To the Instructor

Considering the following **revised GTU syllabus (2019)** carefully and minutely, I have prepared 7 chapters based on the understanding of the level of Students.

Teaching Scheme			Credits	Examination Marks				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE (E)	PA (M)	ESE (V)	PA (I)	
3	2	0	5	70	30	0	0	100

Sr. No.	Content	Total Hrs.	% Weightage	Content covered in Chapters
1.	**Basic Probability:** Experiment, definition of probability, conditional probability, independent events, Bayes' rule, Bernoulli trials, Random variables, discrete random variable, probability mass function, continuous random variable, probability density function, cumulative distribution function, properties of cumulative distribution function, Two dimensional random variables and their distribution functions, Marginal probability function, Independent random variables.	8	20	4, 5
2.	**Some special Probability Distributions:** Binomial distribution, Poisson distribution, Poisson approximation to the binomial distribution, Normal, Exponential and Gamma densities, Evaluation of statistical parameters for these distributions.	10	25	6
3.	**Basic Statistics:** Measure of central tendency: Moments, Expectation, dispersion, skewness, kurtosis, expected value of two dimensional random variable, Linear Correlation, correlation coefficient, rank correlation coefficient, Regression, Bounds on probability, Chebyshev's Inequality	10	20	1, 3, 6
4.	**Applied Statistics:** Formation of Hypothesis, Test of significance: Large sample test for single proportion, Difference of proportions, Single mean, Difference of means, and Difference of standard deviations. Test of significance for Small samples: t- Test for single mean, difference of means, t-test for correlation coefficients, F- test for ratio of variances, Chi-square test for goodness of fit and independence of attributes.	10	25	7
5.	**Curve fitting by the numerical method:** Curve fitting by of method of least squares, fitting of straight lines, second degree parabola and more general curves.	4	10	2

Contents

Chapter 1

Measures of Central Tendency & Dispersion

Before taking introduction of measures of central tendency and dispersion, let us first review following definitions which are useful in the subsequent work.

(1) **Variable (or variate)** It is that quantity which varies from individual to individual.

→ ***For example,*** heights, weights, etc.

(a) **Continuous variable** It is a quantity which can take any numeric value within a certain range.

→ ***For example,*** height, weights, etc.

(b) **Discrete (or discontinuous) variable** It is a quantity which is not capable of taking all possible values within the given range.

→ ***For example,*** number of students in a class.

(2) **Types of series**

(a) **Individual observations** These are observations where frequencies are not given.

→ ***For example,***

$$x_1, x_2, x_3, \ldots, x_n$$

(b) **Discrete series** It is a series of observations of the following form.

x :	x_1	x_2	x_3	...	x_n
f :	f_1	f_2	f_3	...	f_n

(c) **Continuous series** It is a series of observations of the following form.

Class-interval :	$c_1 - c_2$	$c_2 - c_3$	...	$c_n - c_{n+1}$
f :	f_1	f_2	...	f_n

1.1 Introduction to Measures of Central Tendency

The measures of central tendency is called so because it represents, in general, ***the central part*** of the distribution. Different types of commonly used measures of central tendency (or averages) are represented in Figure 1.1.

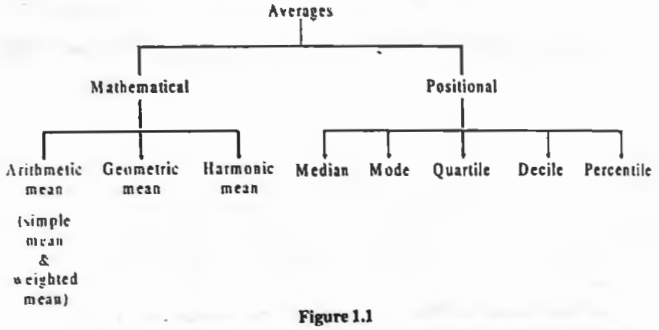


Figure 1.1

Before proceeding further, it should be noted that ***mean*** represents a single number to characterize set of numbers. It is useful in making comparisons.

1.1.1 Simple Arithmetic Mean (A.M.)

➢ **Mean of Raw Data or Ungrouped Data** If $\{x_1, x_2, x_3, \ldots, x_n\}$ is a set of n-numbers (observations or items or values), then the mean of these numbers is denoted and defined as follows.

$$\bar{x} = \frac{\text{sum of all the numbers}}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

or

$$\bar{x} = \frac{\Sigma x}{n}, \qquad \ldots(1.1)$$

Example 1.1

Find the mean of $-1.5, 0, 1, 0.8$.

Solution

Here, $n = 4$. Let $x_1 = -1.5, x_2 = 0, x_3 = 1, x_4 = 0.8$.

Therefore, the mean of the given data using (1.1) is

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i}{4} = \frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{-1.5 + 0 + 1 + 0.8}{4} = 0.075. \quad \textit{Answer}$$

➢ **Mean of Grouped Data** If the values $x_1, x_2, x_3, \ldots, x_n$ occur with corresponding frequencies $f_1, f_2, f_3, \ldots, f_n$, then the mean of these numbers is denoted and defined as follows.

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_n f_n}{f_1 + f_2 + f_3 + \ldots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

or

$$\bar{x} = \frac{\Sigma xf}{\Sigma f}. \qquad \ldots(1.2)$$

Example 1.2

Find mean from the following data.

x_i	:	1	3	5	7	9
f_i	:	4	2	3	1	5

Solution

Here, $n = 5$. Let

$$x_1 = 1, x_2 = 3, x_3 = 5, x_4 = 7, x_5 = 9,$$

$$f_1 = 4, f_2 = 2, f_3 = 3, f_4 = 1, f_5 = 5.$$

Therefore,

$$\sum_{i=1}^{5} x_i f_i = x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4 + x_5 f_5$$
$$= (1)(4) + (3)(2) + (5)(3) + (7)(1) + (9)(5)$$
$$= 4 + 6 + 15 + 7 + 45$$
$$= 77.$$

$$\sum_{i=1}^{5} f_i = f_1 + f_2 + f_3 + f_4 + f_5$$
$$= 4 + 2 + 3 + 1 + 5$$
$$= 15.$$

Using (1.2),

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i f_i}{\sum_{i=1}^{5} f_i} = \frac{77}{15} \approx 5.13.$$ *Answer*

➢ **Short-cut Method for Computing A.M.** This method is applicable when the frequencies and the values of the variable are quite large in quantities and it become difficult to compute arithmetic mean because of their large multiplication. In this case a ***provisional mean*** or ***assumed mean*** is chosen as that value of x which corresponds to the highest frequency or which is located near the middle of the frequency distribution.

● **Method for Ungrouped Data** In this case, mean is defined as follows.

$$\bar{x} = a + \frac{\sum d}{n}, \quad ...(1.3)$$

where

a is assumed mean,

n is number of items,

$d = x - a$ is the deviation of each value from the assumed mean a.

● **Method for Grouped Data** In this case, mean is defined as follows.

$$\bar{x} = a + \frac{\sum fd}{N}, \quad ...(1.4)$$

where

a is assumed mean,

$d = x - a$ is the deviation of each value from the assumed mean a,

N is the sum of all the frequencies; that is, $N = \sum f$.

Example 1.3

Calculate the arithmetic mean by short-cut method for the following data.

x :	0	1	2	3	4	5	6	7	8	9	10
f :	2	8	43	133	207	260	213	120	54	9	1

Solution

Let assumed mean be $a = 5$.

x	f	$d = x - 5$	fd
0	2	−5	−10
1	8	−4	−32
2	43	−3	−129
3	133	−2	−266
4	207	−1	−207
5	260	0	0
6	213	1	213
7	120	2	240
8	54	3	162
9	9	4	36
10	1	5	5
$\sum$	1050		12

Using (1.4),

$$\bar{x} = a + \frac{\sum fd}{\sum f} = 5 + \frac{12}{1050} \approx 5 + 0.0114 = 5.0114.$$ *Answer*

When class-intervals are given, then mean can be calculated by following way.

Example 1.4

For the following frequency table, find mean by short-cut method.

Class :	100-120	120-140	140-160	160-180	180-200	200-220	220-240
Frequency :	10	8	4	4	3	1	2

Solution

Let assumed mean be $a = 170$.

Class	Frequency (f)	Mid value (x)	$d = x - a$ $= x - 170$	fd
100-120	10	110	–60	–600
120-140	8	130	–40	–320
140-160	4	150	–20	–80
160-180	4	170	0	0
180-200	3	190	20	60
200-220	1	210	40	40
220-240	2	230	60	120
Σ	32			–780

Using (1.4),

$$\bar{x} = a + \frac{\Sigma fd}{\Sigma f} = 170 - \frac{780}{32} = 170 - 24.375 \approx 145.6.$$ ***Answer***

Step Deviation Method For a grouped data when equal class-intervals are given, then the calculations of mean can be further simplified. In this case, common factor from the deviations can be taken out, which is same as width of the class-interval. Then the deviation of variate x from the assumed mean a are divided by the common factor. The arithmetic mean then obtained by the following formula.

$$\bar{x} = a + \frac{\Sigma fd}{\Sigma f} \times i, \quad \text{...(1.5)}$$

where

a is assumed mean,

i is the width of the class-interval,

$d = (x - a)/i$ is the deviation of any variate from a.

Example 1.5

Calculate the mean by step deviation method for the following data.

x	:	5	10	15	20	25	30
f	:	21	44	70	65	71	40

Solution

Here, $i = 5$. Let $a = 20$.

x	f	$x - a = x - 20$	$d = \frac{x-a}{i}$	fd
5	21	–15	–3	–63
10	44	–10	–2	–88
15	70	–5	–1	–70
20	65	0	0	0
25	71	5	1	71
30	40	10	2	80
Σ	311			–70

Using (1.5),

$$\bar{x} = a + \frac{\Sigma fd}{\Sigma f} \times i = 20 + \frac{(-70)}{311} \times 5 \approx 18.87.$$ ***Answer***

Example 1.6

Calculate the average marks of the students by step deviation method for the following data.

Marks	:	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	:	40	41	55	30	21	16

Solution

Here, $i = 10$. Let $a = 35$.

Marks	No. of students (f)	Mid value (x)	$d = \frac{x-35}{10}$	fd
0-10	40	5	–3	–120
10-20	41	15	–2	–82
20-30	55	25	–1	–55
30-40	30	35	0	0
40-50	21	45	1	21
50-60	16	55	2	32
Σ	203			–204

Using (1.5),

$$\bar{x} = a + \frac{\sum fd}{\sum f} \times i = 35 + \frac{(-204)}{203} \times 10 = 24.95.$$ **Answer**

Before proceeding futher, let us first understand following definitions.

(1) **Exclusive class** When the class-interval exclude the upper limit of the class, then it is known as exclusive class.

→ ***For example,*** the class-interval of the following type is called exclusive.

$$\{x / a \le x < b\} = [a, b)$$

(2) **Inclusive class** When the class-interval include the upper limit of the class, then it is known as inclusive class.

→ ***For example,*** the class-interval of the following type is called inclusive.

$$\{x / a \le x \le b\} = [a, b]$$

It should be noted that in order to ensure the continuity of the class limits and to get correct class limits, exclusive method of classification should be adopted. To convert the given inclusive class-intervals into exclusive type, adjustment should be done as follows.

(*a*) Find the difference between the lower limit of the second class and the upper limit of the first class.

(*b*) Divide the difference found in (*a*) by 2.

(*c*) Subtract the value obtained in (*b*) from all the lower limits and add to all upper limits.

→ ***For example,*** suppose that the inclusive class in some example is given as follows.

Marks	No. of students
⋮	⋮
30-39	10
40-50	20
⋮	⋮

The adjustment for exclusive data can be done as follows.

Marks	No. of students
⋮	⋮
29.5-39.5	10
39.5-50.5	20
⋮	⋮

When the class-interval in some example given as

Marks	No. of students
⋮	⋮
30-40	10
40-50	20
⋮	⋮

Then it is understood that a student having 40 marks is included in the second class; that is, in the class-interval 40-50.

Note For the class 30 - 39, 30 and 39 are called ***stated lower and upper limits***, respectively, whereas for the same class in exclusive form 29.5 - 39.5, 29.5 and 39.5 are called ***actual (or exact or real) lower and upper limits***, respectively.

Example 1.7

For the following data, find mean.

Class	:	10-19	20-29	30-39	40-49	50-59
Frequency	:	1	1	15	10	20

Solution

Here, the given data is presented in inclusive form. Let us first convert it to exclusive form.

Here, $i = 10$. Let $a = 34.5$.

Class	Frequency (f)	Mid-value (x)	$d = \frac{x - 34.5}{10}$	fd
9.5-19.5	1	14.5	–2	–2
19.5-29.5	1	24.5	–1	–1
29.5-39.5	15	34.5	0	0
39.5-49.5	10	44.5	1	10
49.5-59.5	20	54.5	2	40
Σ	47			47

Using (1.5),

$$\bar{x} = a + \frac{\sum fd}{\sum f} \times i = 34.5 + \frac{47}{47} \times 10 = 44.5.$$ **Answer**

➢ **Weighted Arithmetic Mean** Suppose that $w_1, w_2, \ldots, w_n$ are the weights assigned to the values $x_1, x_2, \ldots, x_n$, respectively, then the weighted arithmetic

mean is defined as follows.

$$\text{Weighted arithmetic mean} = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}. \quad \ldots(1.6)$$

When the assumed mean is used for calculation, then the weighted arithmetic mean can be defined as follows.

$$\text{Weighted arithmetic mean} = a + \frac{\Sigma dw}{\Sigma w}, \quad \ldots(1.7)$$

where

a is assumed mean,

$d = x - a$ is the deviation of each value from the assumed mean a.

When the step deviation method is used for calculation, then the weighted arithmetic mean can be defined as follows.

$$\text{Weighted arithmetic mean} = a + \frac{\Sigma dw}{\Sigma w} \times i, \quad \ldots(1.8)$$

where

a is assumed mean,

i is the width of the class interval,

$d = (x - a)/i$ is the deviation of any variate from a.

Example 1.8

The following table gives the number of students and the average marks obtained by them in different subjects of civil engineering class of an institute. Find the average marks obtained per student.

Subjects	Average marks obtained	No. of students
Mathematics	40	25
SM	70	30
AM	63	15
EME	78	42
EEE	55	13

Solution

Here, the values of x are 40, 70, 63, 78, 55 and their weights w are 25, 30, 15, 42, 13, respectively.

Using (1.6),

$$\bar{x} = \frac{\sum_{i=1}^{5} w_ix_i}{\sum_{i=1}^{5} w_i}$$

$$= \frac{w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5}{w_1 + w_2 + w_3 + w_4 + w_5}$$

$$= \frac{(25)(40) + (30)(70) + (15)(63) + (42)(78) + (13)(55)}{25 + 30 + 15 + 42 + 13}$$

$$= \frac{8036}{125}$$

$$= 64.288.$$

Answer

➢ **Properties of Arithmetic Mean**

(1) The algebraic sum of deviations from the mean is zero; that is, if $\bar{x}$ is the mean for n - observations $x_1, x_2, \ldots x_n$, then

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0 \text{ or } \sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

(2) The sum of squares of the deviations is minimum when taken about the mean.

(3) If $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ are the means of k series of sizes $n_1, n_2, \ldots, n_k$, respectively, then the mean $\bar{x}$ of the composite series is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}.$$

1.1.2 Median For this measure, the observations are arranged either in ascending or in descending order of their magnitudes. Then median is defined as the middle

most value (or obsevation) in a set of observations. Median divides the arranged observations into two equal parts. When the number of observations are odd, then median is the central observation (which is original as given) in a set of observations. When the number of observations are even, then the median is the arithmetic mean of the two central observations. The median is generally denoted by M.

➢ **Median of Ungrouped Data** First arrange the given observations either in ascending or in descending order of their magnitudes.

Case 1 When number of observations are odd (say n), then median is defined as follows.

$$M = \left(\frac{n+1}{2}\right)^{th} \text{observation} \qquad ...(1.9)$$

Case 2 When number of observations are even (say n), then median is defined as follows.

$$M = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2} \qquad ...(1.10)$$

Example 1.9

The number of runs scored by eleven players of a cricket team are 6, 20, 43, 50, 19, 53, 0, 37, 78, 1, 15. Find median.

Solution

Here, $n = 11$. Let us first arrange the given data in ascending order, so that

0, 1, 6, 15, 19, 20, 37, 43, 50, 53, 78.

The number of observations are odd, therefore, using (1.9)

$$M = \left(\frac{n+1}{2}\right)^{th} \text{observation} = \left(\frac{11+1}{2}\right)^{th} \text{observation}$$

$$= 6^{th} \text{observation}$$

$$= 20 \text{ runs.} \qquad \textit{Answer}$$

Example 1.10

Calculate the median marks from the following data.

Roll No.	:	1	2	3	4	5	6	7	8	9	10
Marks	:	10	34	27	24	12	27	20	18	15	30

Solution

Here, $n = 10$. Let us first arrange the given marks in ascending order, so that

10, 12, 15, 18, 20, 24, 27, 27, 30, 34.

The number of observations are even, therefore, using (1.10)

$$M = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$$

$$= \frac{\left(\frac{10}{2}\right)^{th} \text{observation} + \left(\frac{10}{2}+1\right)^{th} \text{observation}}{2}$$

$$= \frac{5^{th} \text{observation} + 6^{th} \text{observation}}{2}$$

$$= \frac{20+24}{2}$$

$$= 22 \text{ marks.} \qquad \textit{Answer}$$

➢ **Median of Grouped Data**

Case 1 When the series is discrete In this case first arrange the values of the variable either in ascending or in descending order of magnitudes. Find cumulative frequencies from the given frequencies. Then calculate the median using the following formula.

$$M = \left(\frac{N+1}{2}\right)^{th} \text{observation,} \qquad ...(1.11)$$

where N is sum of all frequencies.

Example 1.11

Calculate the median for the following data.

No. of students	:	6	4	16	7	8	2
Marks	:	20	9	25	50	40	80

Solution

Let us first arrange marks in ascending order and prepare following table.

Marks	No. of students (f)	Cumulative frequency
9	4	4
20	6	10
25	16	26
40	8	34
50	7	41
80	2	43
Σ	43	

Here, $N = 43$.

Using (1.11),

$$M = \left(\frac{N+1}{2}\right)^{th} \text{ observation}$$

$$= \left(\frac{43+1}{2}\right)^{th} \text{ observation}$$

$$= 22^{nd} \text{ observation.}$$

From table, it is observed that all items from 11 to 26 have their values 25. Since 22nd item lies in this interval, therefore, its value is 25.

Hence,

Median = 25 marks. ***Answer***

Case 2 When the series is continuous In this case data is given in the form of a class-interval with frequency distribution. Find cumulative frequencies from the given frequencies. Then median can be calculated using the following formula.

$$M = L + \frac{\left(\frac{N}{2}\right) - c}{f} \times i, \quad \text{...(1.12)}$$

where

L is the lower limit of the class-interval where median lies,

$N = \Sigma f$, sum of all frequencies,

f is frequency of the class where median lies,

c is cumulative frequency of the class preceding the median class,

i is width of the class-interval of the class where median lies.

Example 1.12

The following table gives the marks obtained by 50 students in statistics. Find the median.

Marks	:	0-10	10-20	20-30	30-40	40-50
No. of students	:	16	12	18	3	1

Solution

Marks	No. of students (f)	Cumulative frequency
0-10	16	16
10-20	12	28
20-30	18	46
30-40	3	49
40-50	1	50
Σ	50	

Here, $N = 50$ or $N/2 = 25$.

Therefore, median class is 10-20, so that

$$L = 10, f = 12, c = 16, i = 10.$$

Using (1.12),

$$M = L + \frac{\left(\frac{N}{2}\right) - c}{f} \times i$$

$$= 10 + \frac{25-16}{12} \times 10$$

$$= 10 + 7.5$$

$$= 17.5 \text{ marks.}$$ ***Answer***

Example 1.13

The following data represents the number of foreign visitors in a multinational company in every 10 days during last 2 months. Use the data to find median.

x	:	0-10	10-20	20-30	30-40	40-50	50-60
No. of visitors f	:	12	18	27	20	17	06

[GTU, June 2016]

Solution

r	No. of visitors (f)	Cumulative frequency
0-10	12	12
10-20	18	30
20-30	27	57
30-40	20	77
40-50	17	94
50-60	06	100
Σ	100	

Here, $N = 100$ or $N/2 = 50$.

Therefore, median class is 20-30, so that

$$L = 20, f = 27, c = 30, I = 10.$$

Using (1.12),

$$M = L + \frac{\left(\frac{N}{2}\right) - c}{f} \times i$$

$$= 20 + \frac{50 - 30}{27} \times 10$$

$$= 20 + \frac{200}{27}$$

$$\approx 20 + 7.407$$

$$= 27.407.$$ ***Answer***

Example 1.14

The following table gives the marks obtained by 50 students in statistics. Find the median.

Marks	:	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
No. of students	:	4	6	10	5	7	3	9	6

Solution

Marks	No. of students (f)	Cumulative frequency
10-14	4	4
15-19	6	10
20-24	10	20
25-29	5	25
30-34	7	32
35-39	3	35
40-44	9	44
45-49	6	50
Σ	50	

Here, $N = 50$ or $N/2 = 25$.

Therefore, median class is 25-29.

Now, the given series is inclusive. If we convert it to exclusive, then

$$L = \frac{24 + 25}{2} = 24.5.$$

Also,

$$f = 5, c = 20, i = 5.$$

Using (1.12),

$$M = L + \frac{\left(\frac{N}{2}\right) - c}{f} \times i$$

$$= 24.5 + \frac{25 - 20}{5} \times 5$$

$$= 29.5 \text{ marks.}$$ ***Answer***

1.1.3 Mode Mode or model value is defined as that value in a series of observations which occurs most frequently. When frequency distribution is given, then mode is that variate which has maximum frequency.

→ ***For example,*** in series 2, 4, 2, 5, 7, 2, 8, 9, it is observed that 2 occurs most frequently and therefore mode of the series is 2.

In practical life we often hear the following.

(a) Average height of a person is 150 *cm*.

(b) Average size of the foot wear is 6.

In all these cases, the average indicates the mode.

Example 1.15

For the following data, find mode.

Variable (x)	:	3	4	5	6
Frequency (f)	:	15	20	19	10

Solution

Here, maximum frequency is 20 for $x = 4$. Therefore, mode is 4. *Answer*

Sometimes there may be two or more than two values which occur with equal frequency, the distribution is then called ***bimodel*** or ***multimodel***.

→ ***For example,*** for the series 40, 31, 42, 35, 31, 40, 65, 22, the distribution is bimodel and modes are 31, 40.

➢ **Mode of Continuous Frequency Distribution** In this case mode can be calculated by the following formula.

$$\text{Mode} = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i \quad \ldots(1.13)$$

where

L is the lower limit of the model class,

f_1 is the frequency of the class preceding the model class,

f_2 is the frequency of the class succeeding the model class,

f_m is the frequency of the model class,

i is the width of the model class.

It should be noted here that model class is that class in which mode lies.

Example 1.16

Calculate mode for the following data.

Marks	:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	:	5	15	20	20	32	14	14

Solution

From the table, the maximum frequency is 32 and it lies in the class 40-50.

Thus, model class is 40-50. Here,

$$L = 40,\ f_1 = 20,\ f_2 = 14,\ f_m = 32,\ i = 10.$$

Using (1.13),

$$\text{Mode} = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

$$= 40 + \frac{32 - 20}{2(32) - 20 - 14} \times 10$$

$$= 40 + \frac{12}{30} \times 10$$

$$= 44. \quad \textit{Answer}$$

➢ **Asymmetrical Distribution** A distribution in which mean, mode and median coincide is called ***asymmetrical distribution***. If the distribution is moderately asymmetrical then mean, mode and median are connected by the formula

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean} \quad \ldots(1.14)$$

1.2 Measures of Dispersion

Although the mean of the given data indicates where the centre of the data lies, it gives no measure about the spread of the data.

→ ***For example,*** –5, 0, 5 and –50, 0, 50 both have the same mean 0 but clearly the data given in the second case are much more widely dispersed than those in the first case. So, measures of central tendency are not sufficient for having some idea about dispersion. Measures of dispersion gives the idea about the degree to which numerical data tend to spread about an average value.

Measures of dispersion may be either absolute or relative.

Measures of dispersion

- **Absolute**
  - → Range
  - → Inter-quartile range
  - → Quartile deviation (Semi-inter quartile range)
  - → Mean deviation (Average deviation)
  - → Standard deviation (Root mean square deviation)
- **Relative**
  - → Coefficient of variation
  - → Coefficient of quartile deviation
  - → Coefficient of mean deviation

The relative measures are used only for the purpose of comparison between two or more series with varying size or number of items or varying central values or varying units of calculations.

A commonly used measure of dispersion is the ***standard deviation***.

1.2.1 Deviation, Variance and Standard Deviation Let $x_1, x_2, \ldots, x_n$ be n-observations of the given data with mean $\bar{x}$. Then $x_i - \bar{x}$ is the amount by which x_i differs from the mean. The quantity $x_i - \bar{x}$ is called ***deviation*** of x_i from the mean. It is clear that some of these deviations will be positive and some will be negative. Moreover, the mean of these deviations is always zero and so this is not helpful in measuring the dispersion of the data. To avoid this situation squared deviation is taken; that is, $(x_i - \bar{x})^2$. Then the variance and standard deviation are defined as follows.

$$\text{Variance} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \quad \ldots(1.15)$$

$$\text{Standard deviation (S.D.)} = \sqrt{\text{Variance}} \quad \ldots(1.16)$$

Simplifying formula (1.15) implies

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2 \quad \ldots(1.17)$$

The above formula is useful when ***ungrouped data*** are given.

Example 1.17

Find the standard deviation of 3, 4, 6, 7, 9, 15.

Solution

Here, $n = 6$. Let $x_1 = 3, x_2 = 4, x_3 = 6, x_4 = 7, x_5 = 9, x_6 = 15$.

Using (1.1),

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{3+4+6+7+9+15}{6} = \frac{44}{6}.$$

$$\sum_{i=1}^{6} x_i^2 = 9+16+36+49+81+225 = 416.$$

Using (1.17),

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{6} x_i^2}{n} - \bar{x}^2}$$

$$= \sqrt{\frac{416}{6} - \left(\frac{44}{6}\right)^2}$$

$$\approx 3.94. \qquad \textit{Answer}$$

➢ **Calculation of Standard Deviation for Grouped Data** Let $x_1, x_2, \ldots, x_n$ be n-observations of the given data with corresponding frequencies $f_1, f_2, \ldots, f_n$, then the variance and standard deviation is defined as follows.

$$\text{Variance} = \frac{f_1(x_1-\bar{x})^2+f_2(x_2-\bar{x})^2+\ldots+f_n(x_n-\bar{x})^2}{f_1+f_2+\ldots+f_n} = \frac{\sum_{i=1}^{n} f_i(x_i-\bar{x})^2}{\sum_{i=1}^{n} f_i} \quad \ldots(1.18)$$

where $\bar{x}$ is mean.

$$\text{Standard deviation} = \sqrt{\text{Variance}} \quad \ldots(1.19)$$

Simplifying formula (1.18) implies

$$\text{Variance} = \frac{\sum_{i=1}^{n} f_i x_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i x_i}{N}\right)^2, \quad N = \sum_{i=1}^{n} f_i \quad \ldots(1.20)$$

Example 1.18

Find standard deviation for the following data.

x :	5	10	15	20	25
f :	7	4	6	3	5

Solution

Here,

$$\sum_{i=1}^{5} f_i = 7+4+6+3+5 = 25.$$

$$\sum_{i=1}^{5} f_i x_i = (7)(5)+(4)(10)+(6)(15)+(3)(20)+(5)(25)$$
$$= 35+40+90+60+125$$
$$= 350.$$

$$\sum_{i=1}^{5} f_i x_i^2 = (7)(25)+(4)(100)+(6)(225)+(3)(400)+(5)(625)$$
$$= 175+400+1350+1200+3125$$
$$= 6250.$$

Using (1.18),

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{5} f_i x_i^2}{\sum_{i=1}^{5} f_i} - \left(\frac{\sum_{i=1}^{5} f_i x_i}{\sum_{i=1}^{5} f_i}\right)^2}$$

$$= \sqrt{\frac{6250}{25} - \left(\frac{350}{25}\right)^2}$$

$$= \sqrt{250-196}$$

$$= \sqrt{54}$$

$$\approx 7.348.$$

Answer

➢ **Short-cut Method for Calculating Standard Deviation**

Case 1 For ungrouped data When the mean of the given data comes out to be a fraction, then this method is applicable. The formula used in this case is given as follows.

$$\text{Standard deviation} = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}. \qquad ...(1.21)$$

where

$d = x_i - a$, a is assumed mean, n is the total number of observations.

Example 1.19

Find mean and standard deviation of 48, 43, 65, 57, 31, 60, 37, 48, 59, 78.

Solution

Let $a = 50$.

x	$d = x - a$	d^2
48	−2	4
43	−7	49
65	15	225
57	7	49
31	−19	361
60	10	100
37	−13	169
48	−2	4
59	9	81
78	28	784
Σ	26	1826

Using (1.3),

$$\bar{x} = a + \frac{\sum d}{n} = 50 + \frac{26}{10} = 52.6.$$

Using (1.21),

$$\text{S.D.} = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$= \sqrt{\frac{1826}{10} - \left(\frac{26}{10}\right)^2}$$

$$= \sqrt{182.6 - 6.76}$$

$$= \sqrt{175.84}$$

$$\approx 13.26.$$ Answer

Example 1.20

The pH solution is measured eight times using the same instrument and the data obtained are as follows.

7.15, 7.20, 7.18, 7.19, 7.21, 7.20, 7.16, 7.18

Calculate the mean, variance and standard deviation. *[GTU, May 2016]*

Solution

Let $a = 7.17$.

x	$d = x - a$	d^2
7.15	−0.02	0.0004
7.20	0.03	0.0009
7.18	0.01	0.0001
7.19	0.02	0.0004
7.21	0.04	0.0016
7.20	0.03	0.0009
7.16	−0.01	0.0001
7.18	0.01	0.0001
Σ	0.11	0.0045

Using (1.3),

$$\bar{x} = a + \frac{\Sigma d}{n} = 7.17 + \frac{0.11}{8} \approx 7.184.$$

Using (1.21),

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

$$= \sqrt{\frac{0.0045}{8} - \left(\frac{0.11}{8}\right)^2}$$

$$= \sqrt{0.00056 - 0.00019}$$

$$= \sqrt{0.00037}$$

$$\approx 0.01924.$$

Using (1.16),

$$\text{Variance} = (\text{S.D})^2$$

$$= (0.01924)^2$$

$$\approx 0.00037.$$ Answer

Case 2 For grouped data Here, standard deviation is defined as follows.

$$\text{S.D.} = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \quad \ldots(1.22)$$

where

$d = x_i - a$, a is assumed mean and $N = \Sigma f_i$.

Example 1.21

Find the standard deviation for the following data.

x	:	5	10	15	20	25
f	:	7	4	6	3	5

Solution

Let $a = 15$.

x	f	$d = x - a$	fd	d^2	fd^2
5	7	−10	−70	100	700
10	4	−5	−20	25	100
15	6	0	0	0	0
20	3	5	15	25	75
25	5	10	50	100	500
Σ	25		−25		1375

Using (1.22),

$$\text{S.D.} = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

$$= \sqrt{\frac{1375}{25} - \left(\frac{-25}{25}\right)^2}$$

$$= \sqrt{55 - 1}$$

$$\simeq 7.348.$$ *Answer*

Step Deviation Method for Calculating Standard Deviation In this case, standard deviation is defined as follows.

$$\text{S.D.} = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i, \qquad ...(1.23)$$

where

$d = (x_i - a)/i$,

a is assumed mean,

$N = \Sigma f$,

i is the width of the class interval.

1.3 Moments - Mean, Variance, Skewness, Kurtosis

➢ **Moments About Arbitrary Origin** Let $x_1, x_2, ..., x_n$ be n-observations of the given data and let $f_1, f_2, ..., f_n$ be their corresponding frequencies. Let A be any arbitrary origin, then the r^{th}-moment about A (also called ***raw moments***) is denoted and defined as

$$\mu'_r \text{ (or } m'_r) = \frac{\sum_{i=1}^{n} (x_i - A)^r}{n} \qquad ...(1.24)$$

- For a frequency distribution, the r^{th}-moment about A is defined by

$$\mu'_r = \frac{\sum_{i=1}^{n} f_i (x_i - A)^r}{N}; N = \sum_{i=1}^{n} f_i \qquad ...(1.25)$$

- If $x_i - A = d_i$, then (1.24) can be written as

$$\mu'_r = \frac{\sum_{i=1}^{n} f_i d_i^r}{N}; N = \sum_{i=1}^{n} f_i \qquad ...(1.26)$$

- **For equal classs intervals If**

$$\frac{x_i - A}{i} = d_i,$$

then (1.24) becomes

$$\mu'_r = \frac{\sum_{i=1}^{n} f_i d_i^r}{N} \times i^r \qquad ...(1.27)$$

where i is the width of the class interval.

In particular, when $A = 0$, then we obtain r^{th}-moment about zero.

The first moment about zero using (1.24) and (1.25) can be written respectively as

$$\mu'_1 = \text{first moment about zero} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \text{ (Mean)},$$

$$\mu'_1 = \text{first moment about zero} = \frac{\sum_{i=1}^{n} x_i f_i}{N} = \bar{x} \text{ (Mean)}.$$

Similarly, we can define 2nd, 3rd, and 4th-moments.

➢ **Moments About Actual Mean** When $A = \bar{x}$ (mean) in the above definition (1.24), then the r^{th}-moment about mean (or r^{th}-central moment) is denoted and defined as

$$\mu_r \text{ (or } m_r) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n}; \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad ...(1.28)$$

Therefore,

First moment about mean is given by

$$\mu_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = 0.$$

Second moment about mean is given by

$$\mu_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \text{Variance}(\sigma^2).$$

μ_3 is called third moment about mean and it measures skewness.

μ_4 is called fourth moment about mean and it measures kurtosis.

● For a frequency distribution, the r^{th}-moment about mean (using (1.25)) is defined as

$$\mu_r = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})}{N}, \quad N = \sum_{i=1}^{n} f_i, \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N} \qquad ...(1.29)$$

Therefore,
First moment about mean is given by

$$\mu_1 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})}{N} = 0.$$

Second moment about mean is given by

$$\mu_2 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{N} = \text{Variance}(\sigma^2).$$

μ_3 is called third moment about mean and it measures skewness.

μ_4 is called fourth moment about mean and it measures kurtosis.

Note 1 The moments are used to describe the various characteristics of a frequency distribution like central tendency, variation, skewness and kurtosis.

Note 2 When the actual mean is in fraction, then the moments about A can be calculated and then it will be converted about the actual mean. This process can be done by following relations.

$$\mu_1 = \mu_1' - \mu_1' = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

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

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

Note 3 Because of convenience in obtaining measures of various characteristics (mean, variance, skewness, kurtosis), the calculation of the first four moments about the mean are sufficient.

Note 4 From μ_2, μ_3 and μ_4, two important constants of a distribution can be defined as follows.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ (Skewness)}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ (Kurtosis)}$$

Note 5 In symmetrical distribution, odd moments are always zero, but this rule does not hold true for asymmetrical distribution.

Example 1.22

Find first four central moments of the observations 1, 3, 5, 7, 9.

Solution

Here, $n = 5$ and let $x_1 = 1, x_2 = 3, x_3 = 5, x_4 = 7, x_5 = 9$.
Therefore,

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{1+3+5+7+9}{5} = \frac{25}{5} = 5.$$

	x_i	$x_i - 5$	$(x_i-5)^2$	$(x_i-5)^3$	$(x_i-5)^4$
	1	−4	16	−64	256
	3	−2	4	−8	16
	5	0	0	0	0
	7	2	4	8	16
	9	4	16	64	256
Σ	25	0	40	0	544

Using (1.28),

$$\mu_1 = \frac{\sum_{i=1}^{5}(x_i - 5)}{5} = \frac{0}{5} = 0.$$

$$\mu_2 = \frac{\sum_{i=1}^{5}(x_i - 5)^2}{5} = \frac{40}{5} = 8.$$

$$\mu_3 = \frac{\sum_{i=1}^{5}(x_i - 5)^3}{5} = \frac{0}{5} = 0.$$

$$\mu_4 = \frac{\sum_{i=1}^{5}(x_i - 5)^4}{5} = \frac{544}{5} = 108.8.$$ *Answer*

Example 1.23

Calculate the first four moments about the mean for the following data.

Marks :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students :	8	12	20	30	15	10	5

Solution

Here,

Marks (x_i)	No. of students (f_i)	Mid point of x_i	$f_i x_i$	$d_i = \frac{x_i - 35}{10}$	$f_i d_i$	$f_i d_i^2$	$f_i d_i^3$	$f_i d_i^4$
0-10	8	5	40	-3	-24	72	-216	648
10-20	12	15	180	-2	-24	48	-96	192
20-30	20	25	500	-1	-20	20	-20	20
30-40	30	35	1050	0	0	0	0	0
40-50	15	45	675	1	15	15	15	15
50-60	10	55	550	2	20	40	80	160
60-70	5	65	325	3	15	45	135	405
Σ	100	245	3320	0	-18	240	-102	1440

Here,

$$\bar{x} = \frac{\sum_{i=1}^{7} x_i f_i}{\sum f_i} = \frac{3320}{100} = 33.20.$$

In this case the mean is fraction, so referring to Note 2, we will use the formula for moments about arbitrary origin. Let $A = 35$.

Using (1.27),

$$\mu_1' = \frac{\sum_{i=1}^{7} f_i d_i}{\sum_{i=1}^{7} f_i} \times i = \frac{-18}{100} \times 10 = -1.8.$$ (Since $i = 10$)

$$\mu_2' = \frac{\sum_{i=1}^{7} f_i d_i^2}{\sum_{i=1}^{7} f_i} \times i^2 = \frac{240}{100} \times (10)^2 = \frac{240}{100} \times 100 = 240.$$

$$\mu_3' = \frac{\sum_{i=1}^{7} f_i d_i^3}{\sum_{i=1}^{7} f_i} \times i^3 = \frac{-102}{100} \times (10)^3 = \frac{-102}{100} \times 1000 = -1020.$$

$$\mu_4' = \frac{\sum_{i=1}^{7} f_i d_i^4}{\sum_{i=1}^{7} f_i} \times i^4 = \frac{1440}{100} \times (10)^4 = \frac{1440}{100} \times 10000 = 144000.$$

Let us now convert the moments about arbitrary origin to moments about the mean using relations given in Note 2.

$$\mu_1 = 0.$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 240 - (-1.8)^2 = 236.76.$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 = -1020 - 3(-1.8)(240) + 2(-1.8)^3 = 264.336.$$

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

$$= 144000 - 4(-1.8)(-1020) + 6(-1.8)^2(240) - 3(-1.8)^4$$

$$= 141290.11.$$

1.4 Skewness

When frequencies are symmetrically distributed about the mean, then the frequency distribution is said to be ***symmetrical***, otherwise it is said to be ***skew*** (***asymmetrical***). That is, ***skewness*** (***lack of symmetry***) is a measure that refers to any departure from symmetry or asymmetry in the distribution. A symmetrical distribution has ***zero skewness*** and identical mean (μ), median (M) and mode (Z). That is,

$$\text{Mean} = \text{Median} = \text{Mode}.$$

Therefore, when plotted on a graph, it will give a perfectly bell-shaped curve, known as normal curve. (refer Figure 1.2)

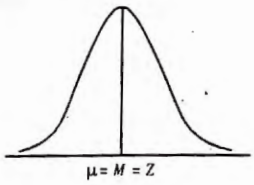


Figure 1.2 Symmetrical distribution (zero skewness)

In skewed distribution mean, median and mode are not identical; that is,

$$\text{Mean} \neq \text{Median} \neq \text{Mode}.$$

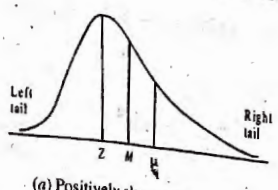
Therefore, when plotted on a graph, there is a long tail or steepness of the curve on one side without its counterpart on the otherside.

When longer tail of the skewed distribution is towards the right, then ***skewness*** is said to be ***positive*** (refer Figure1.3(a)) and the distribution is said to be ***positively skewed distribution***. In this case,

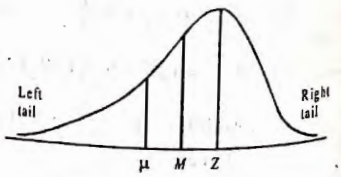
$$\text{Mean} > \text{Median} > \text{Mode}.$$

Similarly, when longer tail of the skewed distribution is towards the left, then ***skewness*** is said to be ***negative*** (refer Figure 1.3(b)) and the distribution is said to be ***negatively skewed distribution***. In this case,

$$\text{Mean} < \text{Median} < \text{Mode}.$$



(*a*) Positively skewed distribution



(*b*) Negatively skewed distribution

Figure 1.3

➢ **Measures of Skewness** Measures of skewness tell us the direction and extent of asymmetry in a distribution and facilitate to compare two distributions. Measures of skewness can either be ***absolute*** or ***relative***.

- **Absolute measures of skewness** These measures are also called ***first measures of skewness***. These measures do not have practical utility because they are expressed in terms of units of distribution and so they are not used for comparative study.
- **Relative measures of skewness** These measures are also known as ***coefficient of skewness***, which are independent of units of distribution and are pure numbers. The commonly used relative measures of skewness are as follows.
  (1) Karl Pearson's coefficient of skewness
  (2) Bowley's coefficient of skewness
  (3) Kelly's coefficient of skewness
  (4) Measures of skewness based on moments
  Out of above four, we will study (1).

1.4.1 Karl Pearson's Coefficient of Skewness (Pearson's First Measure of Skewness) This measure is denoted and defined as follows.

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} \qquad \ldots(1.30)$$

When the mode is ***ill-defined***, then the measure of coefficient of skewness (also known as ***Pearson's second measure of skewness***) is defined as follows.

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}} \qquad \ldots(1.31)$$

Note 1 The coefficient of skewness lies between (–1) and 1.

Note 2 For a positively skewed distribution $S_k > 0$, for a negatively skewed distribution $S_k < 0$ and for a symmetrical distribution $S_k = 0$.

Note 3 When the mean moves more away from the mode, then the skewness will be larger.

Example 1.24

Find Karl Pearson's coefficient of skewness for the following data.

3, 4, 8, 2, 4, 4, 3, 4, 7, 8, 9, 11

Solution

Here, $n = 12$ and let $x_1 = 3, x_2 = 4, x_3 = 8, x_4 = 2, x_5 = 4, x_6 = 4, x_7 = 3, x_8 = 4, x_9 = 7, x_{10} = 8, x_{11} = 9, x_{12} = 11$.

Using (1.1),

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{3+4+8+2+4+4+3+4+7+8+9+11}{12}$$

$$= \frac{67}{12}.$$

$$\sum_{i=1}^{12} x_i^2 = 9+16+64+4+16+16+9+16+49+64+81+121$$

$$= 465.$$

Using (1.17),

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{12} x_i^2}{n} - \bar{x}^2}$$

$$= \sqrt{\frac{465}{12} - \left(\frac{67}{12}\right)^2}$$

$$= \sqrt{\frac{5580-4489}{144}}$$

$$= \sqrt{\frac{1091}{144}}$$

$$= \frac{\sqrt{1091}}{12}.$$

Here, maximum frequency is 4 for $x = 4$. Therefore, mode is 4.
Using (1.30), coefficient of skewness is given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

$$= \frac{\frac{67}{12} - 4}{\frac{\sqrt{1091}}{12}}$$

$$= \frac{67-48}{\sqrt{1091}}$$

$$= \frac{19}{\sqrt{1091}}$$

$$\approx 0.575.$$ *Answer*

Example 1.25

Find Karl Pearson's coefficient of skewness for the following data.

Marks (x)	:	20	21	22	23	24	25	26	27	28
No. of students (f)	:	7	12	25	10	8	6	8	3	1

Solution

Here,

$$\sum_{i=1}^{9} f_i = 7+12+25+10+8+6+8+3+1 = 80.$$

$$\sum_{i=1}^{9} f_i x_i = (7)(20)+(12)(21)+(25)(22)+(10)(23)+(8)(24)+(6)(25)+(8)(26)+(3)(27)+(1)(28)$$

$$= 140 + 252+550+230+192+150+208+81+28$$

$$= 1831.$$

$$\sum_{i=1}^{9} f_i x_i^2 = (7)(400)+(12)(441)+(25)(484)+(10)(529)+(8)(576)+(6)(625)+(8)(676)+(3)(729)+(1)(784)$$

$$= 2800 + 5292+12100+5290+4608+3750+5408+2187+784$$

$$= 42219.$$

Using (1.2),

$$\bar{x} = \frac{\sum_{i=1}^{9} x_i f_i}{\sum_{i=1}^{9} f_i} = \frac{1831}{80}.$$

Using (1.20),

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{9} f_i x_i^2}{\sum_{i=1}^{9} f_i} - \left(\frac{\sum_{i=1}^{9} f_i x_i}{\sum_{i=1}^{9} f_i}\right)^2}$$

$$= \sqrt{\frac{42219}{80} - \left(\frac{1831}{80}\right)^2}$$

$$= \frac{1}{80}\sqrt{3377520 - 3352561}$$

$$= \frac{\sqrt{24959}}{80}.$$

Here, observation having highest frequency 22. Therefore, mode is 22.

Using (1.30), coefficient of skewness is given by

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

$$= \frac{\frac{1831}{80} - 22}{\frac{\sqrt{24959}}{80}}$$

$$= \frac{1831 - 1760}{\sqrt{24959}}$$

$$= \frac{71}{\sqrt{24959}}$$

$$\approx 0.45.$$ *Answer*

Example 1.26

Calculate coefficient of skewness by Karl Pearson's method for the following data.

x :	0-100	100-200	200-300	300-400	400-500	500-600	600-700	700-800
f :	6	10	18	20	15	12	10	9

Solution

Here, $i = 100$. Let $a = 350$.

x	f	Mid value (x)	$d = \frac{x-350}{100}$	d^2	fd	fd^2
0-100	6	50	−3	9	−18	54
100-200	10	150	−2	4	−20	40
200-300	18	250	−1	1	−18	18
300-400	20	350	0	0	0	0

400-500	15	450	1	1	15	15
500-600	12	550	2	4	24	48
600-700	10	650	3	9	30	90
700-800	9	750	4	16	36	144
Σ	100				49	409

Using (1.5),

$$\bar{x} = a + \frac{\Sigma fd}{\Sigma f} \times i = 350 + \frac{49}{100} \times 100 = 399.$$

Using (1.23),

$$\text{S.D.} = \sqrt{\frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2} \times i$$

$$= \sqrt{\frac{409}{100} - \left(\frac{49}{100}\right)^2} \times 100$$

$$\approx 196.21.$$

From the table, the maximum frequency is 20 and it lies in the class 300-400. Thus, model class is 300-400. Here,

$$L = 300, f_1 = 18, f_2 = 15, f_m = 20$$

Using (1.13),

$$\text{Mode} = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

$$= 300 + \frac{20 - 18}{40 - 18 - 15} \times 100$$

$$= 300 + \frac{2}{7} \times 100$$

$$= 328.57.$$

Using (1.30), coefficient of skewness is given by

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

$$\approx \frac{399 - 328.57}{196.21}$$

$$\approx 0.36.$$ *Answer*

Example 1.27

Calculate Karl Pearson's coefficient of skewness from the following data.

x :	40-50	50-60	60-70	70-80	80-90	90-100	100-110	110-120	120-130	130-140
f :	5	6	8	10	25	30	36	50	60	70

Solution

Here, the maximum frequency is 70 and which occurs at the end of the frequency distribution. Hence, the mode is ill-defined. Therefore, formula (1.31) is useful here for finding coefficient of skewness.

Here, $i = 10$. Let $a = 85$.

x	f	Mid value (x)	$d = \frac{x-85}{10}$	d^2	fd	fd^2	c
40-50	5	45	–4	16	–20	80	5
50-60	6	55	–3	9	–18	54	11
60-70	8	65	–2	4	–16	32	19
70-80	10	75	–1	1	–10	10	29
80-90	25	85	0	0	0	0	54
90-100	30	95	1	1	30	30	84
100-110	36	105	2	4	72	144	120
110-120	50	115	3	9	150	450	170
120-130	60	125	4	16	240	960	230
130-140	70	135	5	25	350	1750	300
Σ	300(N)				778	3510	

Using (1.5),

$$\bar{x} = a + \frac{\Sigma fd}{\Sigma f} \times i = 85 + \frac{778}{300} \times 10 \approx 110.93.$$

Using (1.23),

$$\text{S.D.} = \sqrt{\frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2} \times i$$

$$= \sqrt{\frac{3510}{300} - \left(\frac{778}{300}\right)^2} \times 10$$

$$\approx 22.30.$$

Here, $N = 300$ or $N/2 = 150$.

Therefore, median class is 110–120, so that

$L = 110, f = 50, c = 120.$

Using (1.12),

$$\text{Median} = L + \frac{\left(\frac{N}{2}\right) - c}{f} \times i$$

$$= 110 + \frac{150 - 120}{50} \times 10$$

$$= 110 + \frac{300}{50}$$

$$= 116.$$

Using (1.31), coefficient of skewness is given by

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

$$\approx \frac{3(110.93 - 116)}{22.30}$$

$$\approx -0.68.$$ **Answer**

Example 1.28

Using following data, determine which section is more skewed.

Section A: $\bar{x} = 35$ S.D. = 14.3 Mode = 37.12

Section B: $\bar{x} = 32$ S.D. = 14.3 Mode = 27.8

Solution

Using (1.30),

For Section A

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{35 - 37.12}{14.3} \approx -0.15.$$

For Section B

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{32 - 27.8}{14.3} \approx 0.29.$$

Here,

$$|0.29| > |-0.15|$$

and so the distribution in section B is more skewed (refer Note 3). **Answer**

1.5 Short Questions

Example 1.29

Define mode and also give the relationship between mean, median and mode.

[GTU, May 2016]

Solution

Mode is defined as that value in a series of observations which occurs most frequently.
The relationship between mean, median and mode is (using (1.14))

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}.$$

Answer

Example 1.30

Find the arithmetic mean of the following frequency distribution.

x :	1	2	3	4
f :	4	5	2	1

[GTU, June 2016]

Solution

Here, the arithmetic mean by using formula (1.2) becomes

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i f_i}{\sum_{i=1}^{4} f_i} = \frac{(1)(4)+(2)(5)+(3)(2)+(4)(1)}{4+5+2+1}$$

$$= \frac{4+10+6+4}{12}$$

$$= \frac{24}{12}$$

$$= 2.$$

Answer

Example 1.31

What is the mode of the following frequency distribution?

x :	1	2	3	4
f :	4	7	10	8

[GTU, June 2016]

Solution

Mode is defined as that value which has maximum frequency. Therefore, mode is 3.

Answer

Example 1.32

Which measures are called measures of central tendency?

[GTU, June 2017 - Comp.]

Solution

Mean, median and mode are called the measures of central tendency.

Review Exercises

01. Calculate mean, median, mode, standard deviation and variance for the following data.

10.2, 9.5, 8.3, 9.7, 9.5, 11.1, 7.8, 8.8, 9.5, 10

[GTU, May 2017]

02-06. Find Karl Pearson's coefficient of skewness for the following data.

02. 35, 37, 28, 26, 35, 35, 38, 40

03. 8, 12, 8, 12, 16, 15, 7, 3, 14, 15, 11, 10 (Here, mode is ill-defined)

04.

x :	0	1	2	3	4	5	6	7
f :	12	17	29	19	8	4	1	0

05.

x :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f :	10	15	24	25	10	10	6

06.

x :	20	30	40	50	60	70	80
f :	10	25	40	65	80	95	100

07. Calculate the first four moments about the mean for the following data.

x :	2	3	4	5	6
f :	1	3	7	3	1

08. Find the third moment about mean for the following frequency distribution.

x :	5	7	10	18	26
frequency (f) :	5	14	22	6	3

[GTU, June 2017 - Comp.]

Answers

Review Exercises

01. 9.44, 9.5, 9.5, 0.95, 0.9025 **02.** −0.17 **03.** −0.47 **04.** 0.08 **05.** 0.05 **06.** −0.089
07. 0, 0.933, 0, 2.533 **08.** mean = 10.58, 238.71

Chapter 2

Curve Fitting

2.1 Introduction

In practical statistics, we come across many situations where we often require to find a relationship between two or more variables.

→ ***For example,*** weight and height of a person, demand and supply, expenditure depends on income, etc.

These relations, in general, may be expressed by polynomial or they may have exponential or logarithmic relationship. In order to determine such relationship, first it is require to collect the data showing corresponding values of the variables under consideration. Suppose

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \qquad \ldots(2.1)$$

be the data showing corresponding values of the variables x and y under consideration. If we plot the above data points on a rectangular coordinate system, then the set of points so plotted form a ***scatter diagram***. From this diagram, it is sometimes possible to visualise a smooth curve approximating the data. Such a curve is called an ***approximating curve***.

In particular, if the data approximate well to a straight line, we say that a linear relationship exists between the variables. It is quite possible that the relationship may be nonlinear. Thus, the general problem of finding a functional relationship of the form $y = f(x)$ between two variables x and y, giving the approximating curve and which approximately fit the given data (2.1) of x and y, is called ***curve fitting***.

The fitting of curves to a set of numerical data is of considerable importance from theoretical as well as practical statistics point of view. Theoretically, it is useful in the study of correlation and regression (lines of regression can be regarded as fitting of linear curves to a given bivariate frequency or probability distributions). In practical statistics, curve fitting enables us to represent a close functional relationship between two variables by polynomials, exponentials or logarithmic functions using the principle of least squares.

2.2 The Method of Least Squares

The method of least squares assumes the best-fit curve of a given type that has the minimum sum of the square of the deviations (least square error) from a given set of data.

Suppose that the data points are $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x is independent and y is dependent variable. Let the fitting curve $f(x)$ has the following ***deviations*** (or ***errors*** or ***residuals***) from each data points.

$$d_1 = y_1 - f(x_1),\ d_2 = y_2 - f(x_2),\ \ldots,\ d_n = y_n - f(x_n)$$

Clearly, some of the deviations will be positive and others negative. Thus, to give equal weightage to each error, we square each of these and form their sum; that is,

$$D = d_1^2 + d_2^2 + \ldots + d_n^2.$$

Now, according to the method of least squares, the best fitting curve has the property that

$$D = d_1^2 + d_2^2 + \ldots + d_n^2 = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \text{a minimum.}$$

Note The principle of least squares does not help in determining the form of the appropriate curve which can fit a given data, but it helps only in determining the best possible values of the constants of the resulting equations when the approximate form of the curve is known in advance.

2.2.1 Fitting of a Straight Line Suppose the equation of a straight line of the form $y = a + bx$ is to be fitted to the n-data points

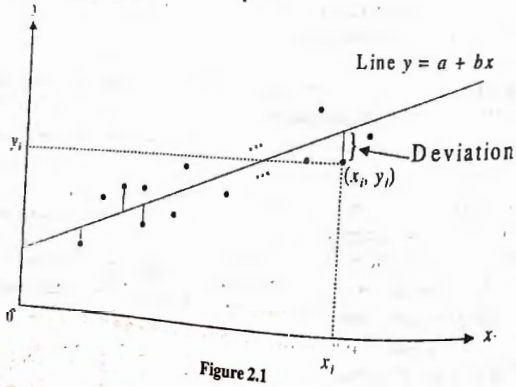


Figure 2.1

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n);\ n \geq 2,$$

where a is y-intercept and b is its slope (refer Figure 2.1).

For the general point (x_i, y_i), if the vertical distance of this point from the line $y = a + bx$ is the deviation d_i, then

$$d_i = y_i - f(x_i) = y_i - a - bx_i.$$

Applying method of least squares, the values of a and b are so determined that they minimise

$$D = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

This will be so, if

$$\frac{\partial D}{\partial a} = 0 \Rightarrow -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0,$$

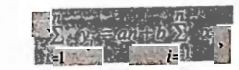
$$\frac{\partial D}{\partial b} = 0 \Rightarrow -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0.$$

Simplifying and expanding the above equations, we have

$$\sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} 1 + b \sum_{i=1}^{n} x_i,$$

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2,$$

which implies



...(2.2)

...(2.3)

Equations (2.2) and (2.3) are known as ***normal equations*** or ***least square equations***. From these equations, we have

$$a = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}, \quad \ldots(2.4)$$

$$b = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \ldots(2.5)$$

Substituting these values of a and b in the equation of a straight line $y = a + bx$, the required best straight line fit to the given data is obtained.

Example 2.1

Using method of least squares, find the best fitting straight line to the following data.

x :	1	2	3	4	5
y :	1	3	5	6	5

Solution

Here, $n = 5$. Let $y = a + bx$ be the required straight line fit. In order to find a and b, let us first calculate the following table.

	x_i	y_i	x_i^2	$x_i y_i$
(x_1, y_1)	1	1	1	1
(x_2, y_2)	2	3	4	6
(x_3, y_3)	3	5	9	15
(x_4, y_4)	4	6	16	24
(x_5, y_5)	5	5	25	25
Σ	15	20	55	71

Using (2.4),

$$a = \frac{\left(\sum_{i=1}^{5} x_i^2\right)\left(\sum_{i=1}^{5} y_i\right) - \left(\sum_{i=1}^{5} x_i\right)\left(\sum_{i=1}^{5} x_i y_i\right)}{(5)\left(\sum_{i=1}^{5} x_i^2\right) - \left(\sum_{i=1}^{5} x_i\right)^2}$$

$$= \frac{(55)(20) - (15)(71)}{(5)(55) - (15)^2}$$

$$= 0.7.$$

Using (2.5),

$$b = \frac{(5)\left(\sum_{i=1}^{5} x_i y_i\right) - \left(\sum_{i=1}^{5} x_i\right)\left(\sum_{i=1}^{5} y_i\right)}{(5)\left(\sum_{i=1}^{5} x_i^2\right) - \left(\sum_{i=1}^{5} x_i\right)^2}$$

$$= \frac{(5)(71) - (15)(20)}{(5)(55) - (15)^2}$$

$$= 1.1.$$

Therefore, the best fitted straight line is

$$y = a + bx = 0.7 + 1.1x. \qquad \textit{Answer}$$

The following Figure 2.2 shows plot of the given data and the corresponding fitted straight line.

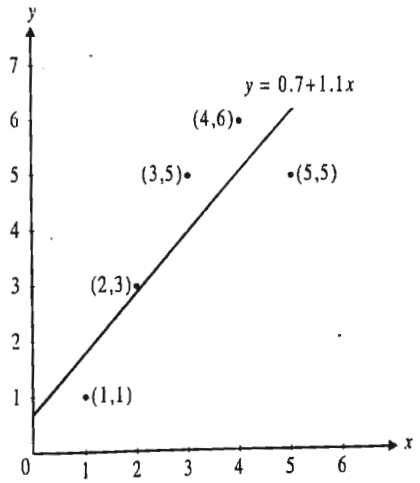


Figure 2.2

2.2.2 Fitting of a Second Degree Curve Suppose the equation of a second degree curve of the form $y = a + bx + cx^2$ is to be fitted to the n-data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n); n \geq 3,$$

where a, b, c are unknown coefficients.

Applying method of least squares, the values of a, b and c are so determined that they minimise

$$D = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2.$$

This will be so, if

$$\frac{\partial D}{\partial a} = 0, \frac{\partial D}{\partial b} = 0, \frac{\partial D}{\partial c} = 0,$$

which implies (by following the similar procedure as in Section 2.2.1) ***normal equations*** as

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \quad \ldots(2.6)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \quad \ldots(2.7)$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \quad \ldots(2.8)$$

Solving these equations, we get a, b and c. Substituting these values in the equation of a second degree curve $y = a + bx + cx^2$, the required best second degree curve fit to the given data is obtained.

Example 2.2

Using method of least squares, find the best fitting second degree curve to the following data.

x :	1	2	3	4
y :	6	11	18	27

Solution

Here, $n = 4$. Let $y = a + bx + cx^2$ be the required second degree curve. In order to find a, b, c, let us first calculate the following table.

	x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
	1	6	1	1	1	6	6
	2	11	4	8	16	22	44
	3	18	9	27	81	54	162
	4	27	16	64	256	108	432
Σ	10	62	30	100	354	190	644

Using (2.6) – (2.8),

$$62 = 4a + 10b + 30c$$
$$190 = 10a + 30b + 100c$$
$$644 = 30a + 100b + 354c$$

By partial pivoting procedure, the given system can be rewritten as

$$30a + 100b + 354c = 644$$
$$10a + 30b + 100c = 190$$
$$4a + 10b + 30c = 62$$

The augmented matrix is

$$\begin{bmatrix} 30 & 100 & 354 & : & 644 \\ 10 & 30 & 100 & : & 190 \\ 4 & 10 & 30 & : & 62 \end{bmatrix}$$

Operating $R_1(1/2), R_2(1/2), R_3(1/2)$, we get

$$\sim \begin{bmatrix} 15 & 50 & 177 & : & 322 \\ 5 & 15 & 50 & : & 95 \\ 2 & 5 & 15 & : & 31 \end{bmatrix}$$

Operating $R_{12}(-1/3), R_{13}(-2/15)$, we get

$$\sim \begin{bmatrix} 15 & 50 & 177 & : & 322 \\ 0 & -\frac{5}{3} & -9 & : & -\frac{37}{3} \\ 0 & -\frac{5}{3} & -\frac{129}{15} & : & -\frac{179}{15} \end{bmatrix}$$

Operating $R_{23}(-1)$, we get

$$\sim \begin{bmatrix} 15 & 50 & 177 & : & 322 \\ 0 & -\frac{5}{3} & -9 & : & -\frac{37}{3} \\ 0 & 0 & \frac{2}{5} & : & \frac{2}{5} \end{bmatrix}$$

By back substitution,

$$\frac{2}{5}c=\frac{2}{5}\Rightarrow c=1.$$

$$-\frac{5}{3}b-9c=-\frac{37}{3}\Rightarrow -\frac{5}{3}b-9(1)=-\frac{37}{3}$$

$$\Rightarrow -\frac{5}{3}b=-\frac{37}{3}+9$$

$$\Rightarrow -\frac{5}{3}b=-\frac{10}{3}$$

$$\Rightarrow b=2.$$

$$15a+50b+177c=322\Rightarrow 15a+50(2)+177(1)=322$$

$$\Rightarrow 15a=45$$

$$\Rightarrow a=3.$$

Therefore, the best fitted second degree curve is

$$y=3+2x+x^2.$$ ***Answer***

2.2.3 Fitting of an Exponential Curve Suppose the exponential curve of the form $y=ae^{bx}$ is to be fitted to the n-data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

where a, b are unknown coefficients.

Taking common logarithm (base 10) on both sides of $y=ae^{bx}$, we get

$$\log_{10} y=\log_{10} a+bx\log_{10} e.$$

Let

$$Y=\log_{10} y,\ A=\log_{10} a \text{ and } B=b\log_{10} e,$$

then above equation becomes

$$Y=A+Bx.$$

Above equation is in linear form of x and y since $Y=\log_{10} y$ is known. The normal equations using (2.2), (2.3) becomes

$$\sum_{i=1}^{n} Y_i = An+B\sum_{i=1}^{n} x_i \qquad \ldots(2.9)$$

$$\sum_{i=1}^{n} x_iY_i = A\sum_{i=1}^{n} x_i+B\sum_{i=1}^{n} x_i^2 \qquad \ldots(2.10)$$

From the above equations, A, B can be found and consequently

$$a=\text{antilog}\,A,\ b=\frac{B}{\log_{10} e}$$

can be calculated, which on substituting in the equation of an exponential curve $y=ae^{bx}$, the required best exponential fit to the given data is obtained.

Note One can similarly fit the curve $y=ab^x$.

Example 2.3

Fit the curve $y=ae^{bx}$ to the following data.

x	:	0	2	4
y	:	5.012	10	31.62

Solution

Here, $n=3$. Let us first calculate the following table.

	x_i	y_i	$Y_i=\log_{10} y_i$	x_i^2	x_iY_i
	0	5.012	0.70	0	0
	2	10	1	4	2
	4	31.62	1.50	16	6
Σ	6	-	3.2	20	8

Using equations (2.9) and (2.10),

$$3.2=3A+6B \qquad \ldots(i)$$

$$8=6A+20B \qquad \ldots(ii)$$

Multiplying (i) by (–2) and add to (ii), we get

$$1.6=8B\Rightarrow B=0.2.$$

Using (ii),

$$8=6A+20(0.2)\Rightarrow A=0.67.$$

Therefore,

$$a=\text{antilog}\,A\approx\text{antilog}\,(0.67)\approx 4.68.$$

$$b=\frac{B}{\log_{10} e}\approx\frac{0.2}{0.4343}\approx 0.46.$$

Therefore, the best fitted exponential curve is

$$y = 4.68e^{0.46x}.$$ **Answer**

2.2.4 Fitting of a Geometric (Power) Curve Suppose the geometric curve of the form $y = ax^b$ is to be fitted to the n-data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

where a and b are unknown coefficients.

Taking common logarithm (base 10) on both sides of $y = ax^b$, we get

$$\log_{10} y = \log_{10} a + b \log_{10} x.$$

Let

$$Y = \log_{10} y,\ A = \log_{10} a \text{ and } X = \log_{10} x,$$

then above equation becomes

$$Y = A + bX.$$

Above equation is in linear form of x and y since $Y = \log_{10} y$ and $X = \log_{10} x$ are known. The normal equations using (2.2) and (2.3) becomes

$$\sum_{i=1}^{n} Y_i = An + b\sum_{i=1}^{n} X_i, \quad \ldots(2.11)$$

$$\sum_{i=1}^{n} X_i Y_i = A\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \quad \ldots(2.12)$$

From the above equations, A, b can be found and consequently

$$a = \text{antilog } A$$

can be calculated. Substituting these values of a and b in the equation of a geometric curve $y = ax^b$, the required best geometric fit to the given data is obtained.

Example 2.4

Fit the curve $y = ax^b$ to the following data.

x :	61	26	7	2.6
y :	350	400	500	600

Solution

Here, $n = 4$. Let us first calculate the following table.

x_i	y_i	$X_i = \log_{10} x_i$	$Y_i = \log_{10} y_i$	X_i^2	X_iY_i
61	350	1.7853	2.5441	3.187	4.542
26	400	1.4150	2.6021	2.002	3.682
7	500	0.8451	2.6990	0.714	2.281
2.6	600	0.4150	2.7782	0.172	1.153
Σ		4.4604	10.6234	6.075	11.658

Using equations (2.11) and (2.12),

$$10.6234 = 4A + 4.4604b \quad \ldots(i)$$

$$11.658 = 4.4604A + 6.075b \quad \ldots(ii)$$

Solving these equations, we get

$$A \approx 2.845,\ b \approx -0.1697.$$

Therefore,

$$a = \text{antilog } A \approx \text{antilog } (2.845) \approx 699.8.$$

Therefore, the best fitted geometric curve is

$$y = 699.8x^{-0.1697}.$$ **Answer**

➢ **Special Case** When the data is very large, then for the sake of convenience and ease of calculations, it is sometimes advisible to change the origin and scale using substitution

$$X = \frac{x - A}{h} \text{ and } Y = \frac{y - B}{h},$$

where A and B are the assumed means (or middle values) of x and y series, respectively and h is the width of the interval.

Example 2.5

Determine the equation of a straight line which best fits the following data.

x :	2003	2004	2005	2006	2007
y :	35	56	79	80	40

Solution

Here, $n = 5$. Let $y = a + bx$ be the required straight line fit. Let us first calculate the following table.

x_i	y_i	$X_i = \frac{x_i - 2005}{1}$	X_i^2	X_iY_i
2003	35	−2	4	−70
2004	56	−1	1	−56

2005	79	0	0	0
2006	80	1	1	80
2007	40	2	4	80
Σ	290	0	10	34

Using equations (2.2) and (2.3),

$$5a + 0b = 290 \quad \text{...(i)}$$
$$0a + 10b = 34 \quad \text{...(ii)}$$

Solving these equations, we get

$$a = \frac{290}{5} = 58 \text{ and } b = \frac{34}{10} = 3.4.$$

Therefore, the best fitted straight line is

$$y = a + bX = 58 + 3.4X.$$

Replacing X by $(x - 2005)$, we get

$$y = 58 + 3.4(x - 2005) = -6759 + 3.4x. \quad \textit{Answer}$$

2.3 Short Questions

Example 2.6

Define curve fitting. *[GTU, May 2016]*

Solution

The process of finding the equation of the curve of best fit to the given data points, which may be most suitable for predicting the unknown values, is known as curve fitting. *Answer*

Example 2.7

What is meant by the curve of best fit ?

[GTU, June 2017 - Comp.]

Solution

The curve of best fit is that curve for which the sum of squares of errors is minimum.

Answer

Review Exercises

01. Fit a straight line to the following data.

x	:	1	2	3	4	5
y	:	3	4	5	6	8

02. Fit a straight line to the following data. Using this equation find the value of y when $x = 2.4$.

x	:	1	2	3	4	5	6	7
y	:	0.5	2.5	2.0	4.0	3.5	6.0	5.5

[GTU, May 2017]

03. If P is the pull required to lift a load W by means of a pulley block, find a linear approximation of the form $P = mW + c$ connecting P and W, using the following data.

P	:	13	18	23	27
W	:	51	75	102	119

[GTU, June 2017 - Comp.]

04. Fit a second degree curve to the following data.

x	:	1	2	3	4	5	6	7	8	9
y	:	2	6	7	8	10	11	11	10	9

05. Fit a least square geometric curve $y = ax^b$ to the following data.

x	:	1	2	3	4	5
y	:	0.5	2	4.5	8	12.5

06. Fit a second degree parabola $y = a + bx + cx^2$ to the following data.

x	:	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	:	1.2	1.4	1.9	2.4	2.8	3.3	4.2

[GTU, June 2017 - Comp.]

Answers

Review Exercises

01. $y = 1.6 + 1.2x$ **02.** $y = 0.0714 + 0.8392x$, $y(2.4) = 2.0854$.

03. $P = 0.2309\,W + 0.2186$. **04.** $y = -0.2673x^2 + 3.523x - 0.9283$ **05.** $a = 0.5012$, $b = 1.9977$

06. $y = 0.8353 + 0.1932\,x + 0.1571\,x^2$.

Chapter 3
Correlation and Regression

3.1 Introduction

Correlation and regression are statistical methods that are commonly used to compare two or more variables.

→ *For example,* comparison between income and expenditure, price and demand, etc.

Correlation measures the ***association*** between two (or more) variables and quantitates the ***strength*** of their relationship. It evaluates only the existing data.

Regression means average relationship between two (or more) variables and this relationship is used to ***estimate*** (or ***predict***) the most likely values of one variable for specified values of the other variable(s). Mathematically, regression uses the existing data to define a mathematical equation which can be used to ***estimate*** (or ***predict***) the value of one variable based on the value of one or more other variables. Therefore, it can be used to extrapolate between the existing data. The regression equation can therefore be used to predict the outcome of observations not previously seen or tested.

The following articles devoted to the detailed study of correlation and regression.

3.2 Correlation

Correlation is a statistical measure for finding out ***degree*** or ***strength*** of association between two (or more) variables. The relationship between the two variables may be linear or nonlinear. If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable then correlation is said to be linear, otherwise it is called nonlinear. The linear correlation is measured by ***correlation coefficient*** or ***coefficient of correlation*** (r), which measures their degree (or strength) of linear relationship between two quantitative variables. A significant advantage of the correlation coefficient is that it does not depend on the units of the variables and can therefore be used to compare any two variables

regardless of their units. Two variables are said to be uncorrelated, if one variable remain unaffected by any change in the other variable.

➢ **Measures of Correlation** Once again two variables X and Y are said to be correlated, if they are so related that variations in the magnitude of one variable tend to be accompained by variations in the magnitude of the other variable. Correlation between two variables X and Y can be calculated by any one of the following methods.

1. Two-way frequency table
2. Scatter diagram
3. Covariance method of Karl Pearson's product moment method
4. Spearman's rank correlation method
5. Concurrent deviation method.

Here, we will take the idea of methods (2), (3) and (4).

3.2.1 Scatter Diagram In this case the first essential step for calculating correlation coefficient is to plot the observations in a ***scattergram*** or ***scatter plot*** to visually evaluate the data for a patential relationship or the presence of outlying values. The given statistical data is plotted as points on a rectangular Cartesian coordinate system taking one independent variable (say X) along horizontal axis while other dependent variable (say Y) along the vertical axis. Then we have the following interpretation regarding their nature and strength of correlation.

(a) **Positive correlation** A positive correlation between X and Y occurs when increase in X tend to increase in Y. The line corresponding to the scatter plot is an increasing line. In this case, $0 < r < 1$ (refer Figure 3.1).

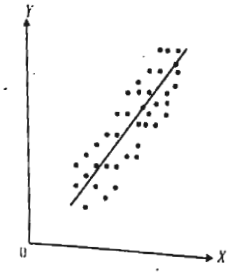


Figure 3.1 Positive correlation $(0 < r < 1)$

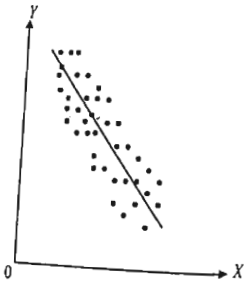


Figure 3.2 Negative correlation $(-1 < r < 0)$

(b) **Negative correlation** A negative correlation between X and Y occurs when increase in X tend to decrease in Y. The line corresponding to the scatter plot is a decreasing line. In this case, $-1 < r < 0$ (refer Figure 3.2).

(c) **Perfect positive correlation** A perfect positive correlation between X and Y occurs when there is a functional dependency between them. Moreover, as X increases, Y increases by the same amount as X and it would be concluded that X is responsible for 100% of the change in Y. In this case, all the points are in a straight line and $r = 1$ (refer Figure 3.3). Such correlation is seldom encountered.

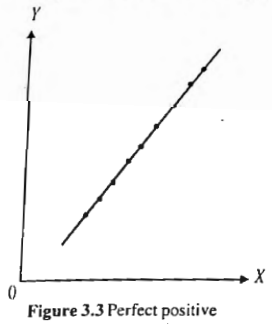


Figure 3.3 Perfect positive correlation $(r = 1)$

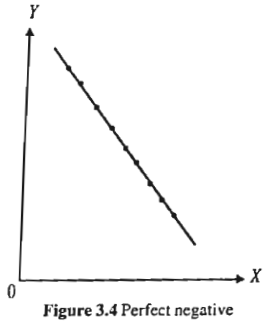


Figure 3.4 Perfect negative correlation $(r = -1)$

(d) **Perfect negative correlation** A perfect negative correlation between X and Y occurs when there is a functional dependency between them. Moreover, as X increases, Y decreases by the same amount as X and it would be concluded that X is responsible for 100% of the change in Y. In this case, all the points are in a straight line and $r = -1$ (refer Figure 3.4). Such correlation is seldom encountered.

(e) **Strong positive correlation** A strong positive correlation between X and Y occurs when the data points are located closer to one another on the line (refer Figure 3.5). In this case r is close to 1.

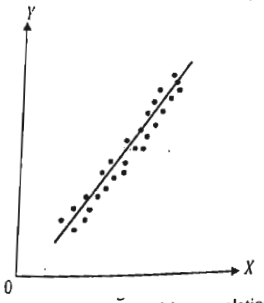


Figure 3.5 Strong positive correlation

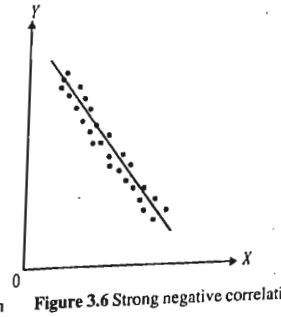


Figure 3.6 Strong negative correlation

Similarly, we can define strong negative correlation (refer Figure 3.6). In this case r is close to -1.

(f) **Weak positive correlation** A weak positive correlation between X and Y occurs when the data points are located farther apart to one another on the line (refer Figure 3.7).

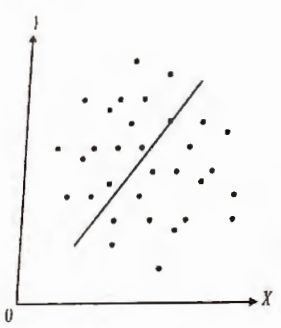


Figure 3.7 Weak positive correlation

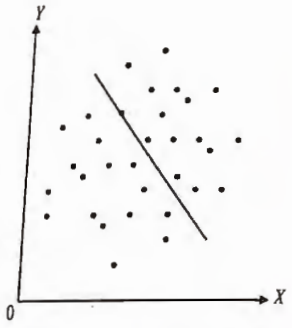


Figure 3.8 Weak negative correlation

Similarly, we can define weak negative correlation (refer Figure 3.8).

(g) **No correlation** If X and Y are not related at all; that is, when there is no linear dependency between the variables. In this case $r = 0$ or close to 0. (refer Figure 3.9).

Figure 3.9 No correlation

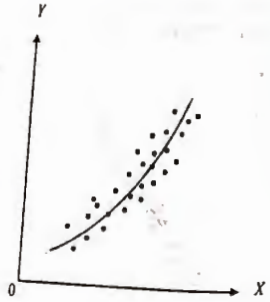


Figure 3.10 Nonlinear correlation

(h) **Nonlinear correlation** A nonlinear correlation between X and Y occurs when the data points assume a nonlinear or curved association, and will have a falsely low r value (refer Figure 3.10).

Before proceeding further, let us note some other properties of linear correlation coefficient.

(i) The linear correlation coefficient is always between (–1) and +1 (both inclusive). That is, $-1 \le r \le 1$.

(ii) The linear correlation coefficient is a unitless measure of association. So the units of measure of X and Y play no role in the interpretation of r.

3.2.2 Karl Pearson's Method (Covariance Method) Before we introduce this method, let us define the concept of ***covariance***.

➢ **Definition of Covariance** Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n-pair of observations on two variables X and Y, then the covariance of X and Y is denoted and defined as

$$\text{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad \ldots(3.1)$$

where $\bar{x}$ and $\bar{y}$ are arithmetic means of X and Y series, respectively; that is,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}.$$

Covariance indicates the join variations between the two variables X and Y.

➢ **Karl Pearson's Method** It is a mathematical measure to find the degree of linear relationship between two variables. The formula for calculating correlation coefficient (r) between X and Y by this method is given by

$$r = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} \qquad \ldots(3.2)$$

where σ_x, σ_y are the standard deviations of X and Y, respectively; that is,

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n} - \bar{x}^2},$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2}.$$

...(3.3)

The above formula (3.2) for correlation coefficient may be written in different forms.

(a) **Direct method** If

$x = x_i - \bar{x}$ is deviation of x_i from its mean $\bar{x}$ and

$y = y_i - \bar{y}$ is deviation of y_i from its mean $\bar{y}$,

then (3.2) becomes

$$r = \frac{\sum_{i=1}^{n} xy}{\sqrt{\sum_{i=1}^{n} x^2 \times \sum_{i=1}^{n} y^2}}.$$

...(3.4)

From (3.1),

$$\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})$$

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

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

$$= \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}.$$

Therefore,

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right).$$

...(3.5)

Again, (3.2) becomes

$$r = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2}\sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2}}$$

or

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}.$$

...(3.6)

Example 3.1

Find the correlation coefficient from the following data.

X	:	1	2	3	4	5	6	7
Y	:	6	8	11	9	12	10	14

Solution

Here,

x_i	y_i
1	6
2	8
3	11
4	9
5	12

	6	10
	7	14
Σ	28	70

Using (1.1),

$$\bar{x} = \frac{\sum_{i=1}^{7} x_i}{7} = \frac{28}{7} = 4, \quad \bar{y} = \frac{\sum_{i=1}^{7} y_i}{7} = \frac{70}{7} = 10. \qquad (\text{Since } n = 7)$$

x_i	y_i	$x = x_i - \bar{x}$ $= x_i - 4$	$y = y_i - \bar{y}$ $= y_i - 10$	x^2	y^2	xy
1	6	−3	−4	9	16	12
2	8	−2	−2	4	4	4
3	11	−1	1	1	1	−1
4	9	0	−1	0	1	0
5	12	1	2	1	4	2
6	10	2	0	4	0	0
7	14	3	4	9	16	12
Σ	28	70		28	42	29

Using (3.4),

$$r = \frac{\sum_{i=1}^{7} xy}{\sqrt{\sum_{i=1}^{7} x^2 \times \sum_{i=1}^{7} y^2}}$$

$$= \frac{29}{\sqrt{28 \times 42}}$$

$$\approx 0.846.$$

Thus, there is a strong positive correlation between X and Y. ***Answer***

(b) **Short-cut method** The above direct method for calculating r is not convenient in the following cases.

(i) When the terms of the series x and y are big and the calculations of $\bar{x}$ and $\bar{y}$ become difficult, or

(ii) When the means $\bar{x}$ and $\bar{y}$ are not integer.

In this case the following formula of assumed mean can be applicable.

$$r = \frac{\Sigma dx\,dy - \frac{(\Sigma dx)(\Sigma dy)}{n}}{\sqrt{\Sigma dx^2 - \left(\frac{\Sigma dx}{n}\right)^2}\sqrt{\Sigma dy^2 - \left(\frac{\Sigma dy}{n}\right)^2}} \qquad \ldots(3.7)$$

where

$dx = x - A$, A is assumed mean of x-series,
$dy = y - B$, B is assumed mean of y-series,
n is number of observations of x and y.

Example 3.2

Find the coefficient of correlation from the following data.

X :	1	2	3	4	5	6	7	8	9	10
Y :	46	42	38	34	30	26	22	18	14	10

Solution

Here,

x_i	y_i
1	46
2	42
3	38
4	34
5	30
6	26
7	22
8	18
9	14
10	10
Σ 55	280

Using (1.1),

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{55}{10} = 5.5, \quad \bar{y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{280}{10} = 28. \quad \text{(Since } n = 10\text{)}$$

Let $A = 5$, $B = 30$.

X	Y	$dx = X - 5$	$dy = Y - 30$	$(dx)^2$	$(dy)^2$	$dx\,dy$
1	46	−4	16	16	256	−64
2	42	−3	12	9	144	−36
3	38	−2	8	4	64	−16
4	34	−1	4	1	16	−4
5	30	0	0	0	0	0
6	26	1	−4	1	16	−4
7	22	2	−8	4	64	−16
8	18	3	−12	9	144	−36
9	14	4	−16	16	256	−64
10	10	5	−20	25	400	−100
Σ 55	280	5	−20	85	1360	−340

Using (3.7),

$$r = \frac{\sum dx\,dy - \frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\sum dx^2 - \left(\frac{\sum dx}{n}\right)^2}\sqrt{\sum dy^2 - \left(\frac{\sum dy}{n}\right)^2}}$$

$$= \frac{-340 - \frac{(5)(-20)}{10}}{\sqrt{85 - \left(\frac{5}{10}\right)^2}\sqrt{1360 - \left(\frac{-20}{10}\right)^2}}$$

$$= \frac{-330}{\sqrt{85 - \frac{1}{4}}\sqrt{1360 - 4}}$$

$$= \frac{-330}{\frac{\sqrt{339}}{2}\sqrt{1356}}$$

$$= -\frac{660}{678}$$

$$\approx -0.97. \qquad \textit{Answer}$$

3.2.3 Spearman's Rank Correlation Method The method we have so far discussed depends on the magnitudes of the variables. But there are situations where magnitude of the variable is not possible.

→ ***For example,*** we cannot measure beauty and intelligence quantitatively, but it is possible to rank the individuals in order.

Rank correlation is based on the rank or the order of the variables and not on the magnitude of the variables. Here, the individuals are arranged in order of proficiency. If the ranks are assigned to the individuals range from 1 to n, then the correlation coefficient between two series of ranks is called ***rank correlation coefficient***. Edward Spearman's formula for rank coefficient of correlation (R) is given by

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\sum d^2}{n^3 - n}, \qquad \text{...(3.8)}$$

where d is the difference between the ranks of the two series and n is the number of individuals in each series.

Example 3.3

Calculate the rank correlation coefficient, if two judges in a beauty contest ranked the entries as follows.

Judge X	:	1	2	3	4	5
Judge Y	:	5	4	3	2	1

Solution

X (Rank R_1)	Y (Rank R_2)	$d = R_1 - R_2$	d^2
1	5	-4	16
2	4	-2	4
3	3	0	0
4	2	2	4
5	1	4	16
Σ			40

Using (3.8),

$$R = 1 - \frac{6\sum d^2}{n^3 - n}$$

$$= 1 - \frac{6(40)}{(5)^3 - 5} \qquad \text{(Since } n = 5\text{)}$$

$$= 1 - \frac{240}{120}$$

$$= -1.$$

Thus, there is perfect negative correlation between two judges. ***Answer***

Example 3.4

Ten students got the following percentage of marks in mathematics and statistics.

Student (Roll No.)	1	2	3	4	5	6	7	8	9	10
Marks in mathematics	78	36	98	25	75	82	90	62	65	39
Marks in statistics	84	51	91	60	68	62	86	58	53	47

Calculate the correlation coefficient.

Solution

Roll No.	Mathematics		Statistics		$d = R_1 - R_2$	d^2
	Marks	Rank (R_1)	Marks	Rank (R_2)		
1	78	4	84	3	1	1
2	36	9	51	9	0	0
3	98	1	91	1	0	0
4	25	10	60	6	4	16
5	75	5	68	4	1	1
6	82	3	62	5	-2	4
7	90	2	86	2	0	0
8	62	7	58	7	0	0
9	65	6	53	8	-2	4
10	39	8	47	10	-2	4
Σ						30

Using (3.8),

$$R = 1 - \frac{6\sum d^2}{n^3 - n} \qquad \text{(Since } n = 10\text{)}$$

$$= 1 - \frac{6(30)}{(10)^3 - 10}$$

$$= 1 - \frac{180}{990}$$

$$\approx 0.82.$$ ***Answer***

Example 3.5

In a college, IT department has arranged one competition for students to develop an efficient program to solve a problem. Ten students took part in the competition and ranked by two judges given in the following table. Find the degree of agreement between the two judges using rank correlation coefficient.

First Judge	:	3	5	8	4	7	10	2	1	6	9
Second Judge	:	6	4	9	8	1	2	3	10	5	7

[GTU, June 2016]

Solution

First Judge (Rank R_1)	Second Judge (Rank R_2)	$d = R_1 - R_2$	d^2
3	6	−3	9
5	4	1	1
8	9	−1	1
4	8	−4	16
7	1	6	36
10	2	8	64
2	3	−1	1
1	10	−9	81
6	5	1	1
9	7	2	4
Σ			214

Using (3.8),

$$R = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

$$= 1 - \frac{6(214)}{(10)^3 - 10} \qquad \text{(Since } n = 10\text{)}$$

$$= 1 - \frac{1284}{990}$$

$$\approx -0.297. \qquad \textit{Answer}$$

3.3 Regression

In introduction Section 3.1, we have discussed little about regression.

Simple linear regression is the most commonly used technique for determining about how one variable of interest (the ***response variable*** or ***regressed variable*** or ***explained variable*** or ***dependent variable***) is affected by changes in another variable (the ***explanatory variable*** or ***regressor*** or ***predictor*** or ***independent variable***). This can be done with the help of a regression line, which shows the average value of one variable for a given value of other variable. The best average value of one variable associated with the given value of the other variable may be estimated or predicted by means of an equation known as ***regression equation***.

Mathematically, regression describes the dependence of the Y variable on the X variable and construct an equation which can be used to predict any value of Y for any value of X.

Unlike correlation, however, regression is not scale independent and the derived regression equation depends on the units of each variable involved. As with correlation, regression assumes that each of the variables is normally distributed with equal variance. In deriving the regression equation, the best fit line through the scatter plot data is considered such that the sum of the squared residuals is minimized; that is, least square method is adopted.

The regression line may be linear in which case the relationship between the variables fits a straight line, or nonlinear in which case a polynomial equation is fitted.

➢ **Types of Regression**

(a) **Simple regression** When the regression analysis confined the study of only two variables at a time, then it is called simple regression.

(b) **Multiple regression** When the regression analysis confined the study of more than two variables at a time, then it is called multiple regression.

3.3.1 Lines of Regression

➢ **Regression Equation of Y on X** In linear regression, if we fit a straight line of the type $Y = a + bX$ to the given data by the method of least squares, then we get regression equation of Y on X.

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the n-pair of observations on two variables X and Y, and let the straight line to be fitted to these data is

$$Y = a + bX \qquad \ldots(3.9)$$

Then by following method of least squares, we get normal equations (refer equations (2.2) and (2.3)) as

$$\Sigma Y = na + b\Sigma X, \qquad \ldots(3.10)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \qquad \ldots(3.11)$$

Divide (3.10) by n, we get

$$\frac{\Sigma Y}{n} = a + b\frac{\Sigma X}{n}$$

or

$$\bar{Y} = a + b\bar{X}, \qquad \ldots(3.12)$$

Subtracting equation (3.12) from equation (3.9), we obtain

$$Y-\bar{Y}=b(X-\bar{X}). \qquad ...(3.13)$$

Multiplying (3.10) by $\sum X$, and (3.11) by n and then subtracting, we have

$$(\sum X)(\sum Y)-n\sum XY=b(\sum X)^2-nb\sum X^2$$

$$\Rightarrow b[n\sum X^2-(\sum X)^2]=n\sum XY-(\sum X)(\sum Y)$$

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

$$\Rightarrow b=\frac{\frac{\sum XY}{n}-\frac{(\sum X)(\sum Y)}{n^2}}{\frac{\sum X^2}{n}-\left(\frac{\sum X}{n}\right)^2}. \qquad \text{(Dividing by } n^2\text{)}$$

Therefore, using (3.5) and (3.3), we obtain

$$b=\frac{\text{cov}(X,Y)}{\sigma_x^2}. \qquad ...(3.14)$$

Replacing b by b_{yx} in (3.13), the regression equation of Y on X is

$$Y-\bar{Y}=b_{yx}(X-\bar{X}), \qquad ...(3.15)$$

where b_{yx} is given by (3.14).

Equation (3.15) is the regression equation of Y on X and it is used to estimate the most likely values of Y for given values of X.

➢ **Regression Equation of X on Y** By following above procedure, the regression equation of X on Y obtained as

$$X-\bar{X}=b_{xy}(Y-\bar{Y}), \qquad ...(3.16)$$

where

$$b_{xy}=\frac{\text{cov}(X,Y)}{\sigma_y^2};\ \sigma_y^2=\frac{\sum Y^2}{n}-\left(\frac{\sum Y}{n}\right)^2. \qquad ...(3.17)$$

Equation (3.16) is the regression equation of X on Y and it is used to estimate the values of X when the values of Y are known.

Note The numbers b_{yx} and b_{xy} are called ***regression coefficients*** of the two regression lines (3.15) and (3.16), respectively. Using (3.2), it can also be presented as

$$b_{yx}=r\frac{\sigma_y}{\sigma_x} \text{ and } b_{xy}=r\frac{\sigma_x}{\sigma_y}, \qquad ...(3.18)$$

where σ_x and σ_y are standard deviations of X and Y series, respectively.

➢ **Properties of Regression Coefficients**

(1) The coefficient of correlation is the geometric mean of the coefficients of regression; that is,

$$r=\sqrt{b_{yx}\cdot b_{xy}} \qquad ...(3.19)$$

(2) If $b_{yx}=\frac{1}{b_{xy}}$, then the two lines are coincide.

(3) If one of the regression coefficient is greater than unity, then the other is less than unity; that is, if $b_{yx}>1$ then $b_{xy}<1$ and vice versa.

(4) Arithmetic mean of the regression coefficients is greater than the correlation coefficient.

(5) Regression coefficients are independent of change of origin but not of scale.

(6) Both regression coefficients will have the same sign.

(7) The sign of correlation coefficient is same as that of regression coefficients. If $r=0$, then both b_{yx} and b_{xy} are also zero.

(8) The regression lines always intersect at $(\bar{X},\bar{Y})$. The slope of the lines are respectively b_{yx} and $1/b_{xy}$.

(9) If $r=0$, then regression lines are perpendicular.

Example 3.6

The scores of 12 students in their mathematics (X) and statistics (Y) classes are as follows.

Mathematics (X)	:	2	3	4	4	5	6	6	7	7	8	10	10
Statistics (Y)	:	1	3	2	4	4	4	6	4	6	7	9	10

Find the regression line of Y on X.

Solution

Here,

x_i	y_i	x_i^2	$x_i y_i$
2	1	4	2
3	3	9	9
4	2	16	8
4	4	16	16
5	4	25	20
6	4	36	24
6	6	36	36
7	4	49	28
7	6	49	42
8	7	64	56
10	9	100	90
10	10	100	100
Σ	72	60	504	431

Using (1.1),

$$\bar{X} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{72}{12} = 6, \quad \bar{Y} = \frac{\sum_{i=1}^{12} y_i}{12} = \frac{60}{12} = 5.$$

Using (3.5),

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{12} x_i y_i}{12} - (\bar{X})(\bar{Y})$$

$$= \frac{431}{12} - (6)(5)$$

$$= \frac{71}{12}.$$

Using (3.3),

$$\sigma_x^2 = \frac{\sum_{i=1}^{12} x_i^2}{12} - (\bar{X})^2$$

$$= \frac{504}{12} - (6)^2$$

$$= 6.$$

Therefore, using (3.14)

$$b_{yx} = \frac{\text{cov}(X, Y)}{\sigma_x^2} = \frac{\frac{71}{12}}{6} = \frac{71}{72}.$$

Using (3.15),

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\Rightarrow Y - 5 = \left(\frac{71}{72}\right)(X - 6)$$

$$\Rightarrow Y = \frac{71}{72}X - \frac{11}{12},$$

which is the required regression line of Y on X. ***Answer***

Example 3.7

A study of the amount of rainfall and the quantity of air pollution removed produced the following data.

Daily rainfall x (0.01 cm) :	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1	7.5
Particulate removed y ($\mu g/m^3$) :	126	121	116	118	114	118	132	141	108

(a) Find the equation of the regression line to predict the particulate removed from the amount of daily rainfall.

(b) Find the amount of particulate removed when daily rainfall is $x = 4.8$ units.

[GTU, May 2016]

Solution

(a) Here,

x	y	x_i^2	y_i^2	x_iy_i
4.3	126	18.49	15876	541.8
4.5	121	20.25	14641	544.5
5.9	116	34.81	13456	684.4
5.6	118	31.36	13924	660.8
6.1	114	37.21	12996	695.4
5.2	118	27.04	13924	613.6
3.8	132	14.44	17424	501.6
2.1	141	4.41	19881	296.1
7.5	108	56.25	11664	810
Σ 45	1094	244.26	133786	5348.2

Using (1.1),

$$\overline{X} = \frac{\sum_{i=1}^{9} x_i}{9} = \frac{45}{9} = 5,\ \overline{Y} = \frac{\sum_{i=1}^{9} y_i}{9} = \frac{1094}{9} \approx 121.56.$$

Using (3.5),

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{9} x_i y_i}{9} - (\overline{X})(\overline{Y})$$

$$\approx \frac{5348.2}{9} - (5)(121.56)$$

$$\approx 594.24 - 607.8$$

$$= -13.56.$$

Using (3.3),

$$\sigma_x^2 = \frac{\sum_{i=1}^{9} x_i^2}{9} - (\overline{X})^2$$

$$= \frac{244.26}{9} - (5)^2$$

$$= 27.14 - 25$$

$$= 2.14.$$

Therefore, using (3.14)

$$b_{yx} = \frac{\text{cov}(X, Y)}{\sigma_x^2} = \frac{-13.56}{2.14} \approx -6.336.$$

Using (3.15),

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

$$\Rightarrow Y - 121.56 = (-6.336)(X - 5)$$

$$\Rightarrow Y - 121.56 = -6.336X + 31.68$$

$$\Rightarrow Y = -6.336X + 153.24. \quad \text{...(i)}$$

Answer (a)

(b) At $x = 4.8$, we have using (i)

$$Y = -6.336(4.8) + 153.24 = 122.83.$$

The amount of particulate removed is approximately 122.83 units when daily rainfall is $x = 4.8$ units. ***Answer (b)***

Example 3.8

From the following data, obtain the two regression lines and the correlation coefficients.

x	:	100	98	78	85	110	93	80
y	:	85	90	70	72	95	81	74

[GTU, June 2016]

Solution

Here,

x_i	y_i	x_i^2	y_i^2	x_iy_i
100	85	10000	7225	8500
98	90	9604	8100	8820
78	70	6084	4900	5460
85	72	7225	5184	6120

	110	95	12100	9025	10450
	93	81	8649	6561	7533
	80	74	6400	5476	5920
Σ	644	567	60062	46471	52803

Using (1.1),

$$\overline{X} = \frac{\sum_{i=1}^{7} x_i}{7} = \frac{644}{7} = 92, \quad \overline{Y} = \frac{\sum_{i=1}^{7} y_i}{7} = \frac{567}{7} = 81.$$

Using (3.5),

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{7} x_i y_i}{7} - (\overline{X})(\overline{Y})$$

$$= \frac{52803}{7} - (92)(81)$$

$$\approx 91.29.$$

Using (3.3),

$$\sigma_x^2 = \frac{\sum_{i=1}^{7} x_i^2}{7} - (\overline{X})^2$$

$$= \frac{60062}{7} - (92)^2$$

$$\approx 116.29.$$

$$\sigma_y^2 = \frac{\sum_{i=1}^{7} y_i^2}{7} - (\overline{Y})^2$$

$$= \frac{46471}{7} - (81)^2$$

$$\approx 77.71.$$

Therefore, using (3.14) and (3.17)

$$b_{yx} = \frac{\text{cov}(X, Y)}{\sigma_x^2} \approx \frac{91.29}{116.29} \approx 0.79,$$

$$b_{xy} = \frac{\text{cov}(X, Y)}{\sigma_y^2} \approx \frac{91.29}{77.71} \approx 1.17.$$

Using (3.15), the regression line of Y on X is given by

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

$$\Rightarrow Y - 81 = 0.79(X - 92)$$

$$\Rightarrow Y = 0.79X + 8.32.$$

Using (3.16), the regression line of X on Y is given by

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

$$\Rightarrow X - 92 = 1.17(Y - 81)$$

$$\Rightarrow X = 1.17Y - 2.77.$$

Using (3.19), the correlation coefficient becomes

$$r = \sqrt{b_{yx} \times b_{xy}}$$

$$\approx \sqrt{0.79 \times 1.17}$$

$$\approx 0.9614.$$

3.4 Short Questions

Example 3.9

If the value of the coefficient of correlation is negative, then what does it signify about the relationship of two variables? *[GTU, May 2016]*

Solution

If the value of the coefficient of correlation is negative, then it signifies that the two variables are inversely proportional to each other; that is, increase in one variable tend to decrease in another. ***Answer***

Example 3.10

Which is the point of intersection of the regression line of y on x and regression line of x on y. *[GTU, May 2016 & May 2017]*

Solution

The regression lines (y on x, or x on y) always intersect at $(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ indicates mean. *Answer*

Example 3.11

If the values of the regression coefficients $b_{xy} = 0.8$ and $b_{yx} = 0.4$, then find the correlation coefficient r.

[GTU, May 2017]

Solution Using (3.19)

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0.4 \times 0.8} \approx 0.5657.$$ *Answer*

Example 3.12

Rank correlation method is used for which type of data?

[GTU, May 2017]

Solution

Rank correlation method is used for qualitative type data. *Answer*

Example 3.13

What is the range of linear correlation coefficient?

Solution

The range of linear correlation coefficient r is always $-1 \le r \le 1$. *Answer*

Example 3.14

For the two data sets represented by x and y, write the regression coefficient of y on x.

Solution

The regression coefficient of y on x is

$$r\frac{\sigma_y}{\sigma_x},$$

where r is coefficient of correlation, σ_x and σ_y are standard deviations of x and y, respectively. *Answer*

Review Exercises

01. Find the correlation coefficient from the following data.

x	:	65	66	67	68	69	70	71
y	:	67	68	66	69	72	72	69

02. Using short-cut method, calculate the coefficient of correlation between x and y for the following data.

x	:	–3	–2	–1	0	1	2	3
y	:	9	4	1	0	1	4	9

03. For the following observations, find regression coefficients b_{yx} and b_{xy}, and hence find the correlation coefficient between x and y.

(4, 2), (2, 3), (3, 2), (4, 4), (2, 4)

04. The following data represents the rank of 10 students in two subject, Environmental studies (ES) and Mechanics of Solids (MoS). Find the rank correlation.

ES	:	4	5	7	8	10	1	3	6	2	9
MoS	:	3	4	7	9	10	8	6	5	2	1

[GTU, May 2017 - Civil]

05. Compute correlation coefficient between X and Y using the following data.

X	:	2	4	5	6	8	11
Y	:	18	12	10	8	7	5

[GTU, May 2017 - Civil]

06. Calculate the coefficient of correlation between the given series of data for x and y in the following table.

x	:	54	57	55	57	56	52	59
y	:	36	35	32	34	36	38	35

[GTU, June 2017 - Comp.]

Answers

Review Exercises

01. 0.67 **03.** $b_{yx} = -1/4, b_{xy} = -1/4, r = 1/4$ **04.** 0.24 **05.** –0.9203 **06.** 0.00041

Chapter 4

Probability Theory

4.1 Introduction

Let us first look at the following terminology related with the basic probability theory.

(1) **Experiment** It is defined as an operation which can produce some well defined outcomes (or results).

(2) **Random experiment** It is defined as those experiment about whose outcomes cannot be successfully predicted. Of course, we know all possible outcomes in advance.

→ ***For example,*** tossing a coin, throwing a dice.

(3) **Sample space** The set of all possible outcomes of a random experiment is called a sample space. It is generally denoted by S. The elements of a sample space are called ***sample points***.

→ ***For example,*** the sample space of a random experiment of tossing a coin is

$$S = \{H, T\},$$

where H represents head and T represents tail.

(4) **Event** Any subset of a sample space is called an event.

→ ***For example,*** $\{H\}$ is a subset of $S = \{H, T\}$, therefore it is called event.

(5) **Impossible event** The subset ϕ of a sample sapce S is called the impossible event.

(6) **Certain (or Sure) event** The subset S of a sample space S is called certain event.

(7) **Primary (or Elementary or Simple) event** A single term subset of a sample space S is called primary event; that is, the event has single possible outcome of a random experiment.

(8) **Compound event** The joint occurrence of two or more simple events is called a compound event.

(9) **Complementary event** Let S be a sample space and A be some event of S, then the subset of S consisting of those elements which are in S but not in A is called complementary event of the event A. It is denoted by the symbol A'. Thus,

$$A' = \{x \in S : x \notin A\}.$$

It should be noted that event A' occurs means event A does not occur and vice versa.

(10) **Union event** Let S be a sample space, and A and B be any two events of S, then the subset of S consisting of those elements which are either in A or in B is called the union of events A and B. It is denoted by the symbol $A \cup B$. Thus,

$$A \cup B = \{x \in S : x \in A \text{ or } x \in B\}.$$

It should be noted that event $A \cup B$ occurs means out of events A and B at least one occur.

(11) **Intersection event** Let S be a sample space, and A and B be any two events of S, then the subset of S consisting of those elements which are in A as well as in B is called the intersection of events A and B. It is denoted by the symbol $A \cap B$. Thus,

$$A \cap B = \{x \in S : x \in A \text{ and } x \in B\}.$$

It should be denoted that event $A \cap B$ occurs means both the events A and B occur.

(12) **Difference event** Let S be a sample space, and A and B be any two events of S, then the subset of S consisting of those elements which are in A but not in B is called the difference event of events A and B. It is denoted by the symbol $A - B$. Thus,

$$A - B = \{x \in S : x \in A \text{ and } x \notin B\}.$$

It should be noted that event $A - B$ occurs means A occur but B does not occur.

(13) **Mutually exclusive events** Let S be a sample space, and A and B be any two events of S. Then A and B are called mutually exclusive events, if $A \cap B = \phi$. It means that occurance of A prevents the occurance of B and vice versa.

(14) **Exhaustive events** Let S be a sample space, and A and B be any two events of S. Then A and B are called exhaustive events, if $A \cup B = S$.

(15) **Equiprobable events (Equally likely events)** Let $S = \{A_1, A_2, \ldots, A_n\}$ be a finite sample space. If

$$P(\{A_1\}) = P(\{A_2\}) = \ldots = P(\{A_n\}),$$

then the elementary events $\{A_1\}, \{A_2\}, \ldots, \{A_n\}$ are called equiprobable events.

→ ***For example,*** if we toss a coin, then head H and tail T will appear about equally often and we say that H and T are equally likely.

(16) **Favourable events or cases** The number of cases favourable to an event in an experiment is the number of outcomes which entail the happening of the event.

→ ***For example,*** in throwing a fair die (that is, regular shaped die of homogeneous material), the number of cases favourable to get a sum 5 is {(1, 4), (4, 1), (2, 3), (3, 2)}; that is, there are four elements in this subset.

4.2 Definition of Probability

The probability can be defined in following different ways.

(a) Classical or mathematical (or a priori) probability

(b) Statistical or a posteriori probability

(c) Subjective approach to probability

(d) Axiomatic or modern approach to probability

Out of above four ways, we will discuss here (a) and (d).

4.2.1 Classical or Mathematical or a Priori Probability If any experiment results in n exhaustive, mutually exclusive and equally likely outcomes, and if m of them are favourable to the occurrence of an event A, then the probability of the occurrence of an event A is denoted and defined as follows.

$$P(A) = \frac{\text{Number of favourable outcomes to the occurance of an event } A}{\text{Total number of exhaustive, mutually exclusive and equally likely outcomes}}$$

or

$$P(A) = \frac{m}{n} \qquad \ldots(4.1)$$

As number of outcomes m favourable to the event A cannot be greater than the number of exhaustive outcomes n, therefore

$$0 \le m \le n \Rightarrow 0 \le \frac{m}{n} \le 1 \Rightarrow 0 \le P(A) \le 1.$$

Case 1 When $m = 0$, then $P(A) = 0$; that is, the event A is impossible as none of the n exhaustive outcomes is favourable to the event A.

Case 2 When $m = n$, then $P(A) = 1$; that is, the event A is certain to happen in any trial as all the n exhaustive outcomes is favourable to the event A.

Example 4.1 (Fair Die)

A fair die is thrown, then find the probability of obtaining at least a 4.

Solution

Here, the six outcomes {1, 2, 3, 4, 5, 6} of a fair die are exhaustive, mutually exclusive and equally likely elementary, therefore, each outcome has probability 1/6. Let A be the event of obtaining at least a 4. Therefore,

$$A = \{4, 5, 6\}$$

and using (4.1),

$$P(A) = \frac{3}{6} = \frac{1}{2}.$$ **Answer**

Example 4.2

A bag contains 5 white and 10 black balls. Three balls are taken out at random. Find the probability that all three balls drawn are black.

Solution

There are total of 5 + 10 = 15 balls in a bag. Three balls out of 15 balls can be taken in following ways.

$$^{15}C_3 = \frac{15!}{(15-3)!\,3!} = \frac{15!}{12!\,3!} = \frac{13\times14\times15}{2\times3} = 455.$$

Now, balls are taken out at random, therefore, all these 455 ways are equally likely. Hence, the experiment results in $n = 455$ exhaustive, mutually exclusive and equally likely cases.

Let A be the event that all three balls drawn are black. There are 10 black balls and if all the three balls drawn are to be black, then it is obvious that these must be drawn from these 10 black balls. Hence, the number of cases favourable to event A is

$$^{10}C_3 = \frac{10!}{(10-3)!\,3!} = \frac{10!}{7!\,3!} = \frac{8\times9\times10}{2\times3} = 120 \text{ ways.}$$

Thus, using (4.1),

$$P(A) = \frac{120}{455} = \frac{24}{91}.$$ **Answer**

Example 4.3 (Fair Coin)

An unbiased coin is tossed 3 times. What is the probability of obtaining two heads (exactly)? ***[GTU, May 2016]***

Solution

Here, following situation happens when an unbiased coin is tossed 3 times.

First tossing	Second tossing	Third tossing	Sample points	Whether two heads exactly?
H	*H*	*H*	*HHH*	No
H	*H*	*T*	*HHT*	Yes
H	*T*	*H*	*HTH*	Yes
H	*T*	*T*	*HTT*	No
T	*H*	*H*	*THH*	Yes
T	*H*	*T*	*THT*	No
T	*T*	*H*	*TTH*	No
T	*T*	*T*	*TTT*	No

where H represents head and T represents tail.

Therefore,

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Let A be the given event of obtaining exactly two heads, then

$$A = \{HHT, HTH, THH\}.$$

Hence, using (4.1),

$$P(A) = \frac{3}{8}.$$ **Answer**

Example 4.4

Write sample space of random experiment of tossing three coins together and obtain the probability of the event that one head and two tails obtained.

Solution

Here,

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\},$$

where H represents head and T represents tail.

Let A be the event that one head and two tails obtained, then

$$A = \{HTT, THT, TTH\}.$$

Therefore, using (4.1),

$$P(A) = \frac{3}{8}.$$ **Answer**

Example 4.5

There are 5 yellow, 2 red and 3 white balls in the box. Three balls are randomly selected from the box. Find the probabilities of the following events.

(*a*) All are of different colour

(*b*) 2 yellow and 1 red colour

(*c*) All are of same colour

Solution

There are total of 10 balls in the box.

Let n be the number of ways of selecting 3 balls out of 10 balls. Therefore,

$$n = {}^{10}C_3 = \frac{10!}{(10-3)!\,3!} = \frac{10!}{7!\,3!} = \frac{8\times9\times10}{2\times3} = 120.$$

(*a*) Let A be the event defined as follows.

A : Selected all 3 balls are of different colour

Therefore, using (4.1),

$$P(A) = \frac{m}{n} = \frac{{}^5C_1 \times {}^3C_1 \times {}^2C_1}{{}^{10}C_3} = \frac{5\times3\times2}{120} = \frac{1}{4}.$$ ***Answer (a)***

(*b*) Let B be the event defined as follows.

B : Out of 3 selected balls 2 yellow and 1 red colour

Therefore, using (4.1),

$$P(B) = \frac{m}{n} = \frac{{}^5C_2 \times {}^2C_1}{{}^{10}C_3} = \frac{4\times5\times2}{2\times120} = \frac{1}{6}.$$ ***Answer (b)***

(*c*) Let C be the event defined as follows.

C : Selected all 3 balls are of same colour

Therefore, using (4.1),

$$P(C) = \frac{m}{n} = \frac{{}^5C_3 + {}^3C_3}{{}^{10}C_3} = \frac{\left(\frac{3\times4\times5}{2\times3}\right)+1}{120} = \frac{11}{120}.$$ ***Answer (c)***

4.2.2 Axiomatic Approach to Probability Let S be a finite sample space and C be a class of all subsets of S (that is, C is a power set). Let $\mathbb{R}$ be a set of real numbers. Suppose that a set function $P : C \to \mathbb{R}$ satisfies the following axioms.

Axiom 1 For every $A \in C$, $0 \le P(A) \le 1$.

Axiom 2 $P(S) = 1$.

Axiom 3 For every pair of mutually exclusive events A and B in C (that is, $A \cap B = \phi$),

$$P(A \cap B) = P(A) + P(B).$$

Then such a function P is called probability function. For event $A \in C$, $P(A)$ is called the probability of the event A.

The triplet (S, C, P) is called probability space.

Let us now review following basic theorems of probability.

Theorem 4.1 (Complementation Rule) Let S be a sample space and A be any event in S, then

$$P(A') = 1 - P(A).$$ ...(4.2)

➢ **Corollary** Probability of impossible event is zero; that is, $P(\phi) = 0$.

Theorem 4.2 Let A and B be any two events of a sample space S with $A \subset B$, then

(1) $P(B - A) = P(B) - P(A)$, (2) $P(B) \ge P(A)$. ...(4.3)

➢ **Corollary** For any two events A, B of a sample space S,

$$P(A \cap B') = P(A) - P(A \cap B).$$ ...(4.4)

Theorem 4.3 (Addition Rule for Arbitrary Events) Let S be a sample space, and A and B be any two events in S, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$ ...(4.5)

➢ **Corollary** For any three events A, B, C, it can be proved that

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$$ ...(4.6)

Theorem 4.4 (Addition Rule for Mutually Exclusive Events) Let S be a sample space and $A_1, A_2, \ldots, A_n$ are mutually exclusive events in S, then

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n) \quad \ldots(4.7)$$

Example 4.6

If A and B are mutually exclusive events and $P(A) = 0.30$, $P(B) = 0.45$, then find the probability of the following events.

(*a*) A' (*b*) $A \cap B$ (*c*) $A \cup B$ (*d*) $A' \cap B'$

Solution

(*a*) Using (4.2),

$$P(A') = 1 - P(A) = 1 - 0.30 = 0.70. \qquad \textit{Answer (a)}$$

(*b*) Since A and B are given to be mutually exclusive events, therefore $A \cap B = \phi$. Thus,

$$P(A \cap B) = P(\phi) = 0. \qquad \textit{Answer (b)}$$

(*c*) Since A and B are given to be mutually exclusive events, therefore using (4.7)

$$P(A \cup B) = P(A) + P(B) = 0.30 + 0.45 = 0.75. \qquad \textit{Answer (c)}$$

(*d*) Using (4.2),

$$P(A' \cap B') = 1 - P((A' \cap B')')$$
$$= 1 - P(A \cup B)$$
$$= 1 - 0.75 \qquad \text{(Using (c))}$$
$$= 0.25. \qquad \textit{Answer (d)}$$

4.3 Conditional Probability

Let A and B be any two events of a sample space S. Then the probability of the occurrence of event A when it is given that B has already occurred is expressed by the symbol $P(A|B)$ and is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \; P(B) > 0 \quad \ldots(4.8)$$

which is known as ***conditional probability*** of the event A relative to event B.

Similarly, conditional probability of the event B relative to event A is defined as follows.

$$P(B|A) = \frac{P(B \cap A)}{P(A)}, \; P(A) > 0 \quad \ldots(4.9)$$

➢ Properties of Conditional Probability

(1) Let A_1, A_2 and B be any three events of a sample space S, then

$$P((A_1 \cup A_2)|B) = P(A_1|B) + P(A_2|B) - P((A_1 \cap A_2)|B), \; P(B) > 0. \quad \ldots(4.10)$$

(2) Let A and B be any two events of a sample space S, then

$$P(A'|B) = 1 - P(A|B), \; P(B) > 0 \quad \ldots(4.11)$$

Once again let us review further theorems on probability.

Theorem 4.5 (Multiplication Rule) Let S be a sample sapce, and A and B be any two events in S, then

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B); \; P(A) > 0, \; P(B) > 0 \quad \ldots(4.12)$$

➢ **Corrollary** Let S be a sample space, and A, B and C be any three events in S, then

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|(A \cap B)) \quad \ldots(4.13)$$

Theorem 4.6 (Total Probability) Let B_1 and B_2 be two mutually exclusive and exhaustive events of a sample space S out of which one definitely occurs and then the event A occur. Then the probability of the occurrence of the event A is

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) \quad \ldots(4.14)$$

Example 4.7

Let A and B be two events with $P(A) = 3/8$, $P(B) = 7/8$ and $P(A \cup B) = 3/4$. Find $P(A|B)$ and $P(B|A)$.

Solution

For finding $P(A|B)$ and $P(B|A)$, we require to find $P(A \cap B)$.

Using (4.5),

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\Rightarrow \frac{3}{4} = \frac{3}{8} + \frac{7}{8} - P(A \cap B)$$

$$\Rightarrow P(A \cap B) = \frac{10}{8} - \frac{3}{4} = \frac{4}{8} = \frac{1}{2}.$$

Using (4.8),

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/2}{7/8} = \frac{4}{7}.$$

Using (4.9),

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{1/2}{3/8} = \frac{4}{3}.$$

Example 4.8

In producing screws, let A mean "screw too slim" and B mean "screw too short". Let $P(A) = 0.1$ and let the conditional probability that a slim screw is also too short be $P(B \mid A) = 0.2$. What is the probability that the screw that we pick randomly from the lot produced will be both too slim and too short?

[GTU, May 2017 - Civil]

Solution

Given that

$$P(A) = 0.1, \; P(B \mid A) = 0.2.$$

The required probability (using (4.9)) is given by

$$P(A \cap B) = P(A)\, P(B \mid A)$$
$$= (0.1)\,(0.2)$$
$$= 0.02.$$ ***Answer***

Example 4.9

In a box, 100 bulbs are supplied out of which 10 bulbs have defects of type A, 5 bulbs have defects of type B and 2 bulbs have defects of both the types. Find the probabilities that

(a) a bulb to be drawn at random has a B type defect under the condition that it has an A type defect, and

(b) a bulb to be drawn at random has no B type defect under the condition that it has no A type defect.

Solution

Let us first define the follo wing events.

A : The bulb has A type defect

B : The bulb has B type defect

Then by given information

$$P(A) = \frac{10}{100} = 0.10, \; P(B) = \frac{5}{100} = 0.05, \; P(A \cap B) = \frac{2}{100} = 0.02.$$

(a) The required probablity (using (4.9)) is given by

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{0.02}{0.10} = 0.2.$$ ***Answer (a)***

(b) The required probability is given by

$$P(B' \mid A') = \frac{P(B' \cap A')}{P(A')}$$

$$= \frac{P((A \cup B)')}{P(A')}$$ (De Morgan's law)

$$= \frac{1 - P(A \cup B)}{1 - P(A)}$$ (Using (4.2))

$$= \frac{1 - [P(A) + P(B) - P(A \cap B)]}{1 - P(A)}$$ (Using (4.5))

$$= \frac{1 - (0.10 + 0.05 - 0.02)}{1 - 0.10}$$

$$= \frac{1 - 0.13}{0.90}$$

$$= \frac{0.87}{0.90}$$

$$= \frac{87}{90}$$

$$= \frac{29}{30}.$$ ***Answer (b)***

Example 4.10

In a certain assembly plant, three machines B_1, B_2 and B_3 produce 30%, 45% and 25% of the products, respectively. It is known from past experience that 2%, 3% and 2% of the products made by each machine, respectively, are defective. Suppose that a finished product is randomly selected. What is the probability that it is defective? *[GTU, May 2016]*

Solution

Let us first define the following events.

A : The product is defective

B_1 : The product is made by machine B_1

B_2 : The product is made by machine B_2

B_3 : The product is made by machine B_3

Given that

$$P(B_1)=\frac{30}{100}=0.3,\ P(B_2)=\frac{45}{100}=0.45,\ P(B_3)=\frac{25}{100}=0.25.$$

$$P(A\mid B_1)=\frac{2}{100}=0.02,\ P(A\mid B_2)=\frac{3}{100}=0.03,\ P(A\mid B_3)=\frac{2}{100}=0.02.$$

The required probability using Theorem 4.6 becomes

$$P(A)=P(B_1)P(A\mid B_1)+P(B_2)P(A\mid B_2)+P(B_3)P(A\mid B_3)$$
$$=(0.3)(0.02)+(0.45)(0.03)+(0.25)(0.02)$$
$$=0.006+0.0135+0.005$$
$$=0.0245.$$

Answer

4.3.1 Independent Events Let A and B be any two events of a sample space S, then A and B are called independent, if

$$P(A\cap B)=P(A)\cdot P(B) \qquad ...(4.15)$$

Using (4.8),

Using (4.9),

$$P(A\mid B)=P(A).$$

$$P(B\mid A)=P(B).$$

This means that the probability of A does not depend on the occurrence or nonoccurrence of B and conversely.

Example 4.11

If A and B are independent events, where $P(A)=1/4$, $P(B)=2/3$. Find $P(A\cup B)$.

Solution

Using (4.5),

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$
$$=P(A)+P(B)-P(A)\cdot P(B)$$

(Since A and B are independent, so using (4.15))

$$=\frac{1}{4}+\frac{2}{3}-\frac{1}{4}\frac{2}{3}$$

$$=\frac{3+8}{12}-\frac{2}{12}$$

$$=\frac{11}{12}-\frac{2}{12}$$

$$=\frac{9}{12}$$

$$=\frac{3}{4}.$$

Answer

Example 4.12

A problem of statistics is given to three students A, B and C whose chances of solving it are 1/3, 1/4 and 1/2, respectively. What is the probability that the problem will be solved ?

Solution

Let us first define the following events.

A : Student A can solve the problem

B : Student B can solve the problem

C : Student C can solve the problem

Given that

$$P(A)=\frac{1}{3},\ P(B)=\frac{1}{4},\ P(C)=\frac{1}{2}.$$

When out of three students at least one solve the problem, then the problem is said to be solved. Therefore, required probability is

$$P(A\cup B\cup C)=1-P((A\cup B\cup C)') \qquad \text{(Using (4.2))}$$

$$=1-P(A'\cap B'\cap C') \qquad \text{(De Morgan's law)}$$

$$=1-P(A')\cdot P(B')\cdot P(C')$$

(Since A, B, C are independent events, so using (4.15))

$$=1-[1-P(A)][1-P(B)][1-P(C)]$$

(Using (4.2))

$$=1-\left(1-\frac{1}{3}\right)\left(1-\frac{1}{4}\right)\left(1-\frac{1}{2}\right)$$

$$=1-\left(\frac{2}{3}\right)\left(\frac{3}{4}\right)\left(\frac{1}{2}\right)$$

$$=1-\frac{1}{4}$$

$$=\frac{3}{4}.$$ ***Answer***

4.4 Bayes' Theorem

Theorem 4.7 (*Bayes' Theorem*) Let $A_1, A_2, \ldots, A_n$ be n-mutually exclusive and exhaustive events of a sample space S and let B be any event, then

$$P(A_i\mid B)=\frac{P(A_i)P(B\mid A_i)}{P(B)};\ P(B)>0,\ i=1,2,\ldots,n.$$

where

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

$$=\sum_{i=1}^{n}P(A_i)(B\mid A_i).$$

Example 4.13

Three boxes contain 10%, 20% and 30% of defective finger joints. A finger joint is

selected at random which is defective. Determine the probability that it comes from (*a*) First box (*b*) Second box (*c*) Third box.

[GTU, May 2017 - Civil]

Solution

Let us first define the following events.

A : Selection of defective finger joint

B_1 : Defective finger joint is from box-1

B_2 : Defective finger joint is from box-2

B_3 : Defective finger joint is from box-3

Given that

$$P(B_1)=\frac{1}{3},\quad P(B_2)=\frac{1}{3},\quad P(B_3)=\frac{1}{3},$$

$$P(A\mid B_1)=0.1,\quad P(A\mid B_2)=0.2,\quad P(A\mid B_3)=0.3.$$

Therefore, using Theorem 4.6

$$P(A)=P(B_1)P(A\mid B_1)+P(B_2)P(A\mid B_2)+P(B_3)P(A\mid B_3)$$

$$=\left(\frac{1}{3}\right)(0.1)+\left(\frac{1}{3}\right)(0.2)+\left(\frac{1}{3}\right)(0.3)$$

$$=0.2.$$

(*a*) The required probability using Throerem 4.7 becomes

$$P(B_1\mid A)=\frac{P(B_1)P(A\mid B_1)}{P(A)}$$

$$=\frac{\left(\frac{1}{3}\right)(0.1)}{0.2}$$

$$\approx 1.6667.$$ ***Answer (a)***

(*b*) The required probability using Throerem 4.7 becomes

$$P(B_2\mid A)=\frac{P(B_2)P(A\mid B_2)}{P(A)}$$

$$= \frac{\left(\frac{1}{3}\right)(0.2)}{0.2}$$

$= 0.3333.$ **Answer (b)**

(c) The required probability using Throerem 4.7 becomes

$$P(B_3 \mid A) = \frac{P(B_3)P(A \mid B_3)}{P(A)}$$

$$= \frac{\left(\frac{1}{3}\right)(0.3)}{0.2}$$

$= 0.5.$ **Answer (c)**

Review Exercises

01. For independent events A and B, it is given that $P(A) = 1/2$ and $P(B) = 1/3$. Find the probability of the following events.
    (a) Neither of the events A and B occur
    (b) Out of A and B at least one occur
    (c) A occur but B does not occur
02. Three cards are selected one by one from 52 cards. What is the probability that all the three selected cards will be an ace. Find the probability for the following two cases.
    (a) With replacement of cards (b) Without replacement of cards
03. If A and B are independent events and $P(A) = 2/3$. Find $P(A \cup B)$, $P(A' \mid B)$ and $P(A' \cap B)$.
04. Two cards are drawn from a well shuffled pack of 52 cards. Find the probability of the following events.
    (a) Either both are spade or both are kings
    (b) Either both are black cards or both are kings
    (c) Both are of the same colour
05. Suppose A and B are two events with $P(A) = 0.6$, $P(B) = 0.3$. Find the probability of the following events.
    (a) A does not occur (b) B does not occur
    (c) A or B occurs (d) neither A nor B occurs
06. What is the probability that a number selected from 1 to 25 is a prime number when each of the numbers is equally likely to be selected.

07. Three coins are tossed. What is the probability of getting the following events.
    (a) All heads (b) Two heads
    (c) At least one head (d) At least two heads

Answers

Review Exercises

01. (a) 1/3 (b) 2/3 (c) 1/3 **02.** (a) 1/2197 (b) 1/5525 **03.** 13/15, 1/3, 1/5
04. (a) 14/221 (b) 55/221 (c) 25/51 **05.** (a) 0.4 (b) 0.7 (c) 0.7 (d) 0.3
06. 9/25 **07.** (a) 1/8 (b) 3/8 (c) 7/8 (d) 1/2

Chapter 5
Probability Distributions

5.1 Concept of a Random Variable

To understand the concept of a random variable, let us first consider an example of tossing two coins together. In this case, sample space S is defined as follows.

$$S = \{HH, HT, TH, TT\}$$

Suppose we associate with each sample point, a real number as 2, 1, 1, 0, respectively. Then it is a obvious question that on which basis different real number is associated to each sample point ? It is obvious that each real number indicates the number of heads. That is, in this case number of heads is a rule by which we have associated a real number to each sample point and this rule is nothing but a real valued funciton defined on a sample space. This real valued function is known as ***random variable*** (***r.v.***).

Thus, if by some rule (or some function) we can associate a unique real number with each sample point of a sample space, then we have defined some function on sample space. This function, itself, is called ***random variable*** (also known as ***chance variable*** or ***stochastic variable*** or a ***variate***). The random variable is generally denoted by capital letters such as $X, Y, Z, \ldots$, and its particular values are denoted by small letters such as $x, y, z, \ldots$.

Alternatively, $X : S \to \mathbb{R}$ is a random variable.

The random variable is called so because the value it will assume depends on chance.

Let us once again recall above example, where

$$S = \{HH, HT, TH, TT\}.$$

Let us denote each sample point by w_1, w_2, w_3 and w_4, respectively. Also, let us define a function X on S such that

$$X(w) = \text{Number of heads in } w;\ w \in S.$$

Therefore,

$$X(w_1) = X(HH) = 2 = x_1$$

$$X(w_2) = X(HT) = 1 = x_2$$

$$X(w_3) = X(TH) = 1 = x_3$$

$$X(w_4) = X(TT) = 0 = x_4$$

This X is known as random variable and the set $\{2, 1, 0\}$ is called ***range*** of a random variable X.

Note On a sample space more than one random variable can be defined.

5.1.1 Discrete and Continuous Random Variables There are two types of random variables in statistics.

(a) **Discrete random variable** If the range of a random variable X is finite or an infinite sequence of real numbers, then it is called a discrete random variable.

→ ***For example,*** when we deal with counting cars on a road, number of students in a class, etc., then we have discrete random variable.

(b) **Continuous random variable** If the range of a random variable X contains interval of $\mathbb{R}$, then it is called a continuous random variable.

→ ***For example,*** when we deal with measuring height of students in a class, hardness of steel, etc., then we have continuous random variable.

In most practical problems, discrete r.v. COUNT data, whereas continuous r.v. represent MEASURED data.

5.2 Discrete Probability Distributions - Probability Mass Function

A graph, table or formula that specifies all possible values that a discrete random variable can take on, together with their associated probabilities, is called a ***discrete probability distribution.***

Let a random variable X assume values $x_1, x_2, \ldots, x_n$ and let the respective probabilities be $p_1, p_2, \ldots, p_n$ such that

$$\sum_{i=1}^{n} p_i = 1 \text{ and every } p_i \geq 0 \qquad \ldots(5.1)$$

Then set

$$\{p_1 = P(X = x_1),\ p_2 = P(X = x_2),\ \ldots,\ p_n = P(X = x_n)\}$$

is called the ***probability distribution*** of a random variable X.

We may regard this probability distribution in terms of a general mathematical expression $f(x)$, where $f(x_1) = p_1,\ f(x_2) = p_2, \ldots, f(x_n) = p_n$ satisfying

$$\sum_{i=1}^{n} f(x_i) = 1 \text{ and every } f(x_i) \geq 0 \qquad \ldots(5.2)$$

This is thus a discrete function defined for values which the random variable X can assume. This function $f(x)$ is known as ***probability mass function*** (***p.m.f.***) or the probability function of the discrete random variable X.

If $f(x)$ is known, then it is not necessary to write the probability $p_1, p_2, \ldots, p_n$ because as mentioned above p_i may be obtained from $f(x)$ on putting $x = x_i$; that is, $p_i = f(x_i)$. The discrete probability distribution is then defined simply by stating the mathematical expression for $f(x)$ along with the set of possible values $x_1, x_2, \ldots, x_n$ as follows.

X	:	x_1	x_2	...	x_n
$p = f(x)$	:	$f(x_1)$	$f(x_2)$	...	$f(x_n)$

Example 5.1

Is $f(x) = x/6;\ x = 0, 1, 2, 3, 4$ define probability distribution? Justify your answer.

[GTU, May 2017]

Solution

Here, $n = 5$. Let $x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 3, x_5 = 4$.

$$\sum_{i=1}^{5} f(x_i) = f(x_1) + f(x_2) + f(x_3) + f(x_4) + f(x_5)$$

$$= \frac{0}{6} + \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6}$$

$$= \frac{10}{6}$$

$$\neq 1.$$

Therefore, using (5.2), we say that the given function $f(x)$ is not defined probability distribution.

Answer

Example 5.2

A discrete random variable X has the following probability distributions.

X	:	0	1	2	3	4	5
$f(X)$	:	0	k	0.2	$2k$	0.3	$2k$

(*a*) Find k. (*b*) Compute $P(X<3)$, $P(X\geq 3)$, $P(2<X<5)$, $P(X\leq 4)$.

Solution

(*a*) Let $x_1=0, x_2=1, x_3=2, x_4=3, x_5=4, x_6=5$. Since X is given to be discrete random variable, therefore, using (5.2)

$$\sum_{i=1}^{6} f(x_i)=1$$

$$\Rightarrow f(x_1)+f(x_2)+f(x_3)+f(x_4)+f(x_5)+f(x_6)=1$$

$$\Rightarrow 0+k+0.2+2k+0.3+2k=1$$

$$\Rightarrow 5k+0.5=1$$

$$\Rightarrow 5k=0.5$$

$$\Rightarrow k=\frac{0.5}{5}=\frac{5}{50}=\frac{1}{10}.$$

Answer (a)

(*b*)

$$P(X<3)=P(X=0)+P(X=1)+P(X=2)$$

$$=0+k+0.2$$

$$=k+0.2$$

$$=\frac{1}{10}+0.2 \qquad (\text{Since } k=1/10)$$

$$=\frac{1}{10}+\frac{2}{10}$$

$$=\frac{3}{10}$$

$$=0.3.$$

$$P(X\geq 3)=P(X=3)+P(X=4)+P(X=5)$$

$$=2k+0.3+2k$$

$$=4k+0.3$$

$$=4\left(\frac{1}{10}\right)+0.3 \qquad (\text{Since } k=1/10)$$

$$=\frac{4}{10}+\frac{3}{10}$$

$$=\frac{7}{10}$$

$$=0.7.$$

$$P(2<X<5)=P(X=3)+P(X=4)$$

$$=2k+0.3$$

$$=2\left(\frac{1}{10}\right)+0.3 \qquad (\text{Since } k=1/10)$$

$$=\frac{2}{10}+\frac{3}{10}$$

$$=\frac{5}{10}$$

$$=0.5.$$

$$P(X\leq 4)=1-P(X=5)$$

$$=1-2k$$

$$=1-2\left(\frac{1}{10}\right) \qquad (\text{Since } k=1/10)$$

$$=1-\frac{2}{10}$$

$$=\frac{8}{10}$$

$$=0.8.$$

5.2.1 Distribution Function If X is a discrete random variable with probability mass function f, then the ***distribution function*** F is defined as

$$F(x) = \sum_{x_i \le x} f(x_i); \quad \text{...(5.3)}$$

that is, $F(x)$ is a sum of probabilities of all x_i whose values are less than or equal to x.

Thus,

$$F(x) = \sum_{x_i \le x} f(x_i) = P(x_i \le x). \quad \text{...(5.4)}$$

→ ***For example***, if random variable X assume the values x_1, x_2, x_3, x_4, x_5, then

$$F(x_5) = \sum_{x_i \le x_5} f(x_i) = f(x_1) + f(x_2) + f(x_3) + f(x_4) + f(x_5).$$

Note Distribution function is also known as ***cumulative distribution function***.

Example 5.3

Determine the distribution function of the probability distribution given in Example 5.1.

Solution

The distribution function for each x is given below.

$$F(X = 0) = f(0) = 0.$$

$$F(X = 1) = f(0) + f(1) = 0 + k = \frac{1}{10}. \quad \text{(Since } k = 1/10\text{)}$$

$$F(X = 2) = f(0) + f(1) + f(2) = 0 + k + 0.2 = 0 + \frac{1}{10} + \frac{2}{10} = \frac{3}{10}.$$

$$F(X = 3) = f(0) + f(1) + f(2) + f(3)$$

$$= 0 + k + 0.2 + 2k$$

$$= 0 + \frac{1}{10} + \frac{2}{10} + \frac{2}{10}$$

$$= \frac{5}{10}$$

$$= 0.5.$$

$$F(X = 4) = P(x_i \le 4) = 0.8. \quad \text{(Calculated in Example 5.1)}$$

$$F(X = 5) = f(0) + f(1) + f(2) + f(3) + f(4) + f(5) = 1.$$

(Since total probability is 1)

5.2.2 Mathematical Expectations Let X be a discrete random variable assuming values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, respectively. Then the ***expected value*** or ***expectation*** of X is denoted and defined as

$$E(X) = \sum_{i=1}^{n} p_i x_i \quad \text{...(5.5)}$$

If X is a discrete random variable with probability mass function $f(x)$, then

$$E(X) = \sum_{i=1}^{n} x_i f(x_i) \quad \text{...(5.6)}$$

Similarly, the expected value of X^2 is denoted and defined as follows.

$$E(X^2) = \sum_{i=1}^{n} p_i x_i^2 \quad \text{...(5.7)}$$

and

$$E(X^2) = \sum_{i=1}^{n} x_i^2 f(x_i) \quad \text{...(5.8)}$$

when probability mass function $f(x)$ is given.

In general, the expected value of any function $g(X)$ is defined as

$$E[g(X)] = \sum_{i=1}^{n} p_i g(x_i) \quad \text{...(5.9)}$$

It should be noted that the expectation $E(X)$ is also known as ***mathematical expectation*** of X or the ***mean*** (μ) of X.

- **Properties of Mathematical Expectations**

(1) The expectation of a constant equal to that constant; that is,

$$E(k) = k, \quad \text{...(5.10)}$$

where k is a constant.

(2) If k is a constant, then

$$E(kX) = kE(X). \quad ...(5.11)$$

(3) The expectation of the sum (or difference) of two discrete random variables X and Y is equal to the sum (or difference) of their expectations; that is,

$$E(X \pm Y) = E(X) \pm E(Y). \quad ...(5.12)$$

(4) The expectation of the product of two independent random variables X and Y is equal to the product of their expectations; that is,

$$E(XY) = E(X) \cdot E(Y). \quad ...(5.13)$$

(5) The expectation operator E is a linear operator.

(6) Let $g_1(X)$ and $g_2(X)$ be functions of random variable X, then

$$E[k_1 g_1(X) + k_2 g_2(X)] = k_1 E[g_1(X)] + k_2 E[g_2(X)], \quad ...(5.14)$$

where k_1 and k_2 are constants.

➢ **Variance** It is denoted and defined as follows.

$$V(X) = \sigma^2 = E(X-\mu)^2 = E(X^2) - [E(X)]^2, \quad ...(5.15)$$

where μ is mean and σ is standard deviation.

Example 5.4

The probability distribution of a commodity is given below.

Demand (x_i)	:	5	6	7	8	9	10
Probability (p_i)	:	0.05	0.10	0.30	0.40	0.10	0.05

Find expected demand and its variance.

Solution

Let $x_1 = 5, x_2 = 6, x_3 = 7, x_4 = 8, x_5 = 9, x_6 = 10$. The expected demand is given by using (5.5) as

$$E(X) = \sum_{i=1}^{6} p_i x_i$$

$$= p_1x_1 + p_2x_2 + p_3x_3 + p_4x_4 + p_5x_5 + p_6x_6$$
$$= (0.05)(5) + (0.10)(6) + (0.30)(7) + (0.40)(8) + (0.10)(9) + (0.05)(10)$$
$$= 0.25 + 0.60 + 2.10 + 3.20 + 0.90 + 0.50$$
$$= 7.55.$$

Using (5.7),

$$E(X^2) = \sum_{i=1}^{6} p_i x_i^2$$

$$= p_1x_1^2 + p_2x_2^2 + p_3x_3^2 + p_4x_4^2 + p_5x_5^2 + p_6x_6^2$$
$$= (0.05)(5)^2 + (0.10)(6)^2 + (0.30)(7)^2 + (0.40)(8)^2 + (0.10)(9)^2 + (0.05)(10)^2$$
$$= 1.25 + 3.60 + 14.70 + 25.60 + 8.10 + 5.00$$
$$= 58.25.$$

Using (5.15), variance is given by

$$\sigma^2 = E(X^2) - [E(X)]^2$$
$$= 58.25 - (7.55)^2$$
$$\approx 1.25. \quad \textit{Answer}$$

Example 5.5

Three coins are tossed together and let random variable X be the number of heads in each outcome. Then find (*a*) Probability distribution, (*b*) Mean, (*c*) Standard deviation.

Solution

Here,

$$S = \{HHH, HHT, HTT, HTH, THH, THT, TTH, TTT\}.$$

(*a*)

Sample points (w)	:	TTT	HTT, THT, TTH	HHT, HTH, THH	HHH
Random variable (X)	:	0	1	2	3
$P(X = x)$ or $f(x)$	:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Answer (a)

(*b*) Using (5.5),

$$\text{Mean } \mu = E(X)$$

$$= \sum_{i=1}^{4} p_i x_i$$

$$= p_1x_1 + p_2x_2 + p_3x_3 + p_4x_4$$

$$= \left(\frac{1}{8}\right)(0) + \left(\frac{3}{8}\right)(1) + \left(\frac{3}{8}\right)(2) + \left(\frac{1}{8}\right)(3)$$

$$= \frac{3}{8} + \frac{6}{8} + \frac{3}{8}$$

$$= \frac{12}{8}$$

$$= \frac{3}{2}.$$

Answer (b)

(*c*) Using (5.7),

$$E(X^2) = \sum_{i=1}^{4} p_i x_i^2$$

$$= p_1x_1^2 + p_2x_2^2 + p_3x_3^2 + p_4x_4^2$$

$$= \left(\frac{1}{8}\right)(0)^2 + \left(\frac{3}{8}\right)(1)^2 + \left(\frac{3}{8}\right)(2)^2 + \left(\frac{1}{8}\right)(3)^2$$

$$= \frac{3}{8} + \frac{12}{8} + \frac{9}{8}$$

$$= \frac{24}{8}$$

$$= 3.$$

Using (5.15), variance is given by

$$V(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

$$= 3 - \left(\frac{3}{2}\right)^2$$

$$= 3 - \frac{9}{4}$$

$$= \frac{12-9}{4}$$

$$= \frac{3}{4}.$$

Hence, standard deviation is

$$\sigma = \sqrt{\text{Variance}} = \sqrt{\frac{3}{4}} = \frac{\sqrt{3}}{2}.$$

Answer (c)

5.3 Continuous Probability Distributions - Probability Density Function

The probability distribution of a continuous random variable is known as ***continuous probability distribution***.

As mentioned in Section 5.1, for continuous random variable X, range of X contains interval of $\mathbb{R}$; that is, it can assume any value in the interval $[a, b]$. Therefore in continuous probabilitydistributions, we cannot associate a finite nonzero probability with each variate value and we have to assign probabilities to intervals and not to individual values.

Consider a very small interval

$$\left[x - \frac{dx}{2}, x + \frac{dx}{2}\right] \qquad ...(5.16)$$

of infinitesimal length dx around the point x. Let f be any continuous function of a random variable X such that $f(x)dx$ represents the probability distribution of X lies in the interval (5.16); that is,

$$P\left(x - \frac{dx}{2} \le X \le x + \frac{dx}{2}\right) = f(x)dx. \qquad ...(5.17)$$

Then such a function f is called ***probability density function*** (***p.d.f.***) of a continuous random variable X, where f has to satisfy following two conditions.

(i) $f(x) \ge 0$ for all x

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

Here. $f(x)dx$ is called probability differential.

Note The probability of x lying in any given interval $[a, b]$ can be obtained by

$$P(a \le x \le b) = \int_a^b f(x)dx .$$

Example 5.6

Check whether the function defined as

$$f(x) = 0 \quad ; x < 2$$

$$= \frac{1}{18}(3+2x) \quad ; 2 \le x \le 4$$

$$= 0 \quad ; x > 4$$

is a probability density function ? If yes, find $P(2 \le X \le 3)$.

Solution

By definition of the given function, it is clear that $f(x) \ge 0$ for all x.

Now,

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{2} 0\,dx + \int_2^4 \frac{1}{18}(3+2x)dx + \int_4^{\infty} 0\,dx$$

$$= \frac{1}{18}\left(3x + \frac{2x^2}{2}\right)_2^4$$

$$= \frac{1}{18}\left(3x + x^2\right)_2^4$$

$$= \frac{1}{18}(12+16-6-4)$$

$$= 1.$$

Hence, the given function $f(x)$ satisfy both the requirements of probability density

function.

$$P(2 \le X \le 3) = \int_2^3 \frac{1}{18}(3+2x)dx$$

$$= \frac{1}{18}\left(3x + \frac{2x^2}{2}\right)_2^3$$

$$= \frac{1}{18}\left(3x + x^2\right)_2^3$$

$$= \frac{1}{18}(9+9-6-4)$$

$$= \frac{1}{18}(8)$$

$$= \frac{4}{9}.$$ *Answer*

5.21 Distribution Function If X is a continuous random variable with probability density function f, then the ***distribution function*** F is defined as

$$F(x) = \int_{-\infty}^{x} f(x)dx \qquad ...(5.18)$$

thatis, $F(x)$ represents the probability that the random variable having a value less than or equal to a specified value of x.

Thus,

$$F(x) = \int_{-\infty}^{x} f(x)dx = P(X \le x). \qquad ...(5.19)$$

Example 5.7

Findthe distribution function $F(x)$ for the following probability density function of a random variable X.

$$f(x) = \frac{1}{x^2+1}; -\infty < x < \infty$$

Solution

Using (5.18),

$$F(x) = \int_{-\infty}^{x} \frac{1}{u^2+1}\,du$$

$$= \int_{-\infty}^{x} \frac{1}{u^2+1}\,du$$

$$= (\tan^{-1} u)_{-\infty}^{x}$$

$$= \frac{\pi}{2} + \tan^{-1} x.$$ ***Answer***

5.3.2 Mathematical Expectations Let X be a continuous random variable with probability density function f. Then the ***expected value*** or ***expectation*** of X is denoted and defined as

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx \qquad \ldots(5.20)$$

Similarly, the expected value of X^2 is denoted and defined as

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx \qquad \ldots(5.21)$$

Here, also $E(X)$ is known as ***mathematical expectation*** of X or the ***mean*** (μ) of X.

➢ **Variance** It is denoted and defined as follows.

$$V(X) = \sigma^2 = E(X-\mu)^2 = E(X^2) - [E(X)]^2, \qquad \ldots(5.22)$$

where μ is mean and σ is standard deviation.

Example 5.8

Check whether the function defined as

$$f(x) = |x| \quad ; -1 < x < 1$$
$$= 0 \quad ; \text{otherwise}$$

is a probability density function? If yes, find mean and variance.

Solution

By definition of the given function, it is clear that $f(x) \geq 0$ for all x.

Now,

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{-1} 0\,dx + \int_{-1}^{1} |x|\,dx + \int_{1}^{\infty} 0\,dx$$

$$= \int_{-1}^{1} |x|\,dx$$

$$= 2\int_{0}^{1} x\,dx \qquad \text{(Since } |x| \text{ is even function)}$$

$$= 2\left(\frac{x^2}{2}\right)_0^1$$

$$= 1.$$

Hence, the given function $f(x)$ satisfy both the requirements of probability density function.

Using (5.20),

$$\text{Mean } \mu = E(X)$$

$$= \int_{-\infty}^{\infty} x f(x)\,dx$$

$$= \int_{-1}^{1} x|x|\,dx$$

$$= 0. \qquad \text{(Since } x|x| \text{ is an odd function)}$$

Using (5.21),

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx$$

$$= \int_{-1}^{1} x^2|x|\,dx$$

$$= 2\int_0^1 x^3\,dx \text{ (Since } x^2|x| \text{ is even function)}$$

$$= 2\left(\frac{x^4}{4}\right)_0^1$$

$$= \frac{1}{2}.$$

Using (5.22), variance is given by

$$V(X) = \sigma^2 = E(X^2) - [E(x)]^2$$

$$= \frac{1}{2} - (0)^2$$

$$= \frac{1}{2}. \qquad \textit{Answer}$$

5.4 Joint Probability Distribution

5.4.1 Joint Probability Mass function - Marginal Distributions There are many experiments where we have to observe two random variables simultaneously is order to determine not only their individual behaviour but also the degree of relationship between them.

→ *For example*, X : Age Y : Blood pressure of a person

Let X and Y be two discrete random variables defined on the same sample space S. Then the joint probability of mass function of two discrete random variables X and Y is a function $f(x, y)$ defined on the product of set

$$X \times Y = \{(x_1, y_1), (x_1, y_2), \ldots, (x_m, y_n)\}$$

assigning probability to each of the ordered pairs (x_i, y_j), where X can assume any one of m-values $x_1, x_2, \ldots, x_m$ and Y can assume any one of n-values $y_1, y_2, \ldots, y_n$.

Thus,

$$f(x_i, y_j) = P(X = x_i, Y = y_j)$$

gives the probability for the simultaneous occurrences of the outcomes x_i and y_j.

Further, $f(x, y)$ satisfies the following.

(1) $f(x_i, y_j) \geq 0$

(2) $\sum_{j=1}^{n} \sum_{i=1}^{m} f(x_i, y_j) = 1$

The set of triplets $(x_i, y_j, f(x_i, y_j))$ for $i = 1, 2.., m; j = 1, 2, \ldots, n$, is known as the joint probability distribution of X any Y. The following Table 5.1 shows representation of joint probability distributions.

X	Y					Row sums
	y_1	y_2	y_3	...	y_n	
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	$f(x_1, y_3)$	...	$f(x_1, y_n)$	$f(x_1)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	$f(x_2, y_3)$	...	$f(x_2, y_n)$	$f(x_2)$
x_3	$f(x_3, y_1)$	$f(x_3, y_2)$	$f(x_3, y_3)$	...	$f(x_3, y_n)$	$f(x_3)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	$f(x_m, y_3)$	⋯	$f(x_m, y_n)$	$f(x_m)$
Column sums	$g(y_1)$	$g(y_2)$	$g(y_3)$	⋯	$g(y_n)$	

Table 5.1 Joint Probability Distribution

➢ **Marginal Distributions** Marginal distribution $f(x)$ of X is given by

$$f(x_i) = f(x_i, y_1) + f(x_i, y_2) + \ldots + f(x_i, y_n)$$

$$= \sum_{j=1}^{n} f(x_i, y_j).$$

Similarly, marginal distribution $g(y)$ of Y is given by

$$g(y_j) = f(x_1, y_j) + f(x_2, y_j) + \ldots + f(x_m, y_j)$$

$$= \sum_{i=1}^{m} f(x_i, y_j).$$

Example 5.9

The joint probability mass function of two discrete random variables X and Y is given by

$$f(x, y) = c(2x + y); \quad 0 \le x \le 2, \ 0 \le y \le 3, \ x \text{ and } y \text{ are integers}$$
$$= 0 \quad ; \text{otherwise.}$$

Find

(a) value of constant c, (b) $P(X = 2, Y = 1)$, (c) $P(X \ge 2, Y \le 1)$.

Also, find marginal distribution $f(x)$ of X.

Solution

Here,

X \ Y	0	1	2	3	Total
0	$f(0,0)$ 0	$f(0,1)$ c	$f(0,2)$ $2c$	$f(0,3)$ $3c$	$6c$
1	$f(1,0)$ $2c$	$f(1,1)$ $3c$	$f(1,2)$ $4c$	$f(1,3)$ $5c$	$14c$
2	$f(2,0)$ $4c$	$f(2,1)$ $5c$	$f(2,2)$ $6c$	$f(2,3)$ $7c$	$22c$
Total	$6c$	$9c$	$12c$	$15c$	Grand total $42c$

(a) Since

$$\sum_{j=1}^{3} \sum_{i=1}^{2} f(x_i, y_j) = 1$$

$$\Rightarrow 42c = 1$$

$$\Rightarrow c = \frac{1}{42}$$

(b) **Answer (a)**

$$P(X = 2, Y = 1) = 5c = 5\left(\frac{1}{42}\right) = \frac{5}{42}.$$

Answer (b)

(c)

$$P(X \ge 1, Y \le 2) = P(X = 1, Y = 0) + P(X = 1, Y = 1)$$
$$+ P(X = 1, Y = 2) + P(X = 2, Y = 0)$$
$$+ P(X = 2, Y = 1) + P(X = 2, Y = 2)$$
$$= 2c + 3c + 4c + 4c + 5c + 6c$$
$$= 24c$$
$$= 24\left(\frac{1}{42}\right)$$
$$= \frac{4}{7}.$$

Answer (c)

The marginal distribution of X is given by

$$f(x) = 6c = \frac{6}{42} = \frac{1}{7} \text{ for } X = 0$$

$$= 14c = \frac{14}{42} = \frac{1}{3} \text{ for } X = 1$$

$$= 22c = \frac{22}{42} = \frac{21}{21} \text{ for } X = 2.$$

5.4.2 Joint Probability Density Function - Marginal Distributions There are many situations where we have to describe an outcome by giving the values of several continuous random variables.

→ ***For example***, we may measure weight and hardness of a rock.

Let X and Y be two continuous random variables and let $f(x, y)$ be a continuous function of random variables X and Y which enables to compute a probability and where

(1) $f(x, y) \ge 0$ for all x, y,

(2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1.$

Then such a function f is called joint probability density function.

Note 1 The probability that $a < X < b$, $c < Y < d$ is given by

$$P(a < X < b, c < Y < d) = \int_{c}^{d} \int_{a}^{b} f(x, y)dxdy.$$

➢ **Marginal Distributions** Marginal distribution $f(x)$ of X is given by

$$f(x) = \int_{-\infty}^{\infty} f(x, y)dy .$$

Similarly, marginal distribution $g(y)$ of Y is given by

$$g(y) = \int_{-\infty}^{\infty} f(x, y)dx .$$

Example 5.10

The joint probability density function of two continuous random variables X and Y is given by

$$f(x, y) = c\,xy \; ; 0 < x < 4, \; 1 < y < 5$$
$$= 0 \quad ; \text{otherwise}$$

Find

(*a*) value of constant c, (*b*) $P(X \geq 3, Y \leq 2)$, (*c*) $P(1 < X < 2, 2 < Y < 3)$.

Solution

(*a*) Since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dx\,dy = 1$$

$$\Rightarrow \int_{1}^{5}\int_{0}^{4} c\,xy\,dx dy = 1$$

$$\Rightarrow \int_{1}^{5} cy\left(\frac{x^2}{2}\right)_0^4 dy = 1$$

$$\Rightarrow \int_{1}^{5} 8cy\,dy = 1$$

$$\Rightarrow 8c\left(\frac{y^2}{2}\right)_1^5 = 1$$

$$\Rightarrow 4c(25-1) = 1$$

$$\Rightarrow 96c = 1$$

$$\Rightarrow c = \frac{1}{96}.$$ ...(i) ***Answer (a)***

(*b*)

$$P(X \geq 3, Y \leq 2) = \int_{1}^{2}\int_{3}^{4} c\,xy\,dx\,dy$$

$$= \int_{1}^{2} cy\left(\frac{x^2}{2}\right)_3^4 dy$$

$$= \int_{1}^{2} \frac{cy}{2}(16-9)\,dy$$

$$= \int_{1}^{2} \frac{7}{2}cy\,dy$$

$$= \frac{7}{2}c\left(\frac{y^2}{2}\right)_1^2$$

$$= \frac{7}{4}c(4-1)$$

$$= \frac{21}{4}c$$

$$= \frac{21}{4}\left(\frac{1}{96}\right)$$ Using (i)

$$= \frac{7}{128}.$$ ***Answer (b)***

(*c*)

$$P(1 < X < 2, 2 < Y < 3) = \int_{2}^{3}\int_{1}^{2} c\,xy\,dx\,dy$$

$$= \int_2^3 cy \left(\frac{x^2}{2} \right)_1^2 dy$$

$$= \int_2^3 \frac{3}{2} cy \, dy$$

$$= \frac{3}{2} c \left(\frac{y^2}{2} \right)_2^3$$

$$= \frac{3}{4} c (9-4)$$

$$= \frac{15}{4} c$$

$$= \frac{15}{4} \left(\frac{1}{96} \right) \qquad \text{(Using (i))}$$

$$= \frac{5}{128}. \qquad \textit{Answer (c)}$$

5.4.3 Independent Random Variables Let X and Y be two random variables. X and Y are said to be ***independent***, if

$$f(x, y) = f(x) \cdot g(y) \text{ for all } (x, y);$$

that is,

$$f(x_i, y_j) = P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

$$= f(x_i) g(y_j);$$

for all $i = 1, 2, \ldots, m; j = 1, 2, \ldots, n$.

In other words, when X and Y are independent, each entry $f(x_i, y_j)$ is obtained as the product of its marginal entries.

Example 5.11

For the point probability density function of two continuous random variables X and Y,

$$f(x, y) = 4xye^{-(x^2+y^2)}; \; x \geq 0, \; y \geq 0,$$

find the marginal distributions $f(x)$ and $g(y)$ and test whether X and Y are independent or not.

Solution

Marginal distribution

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$= \int_0^{\infty} 4xye^{-(x^2+y^2)} dy$$

$$= 4xe^{-x^2} \int_0^{\infty} e^{-y^2} \cdot y \, dy$$

$$= 4xe^{-x^2} \int_0^{\infty} e^{-t} \frac{dt}{2} \qquad \text{(Putting } y^2 = t\text{)}$$

$$= 2xe^{-x^2} \left(\frac{e^{-t}}{-1} \right)_0^{\infty}$$

$$= -2xe^{-x^2} (0-1)$$

$$= 2xe^{-x^2}$$

Similarly, we get

$$g(y) = \int_0^{\infty} 4xye^{-(x^2+y^2)} dx = 2ye^{-y^2}.$$

Now,

$$f(x, y) = 4xy \, e^{-(x^2+y^2)}$$

$$= 4xy \, e^{-x^2} e^{-y^2}$$

$$= (2xe^{-x^2})(2ye^{-y^2})$$

$$= f(x)g(y).$$

Therefore, variables X and Y are independent. *Answer*

5.5 Short Questions

Example 5.29

What is mathematical expectation, if we stand to win \$8 if and only if a balanced coin comes up heads? *[GTU, May 2017]*

Solution

$$E(X) = 8 \cdot \frac{1}{2} = 4\$.$$

Answer

Chapter 6

Some Special Probability Distributions

6.1 Theoretical Discrete Distributions : Binomial and Poisson Distributions

Let us now discuss two important theoretical discrete distributions, namely binomial and poisson.

6.1.1 Binomial Distribution This distribution is associated with repeated trials of an experiment.

➢ **Bernoulli Trials** Suppose a random experiment has two possible outcomes which are complementary, say success (S) and failure (F). If the probability $p(0 < p < 1)$ of getting success at each of the n trials of this experiment is constant, then the trials are called Bernoulli trials.

➢ **Binomial Distribution** Suppose an experiment E consists of finite number of n independent trials. Each trial has only two outcomes, say success and failure. Let probability of success and failure in each trial is constant, then this experiment will generate a random variable X (say) as number of successful trials which assume the values 0, 1, 2, ... , n. Let p be the probability of success and q be the probability of failure in each trial, where $p + q = 1$.

The problem is to find the probability of success (which occurs x times) and failure (which occurs $(n-x)$ times).

Suppose in the first x trials success occurs while in the remaining $(n-x)$ trials failure occurs, then the probability of these particular distribution is

$$p^x q^{n-x}.$$

Since there are ${}^nC_x$ distinct distributions possible, then the probability of x successes is given by

$$P(X = x) = p(x) = {}^nC_x p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n \qquad \ldots(6.1)$$

This is called the *binomial distribution*.

More formally,

A discrete random variable X is said to follow a ***binomial distribution***, if it assume only nonnegative values and its probability mass function is given by

$$p(x) = {}^nC_x p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n$$
$$= 0, \quad \text{otherwise} \qquad \ldots(6.2)$$

where p is the probability of success and q is the probability of failure, so $p + q = 1$.

It is clear from the definition of $p(x)$ that

(1) $p(x) \geq 0$ for $x = 0, 1, 2, \ldots, n$.

(2) $$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n} {}^nC_x p^x q^{n-x} \qquad \text{(Using (6.1))}$$

$$= (q + p)^n$$

$$= 1^n$$

$$= 1.$$

Thus, $p(x)$ satisfies both the requirements for a function to be a probability mass function.

The two constants n and p appearing in (6.1) are called ***parameters*** of the binomial distribution. If n and p are known, the distribution is said to be completely known (since $q = 1 - p$ is also then known). Note that n is a discrete parameter whereas p is a continuous parameter $(0 \leq p \leq 1)$.

Note The mean and variance of the binomial distribution with parameters n and p are defined as follows.

$$\text{Mean } \mu = E(X) = np \qquad \ldots(6.3)$$

$$\text{Variance } V(X) = \sigma^2 = npq \qquad \ldots(6.4)$$

Example 6.1

If 20% of the bolts produced by a machine are defective, determine the probability

that out of 4 bolts chosen, at most 2 bolts will be defective.

Solution

The probability of a defective bolt is

$$p = \frac{20}{100} = 0.2.$$

Therefore, the probability of a nondefective bolt is

$$q = 1 - p = 1 - 0.2 = 0.8.$$

For the bolts to be chosen $n = 4$, the p.m.f. of binomial distribution using (6.2) is

$$P(X = x) = p(x) = {}^4C_x (0.2)^x (0.8)^{4-x}; \quad x = 0, 1, 2, 3, 4.$$

The probability of at most 2 defective bolts is given by

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= {}^1C_0 (0.2)^0 (0.8)^4 + {}^4C_1 (0.2)^1 (0.8)^3 + {}^4C_2 (0.2)^2 (0.8)^2$$

$$= 0.4096 + 0.4096 + 0.1536$$

$$= 0.9728. \qquad \textit{Answer}$$

Example 6.2

A dice is thrown 6 times. If getting an odd number is a success, find the probability of (*a*) five successes, (*b*) at least five successes, (*c*) at the most five successes.

Solution

Let X be the number of successes, then X assume the values 0, 1, 2, 3, 4, 5, 6 and $n = 6$.

Let p be the probability of getting an odd number, then

$$p = \frac{1}{2},$$

which implies

$$q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}.$$

The corresponding binomial distribution using (6.2) is

$$P X = x) = {}^6C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{6-x}; \quad x = 0, 1, 2, 3, 4, 5, 6.$$

$$= {}^6C_x\left(\frac{1}{2}\right)^6.$$

(*a*) Probability of five successes is given by

$$P(X=5) = {}^6C_5\left(\frac{1}{2}\right)^6 = 6\times\frac{1}{64} = \frac{3}{32}.$$ **Answer (*a*)**

(*b*) Probability of at least five successes is given by

$$P(X \geq 5) = P(X=5) + P(X=6)$$

$$= {}^6C_5\left(\frac{1}{2}\right)^6 + {}^6C_6\left(\frac{1}{2}\right)^6$$

$$= \frac{3}{32} + \frac{1}{64}$$ (Using (*a*))

$$= \frac{7}{64}.$$ **Answer (*b*)**

(*c*) Probability of at most five successes is given by

$$P(X \leq 5) = 1 - P(X=6) = 1 - {}^6C_6\left(\frac{1}{2}\right)^6$$

$$= 1 - \frac{1}{64}$$

$$= \frac{63}{64}.$$ **Answer (*c*)**

Example 6.3

If the probability of a defective bolt is 0.1. Find mean and standard deviation of the distribution of defective bolts in a total of 400.

Solution

Let p be the probability of a defective bolt. Therefore, $p = 0.1$.

Also, given that $n = 400$.

Using (6.3),

$$\text{Mean } \mu = np = (400)(0.1) = 40.$$

Therefore, we can expect 40 bolts out of 400 bolts to be defective.

Using (6.4),

$$\text{Variance} = \sigma^2 = npq = (400)(0.1)(0.9)$$

(Since $q = 1 - p = 1 - 0.1 = 0.9$)

$$= 36.$$

Therefore,

$$\text{Standard deviation} = \sigma = \sqrt{36} = 6.$$ **Answer**

Example 6.4

A multiple choice test in mathematics with 40 questions, each having 5 options, is given to a student. If the student guess all 40 questions, what are the mean and stardard deviation of the number of correct answers? ***[GTU, May 2016]***

Solution

Let p be the probability of the correct answer. Therefore, $p = 1/5$.

Also, given that $n = 40$.

Using (6.3),

$$\text{Mean } \mu = np = (40)\left(\frac{1}{5}\right) = 8.$$

Therefore, we can expect 8 correct answers out of 40 questions.

Using (6.4),

$$\text{Variance} = \sigma^2 = npq = (40)\left(\frac{1}{5}\right)\left(\frac{4}{5}\right)$$

$$\left(\text{Since } q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}\right)$$

$$= \frac{32}{5}.$$

Therefore,

$$\text{Standard deviation} = \sigma = \sqrt{\frac{32}{5}} = 2.53. \qquad \textit{Answer}$$

6.1.2 Poisson Distribution This distribution is associated with the name of french mathematician S. D. Poisson. It is obtainable from the binomial distribution by putting $p = \lambda / n$, where λ is a constant and letting n increases indefinitely. Here, the number of trials in the series becomes very large but the probability of success in a trial p very small (that is, $n \to \infty$, $p \to 0$ so that λ is constant). It can be shown that

$$\lim_{n\to\infty} {}^nC_x p^x q^{n-x} = \frac{\lambda^x e^{-\lambda}}{x!} \; ; x = 0, 1, 2, 3, \ldots .$$

Thus, in the limiting form of the binomial distribution, the probability of x successes in an infinite series of trials is

$$\frac{\lambda^x e^{-\lambda}}{x!}.$$

More formally,

A discrete random variable X is said to follow a ***Poisson distribution***, if it assume only nonnegative values and its probability mass function is given by

$$p(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & x = 0, 1, 2, 3, \ldots \\ = 0, & \text{otherwise} \end{cases} \qquad \ldots(6.5)$$

where $\lambda > 0$ is a finite constant and it is called ***parameter*** of the Poisson distribution.

It is clear from the definition of $p(x)$ that

(1) $p(x) \geq 0$ for $x = 0, 1, 2, 3, \ldots$.

(2)

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{(Using (6.5))}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$$

$$= e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right)$$

$$= e^{-\lambda} \cdot e^{\lambda}$$

$$= e^0$$

$$= 1.$$

Thus, $p(x)$ satisfies both the requirements for a function to be a probability mass function.

Note The mean and variance of the Poisson distribution with parameter λ are defined as follows.

$$\text{Mean } \mu = E(X) = \lambda \qquad \ldots(6.6)$$

$$\text{Variance } V(X) = \sigma^2 = \lambda \qquad \ldots(6.7)$$

Example 6.5

In a bolt manufacturing company, it is found that there is a small chance of 1/500 for any bolt to be defective. The bolts are supplied in a packet of 30 bolts. Use Poisson distribution to find approximate number of packets containing

(*a*) no defective bolt,

(*b*) two defective bolts

in the consignment of 10000 packets.

Solution

Let X be the number of defective bolts in a packet, then X assume the values 0, 1, ..., 30 and $n = 30$. Also, let $N = 10000$.

Let p be the probability of any bolt to be defective, then $p = 1/500$.

Therefore,

$$\lambda = np = 30 \cdot \frac{1}{500} = 0.06.$$

The corresponding Poisson distribution using (6.5) is

$$P(X = x) = \frac{e^{-0.06}(0.06)^x}{x!}; \; x = 0, 1, \ldots, 30.$$

(*a*) Probability of a packet containing no defective bolt is

$$P(X = 0) = \frac{e^{-0.06}(0.06)^0}{0!}$$

$$= e^{-0.06}$$

$$\approx 0.942.$$

Therefore, approximate number of packets out of 10000 packets containing no defective bolt are given by

$$N \times P(X=0) = 10000 \times 0.942 \approx 9420. \qquad \textit{Answer (a)}$$

(*b*) Probability of a packet containing two defective bolts is

$$P(X=2) = \frac{e^{-0.06}(0.06)^2}{2!}$$

$$= \frac{0.942 \times 0.0036}{2}$$

$$\approx 0.00170.$$

Therefore, appoximate number of packets out of 10000 packets containing two defective bolts are given by

$$N \times 0.00170 = 10000 \times 0.00170 \approx 17. \qquad \textit{Answer (b)}$$

Example 6.6

In a company, there are 250 workers. The probability of a worker remain absent on any one day is 0.02. Find the probability that on a day seven workers are absent.

Solution

Let X be the number of workers remain absent on any one day, then X assume the values 0, 1, ..., 250 and $n = 250$.

Let p be the probability of a worker remain absent on any one day, then $p = 0.02$.

Therefore,

$$\lambda = np = 250 \times 0.02 = 5.$$

The corresponding Poisson distribution using (6.5) is

$$P(X=x) = \frac{e^{-5}(5)^x}{x!}; x = 0, 1, 2, \ldots, 250.$$

The required probability that on a day seven workers remain absent is given by

$$P(X=7) = \frac{e^{-5}(5)^7}{7!} \approx 0.104. \qquad \textit{Answer}$$

Example 6.7

Potholes on a highway can be a serious problem. The past experience suggests that there are, on the average, 2 potholes per mile after a certain amount of usage. It is assumed that the Poisson process applies to the random variable "number of potholes". What is the probability that no more than 4 potholes will occur in a given section of 5 miles? *[GTU, May 2016]*

Solution

Let X be the number of potholes and λ be the average number of potholes.

Given that $\lambda = 10$.

Therefore, the corresponding Poisson distribution using (6.5) is

$$P(X=x) = \frac{e^{-10}(10)^x}{x!}; x = 0, 1, 2, \ldots.$$

The required probability that no more than 4 potholes occur in a given section of 5 miles is given by

$$P(X \le 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

$$= e^{-10}\left(\frac{10^0}{0!} + \frac{10^1}{1!} + \frac{10^2}{2!} + \frac{10^3}{3!} + \frac{10^4}{4!}\right)$$

$$\approx 0.02925. \qquad \textit{Answer}$$

6.2 The Normal Distribution (or Gaussian Distribution)

It is one of the most important and widely used continuous probability distribution. It is used to calculate the probable values of continuous variables like height, weight, length, rainfall studies, errors in scientific measurements, etc. Further, it is useful approximation of more complicated distributions.

➢ **Definition** A continuous random variable X is said to follow a ***normal distribution***, if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]; -\infty < x < \infty, \sigma > 0, \qquad \ldots(6.8)$$

where x denotes values of a random variable X, μ is expected (mean) value of X and σ is standard deviation of X.

Mean μ and variance σ^2 are called ***parameters*** of the distribution. X is called ***normal random variable***.

A random variable X follow a normal distribution with mean μ and variance σ^2 is expressed by the symbol

$$X \sim N(\mu, \sigma^2).$$

In this case, the curve for $y = f(x)$ is called a ***normal probability curve*** (or ***normal curve***), which is of bell-shaped (refer Figures 6.1, 6.2) and symmetric about the ordinate at $x = \mu$. Once μ and σ are specified, then the normal curve is completely determined.

The following are some normal probability curves.

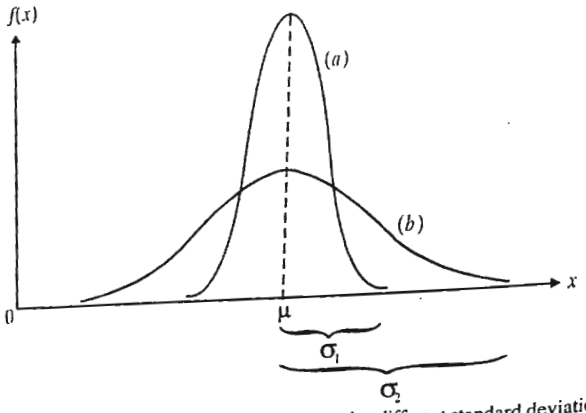


Figure 6.1 Normal probability curves with same mean but different standard deviations

The normal curve in Figure 6.1(*a*) shows the values of X, which are grouped very closely to the mean μ and so the distribution has low standard deviation (say σ_1). In Figure 6.1(*b*), the values of X are spread widely about the same mean μ and so the distribution has a high standard deviation (say σ_2, $\sigma_1 < \sigma_2$). Thus, the two curves are centered at exactly the same position on horizontal axis, but with different standard deviations.

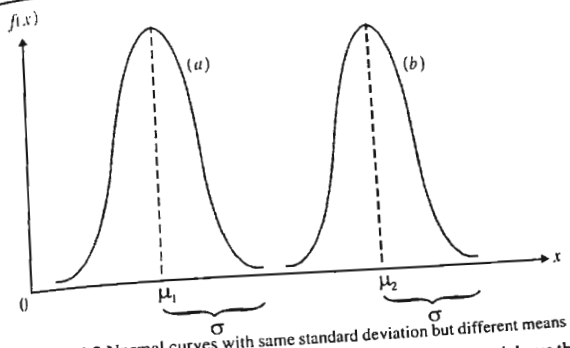


Figure 6.2 Normal curves with same standard deviation but different means

Figure 6.2 shows two normal distributions (*a*) and (*b*). Both have the same standard deviation σ but different means μ_1 and μ_2 ($\mu_1 < \mu_2$), respectively. These two curves are identical in form but are centered at different positions along the horizontal axis.

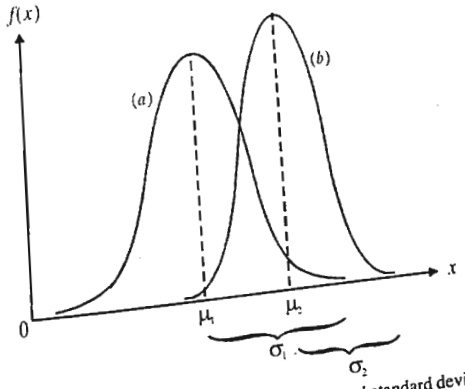


Figure 6.3 Normal curves with different means and standard deviations

Figure 6.3 shows two normal distributions (*a*) and (*b*). Both have different means μ_1, μ_2 ($\mu_1 < \mu_2$) and standard deviations σ_1, σ_2 ($\sigma_1 > \sigma_2$). These two curves are centered at different positions on the horizontal axis with different standard

➢ **Properties of a Normal Curve**

1. The curve is bell-shaped and symmetric about a vertical axis through the mean μ.
2. Since the ordinate at $x = \mu$ devides the area under the normal curve into two equal parts, so the median of the distribution coincides with the mean and the mode. The mode is the point on the horizontal axis where the curve has maximum height and which occurs at $x = \mu$.
3. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.
4. Since $f(x)$ being a probability and can never be negative and hence no portion of the curve lies below x-axis.
5. The total area under the curve and above the horizontal axis is one, which is due to the constant factor $1/\sigma\sqrt{2\pi}$.
6. The curve has its points of inflection at $x = \mu \pm \sigma$ at it is concave downward if $\mu - \sigma < X < \mu + \sigma$ and is concave upward otherwise.

6.2.1 Standard Normal Distribution Suppose we want to find the probability for that X which assumes a value between $x = x_1$ and $x = x_2$ (that is, we want to find the area under the normal curve $y = f(x)$ bounded by the two ordinates $x = x_1$ and $x = x_2$), then we have

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx, \qquad ...(6.9)$$

where $f(x)$ is given by (6.8).

The above integral (6.9) must be computed numerically as analytical integration is impossible. Moreover, as the value of mean μ or standard deviation σ altered, the value of integral (6.9) also changes (That means the area under the normal curve $y=f(x)$ between two ordinates $x=x_1$ and $x=x_2$, in general, dependent on the values of μ and σ). Before proceeding with the example (discussed below) which shows this change, it should be noted that different distributions have different mean or standard deviation. Moreover, the shape of the normal curve is dependent on these mean and standard deviation as may be seen from Figures 6.1 - 6.3.

→ *For example*, Figure 6.4 shows normal curves for two different distributions X_1 and X_2 with different means and standard deviations.

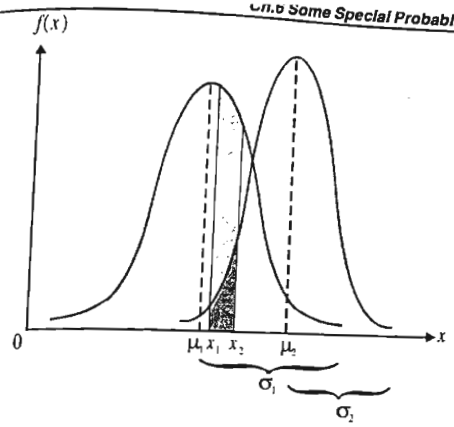


Figure 6.4

Suppose that distribution X_1 is having mean μ_1 and standard deviation σ_1, whereas distribution X_2 is having mean μ_2 ($\mu_1 < \mu_2$) and standard deviation σ_2 ($\sigma_1 > \sigma_2$).

It is obvious from the figure that the size of the shaded region for distribution X_2 (darkly shaded) is different than the size of the shaded region for distribution X_1 (entirely shaded). Therefore, the probability associated with each distribution will be different between $x = x_1$ and $x = x_2$. Because of this discussed reason, some standardization is required, which transforms observations of any normal random variable X into a new set of observations of a normal random variable Z with mean 0 and variance 1. For this, we take transformation

$$\frac{x-\mu}{\sigma} = z$$

in (6.9), which implies

$$P(z_1 < Z < z_2) = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz, \qquad ...(6.10)$$

where

$$z_1 = \frac{x_1 - \mu}{\sigma} \text{ and } z_2 = \frac{x_2 - \mu}{\sigma}.$$

The distribution

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}; \; -\infty < z < \infty \qquad ...(6.11)$$

is known as ***standard or normalized form of the normal distribution.*** In this case, Z is normally distributed with mean 0 and variance 1. Symbolically, it can be expressed as

$$Z \sim N(0,1).$$

Table I given in Appendix A indicates the area under the standard normal curve corresponding to $P(Z<z)$ as shown, in general, in Figure 6.5.

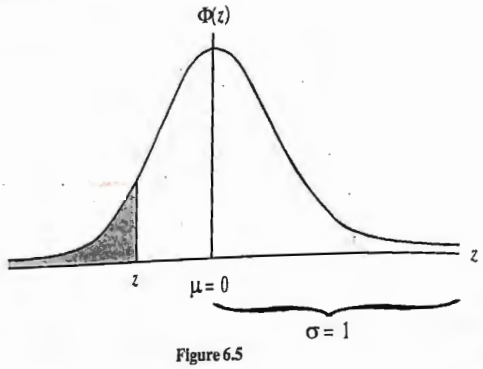


Figure 6.5

➢ **Working with the Normal Table** Let us discuss the following illustrations using Table I given in Appendix A and the symmetric property of the distribution.

Example 6.8

The continuous random variable Z has a standard normal distribution. Calculate the probability of the following.

(*a*) $Z<1.3$ (*b*) $Z>1.3$ (*c*) $Z>-1.3$ (*d*) $Z<-1.3$ (*e*) $-1.37 \le Z \le 2.01$

(*f*) $|Z| \le 0.5$ (*g*) $-1.79 \le Z \le -0.54$

Solution

(*a*) Here, the area under the curve is shown in Figure 6.6.

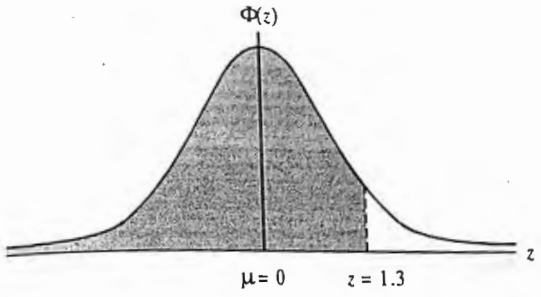


Figure 6.6

Using Table I in Appendix A,

$$P(Z<1.3) \approx 0.90320.$$ ***Answer (a)***

(*b*) Here, the area under the curve is shown in Figure 6.7.

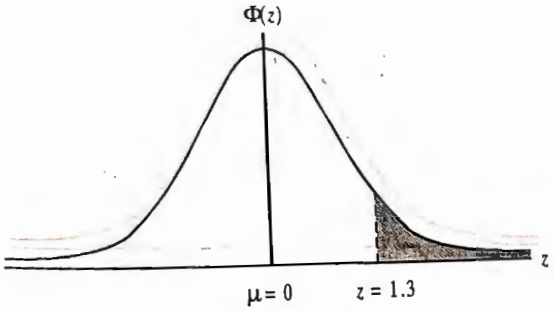


Figure 6.7

Using Table I in Appendix A,

$$P(Z>1.3) = 1 - P(Z<1.3) \approx 1 - 0.90320 = 0.0968.$$

Answer (b)

(c) Here, the area under the curve is shown in Figure 6.8.

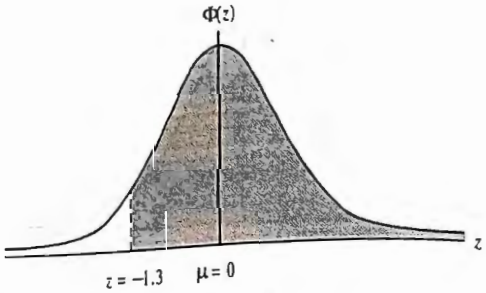


Figure 6.8

By symmetry $P(Z>-1.3)$ is identical to $P(Z<1.3)$, therefore using (a),

$$P(Z>-1.3) = P(Z<1.3) \approx 0.90320.$$ ***Answer (c)***

(d) Here, the area under the curve is shown in Figure 6.9.

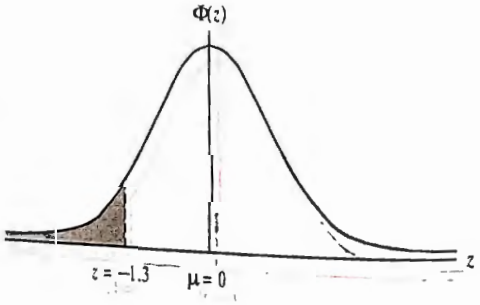


Figure 6.9

$$P(Z<-1.3) = 1 - P(Z<1.3) \approx 1 - 0.90320 = 0.0968.$$

Answer (d)

(e) Here, the area under the curve is shown in Figure 6.10.

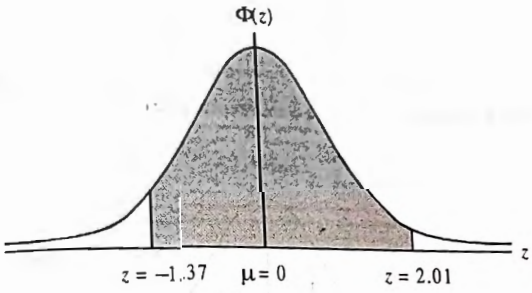


Figure 6.10

$$P(-1.37 \le Z \le 2.01) = P(-1.37 \le Z \le 0) + P(0 < Z \le 2.01)$$
$$= P(0 \le Z \le 1.37) + P(0 < Z \le 2.01)$$
$$\approx (0.91466 - 0.5) + (0.97778 - 0.5)$$
$$= 0.41466 + 0.47778$$
$$= 0.89244.$$ ***Answer (e)***

(f) Here, the area under the curve is shown in Figure 6.11.

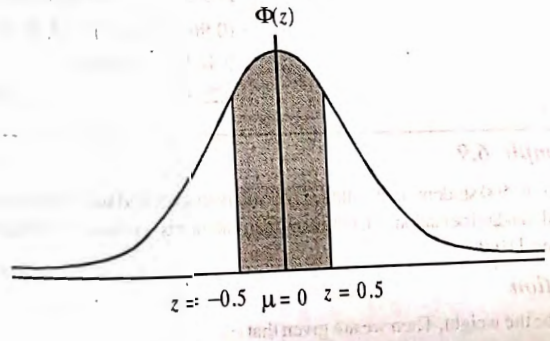


Figure 6.11

$$P(|Z| \leq 0.5) = P(-0.5 \leq Z \leq 0.5)$$
$$= 2P(0 \leq Z \leq 0.5)$$
$$= 2\,(0.69146 - 0.5)$$
$$= 2(0.19146)$$
$$= 0.38292.$$ **Answer (f)**

(g) Here, the area under the curve is shown in Figure 6.12.

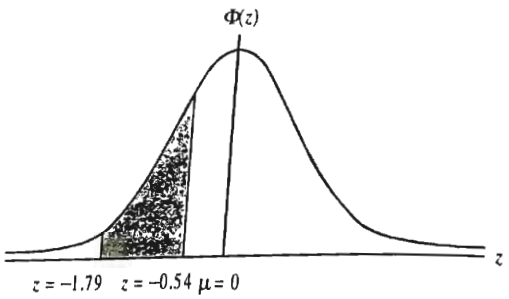


Figure 6.12

$$P(-1.79 \leq Z \leq -0.54) = P(0.54 \leq Z \leq 1.79)$$
$$= P(0 \leq Z \leq 1.79) - P(0 \leq Z < 0.54)$$
$$\approx (0.96327 - 0.5) - (0.70540 - 0.5)$$
$$= 0.46327 - 0.2054$$
$$= 0.25787.$$ **Answer (g)**

Example 6.9

Weighs of 500 students of a college is normally distributed with average weight 95 *lbs* and standard deviation 7.5. Find how many students will have the weight between 100 and 110 *lbs*.

Solution

Let X be the weight. Then we are given that

$$X \sim N[95, (7.5)^2].$$

Probability of the students who have weight between 100 and 110 *lbs* is

$$P(100 < X < 110) = P\left(\frac{100-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{110-\mu}{\sigma}\right)$$
$$= P\left(\frac{100-95}{7.5} < Z < \frac{110-95}{7.5}\right)$$
$$\approx P(0.67 < Z < 2)$$
$$= P(0 \leq Z < 2) - P(0 \leq Z \leq 0.67)$$
$$\approx (0.97725 - 0.5) - (0.74857 - 0.5)$$
$$= 0.47725 - 0.24857$$
$$= 0.22868.$$

Therefore, the number of students out of 500 students having their weights between 100 and 110 *lbs* are given by

$$N \times P(100 < X < 110)$$
$$= 500 \times 0.22868$$
$$\approx 114.$$ **Answer**

Example 6.10

The compressive strength of samples of cement can be modeled by a normal distribution with a mean 6000 kg/cm^2 and a standard deviation 100 kg/cm^2.

(a) What is the probability that a sample's strength is less than 6250 kg/cm^2?

(b) What is the probability, if sample strength is between 5800 and 5900 kg/cm^2?

(c) What strength is exceeded by 95% of the samples?

[GTU, May 2016]

Solution

Let X be the compressive strength of samples of cement. Then we are given that

$$X \sim N[6000, (100)^2].$$

(a) Probability that a sample's strength is less than 6250 kg/cm^2 is

$$P(X < 6250) = P\left(\frac{X-\mu}{\sigma} < \frac{6250-\mu}{\sigma}\right)$$

$$= P\left(Z < \frac{6250-6000}{100}\right)$$

$$= P(Z < 2.5)$$

$$\approx 0.99379.$$ **Answer (a)**

(*b*) Probability of sample's strength between 5800 and 5900 kg/cm^2 is

$$P(5800 < X < 5900) = P\left(\frac{5800-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{5900-\mu}{\sigma}\right)$$

$$= P\left(\frac{5800-6000}{100} < \frac{X-\mu}{\sigma} < \frac{5900-6000}{100}\right)$$

$$= P(-2 < Z < -1)$$

$$= P(1 < Z < 2)$$ (By symmetry)

$$= P(Z < 2) - P(Z \le 1)$$

$$\approx 0.97725 - 0.84134$$

$$= .0.13591.$$ **Answer (b)**

(*c*)

$$P(X < x) = P\left(\frac{X-\mu}{\sigma} < \frac{x-6000}{\sigma}\right)$$

$$= P\left(Z < \frac{x-6000}{100}\right)$$

$$= 0.95.$$

$$\Rightarrow \frac{x-6000}{100} \approx 1.65$$

$$\Rightarrow x = 6165.$$ **Answer (c)**

Exercises 6.1

01. In Poisson distribution, it is given that $P(X = 1) = P(X = 2)$. Find mean and standard deviation of this distribution and compute the value of $P(X = 3)$.

02. A workshop has several machines. During a typical month two machines will break down. What are the probabilities that in a month (*a*) none, (*b*) one, (*c*) more than two machine(s) will break down ?

03. A machine produces on the average of 10% defectives. Find the probability that in a sample of 10 tools chosen at random, exactly two will be defective.

04. The continuous random variable X has a standard normal distribution. Calculate the probability that (*a*) $0 < X < 1$, (*b*) $-1 < X < 1$, (*c*) $-0.5 \le X \le 2$.

05. If the average height of a certain type of corn is 12 *cm* with standard deviation of 1.8 *cm*. What percentage of these corns exceeds 14 *cms* in height, assuming that the heights are normally distributed ?

06. The following table gives the probabilities that a certain computer will malfunction 0, 1, 2, 3, 4, 5 or 6 times on any one day.

No. of malfunctions(x) :	0	1	2	3	4	5	6
Probabilities $f(x)$:	0.17	0.29	0.27	0.16	0.07	0.03	0.01

Find the mean and standard deviation of this probability distribution.

[GTU, May 2017]

07. The breaking strength $X(kg)$ of a certain type of plastic block is normally distributed with a mean of 1250 *kg* and a standard deviation of 55 *kg*. What is the maximum load such that we can expect no more than 5% of the block to break?

[GTU, May 2017]

08. Find the expectation for the following discrete probability distribution.

x :	10	14	18	25	35
$p(x)$:	0.125	0.225	0.325	0.200	0.125

[GTU, May 2017 - Comp.]

6.3 Chebyshev's Inequality

We know that the standard deviation σ gives us idea about the variability of the observations about the mean, and thus it controls the concentration of probability in the neighbourhood of the mean. For smaller values of σ, there is a high probability of getting values close to the mean. The Chebyshev's inequality, in general, gives us bounds on probability that how far a random variable X is deviated when both mean μ and variance σ^2 of the distribution are known. The inequality is also helpful when the probability distribution (either discrete or continuous) of X is not known.

Moreover, the bounds given by the inequality is universal; that is, it is the same for all random variables X with a given μ and σ^2, with the drawbacks that the bounds are not sharp in general. If there is more information about the distribution of X, then it might be possible to get a better bounds as compared to Chebyshev's inequality (refer Example 6.11). The only restriction with this inequality is that X should have finite σ^2.

Theorem 6.1 (Chebyshev's Inequality) If X is a random variable with mean μ and finite variance σ^2, then for every $k > 0$,

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \text{ or } P(|X-\mu| < k\sigma) \geq 1-\frac{1}{k^2}. \quad ...(6.12)$$

A convenient form Chebyshev's inequality can be obtained by taking $k\sigma = \rho$ and $\rho > 0$. Therefore, (6.12) becomes

$$P(|X-\mu| \geq \rho) \leq \frac{\sigma^2}{\rho^2} \text{ or } P(|X-\mu| < \rho) \geq 1-\frac{\sigma^2}{\rho^2}. \quad ...(6.13)$$

Example 6.11

Suppose that a random variable X is $N(\mu, \sigma^2)$. Compute $P(|X-\mu| \geq 2\sigma)$.

Solution

Here, X is normally distributed with $E(X) = \mu$ and $V(X) = \sigma^2$.

Now,

$$P(|X-\mu| \geq 2\sigma) = P\left(\left|\frac{X-\mu}{\sigma}\right| \geq 2\right)$$

$$= P(|Z| \geq 2)$$

$$\approx 0.0456.$$

(Using Table I in Appendix A) ...(i)

By direct application of (6.13), we have

$$P(|X-\mu| \geq 2\sigma) \leq \frac{\sigma^2}{(2\sigma)^2} = \frac{1}{4} = 0.25,$$

which is substantially large compared to the more exact value 0.0456 obtained in (i).

This example justifies the statement of the Introduction that the bounds obtained by Chebyshev's inequality may not be sharp in general. But a better bounds can be obtained if we know exact distribution of X.

Example 6.12

The number of customers who visit a car dealer's showroom on sunday morning is a random variable with mean 18 and standard deviation 2.5. What is the probability that on sunday morning the customer's will be between 8 to 28.

Solution

Given that

$$\mu = E(X) = 18 \text{ and } \sigma^2 = \text{var}(X) = (2.5)^2 = 6.25.$$

Using (6.13), the required probability is

$$P(|X-18| < 10) \geq 1-\frac{6.25}{100} \qquad (\text{Using } \rho = 10)$$

$$\Rightarrow P(8 < X < 28) \geq 1-\frac{1}{16}$$

$$\Rightarrow P(8 < X < 28) \geq \frac{15}{16}.$$

Answer

Example 6.13

Determine the smallest value of k in the Chebyshev's inequality for which the probability is at least 0.95.

Solution

Using (6.12),

$$P(|X-\mu| < k\sigma) \geq 1-\frac{1}{k^2}.$$

Given that

$$0.95 \geq 1-\frac{1}{k^2} \Rightarrow \frac{1}{k^2} \geq 1-0.95$$

$$\Rightarrow k^2 \leq \frac{1}{0.05}$$

$$\Rightarrow k^2 = \sqrt{20}.$$

Answer

Exercises 6.2

01. Suppose X is a random variable such that $E(X) = 3$ and $E(X^2) = 13$. Calculate a lower bound for the probability that X lies between -2 and 8 using Chebyshev's inequality.
02. The number of items cleared by an assembly line during a week is a random variable with mean 50 and variance 25. What can be said about the probability that this week's clearance will be between 40 to 60 ?

03. A random variable X with unknown probability distribution has a mean $\mu = 8$ and $\sigma = 3$. Find the following.

(a) $P(-4 < X < 20)$ (b) $P(|X - 8| \geq 6)$

6.4 The Exponential Distribution

Suppose a random variable follow a Poisson distribution as follows.

→ ***For example***,

(1) The number of telephone calls that arrive each day over a period of a year and note that the arrivals follow a poisson distribution with an ***average*** of 4 per day.

(2) The number of hits to your website and note that hits follow a Poisson distribution at a ***rate*** of 5 per day.

(3) The number of customers arriving at a service point and note that arrivals follow a Poisson distribution with an ***average*** of 4 per day.

In the above examples, if we consider T as the time between the events, then we have following situations.

In the first example, T indicates waiting time between calls.

In the second example, T indicates time between hits.

In the third example, T indicates time between customers.

Here, T be the time between the events happening is a ***random variable*** which follow an ***exponential distribution***. Thus, exponential distribution is typically used to model time intervals between random events. In other words, exponential distribution describes waiting time between Poisson occurrence.

It should be noted here that the number of events is a discrete variable, whereas the time between events is a continuous variable.

The exponential distribution is having the probability density function (p.d.f.)

$$f(t) = \lambda e^{-\lambda t} \; ; t \geq 0 \\ = 0 \quad ; \text{otherwise} \qquad \ldots(6.14)$$

for $\lambda > 0$.

Sometimes the p.d.f. of exponential distribution can also be specified as

$$f(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}} \; ; t \geq 0, \lambda > 0 \\ = 0 \quad ; \text{otherwise.} \qquad \ldots(6.15)$$

In (6.14) λ is known as the ***rate parameter***, whereas in (6.15) λ is known as the ***mean parameter***.

Note 1 When times between random events follow the exponential distribution with rate λ, then the total number of events in a time period of length t follows the Poisson distribution with parameter λt.

Note 2 The exponential distribution is a ***memoryless*** continuous distribution.

It is clear from the definition (6.14) that

(1) $f(t) \geq 0$ for all $t \geq 0$.

(2)

$$\int_{-\infty}^{\infty} f(t)\,dt = \lambda \int_{0}^{\infty} e^{-\lambda t}\,dt$$

$$= \lambda \left(\frac{e^{-\lambda t}}{-\lambda} \right)_0^{\infty}$$

$$= \left(-e^{-\lambda t} \right)_0^{\infty}$$

$$= 1.$$

Thus, $f(t)$ satisfies both the requirements for a function to be a probability density function.

➢ **Distribution Function** In this case the distribution function $F(t)$ is defined as

$$F(t) = P(T \leq t) = \int_0^t f(t)\,dt \qquad \text{(Using (5.19))}$$

$$= \int_0^t \lambda e^{-\lambda t}\,dt \qquad \text{(Using (6.14))}$$

$$= \lambda \left(\frac{e^{-\lambda t}}{-\lambda} \right)_0^t$$

$$= -\left(e^{-\lambda t} \right)_0^t$$

$$= 1 - e^{-\lambda t}.$$

➢ **Mean and Variance** For any $r \geq 0$,

$$E(T^r) = \int_0^{\infty} t^r f(t)\,dt \qquad \text{(Using (5.21))}$$

$$= \int_0^\infty t^r \cdot \lambda e^{-\lambda t} dt \quad \text{(Using (6.14))} \quad ...(6.15)$$

Using $\lambda t = u$, we get $\lambda\, dt = du$. Therefore, (6.16) becomes

$$E(T^r) = \int_0^\infty \left(\frac{u}{\lambda}\right)^r e^{-\lambda\left(\frac{u}{\lambda}\right)} du$$

$$= \frac{1}{\lambda^r} \int_0^\infty e^{-u} u^{(r+1)-1} du$$

$$= \frac{1}{\lambda^r} \Gamma(r+1).$$

$$\left(\text{Since } \Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx\right) \quad ...(6.17)$$

Therefore,

$$\text{Mean } \mu = E(T) = \frac{\Gamma(2)}{\lambda} \quad \text{(Using (6.17) with } r = 1)$$

$$= \frac{1}{\lambda}.$$

$$\text{(Since } \Gamma(n) = (n-1)!) \quad ...(6.18)$$

$$\text{Variance } V(T) = \sigma^2 = E(T^2) - [E(T)]^2$$

$$= \frac{1}{\lambda^2}\Gamma(3) - \left(\frac{1}{\lambda}\right)^2$$

(Using (6.17) with $r = 2$ and (6.18))

$$= \frac{2!}{\lambda^2} - \frac{1}{\lambda^2} \quad \text{(Since } \Gamma(n) = (n-1)!)$$

$$= \frac{1}{\lambda^2}. \quad ...(6.19)$$

Note Both the mean and standard deviation of the exponential distribution are $1/\lambda$.

Note The exponential distribution occurs most often in applications of Reliability theory and Queuing theory.

Example 6.14 (Lifetime of a Battery)

The lifetime T of an alkaline battery is exponentially distributed with $\lambda = 0.05$ per hour.

(*a*) What are the mean and standard deviation of the batterys lifetime ?

(*b*) What are the probabilities for the battery to last between 10 and 15 hours and to last more than 20 hours ?

Solution

(*a*) As both the mean and standard deviation of the exponential distribution are equal to $1/\lambda$. Therefore,

$$\text{Mean } \mu = \text{S.D. } \sigma = \frac{1}{\lambda} = \frac{1}{0.05} = 20 \text{ hours.} \qquad \textit{Answer (a)}$$

(*b*)

$$P(10 < T < 15) = \int_{10}^{15} \lambda e^{-\lambda t} dt \quad \text{(Using (6.14)}$$

$$= \int_{10}^{15} 0.05 e^{-0.05t} dt$$

$$= 0.05\left(\frac{e^{-0.05t}}{-0.05}\right)_{10}^{15}$$

$$= -\left[e^{-0.05(15)} - e^{-0.05(10)}\right]$$

$$= 0.1341.$$

$$P(T > 20) = \int_{20}^{\infty} \lambda e^{-\lambda t} dt \quad \text{(Using (6.14))}$$

$$= \int_{20}^{\infty} 0.05 e^{-0.05t} dt$$

$$= 0.05\left(\frac{e^{-0.05t}}{-0.05}\right)_{20}^{\infty}$$

$$= -[0 - e^{-0.05(20)}]$$

$$= e^{-0.05(20)}$$

$$= 0.3679.$$

Example 6.15

Accidents occur with a Poisson distribution at an average of 2 per week; that is, $\lambda = 2$.

(*a*) Calculate the probability of more than 3 accidents in any one week.

(*b*) What is the probability that at least two weeks will elapse between accidents ?

Solution

(*a*)

$$P(X > 3) = 1 - P(X \le 3)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$= 1 - \left[\frac{e^{-2}(2)^0}{0!} + \frac{e^{-2}(2)^1}{1!} + \frac{e^{-2}(2)^2}{2!} + \frac{e^{-2}(2)^3}{3!}\right]$$

(Using (6.5))

$$= 1 - \left(e^{-2} + 2e^{-2} + 2e^{-2} + \frac{4}{3}e^{-2}\right)$$

$$= 1 - \left(1 + 2 + 2 + \frac{4}{3}\right)e^{-2}$$

$$= 1 - \frac{19}{3}e^{-2}$$

$$\approx 0.14288.$$

Answer (a)

(*b*)

$$P(X > 2) = \int_2^{\infty} 2e^{-2t}\,dt$$

(Using (6.14))

$$= 2\left(\frac{e^{-2t}}{-2}\right)_2^{\infty}$$

$$= -[0 - e^{-2(2)}]$$

$$= e^{-4}$$

$$\approx 0.01832.$$

Answer (b)

Example 6.16

The time between breakdowns of a particular machine follows an exponential distribution with a mean of 17 days. Calculate the probability that a machine breaks down in a 15 day period.

Solution

The p.d.f. $f(t)$ is given by

$$f(t) = \frac{1}{17}e^{-\frac{t}{17}}; t \ge 0.$$

The required probability is given by

$$P(0 \le T \le 15) = \int_0^{15} f(t)\,dt$$

$$= \int_0^{15} \frac{1}{17}e^{-\frac{t}{17}}\,dt$$

$$= \frac{1}{17}\left[\frac{e^{-\frac{t}{17}}}{\left(-\frac{1}{17}\right)}\right]_0^{15}$$

$$= -e^{-\frac{t}{17}}\Big|_0^{15}$$

$$= -e^{-\frac{15}{17}} + 1$$

$$= 0.5862.$$

Thus, there is a 58.62% chance that the machine will breakdown in a 15 day period.

Answer

Example 6.17

A system contains a certain type of component whose lifetime T is exponentially distributed with mean of 5 years. If 8 such components are installed in different systems, what is the probability that at least 3 are still working at the end of 7 years ?

Solution

The p.d.f. $f(t)$ is given by

$$f(t)=\frac{1}{5}e^{-\frac{t}{5}};\ t\geq 0.$$

Therefore,

$$P(T>7)=\frac{1}{5}\int_7^\infty e^{-\frac{t}{5}}dt$$

$$=\frac{1}{5}\left[\frac{e^{-t/5}}{(-1/5)}\right]_7^\infty$$

$$=-e^{-\frac{t}{5}}\Big|_7^\infty$$

$$=e^{-\frac{7}{5}}$$

$$=0.1827.$$

Let n be the number of components out of 8 working after 7 years of instalment, then

$$P(n\geq 3)=\sum_{n=3}^{8} {}^8C_n(0.1827)^n(1-0.1827)^{8-n}$$

(Using (6.2))

$$=\sum_{n=3}^{8} {}^8C_n(0.1827)^n(0.8173)^{8-n}$$

$$=1-\sum_{n=0}^{2} {}^8C_n(0.1827)^n(0.8173)^{8-n}$$

$$=1-[\ {}^8C_0(0.1873)^8+{}^8C_1(0.1827)(0.8173)^7$$

$$+{}^8C_2(0.1827)^2(0.8173)^6\ ]$$

$$=1-(0.1991+0.3560+0.2786)$$

$$=0.1663.$$

Answer

Example 6.18

The arrival rate of cars at a gas station is $\lambda=40$ customers per hour.
(*a*) What is the probability of having no arrivals in a 5 minute interval ?
(*b*) What are the mean and variance of the number n of arrivals in 5 minutes ?
(*c*) What is the probability for having 3 arrivals in a 5 minute interval ?

Solution

(*a*)

$$P\left(T>\frac{5}{60}\right)=\int_{5/60}^{\infty} 40e^{-40t}dt \qquad \text{(Using (6.14))}$$

$$=40\left(\frac{e^{-40t}}{-40}\right)_{5/60}^{\infty}$$

$$=-\left(e^{-40t}\right)_{5/60}^{\infty}$$

$$=e^{-40}\left(\frac{5}{60}\right)$$

$$\approx 0.03567.$$

Answer (a)

(*b*) Here, the variable n has a Poisson distribution with parameter

$$\mu=\lambda t=40\left(\frac{5}{60}\right)=3.333. \qquad \text{(Using } \textit{Note 1}\text{)}$$

Hence,

Mean $E(n)=3.333$ and Variance $V(n)=3.333$.

since mean and variance of Poisson distribution are same.

Answer (b)

(c)

$$P(N=3)=\frac{e^{-3.333}(3.333)^3}{3!} \qquad \text{(Using (6.5))}$$

$$\approx 0.2202. \qquad \textit{Answer (c)}$$

Exercises 6.3

01. If jobs arrive every 15 seconds on average; that is, $\lambda = 4$ per minute. What is the probability of waiting less than or equal to 30 seconds ?
02. If on the average three trucks arrive per hour to be unloaded at a warehouse, using exponential distribution find the probabilities that the time between the arrival of successive trucks will be
    (*a*) less than 5 minutes, (*b*) at least 45 minutes.
03. The length of time for one person to be served at a cafeteria is a random variable T having an exponential distribution with a mean of 4 minutes. Find the probability that a person is served in less than 3 minutes on at least 4 of the next 6 days.
04. The mean time taken by an engineer to repair an electrical fault in a system is 2.7 hours. Calculate the probability that the engineer will repair a fault in less than the mean time.

6.5 The Gamma Distribution

Suppose that a system consisting of one original and $(\alpha-1)$ spare components such that in the case of failure of original component, one of the $(\alpha-1)$ spare components can be used. The process will continue till we use last component. When last component fails, then the whole system fails. Let $X_1, X_2, \ldots, X_\alpha$ be the lifetimes of the α components. Let each of the random variables $X_1, X_2, \ldots, X_\alpha$ have the same exponential distribution with parameter λ, and also are probabilistically independent. Then the lifetime (time until failure) of the entire system is given by

$$T=\sum_{i=1}^{\alpha} X_i,$$

and which has ***gamma distribution*** with p.d.f.

$$f(t)=\begin{cases}\dfrac{\lambda^\alpha t^{\alpha-1}e^{-\lambda t}}{\Gamma(\alpha)} & ;t\geq 0\\ 0 & ;\text{otherwise}\end{cases} \qquad ...(6.20)$$

for $\lambda, \alpha > 0$. The parameter α is called the ***shape parameter***, and the parameter λ is called the ***rate parameter*** (as in the case of exponential distribution).

Thus, the sum of α independent exponential random variables has a gamma

distribution.

→ ***For example***, if we replace a light bulb immediately after it burns out, then the total time that a successive light bulbs last has a gamma distribution.

Introducing $v=\lambda t$, (6.20) reduces to

$$f(v)=\frac{1}{\lambda}f\left(\frac{v}{\lambda}\right)=\frac{1}{\lambda}\left[\frac{\lambda^\alpha\left(\frac{v}{\lambda}\right)^{\alpha-1}e^{-v}}{\Gamma(\alpha)}\right]$$

$$=\begin{cases}\dfrac{v^{\alpha-1}e^{-v}}{\Gamma(\alpha)} & ;v\geq 0\\ 0 & ;\text{otherwise}\end{cases} \qquad ...(6.21)$$

This p.d.f. of the random variable V is known as the ***standard gamma function*** with parameter α.

When $\alpha = 1$ in (6.20), then

$$f(t)=\lambda e^{-\lambda t} \quad ;t\geq 0$$
$$=0 \quad ;\text{otherwise}$$

which is the p.d.f. of the exponential distribution.

It is clear from the definition (6.21) that

(1 $f(v)\geq 0$ for all $t\geq 0$.

(2

$$\int_{-\infty}^{\infty} f(v)dv=\int_0^\infty \frac{e^{-v}v^{\alpha-1}}{\Gamma(\alpha)}dv$$

$$=\frac{1}{\Gamma(\alpha)}\int_0^\infty e^{-v}v^{\alpha-1}\,dv$$

$$=\frac{1}{\Gamma(\alpha)}\Gamma(\alpha)$$

$$\left(\text{Since } \Gamma(n)=\int_0^\infty e^{-x}x^{n-1}dx\right)$$

$$=1.$$

Thus, $f(v)$ satisfies both the requirements for a function to be a probability density function.

➢ **Mean and Variance** For any $r \geq 0$,

$$E(T^r) = E\left(\frac{V^r}{\lambda^r}\right)$$

$$= \frac{1}{\lambda^r} E(V^r)$$

$$= \frac{1}{\lambda^r} \int_0^\infty v^r \cdot f(v)dv \qquad \text{(Using (5.21))}$$

$$= \frac{1}{\lambda^r} \int_0^\infty v^r \cdot \frac{v^{\alpha-1} e^{-v}}{\Gamma(\alpha)}$$

(Using (6.21)) ...(6.22)

$$= \frac{1}{\lambda^r \Gamma(\alpha)} \int_0^\infty e^{-v} v^{\alpha+r-1} dv$$

$$= \frac{1}{\lambda^r \Gamma(\alpha)} \Gamma(\alpha+r).$$

$$\left(\text{Since } \Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx\right) \quad ...(6.23)$$

Therefore,

$$\text{Mean } \mu = E(T) = \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)}$$

(Using (6.23) with $r = 1$)

$$= \frac{\alpha\Gamma(\alpha)}{\lambda\Gamma(\alpha)}$$

(Using $\Gamma(n+1) = n\Gamma(n)$)

...(6.24)

$$= \frac{\alpha}{\lambda}.$$

$$\text{Variance } \sigma^2 = E(T^2) - [E(T)]^2$$

$$= \frac{\Gamma(\alpha+2)}{\lambda^2\Gamma(\alpha)} - \left(\frac{\alpha}{\lambda}\right)^2$$

(Using (6.23) with $r = 2$ and (6.24))

$$= \frac{\alpha(\alpha+1)\Gamma(\alpha)}{\lambda^2\Gamma(\alpha)} - \frac{\alpha^2}{\lambda^2}$$

(Using $\Gamma(n+1) = n\Gamma(n)$)

$$= \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2}$$

$$= \frac{\alpha}{\lambda^2}. \qquad ...(6.25)$$

Note 1 Gamma distribution is a skewed distribution.

Note 2 In (6.42) $\alpha > 0$ but need not be an integer. If α is a positive integer, than (6.42) is known as Erlang distribution.

Example 6.19

The daily consumption of milk in a city, in excess of 20000 *litres*, is approximately distributed as a gamma variate with parameters $\alpha = 2$ and $\lambda = 1/10000$. The city has a daily stock of 30000 *litres*. What is the probability that the stock is insufficient on a particular day ?

Solution

If X denotes the daily comsumption, in excess of 20000 *litres*, then p.d.f. of X is

$$f(x) = \frac{\left(\frac{1}{10000}\right)^2 x^{2-1} e^{-\frac{1}{10000}x}}{\Gamma(2)} \qquad \text{(Using (6.20))}$$

$$= \frac{x e^{-\frac{1}{10000}x}}{(10000)^2 \Gamma(2)}.$$

The stock of 30000 *litres* will be insufficient on a particular day, if the excess consumption is more than 10000 *litres*.

Therefore,

$$P(X>10000)=\int_{10000}^{\infty}\frac{xe^{-\frac{1}{10000}x}}{(10000)^2\Gamma(2)}dx. \qquad ...(i)$$

Let $$\frac{x}{10000}=t\Rightarrow\frac{dx}{10000}=dt.$$

Therefore, (i) becomes

$$P(X>10000)=\int_{1}^{\infty}te^{-t}dt \qquad (\text{Since } \Gamma(2)=1!)$$

$$=t\left(\frac{e^{-t}}{-1}\right)-1\left(\frac{e^{-t}}{1}\right)\Bigg|_1^{\infty}$$

(By Leibniz rule)

$$=0-\left(-\frac{1}{e}-\frac{1}{e}\right)$$

$$=\frac{2}{e}$$

$$\approx 0.736.$$

Answer

Example 6.20

The daily consumption of electric power (in millions if kW hours) in a certain city is a random variable X having p.d.f.

$$f(x)=\frac{1}{9}xe^{-\frac{x}{3}};\quad x>0$$

$$=0\qquad;\quad x\le 0.$$

Find the probability that the power supply is inadequate on any given day if the daily capacity of the power plant is 12 million kW hours.

Solution

The given p.d.f. is for gamma distribution for $\alpha=2$ and $\lambda=1/3$.

Now, the power supply is inadequate when $X>12$. Therefore,

$$P(X>12)=\int_{12}^{\infty}\frac{1}{9}xe^{-\frac{x}{3}}dx. \qquad \text{(Using (6.20))} ...(i)$$

Let $$\frac{x}{3}=t\Rightarrow\frac{dx}{3}=dt.$$

Therefore, (i) becomes

$$P(X>12)=\int_{4}^{\infty}\frac{t}{3}e^{-t}\cdot 3dt$$

$$=\int_{4}^{\infty}te^{-t}dt$$

$$=t\left(\frac{e^{-t}}{-1}\right)-1\left(\frac{e^{-t}}{1}\right)\Bigg|_4^{\infty}$$

(Using Leibniz rule)

$$=0-(-4e^{-4}-e^{-4})$$

$$=\frac{4}{e^4}+\frac{1}{e^4}$$

$$=\frac{5}{e^4}$$

$$\approx 0.09158.$$

Answer

Exercises 6.4

01. The survival time in weeks of an animal when subjected to certain exposure of gamma radiation has a gamma distribution with $\alpha=5$ and $\lambda=1/10$.
    (*a*) What is the mean survival time of a randomly selected animal of the type used in the experiment ?
    (*b*) What is the probability that an animal survives more than 30 weeks ?
02. Suppose that the reaction time X has a standard gamma distribution with $\alpha=2$. Find (*a*) $P(3\le X\le 5)$ (*b*) $P(X>4)$

Answers

Exercises 6.1

01. 2, $\sqrt{2}$, 0.18 **02.** (*a*) 0.135 (*b*) 0.271 (*c*) 0.323 **03.** 0.1839

04. (*a*) 0.3413 (*b*) 0.6826 (*c*) 0.6687 **05.** 13.35%

06. 1.8, 1.3416 (**Hint** Use formula (6.6) and (6.15)) **07.** 1160 *kg* **08.** 19.625

Exercises 6.2

01. 0.84 **02.** 0.75 **03.** (*a*) $\geq 15/16$ (*b*) $\leq 1/4$

Exercises 6.3

01. 0.86 **02.** (*a*) 0.221 (*b*) 0.105. **03.** $1-e^{-3/4}$, 0.3968 **04.** 0.6321

Exercises 6.4

01. (*a*) 50, (*b*) 0.8155 **02.** (*a*) 0.15872 (*b*) 0.09158

Chapter 7

Concept of Sampling and Testing of Hypothesis

7.1 Population and Sample

(1) Population (or Universe) and sample The group (say of size N) of units or items or individuals (animate or inanimate) or observations or objects forming a subject matter of statistical investigation for their various characteristics is known as ***population*** (refer Figure 7.1). Population refers totality of all relevant data.

The population may be ***finite*** or ***infinite*** depending on the size N being ***finite*** or ***infinite***. It may be ***real*** or ***hypothetical***.

For every inquiry and statistical investigation complete enumeration of entire population may not be feasible, affordable and practicable. Therefore, a finite subset of population is selected by some scientific method with a view of estimating population characteristics is known as ***sample*** (refer Figure 7.1); that is, a part or portion of units selected from population on the basis of some definite norm is called a ***sample***. The size of the sample (number of units contained in the sample) is denoted by n ($< N$). *When $n \geq 30$, the sample size is said to be large and when $n < 30$, the sample size is said to be small.*

→ ***For example***, for the population of India, population of Gujarat is a sample.

The process of drawing samples from a given population is knwon as ***sampling***.

There are many methods of sampling – Random sampling, Simple random sampling, Stratified sampling, Systematic sampling, Purposive sampling.

➢ **Random Sampling** If each unit of the population has the same chance of being selected in the sample, then the sampling is said to be ***random sampling***.

In random sampling ***with replacement***, each unit of the population may be chosen more than once since the unit is replaced in the population. Thus, sampling from finite population with replacement can be considered theoretically as sampling from infinite population. Whereas, in random sampling ***without replacement***, each

unit of the population can be chosen only once since the unit is not replaced in the population.

Thus, we say that sampling is done from infinite population when it is drawn from infinite population or it is drawn from finite population with replacement.

Again, we say that sampling is done from finite population when it is drawn from finite population without replacement.

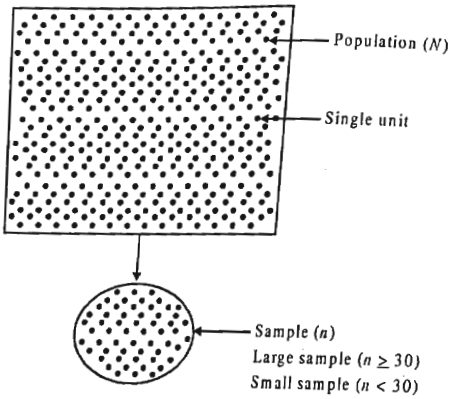


Figure 7.1 Population and Sample

(2) Population parameters and sample statistics Statistical measures or constants obtained from the population (such as population mean μ, population variance σ^2, etc.) are known as ***population parameters*** or simply ***parameters***. It is based on all units of population. Population parameter is denoted by 'θ'.

Statistical measures obtained from sample observations alone (such as sample mean $\bar{x}$, sample variance σ^2, etc.) are known as ***sample statistics*** or simply ***statistics***. It is based only on units of a sample selected from population. Sample statistic is denoted by 't'.

The following Table 7.1 indicates the notations used for different statistical measures for parameters and statistics.

Statistical measure	Parameter (θ)	Statistic (t)
Size	N	n
Mean	μ	$\bar{x}$
S.D.	σ	s
Variance	σ^2	s^2
Proportion	P	p
Correlation	ρ	r

Table 7.1

For the same statistical measure, the statistic t is used to estimate parameters θ.

Note that a statistic varies from sample to sample (and hence it is a random variable) but the parameter remains a constant. This variation in the value of a statistic is known as ***sampling fluctuation***. As statistic is a random variable, so it must have a ***probability distribution***.

7.2 Sampling Distribution

Given a finite population of size N, draw all possible samples each of the same size n. Then the total number of all possible samples each of the same size n is given by

$$^{N}C_{n} = \frac{N!}{(N-n)!n!} = k \text{ (say)}. \qquad ...(7.1)$$

For each of these samples, compute a statistic t (for example, mean, variance, etc.), then t will vary from sample to sample as shown below.

Sample number	1	2	3 ...	k
Statistic t	t_1	t_2	t_3	t_k

The aggregate of all these values of t, together with their relative frequencies or the probabilities with which they occur, constitutes the ***sampling distribution*** of the statistic t.

If the statistic t used is the sample mean, then the distribution is called the sampling distribution of means or the sampling distribution of the mean. The similar meanings we have for distribution when statistic t used are variance, proportion, etc. Refer Example 7.1 for better understanding.

Note When the number of samples each of size n is infinitely large (that is, sampling without replacement), then the probability distribution of the statistic is the sampling distribution of the statistic.

Example 7.1

A population consisting of 5 members 3, 5, 7, 9, 11. If a random sample of size $n = 2$ is selected, find the sampling distribution of the sample mean $\bar{x}$. Also, find mean of 'sample means', variance of 'sample means', population mean and population variance.

Solution

Using (7.1),

$$^5C_2 = \frac{5!}{(5-3)!2!} = \frac{5!}{3!\,2!} = 10.$$

Thus, there are 10 possible equally likely random samples of size $n = 2$. The values of $\bar{x}$ for random sampling when $n = 2$ and $N = 5$ are given by Table 7.2.

Sample	Sample units	Sample mean $\bar{x}$
1	3,5	4
2	3,7	5
3	3,9	6
4	3,11	7
5	5,7	6
6	5,9	7
7	5,11	8
8	7,9	8
9	7,11	9
10	9,11	10

Table 7.2

Hence, sampling distribution of the sample mean $\bar{x}$ is

$\bar{x}$:	4	5	6	7	8	9	10
f :	1	1	2	2	2	1	1
$p(\bar{x})$:	0.1	0.1	0.2	0.2	0.2	0.1	0.1

Now,

$$\text{Mean of sample means} = \frac{(1)(4)+(1)(5)+(2)(6)+(2)(7)+(2)(8)+(1)(9)+(1)(10)}{10}$$

(Using (1.2))

$$= \frac{4+5+12+14+16+9+10}{10}$$

$$= \frac{70}{10}$$

$$= 7.0.$$

$$\text{Variance of sample means} = \frac{(1)(-3)^2+(1)(-2)^2+(2)(-1)^2+(2)(0)^2+(2)(1)^2+(1)(2)^2+(1)(3)^2}{10}$$

(Using (1.18))

$$= \frac{9+4+2+0+2+4+9}{10}$$

$$= \frac{30}{10}$$

$$= 3.0.$$

Population mean,

$$\mu = \frac{3+5+7+9+11}{5} = 7.$$ (Using (1.1))

Population variance,

$$\sigma^2 = \frac{(-4)^2+(-2)^2+(0)^2+(2)^2+(4)^2}{5} = 8.$$

(Using (1.15)) ***Answer***

Note Observe that mean of the sample means is the same as the population mean but variance of the sample means is not the same as the population variance.

7.3 Statistical Inference

The theory of statistical inference (also known as Decision theory) consists of those methods by which one makes inferences or generalisation about a population based on the information provided by samples selected from the population.

The theory of statistical inference is divided into two major areas as follows.

(a) Estimation Here, population parameters are estimated on the basis of sample information.

→ ***For example***, A candidate for a local M.L.A. may wish to estimate the true proportion of voters that favour him by obtaining the opinions from a random sample of 1000 eligible voters. Then the fraction of voters in the sample who favour the candidate could be used as an *estimate* of the true proportion of the population of voters.

→ ***For example***, a manufacturer is interested in estimating the average life of his product, proportion of defective items in his lot, average demand of his product, etc.

(2) Hypothesis Testing Here, some hypothetical statement is made about the population parameter and it is tested at certain level of significance to check whether the made hypothesis is correct or incorrect on the basis of sample information.

→ *For example*, consider a case in which a housewife is interested to find out whether brand *A* detergent cleans brighten than brand *B*. Then she may start with the hypothesis that brand *A* is better and after proper testing (statistical) she accept or reject the hypothesis. In this case, one does not estimate a parameter but instead tries to arrive at a correct conclusion about a presented hypothesis.

7.3.1 Statistical Hypothesis and Its Testing

➢ **Statistical Hypothesis** It is an assumption or statement (which may or may not be true) concerning one or more populations.

→ *For example*,

(1) Population mean $\mu = 35$

(2) The average marks of students of IT classes and CE classes of GTU are same.

➢ **Testing of Hypothesis** The truth and falsity of the statistical hypothesis can be known certainly only if we examine the entire population, and which in most of the cases impractical. Instead, we take a random sample from the population and use the information contained in this sample to decide whether the hypothesis is likely to be true or false.

When the evidence from the sample is inconsistent with the stated hypothesis, then there is a rejection of hypothesis, whereas when evidence supporting the hypothesis, then there is a acceptance of hypothesis. It should be noted here that acceptance of a statistical hypothesis does not necessarily imply that it is true - the fact is that we have not sufficient evidence from the samples drawn to reject it. On the other hand, rejection of a hypothesis means that we have evidence that it is false. So the statisticians usually make hypothesis which they hope to reject. Hypothesis that one formulate with the hope of rejection is called ***null hypothesis*** (H_0). The rejection of null hypothesis leads to the acceptance of an ***alternative hypothesis*** (H_1). A null hypothesis will always be stated with an exact value of the population parameter, whereas the alternative hypothesis allows the possibility of one or several values.

The conventional approach to hypothesis testing is not to construct a single hypothesis about the population parameter but rather to construct two different hypothesis complementary to each other in such a manner that if one is accepted, the other will automatically be rejected and vice versa. These two hypothesis are called null hypothesis and alternative hypothesis.

• **Null Hypothesis (H_0)** It is the statistical hypothesis which is to be actually ***tested for acceptance or rejection***. The null hypothesis is usually a hypothesis of ***no-difference*** and is tested for possible rejection under the assumption that it

is true.

• **Alternative Hypothesis (H_1)** Any hypothesis, which is complementary to the null hypothesis, is called an alternative hypothesis.

There are also other types of hypothesis.

• **Simple Hypothesis** It is the statistical hypothesis which completely specify the population. Null hypothesis is always a simple hypothesis stated as an equality.

→ *For example*,

(1) $H : \mu = 0,\ \sigma = 1$ in the case of standard normal distribution.

(2) $H : n = 5,\ p = 0.6$ in the case of binomial distribution.

• **Composite Hypothesis** It is the statistical hypothesis which does not completely specify the population.

→ *For example*,

(1) $H : \mu = 50,\ \sigma$ is unknown

(2) $H : \mu$ is unknown, $\sigma = 2$

(3) $H : \mu > \mu_0,\ \sigma = \sigma_0$

The example concerning all above types of hypothesis is as follows.

Suppose for any population null hypothesis

$H_0 : \mu = \mu_1$ - simple hypothesis

Then alternative hypothesis H_1 could be any of the following.

(a) $H_1 : \mu \neq \mu_1$ - composite hypothesis

(b) $H_1 : \mu < \mu_1$ - simple hypothesis

(c) $H_1 : \mu > \mu_1$ - simple hypothesis

The alternative hypothesis (a) is known as ***two-tailed hypothesis***, (b) is known as ***left one-tailed hypothesis*** and (c) is known as ***right one-tailed hypothesis*** (defined later)

➢ **Statistical Decisions** Statistical decisions are decisions or conclusions about the population parameters on the basis of a random sample from the population.

➢ **Test Statistic** The statistical decision to accept or reject null hypothesis is made on the basis of a statistic (called test statistic) and it is computed using particular formula depending upon the probability distribution followed by it. The main test statistics are z, χ^2, t, F (which are discussed later)

➢ **Errors in Testing** As the statistical decisions of acceptance or rejection of null hypothesis are based on random sample from population, so there are scopes of committing errors. The following situations arise (including errors) while testing null hypothesis.

Nature of null hypothesis H_0	Accept H_0	Reject H_0
H_0 is true	Correct decision	Type I error
H_0 is false	Type II error	Correct decision

where

Type I error indicates H_0 is true but it is rejected by the test (instead H_1 is accepted) - Rejection error.

Type II error indicates H_0 is false but it is accepted by the test (and reject H_1). - Acceptance error.

Probability of Type I error is denoted by α and probability of Type II error is denoted by β; that is,

$$P(\text{Type I error}) = P(\text{Reject } H_0 / H_0 \text{ is true}) = \alpha$$

$$P(\text{Type II error}) = P(\text{Accept } H_0 / H_0 \text{ is false}) = \beta$$

Note 1 In statistical quality control, while deciding acceptance or rejection of a lot after testing a sample from the population, Type I error rejects a lot when it is good and Type II error accepts a lot when it is bad. Thus, α and β are often referred to as producer's risk and consumer's risk, respectively.

Note 2 In most of decision making problems in business and social science it is more risky to accept a wrong hypothesis than to reject correct one; that is, Type II errors are more serious than Type I error. Therefore, it is advisable to minimize Type II error after fixing up the less serious error Type I.

For a fixed sample, a decrease in the probability of one type of error will usually result in an increase in the probability of the other type of error. Both types of errors can be reduced only by increasing the sample size n.

Note 3 The factor $1-\beta$, the probability of rejecting H_0 when a specific hypothesis H_1 is true is called the ***power of a test***.

➢ **Level of Significance (Size of the Test)(α)** In any test procedure it is advisible to keep both types of errors (Type I and Type II) minimum. But as such both the errors are interreleted practically, so it is not possible to minimize both simultaneously. Hence, in practice, the probability of Type I error is fixed and the probability of Type II error is minimized. The maximum probability with which we would be ready to risk a Type I error is called the ***level of significance*** of the test. It is denoted by α as mentioned above.

In practice a level of significance of $\alpha = 0.01$ or $\alpha = 0.05$ is customary, although other values are used. $\alpha = 0.01$ is used for high precision, whereas $\alpha = 0.05$ is used for moderate precision. Level of significance is also expressed as percentage. Choosing $\alpha = 5\%$ in designing a test of hypothesis means that there are about 5 chances out of 100 that null hypothesis (H_0) is rejected when it is true; that is, we are about 95% ***confident*** that our decision is right. In this case we say that the hypothesis has been rejected at a 5% level of significance, which means that we could be wrong with 0.05 chance.

➢ **Level of Confidence ($1-\alpha$)** As discussed above, level of confidence is complementary to the level of significance. It is $1-\alpha$. If the level of significance is 1%, it implies that level of confidence is 99%.

➢ **Degree of Freedom** Degree of freedom is the number of independent observations of the samples. The number of independent observations is different for different statistics. The degree of freedom is denoted and defined as

$$\nu = n - k;\ \nu > 0,$$

where n is the sample size and k is the number of independent constraints imposed on the observations in the sample.

→ ***For example***, suppose that we asked to select ***any five*** observations, then there is no restriction on the selection of these observations and we are free to select any five observations. Hence, the degree of freedom is

$$\nu = n - k = 5 - 0 = 5.$$

Let us take another situation. Suppose that we asked to select five observations whose sum is 100. then, here, we are free to select any four observations but 5th observation will automatically be selected by virtue of the restriction of total 100. Hence, the degree of freedom for selection is

$$\nu = n - k = 5 - 1 = 4.$$

➢ **Critical Region** In any test of hypothesis, a test statistic t^* calculated from the sample data, is used to accept or reject the null hypothesis of the test. Consider the area under the probability curve of the sampling distribution of the test statistic t^*, which follow some known (given) distribution. The area under the probability curve is divided into two region (by predetermined level of significance) - the region of rejection (critical region of regions of significance) where null hypothesis H_0 is rejected, and the region of acceptance (or non-critical region or region of non-significance) where null hypothesis H_0 is accepted. The area of the critical region is the level of significance of the test α. It should be noted that critical region always lies on the tail(s) of the distribution. Depending on the nature of the alternative hypothesis, critical region may lie on one side or both sides of the tail(s), and

accordingly we have one tailed or two tailed test.

The value of the test statistic t_α^*, which separates the critical region and acceptance region is (are) known as ***critical value(s)***.

The critical value depends on

(a) the test statistic

(b) level of significance

(c) nature of H_1 (one tailed or two tailed)

(d) degrees of freedom

• **One-tailed and Two-tailed Tests** Suppose that under a given hypothesis the sampling distribution of a statistic t is a normal distribution with mean and S.D. Then the distribution of the standard variable, where

$$z = \frac{t - \mu_t}{\sigma_t}$$

is the standard normal distribution with mean 0 and variance 1.

A test of any hypothesis is one-tailed or two-tailed can be determined by the alternative hypothesis H_1 as follows.

(1) If

H_1 carries > sign; that is, $\mu > \mu_1$,

then the test hypothesis is known as ***right one-tailed*** test.

At a level of significance α, the critical region is shown in Figure 7.2.

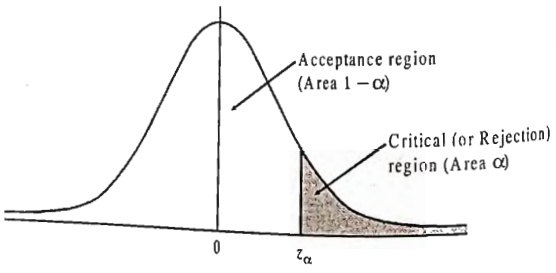


Figure 7.2 Right one-tailed test

(2) If

H_1 carries < sign, that is, $\mu < \mu_1$,

then the test hypothesis is known as ***left one-tailed test***.

At a level of significant α, the critical region is shown in Figure 7.3.

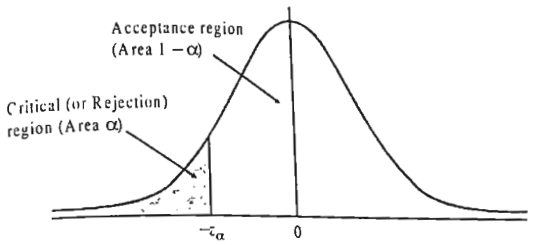


Figure 7.3 Left one-tailed test

(3) If

H_1 carries ≠ sign; that is, $\mu \neq \mu_1$,

then the test hypothesis is known as ***two-tailed test***.

At a level of significance α, the critical region is shown in Figure 7.4.

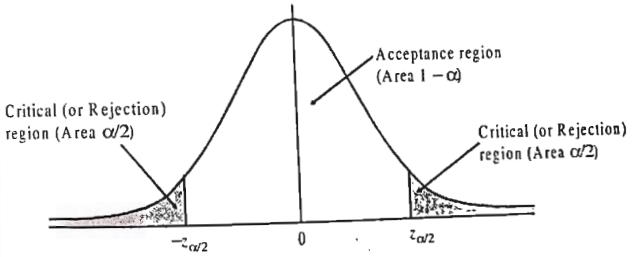


Figure 7.4 Two-tailed test

In all above cases, the total shaded area α is the level of significance of the test. It represents the probability of making Type I error. Thus, we say that the hypothesis is rejected at a α level of significance or z-value of the given sample statistic is significant at a α level of significance.

➢ **Rules of Decision** Reject the hypothesis at a α level of significance, if

(a) positive values of the statistic t lie outside α on the right side tail of the curve (refer Figure 7.2)

or

(b) negative values of the statistic t lie outside $(-\alpha)$ on the left side tail of the curve

(refer Figure 7.3)

(c) the values of the statistic t lie outside the range $-\alpha/2$ to $\alpha/2$ of the curve (refer Figure 7.4.)

➢ **Standard Error (S.E.)** The standard deviation of the sampling distribution of a statistic t is known as its ***standard error***. The S.E. is used to set up confidence limits for population parameters and in tests of significance. As the sample size n increases, S.E. decreases.

→ *For example*, in Example 7.1 S.E. is $\sqrt{3}$.

➢ **Confidence Interval** Let μ_t and σ_t be respectively mean and standard deviation of the sampling distribution of a statistic t. Then $(1-\alpha)$ 100% confidence interval for μ_t is given by

$$t - Z_{\alpha/2}\sigma_t < \mu_t < t + Z_{\alpha/2}\sigma_t \quad \text{...(7.2)}$$

(refer Figure 7.5).

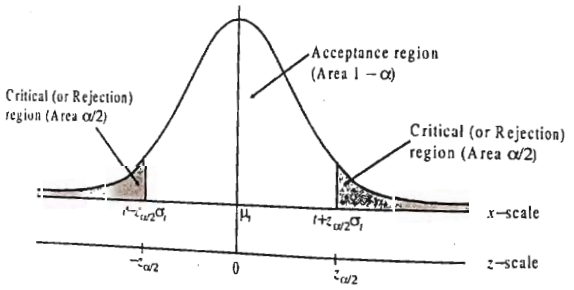


Figure 7.5

7.4 Steps for Hypothesis Testing

The steps for testing of hypothesis (or test of significance or rule of decision) are as follows.

Step 1 Formulate null hypothesis : H_0

Step 2 Formulate alternative hypothesis : H_1

Step 3 Choose level significance : α

Step 4 Determine degrees of freedom if required (This step is required for t-test, χ^2- test and F-test).

Step 5 Determine critical region : It is determined by the critical value t^*_α and by the kind of alternative hypothesis (based on which the test is right one-tailed test or left one-tailed test or two-tailed test).

Step 6 Compute the test statistic (t^*) : It is calculated using the sample data.

Step 7 Decision (or conclusion) : Accept or reject null hypothesis depending on the relation between t^* and t^*_α. The decision will depend on whether the computed value of the test criterion in **Step 5** falls in the region of acceptance or in the region of rejection.

Note The order of the above steps may change.

7.5 Sampling Distribution of Means ($n \geq 30$ or $n < 30$ but σ known) (Distribution of Sample Means)

Sampling distribution of means consists of the mean $\bar{x}$ of every possible sample of same size n drawn from a population having mean μ and standard deviation σ

Theorem 7.1 (Central Limit Theorem) If random samples of size n are drawn from an underlying non-normal population (which is not normally distributed) having mean μ and standard deviation σ, then the sampling distribution of the mean $\bar{x}$ is approximately normally distributed with

$$\text{mean } \mu_{\bar{x}} = \mu$$

and

$$\text{standard deviation } \sigma_{\bar{x}} = \sigma/\sqrt{n}, \quad \text{...(7.3)}$$

provided that the sample size n is large enough (usually $n \geq 30$). Hence, the variable

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \text{...(7.4)}$$

has a standard normal distribution.

If $n < 30$, the approximation is good only if the population is not drastically different from a normal distribution.

If the original population is normally distributed, then the sampling distribution of the mean $\bar{x}$ is also normally distributed regardless of the sample size.

If samples of size n are drawn from a finite population of size N,

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \quad \text{...(7.5)}$$

where $\sqrt{(N-n)/(N-1)}$ is known as ***finite population correction factor***.

Note When σ is unknown, for larger n ($n \geq 30$), σ can be replaced by the sample standard deviation s which is calculated using the sample mean $\bar{x}$. Thus, in (7.3),

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \quad \ldots(7.6)$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \ldots(7.6a)$$

$x_1, x_2, \ldots, x_n$ is random sample of size n.
For small sample size n ($n < 30$), the unknown σ can be substituted by s, provided sample is drawn from a normal population.

7.5.1 Test of Hypothesis Concerning Single (Specified) Population Mean μ with Known Variance σ^2 - Large Sample Test (z -Test) Let a random sample of size n is drawn from a population having mean μ and variance σ^2. Let $\bar{x}$ be the mean of the sample. Then for large samples ($n \geq 30$), it follows from central limit Theorem 7.1 that the sampling distribution of $\bar{x}$ is approximately normally distributed with mean $\mu_{\bar{x}} = \mu$ and $\sigma^2_{\bar{x}} = \sigma^2/n$. Here, the test static for single mean with known variance is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}. \quad \ldots(7.7)$$

To test whether the population mean μ has a specified value μ_1 or not, formulate the test hypothesis as follows.

Step 1 Null hypothesis $H_0 : \mu = \mu_1$ (There is no significant difference between population mean and sample mean)

Step 2 Alternative hypothesis $H_1 : \mu \neq \mu_1$ or $\mu > \mu_1$ *or* $\mu < \mu_1$ (There is significant difference between population mean and sample mean)

Step 3 Level of significance : α

Step 4 Determine critical region : Since H_1 is not equal type, therefore, two-tailed test is considered. For a given α, determine critical values of $-z_{\alpha/2}$ and $z_{\alpha/2}$ from the normal table and then decide about critical region (refer Figure 7.5).

Step 5 Compute the test statistic z by

$$z = \frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}} \quad \ldots(7.7a)$$

Step 6 Decision (or Conclusion) :

Reject H_0, if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

Accept H_0, if $-z_{\alpha/2} < z < z_{\alpha/2}$

that is, reject H_0 if z falls in the critical region.

Note 1 If $H_1 : \mu > \mu_1$, then follow right one-tailed test. The critical region is given by $z > z_\alpha$ (refer Figure 7.2). In this case,

Reject H_0, if $z > z_\alpha$

Accept H_0, if $z < z_\alpha$

Note 2 If $H_1 : \mu < \mu_1$, then follow left one-tailed test. The critical region is given by $z < -z_\alpha$ (refer Figure 7.3). In this case,

Reject H_0, if $z < -z_\alpha$

Accept H_0, if $z > -z_\alpha$

Note 3 Refer following Table 7.3 for critical values.

The following Table 7.3 gives critical values for both one-tailed and two-tailed tests at various level of siginificance.

Level of significance α (α%)	0.15 (15%)	0.1 (10%)	0.05 (5%)	0.04 (4%)	0.01 (1%)	0.005 (0.5%)	0.002 (0.2%)
Critical values for left one-tailed test ($-z_{\alpha/2}$)	–1.04	–1.28	–1.645	–2.6	–2.33	–2.58	–2.88
Critical values for right one-tailed test ($z_{\alpha/2}$)	1.04	1.28	1.645	2.6	2.33	2.58	2.88
Critical values for two-tailed test ($-z_{\alpha/2}$ and $z_{\alpha/2}$)	–1.44 and 1.44	–1.645 and 1.645	–1.96 and 1.96	–2.06 and 2.06	–2.58 and 2.58	–2.81 and 2.81	–3.08 and 3.08

Table 7.3

Note 4 For large sample ($n \geq 30$), if σ is unknown then it is replaced by s (standard deviation of sample).

Example 7.2

Let X be the length of a life of certain computer is approximately normally distributed with mean 800 days and standard deviation 40 days. If a random sample of 30 computers has an average life 788 days, test the null hypothesis that $\mu = 800$ days against the alternative hypothesis that $\mu \neq 800$ days at (*a*) 0.5% (*b*) 15% level of significance.

Solution

(*a*)

Step 1 Null hypothesis $H_0 : \mu = 800$ days

Step 2 Alternative hypothesis $H_1 : \mu \neq 800$ days

Step 3 Level of significance : $\alpha = 0.5\% = 0.005$

Step 4 Critical region : Since H_1 is not equal type, the test is two-tailed and the critical region is

$$z < -2.81 \text{ and } z > 2.81. \qquad \text{(refer Table 7.3)}$$

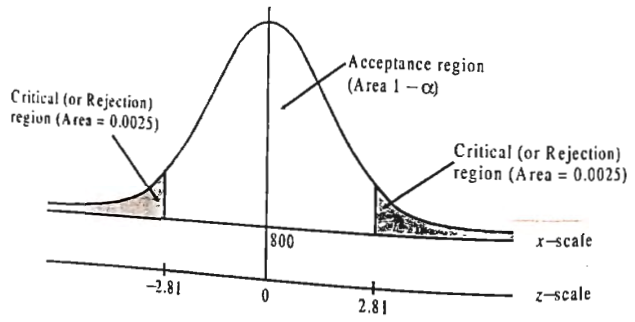


Figure 7.6

Step 5 Calculation of statistic :
Given

$\bar{x}$ = mean of the sample = 788
n = sample size = 30
standard deviation $\sigma = 40$

Therefore, using (7.7a),

$$z = \frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}} = \frac{788 - 800}{40/\sqrt{30}} = -1.643.$$

Step 6 Decision : Accept the null hypothesis H_0 since

$$-2.81 < z = -1.643 < 2.81.$$

(*b*) **Step 1** and **Step 2** remains the same.
Step 3 Level of significance : $\alpha = 15\% = 0.15$
Step 4 Critical region :

$$-1.44 < z < 1.44.$$

Step 5 As above in (*a*),

$$z = -1.643.$$

Step 6 Decision : Reject the null hypothesis H_0 since

$$z = -1.643 < -1.44.$$

Answer

Example 7.3

A college claims that its average class size is 35 students. A random sample of 64 classes has a mean size of 37 students with a standard deviation of 6 students. Test at the $\alpha = 0.05$ level of significance if the claimed value is too low.

Solution

Step 1 Null hypothesis $H_0 : \mu = 35$ students

Step 2 Alternative hypothesis H_1 : Since we suspect that the claim is too low and that the true mean is actually greater than 35, therefore,

$$H_1 : \mu > 35 \text{ students.}$$

Step 3 Level of significance : $\alpha = 0.05$

Step 4 Critical region : Since H_1 is of greater than type, therefore, right one-tailed test is applicable and critical region is

$$z > 1.645. \text{ (refer Table 7.3)}$$

Step 5 Calculation of statistic :
Given

$$\bar{x} = 37, n = 64, s = 6.$$

Since $n = 64 > 30$, the distribution of sample means is approximately normal.
Therefore, using (7.7a),

$$z = \frac{\bar{x} - \mu_1}{s/\sqrt{n}} = \frac{37 - 35}{6/\sqrt{64}} = 2.67.$$

Step 6 Decision : Reject the null hypothesis H_0 since

$$z = 2.67 > 1.645.$$

Thus, we conclude that the true mean class size is likely more than 35. *Answer*

Example 7.4

Sugar is packed in bags by an automatic machine with mean contents of bags as 1.000 *kg*. A random sample of 36 bags is selected and mean mass has been found to be 1.003 *kg*. If a S.D. of 0.01 *kg* is acceptable on all the bags being packed, determine on the basis of sample test whether the machine requires adjustment. (Take level of significance 5%)

Solution

Let the Null hypothesis H_0 be that the machine does not require any adjustment.

Step 1 Null hypothesis $H_0 : \mu = 1.000$ *kg*

Step 2 Alternative hypothesis $H_0 : \mu \neq 1.000$ *kg*

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Critical region :

$$z < -1.96 \text{ and } z > 1.96. \qquad \text{(refer Table 7.3)}$$

Step 5 Calculation of statistic

Given

$$\bar{x} = 1.003, n = 36, \sigma = 0.01.$$

Therefore, using (7.7a),

$$z = \frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}} = \frac{1.003 - 1.000}{0.01/\sqrt{36}} = 1.8.$$

Step 6 Decision : Accept the null hypothesis H_0 since

$$-1.96 < z = 1.8 < 1.96.$$

Thus, we conclude that machine does not require any adjustment. *Answer*

Exercises 7.1

01. A machine runs on an average of 125 hours/year. A random sample of 49 machines has an annual average use of 126.9 hours with standard deviation 8.4 hours. Does this suggest to believe that machines are used on the average more than 125 hours annually at 0.05 level of significance.
02. A company has a computer system that can process 1200 bills per hour. A new system is tested which processes an average of 1260 bills per hour with a standard deviation of 215 bills in a sample of 40 hours. Test if the new system is significantly better than the old one at the 5% level of significance.

03. A manufacturer of tyres guarantees that the average lifetime of its tyres is more than 28000 miles. If 40 tyres of this company tested, yield a mean lifetime of 27463 miles with standard deviation of 1348 miles. Can the guarantee be accepted at 0.01 level of significance ?
04. Record for last several years of applicants for admission into Engineering Colleges for a test showed that their mean score was 115. An administrator is interested in knowing whether the caliber of the recent applicants has changed. For the purpose of testing this hypothesis the score of the last 100 students is obtained from the admission office. The mean of this turned out to be 118 and standard deviation 28, which may be assumed for the population as a whole. Use 5% significance level and draw your conclusion.

7.6 Sampling Distribution of proportions ($n \geq 30$)

For an infinite population, let p be the Probability of occurrence of an event (called its success) and $Q = 1 - P$ be the probability of non-occurrence (failure). Consider all possible samples of size n drawn from the population and for each possible samples determine the proportion p of successes. Then we obtain a sampling distribution of proportions whose

$$\text{mean } \mu_p = P \text{ and standard deviation } \sigma_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{P(1-P)}{n}}. \qquad \text{...(7.8)}$$

While population is binomially distributed, the sampling distribution of proportions is very closely normally distributed for large n ($n \geq 30$).

For finite population of size N in which sampling is without replacement, we have

$$\text{mean } \mu_p = P \text{ and standard deviation } \sigma_p = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}. \qquad \text{...(7.9)}$$

Since, for large n, binomial distribution can be approximated to normal, the statistic z is given by

$$z = \frac{p - \mu_p}{\sigma_p} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \qquad \text{...(7.10)}$$

will be a standard normal variate; that is, $z - N(0,1)$.

→ ***For example***, suppose we need to estimate the proportion P of people in the infinite population who have a specific characteristic, say smoking. If x out of n sampled drawn people have this characteristic, then the sample proportion $p = x/n$ can be taken as an estimate of the population proportion P. In this case, population is binomially distributed, therefore,

$$\mu_p = E(p) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}(nP) \qquad \text{(Using (6.3))}$$

$$= P$$

and

$$\sigma_p^2 = V(p) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2}(nPQ) \qquad \text{(Using (6.4))}$$

$$= \frac{PQ}{n}$$

$$\Rightarrow \sigma_p = \sqrt{\frac{PQ}{n}}.$$

7.6.1 Test of Hypothesis Concerning Single (Specified) Proportion - Large Sample Test (z -Test) For a large sample size n ($n \geq 30$), the bionomial distribution can be approximated by normal distribution with the parameters

$$\text{Mean } \mu = np \text{ and variance } \sigma^2 = npq,$$

where p is the proportion of success and $q = 1 - p$ is the proportion of failure.

The test statistic for testing $p = p_1$ is given by

$$z = \frac{x - np_1}{\sqrt{np_1q_1}}, \qquad \ldots(7.11)$$

where x is the number of successes in a sample of size n and $q_1 = 1 - p_1$. This statistic can also be written as

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}, \qquad \ldots(7.12)$$

where $p = x/n$ is the proportion of success in the sample and P is the actual population porportion of success.

Note The steps for testing the hypothesis remains same as that of Section 7.4 with

$H_0 : P = P_1$ (There is no significant difference between population proportion and sample proportion)

$H_1 : P \neq P_1$ or $P > P_1$ or $P < P_1$ (There is significant difference between population proportion and sample proportion)

Example 7.5

In a sample of 400 parts manufactured by a factory, the number of defective parts found to be 30. The company, however, claims that only 5% of their product is defective. Is the claim tenable? (Take level of significance 5%)

Solution

Let the parts manufatured by a factory being defective be a success.

Step 1 Null hypothesis $H_0 : P = 0.05$ (The claim of the manufacturer is tenable)

Step 2 Alternative hypothesis $H_1 : P > 0.05$ (The claim of the manufacturer is not tenable)

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Critical region : Right one-tailed test and the critical region is

$$z > 1.645. \qquad \text{(refer Table 7.3)}$$

Step 5 Calculation of statistic :

Using (7.12),

$$z = \frac{p - P}{\sqrt{\frac{pq}{n}}}.$$

$$= \frac{\frac{30}{400} - 0.05}{\sqrt{\frac{\frac{30}{400}\left(1 - \frac{30}{400}\right)}{400}}}$$

$$= \frac{0.075 - 0.05}{\sqrt{\frac{(0.075)(0.925)}{400}}}$$

$$= \frac{0.025}{0.013}$$

$$\approx 1.92.$$

Step 6 Decision : Reject the null hypothesis H_0 since

$$z = 1.92 > 1.645.$$

Thus, we conclude that the claims of the manufacturer is not tanable. ***Answer***

Example 7.6

A political party claims that 45% of the voters in an election district prefer its candidate. A sample of 200 voters include 80 who prefer this candidate. Test if the claim is

valid at the 5% significance level.

Solution

Step 1 Null hypothesis $H_0 : P = 0.45$ (The claim of the party is tenable)

Step 2 Alternative hypothesis $H_1 : P \neq 0.45$ (The claim of the party is not tenable)

Step 3 Level of significance : $\alpha = 5\% = 0.05$.

Step 4 Critical region : Two-tailed test and the critical region is

$z < -1.96$ and $z > 1.96$. (refer Table 7.3)

Step 5 Calculation of statistic :

Using (7.12),

$$z = \frac{p - P}{\sqrt{\frac{pq}{n}}}.$$

$$= \frac{\frac{80}{200} - 0.45}{\sqrt{\frac{\frac{80}{200}\left(1 - \frac{80}{200}\right)}{200}}}$$

$$= \frac{0.4 - 0.45}{0.035}$$

$$= -1.43$$

Step 6 Decision : Accept the null hypothesis H_0 since

$$z = -1.43 > -1.96.$$

Thus, the party's claim might be valid. ***Answer***

Exercises 7.2

01. During testing in a sample of 300 chips, 10 have been found to be defective. Can the manufacturers claim that 2% of the chips are defective may be accepted? (Take level of significance 5%)
02. An electric utility survey indicates that 18% of all households in a community own personal computers. A separate study of 80 families with school age children in this community finds that 22 of them own computers. Test whether the proportion of families with school age children owning computers is higher than in the general population in that area. Use $\alpha = 0.02$.

7.7 Sampling Distribution of Differences and Sums of Two Same Statistic

Let μ_{t_1} and σ_{t_1} be the mean and standard deviation of a sampling distribution of statistic t_1 (which may be mean, proportion, etc. as per Table 7.1) obtained by computing t_1 for all possible samples of size n_1 drawn from population N_1. Similarly, let μ_{t_2} and σ_{t_2} be the mean and standard deviation of a sampling distribution of statistic t_2 obtained by computing t_2 for all possible samples of size n_2 drawn from another different population N_2. Now, compute the statistic $t_1 - t_2$ which is the difference of the statistic from all possible combinations of these samples from the two populations N_1 and N_2. Then the mean and standard deviation of the sampling distribution of differences are given as follows.

$$\text{mean } \mu_{t_1 - t_2} = \mu_{t_1} - \mu_{t_2} \text{ and standard deviation } \sigma_{t_1 - t_2} = \sqrt{\sigma_{t_1}^2 + \sigma_{t_2}^2} \quad \ldots(7.13)$$

assuming that the samples are independent.

For larger values of n_1 and n_2, $t_1 - t_2$ can be approximated to a normal variate.

The statistic z is given by

$$z = \frac{(t_1 - t_2) - (\mu_{t_1} - \mu_{t_2})}{\sigma_{t_1 - t_2}}, \quad \ldots(7.14)$$

which is a standard normal variate; that is, $z \sim N(0, 1)$.

Similarly, for statistic $t_1 + t_2$,

$$\mu_{t_1 + t_2} = \mu_{t_1} + \mu_{t_2} \text{ and } \sigma_{t_1 + t_2} = \sqrt{\sigma_{t_1}^2 + \sigma_{t_2}^2}.$$

7.8 Sampling Distribution of Differences of Two Means ($n_1 + n_2 \geq 30$ or $n_1 + n_2 < 30$ but σ_1 and σ_2 are Known)

If the statistics t_1 and t_2 are means $\bar{x}_1$ and $\bar{x}_2$, respectively, in Section 7.7. Then using (7.13) we have following.

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 \quad \text{(Using (7.3))} \ldots(7.15)$$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(Using (7.3)) ...(7.16)

where, μ_1 and μ_2 are means of the two populations N_1 and N_2, respectively. σ_1 and σ_2 are standard deviations of two populations N_1 and N_2, respectively.
The statistic z is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{\bar{x}_1} - \mu_{\bar{x}_2})}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

$$\Rightarrow z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

(Using (7.15) and (7.16))...(7.17)

which is a standard normal variate; that is, $z \sim N(0,1)$.

7.8.1 Test of Hypothesis Concerning Two Population Means μ_1 and μ_2 with Known Variances σ_1^2 and σ_2^2 – Large Sample Test (z -Test) Suppose we want to compare two population means μ_1 and μ_2.

→ **For example**, comparison of the lead levels in drinking water in two different sections of a city.

Let $\bar{x}_1$ be the mean of a random sample of size n_1 drawn from a population having mean μ_1 and variance σ_1^2. Let $\bar{x}_2$ be the mean of an independent random sample of size n_2 drawn from another population having mean μ_2 and variance σ_2^2. For larger values of n_1 and n_2, $\bar{x}_1 - \bar{x}_2$ can be approximated to a normal variate with statistic z is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad ...(7.18)$$

which is a standard normal variate; that is, $N \sim (0, 1)$.

Note 1 When the two variances σ_1^2 and σ_2^2 are unknown, they can be replaced by

sample variances s_1^2 and s_2^2 provided both the samples are large $(n_1, n_2 \geq 30)$. In this case (7.18) becomes

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad ...(7.19)$$

where s_1 and s_2 are standard deviation of samples.

Note 2 If the samples have been drawn from the two populations with common variance σ^2, the (7.18) becomes

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad ...(7.20)$$

When σ is unknown, then

$$\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}, \quad ...(7.21)$$

is taken as an estimate of σ.

Note 3 The steps for testing the hypothesis remains same as that of Section 7.4 only the difference is that, here, we have to check the following with the present z statistic. Under the null and alternative hypothesis as follows.

$H_0 : \mu_1 = \mu_2 = \delta = 0$; that is, there is no significant difference between two population means; that is, difference of two population means is zero.

$H_1 : \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (There is significant difference betwen two population means).

The $H_1 : \mu_1 > \mu_2$ or $\mu_1 < \mu_2$ are used to determine whether one product (or population) is better than (or superior to) the other product.

The statistic (7.18) becomes

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \quad ...(7.22)$$

and hence can be tested accordingly.

OR

(2) When difference of means are given for two populations, then in this case

$H_0 : \mu_1 - \mu_2 = \delta$, difference of two population means is nonzero.

$H_1 : \mu_1 - \mu_2 \neq \delta$ or $\mu_1 - \mu_2 > \delta$ or $\mu_1 - \mu_2 < \delta$.

The statistic z given by (7.18) can be tested.

Example 7.7

In a random sample of 100 light bulbs manufactured by company A, the mean lifetime of light bulb is 1190 hours with standard deviation of 90 hours. Also, in a random sample of 75 light bulbs manufactured by company B, the mean lifetime of light bulb is 1230 hours with standard deviation of 120 hours. Is there a difference between the mean lifetimes of the two brands of lightbulbs at a significance level of (*a*) 0.05 (*b*) 0.01 ?

Solution

Let x_A be the lifetime (in hours) of light bulbs manufactured by company A, and x_B be the lifetime of light bulbs manufactured by company B.

Given that

$$\bar{x}_A = 1190, \ \bar{x}_B = 1230.$$

Let s_A be the standard deviation for light bulbs manufactured by company A, and s_B be the standard deviation for light bulbs manufactured by company B.

Given that

$$s_A = 90, \ s_B = 120.$$

Let n_A be the random sample size of light bulbs manufactured by company A, and n_B be the random sample size of light bulbs manufactured by company B.

Given that

$$n_A = 100, \ n_B = 75.$$

(*a*)

Step 1 Null hypothesis $H_0 : \mu_A = \mu_B$

Step 2 Alternative hypothesis $H_1 : \mu_A \neq \mu_B$

Step 3 Level of significance : $\alpha = 0.05$

Step 4 Critical region : Two-tailed test and the critical region is

$$z < -1.96 \text{ and } z > 1.96. \qquad \text{(refer Table 7.3)}$$

Step 5 Calculation of statistic :

$$\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{(90)^2}{100} + \frac{(120)^2}{75}}$$

$$= \sqrt{\frac{8100}{100} + \frac{14400}{75}}$$

$$= \sqrt{81 + 192}$$

$$= \sqrt{273}$$

$$= 16.5227. \qquad \text{...(i)}$$

Therefore, using (7.19),

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$$= \frac{1190 - 1230}{16.5227} \qquad \text{(Using (i))}$$

$$= -\frac{40}{16.5227}$$

$$\approx -2.421.$$

Step 6 Decision : Reject the null hypothesis H_0 since

$$z = -2.421 < -1.96.$$

Thus, we conclude that there is difference between the mean lifetimes of the light bulbs manufactured by companies A and B.

Answer (a)

(*b*) **Step 1** and **Step 2** remains the same.

Step 3 Level of significance : $\alpha = 0.01$

Step 4 Critical region : Two-tailed test and the critical region is

$$z < -2.58 \text{ and } z > 2.58. \qquad \text{(refer Table 7.3)}$$

Step 5 As above in (*a*),

$$z = -2.421.$$

Step 6 Decision : Accept the null hypothesis H_0 since

$$-2.58 < z = -2.421 < 2.58.$$

Thus, we conclude that there is no difference between the mean lifetimes of the light bulbs manufactured by companies A and B. ***Answer (b)***

Example 7.8

A company A manufactured tube lights and claims that its tube lights are superior than its main competitor company B. The study showed that a sample of 40 tube lights manufactured by company A has a mean lifetime of 647 hours of continuous use with a standard deviation of 27 hours, while a sample of 40 tube lights manufactured by company B had a mean lifetime of 638 hours of continuous use with a standard deviation of 31 hours. Does this substantiate the claim of company A that their tube lights are superior than manufactured by company B at (*a*) 0.05 (*b*) 0.01 level of significance.

Solution

Let x_A be the lifetime (in hours) of tube lights manufactured by company A, and x_B be the lifetime of tube lights manufactured by company B.

Given that

$$\bar{x}_A = 647,\ \bar{x}_B = 638.$$

Let s_A be the standard deviation for tube lights manufactured by company A, and s_B be the standard deviation for tube lights manufactured by company B.

$$s_A = 27,\ s_B = 31.$$

Let n_A be the random sample size of tube lights manufactured by company A, and n_B be the random sample size of tube lights manufactured by company B.

Given that

$$n_A = 40,\ n_B = 40.$$

(*a*)

Step 1 Null hypothesis $H_0 : \mu_A = \mu_B$

Step 2 Alternative hypothesis $H_1 : \mu_A > \mu_B$

Step 3 Level of significance : $\alpha = 0.05$

Step 4 Critical region : Right one-tailed test and the critical region is

$$z > 1.645. \qquad \text{(refer Table 7.3)}$$

Step 5 Calculation of statistic :

$$\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}$$

$$= \sqrt{\frac{729}{40} + \frac{961}{40}}$$

$$= \sqrt{\frac{1690}{40}}$$

$$= 6.5. \qquad \text{...(i)}$$

Therefore, using (7.19),

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$$= \frac{647 - 638}{6.5} \qquad \text{(Using (i))}$$

$$= \frac{9}{6.5}$$

$$= 1.385.$$

Step 6 Decision : Accept the null hypothesis H_0 since

$$z = 1.385 < 1.645.$$

Thus, we conclude that there is no difference between tube lights manufactured by two companies A and B. ***Answer (a)***

(*b*) **Step 1** and **Step 2** remains the same.

Step 3 Level of significance : $\alpha = 0.01$

Step 4 Critical region : Right one-tailed test and the critical region is

$$z > 2.33. \qquad \text{(refer Table 7.3)}$$

Step 5 As above in (*a*),

$$z = 1.385.$$

Step 6 Decision : Accept the null hypothesis H_0 since

$$z = 1.385 < 2.33.$$

Thus, again we have same conclusion as in part (*a*). ***Answer (b)***

Example 7.9

The mean of two large samples of sizes 1000 and 2000 are 67.5 and 68.0, respectively. Test the equality of means of the two populations each with standard deviation 2.5. (Consider level of significance 5%)

Solution

Given that

	Population 1	Population 2
Mean	$\bar{x}_1 = 67.5$	$\bar{x}_2 = 68.0$
S.D.	$s_1 = 2.5$	$s_2 = 2.5$
Sample size	$n_1 = 1000$	$n_2 = 2000$

Step 1 Null hypothesis $H_0 : \mu_1 = \mu_2$ (There is no significant difference between two population means)

Step 2 Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (There is significant difference between two population means)

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Critical region : Two-tailed test and the critical region is

$z < -1.96$ and $z > 1.96$. (refer Table 7.2)

Step 5 Calculation of statistic :

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(2.5)^2}{1000} + \frac{(2.1)^2}{2000}}$$

$$= \sqrt{\frac{6.25}{1000} + \frac{6.25}{2000}}$$

$$= \sqrt{\frac{12.5 + 6.25}{2000}}$$

$$= \sqrt{\frac{18.75}{2000}}$$

$$= 0.097. \quad \text{...(i)}$$

Therefore, using (7.19),

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{67.5 - 68.0}{0.097} \quad \text{(Using (i))}$$

$$= -5.15.$$

Step 6 Decision : Reject the null hypothesis H_0 since

$z = -5.15 < -1.96$.

Thus, we conclude that there is significant difference between two population means.

Answer

Example 7.10

A company claims that alloying reduces resistance of electric wire by more than 0.050 *ohm*. To test this claim samples of 32 standard wire and alloyed wire are tested yielding the following results.

Type of wire	Mean resistance (*ohms*)	S.D.(s) (*ohms*)
Standard	0.136	0.004
Alloyed	0.083	0.005

At the 0.05 level of significance, does this support the claim?

Solution

Step 1 Null hypothesis $H_0 : \mu_1 - \mu_2 = 0.050$

Step 2 Alternative hypothesis $H_1 : \mu_1 - \mu_2 > 0.050$

Step 3 Level of significance : $\alpha = 0.05$

Step 4 Critical region : Right one-tailed test and the critical region is

$z > 1.645$. (refer Table 7.3)

Step 5 Calculation of statistic :

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(0.004)^2}{32} + \frac{(0.005)^2}{32}}$$

$$= \sqrt{\frac{0.000016 + 0.000025}{32}}$$

$$= \sqrt{\frac{0.000041}{32}}$$

$$= 0.0011. \quad \text{...(i)}$$

Therefore, using (7.19),

$$z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

$$=\frac{0.136-0.083-0.050}{0.0011} \quad \text{(Using (i))}$$

$$=2.73.$$

Step 6 Decision : Reject the null hypothesis : H_0 since

$$z=2.73>1.645.$$

Thus, we conclude that data substantiate the claim. *Answer*

Exercises 7.3

01. Test the null hypothesis $\mu_A=\mu_B$ against the alternative hypothesis $\mu_A\neq\mu_B$ at 0.01 level of significance for the following data.

	Sample size (kgs)	Mts.	S.D. (kgs)
Type A	40	247.3	15.2
Type B	30	254.1	18.7

02. A study is conducted to compare the batting averages of players in 1990 and 2015. With the following data test whether there is a difference in batting averages between the players in 1990 and 2015 at the 0.05 level of significance.

Players from	Sample size	Mean batting average	S.D.
1990	35	267	27
2015	40	255	30

03. For the following random sample data of men and women about their daily earnings, test at 0.05 level of significance whether the average income for men and women is same or not.

	Sample size	Mean earnings	S.D.(s) ($\bar{x}$)
Men	42	744.85	397.7
Women	32	516.78	162.523

7.9 Sampling Distribution of Differences of Two Proportions ($n_1+n_2\geq 30$)

If the statistics t_1 and t_2 are proportions p_1 and p_2, respectively in section 7.7. Then using (7.13), we have following.

$$\mu_{p_1-p_2}=\mu_{p_1}-\mu_{p_2}=P_1-P_2 \quad \text{(Using (7.8))...(7.23)}$$

$$\sigma_{p_1-p_2}=\sqrt{\sigma_{p_1}^2+\sigma_{p_2}^2}=\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}} \quad \text{(Using (7.8))...(7.24)}$$

The statistic z is given by

$$z=\frac{(p_1-p_2)-(\mu_{p_1}-\mu_{p_2})}{\sigma_{p_1-p_2}}$$

$$\Rightarrow z=\frac{(p_1-p_2)-(P_1-P_2)}{\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}};\ Q_1=1-P_1, Q_2=1-P_2,$$

(Using (7.23) and (7.24)) ...(7.25)

which is a standard normal variate; that is, $z\sim N(0,1)$.

7.9.1 Test of Hypothesis Concerning Two Proportions - Large Sample Test (z-Test)

Suppose we want to compare two distinct populations N_1 and N_2 with respect to the prevalence of a specific attribute; that is, each item of these two populations belongs to two mutually exclusive classes depending on whether the item has (possess) an attribute c (success) or not (failure).

→ ***For example***, comparison of the prevalence of lung cancer among smokers (Population N_1) and nonsmokers (Population N_2)

Let x_1 and x_2 be the number of items having attribute c in random samples of sizes n_1 and n_2 (independent of n_1) drawn from the two pupulations N_1 and N_2, respectively. Then sample proportions are

$$p_1=\frac{x_1}{n_1} \text{ and } p_2=\frac{x_2}{n_2}.$$

Let P_1 and P_2 be the population proportions of populations N_1 and N_2, respectively. For larger values of n_1 and n_2, p_1-p_2 can be approximated to a normal variate with statistic z is given by

$$z=\frac{(p_1-p_2)-(P_1-P_2)}{\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}};\ Q_1=1-P_1,\ Q_2=1-P_2, \quad ...(7.26)$$

which is a standard normal variate; that is, $N\sim(0,1)$.

Under the null and alternative hypothesis,

$H_0 : P_1 = P_2$: that is, there is no significant difference between two population proportions

$H_1 : P_1 \neq P_2$ or $P_1 > P_2$ or $P_1 < P_2$ (There is significant difference between two population proportions),

the z-statistic (7.26) becomes

$$z = \frac{p_1 - p_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1) \qquad \ldots(7.27)$$

and hence can be tested accordingly.

Note 1 When $P_1 = P_2 = P$ (say), then (7.27) becomes

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}; Q = 1 - P \qquad \ldots(7.28)$$

Note 2 When P is unknown, then (7.26), (7.27) and (7.28) becomes

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}; \; q_1 = 1 - p_1, \; q_2 = 1 - p_2 \qquad \ldots(7.29)$$

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}, \qquad \ldots(7.30)$$

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad \ldots(7.31)$$

where

$$p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}, \qquad \ldots(7.32)$$

which is an unbiased estimate of P.

Note 3 The steps for testing the hypothesis remains same as that of Section 7.4 only the difference is that, here, we have to check the following with the present z-statistic.

(1) $H_0 : P_1 = P_2 = \delta = 0$, difference of two population proportions is zero

$H_1 : P_1 \neq P_2$ or $P_1 > P_2$ or $P_1 < P_2$

OR

(2) When difference of two population proportions are given, then in this case

$H_0 : P_1 - P_2 = \delta$, difference of two population proportions is nonzero

$H_1 : P_1 - P_2 \neq \delta$ or $P_1 - P_2 > \delta$ or $P_1 - P_2 < \delta$

Example 7.11

In a certain city A, 450 persons were considered regular consumer of tea out of a sample of 1000 persons. In another city B, 400 were regular consumers of tea out of sample of 800 persons. Do these facts reveal a significant difference between the two cities as far as tea drinking habit is concerned? (Use level of significance 5%)

Solution

Let a person selected being a tea drinker be a success.

Step 1 Null hypothesis $H_0 : P_1 = P_2$ (There is no significant difference between two cities as far as tea drinking habit is concerned)

Step 2 Alternative hypothesis $H_1 : P_1 \neq P_2$ (There is siginficant difference between two cities as far as tea drinking habit is concerned)

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Critical region : Two-tailed test and the critical region is

$z < -1.96$ and $z > 1.96$. (refer Table 7.3)

Step 5 Calculation of statistic :

Given that

	Population 1 (City A)	Population 2 (City B)
Sample size	$n_1 = 1000$	$n_2 = 800$
Proportion	$p_1 = \frac{450}{1000} = 0.45$	$p_2 = \frac{400}{800} = 0.5$

Using (7.32),

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 n_2}$$

$$= \frac{(1000)(0.45) + (800)(0.5)}{1000 + 800}$$

$$= \frac{450 + 400}{1800}$$

$$= \frac{850}{1800}$$

$$\approx 0.472.$$

Therefore,

$$q = 1 - p = 1 - 0.472 = 0.528.$$

Using (7.31),

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.45 - 0.5}{\sqrt{(0.472)(0.528)\left(\frac{1}{1000} + \frac{1}{800}\right)}}$$

$$= -\frac{0.05}{\sqrt{0.0005607}}$$

$$= -\frac{0.05}{0.0237}$$

$$= -2.11.$$

Step 6 Decision : Reject null hypothesis H_0 since

$$z = -2.11 < -1.96.$$

Thus, we conclude that there is siginificant difference between two cities as far as tea drinking habit is concerned.

Answer

Example 7.12

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 in A? (Use level of significant 5%)

Solution

Step 1 Null hypothesis $H_0 : P_1 = P_2$

Step 2 Alternative hypothesis $H_1 : P_1 < P_2$

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Critical region : Left one-tailed test and the critical region is

$$z < -1.645.$$ (refer Table 7.3)

Step 5 Calculation of statistic :

Given that

	Population 1 (City A)	Population 2 (City B)
Sample size	$n_1 = 400$	$n_2 = 800$
Proportion	$p_1 = \frac{100}{400} = 0.25$	$p_2 = \frac{300}{800} = 0.375$

Using (7.32),

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 n_2}$$

$$= \frac{(400)(0.25) + (800)(0.375)}{400 + 800}$$

$$= \frac{100 + 300}{1200}$$

$$= \frac{400}{1200}$$

$$= 0.33.$$

Therefore,

$$q = 1 - p = 1 - 0.33 = 0.67.$$

Using (7.31),

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.25 - 0.375}{\sqrt{(0.33)(0.67)\left(\frac{\cdot 1}{400} + \frac{1}{800}\right)}}$$

$$= -\frac{0.125}{\sqrt{0.0008291}}$$

$$= -\frac{0.125}{0.029}$$

$$\approx -4.31.$$

Step 6 Decision : Reject the null hypothesis H_0 since

$$z = -4.31 < -1.645.$$

Thus, we conclude that the proportion of smokers is greater in the second city B than in A. ***Answer***

Example 7.13

A question in a true-false quiz is considered to be smart if it discriminates between intelligent person (IP) and average person (AP). Suppose 205 of 250 IP's and 137 of 250 AP's answer a quiz question correctly. Test of 0.01 level of significance whether for the given question, the proportion of correct answers can be expected to be at least 15% higher among IP's than among the AP's.

Solution

Let P_1 and P_2 be the proportion of correct answer by IP's and AP's, respectively. Then

Step 1 Null hypothesis $H_0 : P_1 - P_2 = 0.15$

Step 2 Alternative hypothesis $H_1 : P_1 - P_2 > 0.15$

Step 3 Level of significance : $\alpha = 0.01$

Step 4 Critical region : Right one-tailed test and the critical region is

$z > 2.33$. (refer Table 7.3)

Step 5 Calculation of statistic :

Given that

	Population 1 (IP's)	Population 2 (AP's)
Sample size	$n_1 = 250$	$n_2 = 250$
Proportion	$p_1 = \frac{205}{250} = 0.82$	$p_2 = \frac{137}{250} = 0.548$

Using (7.29),

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$$= \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

$$= \frac{(0.82 - 0.548) - (0.15)}{\sqrt{\frac{0.82(1-0.82)}{250} + \frac{0.548(1-0.548)}{250}}}$$

$$= \frac{0.272 - 0.15}{\sqrt{\frac{0.1476}{250} + \frac{0.2477}{250}}}$$

$$= \frac{0.122}{\sqrt{0.0015812}}$$

$$= \frac{0.122}{0.0398}$$

$$\approx 3.065.$$

Step 6 Decision : Reject the null hypothesis H_0 since

$$z = 3.065 > 2.33.$$

Thus, we conclude that the proportion of correct answers by IP is 15% more than those by AP's. ***Answer***

Exercise 7.4

01. In a political poll, 42 out of 100 randomly selected men surveyed preferred Candidate *A*. Also, 92 out of 200 women preferred Candidate *A*. Test whether there is any difference in the proportions of men and women who prefer Candidate *A* at the $\alpha = 0.05$ level of significance.
02. Suppose that a method *A* results in 20 unacceptable transistors out of 100 produced, whereas another method *B* results in 12 unacceptable transistors out of 100 produced. Can we conclude at 5% level of significance that the two methods are equivalent?
03. In a survey of A.C. machines produced by company *A* it was found that 19 machines were defective in a random sample of 200 while for company *B*, 5 were defective out of 100. At 5% level of significance is there reason to believe that
(*a*) there is significant difference in performance of A.C. machines betwen the two companies *A* and *B*.
(*b*) products of *B* are superior to products of *A*.

7.10 Standard Error (S.E.) Revisited

The standard error of some familiar statistics as discussed in above sections can besummarized as follows.

Statistic	Standard Error	Population size
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$ (refer (7.3))	Either large or the sample is drawn with replacement
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ (refer (7.5))	N is finite or the sample is drawn without replacement
p	$\sqrt{\frac{PQ}{n}}(Q=1-P)$ (refer (7.8))	Either large or the sample is drawn with replacement
p	$\sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$ (refer (7.9))	N is finite or the sample is drawn without replacement

$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ (refer (7.16))	Either large or the sample is drawn with replacement
$p_1 - p_2$	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$ (refer (7.24))	Either large or the sample is drawn with replacement

Note 1 S.E. plays an important role in large sample theory and forms the basis in testing of hypothesis.
Note 2 The magnitude of S.E. gives an index of the precision of the estimate of the parameter. The reciprocal of S.E., in fact, measures the reliability or precision of the sample.
Note 3 S.E. enables us to find limits within which a population parameter is expected to lie.

7.11 Sampling Distribution of Means (σ unknown) : t - Distribution

So far we have discussed large sample testing (that is, problems of inference on a population mean, or equality or differene between two population means) with the assumption that the population standard deviation σ is known. All these studies were based on central limit theorem.

But when σ is unknown (1) for large sample size ($n \geq 30$), σ can be replaced by the sample standard deviation s calculated using (7.6a). (2) for small sample size ($n < 30$), σ can be replaced by the sample standard deviation s, provided sample is drawn from a normal population.

Theorem 7.2 of $\bar{x}$ is the mean of a random sample of size n drawn from a normal population having the mean μ and the variance σ^2, and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \text{(7.33)}$$

then

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \quad \text{(7.34)}$$

is a random variable having the t-distribution with $\nu = n-1$ degrees of freedom with probability density function

$$f(t) = \frac{1}{\sqrt{\nu}B\left(\frac{1}{2},\frac{\nu}{2}\right)}\frac{1}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}}, \quad -\infty < t < \infty \quad \text{(7.35)}$$

The t-statistic defined in (7.34) and its distribution was mathematically derived by W.S. Gosset in 1908.

William Sealy Gosset (1876-1937) was English statistician.

Thus, for small samples ($n < 30$) and with σ unknown, a natural statistic for inference on population μ is given by (7.34), provided samples are drawn from a normal population.

Theorem 7.2 is more general than central limit Theorem 7.1 in the sense that is does not require knowledge of σ. On the other hand it is less general than Theorem 7.1 in the sense that is requires assumption of a normal population.

➢ **Characteristics of the t-statistic**

(1) t-distribution is similar to that of a normal distribution – both are bell shaped and symmetrical about the mean.

(2) Like standard normal distribution, t-distribution has mean 0 but its variance depends on the parameter ν, called the number of degrees of freedom.

(3) The variance of t-distribution exceeds 1, but it approaches 1 as $n \to \infty$.

(4) t-distribution with ν degrees of freedom approaches the standard normal distribution as $\nu \to \infty$.

(5) For samples of size $n \geq 30$, the standard normal distribution provides a good approximation to the t-distribution.

(6) **Critical value** The critical value of t-distribution is denoted by t_α.

The area under the curve to the right of t_α is α. Because of symmetricity of t-distribution,

$$t_{1-\alpha} = -t_\alpha .$$

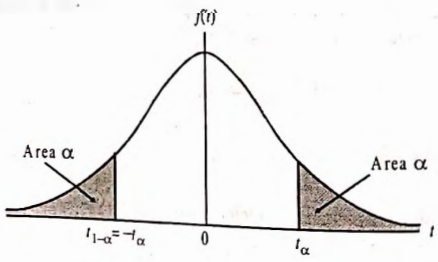


Figure 7.7

(7) Table II and III shows critical values of t-distribution.

7.11.1 Test of Hypothesis Concerning Single (Specified) Population Mean μ with Unknown Variance σ^2 – Small Sample Test (t-Test)

There are many situations where collection of large samples ($n \geq 30$) is uneconomical, impracticable and time consuming.

→ ***For example***, investigation of characteritics of large samples for fighter jet planes, submarines, satellites, super computer, etc.

In such cases, we go for small sample size ($n < 30$). For such type of situation when there is small sample size and unknown σ, the decision criterion is based on the t-distribution with $\nu = n - 1$ degrees of freedom. The test statistic in this case is given by

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \ldots(7.36)$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \ldots(7.37)$$

x_i's are random sample drawn from a normal population and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad \ldots(7.38)$$

This test statistic is known as one sample t-test.

Note 1 The test procedure in this case is same as that for large samples except that t-values are used in place of z-values, and σ is replaced by s.

Note 2 The steps for testing the hypothesis remains same as that of Section 7.4 with null and alternative hypothesis as follows.

$H_0 : \mu = \mu_1$ (There is no significant difference between population mean and sample mean)

$H_1 : \mu \neq \mu_1$ or $\mu > \mu_1$ or $\mu < \mu_1$ (There is significant difference betwen population mean and sample mean)

Example 7.14

A machine is design ed to produce insulating washers for electrical devices of average thickness of 0.025 *cm*. A random sample of 10 washers was found to have an average thickness of 0.024 *cm* with S.D. of 0.002 *cm*. Test the significance of the deviation. (Value o f t for 9 degrees of freedom at 5% level of significance is 2.262)

Solution

Step 1 Null hypot hesis $H_0 : \mu = 0.025$ (There is no significant difference between population mean and sample mean)

Step 2 Alternative hypothesis $H_1 : \mu \neq 0.025$ (There is significant difference between population mean and sample mean)

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Degrees of freedom :

$$\nu = n - 1 = 10 - 1 = 9.$$

Step 5 Critical region : Two-tailed test and the critical region is

$$t < -2.262 \text{ and } t > 2.262. \qquad \text{(Using Table II(a))}$$

Step 6 Calculation of statistic :

Using (7.36),

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{0.024 - 0.025}{\frac{0.002}{\sqrt{10}}}$$

$$= -\frac{0.001}{0.000632}$$

$$= -1.58.$$

Step 7 Decision : Accept null hypothesis H_0 since

$$-2.262 < t = -1.58 < 2.262.$$

Thus, we conclude that there is no significant difference between population mean and sample mean.

Answer

Example 7.15

Ten individuals were chosen at random from a normal population and their heights were found to be in *inches* as 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. Test the hypothesis that the mean height of the population is 66 *inches*. (Take level of significance 0.05)

Solution

Here, $n = 10$.

Step 1 Null hypothesis $H_0 : \mu = 66$

Step 2 Alternative hypothesis $H_1 : \mu \neq 66$

Step 3 Level of significance : $\alpha = 0.05$.

Step 4 Degrees of freedom :

$$\nu = n - 1 = 10 - 1 = 9.$$

Step 5 Critical region : Two-tailed test and the critical region is

$$t < -2.262 \text{ and } t > 2.262. \qquad \text{(Using Table II(a))}$$

Step 6 Calculation of statistic :

Using (7.37),

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n}$$

$$= \frac{63 + 63 + 66 + 67 + 68 + 69 + 70 + 70 + 71 + 71}{10}$$

$$= \frac{678}{10}$$

$$= 67.8.$$

Using (7.38),

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{10-1}\sum_{i=1}^{10}(x_i - \bar{x})^2 \qquad \text{(Since } n = 10\text{)}$$

$$= \frac{1}{10-1}[(-4.8)^2 + (-4.8)^2 + (-1.8)^2 + (-0.8)^2 + (0.2)^2 + (1.2)^2 + (2.2)^2 + (2.2)^2 + (3.2)^2 + (3.2)^2]$$

$$= \frac{1}{9}(23.04 + 23.04 + 3.24 + 0.64 + 0.04 + 1.44 + 4.84 + 4.84 + 10.24 + 10.24)$$

$$= \frac{1}{9}(81.6)$$

$$= 9.067.$$

Therefore,

$$s = 3.011 \text{ inches.}$$

Using (7.36),

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{67.8 - 66}{\frac{3.011}{\sqrt{9}}}$$

$$= \frac{1.8}{1.004}$$

$$= 1.79.$$

Step 7 Decision : Accept null hypothesis H_0 since

$$-2.262 < t = 1.79 < 2.262.$$

Thus, we conclude that there is no significant difference between population mean and sample mean. *Answer*

Example 7.16

Manager of a diet plan advertise that the mean weight loss for people on their plan is at least 45 *pounds* in 6 months. A sample of 28 people on this plan lose an average of 35 *pounds* with a standard deviation of 20 *pounds*. Test at $\alpha = 0.01$ level of significance, if the claim is to high.

Solution

Step 1 Null hypothesis $H_0 : \mu = 45$

Step 2 Alternative hypothesis $H_1 = \mu < 45$

Step 3 Level of significance : $\alpha = 0.01$

Step 4 Degrees of freedom :

$$\nu = n - 1 = 28 - 1 = 27.$$

Step 5 Critical region : Left one-tailed test and the critical region is

$$t < -2.473. \qquad \text{(Using Table II(b))}$$

Step 6 Calculation of statistic : Using (7.36),

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{35 - 45}{\frac{20}{\sqrt{28}}}$$

$$= -\frac{10}{3.78}$$

$$= -2.646.$$

Step 7 Decision : Reject the null hypothesis H_0 since

$$t = -2.646 < -2.473.$$

Thus, we conclude that the claim for an average 45 *pound* weight loss is over stated at 0.01 level of significance. *Answer*

Exercises 7.5

01. The mean weekly sa'es of A.Cs. of a particular brand in company's showrooms was 14.4 A.C. per showroom. After announcing a few incentives the mean weekly sales in 22 stores for a typical week increased to 15.4 with S.D. of 1.7. Were the incentives announced effective in boosting the sale ?

02. An ambulance service company claims that on an average it takes 20 *minutes* between a call for an ambulance and the patient's arrival at the hospital. If in 6 calls the time taken (between a call and arrival at hospital) are 27, 18, 26, 15, 20, 32. Can the company's claim be accepted?

7.12 Test of Hypothesis Concerning Two Population means μ_1 and μ_2 with Unknown Varinaces σ_1^2 and σ_2^2 - Small Sample Test (t - Test)

Let $\bar{x}_1$ and s_1 be the mean and standard deviation of a random sample of size n_1 ($n_1 < 30$) drawn from a normal population having mean μ_1 and variance σ_1^2 (σ_1^2 is unknown). Let $\bar{x}_2$ and s_2 be the mean and standard deviation of a random sample of size n_2 ($n_2 < 30$) drawn from a normal population having mean μ_2 and variance σ_2^2 (σ_2^2 is unknown).

When we are sampling n_1 from the population, there are $n_1 - 1$ degrees of freedom. Similarly, when we are sampling n_2 from the population, there are an additional $n_2 - 1$ degrees of freedom. Therefore, there are total of

$$(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

degrees of freedom for the two samples.

When both the populations have equal variances (that is, $\sigma_1^2 = \sigma_2^2$), then the pooling variance σ^2 is given by

$$\sigma^2 = \frac{\sum_{i=1}^{n_1} (x_{1_i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{2_i} - \bar{x}_2)^2}{n_1 + n_2 - 2} \qquad \ldots(7.39)$$

$$\Rightarrow \sigma^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \quad \text{(Using 7.6a)} \quad ...(7.40)$$

where

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2} \quad ...(7.41)$$

Under the null and alternative hypothesis $H_0 : \mu_1 = \mu_2$; that is, there is no significant difference between two population means $H_1 : \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (There is significant difference between two population means) (also known as two sampled pooled t- test)

The test statistic t is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad ...(7.42)$$

Note The steps for testing the hypothesis remains same as that of Section 7.4 only the difference is that, here, we have to check the following with the present t-statistic.

(1) $H_0 : \mu_1 = \mu_2 = \delta = 0$, difference of two population means is zero

$H_1 : \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

OR

(2) When difference of means are given for two populations, then in this case

$H_0 : \mu_1 - \mu_2 = \delta = 0$, difference of two population means is nonzero

$H_1 : \mu_1 - \mu_2 \neq \delta$ or $\mu_1 - \mu_2 > \delta$ or $\mu_1 - \mu_2 < \delta$

In this case t-test statistic is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad ...(7.43)$$

Example 7.17

Two samples of 6 and 5 items, respectively, gave the following data.

Mean of the first sample = 40

S.D of the first sample = 8

Mean of the second sample = 50

S.D of the second sample = 10

Is the difference of the means significant ? (Test at 5% level of significance)

(The value of t for 9 degrees of freedom at 5% level is 2.262)

Solution

Given that

	Population 1	Population 2
Mean	$\bar{x}_1 = 40$	$\bar{x}_2 = 50$
S.D.	$s_1 = 8$	$s_2 = 10$
Sample size	$n_1 = 6$	$n_2 = 5$

Step 1 Null hypothesis : $H_0 : \mu_1 = \mu_2$ (There is no significant difference between two population means)

Step 2 Alternative hypothesis : $H_1 = \mu_1 \neq \mu_2$ (There is significant difference between two population means)

Step 3 Level of significance : $\alpha = 0.05$

Step 4 Degrees of freedom :

$$\nu = n_1 + n_2 - 2 = 6 + 5 - 2 = 9.$$

Step 5 Critical region : Two-tailed test and the critical region is

$$t < -2.262 \text{ and } t > 2.262. \quad \text{(Using Table II(a))}$$

Step 6 Calculation of statistic :

Using (7.40),

$$\sigma^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$= \frac{(6-1)(8)^2 + (5-1)(10)^2}{6+5-2}$$

$$= \frac{(5)(64) + (4)(100)}{9}$$

$$= \frac{320+400}{9}$$

$$= 80.$$

Therefore,

$$\sigma = 8.94.$$

Using (7.42),

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{40 - 50}{(8.94)\sqrt{\frac{1}{6} + \frac{1}{5}}}$$

$$= -\frac{10}{5.44}$$

$$= -1.838.$$

Step 7 Decision : Accept the null hypothesis H_0 since

$$-2.262 < t = -1.838 < 2.262.$$

Thus, we conclude that there is no significant difference between two population means. *Answer*

Example 7.18

A large group of teachers are trained under QIP (Quality Improvement Programme), where some are trained by institution *A* and some are trained by institution *B*. In a random sample of 10 teachers taken from a large group, the following marks are obtained in an appropriate achivement test.

Institution *A* :	65	69	73	71	75	66	71	68	68	74
Institution *B* :	78	69	72	77	84	70	73	77	75	65

Test the claim that institution *B* is more effective at 0.05 level of significance under the assumption that the two populations are normally distributed with same variances.

Solution

Given that $n_A = n_B = 10$.

Let $\bar{x}_A$ = Average marks obtained in appropriate achievement test by teachers trained by institution A.

Therefore,

$$\bar{x}_A = \frac{65+69+73+71+75+66+71+68+68+74}{10} \quad \text{(Using 7.41)}$$

$$= \frac{700}{10}$$

$$= 70.$$

Similarly,

$$\bar{x}_B = \frac{78+69+72+77+84+70+73+77+75+65}{10} \quad \text{(Using 7.41)}$$

$$= \frac{740}{10}$$

$$= 74.$$

Now,

$$s_A^2 = \frac{\sum_{i=1}^{10} (x_{A_i} - \bar{x}_A)^2}{n_A - 1} \quad \text{(Using (7.6a))}$$

$$= \frac{25+1+9+1+25+16+1+4+4+16}{10-1}$$

$$= \frac{102}{9}$$

$$= 11.33.$$

Therefore,

$$s_A = 3.37.$$

$$s_B^2 = \frac{\sum_{i=1}^{10} (x_{B_i} - \bar{x}_B)^2}{n_B - 1}$$

$$= \frac{16+25+4+9+100+16+1+9+1+81}{10-1}$$

$$= \frac{262}{9}$$

$$= 29.11.$$

Therefore,

$$s_B = 5.40.$$

Therefore,

	Population A	Population B
Mean	$\bar{x}_A = 70$	$\bar{x}_B = 74$
S.D	$s_A = 3.37$	$s_B = 5.40$
Sample size	$n_A = 10$	$n_B = 10$

Step 1 Null hypothesis $H_0 : \mu_1 = \mu_2$ (There is no difference in teaching by institution A and institution B)

Step 2 Alternative hypothesis $H_1 : \mu_1 < \mu_2$ (Institution B is more effective than institution A)

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Degrees of freedom :

$$\nu = n_A + n_B - 2 = 10 + 10 - 2 = 18.$$

Step 5 Critical region: Left one-tailed test and the critical region is

$$t < -1.734. \qquad \text{(Using Table II(a))}$$

Step 6 Calculation of statistic :
Using (7.40),

$$\sigma^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

$$= \frac{(10-1)(11.33) + (10-1)(29.11)}{10+10-2}$$

$$= \frac{(9)(11.33) + (9)(29.11)}{18}$$

$$= \frac{101.97 + 261.99}{18}$$

$$= \frac{363.96}{18}$$

$$= 20.22.$$

Therefore,

Using (7.42),

$$\sigma = 4.50.$$

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

$$= \frac{70 - 74}{(4.50)\sqrt{\frac{1}{10} + \frac{1}{10}}}$$

$$= -\frac{4}{2.012}$$

$$= -1.988.$$

Step 7 Decision : Reject the null hypothesis H_0 since

$$z = -1.988 < -1.734.$$

Thus, we conclude that the institution B is more effective than the institution A in teaching.

Example 7.19

Random samples of specimens of coal from two mines A and B are drawn and their heat producing capacity (in millions of calories per ton) were measured yielding the following results.

Mine A :	8260	8130	8350	8070	8340	
Mine B :	7950	7890	7900	8140	7920	7840

Use the 5% level of significance to test whether the difference between the means of these two samples is significant.

Solution

Given that

$$n_A = 5 \text{ and } n_B = 6.$$

Let $\bar{x}_A$ = Average heat producing capacity of coal from mine A.
Therefore,

$$\bar{x}_A = \frac{8260 + 8130 + 8350 + 8070 + 8340}{5} \qquad \text{(Using (7.41))}$$

$$= \frac{41150}{5}$$

$$= 8230.$$

Similarly,

$$\bar{x}_B = \frac{7950+7890+7900+8140+7920+7840}{6}$$

$$= 7940. \qquad \text{(Using (7.41))}$$

Now,

$$s_A^2 = \frac{\sum_{i=1}^{5}(x_{A_i} - \bar{x}_A)^2}{n_A - 1} \qquad \text{(Using (7.6a))}$$

$$= \frac{900+10000+14400+25600+12100}{5-1}$$

$$= \frac{63000}{4}$$

$$= 15750.$$

Therefore,

$$s_A = 125.50.$$

Similarly,

$$s_B^2 = 10920 \text{ and } s_B = 104.50.$$

Therefore,

	Population A	Population B
Mean	$\bar{x}_A = 8230$	$\bar{x}_B = 7940$
SD	$s_A = 125.50$	$s_B = 104.50$
Sample size	$n_A = 5$	$n_B = 6$

Step 1 Null hypothesis $H_0 : \mu_1 = \mu_2$ (There is no significant difference between the means of the given two populations)

Step 2 Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (There is significant difference between the means of the given two populations)

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Degrees of freedom :

$$v = n_A + n_B - 2 = 5+6-2 = 9.$$

Step 5 Critical region: Two-tailed test and the critical region is

$$t < -2.262 \text{ and } t > 2.262. \qquad \text{(Using Table II(a))}$$

Step 6 Calculation of statistic :

Using (7.40),

$$\sigma^2 = \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A + n_B - 2}$$

$$= \frac{(5-1)(15750)+(6-1)(10920)}{5+6-2}$$

$$= \frac{4(15750)+5(10920)}{9}$$

$$= \frac{63000+54600}{9}$$

$$= \frac{117600}{9}$$

$$= 13066.67.$$

Therefore,

$$\sigma = 114.31.$$

Using (7.42),

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sigma\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}$$

$$= \frac{8230-7940}{(114.31)\sqrt{\frac{1}{5}+\frac{1}{6}}}$$

$$= \frac{290}{(114.31)(0.606)}$$

$$= \frac{290}{69.27}$$

$$= 4.19.$$

Step 7 Decision : Reject the null hypothesis since

$$t = 4.19 > 2.262.$$

Thus, we conclude that the average heat producing capacity of the coal from the two mines is not the same. ***Answer***

Exercises 7.6

01. Samples of two types of electric light bulbs were tested for length of life and the following data were obtained.

	Type I	Type II
Sample size	$n_1 = 8$	$n_2 = 7$
Sample mean	$\bar{x}_1 = 1234$ hrs	$\bar{x}_2 = 1036$ hrs
Sample S.D.	$s_1 = 36$ hrs	$s_2 = 40$ hrs

Does the data support the hypothesis that Type I is superior to Type II regarding length of life ? (Test at 5% level of signigicance)

02. In a statistics examination 9 students of class A and 6 students of class B obtained the following marks. Test at 0.01 level of significance whether the performance in statistics is same or not for the two classes A and B. Assume that the samples are drawn from normal populations having same variance.

A	:	44	71	63	59	68	46	69	54	48
B	:	52	70	41	62	36	50			

7.13 Testing of Hypothesis for Observed Correlation Coefficients

Consider a random sample of n observations (x_i, y_i) from a bivariate normal population. Let r be the observed correlation coefficient and ρ be the population correlation coefficient.

Under the null and alternative hypothesis as follows,

$H_0 : \rho = 0$ (There is no correlation between two variables)

$H_1 : \rho \neq 0$ or $\rho > 0$ or $\rho < 0$ (There is correlation between two variables)

The test statistic t is given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad ...(7.44)$$

is a t-variate with $\nu = n-2$ degrees of freedom.

Note The steps for testing the hypothesis remains same as that of Section 7.5.

Example 7.20

The correlation coefficient between income and food expenditure for sample of 7 household from a low income group is 0.9. Using 1% level of significance, test

whether the correlation coefficient between incomes and food expenditure is positive. Assume that the population of both variables are normally distributed.

Solution

Given $r = 0.9$.

Step 1 Null hypothesis $H_0 : \rho = 0$ (There is no correlation between incomes and food expenditure)

Step 2 Alternative hypothesis $H_1 : \rho \neq 0$ (There is correlation between incomes and food expenditure)

Step 3 Level of significance : $\alpha = 1\% = 0.01$

Step 4 Degrees of freedom :

$$\nu = n - 2 = 7 - 2 = 5.$$

Step 5 Critical region : Two-tailed test and the critical region is

$t < -4.032$ and $t > 4.032$. (Using Table II(b))

Step 6 Calculation of statistic :

Using (7.44),

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{(0.9)\sqrt{7-2}}{\sqrt{1-(0.9)^2}}$$

$$= \frac{(0.9)\sqrt{5}}{\sqrt{1-0.81}}$$

$$= \frac{2.012}{0.436}$$

$$= 4.61.$$

Step 7 Decision : Reject the null hypothesis H_0 since

$t = 4.61 > 4.032$.

Thus, we conclude that there is correlation between incomes and food expenditure. ***Answer***

Example 7.21

A random sample of fifteen paired observations from a bivariate normal population

gives a correlation coefficient of –0.5. Does this signify the existance of correlation in the sampled population ? (Test at 5% level of significance)

Solution

Given $r = -0.5$.

Step 1 Null hypothesis $H_0 : \rho = 0$ (The sampled population is uncorrelated)

Step 2 Alternative hypothesis $H_1 : \rho \neq 0$ (The sampled population is correlated)

Step 3 Level of insignificance : $\alpha = 5\% = 0.05$

Step 4 Degrees of freedom :

$$\nu = n - 2 = 15 - 2 = 13.$$

Step 5 Critical region : Two-tailed test and the critical region is

$$t < -2.160 \text{ and } t > 2.160.$$ (Using Table II(a))

Step 6 Calculation of statistic :

Using (7.44),

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{(-0.5)\sqrt{15-2}}{\sqrt{1-(-0.5)^2}}$$

$$= \frac{(-0.5)\sqrt{13}}{\sqrt{1-0.25}}$$

$$= -\frac{1.803}{\sqrt{0.75}}$$

$$= 2.08.$$

Step 7 Decision : Accept the null hypothesis H_0 since

$$t = 2.08 < 2.160.$$

Thus, we conclude that the sampled population is uncorrelated. ***Answer***

7.14 Chi-Square Distirbution

The Chi-square (denoted as χ^2) distribution is a continuous probability distribution of a continuous random variable X with probability density function is given by

$$f(x) = \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}; \; x > 0,$$

which is a special case of gamma distribution with $\alpha = \nu/2$ and $\beta = 2$. Here, ν is a positive integer (known as degrees of freedom) and is the only single parameter of the distribution. As the χ^2 - distribution dependent only on one parameter ν, so it varies as ν changes. For smaller values of ν, the χ^2 - distribution is highly skewed to the right, and as the degrees of freedom ν increases, the degree of skewness decreases.

In other words,

$$z^2 = \left(\frac{x-\mu}{\sigma}\right)^2$$

is a χ^2 variate with 1 degree of freedom, where μ is mean, σ is standard deviation and x is a normal variate.

If $x_1, x_2, x_3, \ldots, x_n$ be n- independent normal variates with means $\mu_1, \mu_2, \mu_3, \ldots, \mu_n$ and standard deviations $\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_n$, respectively, then

$$\chi^2 = \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 + \left(\frac{x_3-\mu_3}{\sigma_3}\right)^2 + \ldots + \left(\frac{x_n-\mu_n}{\sigma_n}\right)^2 = \sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$$

is a χ^2 variate with n degrees of freedom.

Theorem 7.3 If s^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} \quad (7.45)$$

is a random variable having the χ^2-distribution with the parameter $\nu = n-1$.

➢ Properties of χ^2- distirbution

(1) The probability curve of a χ^2 - distribution is shown in Figure 7.8.

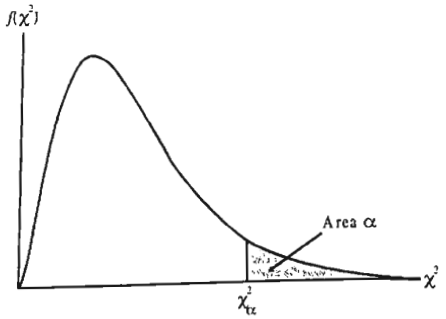


Figure 7.8 χ^2 - distribution

The curve is skewed towards right and its shape varies with the degrees of freedon $\nu = n-1$. It lies completely in the first quadrant. Since χ^2 varies from 0 to ∞; that is, χ^2 - distribution is not symmetrical.

(2) Mean = degree of freedom ν

(3) Mode = $\nu - 2$

(4) S.D. = $\sqrt{2\nu}$

(5) If ${\chi_1}^2$ and ${\chi_2}^2$ are two independent distributions with ν_1 and ν_2 degrees of freedom, then ${\chi_1}^2 + {\chi_2}^2$ will be χ^2 - distribution with $\nu_1 + \nu_2$ degree of freedom.

➢ **Analysis of $r \times c$ Tables (Contigency Tables)** A contigency table consisting of finite number of rows and columns, depending on the classification in which attributes are divided into classes (categories). For simplicity, suppose attribute A is divided into 2 classes A_1, A_2, whereas another attribute B is divided into 3 classes B_1, B_2, B_3. Then, in this case, we have 2×3 contigency table as follows.

A	B			Row Total
	B_1	B_2	B_3	(RT)
A_1	O_{11}	O_{12}	O_{13}	RT_1
A_2	O_{21}	O_{22}	O_{23}	RT_2
Column Total (CT)	CT_1	CT_2	CT_3	N

when O_{ij} known as ***observed frequencies*** and it denotes the number of items possessing both the attributes A_i and B_j $(i = 1, 2; j = 1, 2, 3)$, and N indicates total frequency as

$$N = \sum_{i=1}^{2} A_i = \sum_{i=1}^{2} B_j .$$

Note 1 RT and CT are known as ***marginal frequencies***.

Note 2 A contigency table with r rows (that is, A_i; $i = 1, 2, \ldots, r$) and c columns (that is, B_j; $j = 1, 2, \ldots, c$) is referred to as $r \times c$ table. A contigency table arise in essentially two kinds of problems.

(1) In the problem where we have samples from several populations, with each trial permits more than two possible outcomes.

(2) In the problem where we have samples from one population with each item are classified with respect to two attributes.

7.14.1 χ^2 - Test for Independence of Attributes

Part A Construct a ***contigency table*** on the basis of given information and find ***expected frequency*** for each cell using

$$E_{ij} = \frac{\text{column total} \times \text{row total}}{\text{grand total}} \qquad \ldots(7.46)$$

Part B Testing under the null and alternative hypothesis as follows.

H_0 : Attibutes are independent; that is, there is no significant difference between observed frequencies (O_{ij}) and expected frequencies (E_{ij})

H_1: Attributes are dependent; that is, there is significant difference between observed frequencies (O_{ij}) and expected frequencies (E_{ij})

The test statistic χ^2 for the analysis of $r \times c$ table is given by

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \ldots(7.47)$$

with degree of freedom $(r-1)(c-1)$.

Here, the hypothesis H_0 is tested using right one-tailed test.

Note The steps for testing the hypothesis remains same as that of Section 7.4.

Example 7.22

A company operates three machines on three different shifts daily. The following table presents the data of the machine breakdowns resulted during a 6-month time period.

Shift	Machine			Total
	A	*B*	*C*	
1	12	12	11	35
2	15	25	13	53
3	17	23	10	50
Total	44	60	34	138

Test the hypothesis that for an arbitrary breakdown the machine causing the breakdown and the shift on which the breakdown occurs are independent. (Take level of significance 5%)

Solution

Part A Contigency Table

Let us first prepare the contigency table with the expected frequency for each cell using formula (7.46).

$$E_{ij} = \frac{\text{column total} \times \text{row total}}{\text{grand total}}.$$

Shift	Machine			Total
	A	*B*	*C*	
1	12 (11.16)	12 (15.22)	11 (8.62)	35
2	15 (16.90)	25 (23.04)	13 (13.06)	53
3	17 (15.94)	23 (21.74)	10 (12.32)	50
Total	44	60	34	138

where

$$E_{11} = \frac{44 \times 35}{138} = 11.16,$$

and similarly others shown in table inside ( ).

Part B Testing

Step 1 Null hypothesis H_0 : For an arbitrary breakdown the machine and the shift are independent

Step 2 Alternative hypothesis H_1 : For an arbitrary breakdown the machine and the shift are not independent

Step 3 Level of significance : $\alpha = 5\% = 0.05$

Step 4 Degrees of freedom :

$$\nu = (r-1)(c-1) = (3-1)(3-1) = (2)(2) = 4.$$

Step 5 Critical region : Right one-tailed test and the critical region is

$$\chi^2 > 9.488.$$ (Using Table III(b))

Step 6 Calculation of statistic :

Using (7.47),

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(12-11.16)^2}{11.16} + \frac{(12-15.22)^2}{15.22} + \frac{(11-8.62)^2}{8.62}$$

$$+ \frac{(15-16.90)^2}{16.90} + \frac{(25-23.04)^2}{23.04} + \frac{(13-13.06)^2}{13.06}$$

$$+ \frac{(17-15.94)^2}{15.94} + \frac{(23-21.74)^2}{21.74} + \frac{(10-12.32)^2}{12.32}$$

$$= \frac{0.706}{11.16} + \frac{10.368}{15.22} + \frac{5.664}{8.62}$$

$$+ \frac{3.61}{16.90} + \frac{3.842}{23.04} + \frac{0.004}{13.06}$$

$$+ \frac{1.124}{15.94} + \frac{1.588}{21.74} + \frac{5.382}{12.32}$$

$$= 0.063 + 0.681 + 0.657$$
$$+ 0.214 + 0.167 + 0.0003$$
$$+ 0.071 + 0.073 + 0.437$$

$$= 2.36.$$

Step 7 Decision : Accept the null hypothesis H_0 since

$$\chi^2 = 2.36 < 9.488 .$$

Thus, we conclude that for an arbitrary breakdown the machine and the shift are independent. *Answer*

Example 7.23

Test the hypothesis at 0.05 level of significance that the presence or absence of hypertension (HT) is independent of smoking habits from the following data of 80 persons.

	Non smokers	Moderate smokers	Heavy smokers
HT	21	36	30
No HT	48	26	19

Solution

Part A Contigency Table

The contigency table with the expected frequency for each cell using formula (7.46),

$$E_{ij} = \frac{\text{column total} \times \text{row total}}{\text{grand total}}$$

is as follows.

	Non smokers	Moderate smokers	Heavy smokers	Total
HT	21 (33.35)	36 (29.97)	30 (23.68)	87
No HT	48 (35.65)	26 (32.03)	19 (25.32)	93
Total	69	62	49	180

where

$$E_{11} = \frac{69 \times 87}{180} = 33.35 ,$$

and similarly others shown in table inside ( ).

Part B Testing

Step 1 Null hypothesis H_0 : Hypertension and smoking habits are independent.

Step 2 Alternative hypothesis H_1 : Hypertension and smoking habits are not independent.

Step 3 Level of significance : $\alpha = 0.05$

Step 4 Degrees of freedon :

$$\nu = (r-1)(c-1) = (2-1)(3-1) = (1)(2) = 2.$$

Step 5 Critical region : Right one tailed test and the critical region is

$$\chi^2 = 5.991 .$$ (Using Table III(b))

Step 6 Calculation of statistic :

Using (7.47),

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(21-33.35)^2}{33.35} + \frac{(36-29.97)^2}{29.97} + \frac{(30-23.68)^2}{23.68}$$

$$+ \frac{(48-35.65)^2}{35.65} + \frac{(26-32.03)^2}{32.03} + \frac{(19-25.32)^2}{25.32}$$

$$= 14.46.$$

Step 7 Decision : Reject the null hypothesis H_0 since

$$\chi^2 = 14.46 > 5.991 .$$

Thus, we conclude that Hypertension and smoking habits are not independent. *Answer*

Exercises 7.7

01. The following table shows efficiency in job and academic performance of 400 persons. Test at 0.01 level of significance that whether efficiency in jobs depends on academic performance.

		Acadmic performance			
		Excellent	Good	Satisfactory	Total
Efficiency	Excellent	23	60	29	112
	Good	28	79	60	167
	Satisfactory	9	49	63	121
	Total	60	188	152	400

02. From the following data, use χ^2 - test to conclude whether inoculation is effective in

preventing turberculosis.

	Attacked	Not Attacked	Total
Inoculation	31	469	500
Non inoculation	185	1315	1500
Total	216	1784	2000

7.14.2 Goodness of Fit In many random experiments we are interested to know whether a particular probabilistic model (Poisson, binomial, normal) is appropriate or not. For this it is important to know that how closely the actual distribution approximate the assumed theoretical distributions. The statistical test which compare the observed frequencies (o_i from the sample) with the corresponding values of the expected frequencies (e_i from theoretical frequencies) is known as ***goodness-of-fit*** test.

The statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \qquad ...(7.48)$$

is used to measure the discrepancy existing between the observed and expected frequencies. Here, o_i and e_i are the observed and expected frequencies of the i^{th} cell (or class interval) such that

$$\sum_{i=1}^{k} o_i = \sum_{i=1}^{k} e_i = N \text{, total frequency,}$$

k is the number of cells or class intervals in the given frequency distribution.

When n is large, the statistic χ^2 defined by (7.48) follows chi-square distribution with $\nu = k - m$ degrees of freedom, where m is the number of constraints applied to the observed data to calculate the expected frequencies.

Note 1 If $\chi^2 = 0$, then o_i and e_i agree exactly.

Note 2 When $\chi^2 > 0$ and it is small, then o_i are close to e_i, indicating ***good*** fit.

Note 3 When $\chi^2 > 0$ and it is large, then o_i differs considerably from e_i, indicating ***poor*** fit.

➢ **Conditions for Validity of χ^2 - test**

1. Sample size n should be large ($n \geq 50$).

2. The number of classes k should not be neither too small nor too large. In general, $4 \leq k \leq 16$.
3. When the expected frequency is less than 5, then pool some of the neighbouring data so that expected frequency greater than or equal to 5.

   The degree of freedom changes accordingly.

➢ **χ^2-Test for Goodness of Fit**

Part A Find the expected frequencies using general probability considerations or specific probability model (Poisson, Binomial, Normal) given in the problem itself.

Part B Testing under the null and alternative hypothesis as follows.

H_0 : Given probability distribution fits good with the given data (that is, there is no significant difference between observed frequencies (o_i) and expected frequencies (e_i)

H_1 : Given probability distribution does not fit good with the given data (that is, there is significant difference between observed frequencies (o_i) and expected frequencies (e_i).

The test statistic given by (7.48) is used with $\nu = k - m$ degrees of freedom.

Note The steps for testing the hypothesis remains same as that of Section 7.4.

Example 7.24

Suppose that a dice is tossed 120 times and the recorded data is as follows.

Face (x) :	1	2	3	4	5	6
Observed frequency :	20	22	17	18	19	24

Test the hypothesis that the dice is unbiased at $\alpha = 0.05$.

Solution

Part A On the basis of the null hypothesis that dice is unbiased the probability p_i for the face i is 1/6. Thus, we have following table.

Face (x) :	1	2	3	5	5	6
$p(x)$:	1/6	1/6	1/6	1/6	1/6	1/6
Expected frequency (e) : $= 120 \times p(x)$	20	20	20	20	20	20

Step 1 Null hypothesis H_0 : The dice is unbiased; that is, $p_1 = p_2 = ... = p_6 = 1/6$

Step 2 Alternative hypothesis H_1 : The dice is biased

Step 3 Level of significance : $\alpha = 0.05$

Step 4 Degree of freedom :

$$\nu = k - m = 6 - 1 = 5.$$

(The number of degrees of freedom is $6 - 1 = 5$ since only one quantity, the total frequency of 120, is needed from the observed data to calculate the expected frequencies)

Step 5 Critical region : Right one-tailed test and the critical region is

$$\chi^2 > 11.070. \qquad \text{(Using Table III(b))}$$

Step 6 Calculation of Statistic :

Using (7.48),

$$\chi^2 = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(0-0)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20}$$

$$= 0 + \frac{4}{20} + \frac{9}{20} + \frac{4}{20} + \frac{1}{20} + \frac{16}{20}$$

$$= \frac{4+9+4+1+16}{20}$$

$$= \frac{34}{20}$$

$$= 1.7.$$

Step 7 Decision : Accept the null hypothesis H_0 since

$$\chi^2 = 1.7 < 11.070.$$

Thus, we conclude that the dice may be considered to be unbiased. ***Answer***

Example 7.25

Suppose that during 400 five-minute intervals the air-traffic control of an airport

received 0, 1, 2, ... , or 13 radio messages with respective frequencies of 3, 15, 47, 76, 68, 74, 46, 39, 15, 9, 5, 2, 0 and 1. Test at 0.05 level of significance, the hypothesis that the number of radio messages received during a 5 minute interval follow poisson distribution with $\lambda = 4.6$.

Solution

Part A Let the random variable X be the number of radio messages received during a 5 minute interval and $p(x)$ be the probability of receiving x messages. Set the null hypothesis that X follows a Poisson distribution with parameter 4.6. The Poisson distribution that fits to the data with parameter $\lambda = 4.6$ is

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-4.6}(4.6)^x}{x!}; \; x = 0, 1, 2, \ldots, 13.$$

Thus, we have following table.

Number of radio messages (X)	Observed frequencies (o)	Poisson probabilities $p(x)$	Expected frequencies (e) $= 400 \times p(x)$
0	3	0.010	4.0
1	15	0.046	18.4
2	47	0.107	42.8
3	76	0.163	65.2
4	68	0.187	74.8
5	74	0.173	69.2
6	46	0.132	52.8
7	39	0.087	34.8
8	15	0.050	20.0
9	9	0.025	10.0
10	5	0.012	4.8
11	2	0.005	2.0
12	0	0.002	0.8
13	1	0.001	0.4
	400		400.0

According to condition 3 of validity of χ^2 - test, no expected frequency should be less than 5. So we pool the first two expected frequencies and the last four expected frequencies. Thus, the modified frequencies are as follows.

Observed frequencies (o)	Expected frequencies (e)
18	22.4

47	42.8
76	65.2
68	74.8
74	69.2
46	52.8
39	34.8
15	20.0
9	10.0
8	8.0
400	400.0

Part B Testing

Step 1 Null hypothesis H_0 : Random variable X has Poisson distribution with $\lambda = 4.6$

Step 2 Alternative hypothesis H_1 : Random variable X does not have Poisson distribution with $\lambda = 4.6$

Step 3 Level of significance : $\alpha = 0.05$

Step 4 Degrees of freedom :

$$\nu = k - m = 10 - 1 = 9.$$

(The number of degrees of freedom is $10 - 1 = 9$ since only one quantity, the total freuency of 400, is needed from the observed data to calculate the expected frequencies)

Step 5 Critical region : Right one-tailed list and the critical region is

$$\chi^2 > 16.919.$$ (Using Table III(b))

Step 6 Calculation of statistic :
Using (7.48),

$$\chi^2 = \sum_{i=1}^{14} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(4.4)^2}{22.4} + \frac{(4.2)^2}{42.8} + \frac{(10.8)^2}{65.2} + \frac{(6.8)^2}{74.8}$$

$$+ \frac{(4.8)^2}{69.2} + \frac{(5.8)^2}{52.8} + \frac{(4.2)^2}{34.8} + \frac{(5)^2}{20}$$

$$+ \frac{(1)^2}{10} + \frac{(0)^2}{8}.$$

$$= 6.749.$$

Step 7 Decision : Accept the null hypothesis H_0 since

$$\chi^2 = 6.749 < 16.919.$$

Thus, we conclude that Poisson distribution with $\lambda = 4.6$ provides a good fit.

Answer

Example 7.26

The following table indicates (a) the frequencies of a given distribution with (b) the frequencies of the normal distribution having the same mean, standard deviation and the total frequency as in (a).

(a)	1	5	20	28	42	22	15	5	2
(b)	1	6	18	25	40	25	18	6	1

Apply the χ^2 - test of goodness of fit. (Take level of significance 5%)

Solution

Part A According to condition 3 of validity of χ^2 -test, no expected frequency should be less than 5. So we pool the first two and last two expected frequencies. Thus, the modified frequencies are as follows.

o	e
6	7
20	18
28	25
42	40
22	25
15	18
7	7

Part B Testing

Step 1 Null hypothesis H_0 : There is no significant difference between observed frequencies (o) and expected frequencies (e)

Step 2 Alternative hypothesis H_1 : There is significant difference between observed frequencies (o) and expected frequencies (e)

Step 3 Level of significance : $\alpha = 0.05$

Step 5 Degrees of freedom :

$$\nu = k - m = 7 - 3 = 4.$$

(The number of degrees of freedom is $7 - 3 = 4$ since mean, standard deviation and the total frequency of the original distribution have been used

in calculating the theoretical frequencies)

Step 5 Critical region : Righ one-tailed test and the critical region is

$$\chi^2 > 9.488. \qquad \text{(Using Table III(b))}$$

Step 6 Calculation of statistic :
Using (7.48),

$$\chi^2 = \sum_{i=1}^{7} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(1)^2}{7} + \frac{(2)^2}{18} + \frac{(3)^2}{25} + \frac{(2)^2}{40} + \frac{(-3)^2}{25} + \frac{(-3)^2}{18} + \frac{(0)^2}{7}$$

$$= \frac{1}{7} + \frac{4}{18} + \frac{9}{25} + \frac{4}{40} + \frac{9}{25} + \frac{9}{18} + \frac{0}{7}$$

$$= 0.143 + 0.222 + 0.36 + 0.1 + 0.36 + 0.5 + 0$$

$$= 1.685.$$

Step 7 Decision : Accept the null hypothesis H_0 since

$$\chi^2 = 1.685 < 9.488.$$

Thus, we conclude that the fit is good. ***Answer***

Exercises 7.8

01. Test for goodness of fit of a Poisson distribution at 0.05 level of significance to the following frequency distribution.

x :	0	1	2	3	4	5	6	7	8
f :	52	151	130	102	45	12	5	1	2

02. A large company believes that many of its employees are taking advantage of a liberal absence policy by taking off a disproportionate number of Mondays and Fridays. The following set of data, showing the number of employee absences by the day of the week, is collected.

Day	Monday	Tuesday	Wednesday	Thrusday	Friday
0	57	39	37	54	63

Does this set of data indicate that the company's suspicions are valid ? Test at the $\alpha = 0.05$ level of significance.

7.15 *F*-Distribution (Variance Ratio Distribution)

Let s_1^2 be the variance of an independent sample $(x_1, x_2, ..., x_{n_1})$ of size n_1 drawn from a normal population having mean μ_1 and variance σ_1^2. Similarly, let s_2^2 be the variance of an independent sample $(y_1, y_2, ..., y_{n_2})$ of size n_2 drawn from another normal population having mean μ_2 and variance σ_2^2. Then the sampling distribution of the ratio of the variances of the two independent random sample is defined by

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}, \; s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2, \; s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2, \qquad ...(7.49)$$

which is an F-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

Under the null hypothesis that the population variances σ_1^2 and σ_2^2 are the same; that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the above variance ratio statistic F is given by

$$F = \frac{s_1^2}{s_2^2}. \qquad ...(7.50)$$

More formally, we have following theorem.

Theorem 7.4 If s_1^2 and s_2^2 are the variances of independent random samples 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 parameters $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ (known as degrees of freedom).

F-determines whether the ratio of the two sample variances s_1 and s_2 is t small or too large. When F is close to 1, the two sample variances s_1 and s_2 nearly same. Generally, the greator of the two variances s_1^2 and s_2^2 is taker numerator and ν_1 corresponds to the variance in the numerator.

The probability curve for F distribution is shown in Figure 7.9.

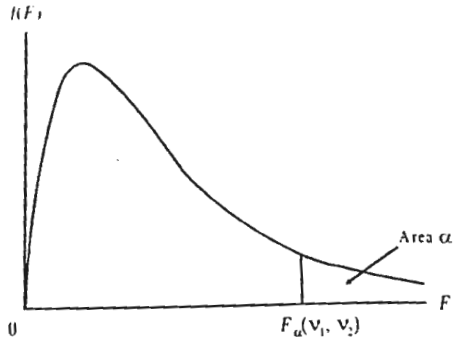


Figure 7.9 F- distribution

The curve is not symmetric and the shape depends on the degrees of freedom ν_1 and ν_2 and their order. F is always a positive number. The F- distribution curve lies entirely in the first quadrant.

For an F variate the reciprocal relation

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_1, \nu_2)} \qquad ...(7.51)$$

also hold.

Note F- distribution forms the backbone of analysis of variance (ANOVA).

Example 7.27

There are two different choices to stimulate a certain chemical process. To test whether the variance of the yield is the same no matter which catalyst is used. A sample of 10 batches is produced using the first catalyst, and of 11 using the second. If the resulting data is $s_1^2 = 0.14$ and $s_2^2 = 0.28$. Test the hypothesis of equal variance at 2% level.

Solution

Step 1 Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$; that is, equality of variances

Step 2 Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$; that is, difference in variances

Step 3 Level of significance : $\alpha = 2\% = 0.02$

Step 4 Calculation of statistic :

Using (7.50),

$$F = \frac{s_2^2}{s_1^2}$$

$$= \frac{0.28}{0.14}$$

$$= 2.$$

Step 5 Degrees of freedom :

$$\nu_1 = n_1 - 1 = 11 - 1 = 10,$$

$$\nu_2 = n_2 - 1 = 10 - 1 = 9.$$

Step 6 Critical region :

$F_{0.2}(10,9)$ two-tailed alternative $= F_{0.1}(10,9)$ at right-tailed alternative.

Therefore, critical region is

$$F > 5.26. \qquad \text{(Using Table IV(d))}$$

Step 7 Decision : Accept the null hypothesis H_0 since

$$F = 2 < 5.26.$$

Answer

Example 7.28

For the following two independent samples of any problem, test the equality of variances at 10% level of significance.

x :	15.0	8.0	3.8	6.4	27.4	19.0	35.3	13.6	
y :	18.8	23.1	10.3	8.0	18.0	10.2	15.2	19.0	20.2

Solution

Step 1 Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$; that is, equality of variances

Step 2 Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$; that is, difference in variances

Step 3 Level of significance : $\alpha = 10\% = 0.10$

Step 4 Calculation of statistic :

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{n_1}$$

$$= \frac{15.0+8.0+3.8+6.4+27.4+19.0+35.3+13.6}{8}$$

$$= \frac{128.5}{8}$$

$$= 16.06.$$

$$\sum_{i=1}^{8}(x_i-\bar{x})^2 = (-1.06)^2+(-8.06)^2+(-12.26)^2+(-9.66)^2$$

$$+(11.34)^2+(2.94)^2+(19.24)^2+(-2.46)^2$$

$$= 1.124 + 64.964 + 150.308 + 93.316$$

$$+ 128.596 + 8.644 + 370.178 + 6.052$$

$$= 823.182. \qquad ...(i)$$

Therefore,

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{8}(x_i-\bar{x})^2$$

$$= \frac{1}{8-1}(823.182) \qquad \text{(Using (i))}$$

$$= \frac{823.182}{7}$$

$$= 117.5974.$$

Similarly,

$$\bar{y} = \frac{\sum_{i=1}^{9} y_i}{n_2}$$

$$= \frac{18.8+23.1+10.3+8.0+18.0+10.2+15.2+19.0+20.2}{9}$$

$$= \frac{142.8}{9}$$

$$= 15.87.$$

$$\sum_{i=1}^{9}(y_i-\bar{y})^2 = (2.93)^2+(7.23)^2+(-5.57)^2+(-7.87)^2$$

$$+(2.13)^2+(-5.67)^2+(-0.67)^2+(3.13)^2+(4.33)^2$$

$$= 8.585 + 52.273 + 31.025 + 61.937$$

$$+ 4.537 + 32.149 + 0.449 + 9.797 + 18.619$$

$$= 219.371. \qquad ...(ii)$$

Therefore,

$$s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{9}(y_i-\bar{y})^2$$

$$= \frac{1}{9-1}(219.371) \qquad \text{(Using (ii))}$$

$$= \frac{219.371}{8}$$

$$= 27.4214.$$

Using (7.50),

$$F = \frac{s_1^2}{s_2^2}$$

$$= \frac{117.5974}{27.4214}$$

$$= 4.29.$$

Step 5 Degrees of freedom :

$$\nu_1 = n_1 - 1 = 8-1 = 7,$$

$$\nu_2 = n_2 - 1 = 9-1 = 8.$$

Step 6 Critical region :

$F_{0.10}(7, 8)$ two-tailed alternative $= F_{0.05}(7, 8)$ at right-tailed alternative.

Therefore, critical region is

$$F > 3.50. \qquad \text{(Using Table IV(a))}$$

Step 7 Decision : Reject the null hypothesis H_0 since

$$F = 4.29 > 3.50.$$

Answer

7.16 Summary of Various Test Statistic and Their Applications

Statistics	Sample size	Area of Application
z	$n \geq 30$ (Large)	1. Specified population mean
	$n_1 + n_2 \geq 30$ (Large)	(Section 7.5.1)
		2. Difference of two population means (Section 7.8.1)
		3. Specified population proportion (Section 7.6.1)
		4. Difference of two population proportions (Section 7.9.1)
t	$n < 30$ (Small)	1. Specified population mean (Section 7.11.1)
	$n_1 + n_2 < 30$ (Small)	2. Difference of two population means (Section 7.12)
		3. Observed correlation coefficients (Section 7.13)
χ^2	Applicable to all sample size	1. Independence of attributes (Section 7.14.1)
		2. Goodness of fit (Section 7.14.2)
F	Applicable to all sample size.	1. Ratio of variances (Section 7.15)

Answers

Exercises 7.1

01. Accept $H_0 : \mu = 125$ hours/year

Decision : Cannot believe that machine works more than 125 hours in an year

02. Reject $H_0 : \mu = 1200$ bills/year

Decision : The new system represents an improvement over the old system

03. Reject $H_0 : \mu = 28000$ miles

Decision : Tyres run for < 28000 miles

04. Accept $H_0 : \mu = 115$ (the caliber of recent applicants has not changed)

Decision : The caliber of recent applicants has not changed

Exercises 7.2

01. Accept $H_0 : P = 0.02$

Decision : Manufacturer's claim may be accepted

02. Reject $H_0 : P = 0.18$

Decision : The proportion of households with school age children that own computers seems higher than the reported proportion of all families in the area

Exercises 7.3

01. Accept $H_0 : \mu_A = \mu_B$

Decision : Thre is no difference between type A and B

02. Accept $H_0 : \mu_1 = \mu_2$ (There is no significant difference between the means of the batting averages between the players in 1990 and 2015)

Decision : There is no difference between the means of the batting averages between the players in 1990 and 2015

03. Reject $H_0 : \mu_1 = \mu_2$ (Average income of men and women is same)

Decision : Average income of men and women is not same

Exercises 7.4

01. Accept H_0

Decision : There is no difference in the proportions of men and women who support Candidate A

02. Accept H_0 : Method A and method B are equivalent

Decision : Method A and method B are equivalent

03. (*a*) Accept H_0 : There is no significant difference between performance of A.C. machines produced by company A and B

(*b*) Accept H_0 : Products of B are not superior than products of A

Exercises 7.5

01. Reject $H_0 : \mu = 14.4$

Decision : Incentives announced were effective

02. Accept $H_0 : \mu = 20$

Decision : The claim of the company is accepted

Exercises 7.6

01. Reject $H_0 : \mu_1 = \mu_2$

02. Accept $H_0 : \mu_1 = \mu_2$ (There is no difference in performance)

Decision : There is no difference between two classes A and B in performance in statistics examination

Exercises 7.7

01. Reject H_0

Decision : There is dependency between efficiency in job and academic performance

02. Reject H_0

Decision : Inoculation is effective in preventing turberculosis

Exercises 7.8

01. Degrees of freedom : 7

Accept H_0 : The data is good fitted with Poisson distribution

02. Degrees of freedom : 4

Reject H_0 : There is no difference in the number of employee absences based on the day of the week