE'S CORRECTION

Due to one degree of freedom one of the four cell frequencies can be arbitrarily given if the row and column totals should remain fixed. Hence Yate has suggested the following correction in calculating the chi-square statistic.

If $ad > bc$, reduce a and d by 0.5 and increase b and c by 0.5,

If $ad < bc$, increase a and d by 0.5 and reduce b and c by 0.5.

$$\therefore \quad \chi^2 \text{ (corrected)} = \frac{N\left\{|ad - bc| - N/2\right\}^2}{R_1 R_2 c_1 c_2}$$

Example 14. *Fit a Poisson distribution to the following data and test the goodness of fit.*

x	0	1	2	3	4
f	112	73	30	4	1

Solution. Here mean, $\bar{x} = \frac{\Sigma xf}{\Sigma f} = \frac{149}{220} = 0.68 = \lambda$

Expected frequencies are obtained by

$$220 . e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, 3$$

Therefore the fitted distribution is

x	0	1	2	3	4
Expected f	111	76	26	6	1

Let H_0 : Poisson distribution is a good fit to the above data.
Statistics :

$$\chi^2 = \Sigma \frac{(O_i - e_i)^2}{e_i}$$

$$= \frac{(112 - 111)^2}{111} + \frac{(73 - 76)^2}{76} + \frac{(30 - 26)^2}{26} + \frac{(4 - 6)^2}{6} + \frac{(1 - 1)^2}{1}$$

$$= 1.41$$

NOTES

Let $\alpha = 0.05$

Degrees of freedom = 5 – 1 – 1 = 3

(1 *d.f.* lost due to $\sum O_i = \sum E_i$

1 *d.f.* lost due to estimate of mean)

Critical value is $\chi^2_{3,0.05} = 7.82$

Since $\chi^2_{cal} < 7.82$

$\Rightarrow$ H_0 is accepted

$\Rightarrow$ Poisson distribution is a good fit to the given data.

Example 15. *Fit a binomial distribution to the following data and test the goodness of fit.*

x	0	1	2	3	4	5
f	8	20	24	14	2	2

Solution. Mean $= \frac{\sum xf}{\sum f} = \frac{128}{70}$, $n = 5$, $N = \sum f = 70$

Therefore $np = \frac{128}{70} \Rightarrow p = \frac{128}{350} = 0.3657$, $q = 0.6343$

The expected frequencies of the fitted binomial distribution can be calculated from $70\ (0.6343 + 0.3657)^5$

Hence we obtain

x	0	1	2	3	4	5
Expected f	7	21	24	14	4	0

Let H_0 : Binomial distribution is a good fit to the above data.

Let us pool the last two expected values (which are less than 5) so that the pooled values in this case are 4 and 4 respectively for observed and expected values.

Now *d.f.* = 6 – 1 – 3 – 1 = 1

(1 *d.f.* lost due to $\sum O_i = \sum E_i$

3 *d.f.* lost due to estimate of p, q and mean

1 *d.f.* lost due to pooling of two expected values)

Let $\alpha = 0.05$, then $\chi^2_{0.05,1} = 3.84$

Also $\chi^2_{cal} = \sum \frac{(O_i - e_i)^2}{e_i} = \sum \frac{(f - F)^2}{F} = 0.19$

Since, $\chi^2_{cal} < 3.84$

$\Rightarrow$ H_0 is accepted

$\Rightarrow$ Binomial distribution is a good fit to the above data.

Example 16. *The results of polls conducted 2 weeks and 4 weeks before an election, are shown in the following table :*

	Two weeks before	*Four weeks before*	*Total*
For candidate A	99	112	211
For candidate B	101	88	189
Total	200	200	400

Use the 0.05 level of significance to test whether there has been a change in opinion between the two polls.

Solution. Let H_0 : Opinion does not change between the two polls.

H_1 : Opinion changes between the two polls.

Since degree of freedom = (2 – 1) (2 – 1) = 1, we have to use Yate's corrected chi-square statistic.

Here N = 400, R_1 = 211, R_2 = 189, $c_1 = c_2$ = 200

$$\therefore \quad \chi^2 = \frac{400\left\{|99 \times 88 - 112 \times 101| - 400/2\right\}^2}{211 \times 189 \times 200 \times 200}$$

$$= \frac{\{2600 - 200\}^2}{211 \times 189 \times 100}$$

$$= 1.44$$

Also $\chi^2_{1, 0.05} = 3.84$

Since $\chi^2_{cal} < 3.84$

$\Rightarrow$ H_0 is accepted

$\Rightarrow$ Opinion does not change between the two polls.

Example 17. *A random sample of 220 students in a college were asked to give opinion in terms of yes or no about the winning of their college cricket team in a tournament. The following data are collected :*

	Class in college		
	Ist year	*IInd year*	*IIIrd year*
Yes	43	20	37
No	23	57	40

Test whether there is any association between opinion and class in college [use 5% level of significance].

Solution. We display the contingency table with both observed frequencies and expected frequencies:

	Class in college			
	Ist year	*IInd year*	*IIIrd year*	*Total*
Yes	43 $\frac{100 \times 66}{220} = 30$	20 $\frac{100 \times 77}{220} = 35$	37 $\frac{100 \times 77}{220} = 35$	100
No	23 $\frac{120 \times 66}{220} = 36$	57 $\frac{120 \times 77}{220} = 42$	40 $\frac{120 \times 77}{220} = 42$	120
Total	66	77	77	220

NOTES

H_0 : There is a association between opinion and class in the college.

H_1 : There is no association between opinion and class in the college.

Here degrees of freedom = (3 – 1) (2 – 1) = 2

Critical value is $\chi^2_{2,0.05} = 5.99$

Computation of statistic:

$$\chi^2 = \frac{(43-30)^2}{30} + \frac{(20-35)^2}{35} + \frac{(37-35)^2}{35} + \frac{(23-36)^2}{36} + \frac{(57-42)^2}{42} + \frac{(40-42)^2}{42}$$
$$= 22.32$$

Since calculated χ^2 > critical value

$\Rightarrow$ H_0 is rejected

$\Rightarrow$ H_1 can be accepted, *i.e.*, there is no association between opinion and class in the college.

SUMMARY

- The decision about the problems in a statistical population on the basis of sample observations are known as a statistical hypothesis.
- The value which separates the vertical region and the acceptance region is called critical value which is set by seeing the alternative hypothesis.

PROBLEMS

(Testing of Mean/Difference of Means)

1. The manufacturer of television tubes knows from the past experience that the average life of tube is 2000 hrs. with a s.d. of 200 hrs. A sample of 100 tubes has an average life of 1950 hrs. Test at the 0.01 level of significance to see if this sample came from a normal population of mean 2000 hrs.
2. The mean lifetime of 100 electric bulbs produced by a manufacturing company is estimated to be 1570 hrs with a s.d. of 120 hrs. If μ is the mean lifetime of all the bulbs produced by the company, test the hypothesis μ = 1600 hrs, against the alternative hypothesis $\mu \neq 1600$ hrs, using 5% level of significance.
3. A sample of 400 students is found to have a mean height of 171.38 cms. Can it be reasonably regarded as a sample from a large population with mean height 171.17 cm and s.d. 3.30 cms ?
4. The mean weight of a random sample of size 100 from a student's population is 65.8 kgs and the standard deviation is 4 kgs. Test at 5% level of significance that the student's population weight is below 72 kgs.
5. The sales manager of a large company conducted a sample survey in states A and B taking 400 sample salesman in each case. The results were :

	State A	*State B*
Average sales	Rs. 2500	Rs. 2200
Standard deviation	Rs. 400	Rs. 550

Test whether the average sales is the same in the two states at 1% level of significance.

6. The mean yield of sunflower seeds from a district A was 200 lbs with s.d. = 10 lbs per acre from a sample of 100 plots. In another district B, the mean yield was 210 lbs with s.d. = 12 lbs from a sample of 120 plots. Assuming that the s.d. of yield in the entire state was 12 lbs, test whether there is any significant difference between the mean yield of crops in the two districts (use 1% level of significance).

7. An investigation of two kinds of machines in a laboratory showed that 52 failures of the first kind of machine took on the average 74 minutes to repair with a standard deviation of 15 minutes, while 68 failures of the second kind of machine took on the average 92 minutes to repair with a standard deviation of 20 minutes. Test the hypothesis that on the average it takes an equal amount of time to repair either kind of machines.

8. The percentage of carbon content of a certain variety of steel has a standard specification 0.05. For 15 samples of steel the percentage of carbon content were found to have an average 0.0482 and standard deviation 0.0012. Do these data reasonably conform to the standard specification ? (Assume that the population of percentages of carbon content is normal)

   [Given $P(|t| > 4.819) < 0.001$ for 11 *d.f.*]

9. The heights of 10 residents of a given locality are found to be 70, 68, 62, 68, 61, 68, 69, 65, 64 and 66 inches. Is it reasonable to believe that the average height is greater than 64 inches ? [Use 5% level of significance].

10. A fertilizer machine is set to give 12 kg of nitrate for every quintal bag of fertilizer. Ten 100 kg bags are examined. The percentages of nitrate are as follows :

    11, 14, 13, 12, 13, 12, 13, 14, 11, 12.

    Is there any reason to believe that the machine is defective ? (Use 5% level of significance).

11. A drug was administered to 10 patients, and the increments in their blood pressure were recorded to be

    6, 3, –2, 4, –3, 4, 6, 0, 0, 2.

    Is it reasonable to believe that the drug has no effect on change of blood pressure ? (Use 5% level of significance).

12. A sample of size 10 is drawn from each of two normal populations having the same variance which is unknown. If the mean and variance of the sample from the first population are 7 and 26 and those of the sample from the second population are 4 and 10, test at 5% significance level if the two populations have the same mean.

13. The sales data of an item in six shops before and after a special promotional campaign are as under :

Shops	*A*	*B*	*C*	*D*	*E*	*F*
Before Campaign	53	28	31	48	50	42
After Campaign	58	29	30	55	56	45

Can the campaign be judged to be a success ? (Use 5% level of significance)

(Testing of Proportion/Difference of Proportions)

14. In a sample of 400 parts manufactured by a factory, the number of defective parts was found to be 30. The company however claims that only 5% of their product is defective. Is the claim tenable ?

15. In a sample of 500 people in a town, 280 are tea drinkers and the rest are coffee drinkers. Can we assume that both coffee and tea are equally popular in the town at 1% level of significance ?

16. A sample survey results show that out of 800 literate people 480 are employed, whereas out of 600 illiterate people only 350 are employed. Can the difference between two proportions of employed persons be ascribed due to sampling fluctuations ?

NOTES

17. In a sample of 600 students of a certain college, 400 are found to use dot pens. In another college from a sample of 900 students 450 were found to use dot pens. Test whether the two colleges are significantly different with respect to the habit of using dot pens. (Use 5% level of significance).

18. In a certain city A, 100 men in a sample of 400 are found to be smokers. In another city B, 300 men in a sample of 800 are found to be smokers. Does this indicate that there is a greater proportion of smokers in B than A ?

19. A transportation company claims that only 7% of all lost luggage is never found. If in a random sample, 18 of 200 pieces of lost luggage are not found, test the null hypothesis $p = 0.07$ against the alternative hypothesis $p > 0.07$ at 5% level of significance.

(Testing of Variances/Chi-square and F-Tests)

20. Weights (in kg.) of 10 students are given below :

38, 40, 45, 53, 47, 43, 55, 48, 52, 49.

Can we say that the variance of the distribution of weights of all students from which the above sample of 10 students was drawn, is equal to 20 square kg. ?

21. A random sample of size 20 from a normal population gives a sample mean of 42 and a sample standard deviation of 6. Test the hypothesis that the population s.d. is 9.

22. If 10 determinations of the specific heat of iron have a standard deviation of 0.0075, test the null hypothesis $\sigma = 0.011$ for such determinations. Use the alternative hypothesis $\sigma \neq 0.011$ at 5% level of significance.

23. Two random samples are drawn from two populations and the following results were obtained :

Sample I	16	17	18	19	20	21	22	24	26	27		
Sample II	19	22	23	25	26	28	29	30	31	32	35	36

Find the variances of the two samples and test whether the two populations have the same variance (use 5% level of significance).

24. The following results were obtained from two independent random samples :

	Sample size	*Mean*	*S.D.*
Sample I	6	24	2.2
Sample II	5	29	3.1

Test whether the two samples may be regarded as drawn from the same normal population (use 5% level of significance).

(Chi-square Test/Goodness of Fit Test)

25. The number of road accidents per week in a certain area were as follows :

12, 8, 20, 2, 14, 10, 15, 6, 9, 4.

Are these frequencies in agreement with the belief that accident conditions were the same during the 10-week period ?

26. A chemical extraction plant processes sea water to collect sodium chloride and magnesium. It is known that sea water contains sodium chloride, magnesium and other elements in the ratio of 62 : 4 : 34. A sample of 200 tonnes of sea water has resulted in 130 tonnes of sodium chloride and 6 tonnes of magnesium. Are these data consistent with the known composition of sea water at 5% level ?

27. The following table gives the number of accounting clerks committing errors and not committing errors among trained and untrained clerks working in an organization:

	Number of clerks		
	Committing errors	*Not committing errors*	*Total*
Trained	70	530	600
Untrained	155	745	900
Total	225	1275	1500

Test the effectiveness of training in preventing the errors.

ANSWERS

NOTES

1. Two tailed test, $Z_{cal} = -2.25$, accept H_0.
2. $Z_{cal} = -2.5$, reject H_0.
3. Two tailed test, $Z_{cal} = 1.27$, accept H_0 at 5% level of significance.
4. $Z_{cal} = -15.5$, accept $H_1 : \mu < 72$ kg.
5. Two tailed test, $Z_{cal} = 8.82$, $H_0 : \mu_1 = \mu_2$ is rejected.
6. Two tailed test, $Z_{cal} = -6.15$, reject H_0.
7. Two tailed test, $Z_{cal} = -5.63$, reject H_0 at 5% level of significance.
8. Two tailed test, $t_{cal} = -5.81$, reject H_0.
9. Right tailed test, $t_{cal} = 0.72$. Average height is not greater than 64 inches.
10. Two tailed test, $t_{cal} = 1.46$, accept H_0.
11. Drug has no effect, $t_{cal} = 0.67$.
12. $t_{cal} = 1.58$, accept $H_0 : \mu_1 = \mu_2$.
13. Left tailed test, $t_{cal} = -2.78$, reject $H_0 : \mu_1 = \mu_2$ *(mean sales are same)*.
14. No. $Z_{cal} = 2.29$, $H_0 : P = 0.05$, $H_1 : P > 0.05$.
15. No. $|Z_{cal}| = 2.70 > 2.58$.
16. $Z_{cal} = 0.6415$, two tailed test, accept $H_0 : P_1 = P_2$ claim is correct.
17. $Z_{cal} = 6.38$, reject $H_0 : P_1 = P_2$, two tailed test.
18. $|Z_{cal}| = 4.33$, $H_0 : P_1 = P_2$ is rejected at 5% level of significance.
19. $Z_{cal} = 1.11$, accept H_0.
20. $\chi^2_{cal} = 14$, accept $H_0 : \sigma^2 = 20$(right tailed test) at 5% level.
21. $\chi^2_{cal} = 8.89$, accept $H_0 : \sigma = 6$ (right tailed test) at 5% level.
22. $\chi^2_{cal} = 4.18$, accept H_0.
23. $F_{cal} = 27.1/14 \approx 1.94$, *d.f.* = (11, 9), accept $H_0 : \sigma_1^2 = \sigma_2^2$, right tailed test.
24. $H_0 : \sigma_1^2 = \sigma_2^2$, $F_{cal} = 2.068$, H_0 accepted ($H_1 : \sigma_2^2 > \sigma_1^2$).
    But $H_0 : \mu_1 = \mu_2$, is rejected against $H_1 : \mu_1 \neq \mu_2$.
    So the samples do not come from the same population.
25. $\chi^2_{cal} = 26.6$, claim is rejected at 5% level of significance.
26. $\chi^2_{cal} = 1.025$, claim is accepted.
27. $\chi^2_{cal} = 8.7147$, H_1 : claim is correct, is accepted at 5% level and *d.f.* = 1.

CHAPTER 12 ANALYSIS OF VARIANCE

NOTES

★ STRUCTURE ★

- Introduction
- Completely Randomised Design (CRD)/One-way Classification
- Randomized Block Design/Two-Way Classification (R.B.D.)
- Summary
- Problems

INTRODUCTION

Analysis of variance (ANOVA) implies a statistical technique on the variations to a group of causes. This technique is closely related to *design of experiment* which may be defined as "the logical construction of the experiment in which the degree of uncertainty with which the inference is drawn may be well defined."

In a comparative experiment various objects of comparisons are termed as *treatments*. Also the whole experimental unit is divided into subgroups, preferably homogeneous are called *blocks*.

Prof. R. A. Fisher has outlined the basic principles of the design of experiments which are

(*i*) Replication – means the repetition of the treatments under investigation.

(*ii*) Randomisation–provides a logical basis for the validity of the statistical tests of significance.

(*iii*) Local control–which is a process of reducing the experimental error by dividing the heterogeneous area into homogeneous blocks and thus increases the efficiency of the design.

COMPLETELY RANDOMISED DESIGN (CRD) /ONE-WAY CLASSIFICATION

Consider the following results of k independent random samples each of size n, from k different populations:

Population 1	:	x_{11},	x_{12},	... x_{1n}	:	Treatment 1
Population 2	:	x_{21},	x_{22}, .	.. x_{2n}	:	Treatment 2
Population k	:	x_{k1},	x_{k2},	... x_{kn}	:	Treatment k

Here x_{ij} refers to the j-th value of the i-th population and the corresponding random variables X_{ij} which are all independent normally distributed with the respective means μ_i and the common variance σ^2.

Consider the model,

$$x_{ij} = \mu + \alpha_i + e_{ij},\ i = 1, 2, ..., k,\ j = 1, 2, ..., n$$

NOTES

Here, μ is referred to as the Grand mean,

α_i is referred to as treatment effects such that $\sum_{i=1}^{k} \alpha_i = 0$

and e_{ij} random errors are identically distributed as N $(0, \sigma^2)$.

Hypothesis:

H_0 : Population means are all equal

i.e., $$\mu_1 = \mu_2 = ... = \mu_k$$

i.e., samples were obtained from k populations with equal means.

Equivalently, $\alpha_i = 0$, $i = 1, 2, ..., k$ *i.e.*, there is no special effect due to any population.

$$H_1 : \alpha_i \neq 0 \text{ for at least one value of } i.$$

Then we construct the following ANOVA TABLE :

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean square*	*F (Test statistic)*
Treatments	$k - 1$	SS (Tr)	MS (Tr)	$\frac{\text{MS (Tr)}}{\text{MSE}}$
Error	$k\ (n - 1)$	SSE	MSE	
Total	$kn - 1$	SST		

where SST = Total sum of squares

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2 - \frac{1}{kn} T_{..}^2$$

SS(Tr) = Treatment sum of squares

$$= \frac{1}{n} \sum_{i=1}^{k} T_{i.}^2 - \frac{1}{kn} T_{..}^2$$

SSE = Error sum of squares,

SST = SS (Tr) + SSE,

$T_{i.}$ = Total of the values obtained for the i-th treatment

$T_{..}$ = Grand total of all nk observations

$$\text{MS (Tr)} = \frac{\text{SS (Tr)}}{k - 1} \text{ (Treatment mean square)}$$

NOTES

$$MSE = \frac{SSE}{k(n-1)} \text{ (Error mean square)}$$

$$F = F_{cal} = \frac{MS(Tr)}{MSE} \text{ which follows F distribution.}$$

Let L = level of significance, then

$$F_{tab} = F_{\alpha,\ \alpha-1,\ k(n-1)}$$

Conclusion:

Reject H_0 if $F_{cal} > F_{tab}$

Accept H_0 if $F_{cal} < F_{tab}$.

Example 1. *A test was given to five students taken at random from the Xth class of three schools of a town. The individual scores are.*

School I	77	81	71	76	80
School II	72	58	74	66	70
School III	76	85	82	80	77

Carry out the analysis of variance and state your conclusions.

Solution. 1. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ (No difference between the schools)

$H_1 : \alpha_i \neq 0$ for at least one value of i.

2. Let $\alpha = 0.01$, Here $k = 3$, $n = 5$

$\therefore \quad F_{tab} = F_{0.01,\ 2,\ 12} = 6.93$

3. Computations :

$$T_{1\bullet} = 77 + 81 + 71 + 76 + 80 = 385$$

$$T_{2\bullet} = 72 + 58 + 74 + 66 + 70 = 340$$

$$T_{3\bullet} = 76 + 85 + 82 + 80 + 77 = 400$$

$$T_{\bullet\bullet} = T_{1\bullet} + T_{2\bullet} + T_{3\bullet} = 1125$$

$$\Sigma\Sigma x_{ij}^2 = \quad 85041$$

$$\therefore SST = 85041 - \frac{1}{1.5}(1125)^2 = 666$$

$$SS(Tr) = \frac{1}{5}\left[(385)^2 + (340)^2 + (400)^2\right] - \frac{1}{1.5}(1125)^2$$

$$= 390$$

$$SSE = SST - SS(Tr) = 666 - 390 = 276$$

NOTES

ANOVA TABLE

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean square*	*F*
Treatments	2	390	$\frac{390}{2} = 195$	$\frac{195}{23} = 8.48$
Error	12	276	$\frac{276}{12} = 23$	
Total	14	666		

4. Conclusion :

Since $F_{cal} > F_{tab} \Rightarrow H_0$ is rejected.

$\Rightarrow$ Students of different schools of class X are not same.

Example 2. *The following are the number of typing mistakes made in five successive weeks by four typists working for a publishing company.*

Typist I	13	16	12	14	15
Typist II	14	16	11	19	15
Typist III	13	18	16	14	18
Typist IV	18	10	14	15	12

Test at the 0.05 level of significance whether the differences among the four sample means can be attributed to chance.

Solution. 1. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : At least two of them are not equal.

2. Here $k = 4$, $n = 5$, $\alpha = 0.05$

$$F_{0\cdot05,\ 3,\ 16} = 3.24$$

3. Computations :

$$T_{1\bullet} = 13 + 16 + 12 + 14 + 15 = 70$$

$$T_{2\bullet} = 14 + 16 + 11 + 19 + 15 = 75$$

$$T_{3\bullet} = 13 + 18 + 16 + 14 + 18 = 79$$

$$T_{4\bullet} = 18 + 10 + 14 + 15 + 12 = 69$$

$$T_{\bullet\bullet} = T_{1\bullet} + T_{2\bullet} + T_{3\bullet} + T_{4\bullet} = 293$$

$$\Sigma\Sigma x_{ij}^2 = 4407$$

$$SST = 4407 - \frac{1}{20}(293)^2 = 114.55$$

$$SS(Tr) = \frac{1}{5}\left[(70)^2 + (75)^2 + (79)^2 + (69)^2\right] - \frac{1}{20}(293)^2 = 12.95$$

NOTES

SSE = SST – SS(Tr) = 101.6

ANOVA TABLE

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean square*	*F*
Treatments	3	12.95	4.32	0.68
Error	16	101.6	6.35	
Total	19	114.55		

4. Conclusions :

Since $F_{cal} < T_{tab}$, H_0 is accepted and we conclude that the typing mistakes can be attributed to chance.

Note. 1. Since variance is independent of change of origin, then the origin can be shifted to an arbitrary data. Final variance ratio is independent of change of scale. Hence change of origin and scale, if necessary, can be employed which help to perform simpler arithmetic.

2. Each observations x_{ij} can be decomposed as

$$x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)$$

Grand mean | Deviation due to treatment | Error

RANDOMIZED BLOCK DESIGN/TWO-WAY CLASSIFICATION (R.B.D.)

Let the measurements pertaining to n treatments distributed over m blocks be available.

Let y_{ij} denotes the observation of the i-th treatment over the j-th block.

We shall use the following layout :

Blocks

	B_1	B_2 ---------- B_j ---------- B_m	*Means*
Treatment 1	y_{11}	y_{12} --------- y_{1j} --------- y_{1m}	$\bar{y}_{1\bullet}$
Treatment 2	y_{21}	y_{22} --------- y_{2j} --------- y_{2m}	$\bar{y}_{2\bullet}$
Treatment i	y_{i1}	y_{i2} --------- y_{ij} --------- y_{im}	$\bar{y}_{i\bullet}$
Treatment n	y_{n1}	y_{n2} --------- y_{nj} --------- y_{nm}	$\bar{y}_{n\bullet}$
Means	$\bar{y}_{\bullet 1}$	$\bar{y}_{\bullet 2}$ --------- $\bar{y}_{\bullet j}$ --------- $\bar{y}_{\bullet m}$	

Consider the model

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, 2, ..., n$$
$$j = 1, 2, ..., m$$

where, μ = Grand mean

α_i = Effect of the i-th treatment with $\sum \alpha_i = 0$

β_j = Effect of the j-th block with $\Sigma\beta_j = 0$

e_{ij} = Random errors which are identically distributed as N (0, σ^2)

Hypothesis :

H_0 : $\alpha_1 = \alpha_2 = ... = \alpha_n = 0$

or H_0 : $\beta_1 = \beta_2 = ... = \beta_n = 0$

H_1 : At least one of the effects is not zero.

Thus we construct the following ANOVA TABLE :

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean square*	*F*
Treatments	$n - 1$	SS (T*r*)	MS (T*r*) = $\frac{SS\ (Tr)}{n-1}$	$F_{Tr} = \frac{MS\ (Tr)}{MSE}$
Blocks	$m - 1$	SS (B*l*)	MS (B*l*) = $\frac{SS\ (Bl)}{m-1}$	$F_{Bl} = \frac{MS\ (Bl)}{MSE}$
Error	$(n-1)\ (m-1)$	SSE	MSE = $\frac{SSE}{(n-1)(m-1)}$	
Total	$nm - 1$	SST		

where $SST = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}^2 - c$

$$SS\ (Tr) = \frac{\sum_i T_{i\bullet}^2}{m} - c$$

$$SS\ (Bl) = \frac{\sum_j T_{\bullet j}^2}{n} - c$$

$$c = \frac{T_{\bullet\bullet}^2}{nm}$$

$T_{i\bullet}$ = Sum of m observations for the i-th treatment

$T_{\bullet j}$ = Sum of n observations for the j-th block

$T_{\bullet\bullet}$ = Grand total of all the observations

and SSE = SST – SS (T*r*) – SS (B*l*)

Set up α

Conclusions:

Reject $H_0 : \alpha_i = 0\ \forall_i$ if $F_{Tr} > F_{\alpha,\ n-1,\ (n-1)(m-1)}$

Reject $H_0 : \beta_j = 0\ \forall_j$ if $F_{Bl} > F_{\alpha,\ m-1,\ (n-1)(m-1)}$

NOTES

NOTES

Example 3. *Three different types of I.Q. test were conducted to four students and the following are the scores which they detained.*

	Students			
	1	2	3	4
Test-I	75	73	59	68
Test-II	83	72	56	69
Test-III	86	61	53	70

Perform a two way analysis of variance to test at the level of significance α = 0.01 whether it is reasonable to treat the three tests as equivalent.

Solution. 1. H_0 : Three tests are equivalent ($\alpha_1 = \alpha_2 = \alpha_3 = 0$)

H_1 : At least one α_i not zero.

Also $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (Student's I.Q. are equivalent), H_1 : At least one $\beta_i \neq 0$

2. $\alpha = 0.01$, Here $n = 3$, $m = 4$

F_{tab} (Tr) = $F_{0.01;\ 2,\ 6} = 10.92$

F_{tab} (Bl) = $F_{0.01,\ 3,\ 6} = 9.78$

3. Computations :

$$T_{1\bullet} = 75 + 73 + 59 + 68 = 275$$

$$T_{2\bullet} = 83 + 72 + 56 + 69 = 280$$

$$T_{3\bullet} = 86 + 61 + 53 + 70 = 270$$

$$T_{\bullet 1} = 75 + 83 + 86 = 244$$

$$T_{\bullet 2} = 73 + 72 + 61 = 206$$

$$T_{\bullet 3} = 59 + 56 + 53 = 168$$

$$T_{\bullet 4} = 68 + 69 + 70 = 207$$

$$T_{\bullet\bullet} = 825$$

$$c = \frac{(825)^2}{12} = 56718.75$$

$$\sum\sum y_{ij}^2 = 57855$$

$$T_{1\bullet}^2 + T_{2\bullet}^2 + T_{3\bullet}^2 = 226925$$

$$T_{\bullet 1}^2 + T_{\bullet 2}^2 + T_{\bullet 3}^2 + T_{\bullet 4}^2 = 173045$$

$$SST = 57855 - 56718.75 = 1136.25$$

$$SS\ (Tr) = \frac{226925}{4} - 56718.75 = 12.5$$

$$SS\ (Bl) = \frac{173045}{3} - 56718.75 = 962.92$$

$$SSE = SST - SS\ (Tr) - SS\ (Bl) = 160.83$$

NOTES

ANOVA TABLE

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean square*	F_{cal}
Treatments	2	12.5	$\frac{12.5}{2} = 6.25$	$\frac{6.25}{26.81} = 0.23$
Blocks	3	962.92	$\frac{962.92}{3} = 320.97$	$\frac{320.97}{26.81} = 11.97$
Error	6	160.83	$\frac{160.83}{6} = 26.81$	
Total	11	1136.25		

4. Conclusions.

Since $F_{cal}\ (Tr) < F_{tab}\ (Tr)$

$\Rightarrow$ $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ is accepted.

$\Rightarrow$ All three tests are equivalent.

But since $F_{cal}\ (Bl) > F_{tab}\ (Bl)$

$\Rightarrow$ $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is rejected.

$\Rightarrow$ I.Q. of the students are not same.

Note. The R.B.D is more accurate than C.R.D. for most types of experimental work. It has greater flexibility, *i.e.*, no restrictions are placed on the number of treatments or the number of replicates. However, R.B.D is not suitable to the problems with large number of treatments or to the wide variable blocks.

SUMMARY

- Analysis of variance (ANOVA) implies a statistical technique on the variations to a group of causes. This technique is closely related to *design of experiment* which may be defined as "the logical construction of the experiment in which the degree of uncertainty with which the inference is drawn may be well defined."
- In a comparative experiment various objects of comparisons are termed as *treatments.* Also the whole experimental unit is divided into subgroups, preferably homogeneous are called *blocks.*

PROBLEMS

1. Three samples below have been obtained from normal populations with equal variances. Test the hypothesis at 5% level of significance that the population means are equal.

NOTES

I	II	III
8	7	12
10	5	9
7	10	13
14	9	12
11	9	14

2. There are three main brands of a certain powder. A set of its 120 sales is examined and found to be allocated among four groups (A, B, C, D) and brands (I, II and III) as shown below :

		Groups		
Brands	A	B	C	D
I	0	4	8	15
II	5	8	13	6
III	18	19	11	13

Is there any significant difference in brands preference ? Answer at 5 per cent level using one way ANOVA.

3. The following table shows the lives (in '000 hrs) of four batches of electric bulbs:

Batches								
I	1.6	1.61	1.65	1.68	1.7	1.72	1.8	
II	1.58	1.64	1.64	1.7	1.75			
III	1.46	1.55	1.6	1.62	1.64	1.66	1.74	1.82
IV	1.51	1.52	1.53	1.57	1.6	1.68		

Perform an analysis of variance of these data and show that a significance test does not reject their homogeneity.

4. Five tests were conducted in Computer Science papers to candidates in three batches selected based on their strengths in Mathematics, English and Numerical computing respectively. The following average test scores were obtained :

Batch–1	Batch–2	Batch–3
86	79	88
77	78	83
81	82	86
84	83	89
69	68	82

Carry out an analysis of variance. Test the hypothesis whether the differences among the means obtained for the three batches are significant at 5% level of significance.

5. Set up a two-way ANOVA table for the data given below :

Prices of Field	*Treatment*			
	A	*B*	*C*	*D*
P	45	40	38	37
Q	43	41	45	38
R	39	39	41	41

(You can shift the origin to 40)

6. In a certain factory, production can be accomplished by four different workers on five different types of machines. A sample study in context of two way design without repeated values, is being made with two fold objectives of examining whether the four workers differ with respect to mean productivity and whether the mean productivity is the same for the five machines. The researcher involved in this study reports while analysing the gathered data as under :

(*i*) Sum of squares for variance between machines = 35.2

(*ii*) Sum of squares for variance between workmen = 53.8

(*iii*) Sum of squares for total variance = 174.2

ANSWERS

1. Reject $H_0 : \mu_1 = \mu_2 = \mu_3$, $F_{cal} = 4$
2. No significant difference in brands
4. Reject H_0, means are different
5.

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean square*	*F*
Between treatments	3	22.917	7.639	1.0
Between fields	2	8.167	4.083	0.534
Error	6	45.833	7.639	
Total	11	76.917		

6. Workers do not differ w.r.t. mean productivity and mean productivity is the same for 5 different machines.

F (machines) = 1.24, F (Workers) = 2.53.

CHAPTER 13 STOCHASTIC PROCESS

NOTES

★ STRUCTURE ★

- Definition
- Markov Process and Markov Chain
- Chapman-kolmogorov Equations
- Classification of States
- Steady-state Probabilities
- First Entrance Probability
- Summary
- Problems

DEFINITION

A stochastic process is defined as an indexed collection of random variables $\{X_t\}$, parameterized on time t, which are defined on a common sample space.

It is to be noted that the distribution of the random variable X_t may not be the same at different points in time t_1 and t_2. If they are the same, we refer to the random variables as identically distributed. To specify a stochastic process completely, we must also distinguish when the samples of the random variable occur in time (the embedding points). The points in time may be equally spaced or their spacing may depend upon the overall behaviour of the physical system in which the stochastic process is embedded.

Example 1. *The Poisson process described in section 2 of Chapter 8 represents a stochastic process with an infinite number of states. Assuming the process starts at time (zero) 0, the random variable X represents the occurrences (arriving of customers) between 0 and t. The states of the system at any time t are thus given by x = 0, 1, 2, ...*

Example 2. *The successive roll of a pair of dice in a game is a stochastic process. Each roll has the same sample space, i.e., from 2 to 12 and each roll occurs at some point in time. Each of the random variables is identically distributed and independent.*

MARKOV PROCESS AND MARKOV CHAIN

NOTES

A stochastic process is said to be Markov process if it satisfies the Markovian property, *i.e.*,

$$P\{X_{t+1} = j \mid X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = k_{t-1}, X_t = i\} = P\{X_{t+1} = j \mid X_t = i\} \text{ for } t = 0, 1, \ldots$$

i.e. the occurrence of a future state depends on the immediately preceding state and only on it.

The conditional probabilities $P\{X_{t+1} = j \mid X_t = i\}$ are called one step transition probabilities as it describes the system between t and $t + 1$.

If for each i and j,

$$P\{X_{t+1} = j \mid X_t = i\} = P\{X_1 = j \mid X_0 = i\} \text{ for all } t = 0,$$

1, ...then the one step transition probabilities are said to be *stationary* and are usually denoted by p_{ij}. These stationary transition probabilities do not change in time.

If, for each i, j and n (= 0, 1, 2,....), we have

$P\{X_{t+n} = j \mid X_t = i\} = P\{X_n = j \mid X_0 = i\}$ for all $t = 0, 1, ..,$ then these conditional probabilities are called *n-step transitional probabilities* and are usually denoted by $p_{ij}^{(n)}$

These probabilities will satisfy the following properties :

(*a*) $p_{ij}^{(n)} \geq 0$ for all i and j and $n = 0, 1, 2, \ldots$

(*b*) $\sum_{j=0}^{M} p_{ij}^{(n)} = 1$ for all i and $n = 0, 1, 2, \ldots$, and M = No. of states.

In Matrix notation, we represent the n-step transition probabilities as

$$P^{(n)} = \begin{bmatrix} p_{00}^{(n)}, \ldots, p_{0M}^{(n)} \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ p_{M0}^{(n)}, \ldots, p_{MM}^{(n)} \end{bmatrix}, n = 0, 1, 2, \ldots$$

and for one step transition probabilities we can take

$$P = \begin{bmatrix} p_{00} & p_{01} \cdots p_{0M} \\ p_{10} & p_{11} \cdots p_{1M} \\ \cdots\cdots\cdots\cdots\cdots \\ p_{M0} & p_{M1} \cdots p_{MM} \end{bmatrix}$$

NOTES

and P is called one step transition matrix, whereas $P^{(n)}$, n step transition matrix.

A stochastic process $\{X_t\}$ $(t = 0,1, 2,...)$ is said to be a *finite state Markov chain* if it has the following properties :

(*a*) A finite number of states,

(*b*) The Markovian property,

(*c*) Stationary transition probabilities

(*d*) A set of initial probabilities $P\{X_0 = i\}$ for all i.

Example 3. *Consider the one step transition matrix with three states :*

$$P = \begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \end{array} \begin{array}{c} \begin{array}{ccc} E_1 & E_2 & E_3 \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{array}$$

A transition diagram is given below. The arrows from each state indicate the possible states to which a process can move from the given state.

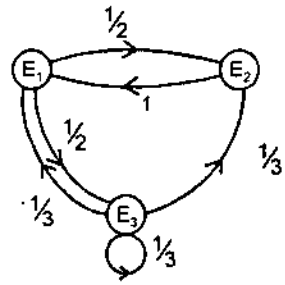


Fig. 13.1

CHAPMAN-KOLMOGOROV EQUATIONS

These are used to compute n-step transition probabilities as given by

$$p_{ij}^{(n)} = \sum_{k=0}^{M} p_{ik}^{(v)} \cdot p_{kj}^{(n-v)} \quad \text{for all } i, j, n \text{ and } 0 \le v \le n.$$

When $v = 1$, then

$$p_{ij}^{(n)} = \sum_{k=0}^{M} p_{ik}\, p_{kj}^{(n-1)} \quad \text{for all } i, j, n$$

When $v = n - 1$, then

$$p_{ij}^{(n)} = \sum_{k=0}^{M} p_{ik}^{(n-1)}\, p_{kj} \quad \text{for all } i, j, n$$

The above equations show that the n-step transition probabilities can be obtained from the one- step transition probabilities recursively.

Let $n = 2$, then we can write

$$p_{ij}^{(2)} = \sum_{k=0}^{M} p_{ik}\, p_{kj} \quad \text{for all } i, j.$$

$$= \text{elements of } P^{(2)}$$

NOTES

Which implies that $P^{(2)} = P.\ P = P^2$

In general,

$$P^{(n)} = P.\ P.\ \ldots\ .P = P^n = P.\ P^{n-1} = P^{n-1}.P$$

i.e., nth power of the one-step transition matrix gives the n–step transition matrix.

CLASSIFICATION OF STATES

A Markov chain is characterized by a set of states and transitions. We define some categories of states.

(*a*) A state is recurrent if the system, once in that state will return to that state through a series of transitions with probability 1.

(*b*) A state is transient if it is not recurrent.

(*c*) A recurrent state is recurrent non null if the mean time to return to the state is finite.

(*d*) A recurrent state is recurrent null if the mean time to return to the state is infinite.

Note. In (*b*) transient does not mean it cannot recur. *It only means that* the state is not guaranteed of recurring.

(*e*) A recurrent state is aperiodic if for some number k, there is a way to return to the state in k, $k + 1$, $k + 2$, ...∞ transitions.

(*f*) A recurrent state is periodic if it is not aperiodic.

(*g*) A Markov chain is irreducible if all states are reachable from all other states.

(*h*) A Markov chain is transient if all its states are transient.

It is recurrent non null if all its states are recurrent non null.

It is recurrent null if all its states are recurrent null.

It is periodic if all its states are periodic.

It is aperiodic if all its states are aperiodic.

(*i*) If the Markov chain is irreducible, recurrent non null and aperiodic, it is called ergodic.

An ergodic Markov chain has the property that it is possible to go from one state to any other state in a finite number of steps, regardless of the present state.

A regular chain is a special type of ergodic chain whose transition matrix P is such that for some power of P, it has only non-zero probability elements.

It follows that all regular chains must be ergodic but all ergodic chains may not be regular.

Note. To test an ergodic chain is regular, continue square the transition matrix P sequentially until all 0's disappear.

NOTES

Example 4. *Consider the discrete time Markov chain as follows :*

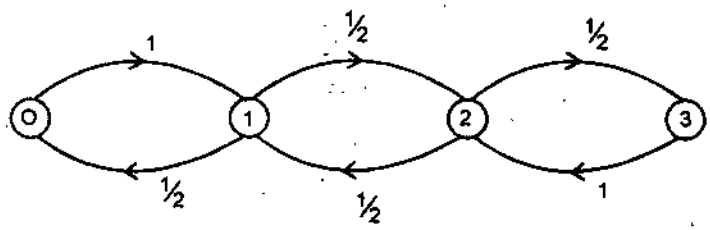


Fig.13.2

(*i*) *Is the state 0 recurrent or transient ?*

(*ii*) *Is the state 0 recurrent null or recurrent non null ?*

(*iii*) *Is the state 0 periodic or aperiodic ?*

Solution.(*i*) It is a recurrent state. There is nowhere to go from state 0 where it cannot return.

(*ii*) It is a recurrent non null simply because there are only a finite number of states *i.e.,* four.

(*iii*) The returning to state 0 in an odd number of steps is not possible. Therefore, there is no number k such that the system can return in $k + 1$ and $k + 2$ steps, since either $k + 1$ or $k + 2$ must be odd. Therefore, the state 0 is periodic with period 2.

Example 5. *Test whether the following Markov chain is ergodic and regular.*

From

$$T_0 \begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \end{array} \begin{array}{c} \begin{array}{ccc} E_1 & E_2 & E_3 \end{array} \\ \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \end{array}$$

Solution. Consider the state diagram.

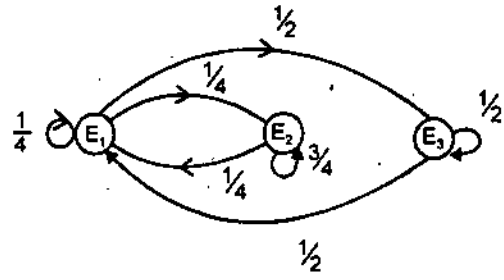


Fig. 13.3

It is possible to go from state E_1 to E_2 or E_3, from state E_2 to E_1 or E_3 from state E_3 to E_1 or E_2. Hence it is ergodic.

Let
$$P = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

Then
$$P^2 = P.P = \begin{bmatrix} \frac{3}{8} & \frac{1}{4} & \frac{3}{8} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{8} \\ \frac{3}{8} & \frac{1}{8} & \frac{1}{2} \end{bmatrix}$$

Since all zeroes disappear, it is regular.

STEADY-STATE PROBABILITIES

There is a long run behaviour of Markov process. For an irreducible ergodic Markov chain it can be shown that

$$\lim_{n \to \infty} p_{ij}^{(n)} = \Pi_j \text{ (i.e., independent of } i)$$

where Π_j satisfies the following steady-state equations :

$$\Pi_j > 0$$

$$\Pi_j = \sum_{i=0}^{M} \Pi_i p_{ij} \text{ for } j = 0, 1, ..., M$$

and
$$\sum_{j=0}^{M} \Pi_j = 1$$

Here Π_j's are called the steady state probabilities of the Markov chain because the probability of finding the process in a certain state, say j, after a large number of transitions tends to the value Π_j, independent of the initial probability distribution defined over the states.

Also we have

$$\Pi_i = \frac{1}{\mu_{jj}}, \text{ for } j = 0, 1, ..., M$$

where μ_{jj} is the expected recurrence time.

NOTES

Example 6.*Find the mean recurrence time for each state of the following Markov chain :*

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.4 & 0.4 \\ 0.1 & 0.5 & 0.4 \end{bmatrix}$$

Solution. We have the steady-state equations

$$\Pi_j = \sum_{i=0}^{2} \Pi_i \, p_{ij}, \; j = 0, 1, 2 \quad \text{and} \quad \sum_{j=0}^{2} \Pi_j = 1$$

$$\Pi_0 = \Pi_0 (0.5) + \Pi_1 (0.2) + \Pi_2 (0.1)$$

$$\Rightarrow \quad \Pi_1 = \Pi_0 (0.3) + \Pi_1 (0.4) + \Pi_2 (0.5)$$

$$\Pi_2 = \Pi_0 (0.2) + \Pi_1 (0.4) + \Pi_2 (0.4)$$

and $$\Pi_0 + \Pi_1 + \Pi_2 = 1$$

$$-0.5\Pi_0 + 0.2\Pi_1 + 0.1\Pi_2 = 0$$

$$\Rightarrow \quad 0.3\Pi_0 - 0.6\Pi_1 + 0.5\Pi_2 = 0$$

$$0.2\Pi_0 + 0.4\Pi_1 - 0.6\Pi_2 = 0$$

and $$\Pi_0 + \Pi_1 + \Pi_2 = 1$$

Solving these equations we obtain

$$\Pi_0 = 0.2353$$

$$\Pi_1 = 0.4118$$

$$\Pi_2 = 0.3529$$

Hence the mean recurrence time for each state is given by

$$\mu_{00} = \frac{1}{\Pi_0} = 4.2499$$

$$\mu_{11} = \frac{1}{\Pi_1} = 2.4284$$

$$\mu_{22} = \frac{1}{\Pi_2} = 2.8337$$

FIRST ENTRANCE PROBABILITY

Let $f_{ij}^{(n)}$ = Probability of arriving at j at time n for the first time, given that the process starts at i

$$= P\,[X_n = j,\, X_{n-1} \neq j,\, X_{n-2} \neq j,\, \ldots X_1 \neq j \mid X_0 = i]$$

Let $T_{ij} = \text{Min}\,\{n : X_n = j \mid X_0 = i\}$. Then

$$f_{ij}^{(n)} = P\left[T_{ij} = n\right] \text{ with } f_{ij}^{(0)} = 0$$

and $f_{ij}^{(1)} = p_{ij}$ (gives the diagonal of the transition matrix)

We can prove the following result :

$$p_{ij}^{(n)} = \sum_{m=1}^{n} p_{ij}^{(m)} f_{jj}^{(n-m)} \text{ for all } m = 1, 2, ..., n.$$

NOTES

A state i of a Markov chain is said to be transient if $f_{ii} < 1$ and recurrent if $f_{ii} = 1$.

Also, the mean recurrence time for $i = \sum_{n=1}^{\infty} n.f_{ii}^{(n)}$.

SUMMARY

- A stochastic process is defined as an indexed collection of random variables $\{X_t\}$, parameterized on time t, which are defined on a common sample space.
- A stochastic process $\{X_t\}$ $(t = 0,1; 2,...)$ is said to be a *finite state Markov chain* if it has the following properties :
  (*a*) A finite number of states,
  (*b*) The Markovian property,
  (*c*) Stationary transition probabilities
  (*d*) A set of initial probabilities P $\{X_0 = i\}$ for all i.

PROBLEMS

1. Test whether the following Markov chains are periodic or aperiodic.

(*a*)

From \ To	E_0	E_1	E_2	E_3
E_0	0	1	0	0
E_1	0	0	0	1
E_2	½	0	½	0
E_3	0	0	1	0

(*b*)

From \ To	E_0	E_1	E_2
E_0	0	0	1
E_1	$\frac{1}{2}$	0	$\frac{1}{2}$
E_2	0	1	0

NOTES

2. Test whether the Markov chain having the following transition matrix is regular and ergodic.

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \end{array}$$

3. Consider the three state Markov chain with transition probability matrix

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix} \end{array}$$

Prove that the chain is irreducible.

4. Find the mean recurrence time for each state of the following Markov chain

$$P = \begin{bmatrix} 0.33 & 0.33 & 0.34 \\ 0.25 & 0.25 & 0.5 \\ 0.2 & 0.6 & 0.2 \end{bmatrix}$$

ANSWERS

1. (*a*) Periodic, (*b*) Aperiodic

2. Ergodic but not regular

4. $\mu_{00} = 3.9604$, $\mu_{11} = 2.5381$, $\mu_{22} = 2.8289$.

APPENDIX-A1

NOTES

SOME STANDARD RESULTS

1. *Permutation* without repetition (n distinct objects taken r at a time)

$$^{n}P_{r} = n(n-1)(n-2)\ldots(n-r+1)$$

Permutation with repetition-arranging n objects among which p are alike, q are alike, r are alike etc. is

$$\frac{n!}{p!\,q!\,r!}.$$

Permutation of n distinct objects in a line is $n!$, in a circle is $n-1!$

2. *Combination.* Selecting r objects out of n distinct objects is given by

$$\binom{n}{r} = \frac{n!}{r!\,n-r!}$$

3. *Sum of Points on the dice.* When n dice are thrown, the number of ways of getting a total of r points is given by coefficient of x^r in $(x + x^2 + x^3 + x^4 + x^5 + x^6)^n$.

4. Gama function is defined by

$$\Gamma n = \int_0^\infty e^{-x} \cdot x^{n-1}\,dx, \quad n > 0$$

with properties

(*i*) $\Gamma n+1 = n\Gamma n$, (*ii*) $\Gamma n+1 = n!$, (*iii*) $\Gamma 1/2 = \sqrt{\Pi}$

5. Beta function is defined by

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

with properties

(*i*) $B(m, n) = B(n, m)$, (*ii*) $B(m, n) = \dfrac{\Gamma m\,\Gamma n}{\Gamma m+n}$, $m, n > 0$

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

(*iv*) $B(m, n) = B(m+1, n) + B(m, n+1)$.

NOTES

6. A box contains x white and y black balls. If $a + b$ balls are drawn at random *without replacement* the probability that among them exactly a are white and b are black is given by

$$\frac{\binom{x}{a}\binom{y}{b}}{\binom{x+y}{a+b}}$$

7. Probability of repeated trials, *i.e.*, drawing with replacement is given by

$$\binom{n}{r} p^r q^{n-r}$$

where n = Repeated trials of the experiment
r = No. of events occur
p = Probability of occurrence of an event
$q = 1 - p$

8. Algebra of Sets :

Commutative Laws : $A \cup B = B \cup A,\ A \cap B = B \cap A$

Associative Law : $(A \cup B) \cup C = A \cup (B \cup C)$

$(A \cap B) \cap C = A \cap (B \cap C)$

Distributive Law : $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Complementary Law : $A \cup \overline{A} = S,\ A \cap \overline{A} = \phi$

Difference Law : $A - B = A \cap \overline{B}$

$A - B = A - (A \cap B) = (A \cup B) - B$

$A - (B - C) = (A - B) \cup (A - C)$

$(A \cup B) - C = (A - C) \cup (B - C)$

$A - (B \cup C) = (A - B) \cup (A - C)$

De Morgan's Law : $\overline{A \cup B} = \overline{A} \cap \overline{B},\ \overline{A \cap B} = \overline{A} \cup \overline{B}.$

APPENDIX-A2

MARGINAL AND CONDITIONAL DISTRIBUTIONS

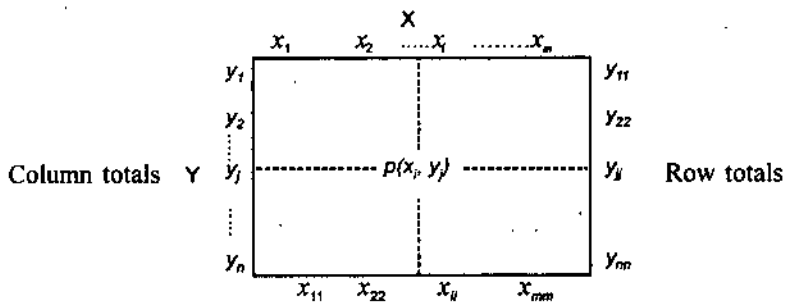
NOTES

1. Marginal Distributions

Let the joint probability distribution of two random variables X and Y be denoted as $p\ (x, y)$ or $f\ (x, y)$

where $\sum_x \sum_y p\ (x, y) = 1$ and $\int_{-\infty}^{\infty} \cdot \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$

(a) Discrete Case



Then marginal distribution of X is defined by

$$g\ (x): \quad \begin{array}{c|cccc} x & x_1 & x_2 & \cdots & x_m \\ \hline p(x) & x_{11} & x_{22} & \cdots & x_{mm} \end{array}$$

and marginal distribution of Y is defined by

$$h\ (y): \quad \begin{array}{c|cccc} y & y_1 & y_2 & \cdots & y_n \\ \hline p(y) & y_{11} & y_{22} & \cdots & y_{nn} \end{array}$$

(b) Continuous Case

Here both the random variables X and Y are continuous

Then $g\ (x)$ = Marginal distribution of X

$$= \int_{-\infty}^{\infty} f(x,y)\, dy \quad \text{for} \quad -\infty < x < \infty$$

$h\ (y)$ = Marginal distribution of Y

$$= \int_{-\infty}^{\infty} f(x,y)\, dx \quad \text{for} \quad -\infty < y < \infty$$

Note

1. Expectation and variance of the marginal distributions can also be calculated.
2. If $p\ (x, y) = g\ (x).h\ (y)$ or $f\ (x, y) = g\ (x)\ h\ (y)$, then the two random variables are said to be independent.

NOTES

2. Conditional Distributions

(a) Discrete Case. Conditional distribution of X given that Y = y is given by

$$f(x|y) = \frac{p(x, y)}{h(y)}, \; h(y) \neq 0 \quad \text{for each } x \text{ within the range of X.}$$

Similarly, conditional distribution of Y given that X = x is

$$w(y|x) = \frac{p(x, y)}{g(x)}, \; g(x) \neq 0, \quad \text{for each } y \text{ within the range of Y.}$$

(b) Continuous Case.

$$f(x|y) = \frac{f(x, y)}{h(y)}, \quad h(y) \neq 0, \quad -\infty < x < \infty$$

and

$$w(y|x) = \frac{f(x, y)}{g(x)}, \quad g(x) \neq 0, \quad -\infty < y < \infty$$

Example 1. *Consider the following joint density function of two random variables.*

		x		
		0	1	2
	0	1/12	1/3	1/6
y	1	0	1/6	2/9
	2	0	0	1/36

(i) Find the marginal distributions of X and Y.

(ii) Find the conditional distribution of X given Y = 1.

(iii) Find the conditional distribution of Y given X = 2.

Solution.

		x			
		0	1	2	
	0	1/12	1/3	1/6	7/12
y	1	0	1/6	2/9	7/18
	2	0	0	1/36	1/36
		1/12	1/2	5/12	1

(i) Marginal distribution of X is given by

x	0	1	2
$g(x)$: $p(x)$	1/12	1/2	5/12

Marginal distribution of Y is given by

$$g(x): \begin{array}{c|ccc} y & 0 & 1 & 2 \\ \hline p(y) & 7/12 & 7/18 & 1/36 \end{array}$$

(*ii*) Given $Y = 1$, then $h(1) = 7/18$

For $x = 0$, $f(0|1) = \dfrac{f(0,1)}{h(1)} = \dfrac{0}{7/18} = 0$

For $x = 1$, $f(1|1) = \dfrac{f(1,1)}{h(1)} = \dfrac{1/6}{7/18} = \dfrac{3}{7}$

For $x = 2$, $f(2|1) = \dfrac{f(2,1)}{h(1)} = \dfrac{2/9}{7/18} = \dfrac{4}{7}$

The conditional distribution of X given Y = 1 is

$$f(x|1): \begin{array}{c|ccc} x & 0 & 1 & 2 \\ \hline p(x) & 0 & 3/7 & 4/7 \end{array}$$

(*iii*) Similarly, the conditional distribution of Y given X = 2 is

$$w(y|2): \begin{array}{c|ccc} y & 0 & 1 & 2 \\ \hline p(y) & 2/5 & 8/15 & 1/15 \end{array}$$

Example 2. *Given the joint density*

$$f(x, y) = x + y, \quad 0 < x < 1, \quad 0 < y < 1$$
$$= 0, \quad \text{elsewhere}$$

(i) Find the marginal distribution of X and Y.

(ii) Find the conditional distribution of X given Y = y.

(iii) Find the conditional distribution of Y given X = x.

Solution. (*i*) Marginal distribution of X is

$$g(x) = \int_0^1 (x + y)\, dy = \left[xy + \frac{y^2}{2}\right]_0^1 = x + \frac{1}{2}, \quad 0 < x < 1$$

Marginal distribution of Y is

$$h(y) = \int_0^1 (x + y)\, dx = \left[\frac{x^2}{2} + xy\right]_0^1 = \frac{1}{2} + y, \quad 0 < y < 1$$

(*ii*) Conditional distribution of X given Y = *y*

$$f(x|y) = \frac{f(x, y)}{h(y)} = \frac{x + y}{\frac{1}{2} + y} = \frac{2(x + y)}{1 + 2y}, \quad 0 < x < 1$$

NOTES

NOTES

(*iii*) Conditional distribution of Y given X = x

$$w(y|x) = \frac{f(x,y)}{g(x)} = \frac{x+y}{x+\frac{1}{2}} = \frac{2(x+y)}{2x+1}, \quad 0<y<1.$$

PROBLEMS

1. Consider the following joint density function

Y \ X	−1	0	1
0	1/15	1/15	2/15
1	2/15	2/15	2/15
2	3/15	1/15	1/15

(*a*) Find the marginal distributions of X and Y.

(*b*) Find the conditional distribution of X given Y = 2.

(*c*) Find the conditional distribution of Y given X = 1.

2. The joint *pdf* of a two random variables X and Y is given by

$$f(x, y) = 2, \qquad 0 < x < 1, \quad 0 < y < x$$
$$= 0, \qquad \text{elsewhere}$$

(*a*) Find the marginal distributions of X and Y.

(*b*) Find the conditional density of Y given X = x.

(*c*) Find the conditional density of X given Y = y.

(*d*) Can you say that X and Y are independent ?

3. Given the joint *pdf* of two random variables as

$$f(x, y) = \frac{-(x+y-6)}{8} \quad 0<x<2, \ 2<y<4$$

Find the conditional density of Y given X = x and conditional density of X given Y = y.

4. Consider the joint *pmf* of two random variables as

X_2 \ X_1	0	1	2
−1	0.1	0.2	0.1
0	0.1	0.1	0.1
1	0.1	0.1	0.1

Find (*a*) $P[X_1 + X_2 \geq 2]$, (*b*) $P[X_1|X_2 = 1]$

ANSWERS

1. (*a*)

x	−1	0	1
$g(x)$	6/15	4/15	5/15

y	0	1	2
$h(y)$	4/15	6/15	5/15

(b)

x	−1	0	1
$p(x/y=2)$	3/5	1/5	1/5

(c)

y	0	1	2
$p(y/x=1)$	2/5	2/5	1/5

2. (a) $g(x) = 2x \quad , 0 < x < 1$

$= 0 \quad$, elsewhere

and $h(y) = 2(1-y) \quad , 0 < y < 1$

$= 0 \quad$, elsewhere

(b) $w\ (y/x) = \dfrac{1}{x}, \quad 0 < x < 1$

(c) $f\ (x/y) = \dfrac{1}{1-y} \qquad 0 < y < 1$

(d) Not independent.

3. $w\ (y/x) = \dfrac{x + y - 6}{2x - 6}, \quad 2 < y < 4$

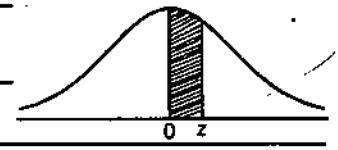
$f(x/y) = \dfrac{x + y - 6}{2y - 10}, \quad 0 < x < 2.$

4. (a) 0.3 (b)

X_1	0	1	2
$p(X_1 \mid X_2 = 1)$	1/3	1/3	1/3

NOTES

APPENDIX-A3



STATISTICAL TABLES

Table I: Area under the Normal curve from 0 to z = Φ (z)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993

Table II : Values of t_α

NOTES

υ	α = 0.10	α = 0.05	α = 0.025	α = 0.01	α = 0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756

NOTES

Table III : Values of χ^2 with level of significance α and degrees of freedom ν

ν \ α	0.99	0.95	0.50	0.30	0.20	0.10	0.05	0.01
1	0.0002	0.004	0.46	1.07	1.64	2.71	3.84	6.64
2	0.020	0.103	1.39	2.41	3.22	4.60	5.99	9.21
3	0.115	0.35	2.37	3.66	4.64	6.25	7.82	11.34
4	0.30	0.71	3.36	4.88	5.99	7.78	9.49	13.28
5	0.55	1.14	4.35	6.06	7.29	9.24	11.07	15.09
6	0.87	1.64	5.35	7.23	8.56	10.64	12.59	16.81
7	1.24	2.17	6.35	8.38	9.80	12.02	14.07	18.48
8	1.65	2.73	7.34	9.52	11.03	13.36	15.51	20.09
9	2.09	3.32	8.34	10.66	12.24	14.68	16.92	21.67
10	2.56	3.94	9.34	11.78	13.44	15.99	18.31	23.21
11	3.05	4.58	10.34	12.90	14.63	17.28	19.68	24.72
12	3.57	5.23	11.34	14.01	15.81	18.55	21.03	26.22
13	4.11	5.89	12.34	15.12	16.98	19.81	22.36	27.69
14	4.66	6.57	13.34	16.22	18.15	21.06	23.68	29.14
15	5.23	7.26	14.34	17.32	19.31	22.31	25.00	30.58
16	5.81	7.96	15.34	18.42	20.46	23.54	26.30	32.00
17	6.41	8.67	16.34	19.51	21.62	24.77	27.59	33.41
18	7.02	9.39	17.34	20.60	22.76	25.99	28.87	34.80
19	7.63	10.12	18.34	21.69	23.90	27.20	30.14	36.19
20	8.26	10.85	19.34	22.78	25.04	28.41	31.41	37.57
21	8.90	11.59	20.34	23.86	26.17	29.62	32.67	38.93
22	9.54	12.34	21.34	24.94	27.30	30.81	33.92	40.29
23	10.20	13.09	22.34	26.02	28.43	32.01	35.01	41.64
24	10.86	13.85	23.34	27.10	29.55	33.20	36.42	42.98
25	11.52	14.61	24.34	28.17	30.68	34.68	37.65	44.31
26	12.20	15.38	25.34	29.25	31.80	35.56	38.88	45.64
27	12.88	16.15	26.34	30.32	32.91	36.74	40.11	46.96
28	13.56	16.93	27.34	31.39	34.03	37.92	41.34	48.28
29	14.26	17.71	28.34	32.46	35.14	39.09	42.56	49.59
30	14.95	18.49	29.34	33.53	36.25	40.26	43.77	50.89

Table IV : Values of $F_{0.05}$

ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																
	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.57
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.43
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.74
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.30
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.01
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.79
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.62
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.49
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.38	2.38
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.30
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.22
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.16
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.11
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.06
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.02
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.95
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.89
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.86
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.74
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.64
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.53

Table V : Values of $F_{0.01}$

ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																
	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60
1	4,052	5,000	5,403	5,625	5,764	5.859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,240	6,261	6,287	6,313
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.57	99.47	99.48
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.32
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.65
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.20
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.06
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.82
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.03
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.48
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.08
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.78
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.54
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.34
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.18
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.05
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.93
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.83
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.75
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.67
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.61
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.55
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.50
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.45
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.40
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.36
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.21
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.02
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.84

Table VI : Random Digits

5294	6695	7471	9235	7132	2330
4339	0587	2009	2353	3545	1175
9821	6851	0854	6413	4368	7996
0393	7985	1709	5643	4697	3800
1933	2685	0093	6201	2533	6148
6780	7309	8118	2292	1642	7949
8016	5775	6688	0102	9684	6589
1129	8237	1451	1006	9988	1821
5635	8142	1143	3833	8731	2938
8527	7400	6440	5748	2444	1093
0652	1488	3884	8103	1157	3980
2825	5571	2717	2184	7843	6814
4398	3132	8710	1246	8319	2208
5395	9559	2227	2701	5367	4534
2920	4973	6841	1884	2137	1207
0111	8260	9023	1368	6122	3001
0272	4702	8030	9239	7092	2201
8414	5198	7896	7026	1111	1331
9005	8021	0205	6855	7342	1548
4600	0560	8892	6515	3237	7916
7631	7361	9031	9749	7000	6032
4114	2228	4595	0277	7193	6515
2478	8899	1901	2176	4140	4482
8522	9041	4748	5044	3897	5606
7755	2031	1191	7745	0124	6341
0456	9977	6923	2539	6678	9906
7196	7275	4971	0110	0220	6817
8252	8006	3957	7149	3576	6591
3026	0014	9368	7262	8134	4141
8245	4972	9200	8898	5225	6855

NOTES