Measures of Central Tendency:

The following are the five measures of central tendency that are in common use:

i. Arithmetic mean or simple mean
ii. Median
iii. Mode
iv. Geometric mean
v. Harmonic mean

Arithmetic Mean:

Arithmetic mean of a set of observations is their sum divided by number of observations, for example, the arithmetic mean $\bar{x}$ of n observations $x_{1,} x_2, x_3, \dots, x_n$ is given by:

$$\boldsymbol{\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

In case the frequency distribution $f_i,\ i = 1,2,3,\dots,n$, where f_i is the frequency of the variable x_i,

$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n}{f_1 + f_2 + f_2 + \cdots + f_n} = \frac{1}{N}\sum_{i=1}^{n} f_ix_i$, where $N = \sum_{i=1}^{n} f_i$

In case of grouped or continuous frequency distribution, x is taken as the as the mid value of the corresponding class.

It may be noted that if the values of x or f are large, the calculation of mean by formula is $\frac{1}{N}\sum_{i=1}^{n} f_ix_i$ is quite time-consuming and tedious. The arithmetic is reduced to a great extent by taking the deviations of the given values from any arbitrary point 'A' as explained below:

Let $d_i = x_i - A$. Then $f_id_i = f_i(x_i - A) = f_ix_i - Af_i$

Summing both sides over i from 1 to n, we get

$$\sum_{i=1}^{n} f_id_i = \sum_{i=1}^{n} f_ix_i - A\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_ix_i - AN$$

$$\frac{1}{N}\sum_{i=1}^{n} f_id_i = \frac{1}{N}\sum_{i=1}^{n} f_ix_i - \frac{1}{N}A\sum_{i=1}^{n} f_i = \frac{1}{N}\sum_{i=1}^{n} f_ix_i - A$$

$$\frac{1}{N}\sum_{i=1}^{n} f_i d_i = \bar{x} - A$$

Where $\bar{x}$ is the arithmetic mean of the distribution.

$$\bar{x} = A + \frac{1}{N}\sum_{i=1}^{n} f_i d_i$$

In case of grouped (or) continuous frequency distribution, the arithmetic is reduced to still greater extent by taking point h is the common magnitude of class interval. In this case, we have $\mathrm{h}d_i = x_i - A$ and proceeding exactly similarly above, we get

$$\bar{x} = A + \frac{\mathrm{h}}{N}\sum_{i=1}^{n} f_i d_i$$

Problem 1:

The intelligence quotient (IQ's) of 10 boys in a class are given below:

70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Then find the mean I.Q.

Solution:

Mean I.Q of 10 boys in a class are given below:

$$\bar{X} = \frac{\sum X}{n} = \frac{70+120+110+101+88+83+95+98+107+100}{10} = \frac{972}{10} = 97.2$$

Problem 2:

The following frequency is distribution of the number of telephone calls received in 245 successive one-minute intervals at an exchange:

No. of calls	0	1	2	3	4	5	6	7
frequency	14	21	25	43	51	40	39	12

Obtain the mean number of calls per minute.

Solution:

No. of Calls (X)	Frequency (f)	fX
0	14	0
1	21	21
2	25	50
3	43	129
4	51	204
5	40	200
6	39	234
7	12	84
	$N = 245$	$\sum fX = 922$

Mean number of calls per minute at the at the exchange is given by

$$\bar{X} = \frac{\sum fX}{N} = \frac{922}{245} = 3.763$$

Problem 3:

Find the arithmetic mean of the following frequency distribution:

X	1	2	3	4	5	6	7
f	5	9	12	17	14	10	6

Solution:

X	f	fX

1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	7	42
Total	**73**	**299**

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{7} f_i x_i = \frac{299}{73} = 4.09$$

Problem 4:

For a certain frequency table which has only been partly reproduced here, the mean was found to be 1.46.

0	1	2	3	4	5	Total
46	?	?	25	10	5	200

Calculate the missing frequencies.

Solution:

Let X denote the number of accidents and let the missing frequencies corresponding to $X = 1$ and $X = 2$ be f_1 and f_2 respectively.

No. of accidents (X)	Frequency (f)	fX
0	46	0
1	f_1	f_1
2	f_2	$2f_2$
3	25	75

4	10	40
5	5	25
	$86 + f_1 + f_2 = 200$	$140 + f_1 + 2f_2$

$$200 = 86 + f_1 + f_2$$

$$f_1 + f_2 = 200 - 86 = 114$$

$$f_1 + f_2 = 114 \qquad (1)$$

$$\bar{X} = \frac{1}{N}\sum fX = \frac{f_1 + 2f_2 + 140}{200} = 1.46$$

$$f_1 + 2f_2 + 140 = 1.46 \times 200 = 292$$

$$f_1 + 2f_2 = 292 - 140 = 152$$

$$f_1 + 2f_2 = 152 \qquad (2)$$

Solving equations (1) and (2), we get

$$f_1 = 38, \; f_2 = 76$$

Problem 5:

Calculate the arithmetic mean of the marks from the following table:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	12	18	27	20	17	6

Solution:

Marks	No. of Students (f)	Mid-point (X)	(f)
0-10	12	5	60

10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330
Total	**100**		**2800**

$$Arithmetic\ mean = \bar{x} = \frac{1}{N}\sum fx = \frac{1}{100} \times 2800 = 28$$

Problem 6:

Calculate the mean for the following frequency distribution:

Class interval	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	8	7	16	24	15	7

Solution:

Here we take $A = 28$, $\mathrm{h} = 8$

Class interval	Mid-value (x)	Frequency (f)	$d = \frac{x - A}{\mathrm{h}}$	(fd)
0-8	4	8	-3	-24
8-16	12	7	-2	-14
16-24	20	16	-1	-16
24-32	28	24	0	0
32-40	36	15	1	15
40-48	44	7	2	14
Total		**77**		**-25**

$$\bar{x} = A + \frac{h\sum fd}{N}$$

$$= 28 + \frac{8 \times (-25)}{77} = 28 - \frac{20}{77} = 25.404$$

Problem 7: **Try this?**

Calculate the mean for the following frequency distribution:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of students	6	5	8	15	7	6	3

Median:

Median of a distribution is the value of the variable which divides it into two equal parts. It is the value which exceeds and is exceeded by the same number of observations. That is, it is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a positional average.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observations, there are two middle terms are median is obtained by taking the arithmetic mean of the middle terms. For example, the median of the values 8,4,7,6,2, i.e., 2,4,6,7,8 is 6 and the median of 10,15,30,70,40,80, i.e., 10,15,30,40,70,80 is $\frac{1}{2}(30 + 40) = 35$.

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

- Find $\frac{N}{2}$, where $N = \sum_{i=1}^{n} f_i$

- See the (less than) cumulative frequency (c.f.) just greater than $\frac{N}{2}$.
- The corresponding value of x is median.

Example 1:

Obtain the median for the following frequency distribution:

x	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6

Solution:

x	f	$c.f.$
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	$N = 120$	

$N = 120$

$$\frac{N}{2} = 60$$

The cumulative frequency (c.f.) just greater than $\frac{N}{2}$ is 65 and the value of x corresponding to 65 is 5. So, median is 5.

Example 2:

Eight coins were tossed together and the number of heads (x) resulting was noted. The operation was repeated 256 times and the frequency distribution of the number of heads is given below:

No. of heads (x)	0	1	2	3	4	5	6	7	8
Frequency (f)	1	9	26	59	72	52	29	7	1

Find the median.

Solution:

x	f	$c.f.$
0	1	1
1	9	10
2	26	36
3	59	95
4	72	167
5	52	219
6	29	248
7	7	255
8	1	256
	N =256	

$$N = 256$$

$$\frac{N}{2} = 128$$

The cumulative frequency (c.f.) just greater than $\frac{N}{2}$ is 128 and the value of x corresponding to 167 is 5. So, median is 4.

Derivation of Median:

Derivation of the median:-

Let us Considering the continuous frequency distribution i.e., $x_1 < x_2 < x_3 < \cdots < x_n < x_{n+1}$

class interval	$x_1 - x_2$	$x_2 - x_3$	$\cdots$	$x_k - x_{k+1}$	$\cdots$	$x_n - x_{n+1}$
frequency	f_1	f_2	$f_3 \cdots$	$f_k - f_{k+1}$	$\cdots$	f_n

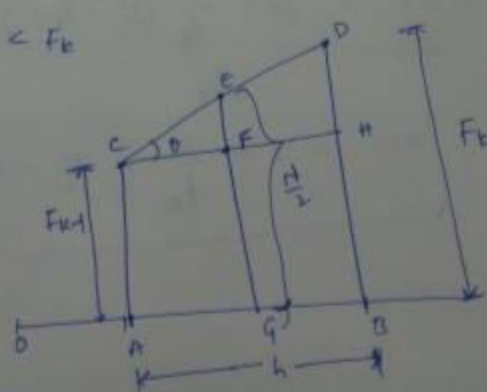
Now, the Cumulative frequency distribution is

class interval	$x_1 - x_2$	$x_2 - x_3$	$\cdots$	$x_k - x_{k+1}$	$\cdots$	$x_n - x_{n+1}$
frequency	F_1	F_2	$\cdots$	$F_k - F_{k+1}$	$\cdots$	F_n

where, $F_i = f_1 + f_2 + \cdots + f_i$

The class $x_k - x_{k+1}$ is the median class

$F_{k-1} < \frac{N}{2} < F_k$



$$\tan\theta = \frac{EF}{FC} = \frac{DH}{HC}$$

$$\Rightarrow \frac{EG - GF}{FC} = \frac{DB - BH}{AB}$$

$$\Rightarrow \frac{\frac{N}{2} - AC}{FC} = \frac{F_k - AC}{AB}$$

$$\Rightarrow \frac{\frac{N}{2} - F_{k-1}}{FC} = \frac{F_k - F_{k-1}}{h}$$

$$\Rightarrow \frac{\frac{N}{2} - F_{k-1}}{FC} = \frac{f_k}{h} \quad (\text{Since } f_k = F_k - F_{k-1})$$

$$\Rightarrow FC = \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right) \quad \text{—(1)}$$

Next, $OG = OA + AG$

$= OA + CF \quad (\because AG = CF)$

$$\Rightarrow \boxed{OG = OA + \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)} \quad (\because \text{eq}(1))$$

(or)

$$\boxed{OG = l + \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)} \quad (\because OA = l,\ F_{k-1} = c)$$

Median for Continuous frequency distribution:

In the case of continuous frequency distribution, the class corresponding to the c.f. just greater than $\frac{N}{2}$ is called the median class and the value of median is obtained by the following formula:

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

Where l is the lower limit of the median class

f is the frequency of the median class

h is the magnitude of the median class

c is the c.f. of the class preceding the median class

$N = \sum f$

Example 1:

Find the median wage of the following distribution:

Wages (in rupees)	No. of workers
2000-3000	3
3000-4000	5
4000-5000	20
5000-6000	10
6000-7000	5

Solution:

Now we have to write the given distribution into continuous frequency distribution.

Wages (in rupees)	No. of workers (f)	$c.f.$
2000-3000	3	3
3000-4000	5	8
4000-5000	20	28
5000-6000	10	38
6000-7000	5	43
	$N = 43$	

$$N = 43$$

$$\frac{N}{2} = 21.5$$

Cumulative frequency is just greater than 21.5 is 28 and the corresponding class is 4000-5000.

Median class is 4000-5000.

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

$$l = 4000,\ h = 1000,\ f = 20,\ N = 43,\ c = 8$$

$Md = 4000 + \frac{1000}{20}\left(\frac{43}{2} - 8\right) = 4675$ rupees.

The median wage is 4675 rupees.

Example 2:

Find the frequency distribution of weight in grams of mangoes of a given variety is given below. Then find the median.

Weight in grams	410-419	420-429	430-439	440-449	450-459	460-469	470-479
No. of mangoes	14	20	42	54	45	18	7

Solution:

Continuous frequency distribution we have to convert the given inclusive class interval series into exclusive class interval series

Weight in grams	No. of mangoes (f)	$c.f.$ (Less than)
409.5-419.5	14	14
419.5-429.5	20	34
429.5-439.5	42	76
439.5-449.5	54	130
449.5-459.5	45	175
459.5-469.5	18	193
469.5-479.5	7	200

	$N = 200$	

$$N = 200$$

$$\frac{N}{2} = 100$$

The c.f. just greater than 100 is 130

So, the corresponding class 439.5-449.5 is the median class.

Now we have to find the median for continuous data

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

$$l = 439.5,\ \text{h} = 10,\ f = 54,\ N = 200,\ c = 76$$

$= 439.5 + \frac{10}{54}\left(\frac{200}{2} - 76\right) = 443.94$ gms.

Example 3:

Find the missing frequency from the following distribution of daily sales of shops, given that the median sale of shops in Rs. 2,400.

Sales (in hundred rupees)	0-10	10-20	20-30	30-40	40-50
No. of shops	5	25	?	8	7

<u>Solution</u>:

Let the frequency be x.

Given that the median sale of shops is 24 hundred.

Sales (in hundred rupees)	No. of shops(f)	$c.f.$
0-10	5	5
10-20	25	30
20-30	x	$30 + x$
30-40	18	$48 + x$
40-50	7	$55 + x$
	$N = 55 + x$	

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

$$l = 20,\ h = 10,\ f = x,\ N = 55 + x,\ c = 30$$

$$24 = 20 + \frac{10}{x}\left(\frac{55 + x}{2} - 30\right)$$

$$x = 25$$

Example 4:

In the frequency distribution of 100 families given below, the number of families corresponding to expenditure groups 20-40 and 60-80 are missing from table. However, the median is known to be 50. Find the missing frequencies.

Expenditure	0-20	20-40	40-60	60-80	80-100
No. of families	14	?	27	?	15

Solution:

Let the missing frequencies for the classes 20-40 and 60-80 be $f_1 and\ f_2$ respectively.

Expenditure (in rupees)	No. of families (f)	$c.f.$
0-20	14	14
20-40	f_1	$14 + f_1$
40-60	27	$41 + f_1$
60-80	f_2	$41 + f_1 + f_2$
80-100	15	$56 + f_1 + f_2$
	$N = 56 + f_1 + f_2$	

The no. of families is 100, therefore

$$N = 100 = 56 + f_1 + f_2$$

$$f_1 + f_2 = 44$$

Given that median is 50, which lies class 40-60.

therefore, 40-60 is the median class.

$$l = 40,\ \mathrm{h} = 20,\ f = 27,\ N = 56 + f_1 + f_2,\ c = 14 + f_1$$

$$Median = l + \frac{\mathrm{h}}{f}\left(\frac{N}{2} - c\right)$$

$$50 = 40 + \frac{20}{27}\left(\frac{100}{2} - (14 + f_1)\right)$$

$$10 = \frac{20}{27}(36 - f_1)$$

$$f_1 = 22.5 \approx 23$$

$$f_2 = 21$$

Try these:

1. Calculate the mean and median from the following distribution

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	4	12	40	41	27	13	9	4

2. A number of particular articles has been classified according to their weights. After drying two weeks the same articles have again been weighted and similarly classified. It is known that the median weight in the first weighing is 20.83 gm, while in the second weighing it was 17.35 gm. Some frequencies a and b in the first weighing and x and y in the second are missing. It is known that $a = \frac{1}{3}x$ and $b = \frac{1}{2}y$. Find out the values of the missing frequencies.

Geometric Mean:

The geometric mean, usually abbreviated as G.M. of a set of n observations is the n^{th} root of their product. Thus, if $X_1, X_2, X_3, \ldots, X_n$ are the n observations then their G.M. is given by

$$G.M = \sqrt[n]{X_1 \times X_2 \times X_3 \times \ \ldots \times X_n} = \left(X_1 \times X_2 \times X_3 \times \ \ldots \times X_n\right)^{\frac{1}{n}} \qquad (1)$$

If $n = 2$ i.e., if we take two observations, then $G.M = \sqrt{X_1 \times X_2}$

If n number of observations, then the n^{th} root is very tedious. In such a case the calculations by making use of the logarithms.

Now take logarithm both sides of eq. (1), we get

$$\log(G.M) = \log\left(X_1 \times X_2 \times X_3 \times \ \ldots \times X_n\right)^{\frac{1}{n}}$$

$$= \frac{1}{n}\log(X_1 \times X_2 \times X_3 \times \ \ldots \times X_n)$$

$$= \frac{1}{n}(\log X_1 + \ log X_2 + \ log X_3 + \ \ldots + log\, X_n)$$

$$\log(G.M) = \frac{1}{n}\sum logX \qquad (2)$$

i.e., the logarithm of the G.M of a set of observations is the arithmetic mean of their logarithms.

Now, taking Antilog on both sides of eq. (2)

$$G.M = Antilog\left(\frac{1}{n}\sum logX\right)$$

In case of frequency distribution $\left(X_i,\ f_i\right)$, $i = 1,2,3, \ldots, n$, where the total no. of observations is $N = \sum f$.

$$\left[\left(X_1 \times X_1 \times X_1 \times \ \ldots \times f_1\, times\right) \times \left(X_2 \times X_2 \times X_2 \times \ \ldots \times f_2\, times\right) \times \ldots \times \left(X_{\mathrm{n}} \times X_{\mathrm{n}} \times X_{\mathrm{n}} \times \ \ldots \times f_n\, times\right)\right]^{\frac{1}{n}} \qquad (3)$$

$$G.M = (X_1 \times X_2 \times X_3 \times \ \ldots \times X_n)^{\frac{1}{N}}$$

Taking logarithm on both sides of eq. (3)

$$\log(G.M) = \frac{1}{n}\left[log\left({X_1}^{f_1} \times {X_2}^{f_2} \times \ldots \times {X_n}^{f_n}\right)\right]$$

$$= \frac{1}{n}\left[log{X_1}^{f_1} + log{X_2}^{f_2} + \cdots + log{X_n}^{f_n}\right]$$

$$= \frac{1}{n}[f_1\, logX_1 + f_2\, logX_2 + \cdots + f_n\, logX_n]$$

$$\log(G.M) = \frac{1}{n}\sum f\, logX$$

$$G.M = Antilog\left[\frac{1}{n}\sum f\, logX\right]$$

Example 1:

Find the geometric mean of 2,4,8,12,16 and 24.

Solution:

X	$logX$
2	0.3010
4	0.6021
8	0.9031
12	1.0792
16	1.2041
24	1.3802
	$\sum logX = 5.4697$

$$\log(G.M) = \frac{1}{n}\sum logX$$

$$= \frac{1}{6}(5.4697) = 0.9116$$

$$= Antilog(0.9116) = 8.158$$

$$G.M = 8.158$$

Example 2:

Find the geometric mean for the following distribution:

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	5	7	15	25	8

Solution:

Marks	Mid-point (X)	No. of students(f)	$logX$	$f\ logX$
0-10	5	5	0.6990	3.4950
10-20	15	7	1.1761	8.2327
20-30	25	15	1.3979	20.9685
30-40	35	25	1.5441	38.6025
40-50	45	8	1.6532	13.2256
		$N = 60$		$\sum f\ logX = 84.5243$

$$G.M = Antilog\left[\frac{1}{N}\sum f\ logX\right]$$

$$= Antilog\left[\frac{1}{60}(84.5243)\right]$$

$$= Antilog(1.4087) = 25.64 \text{ marks}$$

Example 3:

The geometric mean of 10 observations on a certain variable was calculated as 16.2. It was later discovered that one of the observations was wrongly recorded as 12.9; in fact, it was 2.19. Apply approximate correction and calculate the correct geometric mean.

Solution:

Geometric mean of observations is given by

$$G = (X_1 \times X_2 \times X_3 \times \ ... \times X_n)^{\frac{1}{n}} \qquad (1)$$

$$G^n = X_1 \times X_2 \times X_3 \times \ ... \times X_n$$

The product of the numbers is given by $X_1 \times X_2 \times X_3 \times \ ... \times X_n = G^n = (16.20)^{10}$ (2)

If the wrong observation 12.9 is replaced by the correct values 21.9, then the corrected value of the product of 10 numbers is obtained on dividing the expression in (2) by wrong observation and multiplying by the product of correct observation. Thus, the corrected product

$$(X_1 \times X_2 \times X_3 \times \ ... \times X_n) = \frac{(16.20)^{10} \times 21.9}{12.9}$$

The corrected value of G.M, is say G'

$$G' = \left[\frac{(16.20)^{10} \times 21.9}{12.9}\right]^{\frac{1}{10}}$$

$$logG' = log\left[\frac{(16.20)^{10} \times 21.9}{12.9}\right]^{\frac{1}{10}}$$

$$= \frac{1}{10}\left[[log(16.20)^{10} + log(21.9) - log(12.9)]\right]$$

$$= \frac{1}{10}[10 \, log(16.20) + log(21.9) - log(12.9)]$$

$$logG' = \frac{1}{10}[10 \ (1.2095) + 1.3404 - 1.1106]$$

$$logG' = \frac{1}{10}[10 \ (1.2095) + 1.3404 - 1.1106] = 1.2325$$

$$logG' = 1.2325$$

$$G' = Antilog(1.2325) = 17.08$$

TABLE I
LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	·0000	0043	0086	0120	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	·0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	·0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	·1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	·1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	·1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	·2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	·2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	·2553	2577	2601	2625	3648	2672	2695	2618	2742	2765	2	5	7	9	12	14	16	19	21
19	·2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	·3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	·3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	·3424	3444	3464	3483	3502	3522	3541	3562	3579	3598	2	4	6	8	10	12	14	15	17
23	·3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	·3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	·3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	·4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	·4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	·4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	·4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	·4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	·4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	·5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	·5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	·535	5315	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	·5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	·5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	·5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	·5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	·5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	·6021	6031	6042	6053	6065	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	·6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	·6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	·6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	·6435	6444	6454	6465	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	·6532	6542	6551	6561	6571	6380	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	·6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	·6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	·6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	·6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	·6990	6993	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	·7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	·7160	7168	7177	7185	7193	7202	7210	7218	7225	7235	1	2	2	3	4	5	6	7	7
53	·7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	·7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
55	·7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	·7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	·7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	·7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	·7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	·7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	·7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	·7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	·7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	·8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	·8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	·8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	·8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	·8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	·8388	8395	8401	8407	8414	8420	8426	8432	8439	3445	1	1	2	2	3	4	4	5	6
70	·8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	·8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	6
72	·8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	6
73	·8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1		2	2	3	4	5	6
74	·8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	6
75	·8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	6
76	·8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	6
77	·8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	·8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	·9331	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	·9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	·9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	·9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	·9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	·9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	·9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	·9395	9400	9405	9110	9415	9220	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	·9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	·9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	·9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	·9590	9596	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	·9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	·9685	9689	9694	9699	9703	9708	9713	9117	9722	9727	0	1	1	2	2	3	3	4	4
94	·9731	9736	9341	9345	9750	9734	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	·9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	·9823	9827	9832	9836	9841	9845	9850	9854	9359	9863	0	1	1	2	2	3	3	4	4
97	·9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	·9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	·9956	9961	9965	9969	9974	9978	9983	9887	9991	9996	0	1	1	2	2	3	3	3	4

TABLE II

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
·00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
·01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
·02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
·03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
·04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
·05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
·06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
·07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
·08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
·09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
·10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
·11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
·21	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
·13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
·14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
·15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	1	1	1	2		2	2	3	3
·16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
·17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
·18	1514	1717	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
·19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
·20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
·21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
·22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
·23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
·24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
·25	1778	1782	1786	1791	1795	1799	1803	.807	1811	1816	0	1	1	2	2	2	3	3	4
·26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
·27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
·28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
·29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
·30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
·31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
·32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
·33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
·34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
·35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
·36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
·37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
·38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
·39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	5	5
·40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
·41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
·42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
·44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
·45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
·46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
·47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
·48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	5	5	6	6
·49	3090	3097	3105	3112	3119	3126	3133	3141	5148	3155	1	1	2	3	4	4	5	6	6

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3319	3327	3334	3342	3550	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	3	3	4	5	6	6	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	3	4	5	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5165	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
.79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	5353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
.82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7445	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7344	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20

Harmonic Mean:

If X_1, X_2, X_3, ..., X_n is a given n set of observations, then their harmonic mean, abbreviated as H.M.

$$H = \frac{1}{\frac{1}{n}\left[\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \cdots + \frac{1}{X_n}\right]}$$

$$H = \frac{1}{\frac{1}{n}\Sigma\left(\frac{1}{X}\right)}$$

(Or)

Harmonic mean is the value of the variable or the mid-value of the class (in case of grouped or continuous frequency distribution) and f is the corresponding frequency of X.

In case of frequency distribution, we have

$$\frac{1}{H} = \frac{1}{N}\left(\frac{f_1}{X_1} + \frac{f_2}{X_2} + \cdots + \frac{f_n}{X_n}\right)$$

$$= \frac{1}{N}\sum\left(\frac{f}{X}\right)$$

$H = \frac{N}{\sum\left(\frac{1}{X}\right)}$, where, $N = \sum f$

X= mid-value of the variable or mid-value of the class

f= frequency of X

Example 1:

A cyclist pedals from his house to his college at a speed of 10 kmph and back from college to his house at 15 kmph. Find the average speed.

Solution:

Let the distance from house to college be $x\ km$.

In going from house to college, the distance $x\ km$ is covered by $\frac{x}{10}$ hours, while in coming from college to house, the distance covered in $\frac{x}{15}$ hours.

Total distance of $2x\ km$ is covered in $\left(\frac{x}{10} + \frac{x}{15}\right)$ hours.

$$Average\ speed = \frac{total\ distance\ covered}{total\ time\ atken}$$

$$= \frac{2x}{\left(\frac{x}{10} + \frac{x}{15}\right)} = \frac{2x}{\left(\frac{1}{10} + \frac{1}{15}\right)} = 12\ kmph$$

Example 2:

A vehicle when climbing up a gradient, consumes petrol at the rate of 1 liter per 8 km. While coming down it gives 12 km per liter. Find its average consumption for to and fro travel between two places situated at the two ends of a 25 km long gradient. Verify your answer?

Solution:

Since the consumption of petrol is different for upward and downward journeys (at a constant speed of 25 km), the appropriate average consumption for to and fro journey is given by harmonic mean of 8km and 12 km.

$$Average\ consumption = \frac{2}{\left(\frac{1}{8}+\frac{1}{12}\right)} = \frac{48}{5} = 9.6\ km/liter$$

Example 3:

In a certain office, a letter is typed by A in 4 times. The same letter is typed by $B,\ C,\ and\ D$ are 5,6,10 minutes respectively. What is the average time taken in completing one letter? How many letters do you expect to be typed in one day comprising of 8 working hours?

Solution:

The average time taken by each of $A, B,\ C,\ and\ D$ in competing one letter is the harmonic mean of 4,5,6, and 10.

$$\frac{4}{\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{10}} = \frac{4}{\left(\frac{15+12+1+6}{60}\right)} = \frac{240}{43} = 5.5814\ min/letter$$

Expected no. of letters typed by each of $A,\ B,\ C,\ and\ D$ is $\frac{43}{240}$ letters/min.

In a day comprising 8 hours = 8×60 minutes.

Total letters typed all of them $A, B,\ C,\ and\ D = \frac{43}{240} \times 4 \times 8 \times 60 = 344.$

Geometric Mean:

The geometric mean, usually abbreviated as G.M. of a set of n observations is the n^{th} root of their product. Thus, if $X_1,\ X_2,\ X_3,\ \ldots,\ X_n$ are the n observations then their G.M. is given by

$$G.M = \sqrt[n]{X_1 \times X_2 \times X_3 \times \ \ldots \times X_n} = (X_1 \times X_2 \times X_3 \times \ \ldots \times X_n)^{\frac{1}{n}} \qquad (1)$$

If $n = 2$ i.e., if we take two observations, then $G.M = \sqrt{X_1 \times X_2}$

If n number of observations, then the n^{th} root is very tedious. In such a case the calculations by making use of the logarithms.

Now take logarithm both sides of eq. (1), we get

$\log(G.M) = \log(X_1 \times X_2 \times X_3 \times \ldots \times X_n)^{\frac{1}{n}}$

$= \frac{1}{n}\log(X_1 \times X_2 \times X_3 \times \ldots \times X_n)$

$$= \frac{1}{n}(\log X_1 + \log X_2 + \log X_3 + \ldots + \log X_n)$$

$$\log(G.M) = \frac{1}{n}\sum \log X \qquad (2)$$

i.e., the logarithm of the G.M of a set of observations is the arithmetic mean of their logarithms.

Now, taking Antilog on both sides of eq. (2)

$$G.M = Antilog\left(\frac{1}{n}\sum \log X\right)$$

In case of frequency distribution (X_i, f_i), $i = 1,2,3, \ldots, n$, where the total no. of observations is $N = \sum f$.

$[(X_1 \times X_1 \times X_1 \times \ldots \times f_1\, times) \times (X_2 \times X_2 \times X_2 \times \ldots \times f_2\, times) \times \ldots \times$
$(X_n \times X_n \times X_n \times \ldots \times f_n\, times)]^{\frac{1}{N}}$
(3)

$$G.M = (X_1 \times X_2 \times X_3 \times \ldots \times X_n)^{\frac{1}{N}}$$

Taking logarithm on both sides of eq. (3)

$$\log(G.M) = \frac{1}{N}\left[\log\left(X_1^{f_1} \times X_2^{f_2} \times \ldots \times X_n^{f_n}\right)\right]$$

$$= \frac{1}{N}\left[\log X_1^{f_1} + \log X_2^{f_2} + \cdots + \log X_n^{f_n}\right]$$

$$= \frac{1}{N}[f_1 \log X_1 + f_2 \log X_2 + \cdots + f_n \log X_n]$$

$$\log(G.M) = \frac{1}{N}\sum f \log X$$

$$G.M = Antilog\left[\frac{1}{N}\sum f \log X\right]$$

Example 1:

Find the geometric mean of 2,4,8,12,16 and 24.

Solution:

X	$logX$
2	0.3010
4	0.6021
8	0.9031
12	1.0792
16	1.2041
24	1.3802
	$\sum logX = 5.4697$

$$\log(G.M) = \frac{1}{n}\sum logX$$

$$= \frac{1}{6}(5.4697) = 0.9116$$

$$= Antilog(0.9116) = 8.158$$

$$G.M = 8.158$$

Example 2:

Find the geometric mean for the following distribution:

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	5	7	15	25	8

Solution:

Marks	Mid-point (X)	No. of students(f)	$logX$	$f\ logX$
0-10	5	5	0.6990	3.4950
10-20	15	7	1.1761	8.2327
20-30	25	15	1.3979	20.9685
30-40	35	25	1.5441	38.6025
40-50	45	8	1.6532	13.2256
		$N = 60$		$\sum f\ logX = 84.5243$

$$G.M = Antilog\left[\frac{1}{N}\sum f\ logX\right]$$

$$= Antilog\left[\frac{1}{60}(84.5243)\right]$$

$$= Antilog(1.4087) = 25.64 \text{ marks}$$

Example 3:

The geometric mean of 10 observations on a certain variable was calculated as 16.2. It was later discovered that one of the observations was wrongly recorded as 12.9; in fact, it was 2.19. Apply approximate correction and calculate the correct geometric mean.

Solution:

Geometric mean of observations is given by

$$G = (X_1 \times X_2 \times X_3 \times \ ... \times X_n)^{\frac{1}{n}} \qquad (1)$$

$$G^n = X_1 \times X_2 \times X_3 \times \ ... \times X_n$$

The product of the numbers is given by $X_1 \times X_2 \times X_3 \times \ ... \times X_n = G^n = (16.20)^{10}$ (2)

If the wrong observation 12.9 is replaced by the correct values 21.9, then the corrected value of the product of 10 numbers is obtained on dividing the expression in (2) by wrong observation and multiplying by the product of correct observation. Thus, the corrected product

$$(X_1 \times X_2 \times X_3 \times \ \ldots \times X_n) = \frac{(16.20)^{10} \times 21.9}{12.9}$$

The corrected value of G.M, is say G'

$$G' = \left[\frac{(16.20)^{10} \times 21.9}{12.9}\right]^{\frac{1}{10}}$$

$$logG' = log\left[\frac{(16.20)^{10} \times 21.9}{12.9}\right]^{\frac{1}{10}}$$

$$= \frac{1}{10}\left[[log(16.20)^{10} + log(21.9) - log(12.9)]\right]$$

$$= \frac{1}{10}[10\, log(16.20) + log(21.9) - log(12.9)]$$

$$logG' = \frac{1}{10}[10\ (1.2095) + 1.3404 - 1.1106]$$

$$logG' = \frac{1}{10}[10\ (1.2095) + 1.3404 - 1.1106] = 1.2325$$

$$logG' = 1.2325$$

$$G' = Antilog(1.2325) = 17.08$$

TABLE I
LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	·0000	0043	0086	0120	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	·0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	·1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	·1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	·1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	·2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	·2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	·2553	2577	2601	2625	2648	2672	2695	2618	2742	2765	2	5	7	9	12	14	16	19	21
19	·2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	·3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	·3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	·3424	3444	3464	3483	3502	3522	3541	3562	3579	3598	2	4	6	8	10	12	14	15	17
23	·3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	·3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	·3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	·4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	·4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	·4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	·4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	·4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	·4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	·5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	·5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	·535	5315	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	·5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	·5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	·5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	·5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	·5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	·6021	6031	6042	6053	6065	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	·6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	·6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	·6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	·6435	6444	6454	6465	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	·6532	6542	6551	6561	6571	6380	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	·6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	·6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	·6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	·6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	·6990	6993	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	·7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	·7160	7168	7177	7185	7193	7202	7210	7218	7225	7235	1	2	2	3	4	5	6	7	7
53	·7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	·7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7

LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
55	·7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	·7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	·7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	·7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	·7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	·7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	·7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	·7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	·7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	·8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	·8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	·8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	·8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	·8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	·8388	8395	8401	8407	8414	8420	8426	8432	8439	3445	1	1	2	2	3	4	4	5	6
70	·8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	6
72	·8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	6
73	·8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1		2	2	3	4	5	6
74	·8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	6
75	·8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	6
76	·8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	6
77	·8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	·8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	·9331	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	·9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	·9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	·9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	·9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	·9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	·9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	·9395	9400	9405	9110	9415	9220	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	·9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	·9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	·9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	·9590	9596	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	·9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	·9685	9689	9694	9699	9703	9708	9713	9117	9722	9727	0	1	1	2	2	3	3	4	4
94	·9731	9736	9341	9345	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	·9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	·9823	9827	9832	9836	9841	9845	9850	9854	9359	9863	0	1	1	2	2	3	3	4	4
97	·9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	·9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	·9956	9961	9965	9969	9974	9978	9983	9887	9991	9996	0	1	1	2	2	3	3	3	4

TABLE II

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
·00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
·01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
·02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
·03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
·04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
·05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
·06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
·07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
·08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
·09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
·10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
·11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
·12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
·13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
·14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
·15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	1	1	1	2		2	2	3	3
·16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
·17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
·18	1514	1717	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
·19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
·20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
·21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
·22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
·23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
·24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
·25	1778	1782	1786	1791	1795	1799	1803	·807	1811	1816	0	1	1	2	2	2	3	3	4
·26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
·27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
·28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
·29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
·30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
·31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
·32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
·33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
·34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
·35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
·36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
·37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
·38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
·39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	5	5
·40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
·41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
·42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
·43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
·44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
·45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
·46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
·47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
·48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	5	5	6	6
·49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	2	3	4	4	5	6	6

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3319	3327	3334	3342	3550	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	3	3	4	5	6	6	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	3	4	5	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5165	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5889	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
.79	6166	6188	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	5353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
.82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7445	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7344	7261	7278	7295	7311	7328	7345	7362	7339	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20

Harmonic Mean:

If $X_1, X_2, X_3, \ldots, X_n$ is a given n set of observations, then their harmonic mean, abbreviated as H.M.

$$H = \frac{1}{\frac{1}{n}\left[\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \cdots + \frac{1}{X_n}\right]}$$

$$H = \frac{1}{\frac{1}{n}\Sigma\left(\frac{1}{X}\right)}$$

(Or)

Harmonic mean is the value of the variable or the mid-value of the class (in case of grouped or continuous frequency distribution) and f is the corresponding frequency of X.

In case of frequency distribution, we have

$$\frac{1}{H}=\frac{1}{N}\left(\frac{f_1}{X_1}+\frac{f_2}{X_2}+\cdots+\frac{f_n}{X_n}\right)$$

$$=\frac{1}{N}\sum\left(\frac{f}{X}\right)$$

$H=\frac{N}{\sum\left(\frac{1}{X}\right)}$, where, $N=\sum f$

X= mid-value of the variable or mid-value of the class

f= frequency of X

Example 1:

A cyclist pedals from his house to his college at a speed of 10 kmph and back from college to his house at 15 kmph. Find the average speed.

Solution:

Let the distance from house to college be $x\ km$.

In going from house to college, the distance $x\ km$ is covered by $\frac{x}{10}$ hours, while in coming from college to house, the distance covered in $\frac{x}{15}$ hours.

Total distance of $2x\ km$ is covered in $\left(\frac{x}{10}+\frac{x}{15}\right)$ hours.

$$Average\ speed=\frac{total\ distance\ covered}{total\ time\ atken}$$

$$=\frac{2x}{\left(\frac{x}{10}+\frac{x}{15}\right)}=\frac{2x}{\left(\frac{1}{10}+\frac{1}{15}\right)}=12\ kmph$$

Example 2:

A vehicle when climbing up a gradient, consumes petrol at the rate of 1 liter per 8 km. While coming down it gives 12 km per liter. Find its average consumption for to and fro travel between two places situated at the two ends of a 25 km long gradient. Verify your answer?

Solution:

Since the consumption of petrol is different for upward and downward journeys (at a constant speed of 25 km), the appropriate average consumption for to and fro journey is given by harmonic mean of 8km and 12 km.

$$Average\ consumption = \frac{2}{\left(\frac{1}{8}+\frac{1}{12}\right)} = \frac{48}{5} = 9.6\ km/liter$$

Example 3:

In a certain office, a letter is typed by A in 4 times. The same letter is typed by $B,\ C,\ and\ D$ are 5,6,10 minutes respectively. What is the average time taken in completing one letter? How many letters do you expect to be typed in one day comprising of 8 working hours?

Solution:

The average time taken by each of $A, B,\ C,\ and\ D$ in competing one letter is the harmonic mean of 4,5,6, and 10.

$$\frac{4}{\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{10}} = \frac{4}{\left(\frac{15+12+1+6}{60}\right)} = \frac{240}{43} = 5.5814\ min/letter$$

Expected no. of letters typed by each of $A,\ B,\ C,\ and\ D$ is $\frac{43}{240}$ letters/min.

In a day comprising 8 hours $= 8 \times 60$ minutes.

Total letters typed all of them $A, B,\ C,\ and\ D = \frac{43}{240} \times 4 \times 8 \times 60 = 344.$

Measures of variability (dispersion):

Dispersion is the variation of the values which helps one to known as how the variates are closely passed around or widely scattered away from the point of central tendency.

The various measures of dispersion are:

Range
Quartile Deviation
Mean deviation or Absolute Mean deviation
Standard Deviation

Range:

Range is the difference between the greatest (maximum) and the smallest (minimum) observation of the distribution.

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

Example 1:

Find the range of the following distribution

Class interval	0-2	2-4	4-6	6-8	8-10	10-12
Frequency	5	16	13	7	5	4

Solution:

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

$$= 12 - 0 = 12$$

Example 2:

The following table given the age of distribution of a group 50 individuals

Age (in years)	16-20	21-25	26-30	31-36
No. of persons	10	15	17	8

Solution:

Age (in years)	No. of persons
15.5-20.5	10
20.5-25.5	15
25.5-30.5	17

30.5-35.5	8

$$Range = X_{max} - X_{min}$$
$$= 35.5 - 15.5 = 20$$

Quartile deviation:

It is a measure of dispersion based on the upper quartile Q_3 and the lower quartile Q_1.

$$Quartile\ deviation\ (Q.D) = \frac{Q_3 - Q_1}{2}$$

$$where\, Q_i = l + \frac{h}{f}\left(\frac{N*i}{4} - c\right),\ for\ i = 1,2,3$$

$$Q_1 = first\ quartile\ deviation$$
$$Q_2 = second\ quartile\ deviation$$
$$Q_3 = third\ quartile\ deviation$$

Mean Deviation (or) absolute Mean Deviation:

For ungrouped data or raw data:

$$M.D = \frac{1}{n}\sum_i |x_i - \bar{x}|$$

Or frequency distribution:

$$M.D = \frac{1}{N}\sum_i f_i|x_i - \bar{x}|$$

(Or)

If f_i, $i = 1,2,3, \dots, n$ is the frequency distribution corresponding observations x_i , then mean deviation from average (usually mean, median and mode) is given by;

Mean deviation from average $A = \frac{1}{N}\sum_{i=1}^{n} f_i|x_i - A|$, $\sum_i f_i = N$, where $|x_i - A|$ represents modulus or absolute value of the deviation $(x_i - A)$, where negative sign ignored.

Example:

Calculate the Mean deviation from the following data:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	6	5	8	15	7	6	3

Solution:

Mean deviation from average $A = \frac{1}{N}\sum_{i=1}^{n} f_i|x_i - A|$

We have to find the mean deviation from mean is $M.D = \frac{1}{N}\sum f\,|x - \bar{x}|$

So, first we have to find the mean $\bar{x}$.

$$\bar{x} = A + \frac{h}{N}\sum fd$$

Marks	Mid-value (x)	No. of students (f)	$d = \frac{x-35}{h}$	fd
0-10	5	6	-3	-18
10-20	15	5	-2	-10
20-30	25	8	-1	-8
30-40	35	15	0	0
40-50	45	7	1	7
50-60	55	6	2	12
60-70	65	3	3	9
		$N = 50$		$\sum fd = -8$

$$\bar{x} = A + \frac{h}{N}\sum fd$$

$$\bar{x} = 35 + \frac{10}{50}(-8) = 33.4\,marks$$

Marks	Mid-value (x)	No. of students (f)	$d = \frac{x-35}{h}$	fd	$\|x-\bar{x}\| = \|x - 33.4\|$	$f\|x-\bar{x}\|$
0-10	5	6	-3	-18	28.4	170.4
10-20	15	5	-2	-10	18.4	92
20-30	25	8	-1	-8	8.4	67.2

30-40	35	15	0	0	1.6	24
40-50	45	7	1	7	11.6	81.2
50-60	55	6	2	12	21.6	129.6
60-70	65	3	3	9	31.6	94.8
		$N = 50$				$\sum f\,\lvert x-\bar{x}\rvert = 659.2$

$$M.D = \frac{1}{N}\sum f\,|x-\bar{x}| = \frac{659.2}{50} = 13.184$$

Standard Deviation:

$$Variance = \frac{1}{n}\sum (x_i-\bar{x})^2 = \frac{1}{n}\sum {x_i}^2 - \bar{x}^2$$

$$Standard\ Deviation = S.D = \sigma = \sqrt{\frac{1}{n}\sum (x_i-\bar{x})^2}\ \text{(or)}\ \sqrt{\frac{1}{n}\sum {x_i}^2-\bar{x}^2}$$

Coefficient of dispersion $= \frac{Q_3-Q_1}{Q_3+Q_1}$

Coefficient of variation= $C.V = 100\times\frac{\sigma}{\bar{x}}$

Example 1:

Lives of two modes of refrigerators recorded in a received survey are given in the following table. Find the average life of each model. Which model shows more uniformity?

Life (No. of years)	Model A	Model B
0-2	5	2
2-4	16	7
4-6	13	12
8-6	7	19
8-10	5	9
10-12	4	1

Solution:

Life (No. of years)	Mid-value (x)	x^2	Model A (f_A)	Model B (f_B)	$d_A = \frac{x-A}{\text{h}}$	$d_B = \frac{x-B}{\text{h}}$	$d_A f_A$	$d_B f_B$	$f_A\, x^2$	$f_B\, x^2$
0-2	1	1	5	2	-1	-3	-5	-6	5	2
2-4	3	9	16	7	0	-2	0	-14	144	63
4-6	5	25	13	12	1	-1	13	-12	325	300
8-6	7	49	7	19	2	0	14	0	343	931
8-10	9	81	5	9	3	1	15	9	405	729
10-12	11	121	4	1	4	2	16	2	484	121
		$\sum x^2 = 286$	$N = 50$	$N = 50$			$\sum d_A f_A = 53$	$\sum d_B f_B = -21$	1706	2146

$$\bar{x}_A = A + \frac{\text{h}}{N}\left(\sum d_A f_A\right)$$

$$= 3 + \frac{2}{50}(53) = 5.12$$

$$\bar{x}_B = B + \frac{\text{h}}{N}\left(\sum d_B f_B\right)$$

$$= 7 + \frac{2}{50}(-21) = 6.16$$

$$\sigma_A = \sqrt{\frac{1}{N}\sum f_A {x_i}^2 - \bar{x}^2{}_A}$$

$$= \sqrt{\frac{1}{50} \times 286 - (5.12)^2} = 2.8115$$

$$\sigma_B = \sqrt{\frac{1}{N}\sum f_B {x_i}^2 - \bar{x}^2{}_B}$$

$$= \sqrt{\frac{1}{50} \times 2146 - (6.16)^2} = 2.23$$

$$CV_A = 100 \times \frac{\sigma_A}{\bar{x}_A} = 53.95$$

$$CV_B = 100 \times \frac{\sigma_B}{\bar{x}_B} = 36.20$$

$$CV_B < CV_A$$

Therefore, Model B has more uniformity.

Example 2:

Find the range, all three quartiles, quartile deviation, mean deviation, absolute mean deviation and standard deviation for the following distribution

Class Interval	frequency	Cumulative frequency (c. f)
0-2	5	5
2-4	16	21
4-6	13	34
6-8	7	41
8-10	5	46
10-12	4	50

Solution:

Range:

$$Range = X_{max} - X_{min} = 12 - 0 = 12$$

Quartiles:

$$Q_1 = l + \frac{\mathrm{h}}{f}\left(\frac{N}{4} - c\right)$$

$$l = 2,\ \mathrm{h} = 2,\ f = 16,\ c = 5, N = 50$$

$$Q_1 = 2.9375$$

$$Q_2 = l + \frac{\mathrm{h}}{f}\left(\frac{N}{2} - c\right)$$

$$l = 4,\ \mathrm{h} = 2,\ f = 13,\ c = 21, N = 50$$

$$Q_2 = 4.6153$$

$$Q_3 = l + \frac{\mathrm{h}}{f}\left(\frac{3N}{4} - c\right)$$

$$l = 6,\ \mathrm{h} = 2,\ f = 7,\ c = 34,\ N = 50$$

$$Q_3 = 7$$

$$Q.D = \frac{7 - 2.9375}{2} = 2.0312$$

Mean Deviation:

$$M.D = \frac{1}{N}\sum f.\ (x - \bar{x})$$

Absolute Mean Deviation:

$$M.D = \frac{1}{N}\sum_i f_i.|x_i - \bar{x}|$$

Standard Deviation:

$$S.D = \sigma = \sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}$$

Moments, Skewness and Kurtosis:

Moments:

The r^{th} moment of a variable x about any point $x = A$, usually denoted by μ_r' is given by:

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

$\mu_r' = \frac{1}{N}\sum_i f_i d_i^{\,r}, \;\; where\; d_i = x_i - A$ (2)

The r^{th} moment of a variable x about the mean $\bar{x}$, usually denoted by μ_r is given by:

$\mu_r = \frac{1}{N}\sum_i f_i z_i^{\,r}, \;\; where\; z_i = x_i - \bar{x}$ (3)

In particular, $\mu_0 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^0 = 1$, $\mu_1 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^1 = 0$,

$\mu_2 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^2 = \sigma^2$,...

We know that if $d_i = x_i - A$, then

$\bar{x} = A + \frac{1}{N}\sum_i f_i d_i = A + \mu_1'$ (4)

<u>Relation between Moments about Mean in terms of Moments about any point and vice versa</u>:

$\mu_r = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^r = \frac{1}{N}\sum_i f_i(x_i - \mathrm{A} + \mathrm{A} - \bar{x})^r = \frac{1}{N}\sum_i f_i(d_i + \mathrm{A} - \bar{x})^r$, where $d_i = x_i - A$.

Using eq. (4), we get

$$\mu_r = \frac{1}{N}\sum_i f_i(d_i - \mu_1')^r$$

$\mu_r = \frac{1}{N}\sum_i f_i\left(d_i^{\,r} - r_{C_1} d_i^{\,r-1}\mu_1' + r_{C_2} d_i^{\,r-2}\mu_1'^{2} - r_{C_3} d_i^{\,r-3}\mu_1'^{3} + \cdots + (-1)^r \mu_1'^{r}\right)$ (5)

$\mu_r = \mu_r' - r_{C_1}\mu_{r-1}'\mu_1' + r_{C_2}\mu_{r-2}'\mu_1'^{2} - r_{C_3}\mu_{r-3}'\mu_1'^{3} + \cdots + (-1)^r \mu_1'^{r}$ (Using eq. (3))

In particular on putting $r = 2,3$ and 4 in eq. (5) and simplifying, we get

$$\mu_2 = \mu_2' - \mu_1'^{2}$$

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

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

$\mu_r' = \frac{1}{N}\sum_i f_i(x_i - A)^r = \frac{1}{N}\sum_i f_i(x_i - \bar{x} + \bar{x} - A)^r = \frac{1}{N}\sum_i f_i(z_i + \mu_1')^r$,

where $x_i - \bar{x} = z_i$ *and* $\bar{x} = A + \mu_1'$

Thus $\mu_r' = \frac{1}{N}\sum_i f_i\left(z_i^r + r_{C_1} z_i^{r-1}\mu_1' + r_{C_2} z_i^{r-2}\mu_1'^2 + r_{C_3} z_i^{r-3}\mu_1'^3 + \cdots + \mu_1'^r\right)$

$$= \mu_r + r_{C_1}\mu_{r-1}\mu_1' + r_{C_2}\mu_{r-2}\mu_1'^2 + \cdots + \mu_1'^r$$

In particular on putting $r = 2,3,\ 4$ *and noting that* $\mu_1 = 0$, we get

$$\mu_2' = \mu_2 + \mu_1'^2$$

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4$$

These formulae unable to find the moments about any point, once the mean and the moments about mean are known.

Pearson's β and γ Coefficients:

Karl Pearson defined the following four coefficients, based upon the first four moments about mean:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3},\ \gamma_1 = +\sqrt{\beta_1}\ and\ \beta_2 = \frac{\mu_4}{\mu_2^2},\ \gamma_2 = \beta_2 - 3$$

It may be pointed out that these coefficients are pure numbers independent of units of measurement.

Example:

Calculate the first four moments of the following distribution about the mean and hence find β_1 *and* β_2.

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

Solution:

x	f	$d = x - 4$	fd	fd^2	fd^3	fd^4
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
Total	$\boldsymbol{N = 256}$	**0**	$\boldsymbol{\sum fd = 0}$	$\boldsymbol{\sum fd^2 = 512}$	$\boldsymbol{\sum fd^3 = 0}$	$\boldsymbol{\sum fd^4 = 2816}$

Moments about the point $x = 4$ are

$$\mu_1' = \frac{1}{N}\sum fd = 0,\ \mu_2' = \frac{1}{N}\sum fd^2 = \frac{512}{256} = 2,$$

$$\mu_3' = \frac{1}{N}\sum fd^3 = 0,\ \mu_4' = \frac{1}{N}\sum fd^4 = \frac{2816}{256} = 11$$

Moments about mean are

$$\mu_1 = 0,\ \mu_2 = \mu_2' - {\mu_1'}^2 = 2,\ \ \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2{\mu_1'}^3 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'{\mu_1'}^2 - 3{\mu_1'}^4 = 11$$

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} = 0,\quad \beta_2 = \frac{\mu_4}{{\mu_2}^2} = \frac{11}{4} = 2.75$$

Skewness and Kurtosis:

Skewness:

Literally, skewness means lack of symmetry i.e., skewness to have an idea about the shape of the curve which can draw with help of the given data. A distribution is said to be skewed, if

- Mean(M), median M_d and mode M_0 fall at different points i.e., $Mean \neq Median \neq Mode$
- Quartiles are not equidistant from median
- The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Measures of Skewness:

Various measures of skewness (S_k) are:

- $S_k = M - M_d$
- $S_k = M - M_0$
- $S_k = (Q_3 - M_d) - (M_d - Q_1)$

These are the absolute measures of skewness.

1. Prof. Karl Pearson's Coefficient of Skewness:

$S_k = \frac{M - M_0}{\sigma}$, σ is the standard deviation of the distribution.
If mode is ill-defined, then using the empirical relation, $M_0 = 3M_d - 2M$, for a moderately asymmetrical distribution, we get

$$S_k = \frac{3(M - M_d)}{\sigma}$$

$S_k = 0$, if $M = M_0 = M_d$. Hence for a symmetrical distribution all are coincide.

2. Prof. Bowley's Coefficient of Skewness:

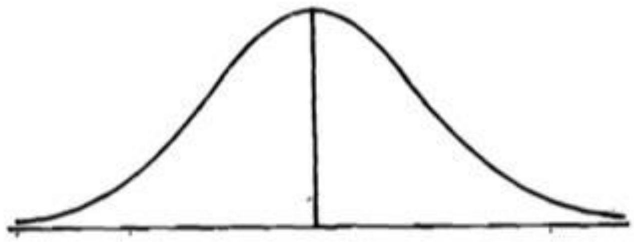
$$S_k = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

$S_k = 0$, if $Q_3 - M_d = M_d - Q_1$. Hence for a symmetrical distribution. Median is equidistant from the upper and lower quartiles.

3. Based upon the moments, coefficient of skewness:

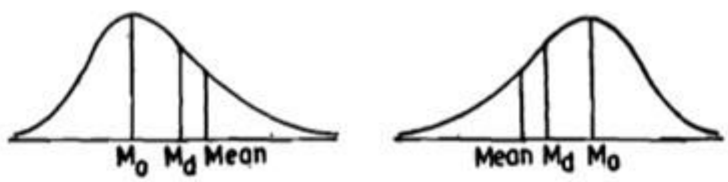
$$S_k = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

$S_k = 0$, if either $\beta_1 = 0$ or $\beta_2 = -3$



$\bar{x}$ (*Mean*) = M_0 = M_d
(Symmetrical Distribution)

e higher values of the variate (the right), i.e., if the curve drawn with the help of
e given data is tretched more to the right than to the left and is negative

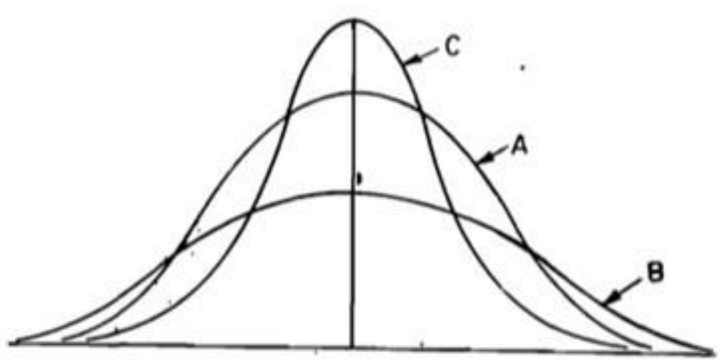


(Positively Skewed Distribution) (Negatively Skewed Distribution)

Kurtosis:

- Prof. Karl Pearson's calls as the 'convexity of the frequency curve' or Kurtosis.
- Kurtosis enables the flatness or peakedness of the frequency curve.
- It is measured by the coefficient β_2 or its derivation is given by

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2}, \gamma_2 = \beta_2 - 3$$



A: Leptokurtic Curve (which is more peaked than the normal curve$\beta_2 > 3$ $i.e.$ $\gamma_2 > 0$)

B: Normal Curve or Mesokurtic Curve (which is neither flat nor peaked $\beta_2 = 3$ $i.e.$ $\gamma_2 = 0$)

C: Platykurtic Curve (which is flatter than the normal curve $\beta_2 < 3$ $i.e.$ $\gamma_2 < 0$)

Example 1:

For a distribution, the mean is 10, variance is 16, γ_1 is +1 and β_2 is 4. Obtain the first four moments about the origin, i.e. zero. Comment upon the nature of distribution.

Solution:

Given that Mean=10, $\mu_2 = 16,$ $then\ S.D = 4, \gamma_1 = \ +1$ and $\beta_2 = 4$

First four moments about the origin is $\mu_1{}', \mu_2{}', \mu_3{}', \mu_4{}'$.

$$\mu_1{}' = first\ moment\ about\ the\ origin = Mean = 10$$

$$\mu_2 = \mu_2{}' - \mu_1{}'^2 \text{ then } \mu_2{}' = \mu_2 + \mu_1{}'^2 = 16 + 100 = 116$$

We have $\gamma_1 = \ +1$ then $\frac{\mu_3}{\mu_2{}^{3/2}} = \frac{\mu_3}{\sigma^3} = 1\ or\ \mu_3 = \sigma^3 = 64$

Therefore , $\mu_3 = \mu_3{}' - 3\mu_2{}'\mu_1{}' + 2\mu_1{}'^3$

$$\mu_3{}' = \mu_3 + 3\mu_2\mu_1{}' + \mu_1{}'^3 = 64 + 3(116)(1) - 2(1000) = 1544$$

Now $\beta_2 = \frac{\mu_4}{\mu_2{}^2} = 4$, then $\mu_4 = 4(256) = 1024$

$$\mu_4 = \mu_4{}' - 4\mu_3{}'\mu_1{}' + 6\mu_2{}'\mu_1{}'^2 - 3\mu_1{}'^4$$

$$\mu_4 = 1024 + 4(1{,}544)(10) - 6(116)(100) + 3(10{,}000) = 23{,}184$$

Comments on nature of the distribution:

Since $\gamma_1 = +1$, the distribution is moderately positively skewed, i.e., if we draw the curve of given distribution, it will have longer tail towards the right. Further since $\beta_2 = 4 > 3$, the distribution is leptokurtic, i.e., it will be slightly more peaked than the normal curve.

Example 2:

For the frequency distribution of scores in mathematics of 50 candidates selected at random from among those appearing at a certain examination, compute the first four moments about the mean of the distribution.

Scores	Frequency
50-60	1
60-70	0
70-80	0
80-90	1
90-100	1
100-110	2
110-120	1
120-130	0
130-140	4
140-150	4
150-160	2
160-170	5
170-180	10
180-190	11
190-200	4
200-210	1
210-220	1
220-230	2

Find also find the correct values of the moments after Sheppard's corrections are applied. Also obtain moment coefficient of skewness and kurtosis and comment on the nature of the distribution.

Solution:

Mid-value (x)	frequency (f)	$d = \frac{x-4}{10}$	fd	fd^2	fd^3	fd^4
55	1	-8	-8	64	-512	4096
65	0	-7	0	0	0	0
75	0	-6	0		0	0

85	1	-5	-5	25	-125	625
95	1	-4	-4	16	-64	256
105	2	-3	-6	18	-54	162
115	1	-2	-2	4	-8	16
125	0	-1	0	0	0	0
135	4	0	0	0	0	0
145	4	1	4	4	4	4
155	2	2	4	8	16	32
165	5	3	15	45	135	405
175	10	4	40	160	640	2560
185	11	5	55	275	1375	6875
195	4	6	24	144	864	5184
205	1	7	7	49	343	2401
215	1	8	8	64	512	4096
225	2	9	18	162	1458	13122
Total	**50**		**150**	**1038**	**4584**	**39834**

The raw moments of variable d (about origin) are computed as

$$\mu_1{}' = \frac{1}{N}\sum fd = \frac{150}{50} = 3,\ \mu_2{}' = \frac{1}{N}\sum fd^2 = \frac{1038}{50} = 20.76,$$

$$\mu_3{}' = \frac{1}{N}\sum fd^3 = \frac{4584}{50} = 91.68,\ \mu_4{}' = \frac{1}{N}\sum fd^4 = \frac{39834}{50} = 796.68$$

The central moments of variable X are then computed as shown below

$$\mu_2 = \left(\mu_2{}' - \mu_1{}'^2\right) \times \mathrm{h}^2 = (20.76 - 9) \times 100 = 1176$$

$$\mu_3 = \left(\mu_3{}' - 3\mu_2{}'\mu_1{}' + 3\mu_1{}'^3\right) \times \mathrm{h}^3 = \left(91.68 - 3 \times 20.76 \times 3 + 2(27)\right) \times 1000 = -41160$$

$$\mu_4 = \left(\mu_4{}' - 4\mu_3{}'\mu_1{}' + 6\mu_2{}'\mu_1{}'^2 - 3\mu_1{}'^4\right) \times \mathrm{h}^4$$
$$= (796.68 - 4 \times 91.68 \times 3 + 6 \times 20.76 \times 9 - 3 \times 91) \times 10000 = 5687091.67$$

Sheppard's corrections for moments:

$$\overline{\mu_2} = \mu_2 - \frac{h^2}{12} = 1176 - \frac{100}{12} = 1167.67$$

$$\overline{\mu_3} = \mu_3 - \frac{h^2}{12} = -41160$$

$$\overline{\mu_4} = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4 = 5745600 - 58800 + 291.67 = 5687091.67$$

Moment coefficient of skewness is given by

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{{\mu_2}^{3/2}} = \frac{-41160}{1176\sqrt{1176}} = -1.02$$

Moment coefficient of kurtosis=$\beta_2 = \frac{\mu_4}{{\mu_2}^2} = \frac{5745600}{(1176)^2} = 4.15$

Comments on nature of the distribution:

Since $\gamma_1 = -1.02$ is negative, the distribution is moderately negatively skewed, i.e., the frequency curve of the given distribution has a longer tail towards the left.

Further, since $\beta_2 = 4.15 > 3$ is positive the distribution is leptokurtic, i.e., the frequency curve is more peaked than the normal curve.

Combined Mean:

Given the sample size n_1 n_2

Sample mean $\overline{x_1}$ $\overline{x_2}$

Combined mean $\bar{x} = \frac{n_1\overline{x_1}+n_2\overline{x_2}}{n_1+n_2}$

Also, sample variance ${\sigma_1}^2$ ${\sigma_2}^2$

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

$$d_1 = x_1 - \bar{x}$$

Example:

A distribution consists of three 25 measurements, it was found that the mean and standard deviation are 36 cm and 12 cm. after these results were calculated, it was noticed that 2 measurements were wrongly recorded as 60 cm and 36 cm, instead of 40 cm and 3 cm. find the corrected values of the mean and standard deviation.

Solution:

Given that $n = 25,\ \bar{x} = 36,\ \sigma = 12$

We know that, $\bar{x} = \frac{\sum x_i}{n}$

$$\sum x_i = n.\,\bar{x} = 25 \times 36 = 900$$

$$\sigma^2 = \frac{1}{n}\sum {x_i}^2 - \bar{x}^2$$

$$\sum {x_i}^2 = n(\sigma^2 + \bar{x}^2) = 25(144 + 296) = 36000$$

Corrected $\sum x_i = 900 - 60 - 36 + 40 + 30 = 874$

Corrected $\sum {x_i}^2 = 36000 - 60^2 - 36^2 + 40^2 + 30^2 = 33604$

Corrected mean= $\frac{corrected\ \sum x_i}{n} = \frac{874}{25} = 34.96$

Corrected variance$(\sigma^2) = \frac{33604}{25} - (34.96)^2 = 121.96$

Corrected standard deviation $(\sigma) = \sqrt{121.96} = 11.04$

Module 2

Probability

Random experiment:

If an experiment is conducted, any number of times, under essentially identical conditions, there is a set of all possible outcomes associated with it. If the result is not certain and is anyone of the several possible outcomes, the experiment is called a random trail or a random experiment. The outcomes are known as elementary events and a set of outcomes is an event. Thus, an elementary event is also an event.

Equally likely events:

Events are said to be equally likely when there is no reason to expect anyone of them rather than anyone of the others.

Example:

When a card is drawn from a pack, any card may be obtained. In this trail, all the 52 elementary events are equally likely.

Exhaustive Events:

All possible events in any trial are known as exhaustive events.

Example:

1. In tossing a coin, there are two exhaustive elementary events, like head and tail.
2. In throwing a die, there are six exhaustive elementary events i.e., getting 1 or 2 or 3 or 4 or 5 or 6.

Mutually exclusive events:

Events are said to be mutually exclusive, if the happening of anyone of the events in a trial excludes the happening of any one of the others i.e., if no two or more of the events can happen simultaneously in the same trial.

Probability:

If a random experiment or a trial results 'n' exhaustive, mutually exclusive and equally likely outcomes, out of which 'm' are favorable to the to the occurrence of event E, then the probability 'p' of occurrence or happening of E, usually denoted by P(E), is given by

$$p = P(E) = \frac{No.\ of\ favourable\ cases}{total\ no.\ of\ exhaustive\ cases} = \frac{m}{n}$$

Note:

1. Since $m \geq 0,\ n > 0\ and\ m \leq n$, we get $P(E) \geq 0$ and $P(E) \leq 1$, then $0 \leq P(E) \leq 1$ or $0 \leq P(\bar{E}) \leq 1$.
2. The non-happening of the event E is called the complementary event of E and is denoted by $\bar{E}\ or\ E^c$. The number of cases favorable to E i.e., non-happening of E is $n - m$. Then the probability q that E will not happen is given by:

   $q = P(\bar{E}) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p.$

   Then, $p + q = 1$

$$q = P(\bar{E}) = 1 - P(E)$$
$$P(E) = 1 - P(\bar{E})$$
$$P(E) + P(\bar{E}) = 1$$

3. If $P(E) = 1,\ E\ is\ called$ certain event and if $P(E) = 0,\ E$ is called impossible event.

Example

1. What is the chance of getting 4 on rolling a die.

Solution:

There are six possible ways in which die can roll.

There is one way of getting 4.

i.e., the required choice of getting 4 is $\frac{1}{6}$.

2. What is the chance of that a leap year selected at random will contain 53 Sundays.

Solution:

Number of days in a leap year is 366.

Number of full weeks, in a leap year 52+ 2 days.

These two days can be any one of the following 7 ways.

Out of these 7 cases the last two are favorable.

Hence the required probability is $\frac{2}{7}$.

Simple Event:

An event in a trial that cannot be further split is called a simple event or an elementary event.

Sample space:

The set of all possible simple events in a trial is called a sample space for the trial. Each element of a sample space is called a sample point.

Example:

Two coins are tossed, then the possible simple events of the trial are HH, HT, TH, TT.

i.e., the sample space is $S = \{\text{HH}, \ \text{HT}, \text{TH}, \text{TT}\}$

Axioms of Probability:

Let E be the random experiment whose sample space is S. If C is a subset of sample space. We define the following three axioms:

1. $P(C) \geq 0$, for every $C \subseteq \mathrm{S}$

2. $P(S) = 1$

3. $P(C_1 \cup C_2 \cup C_3 \cup \dots) = P(C_1) + P(C_2) + P(C_3) + \cdots$, where $C_1, C_2, C_3, \dots$ are subsets of S and they are mutually disjoint. i.e., $C_1 \cap C_2 = \emptyset,\ for\ i \neq j.$

Properties of the probability function:

1. For each $C \subseteq S$, $P(C) = 1 - P(C^*)$, where C^* is the compliment of C in S.

2. The Probability of null set is zero, i.e., $P(\emptyset) = 0$,

3. If C_1 and C_2 are subsets of S such that then $P(C_1) \leq P(C_2)$.

4. $C \subset S,\ 0 \leq P(C) \leq 1$

5. If C_1 and C_2 are subsets of S such then $P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2)$.

Example:

If the sample space is , $S = C_1 \cup C_2$, if $P(C_1) = 0.8,\ P(C_2) = 0.5$ then find $P(C_1 \cap C_2)$.

Solution:

Given that $S = C_1 \cup C_2$,

$$P(C_1) = 0.8,\ P(C_2) = 0.5$$

By the addition law of probability

$$P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2)$$

$$P(C_1 \cap C_2) = P(C_1) + P(C_2) - P(S)$$

$$P(C_1 \cap C_2) = 0.8 + 0.5 - 1$$

$= 1.3 - 1$

$= 0.3.$

Conditional Event:

If C_1, C_2 are events of a sample space S and if C_2 occurs after the occurrence of C_1, then the event of occurrence of C_2 after the event C_1 called conditional event of C_2 given C_1. It is denoted by $\frac{C_2}{C_1}$. Similarly, we define $\frac{C_1}{C_2}$.

Examples:

1. Two coins are tossed. The event of getting two tails given that there is at least one tail is a conditional event.
2. A die is thrown three times. The event of getting the sum of the numbers thrown is 15 when it is known that the first throw was a 5 is a conditional event.

Conditional Probability:

Let S be the sample space of a random experiment. Let $C_1 \subset S$, further let $C_2 \subset C_1$, then the conditional event C_1has already occurred, denoted by $P(C_2/C_1)$ is defined as

$$P(C_2/C_1) = \frac{P(C_2 \cap C_1)}{P(C_1)},\ if\ P(C_1)\ \neq 0$$

Or

$$P(C_2 \cap C_1) = P(C_1)P(C_2/C_1)$$

Note:

1. IfC_1, C_2, C_3 are any three events, then
$$P(C_1 \cap C_2 \cap C_3) = P(C_1)P(C_2/C_1)P(C_3/C_1 \cap C_2)$$
2. For any events $C_1, C_2, C_3, \ldots, C_n$, then
$$P(C_1 \cap C_2 \cap C_3 \cap \ldots \cap C_n) = P(C_1)P(C_2/C_1)P(C_3/C_1 \cap C_2) \ldots P(C_n/C_1 \cap C_2 \cap \ldots \cap C_n)$$

Examples:

1. A box contains 12 items of which 4 are defective the items are drawn at random from the box one after the other. Find the probability that all three are non-defective.

Solution:

There are 8 non-defective items

Total no. of items=12

Let $C_1,\ C_2,\ C_3$ be the events getting non-defective items on first, second and third drawn.

$$P(C_1) = \frac{8}{12}$$

$$P(C_2/C_1) = \frac{7}{11}$$

$$P(C_3/C_1 \cap C_2) = \frac{6}{10}$$

The probability of three are non-defective

$$P(C_1 \cap C_2 \cap C_3) = P(C_1)P(C_2/C_1)P(C_3/C_1 \cap C_2)$$
$$= \frac{8}{12} \times \frac{7}{11} \times \frac{6}{10} = \frac{42}{55}$$

2. A box contains 20 balls of which 5 are red, 15 are white. If 3 balls are selected at random and are drawn in succession without replacement. Find the probability that all three balls selected are red.

Solution:

Total balls=20

Where 5 red and 15 white

$$P(C_1) = \frac{5}{20}$$

$$P(C_2/C_1) = \frac{4}{19}$$

$$P(C_3/C_1 \cap C_2) = \frac{3}{18}$$

$$P(C_1 \cap C_2 \cap C_3) = P(C_1)P(C_2/C_1)P(C_3/C_1 \cap C_2)$$
$$= \frac{5}{20} \times \frac{4}{19} \times \frac{3}{18} = \frac{1}{114}$$

3. The probability that A hits the target is ¼ and the probability B hits is 2/5. What is the probability the target will be hit if A and B each shoot at the target?

Solution:

Given that

$$P(A) = \frac{1}{4}$$

$$P(B) = \frac{2}{5}$$

$$P(A \cup B) = P(A) + \ P(B) - P(A \cap B)$$

$$P(A \cup B) = P(A) + \ P(B) - P(A).P(B)$$

$$P(A \cup B) = \frac{1}{4} + \frac{2}{5} - \frac{1}{4} \times \frac{2}{5} = \frac{11}{20}$$

Bayes theorem:

Let $C_1,\ C_2,\ C_3, \dots,\ C_n$ be a partition of sample space and let C be any event which is a subset of $\bigcup_{i=1}^{n} C_i$ such that

$$\mathrm{P}(C) > 0,\ then\ P(C_\mathrm{i}/\mathrm{C}) = \frac{P(C_i)P(C/C_i)}{\sum_{i=1}^{n} P(C_i)P(C/C_i)}.$$

Proof:

Let S be the sample space.

Let $C_1, C_2, C_3, \dots, C_n$ be a mutually disjoint event.

Let $C \subset \bigcup_{i=1}^{n} C_i$ such that $P(C) > 0$

$$C \subset \bigcup_{i=1}^{n} C_i$$

$$C = C \cap \left(\bigcup_{i=1}^{n} C_i\right)$$

$$C = \bigcup_{i=1}^{n} (C \cap C_i)$$

$P(C) = \sum_{i=1}^{n} P(C \cap C_i)$

$P(C) = \sum_{i=1}^{n} P(C_i)P(C/C_i)$ (1)

Now, $P(C_i/C) = \frac{P(C_i \cap C)}{P(C)}$

$$P(C_i/C) = \frac{P(C_i)P(C/C_i)}{P(C)}$$

$P(C_i/C) = \frac{P(C_i)P(C/C_i)}{\sum_{i=1}^{n} P(C_i)P(C/C_i)}$ (using eq. (1))

Example 1:

A box contains 3 blue, 2 red marbles while another box contains 2 blue, 5 red. A marble drawn at random from one of the boxes terms out to be blue. What is the probability that it come from the first box?

Solution:

Let A_1, A_2 be boxes, then $P(A_1) = \frac{1}{2}$ and $P(A_2) = \frac{1}{2}$

Let C_2 be the event of drawing blue marble $P(C_2/A_1) = \frac{3}{5}$

Let C_2 be the event of drawing blue marble $P(C_2/A_2) = \frac{2}{7}$

We have to require

$$P(A_1/C_2) = \frac{P(A_1)P(C_2/A_1)}{P(A_1)P(C_2/A_1) + P(A_2)P(C_2/A_2)}$$

$$P(A_1/C_2) = \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{1}{2} \times \frac{3}{5} + \frac{1}{2} \times \frac{2}{7}}$$

$$= \frac{\frac{3}{10}}{\frac{3}{10} + \frac{2}{14}} = \frac{21}{31}$$

Example 2:

A bowl one contains 6 red chips and 4 blue chips, 5 of these are selected at random and put in bowl 2 which was originally empty. One chip is drawn at random from bowl 2 relative to the hypothesis that this chip is blue. Find the conditional probability that 2 red chips and 3 blue chips are transferred from bowl 1 to bowl 2.

Solution:

Bowl-1: 6 red and 4 blue

Bowl-2: 2 red and 3 blue

Let E be the event 2 red and 3 blue chips are transferred from bowl 1 to bowl 2.

Then $P(E) = \frac{6_{C_2} \times 4_{C_3}}{10_{C_5}}$

Let B be the event that a blue chip is drawn from bowl 2.

Then $P(B/E) = \frac{3}{5}$

$$P(E/B) = \frac{P(E)P(B/E)}{P(B)}$$

$$P(E/B)=\frac{\frac{6_{C_2}\times 4_{C_3}}{10_{C_5}}\times\frac{3}{5}}{1}=\frac{1}{7}$$

Random Variable:

A random variable is a function that associates a real number with each element in the sample space.

Example:

Suppose that a coin is tossed twice so that the sample space is S={HH, TH, HT, TT}.

Let X represents the number of heads which can come up with each sample point. We can associate a number for X as:

Sample point:	HH	TH	HT	TT
X :	2	1	1	0

There are two types of random variables

(i) Discrete Random Variable

(ii) Continuous Random Variable

(i) Discrete Random Variable: A Random Variable which takes on a finite (or) countably infinite

number of values is called a Discrete Random Variable**.**

(ii) Continuous Random Variable: A Random Variable which takes on non-countable infinite

number of values is called as non- Discrete (or) Continuous Random Variable.

Probability Mass Function (P.M.F):

The set of ordered pairs $\left(x,\ f(x)\right)$ is a probability function of Probability Mass Function of a Discrete Random Variable x.

If for each possible outcome $x,\ f(x)$ must be

(i) $f(x) \geq 0$
(ii) $\sum f(x) = 1$

(iii) $P(X = x) = \ f(x)$

The Probability Mass Function is also denoted by $P_X(x) = P(X = x)$.

Probability Density Function (P.D.F):

The function $f(x)$ is a Probability Density Function for the Continuous Random Variable x defined over the set of real numbers $R,\ if$

(i) $f(x) \geq 0,\ \forall\ x \in R$

(ii) $\int_{-\infty}^{+\infty} f(x)\ dx = 1$

(iii) $P(a < X < b) = \int_a^b f(x)\ dx$

Cumulative Distribution Function $F(x)$:

The Cumulative density distribution function of a discrete random variable X with probability distribution function $f(x)$ as

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

Example:

The Probability function of a random variable X is

X	1	2	3
f(x)	½	1/3	1/6

Find the cumulative distribution of X.

Sol:

The cumulative distribution of X is

X	1	2	3
f(x)	½	5/6	1

Problem:

A shipment of 8 similar computers to a retail outlet contains 3 that are defective. If a school makes a random purchase of 2 computers. Find the probability distribution are the number of defective.

Sol:

Let X be a random variable whose values x at the possible number of defective computers purchased by the school then x maybe 0,1,2

Now, $f(X = x = 0) = P(X = 0)$

$$= \frac{3_{C_0} \times 5_{C_2}}{8_{C_2}} = \frac{10}{28} = \frac{5}{14}$$

$$f(X = x = 1) = P(X = 1)$$

$$= \frac{3_{C_1} \times 5_{C_1}}{8_{C_2}} = \frac{15}{28}$$

$$f(X = x = 2) = P(X = 2)$$

$$= \frac{3_{C_2} \times 5_{C_0}}{8_{C_2}} = \frac{3}{28}$$

The probability distribution function of X is

X	0	1	2
f(x)	10/28	15/28	3/28

Problem:

A random variable X has density function

$$f(x) = \begin{cases} Ce^{-3x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Find

(i) The constant C

(ii) $P(1 < x < 2)$

(iii) $P(X \geq 3)$

(iv) $P(X < 1)$

Sol:

Given that

$$f(x) = \begin{cases} Ce^{-3x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

(i) $\int_0^{\infty} f(x)\, dx = 1$

$$\left\{ since \int_{-\infty}^{+\infty} f(x)\, dx = \int_{-\infty}^{0} f(x)\, dx + \int_{0}^{+\infty} f(x)\, dx \right\}$$

$$\int_0^{\infty} Ce^{-3x}\, dx = 1$$

$$C\left[0 + \frac{1}{3}\right] = 1$$

$$C = 3.$$

(ii) $\int_1^2 3e^{-3x}\, dx$

$$= 3\left[\frac{e^{-3x}}{-3}\right]_1^2$$

$$= \left[\frac{e^{-3x}}{-1}\right]_1^2$$

$$= -[e^{-6} - e^{-3}]$$

$$= e^{-3} - e^{-6}$$

(iii) $\int_3^{\infty} f(x)\, dx = \int_3^{\infty} 3e^{-3x}\, dx$

$$= 3\left[\frac{e^{-3x}}{-3}\right]_3^{\infty}$$

$$= 0 + e^{-9}$$

$$= e^{-9}$$

(iv) $\int_{-\infty}^{1} f(x)\, dx = \int_{-\infty}^{0} f(x)\, dx + \int_0^1 f(x)\, dx$

$$= 0 + \int_0^1 3e^{-3x}\, dx$$

$$= 3\left[\frac{e^{-3x}}{-3}\right]_0^1$$

$$= 1 - e^{-3}$$

Problem:

Let X be a random variable of discrete type having

$$f(x) = \frac{4!}{x!\ (4-x)!}\left(\frac{1}{2}\right)^4,\ x = 0,1,2,3,4.$$

Check whether $f(x)$ is actual a probability density function, if so find $P(A_1)$, where $A_1 = \{0,1\}$.

Solution:

Given the sample space of random variables is $S = \{0,1,2,3,4\}$.

$$P(S) = \sum_{x=0}^{4} f(x)$$

$$= f(0) + f(1) + f(2) + f(3) + f(4)$$

$$= \frac{4!}{4!}\left(\frac{1}{2}\right)^4 + \frac{4!}{1!\,3!}\left(\frac{1}{2}\right)^4 + \frac{4!}{2!\,2!}\left(\frac{1}{2}\right)^4 + \frac{4!}{3!\,1!}\left(\frac{1}{2}\right)^4 + \frac{4!}{4!}\left(\frac{1}{2}\right)^4$$

$$= \left(\frac{1}{2}\right)^4 (1 + 4 + 6 + 4 + 1) = 1$$

$$P(S) = 1$$

We observe that $f(x)$ is probability density function of given random variable of discrete type.

If $A_1 = \{0,1\}$

Then $P(A_1) = f(0) + f(1) = \frac{4!}{4!}\left(\frac{1}{2}\right)^4 + \frac{4!}{1!3!}\left(\frac{1}{2}\right)^4 = \frac{5}{16}$

Problem:

Let x be a random variable of continuous type with probability density function

$$f(x) = \begin{cases} e^{-x}, \; for\; 0 < x < \infty \\ \quad 0\; otherwise \end{cases}$$

Find $(x \in A_1)$, $A_1 : 0 < x < 1$.

.**Solution:**

$$f(x) = \begin{cases} e^{-x}, \; for\; 0 < x < \infty \\ \quad 0\; otherwise \end{cases}$$

We observe that given $f(x)$ is probability density function

$$\int_0^\infty f(x)\, dx = 1$$

$$\int_0^\infty e^{-x}\, dx = 1 \left(since \text{ since } \int_{-\infty}^{+\infty} f(x)\, dx = \int_{-\infty}^{0} f(x)\, dx + \int_0^{+\infty} f(x)\, dx\right)$$

$$1 = 1$$

$$P(A_1) = P(0 < x < 1) = \int_0^1 e^{-x}\, dx = 1 - \frac{1}{e}$$

Problem:

Determine the value of the constant k and the distribution function of the continuous type of random variables x, whose p.d.f. is

$$f(x) = \begin{cases} 0 & for \quad x < 0 \\ kx & for \quad 0 \le x \le 1 \\ k & for \quad 1 \le x \le 2 \\ -kx + 3k & for \quad 2 \le x \le 3 \\ 0 & for \quad x > 3 \end{cases}$$

Solution:

We know that

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{2} f(x)\,dx + \int_{2}^{3} f(x)\,dx + \int_{3}^{\infty} f(x)\,dx = 1$$

$$0 + \int_{0}^{1} kx\,dx + \int_{1}^{2} k\,dx + \int_{2}^{3} (-kx + 3k)\,dx + 0 = 1$$

$$\left[k\frac{x^2}{2}\right]_1^0 + [kx]_1^2 + \left[-k\frac{x^2}{2}\right]_2^3 + [3kx]_2^3 = 1$$

$$\frac{k}{2} + 2k - k + \left(-\frac{9k}{2} + 2k\right) + 9k - 6k = 1$$

$$-4k + 13k - 7k = 1$$

$$-11k + 13k = 1$$

$$k = \frac{1}{2}$$

The cumulative distribution function $F(x)$

$$\int_{0}^{x} kt\,dt = \frac{1}{2}\int_{0}^{x} t\,dt = \frac{1}{2}\frac{x^2}{2} \; for \quad 0 \leq x \leq 1$$

$$\int_{0}^{1} kt\,dt + \int_{1}^{x} k\,dt = \int_{0}^{1} \frac{t}{2}\,dt + \int_{1}^{x} \frac{t}{2}\,dt = \frac{x}{2} - \frac{1}{4}, \; for \quad 1 \leq x \leq 2$$

$$\int_{0}^{1} kt\,dt + \int_{1}^{2} k\,dt + \int_{2}^{x} (-kt + 3k)\,dt$$

$$=\frac{1}{4}+\frac{1}{2}+\left[-\frac{1}{2}\frac{t^2}{2}+\frac{3}{2}t\right]_2^x$$

$$=\frac{1}{4}+\frac{1}{2}-\frac{x^2}{2}+\frac{3}{2}x+1-3$$

$$=-\frac{x^2}{2}+\frac{3}{2}x-\frac{5}{4};\ 2\le x\le 3$$

$$1,\ \ for\ x>3$$

$$F(x)=\begin{cases} 0 & for\quad x<0 \\ \dfrac{x^2}{4} & for\quad 0\le x\le 1 \\ \dfrac{x}{2}-\dfrac{1}{4} & for\quad 1\le x\le 2 \\ -\dfrac{x^2}{2}+\dfrac{3}{2}x-\dfrac{5}{4} & for\quad 2\le x\le 3 \\ 1,\ for\quad x>3 & \end{cases}$$

Mathematical Expectation:

Let X be a random variable with probability distribution $f(x)$, then the mean or mathematical expectation of X is denoted by $E(X)$ and it is denoted by

$E(X)=\sum x\,f(x)$, where $X$is a discrete random variable

$E(X)=\int_{-\infty}^{+\infty} xf(x)dx$,where $X$is a continuous random variable

X be a random variable with pdf $f(x)$ and the mean μ, then the variance of X is

$V(x)=\sigma^2=\mathrm{E}[(X-\mu)^2]=\sum(X-\mu)^2f(x)$, where $X$is a discrete random variable

$V(x)=\sigma^2=\int_{-\infty}^{+\infty}(X-\mu)^2f(x)$, where $X$is a continuous random variable

The positive square root of variance is a standard deviation of X. It is denoted by $\sigma(S.D)$.

Note:

$E(x^2) = \sum x^2\, f(x)$ (discrete)

$E(x^2) = \int_{-\infty}^{+\infty} x^2\, f(x)$ (continuous)

Problem:

If X is a random variable whose pdf is

$f(x) = \begin{cases} \frac{x}{3} & x = 1,2 \\ 0 & otherwise \end{cases}$ find the mathematical expectation of

(i) x (ii) x^2 (iii) $15 - 6x$

Sol:

Given $f(x) = \begin{cases} \frac{x}{3}, & x = 1,2 \\ 0, & otherwise \end{cases}$

We know that $E(X) = \int x\, f(x)\, dx$ (continuous)

(i) $E(x) = \sum_{x=1}^{2} x\, f(x)$ (discrete)

$$= 1\, f(1) + 2\, f\,(2)$$

$$= 1\left(\frac{1}{3}\right) + 2\left(\frac{2}{3}\right) = \frac{5}{3}$$

(ii) $E(x^2) = \sum_{x=1}^{2} x^2\, f(x)$ (discrete)

$$= 1\, f(1) + 4\, f\,(2)$$

$$= 1\left(\frac{1}{3}\right) + 4\left(\frac{2}{3}\right) = \frac{9}{3} = 3$$

(iii) $E(15 - 6x) = \sum_{x=1}^{2} 15 - 6x\, f(x)$

$$= (15 - 6(1))\, f(1) + (15 - 6(2))\, f(2) \qquad \text{(since E(c) = c)}$$

$$= 9\left(\frac{1}{3}\right) + 3\left(\frac{2}{3}\right) = \frac{15}{3} = 5$$

Problem:

If X is a random variable whose pdf is

$f(x) = \begin{cases} 2(1-x), & 0 < x < 1 \\ 0, & otherwise \end{cases}$ find the mathematical expectation of

(i) x (ii) x^2 (iii)$6x - 3x^2$

Sol:

Given $f(x) = \begin{cases} 2(1-x), & 0 < x < 1 \\ 0, & otherwise \end{cases}$

(i) $E(x) = \int_0^1 2x\,(1-x)dx$

$$= \int_0^1 2x\,dx - \int_0^1 2x^2 dx$$

$$= 2\left[\left(\frac{x^2}{2}\right)_0^1 - \left(\frac{x^3}{3}\right)_0^1\right]$$

$$= 2\left[\frac{1}{2} - \frac{1}{3}\right] = \frac{1}{3}$$

(ii) $E(x^2) = \int_0^1 2x^2\,(1-x)dx$

$$= \int_0^1 2x^2\,dx - \int_0^1 2x^3 dx$$

$$= 2\left[\left(\frac{x^3}{3}\right)_0^1 - \left(\frac{x^4}{4}\right)_0^1\right]$$

$$= 2\left[\frac{1}{3} - \frac{1}{4}\right] = \frac{1}{6}$$

(iii) $E(6x) - E(3x^2) = 6\int_0^1 2x\,(1-x)dx - 3\int_0^1 2x^2\,(1-x)dx$

$$= 12\int_0^1 (x - x^2)dx - 6\int_0^1 (x^2 - x^3)dx$$

$$= 12\left[\frac{1}{2} - \frac{1}{3}\right] - 6\left[\frac{1}{12}\right] = \frac{3}{2}$$

Problem:

If X is a random variable whose pdf is

$f(x) = \begin{cases} \frac{1}{\pi}\,\frac{1}{(1+x)^2} & -\infty < x < \infty \\ 0 & otherwise \end{cases}$, then show that $E(x)$ does not exist.

Sol:

$$E(x) = \int_{-\infty}^{+\infty} x\,f(x)\;dx$$

$$= \int_{-\infty}^{+\infty} \frac{x}{\pi}\;\frac{1}{(1+x)^2}\quad dx$$

$$= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{2x}{(1+x)^2}\quad dx$$

$$= \frac{1}{2\pi}(\log(1+x)^2)_{-\infty}^{+\infty}$$

Therefore $E(x)$ does not exist.

Result 1:

If k is a constant $E(k) = k$ itself.

Proof:

We know that

$$E(x) = \int_{-\infty}^{+\infty} x\,f(x)\;dx$$

$$E(k) = \int_{-\infty}^{+\infty} k\, f(k)\ dk$$

$$= k \int_{-\infty}^{+\infty} f(k)\ dk$$

$$= k.1 = k.$$

$$E(k) = k$$

Result 2:

If $a\ and\ b$ are constants and X is a random variable with pdf $f(x)$ then $(ax + b) = aE(x) + b$.

Proof:

We have

$$E(ax + b) = \int_{-\infty}^{+\infty} (ax + b)\, f(x)\ dx$$

$$= \int_{-\infty}^{+\infty} a\, x\, f(x)\ dx + \int_{-\infty}^{+\infty} b\, f(x)\ dx$$

$$= aE(x) + b.1$$

$= aE(x) + b$

Result 3:

The variance of a random variable X is $\sigma^2 = E(x^2) - \mu^2$.

Proof:

Let X be a random variable then variance $\sigma^2 = \mathrm{E}[(X - \mu)^2] = \sum (X - \mu)^2 f(x)$

$$\sigma^2 = \sum (x^2 + \mu^2 - 2x\mu) f(x)$$

$$= \sum x^2 f(x) + \sum \mu^2 f(x) - \sum 2x\mu f(x)$$

$$= E(x^2) + \mu^2 - 2\mu E(x)$$

$$= E(x^2) + \mu^2 - 2\mu^2$$

$$= E(x^2) - \mu^2$$

$$i.e., \ \sigma^2 = E(x^2) - \mu^2$$

Try your self

Find the mean and variance of a random variable whose pdf is

$$f(x) = \begin{cases} \frac{x}{15} & for\ x = 1,2,3,4,5. \\ 0 & otherwise \end{cases}$$

Discrete random variables $X\ and\ Y$:

Joint Probability Distribution function of (X, Y)

The set of triples (X_i, Y_j, P_{ij}), $i = 1,2,3, \ldots, n$; $j = 1,2,3, \ldots, m$ is called the joint probability distribution function of (X, Y) and it can be represented in the form of table as follows:

Y \ X	Y_1	Y_2	Y_3	$\cdots$	Y_m	$P_X(X_i)$
X_1	P_{11}	P_{12}	P_{13}	$\cdots$	P_{1m}	P_{1*}
X_2	P_{21}	P_{22}	P_{23}	$\cdots$	P_{2m}	P_{2*}
X_3	P_{31}	P_{32}	P_{33}		P_{3m}	P_{3*}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_n	P_{n1}	P_{n2}	P_{n3}	$\cdots$	P_{nm}	P_{n*}
$P_Y(Y_j)$	P_{*1}	P_{*2}	P_{*3}	$\cdots$	P_{*m}	1

Marginal Probability Distribution

Let (X, Y) be a two-dimensional discrete random variable. Then the marginal probability function of the random variable X is defined as

$$P(X = x_i) = \sum_{j=1}^{m} P_{ij} = P_{i*}$$

The marginal probability function of the random variable Y is defined as

$$P(Y = y_j) = \sum_{i=1}^{n} P_{ij} = P_{*j}$$

The marginal distribution of X is the coefficient of pairs (x_i, P_{i*}) and of Y is (y_j, P_{*j}).

Conditional Probability Distribution

Let (X, Y) be two-dimensional discrete random variable, then

$$P\left(X = x_i / Y = y_j\right) = \frac{P\left(X = x_i, Y = y_j\right)}{P\left(Y = y_j\right)} = \frac{P_{ij}}{P_{*j}}$$

is called the conditional probability function of X given $Y = y_j$.

and

$$P\left(Y = y_j / X = x_i\right) = \frac{P\left(X = x_i, Y = y_j\right)}{P(X = x_i)} = \frac{P_{ij}}{P_{i*}}$$

is called the conditional probability function of X given $X = x_i$.

Problem 1:

For the bivariate probability distribution of (X,Y) given below, find

$P(X \leq 1), P(Y \leq 3), P(X \leq 1, Y \leq 3\,), P(X \leq 1/Y \leq 3\,)\ and\ P(Y \leq 3/X \leq 1\,)$.

Y \ X	1	2	3	4	5	6
0	0	0	1/32	2/32	2/32	3/32
1	1/16	1/16	1/8	1/8	1/8	1/8
2	1/32	1/32	1/64	1/64	0	2/64

Problem 2:

A random observation on a bivariate population (X,Y) can yield one of the following pairs of values with probabilities noted against them:

For each observation pair	Probability
(1,1); (2,1); (3,3); (4,3)	1/20
(3,1); (4,1); (1,2); (2,2); (3,2); (4,2); (1,3); (2,3)	1/10

Find the probability that $Y = 2$ given that $X = 4$. Also find the probability that $Y = 2$. Examine if the two events $X = 4$ and $Y = 2$ are independent.

Problem 3:

The joint probability distribution of two random variables $X\ and\ Y$ is given by: $P(X = 0, Y = 1) = \frac{1}{3}, P(X = 1, Y = -1) = \frac{1}{3}, and\ P(X = 1, Y = 1) = \frac{1}{3}$.

Find

(i) Marginal distributions of $X\ and\ Y$

(ii) the conditional probability distribution of X given $Y = 1$.

Try Yourself:

(a) A two-dimensional random variable (X, Y) have a bivariate distribution given by:

$P(X = x, Y = y) = \frac{x^2+y}{32}, for\ x = 0,1,2,3\ and\ y = 0,1.$

Find the marginal distribution of $X\ and\ Y$

(b) a two-dimensional random variable (X, Y) have a joint probability mass function: $P(x, y) = \frac{1}{27}(2x + y)$, where $x\ and\ y$ can assume only the integer values 0,1 and 2.

Find the conditional distribution of Y for $X = x$.

Continuous random variables $X\ and\ Y$:

Joint Probability Density function of (X, Y)

Let (X, Y) be a two-dimensional continuous random variable such that

$$P\left(X - \frac{dX}{2} \leq X \leq X + \frac{dX}{2}, \qquad Y - \frac{dY}{2} \leq Y \leq Y + \frac{dY}{2}\right) = \iint f(X, Y)\ dXdY$$

Then $f(X, Y)$ is called the joint density function of (X, Y), if it satisfies the following conditions:

(i) $f(X, Y) \geq 0, for\ all\ (X, Y) \in R$, where R is the range space.

(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(X, Y)\ dXdY = 1$

Moreover, if $(a, b), (c, d) \in R$, then

(iii) $P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(X, Y)\ dXdY = 1$

Marginal Probability Distribution:

When (X, Y) be a two-dimensional continuous random variable, then the marginal density function of the random variable X is defined as

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$$

The marginal density function of the random variable Y is defined as

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$$

Conditional Probability Distribution

Let (X, Y) be two-dimensional continuous random variable, then

$$f(x/y) = \frac{f(x,y)}{f_Y(y)}$$

is called the conditional probability function of X given Y.

and, $f(y/x) = \frac{f(x,y)}{f_X(x)}$

is called the conditional probability function of Y given X.

Problem 1:

Joint distribution of $X\ and\ Y$ is given by $f(x,y) = 4xye^{-(x^2+y^2)}; x \geq 0, y \geq 0$, test whether $X\ and\ Y$ are independent. For the above joint distribution, find the conditional density of $X\ given\ Y = y$.

Solution:

Joint probability distribution function of $X\ and\ Y$ is $f(x,y) = 4xye^{-(x^2+y^2)}; x \geq 0, y \geq 0$

The marginal density of X is given by

$$f_X(x) = \int_0^{\infty} f(x,y)\, dy$$

$$= \int_0^{\infty} 4xye^{-(x^2+y^2)} dy = 4x\, e^{-x^2} \int_0^{\infty} y\, e^{-y^2} dy$$

$$= 4x\, e^{-x^2} \int_0^{\infty} e^{-t} \frac{dt}{2}$$

$$= 2x\, e^{-x^2}; x \geq 0$$

Similarly,

$$f_Y(y) = \int_0^{\infty} f(x,y)\, dx$$

$$= 2y\, e^{-y^2}; y \geq 0$$

Since $f_{XY}(x,y) = f_X(x). f_Y(y), X \text{ and } Y$ are independently distributed.

The conditional distribution of X is given by $Y = y$

$$f_{X/Y}(X = x, Y = y) = \frac{f(x,y)}{f_Y(y)}$$

$$= \frac{4xye^{-(x^2+y^2)}}{2y\, e^{-y^2}} = 2x\, e^{-x^2};\ x \geq 0$$

Problem 2:

Suppose that two-dimensional continuous random variable (X, Y) has joint probability density function given by

$$f(x,y) = \begin{cases} 6x^2y, & 0 < x < 1, 0 < y < 1 \\ 0, & elsewhere \end{cases}$$

(i) Verify that $\int_0^1 \int_0^1 f(x,y) dxdy = 1$

(ii) Find $P\left(0 < X < \frac{3}{4}, \frac{1}{3} < Y < 2\right),\ P(X + Y < 1), P(X > Y) and\ P(X < 1/Y < 2)$

Solution:

(i) $\int_0^1 \int_0^1 f(x,y) dxdy = \int_0^1 \int_0^1 6x^2y\, dxdy = \int_0^1 6x^2 \left|\frac{y^2}{2}\right|_0^1 dx = \int_0^1 3x^2\, dx = 1$

(ii) $P\left(0 < X < \frac{3}{4}, \frac{1}{3} < Y < 2\right) = \int_0^{3/4} \int_{1/3}^{1} 6x^2y\, dxdy + \int_0^{3/4} \int_1^2 0\, dxdy$

$$= \int_0^{3/4} 6x^2 \left(\frac{y^2}{2}\right)_0^1 \, dx = \frac{8}{9}\int_0^{3/4} 3x^2 \, dx = \frac{3}{8}$$

$P(X + Y < 1)$

$$= \int_0^1 \int_0^{1-x} 6x^2 y \, dxdy = \int_0^1 6x^2 \left(\frac{y^2}{2}\right)_0^{1-x} dx = \int_0^1 3x^2(1-x)^2 dx$$

$$= \frac{1}{10}$$

$$P(X > Y) = \int_0^1 \int_0^x 6x^2 y \, dxdy = \int_0^1 3x^2 (y^2)_0^x \, dx = \int_0^1 3x^4 dx = \frac{3}{5}$$

$$P(X < 1/Y < 2) = \frac{P(X < 1 \cap Y < 2)}{P(Y < 2)}$$

$$P(X < 1 \cap Y < 2) = \int_0^1 \int_0^1 6x^2 y \, dxdy + \int_0^1 \int_1^2 0. \, dxdy = 1$$

$$P(Y < 2) = \int_0^1 \int_0^2 f(x, y) \, dxdy = \int_0^1 \int_0^1 6x^2 y \, dxdy + \int_0^1 \int_1^2 0. dxdy = 1$$

$$P(X < 1/Y < 2) = \frac{P(X < 1 \cap Y < 2)}{P(Y < 2)} = 1$$

Try yourself:

1. If X and Y are two random variables having joint density function

$$f(x, y) = \begin{cases} \frac{1}{8}(6 - x - y), & 0 \le x < 2, \quad 2 \le y < 4 \\ 0, & elsewhere \end{cases}$$

Find

(i) $P(X < 1 \cap Y < 3)$
(ii) $P(X + Y < 3)$
(iii) $P(X < 1/Y < 3)$

2. If $f(x_1, x_2) = \begin{cases} 4x_1 x_2, & 0 < x_1 < 1, \ 0 < x_2 < 1 \\ 0 & elsewhere \end{cases}$ is a joint p.d.f. of x_1 and x_2.

   Then find $P\left(0 < x_1 < \frac{1}{2}, \ \frac{1}{4} < x_2 < 1\right)$.

Moments:

The r^{th} moment about the origin of a random variable X denoted by μ_r is $E(X^r)$, i.e.,

$$\mu_0 = E(X^0) = E(1) = 1$$

$$\mu_1 = E(X^1) = E(X) = \mu$$

$$\mu_2 = E(X^2) - \left(E(X)\right)^2 = E(X^2) - \mu^2$$

$$E(X^2) = \sigma^2 + \mu^2$$

Moment Generating function (MGF):

The MGF of the distribution of a random variable completely describes the nature of the distribution.

Let having PDF $f(X)$, then the MGF of the distribution of X is denoted by $M(t)$ and is defined as $M(t) = E(e^{tx})$.

Thus, the MGF $M(t) = \begin{cases} \sum e^{tx} f(x), & \text{if } x \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f(x)dx, & if\ x\ is\ contuous \end{cases}$

We know that $M(t) = E(e^{tx})$

$$M(t) = E\left(1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \cdots\right)$$

$$= E(1) + E(tx) + E\left(\frac{t^2x^2}{2!}\right) + E\left(\frac{t^3x^3}{3!}\right) + \cdots$$

$$= 1 + t.E(x) + \frac{t^2}{2!}.E(x^2) + \frac{t^3}{3!}.E(x^3) + \cdots$$

$$= 1 + t.\mu_1 + \frac{t^2}{2!}.\mu_2 + \cdots$$

$$M(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r{}'$$

The coefficient of $\frac{t^r}{r!}$ is about the origin is $\mu_r{}'$.

If X be a continuous random variable, then MGF is

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

$$M'(t) = \int_{-\infty}^{\infty} x.e^{tx} f(x) dx$$

$$M''(t) = \int_{-\infty}^{\infty} x^2.e^{tx} f(x) dx, \dots$$

Now at $t = 0$

$$M(0) = E(1) = 1$$

$$M'(0) = E(x) = \mu$$

$$M''(0) = E(x^2) = \sigma^2 + \mu^2$$

Mean is $\mu = M'(0)$

Variance is $M''(0) = M''(0) - \left(M'(0)\right)^2$

$$\mu_r{}' = \frac{\partial^r}{\partial t^r} (M(t)) \ ; \quad r = 0,1,2,\dots$$

Example 1:

Obtain the moment generating function of the probability density function is

$f(x) = \begin{cases} xe^{-x}, & 0 < x < \infty \\ 0, & otherwise \end{cases}$.

Solution:

The MGF of the distribution is

$$M(t) = E(e^{tx})$$

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

$$= \int_{0}^{\infty} e^{tx}\, xe^{-x} dx$$

$$= \int_{0}^{\infty} xe^{(t-1)x} dx$$

$$= \int_{0}^{\infty} xe^{-(1-t)x} dx$$

$$= \frac{1}{(1-t)^2}, \qquad for\ t < 1$$

Example 2:

Find the moment generating function of the probability distribution function is

$$f(x) = \begin{cases} \frac{3!}{x!(3-x)!}\left(\frac{1}{2}\right)^3, & x = 0,1,2,3 \\ 0, & otherwise \end{cases}.$$

Solution:

The MGF of the distribution is

$$M(t) = \sum_{x=0}^{3} e^{tx} f(x)$$

$$= \sum_{x=0}^{3} e^{tx} \frac{3!}{x!\,(3-x)!}\left(\frac{1}{2}\right)^3$$

$$= \left(\frac{1}{2}\right)^3 \left[\frac{3!}{3!} + e^t \frac{3!}{1!\,2!} + e^{2t}\frac{3!}{2!\,1!} + e^{3t}\frac{3!}{3!\,0!}\right]$$

$$M(t) = \left(\frac{1}{2}\right)^3 \left[1 + 3e^t + 2e^{2t} + e^{3t}\right]$$

Characteristic function:

The characteristic function is defined as

$$\emptyset_X(t) = E\left(e^{itX}\right) = \begin{cases} \sum_x e^{itX} f(x), & for\ discrete\ probability\ distribution \\ \int e^{itX} f(x)dx, for\ continuous\ probability\ distribution \end{cases}$$

If $F_X(x)$ is the distribution function of a continuous random variable X, then

$$\emptyset_X(t) = \int_{-\infty}^{\infty} e^{itX}\, dF(x)$$

Where, $dF(x) = C\ \frac{1}{(1+x^2)^m}; m > 1, -\infty < x < \infty$

For discrete case, we have

$$\emptyset_X(t) = E\left(e^{itX}\right)$$

$$= E\left(1 + itX + \frac{(it)^2X^2}{2!} + \frac{(it)^3x^3}{3!} + \cdots\right)$$

$$= E(1) + E(itX) + E\left(\frac{(it)^2X^2}{2!}\right) + E\left(\frac{(it)^3X^3}{3!}\right) + \cdots$$

$$= 1 + it.E(X) + \frac{(it)^2}{2!}.E(X^2) + \frac{(it)^3}{3!}.E(X^3) + \cdots$$

$$= 1 + it.\mu_1 + \frac{(it)^2}{2!}.\mu_2 + \cdots$$

$$M(t) = \sum_{r=0}^{\infty} \frac{(it)^r}{r!} \mu_r{}'$$

The coefficient of $\frac{(it)^r}{r!}$ is about the origin is $\mu_r{}'$.

Properties of characteristic function:

1. If the distribution function of a random variable X is symmetrical about zero, i.e., if $f(-x) = f(x)$, then $\emptyset_X(t)$ is real valued function of t.
2. $\emptyset_{cX}(t) = \emptyset_X(ct)$, c being a constant.
3. If X_1 and X_2 are independent random variables, then $\emptyset_{X_1+X_2}(t) = \emptyset_{X_1}(t).\emptyset_{X_2}(t)$
4. $\emptyset_X(-t)$ and $\emptyset_X(t)$ are conjugate functions, i.e., $\emptyset_X(-t) = \overline{\emptyset_X(t)}$.

Covariance:

If X and Y are two random variables, then the Covariance between them is defined as

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

Properties:

1. If X and Y are independent random variables, then $E(XY) = E(X)E(Y)$
2. $Cov(X + a, Y + b) = Cov(X, Y)$

3. $Cov(aX + b, cY + d) = ac\,Cov(X, Y)$

Problem:

Two random variables X and Y have the following joint pdf

$$f(x, y) = \begin{cases} 2 - x - y, & 0 \le x \le 1, \quad 0 \le y \le 1 \\ 0, & otherwise \end{cases}$$

Find the

i. Variance of X
ii. Variance of Y
iii. Covariance of X and Y.

Module-3

Correlation and Regression

Correlation:

In a bivariate distribution we have to find out the if there is any correlation or covariance between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables are deviate in the same direction, that is, if the increase (or decrease) in one result in a corresponding increase (or decrease) in the other, correlation is said to be positive. But, if they are constantly deviate in the opposite directions, that is if increase (or decrease) in one result in corresponding decrease (or increase) in the other, correlation is said to be negative.

Type of Correlation:

(a) Positive and Negative Correlation
(b) Linear and Non-linear Correlation

Positive and Negative Correlation:

If the values of the two variables deviate in the same direction, that is, if the increase of one variable results, on an average, in a corresponding increase in the values of the other variable or if decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, correlation is said to be positive or direct.

Examples:

- Heights and weights
- Price and supply of a commodity
- The family income and expenditure on luxury items, etc.

On the other hand, correlation is said to be negative or inverse if the variables deviate in the opposite direction that is, if the increase or decrease in the values of one variable results, on the average, in a corresponding decrease or increase in the values of the other variable.

Examples:

- Price and demand of a commodity
- Volume and pressure of a perfect gas, etc.

Linear and Non-linear Correlation:

The correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values.

Example:

Let us consider the following data:

x	1	2	3	4	5
y	5	7	9	11	13

Thus, for a c unit change in the variable of x, there is constant change in the corresponding values of y. Mathematically, the above data can be expressed by the relation

$$y = 2x + 3$$

In general, two variables $x \ and \ y$ are said to be linearly related, if there exists a relationship of the form

$$y = a + bx \qquad (1)$$

between them. From eq. (1) of straight line with slope b and which makes an intercept a on the $y - axis$. Hence, if the values of the two variables are plotted as points in the $xy - plane$. Then we get a straight line.

The relationship between two variables is said to be non-linear or curvilinear if corresponding to unit change in one variable, the other variable does not change at a constant rate but at fluctuating rate. In such cases if the data are plotted on the $xy - plane$, we do not get a straight-line curve.

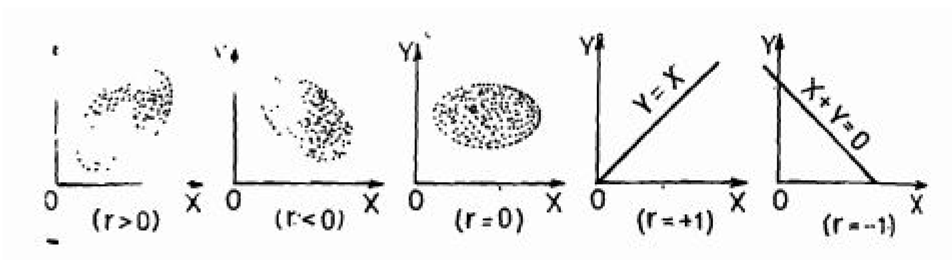
Methods of studying Correlation:

The methods of ascertaining only linear relationship between two variables. The commonly used methods for studying the correlation between two variables are:

(a) Scatter diagram or plot method
(b) Karl Pearson's coefficient of correlation (or Covariance method)
(c) Two-way frequency table (Bivariate correlation method)
(d) Rank correlation method

(a) Scatter diagram method:

If the simplest way of the diagrammatic representation of bivariate data. Thus, for the bivariate distribution $(x_i, y_i); i = 1,2,3 \ldots, n$, if the values of the variables $X \, and \, Y$ are plotted along $x - axis$ and $y - axis$ respectively in the $xy - plane$, diagram of dots so obtained is known as scatter diagram. From the scatter diagram, we can form a fairly good, whether the variables are correlated or not. For example, if the points are very dense, i.e., very close to each other, we should expect a fairly good amount of correlation is expected. This method, however, is not suitable if the number of observations is fairly large.



(b) Karl Pearson's coefficient of Correlation (Covariance method):

As a measure of intensity or degree of linear relationship between two variables, Karl Pearson's, a British Biometrician, developed a formula called Correlation coefficient.

Correlation coefficient between two variables $X\ and\ Y$, usually denoted by $r(X,Y)$ or simply r_{XY} or simply r, is a numerical measure of linear relationship between them and is defined as:

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad (1)$$

Or

It is defined as the ratio of covariance between $X\ and\ Y$ say $Cov(X,Y)$ to the product of the standard deviations $X\ and\ Y$, say

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

If $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$ are n pairs of observations of the variables $X\ and\ Y$ in a bivariate distribution, then

$$Cov(x,y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y});\ \ \sigma_x = \sqrt{\frac{1}{n}\sum(x-\bar{x})^2},\ \sigma_y = \sqrt{\frac{1}{n}\sum(y-\bar{y})^2} \qquad (2)$$

Summation being taken over n pairs of observations.

$$r = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\sum(x-\bar{x})^2 \frac{1}{n}\sum(y-\bar{y})^2}}$$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \qquad (3)$$

Eq. (3) can also be written as

$$r = \frac{\sum dx\ dy}{\sqrt{\sum dx^2\ \sum dy^2}}$$

Where, $dx = x - \bar{x}\ \ and\ \ dy = y - \bar{y}$.

Example 1:

Calculate Karl Pearson’s coefficient of correlation between expenditure on advertising and sales from the data given below

Advertising expenses (thousands Rs.)	39	65	62	90	82	75	25	98	36	78
Sales (lakhs Rs.)	47	53	58	86	62	68	60	91	51	84

Solution:

Let the advertising expenses('000Rs.) be denoted by the variable x and the sales (in lakhs Rs.) be denoted by the variable y.

We have to find the Calculation for correlation coefficient

x	y	$dx = x - \bar{x}$ $= x - 65$	$dy = y - \bar{y}$ $= y - 66$	dx^2	dy^2	$dxdy$
39	47	-26	-19	676	361	494
65	53	0	-13	0	169	0
62	58	-3	-8	9	64	24
90	86	25	20	625	400	500
82	62	17	-4	289	16	-68
75	68	10	2	100	4	20
25	60	-40	-6	1600	36	240
98	91	33	25	1089	625	825
63	51	-29	-15	841	225	435
78	84	13	18	169	324	234
$\sum x$ =650	$\sum y = 660$	$\sum dx = 0$	$\sum dy = 0$	$\sum dx^2 = 5398$	$\sum dy^2 = 2224$	$\sum dxdy = 2704$

$$\sum \bar{x} = \frac{\sum x}{n} = \frac{650}{10} = 65$$

$$\sum \bar{y} = \frac{\sum y}{n} = \frac{660}{10} = 66$$

$$dx = x - \bar{x} = x - 65$$

$$dy = y - \bar{y} = y - 66$$

$$r = \frac{\sum dxdy}{\sqrt{\sum dx^2 \sum dy^2}} = \frac{2704}{\sqrt{5398 \times 2224}} = \frac{2704}{\sqrt{12005152}} = \frac{2704}{3464.8451} = 0.7804$$

Hence, there is a fairly high degree of positive correlation between expenditure on advertising sales. We may, therefore conclude that in general, sales have increased with an increase in the advertising expenditures.

Example 2:

From the following table calculate the coefficient of correlation by Karl Pearson's method

X	6	2	10	4	8
Y	9	11	?	8	7

Arithmetic mean of X and Y series of 6 and 8 respectively.

Solution:

First of all, we shall find the missing value of Y . Let the missing value of Y series be a.Then the mean of $\bar{y}$ is given by:

$$\bar{y} = \frac{\sum y}{n} = \frac{9 + 11 + a + 8 + 7}{5} = \frac{35 + a}{5} = 8\ (given)$$

$$35 + a = 5 \times 8$$

$$a = 40 - 35 = 5$$

Now, we calculate the Correlation coefficient

X	Y	$X - \bar{X} = X - 6$	$Y - \bar{Y} = Y - 8$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})\ (Y - \bar{Y})$
6 2 10 4 8	9 11 5 8 7	0 -4 4 -2 2	1 3 -3 0 -1	0 16 16 4 4	1 9 9 0 1	0 -12 -12 0 1
$\sum X$ =30	$\sum Y$ =40	0	0	$\sum(X - \bar{X})^2$ =40	$\sum(Y - \bar{Y})^2$ =20	$\sum(X - \bar{X})\ (Y - \bar{Y})$ =-26

$$\bar{X} = \frac{\sum x}{5} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum y}{5} = \frac{40}{5} = 8$$

Karl Pearson's correlation coefficient is given by

$$r = \frac{COV(X,Y)}{\sigma_x \sigma_y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}} = \frac{-26}{\sqrt{40 \times 20}} = \frac{-26}{\sqrt{800}} = \frac{-26}{28.2843} = -0.9192$$

$$r \approx -0.92$$

Example 3:

Calculate the coefficient of correlation between $X$and Y series from the following data

	Series	
	X	Y
No. of series observations	15	15
Arithmetic mean	25	18
Standard deviation	3.01	3.03
Sum of squares of deviations from mean	136	138

Summation of product deviation of $X\ and\ Y$ series from their respective arithmetic mean=122.

Solution:

In the usual notations, we are given

$n = 15, \bar{x} = 25, \bar{y} = 18, \sigma_x = 3.01, \sigma_y = 3.03, \sum(x-\bar{x})^2 = 136,\ \sum(y-\bar{y})^2 = 138$ and $\sum(x-\bar{x})(y-\bar{y}) = 122.$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{n\ \sigma_x \sigma_y} = \frac{122}{15 \times 3.01 \times 3.03} = 0.8917$$

Example 4:

A computer while calculating correlation coefficient between two variables $X\ and\ Y$ from 25 pairs of observations obtained the following results:

$$n = 25, \sum X = 125, \sum X^2 = 650, \sum Y = 100, \sum Y^2 = 460, \sum XY = 508$$

It was, however, discovered at the time of checking that two pairs of observations were not correctly copied. They were taken as (6,14) and (8,6) while the correct values were (8,12) and (6,8). Prove that the correct value of the correlation coefficient should be 2/3.

Solution:

Corrected $\sum X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\sum Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\sum X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\sum Y^2 = 460 - 6^2 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\sum XY = 508 - (6 \times 14) - (8 \times 6) + (8 \times 12) + (6 \times 8) = 520$

Corrected value of r is given by

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2] \times [n \sum Y^2 - (\sum Y)^2]}}$$

$$r = \frac{(25 \times 520) - (125 \times 100)}{\sqrt{[25 \times 520 - 125^2] \times [25 \times 436 - 100^2]}} = \frac{2}{3}$$

Properties of correlation coefficient:

1. Pearson coefficient cannot exceed 1 numerically. In other words, it lies between -1 and +1 i.e., $-1 \leq r \leq 1$
2. Correlation coefficient is independent of the change of origin and scale. Mathematically, if X and y are the given variables and they are transformed to the new variables $u \ and \ v$ by the change of origin and scale
   $u = \frac{X-A}{h} \ and \ v = \frac{y-B}{k}, \ h > 0, k > 0.$

   Where, A, B, h and k are constants, $h > 0, k > 0$, then the correlation between $x \ and \ y$ is same the correlation coefficient between $u \ and \ v$ i.e., $r(x, y) = r(u, v)$

$$r_{xy} = r_{uv}$$

$$r_{uv} = \frac{\sum(u-\bar{u})(v-\bar{v})}{\sqrt{\sum(u-\bar{u})^2 \sum(v-\bar{v})^2}}$$

$$r_{uv} = \frac{n\sum uv - (\sum u)(\sum v)}{\sqrt{[n\sum u^2 - (\sum u)^2] \times [n\sum v^2 - (\sum v)^2]}}$$

3. Two independent variables are uncorrelated i.e., $r_{xy} = 0$.
4. $r(aX + b, cY + d) = \frac{a\times c}{|a\times c|}.r(X,Y)$

Example:

Calculate the coefficient of correlation for the ages of husbands and wives

Ages of husbands (years)	23	27	28	29	30	31	33	35	36	39
Ages of wives(years)	18	22	23	24	25	26	28	29	30	32

Solution:

x	y	$u = x - 31$	$v = y - 25$	u^2	v^2	uv
23	18	-8	-7	64	49	56
27	22	-4	-3	16	9	12
28	23	-3	-2	9	4	6
29	24	-2	-1	4	1	2
30	25	-1	0	1	0	0
31	26	0	1	0	1	0
33	28	2	3	4	9	6
35	29	4	4	16	16	16
36	30	5	5	28	25	25
39	32	8	7	64	49	56
$\sum$x =311	$\sum$y =257	$\sum u$ =1	$\sum v$ =7	$\sum$u^2 =203	$\sum$v^2 =163	$\sum$uv =179

Karl Pearson's correlation coefficient between u and v is given by

$$r_{uv} = \frac{n\sum uv - (\sum u)(\sum v)}{\sqrt{[n\sum u^2 - (\sum u)^2] \times [n\sum v^2 - (\sum v)^2]}}$$

$$= \frac{10 \times 179 - 1 \times 7}{\sqrt{[10 \times 203 - (1)^2] \times [10 \times 163 - (7)^2]}}$$

$$= \frac{1790 - 7}{\sqrt{[2030 - 1] \times [1630 - 49]}}$$

$$= \frac{1783}{\sqrt{2029 \times 1581}}$$

$$= \frac{1783}{45.04 \times 39.76}$$

$$= \frac{1783}{1790.79} = 0.9956$$

Since Karl Pearson's correlation coefficient (r) is independent of change of origin, we get

$$r_{xy} = r_{uv} = 0.9956$$

(c) **Rank Correlation method**:

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty, character, morality, etc., which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's coefficient of correlation cannot be used as such. Charles Edward Spearman, a British psychologist, developed a formula in 1904 which consists in obtaining the correlation coefficient between the ranks of n individuals in the two attributes under study.

Suppose we want to find if two characteristics A, say, intelligence and B, say, beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of n individuals in order of merit (ranks) w.r.t. proficiency in the two characteristics. Let the random variables X and Y denote the ranks of the individuals in the characteristics A and B respectively. If we assume that there is no tie, i.e., if no two individuals get the same rank in a characteristic then, obviously, X and Y assume numerical values ranging from 1 to n.

The Pearsonian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characteristics A and B for that group of individuals.

Spearman's rank correlation coefficient, usually denoted by ρ (Rho) is given by the formula

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad (1)$$

Where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.

Computation of rank correlation coefficient:

We shall discuss below the method of computing the Spearman's rank correlation coefficient ρ under the following situations:

I. When actual ranks are given
II. When ranks are not given

Case I: When actual ranks are given:

In this situation the following steps are involved:

i. Compute d, the difference of ranks.
ii. Compute d^2
iii. Obtain the sum $\sum d^2$
iv. Use formula (1) to get the value of ρ.

Example.

The ranks of the same 15 students in two subjects A and B are given below:

the two numbers within the brackets denoting the ranks of the same student in A and B respectively. (1,10), (2,7), (3,2), (4,6), (5,4), (6,8), (7,3), (8,1), (9,11), (10,15), (11,9), (12,5), (13,14), (14,12), (15,13).

Use Spearman's formula to find the rank correlation coefficient.

Solution:

Rank in A (x)	Rank in B (y)	$d = x - y$	d^2
1	10	-9	81
2	7	-5	5
3	2	1	1
4	6	-2	4
5	4	1	1

6	8	-2	4
7	3	4	16
8	1	7	49
9	11	-2	4
10	15	-5	25
11	9	2	4
12	5	7	49
13	14	-1	1
14	12	2	4
15	13	2	4
		$\sum d = 0$	$\sum d^2$ =272

Spearman's rank correlation coefficient ρ (Rho) is given by

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 272}{15(225 - 1)} = 1 - \frac{17}{35} = \frac{18}{35} = 0.51$$

Example:

Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data,

Advertising cost (thousands Rs.)	39	65	62	90	82	75	25	98	36	78
Sales (lakhs Rs.)	47	53	58	86	62	68	60	91	51	84

Solution:

Let X denotes the advertising cost('000Rs.) and Y denotes the Sales (lakhs Rs.).

X	Y	Rank of $X(x)$	Rank of $Y(y)$	$d = x - y$	d^2
39	47	8	10	-2	4
65	53	6	8	-2	4
62	58	7	7	0	0
90	86	2	2	0	0
82	62	3	5	-2	4
75	68	5	4	1	1
25	60	10	6	4	16
98	91	1	1	0	0
63	51	9	9	0	0
78	84	4	1	1	1

				$\sum d$ =0	$\sum d^2$ =30

Here $n = 10$

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 30}{10 \times 99} = 1 - \frac{2}{11} = \frac{9}{11} = 0.82.$$

Example:

Find the rank correlation coefficient from the following data

Ranks in X	1	2	3	4	5	6	7
Ranks in Y	4	3	1	2	6	5	7

Solution:

In this problem ranks are not repeated

x	y	$d_i = x_i - y_i$	${d_i}^2$
1	4	-3	9
2	3	-1	1
3	1	2	4
4	2	2	4
5	6	-1	1
6	5	1	1
7	7	0	0
			$\sum {d_i}^2 = 20$

In this problem ranks are not repeated, so the rank correlation coefficient is

$$r(x,y) = \rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 20}{7(7^2 - 1)} = 0.6429$$

Example:

Calculate the rank correlation coefficient from the following data, which give the ranks of 10 students in Mathematics and Computer Science

Mathematics (x)	1	5	3	4	7	6	10	2	9	8
Computer Science(y)	6	9	1	3	5	4	8	2	10	7

Solution:

x	y	$d_i = x_i - y_i$	${d_i}^2$
1	6	-5	25
5	9	-4	16
3	1	2	4
4	3	1	1
7	5	2	4
6	4	2	4
10	8	2	4
2	2	0	0
9	10	-1	1
8	7	1	1
			$\sum {d_i}^2$ =60

In this problem ranks are not repeated, so the rank correlation coefficient is

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 60}{10(10^2 - 1)} = 0.63636$$

Try yourself:

The ranks of same 16 students in mathematics and physics are as follows. Calculate rank correlation coefficients for proficiency in mathematics and physics

Mathematics (x)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Physics (y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13

Example:

Ten competitors in a beauty contest are ranked by three judges in the following order

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach in common tastes in beauty.

Solution:

Let R_1, R_2 and R_3 denote the ranks given by the first, second and third judges respectively and let ρ_{ij} be the rank correlation coefficient between the ranks given by ith and jth judges $i \neq j = 1,2,3$. Let $d_{ij} = R_i - R_j$, be the difference of ranks of an individual given by the ith and jth judge.

R_1	R_2	R_3	$d_{12} = R_1 - R_2$	$d_{13} = R_1 - R_3$	$d_{23} = R_2 - R_3$	${d_{12}}^2$	${d_{13}}^2$	${d_{23}}^2$
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	1

			$\sum d_{12}$ =0	$\sum d_{13}$ =0	$\sum d_{23}$ =0	$\sum {d_{12}}^2 =$ 200	$\sum {d_{13}}^2 =$ 60	$\sum {d_{23}}^2 =$ 214

We have $n = 10$

Spearman's rank correlation coefficient ρ is given by

$$\rho_{12} = 1 - \frac{6\sum {d_{12}}^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = -\frac{7}{33} = -0.2121$$

$$\rho_{13} = 1 - \frac{6\sum {d_{13}}^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = \frac{7}{11} = 0.6363$$

$$\rho_{23} = 1 - \frac{6\sum {d_{23}}^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -\frac{49}{165} = -0.2970$$

Since ρ_{13} is maximum, the pair of first and third judges has the nearest approach to common tastes in beauty.

Remark, since ρ_{12} and ρ_{23} are negative, the pair of judges (1,2) and (2,3) have opposite (divergent) tastes for beauty.

Case II: When ranks are not given

Spearman's rank correlation formula can also be used even if we are dealing with variables which are measured quantitatively, i.e., when the actual data but not the ranks relating to two variables are given. In such a case we shall have to convert the data into ranks. The highest (smallest) observation is given the rank 1. The next highest (next lowest) observation is given rank 2 and so on. It is immaterial in which way (descending or ascending) the ranks are assigned. However, the same approach should be followed for the all the variables under consideration.

Repeated ranks:

In case of attributes if there is a tie i.e., if any two or more individuals are placed together in any classification with respect to an attribute or if in case of variable data there is more than one item with the same value in either or both the series, then Spearman's formula for calculating the rank correlation coefficient breaks down, since in this case the variables X [the ranks of individuals in characteristic A (1st series)] and Y [the ranks of individuals in characteristic B (2nd series) do not take the values from 1 to n and consequently $\bar{x} \neq \bar{y}$, while Spearman's formula proving we had assumed that $\bar{x} = \bar{y}$.

In this case, common ranks are assigned to the repeated items. These common ranks are the arithmetic mean of the ranks which these items should have got if they are different from each other and the next item will get the rank next to the rank used computing the common rank.

For example, suppose an item is repeated at rank 4. The common rank common rank to be assigned to each item is (4+5)/2 i.e., 4.5 which is the average of 4 and 5, the ranks which these observations would have assigned if they were different. The next item will be assigned the rank 6. If an item is repeated thrice at rank 7, then the common rank to be assigned to each value will e (7+8+9)/3 i.e., 8 which is arithmetic mean of 7, 8 and 9. The ranks these observations would have got if they were different from each other. The next rank to be assigned will be 10.

In the Spearman's formula add the factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m is the number of times is repeated. This correction factor is to be added for each repeated value in both the series.

Problem:

A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so the students in a pair have approximately equal scores on an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below:

Pair	1	2	3	4	5	6	7	8	9	10	11
A	24	29	19	14	30	19	27	30	20	28	11
B	37	35	16	26	23	27	19	20	16	11	21

Find the rank correlation coefficient.

Solution:

In the X-series, we seen that the value 30 occurs twice. The common rank assigned to each of these values is 1.5, the arithmetic mean of 1 and 2, the ranks which these which observations would have taken if they were different. The next value 29 gets the next i.e. rank 3. Again, the value 19 occurs twice. The common rank assigned to it as 8.5, the arithmetic mean of 8 and 9 and the next value, 14 gets the rank 10. Similarly, in the y-series the value 16 occurs twice and the common rank assigned to each is 9.5, the arithmetic mean of 9 and 10, the next value, 11 gets the rank 11.

X	Y	Rank of X (x)	Rank of Y (y)	d=x-y	d^2
24	37	6	1	5	25
29	35	3	2	1	1
19	16	8.5	9.5	-1	1
14	26	10	4	6	36
30	23	1.5	5	-3.5	12.25
19	27	8.5	3	5.5	30.25
27	19	5	8	-3	9
30	20	1.5	7	-5.5	30.25
20	16	7	9.5	-2.5	6.25
28	11	4	11	-7	49
11	21	11	6	5	25
				$\sum d = 0$	$\sum d^2 = 225$

Hence, we see that in the X-series the items 19 and 30 are repeated, each occurring twice and, in the Y-series in the item 16 is repeated. Thus, in each of the three cases $m = 2$. Hence on applying the correction factor $\frac{m(m^2-1)}{12}$ for each repeated item, we get

$$\rho = 1 - \frac{6\left[\sum d^2 + 2\left(\frac{4-1}{12}\right) + 2\left(\frac{4-1}{12}\right) + 2\left(\frac{4-1}{12}\right)\right]}{11(121-1)}, \quad \text{here n=11}$$

$$\rho = 1 - \frac{6 \times 226.5}{11 \times 120} = 1 - 1.0225 = -0.0225$$

Problem:

A sample of 12 fathers and their eldest sons have the following data about their heights in inches.

Fathers (x)	65	63	67	64	68	63	70	66	68	67	69	71
Sons (y)	68	66	68	65	69	66	68	65	71	67	68	70

Calculate the rank correlation coefficient.

Solution:

Fathers (x)	Sons (y)	Rank of x	Rank of y	$d = x - y$	d^2
65	68	9	5.5	3.5	12.25
63	66	11	9.5	1.5	2.25
67	68	6.5	5.5	1	1
64	65	10	11.5	-1.5	2.25
68	69	4.5	3	1.5	2.25
62	66	12	9.5	2.5	6.25
70	68	2	5.5	-3.5	12.25
66	65	8	11.5	-3.5	12.25
68	71	4.5	1	3.5	12.25
67	67	6.5	8	-1.5	2.25
69	68	3	5.5	2.5	6.25
71	70	1	2	-1	1
				$\sum d$ =0	$\sum d^2$ =72.5

Correlation factors

In x, 68 is repeated twice, then $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

In x, 67 is repeated twice, then $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

In y, 67 is repeated 4 times, then $\frac{m(m^2-1)}{12} = \frac{4(4^2-1)}{12} = 5$

In y, 66 is repeated twice, then $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

In y, 65 is repeated twice, then $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

Rank correlation is

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{1}{2} + \frac{1}{2} + 5 + \frac{1}{2} + \frac{1}{2}\right]}{12(144-1)} = 0.722$$

Linear Regression:

If the variables in bivariate distribution are related, will find that the points in the scatter diagram will cluster round some curve called the ‘’curve of regression’’. If the curve is a straight line, it is called the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear.

The lines of regression are the line which gives to be best estimate to the value of one variable for any specific value of the other variable. Thus, the line of regression is the line of ‘best fit’ and is obtained by the principle of least squares.

Let us suppose that the in the bivariate distribution $(x_i, y_i); i = 1,2,3, \ldots, n;$ y is dependent variable and x is independent variable. Let the line of regression is the line of y on x be

$$y = a + bx \tag{1}$$

Eq. (1) represents the family of straight lines for different values of the arbitrary constants $'a' and\ 'b'$. The problem is to determine the $'a' and\ 'b$ so that the line Eq. (1) is the line of best fit.

According to the principle of the principle of least squares, we have to determine $'a' and\ 'b'$.

$$E = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2$$

Is minimum. From the principle of maxima and minima, the partial derivatives of E, with respect to $'a' and\ 'b'$ should vanish separately, i.e.,

$$\frac{\partial E}{\partial a} = 0 = -2\sum_{i=1}^{n}\left(y_i - (a + bx_i)\right)$$

$$\textstyle\sum_{i=1}^{n} y_i = an + \sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i \tag{2}$$

$$\frac{\partial E}{\partial b} = 0 = -2\sum_{i=1}^{n} x_i\left(x_i - (a + bx_i)\right)$$

$$\textstyle\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} {x_i}^2 \tag{3}$$

Dividing on both sides to the Eq. (2) by n, we get

$$\bar{y} = a + b\bar{x} \tag{4}$$

Now, the line of regression of $Y\ on\ X$ passes through the point $(\bar{x}, \bar{y}\ )$.

$$\mu_{11} = Cov(x, y) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

$$\tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\bar{y} \tag{5}$$

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

$$\tfrac{1}{n}\textstyle\sum_{i=1}^{n} {x_i}^2 = {\sigma_x}^2 + \bar{x}^2 \tag{6}$$

Dividing Eq. (3) by n and using Eqs. (5) and (6), we get

$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b({\sigma_x}^2 + \bar{x}^2) \tag{7}$$

Eq. (7)- Eq. (4)$\times\ \bar{x}$, we get

$$\mu_{11} = b{\sigma_x}^2$$

$$b = \frac{\mu_{11}}{{\sigma_x}^2}$$

Since 'b' is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y}\,)$ its equation is

$$Y - \bar{y} = b(x - \bar{x}) = \frac{\mu_{11}}{{\sigma_x}^2}(X - \bar{x})$$

$$Y - \bar{y} = r\,\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

Starting the equation $X = A + BY$ and proceeding similarly, we get

$$X - \bar{x} = r\,\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

Problem:

From the following data, obtain the two regression equations

Sales	91	97	108	121	67	124	51	73	111	57
Purchases	71	75	69	97	70	91	39	61	80	47

Solution:

Let us denote the sales by the variable x and y the purchases by the variable y

x	y	$dx = x - 90$	$dy = y - 70$	dx^2	dy^2	$dx\,dy$
91	71	1	1	1	1	1
97	75	7	5	49	25	35
108	69	18	-1	324	1	-18
121	97	31	27	961	729	837
67	70	-23	0	529	0	0
124	91	34	21	1156	441	714
51	39	-39	-31	1521	961	1209
73	61	-17	-9	289	81	153
111	80	21	10	441	100	210

57	47	-33	-23	1089	529	759
$\sum x = 900$	$\sum y = 700$	$\sum dx = 0$	$\sum dy = 0$	$\sum dx^2 = 6360$	$\sum dy^2 = 2868$	$\sum dx\, dy = 3900$

We have, $\bar{x} = \frac{\sum x}{n} = \frac{900}{10} = 90$

$$\bar{y} = \frac{\sum y}{n} = \frac{700}{10} = 70$$

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum dxdy}{\sum dx^2} = \frac{3900}{6360} = 0.6132$$

$$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2} = \frac{\sum dxdy}{\sum dy^2} = \frac{3900}{2868} = 1.361$$

Equation of regression of $y\ on\ x$ is

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

$$y - 70 = 0.6132(x - 90)$$

$$y - 70 = 0.6132\ x - 0.613 \times 90$$

$$= 0.6132\ x - 55.188$$

$$y = 0.6132\ x - 55.188 + 70$$

$$y = 0.6132\ x + 14.812$$

Equation of regression of $x\ on\ y$ is

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

$$x - 90 = 1.361(y - 70)$$

$$x - 90 = 1.361\ \ y - 1.361 \times 70$$

$$= 1.361\ y - 95.27$$

$$x = 1.361\ y - 95.27 + 90$$

$$x = 1.361\ y - 5.27$$

$$r^2 = b_{yx} \,.\, b_{xy}$$

$$r^2 = 0.6132 \times 1.361 = 0.8346$$

$$r = \mp 0.9135$$

But since, both the regression coefficients are positive, r must be positive.

$$r = 0.9135$$

Problem:

From the data given below find

(a) Two regression coefficients
(b) The two regression equations
(c) The coefficient of correlation between the marks in Economics and Statistics
(d) The most likely marks in Statistics when marks in Economics are 30.

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

Solution:

x	y	$dx = x - 32$	$dy = y - 38$	dx^2	dy^2	$dx\,dy$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
$\sum x = 320$	$\sum y = 380$	$\sum dx = 0$	$\sum dy = 0$	$\sum dx^2 = 140$	$\sum dy^2 = 398$	$\sum dx\,dy = -93$

$$\bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32$$

$$\bar{y} = \frac{\sum y}{n} = \frac{380}{10} = 38$$

(a) **Regression coefficients**:

Coefficient of regression $y\ on\ x$ is given by

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum dxdy}{\sum dx^2} = \frac{-93}{140} = -0.6643$$

$$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2} = \frac{\sum dxdy}{\sum dy^2} = \frac{-93}{398} = -0.2337$$

(b) Equations of the line of regression of $x \; on \; y$ is

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

$$x - 32 = -0.2337(y - 38)$$

$$x - 32 = -0.2337\, y + 38 \times 0.2337$$

$$= -0.2337\, y + 8.8806$$

$$x = -0.2337\, y + 8.8806 + 32$$

$$x = -0.2337\, y + 40.8806$$

Equation of line of regression of $y \; on \; x$ is

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

$$y - 38 = -0.6643(x - 32)$$

$$y - 38 = -0.6643\, x + 0.6643 \times 32$$

$$= -0.6643\, x + 0.6643 \times 32 + 38$$

$$y = -0.6643\, x + 59.2576$$

(c) Correlation coefficient:

$$r^2 = b_{yx} \,.\, b_{xy}$$

$$r^2 = (-0.6643)(-0.2337) = 0.1552$$

$$r = \mp 0.394$$

Since the both regression coefficients are negative. Hence the discarding plus sign, we get

$$r = -0.394$$

(d) In order to estimate the most likely marks in Statistics (y) when marks in Economics (x) are 30, we use the line of regression of $y \; on \; x$.

The equation is

$$y - 38 = -0.6643\;(30) + 59.2576$$

$$y = 39.3286$$

Hence the most likely marks in Statistics when in Economics are 30, are 39.3286$\approx$ 39.

Problem:

The following is an estimated supply regression for sugar:

$y = 0.025 + 1.5\, x$, where y is supply in kilos and x is price in rupees per kilo.

(a) Interpret the coefficient of variable x
(b) Predict the supply when supply when price is Rs. 20 per kilo
(c) Given that $r(x, y) = 1$, interpret the implied relationship between price and quality supplied.

Solution:

The regression equation of y (supply in kgs) on x (price in rupees per kg) is given to be

$$y = 0.025 + 1.5\, x = a + bx \text{ (say)} \qquad (1)$$

(a) The coefficients of variation x
$b = 1.5$ is the coefficient of regression of $y\ on\ x$. It reflects the unit change in the value of y, for a unit change in the corresponding value of x. This means that if the price of sugar goes up by Re. 1 per kg, the estimated supply of sugar goes up by 1.5 kg.

(b) From eq. (1), the estimated supply of sugar when its price is Rs. 20 per kg is given by
$y = 0.025 + 1.5 \times 20 = 30.025$ kg
(c) $r(x, y) = 1$
The relationship between that $x\ and\ y$ is exactly linear. i.e., all the observed values (x, y) lies on straight line.

Problem:

Given that the regression equations of $y\ on\ x$ and of $x\ on\ y$ are respectively $y = x$ and

$4x - y = 3$, and that the second moment of x about the origin is 2, find

(a) The correlation coefficient between $x\ and\ y$
(b) The standard deviation of y

Solution:

Regression equation of $y\ on\ x$ is $y = x$

$$b_{yx} = 1$$

Regression equation of $x \; on \; y$ is $4x - y = 3$

$$x = \frac{1}{4}y + \frac{3}{4}$$

$$b_{xy} = \frac{1}{4}$$

(a) The correlation coefficient between $x \; and \; y$ is

$$r^2 = b_{yx} \,.\, b_{xy}$$

$$r^2 = 1 \times \frac{1}{4} = \frac{1}{4}$$

$$r = \mp 0.5$$

Since the both the regression coefficients are positive $r = 0.5$.

(b) We are given that the second moment of x about origin is 2. i.e., $\frac{\sum x^2}{n} = 2$

Since $(\bar{x}, \bar{y})$ is the point of intersection of the two lines of regression

Solving $y = x$ and $4x - y = 3$, then $x = 1 = y$

$$\bar{x} = 1 \; and \; \bar{y} = 1$$

$$\sigma_x{}^2 = \frac{\sum x^2}{n} - \bar{x}^2 = 2 - 1 = 1$$

Also, $b_{yx} = r \, \frac{\sigma_y}{\sigma_x}$

$$1 = \frac{1}{2}\left(\frac{\sigma_y}{1}\right)$$

$$\sigma_y = 2$$

Coefficient of Determination:

Coefficient of correlation between two variable series is a measure of linear relationship between them and indicates the amount of variation of one variable which is associated with or accounted for by another variable. A more useful and readily comprehensible measure for this purpose is the coefficient of determination which gives the percentage variation in the dependent variable that is accounted for by the independent variable.

In other words, the coefficient of determination gives the ratio of the explained variance to the total variance. The coefficient is given by the square of the correlation coefficient i.e.,

$$r^2 = \frac{explained\ variance}{total\ variance}.$$

Ex:

If the value of $r = 0.8$, we cannot conclude that 80% of the variation in the relative series (dependent variable) is due to the variation in the subject series (independent variable). But the coefficient of determination in this case $r^2 = 0.64$ which implies that only 64% of the variation in the relative series has been explained by the subject series and the remaining 36% of the variation is due to other factors.

Coefficient of Partial correlation:

Sometimes the correlation between **two variables $X_1 and\ X_2$ may be partly due to the correlation of third variable X_3 with both $X_1\ and\ X_2$**. In such a situation, one may want to know what the correlation between $X_1 and\ X_2$would be if the effect of X_3 an each of $X_1 and\ X_2$ were eliminated. **This correlation is called partial correlation and the correlation coefficient between $X_1 and\ X_2$ after the linear effect of X_3 on each of them has been eliminated is called the partial coefficient.**

The partial correlation coefficient between $X_1 and\ X_2$, usually denoted by $r_{12.3}$ is given by

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}{}^2)(1 - r_{23}{}^2)}}$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - r_{12}{}^2)(1 - r_{32}{}^2)}}$$

and

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{(1 - r_{21}{}^2)(1 - r_{31}{}^2)}}$$

Multiple correlation in terms of total and partial correlations:

$$1 - R_{1.23}{}^2 = 1 - \frac{r_{12}{}^2 + r_{13}{}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}{}^2}$$

$$= \frac{1 - r_{23}{}^2 - r_{12}{}^2 - r_{13}{}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}{}^2}$$

Note:

$$1 - R_{1.23}{}^2 = \frac{\omega}{\omega_{11}}$$

Where, $\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}{}^2 - r_{13}{}^2 - r_{23}{}^2 + 2r_{12}r_{13}r_{23}$

and $\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}{}^2$

Problem:

From the data relating to the yield of dry bark (X_1), height (X_2) and girth (X_3) for 18 cinchona plants, the following correlation coefficients were obtained:

$r_{12} = 0.77, r_{13} = 0.72$ *and* $r_{23} = 0.52$. find the partial correlation coefficients $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$.

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}{}^2)(1 - r_{23}{}^2)}}$$

$$= \frac{0.77 - 0.72 \times 0.52}{\sqrt{(1 - 0.72^2)(1 - 0.52^2)}} = 0.62$$

$$R_{1.23}{}^2 = \frac{r_{12}{}^2 + r_{13}{}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}{}^2}$$

$$= \frac{0.77^2 + 0.72^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1 - 0.52^2} = 0.7334$$

$R_{1.23} = +0.8564$ (since multiple correlation is non-negative).

Problem:

In a trivariate distribution $\sigma_1 = 2, \sigma_2 = \sigma_3 = 3, r_{12} = 0.7, r_{23} = r_{31} = 0.5.$

Find (i) $r_{23.1}$ (ii) $R_{1.23}$ (iii) $b_{12.3}$, $b_{13.2}$ and (iv) $\sigma_{1.23}$.

Solution:

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{(1 - r_{21}{}^2)(1 - r_{31}{}^2)}}$$

$$= \frac{0.5 - 0.7 \times 0.5}{\sqrt{(1 - 0.7^2)(1 - 0.5^2)}} = 0.2425$$

(ii)

$$R_{1.23}{}^2 = \frac{r_{12}{}^2 + r_{13}{}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}{}^2}$$

$$R_{1.23}{}^2 = \frac{0.7^2 + 0.5^2 - 2 \times 0.7 \times 0.5 \times 0.5}{1 - 0.5^2} = 0.52$$

$$R_{1.23} = +0.7211$$

(iii)

$$b_{12.3} = r_{12.3} \left(\frac{\sigma_{1.3}}{\sigma_{2.3}}\right) \text{ and } b_{13.2} = r_{13.2} \left(\frac{\sigma_{1.2}}{\sigma_{3.2}}\right) \qquad (1)$$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}{}^2)(1 - r_{23}{}^2)}} = 0.6$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - r_{12}{}^2)(1 - r_{32}{}^2)}} = 0.2425$$

$$\sigma_{1.3=}\, \sigma_1 \sqrt{(1 - r_{13}{}^2)} = 2 \times \sqrt{(1 - 0.5^2)} = 1.7320$$

$$\sigma_{2.3=}\, \sigma_2 \sqrt{(1 - r_{23}{}^2)} = 3 \times \sqrt{(1 - 0.5^2)} = 2.5980$$

$$\sigma_{1.2=}\, \sigma_1 \sqrt{(1 - r_{12}{}^2)} = 2 \times \sqrt{(1 - 0.7^2)} = 1.4282$$

$$\sigma_{3.2=}\, \sigma_3 \sqrt{(1 - r_{32}{}^2)} = 2 \times \sqrt{(1 - 0.5^2)} = 2.5980$$

Eq. (1) gives $b_{12.3} = 0.6 \times \frac{1.7320}{2.5980} = 0.4$ and $b_{13.2} = 0.2425 \times \frac{1.4282}{2.5980} = 0.1333$

(iv) $$\sigma_{1.23} = \sigma_1 \left(\sqrt{\frac{\omega}{\omega_{11}}}\right)$$

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}{}^2 - r_{13}{}^2 - r_{23}{}^2 + 2 r_{12} r_{13} r_{23} = 0.36$$

and $\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}{}^2 = 1 - 0.5^2 = 0.75$

$$\sigma_{1.23} = 2 \times \left(\sqrt{\frac{0.36}{0.75}}\right) = 1.3856$$

Multiple regression:

Bivariate Regression equation:

→ Here we try to study the linear relationship between two variables x and y.

$$Y = a + bX = \beta_0 + \beta_1 X$$

We see that the $a = \beta_0$ is the Y intercept, $b = \beta_1$ is the slope of the linear relationship between the variable X and Y.

Multivariate regression equation

- $Y = a + b_1X_1 + b_2X_2 = \beta_0 + \beta_1X_1 + \beta_2X_2$
- $b_1 = \beta_1$ = partial slope of the linear relationship between the first independent variable and Y, indicates the change in Y for one unit change in X_1.
- $b_2 = \beta_2$ = partial slope of the linear relationship between the second independent variable and Y, indicates the change in Y for one unit change in X_2.

Formulas for finding partial slopes:

$$b_1 = \beta_1 = \frac{S_y}{S_1}\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \beta_2 = \frac{S_y}{S_2}\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

S_y= standard deviation of Y

S_1 = standard deviation of the first independent variable (X_1)

S_2 = standard deviation of the second independent variable (X_2)

r_{y1} = bivariate correlation between Y and X1

r_{y2} = bivariate correlation between Y and X2

r_{12} = bivariate correlation between X_1 and X_2

Example:

1) The salary of a person in an organisation has to be regressed in terms of experience (X_1) and mistakes (X_2). If it is given that the values

$$\bar{Y} = 3.3;\ \overline{X_1} = 2.7;\ \overline{X_2} = 13.7$$
$$S_y = 2.1; S_1 = 1.5;\ S_2 = 2.6$$

and the zero order correlations :

$$r_{y1} = 0.5;\ r_{y2} = -0.3;\ r_{12} = -0.47;$$

Find the linear regression and interpret the results.

So,

$$b_1 = \beta_1 = \frac{S_y}{S_1}\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

$$b_1 = \beta_1 = \frac{2.1}{1.5}\left(\frac{0.50 - (-0.3)(-0.47)}{1 - (-0.47)^2}\right) = 0.65$$

Similarly,

$$b_2 = \beta_2 = \frac{S_y}{S_2}\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \beta_2 = \frac{2.1}{2.6}\left(\frac{0.30 - (0.5)(-0.47)}{1 - (-0.47)^2}\right) = -0.07$$

Calculation of a:

$$a = \bar{Y} - b_1\overline{X_1} - b_2\overline{X_2}$$
$$\beta_0 = \bar{Y} - \beta_1\overline{X_1} - \beta_2\overline{X_2}$$
$$= 3.3 - (0.65)(2.7) - (-0.07)\ 13.7$$
$$= 2.5$$

Interpretation:

1) If a person has no experience and has not done any mistakes, he would get a salary of 2.5 units.
2) If the experience goes up by 1 unit, there would be an increment in the salary by 0.65 units.
3) If he/ she commits a mistake, then the salary would decrease by 0.07 units.

<u>Module-4</u>

DISCRETE PROBABILITY DISTRIBUTIONS

Bernoulli's trials:

Suppose, associated with random trial there is an event called 'success' and the complementary event is called 'failure'. Let the probability for success be p and probability for failure be q. Suppose the random trials are prepared n times under identical conditions. These are called Bernoullian trials.

Bernoulli's Distribution:

A random variable X which takes two values 0 and 1 with probability $q \; and \; p$ respectively. That is $P(X = 0) = q \; and \; P(X = 1) = p, \;\; q = 1 - p$ is called a Bernoulli's discrete random variable. The probability function of Bernoulli's distribution can be written as

$$P(X) = p^X q^{n-X} = p^X (1-p)^{n-X} \;\; ; \; X = 0{,}1$$

Note:

1. Mean of Bernoulli's distribution discrete random variable X

$$\mu = E(X) = \sum X_i . P(X_i) = (0 \times q) + (1 \times p) = p$$

2. Variance of X is

$$V(X) = E(X^2) - E(X)^2 = \sum {X_i}^2 \, P(X_i) - \mu^2$$

$$= (0^2 \times q) + (1^2 \times p) - p^2 = p - p^2 = p(1-p) = pq$$

The standard deviation is $\sigma = \sqrt{pq}$

Probability Binomial Distribution:

Let a random experiment be performed repeatedly and let the occurrence of an event A is any trial be called success and non-occurrence $P(\bar{A})$, a failure (Bernoulli trial). Consider a series of n independent Bernoulli trials (n being finite) in which the probability of success $P(A) = p$ or $P(\bar{A}) = 1 - p = q$ in any trial is constant for each trial.

$$P(X = x) = n_{C_x}\, p^x q^{n-x}, \qquad x = 0,1,2,3, \dots, n$$

Since the probabilities of 0,1,2, 3,…,n successes, namely $q^n, n_{C_1}\ \ q^{n-1}p, n_{C_2}\ \ q^{n-2}p^2, \dots, p^n$ are the successive terms of the Binomial expansion of $(q + p)^n$, the probability distribution so obtained is called Binomial probability distribution.

Definition:

A random variable X is said to be follow Binomial distribution denoted by B (n, p), if it assumes only non-negative values and its probability mass function is given by

$$P(X = x) = \begin{cases} n_{C_x}\, p^x q^{n-x}, & x = 0,1,2,3, \dots, n \\ 0, & otherwise \end{cases}$$

Where $n\ and\ p$ are known as parameters.

Note:

- If n is also sometimes known as the degree of the distribution
- $\sum_{x=0}^{n} n_{C_x}\, p^x q^{n-x} = (q + p)^n = 1$
- The Binomial distribution is important not only because of its wide range applicability, but also because it gives rise to many other probability distributions.
- Any variable which follows Binomial distribution is known as Binomial variate.

Conditions for Binomial Experiment:

The Bernoulli process involving a series of independent trials, is based on certain conditions as under:

- There are only two mutually exclusive and collective exhaustive outcomes of the random variable and one of them is referred to as a success and the other as a failure.
- The random experiment is performed under the same conditions for a fixed and finite (also discrete) number of times, say n. Each observation of the random variable in a random experiment is called a trial. Each trial generates either a success denoted by p or a failure denoted by q.
- The outcome (i.e., success or failure) of any trial is not affected by the outcome of any other trial.
- All the observations are assumed to be independent of each of each other. This means that the probability of outcomes remains constant throughout the process.

Example:

To understand the Bernoulli process, consider the coin tossing problem where 3 coins are tossed. Suppose we are interested to know the probability of two heads. The possible sequence of outcomes involving two heads can be obtained in the following three ways:
HHT, HTH, THH.

Binomial Probability Function:

In general, for a binomial random variable, X the probability of success (occurrence of desired outcome) r number of times in n independent trials, regardless of their order of occurrence is given by the formula:

$$P(X = r\ successes) = n_{C_r} p^r q^{n-r} = \frac{n!}{(n-r)!\,r!} p^r q^{n-r}, r = 0,1,2,3, \dots, n$$

where

n = number of trials (specified in advance) or sample size

p = probability of success

$q = (1 - p)$, probability of failure

x = discrete binomial random variable

r = number of successes in n trials

Relationship between mean and variance:

Mean of a Binomial distribution:

The Binomial probability distribution is given by

$$p(r) = n_{C_r} p^r q^{n-r} \; ; \; r = 0,1,2, \dots, n, \qquad and \; q - 1 - p$$

Mean of X is

$$\mu = E(X) = \sum_{r=0}^{n} r \, p(r) = \sum_{r=0}^{n} r \, n_{C_r} p^r q^{n-r}$$

$$= 0 \times q^n + 1 \times n_{C_1} p q^{n-1} + 2 \times n_{C_2} p^2 q^{n-2} + 3 \times n_{C_3} p^3 q^{n-3} + \cdots + n \, p^n$$

$$= npq^{n-1} + 2. \frac{n(n-1)}{2!} p^2 q^{n-2} + 3. \frac{n(n-1)(n-2)}{3!} p^3 q^{n-3} + \cdots + n \, p^n$$

$$= np \left[q^{n-1} + (n-1) p q^{n-2} + \frac{(n-1)(n-2)}{2!} p^2 q^{n-3} + \cdots + p^{n-1} \right]$$

$$= np(q+p)^{n-1}$$

$$= np$$

$$\mu = E(X) = np$$

Variance of a Binomial distribution:

Variance $V(X) = E(X^2) - E(X)^2$

$$= \sum_{r=0}^{n} r^2 p(r) - \mu^2$$

$$= \sum_{r=0}^{n} [r(r-1) + r]. \, p(r) - \mu^2$$

$$= \sum_{r=0}^{n} r(r-1)\, n_{C_r}\, p^r q^{n-r} + \sum_{r=0}^{n} r\, p(r) - \mu^2$$

$$= \left[2.\, n_{C_2}\, p^2 q^{n-2} + 3.2.\, n_{C_3} p^3 q^{n-3} + \cdots + n(n-1)\, p^n\right] + \mu - \mu^2$$

$$= \left[2.\frac{n(n-1)}{2!} p^2 q^{n-2} + 6.\frac{n(n-1)(n-2)}{3!} p^3 q^{n-3} + \cdots + n(n-1)\, p^n\right] + \mu - \mu^2$$

$$= n(n-1)p^2 \left[q^{n-2} + (n-2)pq^{n-3} + \frac{(n-2)(n-3)}{2!} p^2 q^{n-4} + \cdots + \ p^{n-2}\right] + \mu - \mu^2$$

$$= n(n-1)p^2(q+p)^{n-2} + \mu - \mu^2$$

$$= n(n-1)p^2 + np - (np)^2$$

$$= n^2p^2 - np^2 + np - n^2p^2 = np - np^2 = np(1-p) = npq$$

$$V(X) = npq$$

Problem 1:

A fair coin is tossed six times, then find the probability of getting four heads.

Solution:

p= probability of getting a head=1/2

q= probability of not getting a head=1/2

$n = 6, r = 4$

$p(r) = 6_{C_4} p^r q^{n-r}$

$p(4) = 6_{C_4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2$

$= \frac{6!}{4!\,2!} \left(\frac{1}{2}\right)^6 = \frac{15}{64}$

Problem 2:

The incidence of an occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the probability that out of 6 workers chosen at random, four or more will suffer from disease?

Solution:

The probability of a worker suffering from disease= p=20%=0.2

The probability that of no worker suffering from disease= q=80%=0.8

The probability that four or more workers suffer from disease $= P(X \geq 4)$

$$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6)$$

$$= 6_{C_4}(0.2)^4(0.8)^2 + 6_{C_5}(0.2)^5(0.8) + 6_{C_6}(0.2)^6 = 0.0175$$

Problem 3:

Six dice are thrown 729 times. How many times do you except at least three dice to show a 5 or 6.

Solution:

$p =$ probability of occurrence of 5 or 6 in one throw$= \frac{2}{6} = \frac{1}{3}$

$q = 1 - p = 1 - \dfrac{1}{3} = \dfrac{2}{3}$

$n = 6$

The probability of getting at least three dice to show a 5 or 6

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$

$= 6_{C_3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^3 + 6_{C_4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^2 + 6_{C_5}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^1 + 6_{C_6}\left(\frac{1}{3}\right)^6$

$= \dfrac{1}{(3)^6}(160 + 60 + 12 + 1) = \dfrac{233}{729}$

The expected number of such cases in 729 times

$= 729\left(\dfrac{233}{729}\right) = 233$

Problem 4:

If the probability of a detective bolt is 0.2, find

(i) Mean and

(ii) Standard deviation for the bolts in a total of 400.

Solution:

Given $n = 400, p = 0.2, q = 0.8$

(i) Mean is $np = 400 \times 0.2 = 80$

(ii) Standard deviation is $\sqrt{npq} = \sqrt{80 \times 0.8} = \sqrt{64} = 8$

Problem 5:

Find the maximum n such that the probability of getting no head in tossing a fair coin n times is greater than 0.1.

Solution:

$p =$ probability of getting a head = ½

$q =$ probability of not getting a head = 1- ½ = ½

Probability of getting no head in tossing a fair coin n times is greater than 0.1 is

$$P(X = 0) > 0.1$$

$$n_{C_0} p^0 q^n > 0.1$$

$$q^n > 0.1$$

$$\left(\frac{1}{2}\right)^n > 0.1$$

$$2^n < 10, then\ n > 3.$$

Hence the required maximum $n = 3$.

Problem 6:

Fit a binomial distribution to the following frequency distribution

x	0	1	2	3	4	5	6
f	13	25	52	58	32	16	4

Solution:

The number of trials is $n = 6$

$N = \sum f_i =$ total frequency

Mean $= \frac{\sum f_i x_i}{\sum f_i} = \frac{25+104+174+128+8+24}{200} = 2.675$

Mean $= np$

$$np = 6p, \quad then\ p = \frac{2.675}{6} = 0.446$$

$$q = 1 - 0.446 = 0.554$$

Binomial distribution to be fitted is $N(q+p)^n = 200(0.554 + 0.446)^6$

$$= 200\big[6_{C_0}(0.554)^6 + 6_{C_1}(0.554)^5(0.446) + 6_{C_2}(0.554)^4(0.446)^2 + 6_{C_3}(0.554)^3(0.446)^3 + 6_{C_4}(0.554)^2(0.446)^4 + 6_{C_5}(0.554)^1(0.446)^5 + 6_{C_6}(0.446)^6\big]$$

$$= 200[0.02891 + 0.1396 + 0.2809 + 0.3016 + 0.1821 + 0.05864 + 0.007866]$$

$$= 5.782 + 27.92 + 56.18 + 60.32 + 36.42 + 11.728 + 1.5732$$

The successive terms in the expansion give the expected or theoretical frequencies which are

x	0	1	2	3	4	5	6
f (expected or theoretical frequencies)	6	28	56	60	36	12	2

Home Work:

1. A die is tossed thrice. A success is getting 1 or 6 on a toss. Find the mean and variance of the number of successes.
2. The mean and variance of a binomial distribution are 4 and 4/3 respectively. Then find $P(X \geq 1)$.
3. Fit a binomial distribution to the following frequency distribution

x	0	1	2	3	4	5
f	2	14	20	34	22	8

Moment Generating Function of Binomial distribution:

Let $X \sim B(n, p)$, then

$$M(t) = M_X(t) = E(e^{tx})$$

$$= \sum_{x=0}^{n} e^{tx}\, n_{C_x}\, p^x q^{n-x}$$

$$= \sum_{x=0}^{n} n_{C_x}\, (pe^t)^x\, q^{n-x} = (q + pe^t)^n$$

Characteristic Function of Binomial distribution:

$$\emptyset_X(t) = E\left(e^{itx}\right)$$

$$= \sum_{x=0}^{n} e^{itx}\, n_{C_x}\, p^x q^{n-x}$$

$$= \sum_{x=0}^{n} n_{C_x}\, (pe^{it})^x\, q^{n-x} = (q + pe^{it})^n$$

Problem:

If the moment generating function of a random variable X is of the form $(0.4\, e^{t} + 0.6)^{8}$, find the moment generating function of $3X + 2$.

Solution:

Moment generating function of a random variable X is

$$M_X(t) = E(e^{tx}) = (q + pe^{t})^{n} = (0.6 + 0.4\, e^{t})^{8}$$

X follows the Binomial distribution with $q = 0.4, p = 0.6, n = 8$

MGF of $3X + 2$ is

$$\begin{aligned} M_{3X+2}(t) &= E\left(e^{t(3x+2)}\right) = E(e^{t3x}e^{2t}) = e^{2t}E(e^{t3x}) \\ &= e^{2t}E\left(e^{(3t)x}\right) \\ &= e^{2t}(0.4\, e^{3t} + 0.6)^{8} \end{aligned}$$

Cumulative Binomial distribution:

$$B(x; n, p) = P(X \le x) = \sum_{k=0}^{x} B(k; n, p) = \sum_{k=0}^{x} n_{C_k}\, p^{k} q^{n-k}\ , \qquad x = 0,1,2,3, \dots, n$$

The Binomial probabilities can be obtained from cumulative distribution as follows

$$B(x; n, p) = B(x; n, p) - B(x - 1; n, p)$$

Note: B(-1)=0

By using the Binomial table these can also be obtained.

Example:

The manufacture of large high-definition LCD panels is difficult, and a moderately high proportion have too many defective pixels to pass inspection. If the probability is 0.3 that an

LCD panel will not pass inspection, what is the probability that 6 of 18 panels, randomly selected from production will not pass inspection?

Solution:

X: LCD panel not pass in inspection.

n=18, p=0.30 and x=6

$$b(x;n,p) = B(x;n,p) - B(x-1;n,p)$$

$$b(6;18,0.30) = B(6;18,0.30) - B(5;18,0.30) = 0.7217 - 0.5344 = 0.1873$$

Table A.1 (continued) Binomial Probability Sums $\sum_{x=0}^{r} b(x;n,p)$

		p									
n	**r**	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
17	**0**	0.1668	0.0225	0.0075	0.0023	0.0002	0.0000				
	1	0.4818	0.1182	0.0501	0.0193	0.0021	0.0001	0.0000			
	2	0.7618	0.3096	0.1637	0.0774	0.0123	0.0012	0.0001			
	3	0.9174	0.5489	0.3530	0.2019	0.0464	0.0064	0.0005	0.0000		
	4	0.9779	0.7582	0.5739	0.3887	0.1260	0.0245	0.0025	0.0001		
	5	0.9953	0.8943	0.7653	0.5968	0.2639	0.0717	0.0106	0.0007	0.0000	
	6	0.9992	0.9623	0.8929	0.7752	0.4478	0.1662	0.0348	0.0032	0.0001	
	7	0.9999	0.9891	0.9598	0.8954	0.6405	0.3145	0.0919	0.0127	0.0005	
	8	1.0000	0.9974	0.9876	0.9597	0.8011	0.5000	0.1989	0.0403	0.0026	0.0000
	9		0.9995	0.9969	0.9873	0.9081	0.6855	0.3595	0.1046	0.0109	0.0001
	10		0.9999	0.9994	0.9968	0.9652	0.8338	0.5522	0.2248	0.0377	0.0008
	11		1.0000	0.9999	0.9993	0.9894	0.9283	0.7361	0.4032	0.1057	0.0047
	12			1.0000	0.9999	0.9975	0.9755	0.8740	0.6113	0.2418	0.0221
	13				1.0000	0.9995	0.9936	0.9536	0.7981	0.4511	0.0826
	14					0.9999	0.9988	0.9877	0.9226	0.6904	0.2382
	15					1.0000	0.9999	0.9979	0.9807	0.8818	0.5182
	16						1.0000	0.9998	0.9977	0.9775	0.8332
	17							1.0000	1.0000	1.0000	1.0000
18	**0**	0.1501	0.0180	0.0056	0.0016	0.0001	0.0000				
	1	0.4503	0.0991	0.0395	0.0142	0.0013	0.0001				
	2	0.7338	0.2713	0.1353	0.0600	0.0082	0.0007	0.0000			
	3	0.9018	0.5010	0.3057	0.1646	0.0328	0.0038	0.0002			
	4	0.9718	0.7164	0.5187	0.3327	0.0942	0.0154	0.0013	0.0000		
	5	0.9936	0.8671	0.7175	0.5344	0.2088	0.0481	0.0058	0.0003		
	6	0.9988	0.9487	0.8610	0.7217	0.3743	0.1189	0.0203	0.0014	0.0000	
	7	0.9998	0.9837	0.9431	0.8593	0.5634	0.2403	0.0576	0.0061	0.0002	
	8	1.0000	0.9957	0.9807	0.9404	0.7368	0.4073	0.1347	0.0210	0.0009	
	9		0.9991	0.9946	0.9790	0.8653	0.5927	0.2632	0.0596	0.0043	0.0000
	10		0.9998	0.9988	0.9939	0.9424	0.7597	0.4366	0.1407	0.0163	0.0002
	11		1.0000	0.9998	0.9986	0.9797	0.8811	0.6257	0.2783	0.0513	0.0012
	12			1.0000	0.9997	0.9942	0.9519	0.7912	0.4656	0.1329	0.0064
	13				1.0000	0.9987	0.9846	0.9058	0.6673	0.2836	0.0282
	14					0.9998	0.9962	0.9672	0.8354	0.4990	0.0982
	15					1.0000	0.9993	0.9918	0.9400	0.7287	0.2662
	16						0.9999	0.9987	0.9858	0.9009	0.5497
	17						1.0000	0.9999	0.9984	0.9820	0.8499
	18							1.0000	1.0000	1.0000	1.0000

Poisson Distribution:

S.D. Poisson (1837) introduced Poisson distribution as a rare distribution of rare events.

i.e. The events whose probability of occurrence is very small but the no. of trials which could lead to the occurrence of the event, are very large.

Ex:

1. The no. of printing mistakes per page in a large text
2. Number of suicides reported in a particular city
3. Number of air accidents in some unit time
4. Number of cars passing a crossing per minute during the busy hours of a day, etc.

Definition:

A random variable X taking on one of the non-negative values 0,1,2,3,4, … (i.e. which do not have a natural upper bound) with parameter λ, λ >0, is said to follow Poisson distribution if its probability mass function is given by

$$P(x;\lambda) = P(X = x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0,1,2,3,\ldots \\ 0, & otherwise \end{cases}$$

Then X is called the Poisson random variable and the distribution is known as Poisson distribution.

And the Poisson parameter, $\lambda = \text{np} > 0$

Conditions to follow in Poisson Distribution:

- The no. of trials ‘n’ is very large
- The probability of success ‘p’ is very small
- $\lambda = \text{np}$ is finite.

Mean of Poisson distribution:

$$\mu = E(X) = \sum_{x=0}^{\infty} x\, P(x;\lambda) = \sum_{x=0}^{\infty} x\, \frac{\lambda^x e^{-\lambda}}{x!}$$

$$\mu = E(X) = \lambda = \mathrm{np}$$

Variance of Poisson distribution:

Variance $V(X) = E(X^2) - E(X)^2$

$$= \sum_{x=0}^{\infty} x^2 p(x;\boldsymbol{\lambda}) - \mu^2$$

$$V(X) = \sigma^2 = \lambda$$

Cumulative Poisson distribution:

$$F(x;\lambda) = P(X \le x) = \sum_{k=0}^{x} P(k;\lambda) = \sum_{k=0}^{x} \frac{\lambda^k e^{-\lambda}}{k!}$$

Moment generating function:

$$M(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx}.P(x;\lambda) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{\lambda(e^t - 1)}$$

$$M(t) = e^{\lambda(e^t - 1)}$$

Characteristic function:

$$\emptyset(t) = E\left(e^{itX}\right) = \sum_{x=0}^{\infty} e^{itx}.P(x;\lambda) = \sum_{x=0}^{\infty} e^{itx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{\lambda(e^{it} - 1)}$$

$$\emptyset(t) = e^{\lambda(e^{it} - 1)}$$

Problem 1:

A hospital switch board receives an average of 4 emergency calls in a 10-minute interval. What is the probability that

i. there at most 2 emergency calls in a 10-minute interval

ii. there are exactly 3 emergency calls in a 10-minute interval.

Solution:

Mean =λ $= 4$

$$P(X = x) = p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

i. P(at most 2 calls) = $P(X \leq 2)$

$$= P(X = 0) + P(X = 1) + P(X = 2)$$

$$= \frac{1}{e^4} + 4.\frac{1}{e^4} + 8.\frac{1}{e^4}$$

$$= \frac{1}{e^4}(1 + 4 + 8) = 0.2381$$

ii. P(Exactly 3 calls) = $P(X = 3) = \frac{1}{e^4}.\frac{16}{3!} = 0.1954$

Problem 2:

If a random variable has a Poisson distribution such that P (1) =P (2). Find

i. Mean of the distribution

ii. P(4)

iii. $P(X \geq 1)$

iv. $P(1 < X < 4)$

Solution:

$$\frac{e^{-\lambda}\lambda^1}{1!} = \frac{e^{-\lambda}\lambda^2}{2!}$$

$$\lambda^2 = 2\lambda$$
$$\lambda = 0 \text{ or } 2$$
$$\text{But } \lambda \neq 0 \text{ or } 2$$
$$\text{Therefore } \lambda = 2$$

i. Mean of the distribution is $\lambda = 2$

ii. $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$$p(4) = \frac{e^{-2}2^4}{4!} = 0.09022$$

iii. $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0)$

$$= 1 - \frac{e^{-2}2^0}{0!} = 0.8647$$

iv. $P(1 < X < 4) = P(X = 2) + P(X = 3)$

$$= \frac{e^{-2}2^2}{2!} + \frac{e^{-2}2^3}{3!} = 0.4511$$

Problem 3:

Fit a Poisson distribution to the following data

x	0	1	2	3	4	5
f	142	156	69	27	5	1

Solution:

Mean $= \frac{\sum f_i x_i}{\sum f_i} = \frac{0+156+138+81+20+5}{400} = 1$

Mean of the distribution is $\lambda = 1$

So, theoretical frequency for x successes are given by $N\,P(x)$.

$$N\,P(x) = 400 \times \frac{e^{-1}1^x}{x!}, \quad x = 0,2,3,4,5$$

i.e., $400 \times e^{-1}, 400 \times e^{-1}, 200 \times e^{-1}, 66.67 \times e^{-1}, 16.67 \times e^{-1}, 3.33 \times e^{-1}$

i.e., $147.15, 147.15, 73.58, 24.53, 6.13, 1.23$

The expected frequencies are

x	0	1	2	3	4	5
Theoretical frequency	142	156	69	27	5	1
Expected frequency	147	147	74	25	6	1

Problem 4:

If the moment generating function of the random variable is $e^{4(e^t-1)}, find\ P(X=\mu+\sigma)$, where μ and σ^2 are the mean and variance of the Poisson random variable X.

Solution:

$$M(t)=e^{\lambda(e^t-1)}=e^{4(e^t-1)}$$

Mean=Variance= $\lambda=4$

Standard deviation = $\sqrt{4}$ =2

$$P(X=\mu+\sigma)=P(X=4+2)=P(6)$$

$$P(X=x)=P(X=6)=\frac{e^{-4}4^6}{6!}=0.1042$$

Try yourself:

1. The distribution of typing mistakes committed by a typist is given below. Assuming the distribution to be Poisson, find the expected frequencies

x	0	1	2	3	4	5

f	42	33	14	6	4	1

Hypergeometric Distribution:

A discrete random variable X is said to follow the hypergeometric distribution with parameters N, M *and* n, if it assumes only non-negative values and its probability mass function is given by

$$P(X = x) = h(k; N, M, n) = \begin{cases} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, & k = 0,1,2,3,\dots, \quad \min(n, M) \\ 0, & otherwise \end{cases}$$

Where N is a positive integer, M is a positive integer not exceeding N and n is a positive integer that is at most N.

(Or)

$$P(X = x) = h(k; N, M, n) = \frac{M_{C_k} \times N - M_{C_{n-k}}}{N_{C_n}}$$

Mean and Variance of Hypergeometric Distribution:

Mean is $E(X) = np = \frac{nM}{N}, where\ p = \frac{M}{N}$

Variance is $var(X) = npq = \frac{nM(N-M)(N-n)}{\mathrm{N}^2(N-1)}$

Problem 1:

A batch of 10 rocker cover gaskets contains 4 defective gaskets. If we draw samples of size 3 without replacement, from the batch of 10, find the probability that a sample contains 2 defective gaskets. Also find mean and variance.

Solution:

$$P(X = x) = h(k; N, M, n) = \frac{M_{C_k} \times \ N - M_{C_{n-k}}}{N_{C_n}}$$

Here $N = 10, M = 4, n = 3, k = 2$

$$P(X = 2) = h(k; N, M, n) = \frac{4_{C_2} \times \ 6_{C_1}}{10_{C_3}} = 0.3$$

Mean is $E(X) = np = \frac{nM}{N} = \frac{3\times4}{10} = \frac{12}{10} = 1.2$

Variance is $var(X) = npq = \frac{nM(N-M)(N-n)}{\mathrm{N}^2(N-1)} = \frac{3\times4(10-4)(10-3)}{10^2(10-1)} = 0.56$

Problem 2:

In the manufacture of car tyres, a particular production process is known to yield 10 tyres with defective walls in every batch of 100 tyres produced. From a production batch of 100 tyres, a sample of 4 is selected for testing to destruction. Find

i. the probability that the sample contains 1 defective tyre
ii. the expectation of the number of defectives in samples of size 4
iii. the variance of the number of defectives in samples of size 4.

Solution:

Sampling is clearly without replacement and we use the hypergeometric distribution with $N = 100, M = 10, n = 4, k = 1$

i. $P(X = x) = \frac{M_{C_k} \times N - M_{C_{n-k}}}{N_{C_n}}$

$$P(X = x) = \frac{10_{C_1} \times \ 100 - 10_{C_{4-1}}}{100_{C_4}} = \frac{10 \times 117480}{3921225} = 0.299 \approx 0.3$$

ii. The expectation of the number of defectives in samples of size 4

$$E(X) = \frac{nM}{N} = \frac{4 \times 10}{100} = 0.4$$

iii. The variance of the number of defectives in samples of size 4

$$var(X) = \frac{nM(N - M)(N - n)}{\mathrm{N}^2(N - 1)} = \frac{4 \times 10(100 - 10)(100 - 4)}{100^2(100 - 1)} = 0.349$$

Multinomial distribution:

- This distribution can be regarded as a generalization of Binomial distribution.
- When there are more than two mutually exclusive outcomes of a trial, the observations lead to multinomial distribution. Suppose $E_1, E_2, \ldots, E_k$ are k mutually exclusive outcomes of a trial with respective probabilities.
- Th probability that E_1 occurs x_1 times, E_2 occurs x_2 times,… and E_k, occurs x_k times n independent observations, is given by $p(x_1, x_2, \ldots, x_k) = c{p_1}^{x_1}{p_2}^{x_2} \ldots {p_k}^{x_k}, where \sum x_i = n \; and \; c$ is the number of permutations of the events $E_1, E_2, \ldots, E_k$.
- To determine c, we have to find the number of permutations of n objects of which x_1 are of one kind, x_2 of another kind, …, x_k of another kind k^{th} kind, which is given by

$$c = \frac{n!}{x_1!\, x_2! \ldots x_k!}$$

Hence $p(x_1, x_2, \ldots x_k) = \frac{n!}{x_1!x_2!\ldots x_k!}{p_1}^{x_1}{p_2}^{x_2} \ldots {p_k}^{x_k}, \;\; 0 \le x_i \le n$

$$p(x_1, x_2, \ldots x_k) = p = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} {p_i}^{x_i} \tag{1}$$

Which is the required probability function of the multinomial distribution. Eq. (1) is called in multinomial expansion

$$(p_1 + p_2 + \ldots + p_k)^n, \quad \sum_{i=1}^{k} p_i = 1$$

Since, the total probability is 1, we have

$$\sum_x p(x) = \sum_x \left[\frac{n!}{x_1!\,x_2!\ldots x_k!} {p_1}^{x_1} {p_2}^{x_2} \ldots {p_k}^{x_k}\right]$$
$$= (p_1 + p_2 + \ldots + p_k)^n = 1$$

Moments of multinomial distribution:

$$M(t) = M_{X_1, X_2, \ldots, X_k}(t_1, t_2, \ldots, t_k) = E\left(e^{\sum_{i=1}^{k} t_i X_i}\right)$$
$$= \left(p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k}\right)^n$$

Moments:

$$E(X_i) = np_i$$
$$Var(X_i) = np_i(1 - p_i)$$

Problem 1:

Suppose we have a bowl with 10 marbles 2 red marbles, 3 green marbles, and 5 blue marbles. We randomly select 4 marbles from the bowl, with replacement. What is the probability of selecting 2 green marbles and 2 blue marbles?

Solution:

To solve this problem, we apply the multinomial formula. We know the following:

The experiment consists of 4 trials, i.e., $n = 4$.

4 trials produce 0 red marbles, 2 green marbles, and 2 blue marbles;

i.e., $x_{red} = x_1 = 0$, $x_{green} = x_2 = 2$, and $x_{blue} = x_3 = 2$

On any particular trial, the probability of drawing a red, green, and blue marble is 0.2, 0.3, and 0.5, respectively.

i.e., $p_{red} = 0.2, p_{green} = 0.3,\ p_{blue} = 0.5$

The multinomial formula is

$$p(x_1, x_2, \dots x_k) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} {p_i}^{x_i}$$

$$\text{i.e., } p = \frac{n!}{x_1! x_2! x_3!} {p_1}^{x_1} {p_2}^{x_2} {p_3}^{x_3}$$

$$= \frac{4!}{0!\,2!\,2!}\,(0.2)^0 (0.3)^2 (0.5)^2 = 0.135$$

$$p = 0.135$$

Thus, if we draw 4 marbles with replacement from the bowl, the probability of drawing 0 red marbles, 2 green marbles, and 2 blue marbles is 0.135.

Problem 2:

In India, 30% of the population has a blood type of O+, 33% has A+, 12% has B+, 6% has AB+, 7% has O-, 8% has A-, 3% has B-, and 1% has AB-. If 15 Indian citizens are chosen at random, what is the probability that 3 have a blood type of O+, 2 have A+, 3 have B+, 2 have AB+, 1 has O-, 2 have A-, 1 has B-, and 1 has AB-?

Solution:

$$n = 15 \text{ trials}$$

$$p_1 = 30\% = 0.30 \ \text{(probability of O+)}$$

$$p_2 = 33\% = 0.33 \text{ (probability of A+)}$$

$$p_3 = 12\% = 0.12 \text{ (probability of B+)}$$

$$p_4 = 6\% = 0.06 \text{ (probability of AB+)}$$

$$p_5 = 7\% = 0.07 \text{ (probability of O}-\text{)}$$

$$p_6 = 8\% = 0.08 \text{ (probability of A}-\text{)}$$

$$p_7 = 3\% = 0.03 \text{ (probability of B}-\text{)}$$

$$p_8 = 1\% = 0.01 \text{ (probability of AB}-)$$

$$x_1 = 3 \text{ (3 O+)}$$

$$x_2 = 2 \text{ (2 A+)}$$

$$x_3 = 3 \text{ (3 B+)}$$

$$x_4 = 2 \text{ (2 AB+)}$$

$$x_5 = 1 \text{ (1 O}-)$$

$$x_6 = 2 \text{ (2 A}-)$$

$$x_7 = 1 \text{ (1 B}-)$$

$$x_8 = 1 \text{ (1 } A\text{B}-)$$

$$k = 8 \text{ (8 } possibilities)$$

$$p = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!\,x_5!\,x_6!\,x_7!\,x_8!} {p_1}^{x_1} {p_2}^{x_2} {p_3}^{x_3} {p_4}^{x_4} {p_5}^{x_5} {p_6}^{x_6} {p_7}^{x_7} {p_8}^{x_8}$$

$$= \frac{15!}{3!\,2!\,3!\,2!\,1!\,2!\,1!\,1!} \times 0.30^3 \times 0.33^2 \times 0.12^3 \times 0.06^2 \times 0.07^1 \times 0.08^2 \times 0.03^1 \times 0.01^1$$

$$p = 0.000011$$

Discrete Bivariate Distributions:

Covariance:

We are often interested in the inter-relationship, or association, between two random variables. The covariance of two random variables X and Y is

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

Or

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

Note:

Covariance is an obvious extension of variance

$$Cov(X,X) = Var(X)$$

Moments:

1. **The moments for the Binomial Distribution**:

By the definition of a moment $\mu_r = E[X - E(X)]^r$

The first four central moments of the Binomial distribution are:

we know that $\mu_0 = 1, \mu_1 = 0$

$$Mean = np$$

$$\mu_2 = npq$$

$$\mu_3 = npq(q - p)$$

$$\mu_4 = npq[1 + 3pq(n - 2)]$$

2. **The moments for the Poisson Distribution**:

The first four central moments of the Poisson distribution are:

$$\mu_1 = 0$$

$$\mu_2 = \lambda$$

$$\mu_3 = \lambda$$

$$\mu_4 = 3\lambda^2 + \lambda$$

Module 5

Continuous Probability Distribution

Uniform Distribution:

A random variable X is said to follow uniform distribution over an interval (a, b), if its probability density function is constant = k (say), over the entire range of X,

$$f(X) = \begin{cases} k, & a < X < b \\ 0, & otherwise \end{cases}$$

Since the total probability is always unity, we have

$$\int_a^b f(X)dX = 1$$

$$\int_a^b k\,dX = 1$$

$$k(b-a) = 1$$

$$k = \frac{1}{(b-a)}$$

$$\boldsymbol{f(X) = \begin{cases} \dfrac{1}{(b-a)}, & a < X < b \\ 0, & otherwise \end{cases}}$$

Note:

1. $\int_{-\infty}^{\infty} f(X)dX = \int_a^b \frac{1}{(b-a)}\,dX = 1,\ a < b,$ a and b are two parameters of the uniform distribution on (a, b).
2. The distribution is also known as rectangular distribution, since the curve $y = f(x)$ describes a rectangle over the X-axis and between the coordinates at $x = a$ and $x = b$.
3. The distribution function $F(x)$ is given by
4. $f(X) = \begin{cases} 0, & if -\infty < X < a \\ \frac{X-a}{b-a}, & a \le X \le b \\ 1, & b < X < \infty \end{cases}$

Since $F(X)$ is not continuous at $x = a$ and $x = b$, it is not differentiable at these points. Thus $\frac{d}{dX}\,F(X) = f(X) = \frac{1}{(b-a)} \neq 0$ exists everywhere except the points $x = a$ and $x = b$.

Moments:

$$\mu_r' = \int_a^b X^r\, f(X)dX$$

$$= \frac{1}{(b-a)}\int_a^b X^r\,dX = \frac{1}{(b-a)}\left[\frac{b^{r+1} - a^{r+1}}{r+1}\right]$$

In particular,

$$\textbf{Mean}=\boldsymbol{\mu_1}' = \frac{\boldsymbol{b+a}}{\boldsymbol{2}}$$

$$\mu_2' = \frac{1}{3}(b^2 + ab + a^2)$$

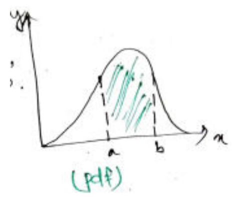
$$\boldsymbol{variance = \mu_2 = \mu_2' - (\mu_1')^2 = \frac{(b-a)^2}{12}}$$

Normal probability distribution:

Probability density or distribution function (PDF):

Let X be a continuous random variable, then the probability density function of X is a function of $f(X)$ such that for any two numbers a and b with $a \leq b$.

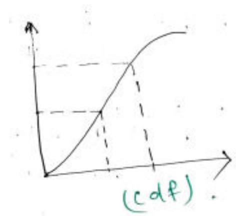
$P(a \leq X \leq b) = \int_a^b f(X)dX$



Cumulative distribution function (CDF):

$$F(X) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

$$P(a \leq X \leq b) = F(b) - F(a)$$



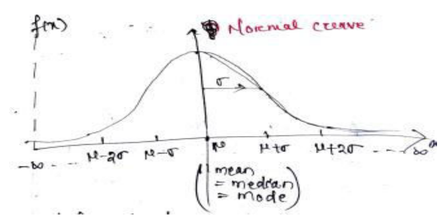
Normal Distribution:

A random variable X is said to have a normal distribution, if its density function or probability distribution is given by

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty, -\infty < \mu < \infty, \qquad \sigma > 0.$$

Where, μ is the mean and σ is the standard deviation of x.

- As can be seen, the function, called probability density function of the normal distribution, depends on two values μ and σ. These are referred as the two parameters of the normal distribution.
- The curve has maximum value at μ and tapers off on either side but never touches the horizontal line.
- The curve on the left side goes up to $-\infty$, and on the right side it goes up to $+\infty$. However, as much as 99.73% of the area under the curve lies between $(\mu - 3\sigma)$ and $(\mu + 3\sigma)$ and only 0.277% of the area lies beyond these points.
- The random variable x is then said to a normal random variable or normal variate. The curve representing the normal distribution is called the normal curve and the total area bounded by the curve and the x -axis is one. i.e., $\int f(x)dx = 1$



Normal distribution is applicable in the following situations:

1. Life of items subjected to wear and tear like tyres, batteries, bulbs, currency notes, etc.
2. Length and diameter of certain products like pipes, screws and discs.
3. Height and weight of baby at birth.
4. Aggregate marks obtained by students in an examination.

5. Weekly sales of an item in store.

Standard Normal Distribution:

The Normal Distribution with mean (μ) =0 and S.D. (σ)=1, is known as Standard Normal Distribution.

The random variable that follows this distribution is denoted by z. If a variable x follows normal distribution with mean μ and s.d. σ, the variable z defined as

$$z=\frac{x-\mu}{\sigma}$$

has standard normal distribution with mean 0 and s.d. as 1. This is also referred as z-score.

Uses of Normal distribution:

1. The Normal distribution can be used to approximate Binomial and Poisson distributions.
2. It has extensive use in sampling theory. It helps us to estimate parameter from statistic and to find confidence limits of the parameter.
3. It has a wide use in testing Statistical Hypothesis and Tests of significance in which it is always assumed that the population from which the samples have been drawn should have normal distribution.
4. It serves as a guiding instrument in the analysis and interpretation of statistical data.

Mean of Normal Distribution:

Consider the Normal Distribution with b, σ as the parameters. Then

$$f(x; b, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(x-b)^2}{2\sigma^2}}$$

The mean μ= E(X) is given by

$$\mu = \int_{-\infty}^{\infty} x\, f(x)dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(x-b)^2}{2\sigma^2}}\, dx.$$

put $z = \frac{x-b}{\sigma}$ => $\sigma z + b = x$

$dz = \frac{dx}{\sigma}$

$$\mu = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z + b) e^{\frac{-z^2}{2}} dz.$$

$$\mu = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z) e^{\frac{-z^2}{2}} dz + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (b) e^{\frac{-z^2}{2}} dz$$

here, $ze^{\frac{-z^2}{2}}$ *is odd function so* $\int_{-\infty}^{\infty} (z) e^{\frac{-z^2}{2}} dz = 0$

$$\mu = 0 + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (b) e^{\frac{-z^2}{2}} dz.$$

$$\mu = \frac{b}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz.$$

$e^{\frac{-z^2}{2}}$ *is even function so* $\int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz = 2\int_{0}^{\infty} e^{\frac{-z^2}{2}} dz$

$$\mu = \frac{2b}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{-z^2}{2}} dz.$$

We know that, $\int_{0}^{\infty} e^{\frac{-z^2}{2}} dz = \sqrt{\frac{\pi}{2}}$

$$\mu = \frac{2b}{\sqrt{2\pi}} \times \sqrt{\frac{\pi}{2}}$$

$$\mu = b$$

Variance of N.D:

$$\sigma_X{}^2 = Var(X) = \int_{-\infty}^{\infty} (X-\mu)^2 dX = E[(X-\mu)^2]$$

Let X has a normal distribution i.e., $X \sim N(\mu, \sigma^2)$ with mean μ and standard deviation σ, we can standardize to a standard normal random variable

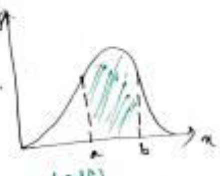
$$Z = \frac{X-\mu}{\sigma}$$

Normal Distribution

Probability density / distribution function :→

Let X be a cont. r.v. Then the probability density function (pdf) of X is a function 'f(x)' such that for any two numbers 'a' & 'b' with a≤b.

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

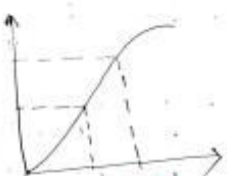


(pdf)

CDF

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)\,dy$$

$$P(a \le X \le b) = F(b) - F(a)$$



(cdf)

Mean $\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$

Variance $\sigma_X^2 = V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx = E\left[(X-\mu)^2\right]$.

Normal distribution

A cont. r.v. 'X' is said to have the normal distribution with parameter 'μ' and 'σ' if its pdf is given by

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}}, \quad -\infty < x < \infty$$

where $\mu = E(X)$, σ = std. dev.

- Thus the normal distribution is characterised by mean 'μ' & std dev 'σ'

*. It is used to study the height of person, the velocity in any direction of a gas molecule & error made in measuring physical quantity. etc.

Properties

→ Applied to single variable cont. data. i.e. height, weight, length etc.

→ ~~The normal curve is bell~~ used to calculate the probability 'less than' 'greater than', & in bet".

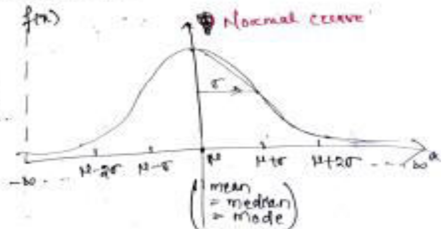
→ Since f(x) being the probability, can never be negative, no portion of the curve lies below x-axis.

graphical property

→ The normal density curve is bell shaped.

→ It is symmetric about mean.

(mean = median = mode)



→ Spread of the curve is determined by the std deviation 'σ'.

→ Location is determined by the mean 'μ'.

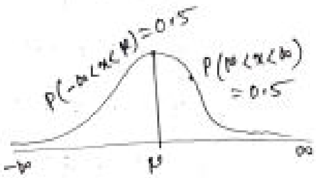
IBM.

The normal curve is symmetric about mean.

The total area under the curve is 1.

i.e. $P(-\infty < X < \infty) = 1$, also $P(-\infty < Z < \infty) = 1$

and since the curve is symmetric about mean μ so half is above (right) the mean and half is below (left) the mean.

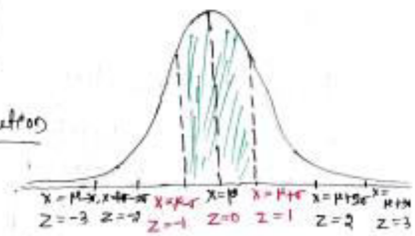


$P(-\infty < X < \infty) = 1$

Standardizing Normal RV.

If 'X' has a normal distribution (i.e. $X \sim N(\mu, \sigma^2)$) with mean μ and standard deviation σ, we can standardize to a standard normal r.v.

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$



Standard normal distribution

$(\mu = 0, \sigma^2 = 1)$

then,

$$\boxed{f(z) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{z^2}{2}}}$$

or

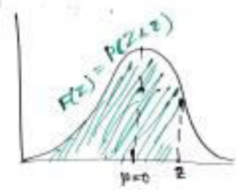
$f(z; 0, 1)$ or $\phi(z)$

CDF of Z

$$F_Z(z) = P(Z \le z) = \int_{-\infty}^{z} f(y; 0, 1)\, dy$$

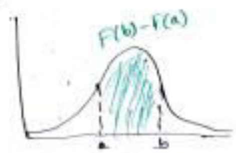
or $\Phi(z)$

$$= \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{y^2}{2}}\, dy$$



The standard normal probability in the form of CDF.

$$P(a \le Z \le b) = F(b) - F(a)$$



* $F(-z) = 1 - F(z)$.

Note: CDF of n.d. does not have any analytic soln & its values must be looked up in $N(0,1)$ table.

MGF

$$\boxed{M(t) = \exp\left(\mu t + \frac{t^2\sigma^2}{2}\right)}$$

Normal probabilities

To find probabilities concerning X, we need to convert its values to z scores using

$$\boxed{Z = \frac{X-\mu}{\sigma}}$$

*→ when X has the normal distribution with mean μ & std. deviation σ.

$$P(a < X \le b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right)$$

$$= P\left(\frac{a-\mu}{\sigma} < Z \le \frac{b-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Ex. Find the probabilities that a r.v. having the std. normal distribution will take on a value

(i) between 0.87 and 1.28

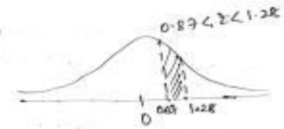
(ii) between −0.34 & 0.62

(iii) greater than 0.85

(iv) greater than −0.65.

Sol

(i) $P(0.87 < Z < 1.28)$

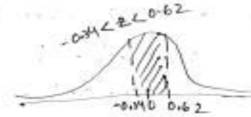


$= \Phi(1.28) - \Phi(0.87)$

$= 0.8997 - 0.8078$

$= 0.0919.$

(ii) $P(-0.34 < Z < 0.62)$



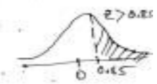
$= \Phi(0.62) - \Phi(-0.34)$

$= \Phi(0.62) - [1 - \Phi(0.34)]$

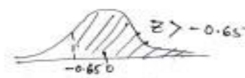
$= 0.7324 - 0.3669 = 0.3655$

(iii) $P(Z > 0.85) = 1 - P(Z < 0.85)$

$= 1 - \Phi(0.85)$



(iv) $P(Z > -0.65)$



$= 1 - P(Z < -0.65)$

$= 1 - \Phi(-0.65)$

$= 1 - [1 - \Phi(0.65)]$

$= \Phi(0.65) = 0.7422.$

Ex-2 X is normally distributed and the mean of X is 12 and S.D. is 4.

(a) Find out the probability of the following

(i) $X \geq 20$ (ii) $X \leq 20$ (iii) $0 \leq X \leq 12$

(b) Find x', when $P(X > x') = 0.24$

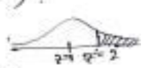
(c) Find x_0' & x_1', when $P(x_0' < X < x_1') = 0.50$ and $P(X > x_1') = 0.25$

Sol we have $\mu = 12$, $\sigma = 4$, i.e. $X \sim N(\underset{\mu}{12}, \underset{\sigma^2}{16})$

(a)

(i) $P(X \geq 20)$

$\Rightarrow P\left(\frac{X-\mu}{\sigma} \geq \frac{20-\mu}{\sigma}\right) = P\left(Z \geq \frac{20-12}{4}\right)$



$\Rightarrow P(Z \geq 2) = 1 - P(Z \leq 2)$
$= 1 - 0.9772 = 0.0228$

(ii) $P(X \leq 20) = 0.9772.$
$= P(Z \leq 2)$

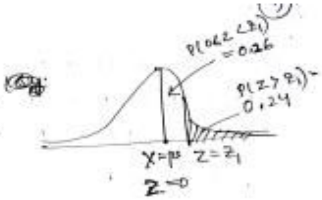
(iii) $P(0 \leq X \leq 12) = P(-3 \leq Z \leq 0)$
$= \Phi(0) - \Phi(-3)$
$= \Phi(0) - (1 - \Phi(3))$
$= 0.5 - (1 - 0.9987)$
$= 0.4987.$

(b) Given $P(X > x') = 0.24$

$\Rightarrow P\left(\frac{X-\mu}{\sigma} > \frac{x'-\mu}{\sigma}\right) = P(Z > z_1) = 0.24$

Since, $P(Z > z_1) = 0.24$

then $P(0 < Z < z_1) = 0.26.$



$\Rightarrow \Phi(z_1) - \Phi(0) = 0.26$

$\Rightarrow \Phi(z_1) = 0.26 + 0.5$

$\Rightarrow \Phi(z_1) = 0.76.$

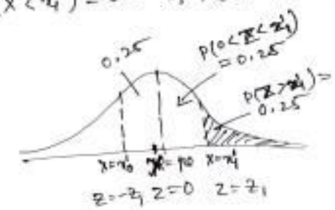
$\Rightarrow z_1 \sim .71$ (from normal table).

hence, $\frac{x'-12}{4} = 0.71 \Rightarrow x' = 12 + 4 \times 0.71 = 14.84$

(c) we are given $P(x_0' < X < x_1') = 0.5$ & $P(X > x_1') = 0.25$

when $X = x_1'$

$Z = \frac{x_1' - 12}{4} = z_1'$ (say)



$X = x_0'$,

$Z = \frac{x_0' - 12}{4} = -z_1'$

we have $P(Z > z_1) = 0.25 \Rightarrow P(0 < Z < z_1) = 0.25$

$z_1 = 0.67$ (From normal table).

hence $\frac{x_1' - 12}{4} = 0.67 \Rightarrow x_1' = 12 + 4 \times 0.67 = 14.68$

$\frac{x_0' - 12}{4} = -0.67 \Rightarrow x_0' = 12 - 4 \times 0.67 = 9.32.$

HW $X \sim N(\underset{\mu}{30}, \underset{\sigma^2}{25})$, Find the probabilities that

(i) $26 \leq X \leq 40$ (ii) $X \geq 45$ & (iii) $|X - 30| > 5$.

Exponential Probability Distribution:

A continuous random variable X is said to follow an exponential distribution with parameter $\lambda > 0$, if its probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & otherwise \end{cases}$$

The general form of the exponential distribution is

$f(x) = \frac{1}{a} e^{-\frac{x}{a}},\ \ a > 0,\ x \geq 0$ with parameter a.

Momentum generating Function (MGF) of Exponential Distribution:

The MGF is $M_X(t) = \frac{\lambda}{\lambda - t}$

Mean $= \frac{1}{\lambda}$

Variance$= \frac{1}{\lambda^2}$

The cumulative distribution function is

$$F(x) = P(X \leq x) = \int_0^x f(x)dx = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$$

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0\ , & x < 0 \end{cases}$$

Exponential Distribution possesses memoryless property:

$P(X > s + t\ /X > t) = P(X > s)$, for any $s, t > 0$

The probability density function of X is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$P(X > k) = \int_k^\infty \lambda e^{-\lambda x} dx = e^{-\lambda k}$

$$P(X > s + t/\ X > t) = \frac{P(X > s + t\ \cap\ X > t)}{P(X > t)} = \frac{P(X > s + t)}{P(X > t)}$$

$$= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-st} = P(X > s)$$

Therefore, exponential distribution possesses memoryless property.

Problem 1:

If X has an exponential distribution with mean is 2, find $P(X < 1/\ X < 2)$.

Solution:

Mean of the exponential distribution is

Mean $= \frac{1}{\lambda} = 2$

$$\lambda = \frac{1}{2} = 0.5$$

The probability density function is

$$f(x) = \lambda e^{-\lambda x} = 0.5 e^{-0.5\,x}, x \geq 0$$

$$P(X < 1/\ X < 2) = \frac{P(X < 1 \cap X < 2)}{P(X < 2)}$$

$$= \frac{P(X < 1)}{P(X < 2)}$$

$$P(X < 1) = \int_{-\infty}^1 f(x)dx = \int_0^1 0.5 e^{-0.5x} dx = 0.5 \left[\frac{e^{-0.5} - 1}{-0.5}\right] = 0.393\,4$$

$$P(X < 2) = \int_0^x f(x)dx = \int_0^2 0.5 e^{-0.5x} dx = 0.5 \left[\frac{e^{-1} - 1}{-0.5}\right] = 0.6321$$

$P(X < 1/X < 2) = \frac{0.3934}{0.6321} = 0.6223$

Problem 2:

The time (in hours) required to repair a watch is exponentially distributed with parameter $\lambda = \frac{1}{2}$

i. What is the probability that the repair the time exceeds 2 hours?

ii. What is the probability that a repair takes 11 hours given that duration exceeds 8 hours?

with mean 120 days, find the probability that such a watch

iii. Will have to set in less than 24 days, and

iv. Not have to reset in a least 180 days.

Solution:

Let X be the random variable which denotes the time to repair the watch.

The probability density function of the exponential distribution is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Given that $\lambda = \frac{1}{2}$

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{2} e^{-\frac{1}{2}x}, x \geq 0$$

i. $P(X > 2) = \int_2^\infty \frac{1}{2} e^{-\frac{1}{2}x} dx = e^{-1}$

ii. Using the memoryless property, we have

$$P(X \geq 11/\, X > 8) = P(X > 3)$$

$P(X > 3) = \int_3^\infty \frac{1}{2} e^{-\frac{1}{2}x} dx = e^{-1.5}$

In the second case, given

Mean=120 i.e., Mean $= \frac{1}{\lambda} = 120$

$$\lambda = \frac{1}{120}$$

The probability density function is given by

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{120} e^{-\frac{1}{120}x}, x \geq 0$$

iii. $P(X < 24) = \int_0^{24} \frac{1}{120} e^{-\frac{1}{120}x} dx = 1 - e^{-0.2} = 0.1813$

iv. $P(X > 180) = \int_{180}^\infty \frac{1}{120} e^{-\frac{1}{120}x} dx = e^{-1.5} = 0.2231$

Problem 3:

The time line in hours required to repair a machine is exponentially distributed with parameter $\lambda = \frac{1}{2}$. What is the probability that the required time

i. Exceeds 2 hours

ii. Exceed 5 hours

Solution:

Let X be the random variable which denotes the time to repair the machine. Then the density function of X is given by

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{2} e^{-\frac{1}{2}x}, \qquad x > 0$$

i. $P(X > 2) = \int_2^\infty \frac{1}{2} e^{-\frac{1}{2}x} dx = e^{-1}$

ii. $P(X > 5) = \int_5^\infty \frac{1}{2} e^{-\frac{1}{2}x} dx = e^{-\frac{5}{2}} = 0.082$

Try yourself:

Problem 4:

A component has an exponentially time of failure distribution with mean 10,000 hours

i. The component has already been in operation for its mean life. What is the probability that it will fail by 15,000 hours?

ii. At 15,000 hours the component is still in operation life. What is the probability that it operates for another 5,000 hours?

Problem 5:

The mileage which car owners get with certain kind of radial tyre is a random variable having an exponential distribution with mean 40,000 km. Find the probabilities that one of these tyres will last

i. At least 20,000 km

ii. At most 30,000 km.

Gamma Distribution:

A continuous random variable X is said to follow general Gamma distribution with two parameters $\boldsymbol{\lambda > 0}$ **and** $\boldsymbol{k > 0}$**,** if its probability density function is given by

$$\boldsymbol{f(x) = \begin{cases} \dfrac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}, x \geq 0 \\ 0, \quad\quad\quad otherwise \end{cases}}$$

Note:

1. When $\boldsymbol{k = 1}$, the distribution is called exponential distribution
2. $\int_{-\infty}^{\infty} f(x)dx = 1$ $\left(\text{Since, } \int_0^{\infty} x^{k-1}e^{-ax}\, dx = \frac{\Gamma(k)}{a^k}\right)$

Momentum generating Function of Gamma Distribution:

The probability density function of the general Gamma random variable X is

$$f(x) = \begin{cases} \dfrac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}, x \geq 0 \\ 0, \quad\quad\quad otherwise \end{cases}$$

Where $\lambda\ and\ k$ are the parameters.

The MGF is

$$\boldsymbol{M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^k}$$

Mean $= \frac{\boldsymbol{k}}{\boldsymbol{\lambda}}$

Variance $= \frac{\boldsymbol{k}}{\boldsymbol{\lambda^2}}$

Problem 1:

The lifetime (in hours) of a certain piece of equipment is a continuous random variable having range $0 < x < \infty$ and the PDF is $f(x) = \begin{cases} xe^{-kx}, & 0 < x < \infty \\ 0, & otherwise \end{cases}$. Determine the constant k and evaluate the probability that the lifetime exceeds 2 hours.

Solution:

Let X denote the lifetime of a certain piece of equipment with PDF

$$f(x) = \begin{cases} xe^{-kx}, & 0 < x < \infty \\ 0, & otherwise \end{cases}$$

Now we have to find k

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_0^{\infty} xe^{-kx}dx = \int_0^{\infty} x^{2-1}\, e^{-kx}dx = 1$$

Using

$$\int_0^{\infty} x^{n-1}e^{-ax}dx = \frac{\Gamma(n)}{a^n}$$

$$\frac{\Gamma(2)}{k^2} = 1$$

$$k^2 = 1$$

$$k = 1$$

Then

$$f(x) = \begin{cases} xe^{-x}, & 0 < x < \infty \\ 0, & otherwise \end{cases}$$

P (lifetime exceeds 2 hours) $= P(X > 2) = \int_2^{\infty} f(x)dx = \int_2^{\infty} xe^{-x}dx = 0.4060$

Problem 2:

The daily consumption of milk in a city, in excess of 20,000 liters, is approximately distributed as a Gamma variate with parameters $k = 2$ and $\lambda = \frac{1}{10,000}$. The city has a daily stock of 30,000 liters. What is the probability that the stock is insufficient on a particular day?

Solution:

If the random variable X denotes the daily consumption of milk (in liters) in a city, then the random variable $Y = X - 20,000$ has a Gamma distribution with probability density function

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}, x \geq 0$$

$$f(y) = \frac{\lambda^2 y^{2-1} e^{-\lambda y}}{\Gamma(2)}, y \geq 0$$

$$= \frac{\left(\frac{1}{10,000}\right)^2 y^{2-1} e^{-\left(\frac{1}{10,000}\right)y}}{\Gamma(2)}, y \geq 0$$

Since the daily stock of the city is 30,000 liters, the required probability that the stock is insufficient on a particular day is given by

$$P(X > 30,000) = P(Y > 10,000) = \int_{10,000}^{\infty} f(y)dy$$

$$= \int_{10,000}^{\infty} \left(\frac{1}{10,000}\right)^2 \frac{y^{2-1} e^{-\frac{y}{10,000}}}{\Gamma(2)} dy = \int_{1}^{\infty} ze^{-z} dz$$

Taking $z = \frac{y}{10,000}$

$$\int_{1}^{\infty} ze^{-z} dz = e^{-1} + e^{-1} = 2e^{-1} = 0.7357$$

Problem 3:

In a certain city, the daily consumption of electric power (in millions of kilowatt-hours) can be treated as a random variable having Gamma distribution with parameters $\lambda = \frac{1}{2}$ and $k = 3$. If the power plant of this city has a daily capacity of 12 million kilowatt hours, what is the probability that this power supply will be adequate on any day?

Solution:

Let X be the random variable denoting the daily consumption of electric power (in millions of kilowatt hours).

Also, given $\lambda = \frac{1}{2}$ and $k = 3$.

Gamma distribution with probability distribution function is

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$$

$$= \frac{\left(\frac{1}{2}\right)^3 x^{3-1} e^{-\frac{x}{2}}}{\Gamma(3)}, x \geq 0$$

The daily capacity of the power plant is 12 million kilowatt hours. The power supply is more than 12 million on any day.

$$P(X > 12) = \int_{12}^{\infty} f(x)dx = \int_{12}^{\infty} \frac{\left(\frac{1}{8}\right) x^2 e^{-\frac{x}{2}}}{\Gamma(3)} dx$$

$$= \int_{12}^{\infty} \frac{\left(\frac{1}{8}\right) x^2 e^{-\frac{x}{2}}}{\Gamma(3)} dx = 0.0625$$

Try yourself:

Problem 4:

Consumer demand for milk in a certain locality per month is known to be a general Gamma distribution random variable. If the average demand is a liters and the most likely demand is b liters $(b < a)$, what is the variance of the demand?

Beta Distribution:

A continuous random variable X takes on values in the interval from 0 to 1. It has to follow the Beta distribution, if its probability density is given as

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, for\ 0 < x < 1, \alpha > 0, \beta > 0 \\ 0, \qquad otherwise \end{cases}$$

The mean and variance of this distribution are given by

$$\mu = \frac{\alpha}{\alpha+\beta} \ \ and\ \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Note:

If $\alpha = 1\ and\ \beta = 1$, we obtain as special case the uniform distribution.

Problem:

In a certain country, the proportion of highway sections requiring repairs in any given year is a random variable having the Beta distribution with $\alpha = 3\ and\ \beta = 2$

i. On average, what percentage of the highway sections require in any given year?
ii. Find the probability that at most half of the highway sections will require repairs in any given year.

Solution:

i. $\mu = \frac{\alpha}{\alpha+\beta} = \frac{3}{5} = 0.60$

That is on the average 60% of the highway sections require repairs in any given year.

ii. For $\alpha = 3\ and\ \beta = 2$

$$\Gamma(\alpha) = \Gamma(3) = (3-1)! = 2! = 2$$
$$\Gamma(\beta) = \Gamma(2) = (2-1)! = 1! = 1$$
$$\Gamma(\alpha+\beta) = \Gamma(5) = (5-1)! = 4! = 24$$
$$f(x) = \begin{cases} 12x^2(1-x), & for\ 0 < x < 1 \\ 0, & otherwise \end{cases}$$

Thus, the desired probability as given by

$$\int_0^{1/2} 12x^2(1-x)dx = \frac{5}{16}$$

Weibull distribution:

The random variable X is said to follow Weibull distribution, if its probability distribution is given by

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, & x > 0 \\ 0, & otherwise \end{cases}$$

Where $\alpha > 0\ and\ \beta > 0$ are two parameters of the Weibull distribution.

Note:

When $\beta = 1$, the Weibull distribution reduces to the exponential distribution with parameter α.

Mean and Variance of Weibull distribution:

The probability density function of Weibull distribution is given by

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, & x > 0 \\ 0, & otherwise \end{cases}$$

Where $\alpha > 0\ and\ \beta > 0$ are two parameters.

Mean = $E(X) = \mu = \alpha^{-\frac{1}{\beta}}\Gamma\left(1+\frac{1}{\beta}\right)$

Variance $= \sigma^2 = \alpha^{-2/\beta}\left\{\Gamma\left(1+\frac{2}{\beta}\right) - \left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right\}$

Cumulative distribution function:

$$\boldsymbol{F(x;\alpha,\beta) = \begin{cases} 1 - e^{-\alpha x^{\beta}}, & x \geq 0 \\ 0, & x < 0 \end{cases}}$$

Problem 1:

Suppose that the lifetime of a certain kind of an emergency backup battery (in hours) is a random variable X having Weibull distribution with $\alpha = 0.1$ *and* $\beta = 0.5$. Find

i. The mean lifetime of these batteries

ii. The probability that such battery will last more than 300 hours.

Solution:

i. Mean is $\mu = \alpha^{-\frac{1}{\beta}}\Gamma\left(1+\frac{1}{\beta}\right)$

$$= (0.1)^{-\frac{1}{0.5}}\Gamma\left(1+\frac{1}{0.5}\right) = \frac{2}{\left(\frac{1}{10}\right)^2} = 200\ hours$$

ii. $P(X > 300) = \int_{300}^{\infty} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}dx$

$$= \int_{300}^{\infty} (0.1)(0.5)x^{-0.5}e^{-0.1(x)^{0.5}}dx$$

$$= 0.1769$$

Problem 2:

Suppose that the time to failure (in minutes) of certain electronic components subjected to continuous vibrations may be looked upon as a random variable having the Weibull distribution with $\alpha = \frac{1}{5}$ *and* $\beta = \frac{1}{3}$

i. How long can such a component be expected to last?

ii. What is the probability that such a component will fail in less than 5 hours.

Solution:

i. Mean is $\mu = \alpha^{-\frac{1}{\beta}}\Gamma\left(1+\frac{1}{\beta}\right)$

$$= \left(\frac{1}{5}\right)^{-\frac{1}{\left(\frac{1}{3}\right)}}\Gamma\left(1+\frac{1}{\left(\frac{1}{3}\right)}\right) = 5^3 3! = 750\ minutes$$

ii. $P(X < 5\ hours) = P(X < 300\ minutes)$

$$= \int_0^{300} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}dx = \int_0^{300}\left(\frac{1}{5}\right)\left(\frac{1}{3}\right)x^{\frac{1}{3}-1}e^{-\left(\frac{1}{5}\right)(x)^{\frac{1}{3}}}dx = 0.7379$$

Or

$$F\left(300;\frac{1}{5},\frac{1}{3}\right) = \begin{cases} 1 - e^{-\alpha x^{\beta}}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$= 1 - e^{-\frac{1}{5}(300)^{\frac{1}{3}}} = 0.7379$$

The failure rate of the Weibull distribution:

- The Weibull distribution helps to determine the failure rate (or hazard rate) in order to get a sense of deterioration of the component.
- Consider the reliability of a component or product as the probability that it will function property for at least a specified time to under specified experimental conditions.
- Then the **failure rate** at time 't' for the Weibull distribution is given by

$$\boldsymbol{Z(t) = \alpha\beta t^{\beta-1}, t > 0}$$

Interpretation of the failure rate;

1. If $\beta = 1$, the failure rate=α, a constant. This is the special case of the exponential distribution in which lack of memoryless property.
2. If $\beta > 1$, $Z(t)$ is an increasing function of time ‘t’, which indicates that the component **wears over time**.
3. If $\beta < 1$, $Z(t)$ is decreasing function of time ‘t’ and hence the component strengthens or **hardness over time**.

Problem:

The length of life X, in hours of an item in a machine shop has a Weibull distribution with $\alpha = 0.01 \; and \; \beta = 2$

i. What is the probability that it fails before eight hours of usage?

ii. Determine the failure rates.

$$\boldsymbol{F(x; \alpha, \beta) = \begin{cases} 1 - e^{-\alpha x^{\beta}}, & x \geq 0 \\ 0, & x < 0 \end{cases}}$$

Solution:

i. $P(X < 8) = F(8) = 1 - e^{-(0.01)8^2} = 1 - 0.257 = 0.473$

ii. Here $\beta = 2$, and hence it wears over time and the failure rate is given by

$$Z(t) = 0.02\; t$$

If $\beta = \frac{3}{4} \; and \; \alpha = 2$, then

$$Z(t) = 1.5\; t^{1/4}$$

Hence the component gets stronger over time.

Module-6

Hypothesis Testing-I

Introduction:

- Population are often described by the distribution of their values.
- Sample is a part of population.
- Statistical measures (such as mean and variance) calculated on the basis of population are called parameters.
- Corresponding measures computed on the basis of sample observations are called statistics
- Sampling distribution: The distribution of a statistics calculated on the basis of a random sample is basic to all of statistical inference

Or

The probability distribution of the statistics that would be obtained, if the number of samples, each of same size, were infinitely large is called the sampling distribution of the statistics.

Notation:

Population Parameters	Sample Statistics
Population mean (μ)	Sample mean $(\bar{X})$
Population standard deviation (σ)	Sample standard deviation (S)
Population size (N)	Sample size (n)
Population proportion (P)	Sample proportion (p)

Standard error:

The standard deviation of the sampling distribution of a statistic is called the standard error of the statistics. It has most important in tests of hypothesis.

1. Let $X_1, X_2, X_3, \dots, X_n$ be a sample of values from a population, then the sample mean is defined by

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

and sample variance is defined by

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

Since the values of the sample mean $\overline{X}$ is determined by the values of the random variables in the sample, if it follows that $\overline{X}$ is also a random variable

$$\mu_{\overline{X}} = E\left(\frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}\right)$$

$$= \frac{1}{n}(E(X_1) + E(X_2) + \cdots + E(X_n))$$

$$= \mu$$

$$\text{Var } (\bar{X}) = \sigma_{\overline{X}}{}^2 = Var\left(\frac{X_1+X_2+X_3+\cdots+X_n}{n}\right)$$

$$= \frac{1}{n^2}(Var(X_1) + Var(X_2)+\ldots+Var(X_n))$$

$$= \frac{n\sigma^2}{n^2}$$

$$= \frac{\sigma^2}{n}$$

2. If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $\bar{X}$ is a random variable whose distribution has the mean μ and variance

$\frac{\sigma^2}{n}$ for samples from infinite population

$\frac{\sigma^2}{n}\frac{N-n}{n-1}$ for samples from a finite population of size N

Standard error of the mean

$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ (infinite population)

$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{n-1}}$ (finite population)

Sampling distribution of mean (σ *known*):

If $\overline{X}$ is the mean of a random sample of size n taken from a population having the mean μ and the finite variance σ^2, then

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

is a random variable whose distribution function approaches that of the standard normal distribution as $n \to \infty$.

Sampling distribution of mean ($\boldsymbol{\sigma\ unknown}$):

If $\overline{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the finite variance σ^2, and $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ then

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

is a random variable having **t-distribution** with parameter $\nu = n - 1$.

Sampling distribution of the variance:

If s^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$\boldsymbol{\chi^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}}$$

is a random variable having $\boldsymbol{\chi^2}$ **− distribution** with the parameter $\nu = n - 1$.

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 parameter $\nu_1 = n_1 - 1\ and\ \nu_2 = n_2 - 1$.

Hypothesis testing:

Hypothesis testing is a method for testing a claim/hypothesis about a parameter in a population using data measured in a sample.

Statistical hypothesis:

1. It is a statement or claim about one or more population parameters.
2. Hypothesis testing is formulated in terms of two hypothesis:
   H_0: The null hypothesis (this is the negation of the claim)
   H_1: The alternative hypothesis (this is the claim we wish to establish)
3. In H_0, a statement involving equality $(=, \geq, \leq)$
   In H_1, a statement involving equality $(\neq, >, <)$

The hypothesis we want to test, if H_1 is "likely" two

So, there are two possible outcomes

- Reject H_0 and accept H_1because of sufficient evidence in the sample in favor of H_1
- Do not reject H_0because of insufficient evidence to support H_1.

Critical region:

The region of rejection or critical region as the region beyond a critical value in a hypothesis test. When the value of a test statistic is in the rejection region, we decide to reject the H_0. Otherwise, will accept it.

Level of significance (LOS):

The probability that a random value $(\alpha)$of the statistic lies in the critical region is called level of significance and is usually expressed as a percentage.

Or

The total area of the critical region expressed as α % is the loss of significance.

Types of test:

Suppose we test for population mean, then

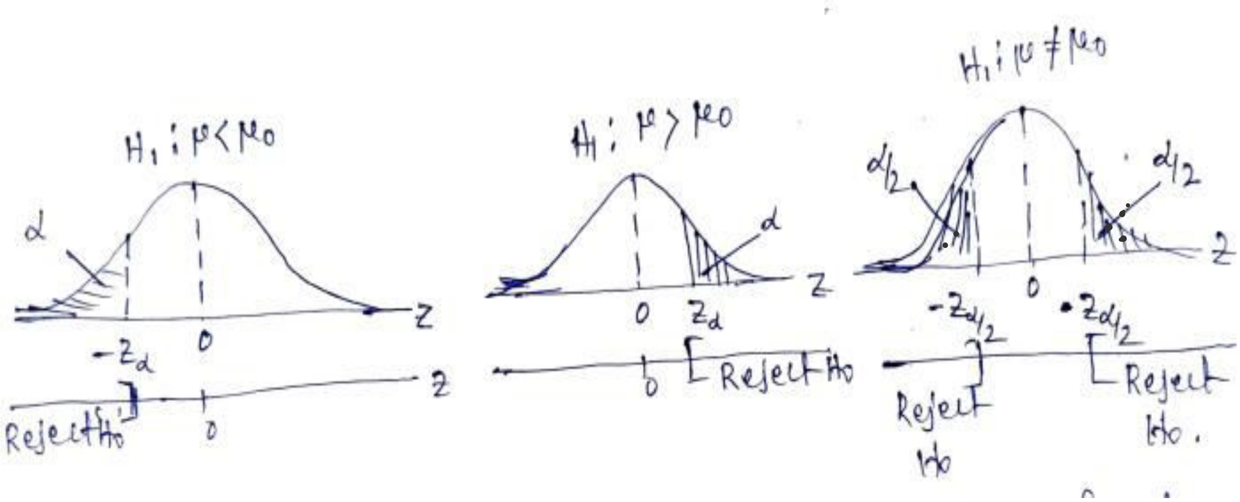
Null hypothesis $H_0: \mu = \mu_0$

Alternative hypothesis $H_1: \mu \neq \mu_0 \ or \ \mu > \mu_0 \ or \ \mu < \mu_0$

If $\mu \neq \mu_0$, then the test is called two-tailed test.

If $\mu > \mu_0$, then the test is called right-tailed test (One-tailed)

If $\mu < \mu_0$, then the test is called left-tailed test (One-tailed)



A critical value is a cutoff value that of the boundaries beyond which less than 5% of sample means can be obtained, if the Null hypothesis is true. Same **means obtained** beyond a critical value will result in a decision to reject the null hypothesis.

Level of significance (α).	**Types of test**	
	One-tailed	**Two-tailed**
5% (0.05)	+1.645 or - 1.645	$\pm$ 1.96
1% (0.01)	+2.33 or - 2.33	$\pm$ 2.58
10% (0.1)	+3.09 or - 3.09	$\pm$ 3.30

Types of error and their probabilities:

	H_0 is true	H_0 is false
Reject H_0	Type-I error	Correct decision
Accept H_0	Correct decision	Type-II error

α -Level of significance- probability of making type-I error

Or

P (type-I error) = α

Similarly, P (type-II error) = $\boldsymbol{\beta}$

Steps involved in hypothesis testing:

1. Formulate Null and alternate hypothesis
2. Identify the level of significance
3. Set the criteria for a decision
4. Compute test statistics
5. Critical or rejection region
6. Draw a conclusion or make decision

If $\boldsymbol{n \geq 30}$ is a large sample

If $\boldsymbol{n < 30}$ is a small sample

Hypothesis concerning one mean or test of single mean condition:

- Either a population is a normally distributed sample size should be large i.e., $n \geq 30$
- Population standard deviation σ should be known. If it is not known, then we can use a sample standard deviation $'s'$, instead of provided $n \geq 30$

Test of single mean condition:

- $H_0: \mu = \mu_0$

  Test statistic: Statistic for test concerning mean (σ known) is

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

  Which follows standard normal distribution

- Critical regions for testing $\mu = \mu_0$ (standard normal distribution and σ be known)

Alternative hypothesis	H_0 Reject Null hypothesis
$\mu < \mu_0$	$Z < -Z_\alpha$
$\mu > \mu_0$	$Z > Z_\alpha$
$\mu \neq \mu_0$	$Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$

Example 1:

Suppose, for instance, we want to establish that the thermal conductivity of a certain kind of cement brick differs from 0.340, the value claimed. We will test on the basis of $n = 35$ determinations and at the 0.05 level of significance. From information gathered in similar studies, we can except that the variability of such determinations is given by $\sigma = 0.01$ and mean=0.343.

Solution:

$$n = 35,\ \overline{X} = 0.340, \sigma = 0.01$$

1. Null hypothesis: $H_0: \mu = 0.340$

   Alternative hypothesis: $H_1: \mu \neq 0.340$

2. The level of significance is $\alpha = 0.05$
3. **Test Statistic:**

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

$$Z = \frac{0.340 - 0.343}{\frac{0.01}{\sqrt{35}}} = -1.77$$

4. **Critical region:**

   Since it is a two-tailed test, the critical value is $Z_{\alpha/2} = 1.96$
5. **Decision:** Since $Z = -1.77$ falls on the interval from -1.96 to 1.96, the null hypothesis can not be rejected. i.e., null hypothesis is accepted.

Example 2:

The mean lifetime of a sample of 100 tube lights produced by a company is found to be 1580 hours with standard deviation of 90 hours. Test the hypothesis at 1% loss of significance, that the mean lifetime of the tubes produced by the company is 1600 hours.

Solution:

$$n = 100,\ \overline{X} = 1580, \sigma = 90$$

1. $H_0: \mu = 1600$ against $H_1: \mu \neq 1600$ (two tailed test)

2. The level of significance is $\alpha = 0.01$
3. **Test Statistic**:

Since $n \geq 30$, Z follows standard normal distribution

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

$$= \frac{1580 - 1600}{\frac{90}{\sqrt{100}}} = -2.22$$

$$Z = -2.22$$

4. **Critical region:**

   Since it is a two-tailed test, the critical value is $Z_{\alpha/2} = 2.58$
5. **Decision:** Since $Z = -2.22$ falls on the interval from -2.58 to 2.58, the null hypothesis cannot be rejected. So, we accept the null hypothesis.

   We conclude that the mean lifetime of the tubes produced by the company is 1600 hours.

Example 3:

In a random sample of 60 workers, the average time taken by them to get to work is 33.8 minutes with a standard deviation of 6.1 minutes. Can we reject the null hypothesis $\mu = 32.6$ minutes in favor of alternative null hypothesis $\mu > 32.6$ at $\alpha = 0.025$ level of significance?

Solution:

$$n = 160,\ \overline{X} = 33.8, \sigma = 6.1$$

1. H_0: $\mu = 32.6$ against
   H_1: $\mu > 32.6$ (one-tailed test)
2. The level of significance is $\alpha = 0.025$
3. **Test Statistic**:

Since $n \geq 30$, Z follows standard normal distribution

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

$$= \frac{33.8 - 32.6}{\frac{6.1}{\sqrt{60}}} = 1.52$$

4. **Critical region:**

   Since it is a one-tailed test, the critical value is $Z_{\alpha/2} = 1.645$

5. **Decision:**

   Since $Z = 1.52$ falls on the interval from -1.645 to 1.645, so the null hypothesis cannot be rejected. So, we accept the null hypothesis.

Example 4:

A sample of 400 items is taken from a population whose standard deviation is 10. The mean of the sample is 40. Test whether the sample has come from a population with mean 38. Also calculate 95% confidence interval for the population.

Solution:

$$n = 400, \overline{X} = 40, \qquad \mu = 38, \sigma = 10$$

1. H_0: $\mu = 32.6$ against
   H_1: $\mu \neq 32.6$ (two- tailed test)
2. The level of significance is $\alpha = 0.05$
3. **Test Statistic**:

   Since $n \geq 30$, Z follows standard normal distribution

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

$$= \frac{40 - 38}{\frac{10}{\sqrt{400}}} = 4$$

4. **Critical region:**

   Since it is a two-tailed test, the critical value is $Z_{\alpha/2} = 1.96$

5. **Decision:**

Since $Z = 4$ is out of the interval from -1.96 to 1.96, so the null hypothesis is rejected.

That is, the sample is not from the population whose mean is 38.

Next, we have to find the 95% confidence interval

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$= \left(40 - 1.96.\frac{10}{\sqrt{400}}, 40 + 1.96.\frac{10}{\sqrt{400}}\right)$$

$$= (40 - 0.98, 40 + 0.98)$$

$$= (39.02, 40.98)$$

Example 5:

An insurance agent has claimed that the average age of policy holders who issue through him the average for all agents which is 30.5 years. A random sample of 100 policy holders who had issued through him gave the following age distribution.

Age	16-20	21-25	26-30	31-35	36-40
No. of persons	12	22	20	30	16

Calculate the arithmetic mean and standard deviation of this distribution and use the values to test his claim at 5% level of significance.

Solution:

Take $A = 28, d_i = X_i - A, h = 5, N = 100$

$$\overline{X} = A + \frac{h \sum f_i d_i}{N}$$

$$= 28 + \frac{5(16)}{100} = 28.8$$

$$\overline{X} = 28.8$$

The standard deviation is

$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

$$= 5\left(\sqrt{\frac{164}{100} - \left(\frac{16}{100}\right)^2}\right) = 6.35$$

$$s = 6.35$$

1. Null hypothesis H_0: The sample is drawn from a population with mean μ i.e., $\overline{X}$ and μ do not differ significantly where $\mu = 30.5$ years.
   Alternative hypothesis H_1: $\mu < 32.6$ (one- tailed test i.e., left tail test)

$$n = 100,\ \overline{X} = 28.8, \mu = 30.5, s = 6.35$$

2. The level of significance is $\alpha = 5\% = 0.05$
3. **Test Statistic**:

   Since $n \geq 30$, Z follows standard normal distribution

$$Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{28.8 - 30.5}{\frac{6.35}{\sqrt{100}}} = -2.677 \approx -2.68$$

4. **Critical region:**

   Since it is a one-tailed test, the critical value is $Z_{\alpha/2} = 1.645$
5. **Decision:**

Since $Z = -2.645$ is out of the interval from -1.645 to 1.645, so the null hypothesis is rejected at 5% level of significance.

Try yourself:

6. The mean and standard deviation of a population are 11795 and 41054, respectively. If $n = 50$, find 95% confidence interval for the mean.

7. It is claimed that a random sample of 49 tyres has a mean life of 15200 km. this sample was drawn from a population whose mean is 15150 kms and a standard deviation of 1200 km. test the significance at 0.05 level.

8. An ambulance service claims that it takes on average less than 10 minutes to reach its destination in emergency calls. A sample of 36 calls has a mean of 11 minutes and the variance of 16 minutes. Test the significance of 0.05 level.

Example 9:

Suppose that a consumer agency wishes to establish that that the population mean is less than 71 pounds, the target amount established for this product. There are $n = 80$ observations and a computer calculation give $\overline{x} = 68.45$ and $s = 9.583$.what can it conclude if the probability of a type I error is to be at most 0.01?

Solution:

H_0: $\mu \geq 71$ pounds

H_1: $\mu < 71$ pounds

The level of significance is $\alpha \leq 0.01$

Criterion:

Since the probability of a type I error is greatest when $\mu = 71$ pounds we proceed as if we were testing the H_0: $\mu = 71$ pounds against H_1: $\mu < 71$ pounds at the 0.01 level of significance.

Thus H_0 must be rejected, if $Z < -Z_\alpha$ i.e., $Z < -2.33$
where

$$Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{68.45 - 71}{\frac{9.583}{\sqrt{80}}} = -2.38$$

$$Z = -2.38$$

Decision:

Since $Z = -2.38$ is less than – 2.33, H_0 must be rejected at level of significance 0.01 or we can say, the suspicion that $\mu < 71$ pounds confirmed.

Two independent large samples($\boldsymbol{n_1 \geq 30, n_2 \geq 30}$):

1. Let $\boldsymbol{X_1, X_2, X_3, \ldots, X_n}$ is a random sample of size n_1 from population 1 which has mean=μ_1 and variance=${\sigma_1}^2$.
2. Let $\boldsymbol{Y_1, Y_2, Y_3, \ldots, Y_n}$ is a random sample of size n_2 from population 2 which has mean=μ_2 and variance=${\sigma_2}^2$.
3. Two samples $\boldsymbol{X_1, X_2, X_3, \ldots, X_n}$ and $\boldsymbol{Y_1, Y_2, Y_3, \ldots, Y_n}$ are independent
   $\boldsymbol{E}(\overline{X}) = \mu_1$ *and* $\boldsymbol{E}(\overline{Y}) = \mu_2$

$$\boldsymbol{Var}(\overline{X}) = \frac{{\sigma_1}^2}{{n_1}^2} \ \boldsymbol{and}\ \boldsymbol{Var}(\overline{Y}) = \frac{{\sigma_2}^2}{{n_2}^2}$$

$\boldsymbol{E}(\overline{X} - \overline{Y}) = \boldsymbol{\mu_1} - \boldsymbol{\mu_2} = \boldsymbol{\delta}$ (say)

$$\boldsymbol{Var}(\overline{X} - \overline{Y}) = \frac{{\sigma_1}^2}{{n_1}^2} + \frac{{\sigma_2}^2}{{n_2}^2}$$

Two-sample Z statistics is

$$Z = \frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

When the sample sizes n_1 *and* n_2 , then the statistic for large samples inferences concerning difference between two means will be

$$Z = \frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

Hypothesis test concerning two mean:

Formulation:

- If we consider $H_0: \mu_1 - \mu_2 = \delta_0$ tests of their H_0 against each of the H_1: $\mu_1 - \mu_2 < \delta_0$, $\mu_1 - \mu_2 > \delta_0$, $\mu_1 - \mu_2 \neq \delta_0$.
- The test itself will depend on the distance measured in estimated standard deviation units, from the difference in sample means $\overline{X} - \overline{Y}$ to the hypothesized value δ_0.

Test statistic for large samples concerning a difference between two means:

When $n_1 \geq 30, n_2 \geq 30$, to test $H_0 = \mu_1 - \mu_2 = \delta_0$, we will use the Z-statistic

$$Z = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Critical regions for testing $\mu_1 - \mu_2 = \delta_0$ (normal population and σ_1, σ_2 are known or large samples $n_1 \geq 30, n_2 \geq 30$)

Alternative hypothesis H_1	**H_0 Reject Null hypothesis**
$\mu_1 - \mu_2 < \delta_0$	$Z < -Z_\alpha$
$\mu_1 - \mu_2 > \delta_0$	$Z > Z_\alpha$
$\mu_1 - \mu_2 \neq \delta_0$	$Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$

Note: Somewhere H_0 can be $\mu_1 = \mu_2$, as δ_0 can be any constant.

Problem:

To test the claim that the resistance of electric wire can be reduced by more than 0.050 Ohm by alloying, 32 values obtained for standard wire yielded $\overline{X} = 0.136$ Ohm and $s_1 = 0.004$ Ohm and 32 values obtained for alloyed wire yielded $\overline{Y} = 0.083$ Ohm and $s_2 = 0.005$ Ohm. At the 0.05 level of significance, does this support the claim?

Solution:

$$H_0: \mu_1 - \mu_2 = 0.050$$

$$H_1: \mu_1 - \mu_2 > 0.050$$

The level of significance: $\alpha = 0.05$

Test Statistic:

$$Z = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

$$= \frac{0.136 - 0.083 - 0.050}{\sqrt{\frac{(0.004)^2}{32} + \frac{(0.005)^2}{32}}} = 2.65$$

Decision:

Since $Z = 2.65$ exceeds 1.96, so H_0 must be rejected.

Large sample test of the H_0 at the equality of two means:

Problem 1:

The means of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68 inches respectively. Can be the samples regarded as drawn from the same population of standard deviation 2.5 inches.

Solution:

Given that $n_1 = 1000, n_2 = 2000$, $\overline{X} = 67.5$, $\overline{Y} = 68$

Null hypothesis: the samples have drawn from the sample population of standard deviation $\sigma = 2.5$ inches. i.e., $H_0: \mu_1 = \mu_2$.

Alternative hypothesis: $H_1: \mu_1 \neq \mu_2$.

The level of significance: $\alpha = 5\%$

Test Statistic:

$$Z = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

$$= \frac{67.5 \; - \; 68 - 0}{\sqrt{\frac{(2.5)^2}{1000} + \frac{(2.5)^2}{2000}}} = -\,5.16$$

$$Z = -5.16$$

Decision:

Since $Z = -5.16$ exceeds 1.96.

Therefore, the null hypothesis is rejected at 5% level of significance.

i.e., the samples are not drawn from the same population of standard deviation 2.5 inches.

Problem 2:

In a survey of buying habits, 400 women shoppers are chosen at random in super market A located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in super market B in another section of the city, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test a 10% level of significance whether the average weekly food expenditure of the two populations of the shoppers are equal.

Solution:

$$n_1 = 400, n_2 = 400,\ \overline{X} = 250,\ \overline{Y} = 220, s_1 = 40, s_2 = 55$$

Null hypothesis: Assume that the average weekly food expenditure of the two populations of the shoppers are equal i.e., $H_0: \mu_1 = \mu_2$.

Alternative hypothesis: $H_1: \mu_1 \neq \mu_2$.

The level of significance: $\alpha = 10\%$

Test Statistic:

$$Z = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

$$= \frac{250 - 220 - 0}{\sqrt{\frac{(40)^2}{400} + \frac{(55)^2}{400}}} = 8.82$$

$$Z = 8.82$$

Decision:

Since $Z = 8.82$ exceeds 3.30.

Therefore, the null hypothesis is rejected at 10% level of significance.

i.e., the average weekly food expenditure of the two populations are not equal.

Problem 3:

Samples of students were drawn from two universities and from their weights in kilograms, mean and standard deviations are calculated and shown in below. Make a large sample test to test the level of significance between the means

	Mean	Standard deviation	Size of the sample
University A	55	10	400
University B	57	15	100

Inferences concerning Proportions:

Proportion:

A proportion refers to the fraction of the total population possesses a certain attribute.

Example:

Suppose we have a sample of four pets; a bird, a fish, a dog and a cat. If we ask what proportion has four legs, then only two pets (the dog and the cat) have four legs, therefore the proportion of pets with four legs is '$\frac{2}{4}$'or 0.5.

A proportion, denoted by $'p'$ is a parameter that describes a percentage value associated with a population.

Example:

A survey showed 83% of women in a village are illiterate, the value 0.83 is a population proportion.

Finding of sample proportion:

$X \sim B(n,p)$with mean $E(x) = nP$, variance $V(X) = nPQ$, where $Q = 1 - P$

When n is very large, then $X \sim N(nP, nPQ)$, i.e., a Normal distribution with mean 'nP' and standard deviation is $\sqrt{nPQ}$.

To form a sample proportion, divide the random variable X for the number of successes by the number of trials (n).

i.e., $p = \frac{X}{n}$ is a sample proportion in a random sample of size n.

Now,

$$\frac{X}{n} \sim N\left\{\frac{nP}{n}, \sqrt{\frac{nPQ}{n^2}}\right\}$$

$$X \sim N\left\{P, \sqrt{\frac{PQ}{n}}\right\}$$

Therefore, the test statistics $'Z'$ is given by

$$\boldsymbol{Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}}$$

Test for single Proportion (Large sample):

Condition:

$nP \geq 5 \; and \; n(1 - P) \geq 5.$

Null hypothesis: $H_0 \colon p = P_0$

Test statistics: $\boldsymbol{Z = \frac{X - nP_0}{\sqrt{P_0(1 - P_0)}}}$

$$\boldsymbol{= \frac{p - nP_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}}$$

Which is a random variable having approximately the standard normal distribution.

Critical region for testing $\boldsymbol{P = P_0}$ (Large sample):

Alternative hypothesis $\boldsymbol{H_1}$	$\boldsymbol{H_0}$ Reject Null hypothesis
$p < P_0$	$Z < -Z_\alpha$
$p > P_0$	$Z > Z_\alpha$
p$\neq P_0$	$Z < -\, Z_{\alpha/2}$ or $Z > \; Z_{\alpha/2}$

Example 1:

In a sample of 1000 people Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that the both rice and wheat are equally popular in this state at 1% level of significance?

Solution:

$$X = 540, n = 1000$$

$$p = \frac{X}{n} = \frac{540}{1000} = 0.54$$

$$P = \frac{1}{2} = 0.5$$

$$Q = 1 - P = 0.5$$

Null hypothesis: H_0: Both rice and wheat are equally popular in the state.

Alternative hypothesis: H_1: $p \neq 0.5$ (two- tailed test)

Loss of significance: $\alpha = 0.01$

Test Statistics:

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{1000}}} = 2.532$$

$$Z = 2.532$$

The tabulated value of Z at 1% level of significance for two tailed test is 2.58. Since calculated $Z = 2.532 < 2.58$.

So, we accept null hypothesis H_0 at 1% level of significance. i.e., both rice and wheat are equally likely popular in the state.

Example 2:

40 people were attacked by a disease and only 36 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease is 85% in favor of the hypothesis that it is more at 5% level of significance.

Solution:

Let X denotes the number of people attacked by disease and survived

Here $X = 36, n = 40, P = 0.85, Q = 0.15, p = 0.9$

Sample proportion is $p = \frac{X}{n} = \frac{36}{40} = 0.9$

Null hypothesis: $H_0: p = 0.85$

Alternative hypothesis: $H_1: p > 0.85$ (Right tailed test)

Loss of significance: $\alpha = 0.05$

Test Statistics: consider the conditions $nP = 40 \times 0.85 = 34 > 5$

$$n(1 - P) = 40 \times 0.15 = 6 > 5$$

$$\boldsymbol{Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.9 - 0.85}{\sqrt{\frac{0.85 \times 0.15}{40}}} = 0.8856}$$

Critical region: $\boldsymbol{Z > Z_\alpha}$

Since calculation of $Z = 0.8856 < 1.645$ i.e., $-1.645 < 0.8856 < 1.645$

Hence, we fail to reject H_0. i.e., accept H_0.

i.e., there is no statistical evidence to prove that more than 85% of the people are attacked by a disease and survived.

Example 3:

Experience had shown that 20% of a manufactured product is of the top quantity. In one day's, production of 400 articles only 50 are of top quality. Test the hypothesis at 0.05 level.

Example 4:

In a random sample of 125 cool drinkers, 68 said they prefer thumps up to Pepsi. Test the hypothesis $p = 0.5$ against the alternative hypothesis $p > 0.5$.

Test for the difference between two sample Proportions:

Let P_1 *and* P_2 be the proportions of successes in two large samples of size n_1 *and* $n_2$respectively, drawn from the sample population or from two populations with the same proportion $P_1 = P_2 = P$.

$$P_1 \sim N\left\{P, \sqrt{\frac{PQ}{n_1}}\right\} \text{ and } P_2 \sim N\left\{P, \sqrt{\frac{PQ}{n_2}}\right\}$$

Then, $P_1 - P_2 \sim N\left\{0, \sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right\}$

With mean $E(P_1 - P_2) = E(P_1) - E(P_2) = P - P = 0$

$V(P_1 - P_2) = V(P_1) + V(P_2) = PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ (since the two samples are independent)

Test statistics:

$$\boldsymbol{Z} = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where population proportion mean P is known.

If, P is not known, an unbiased estimate of P based on the both samples, given by

$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ is used in the place of P.

Critical region for two sample proportion (Large sample):

Alternative hypothesis H_1	**H_0 Reject Null hypothesis**
$p_1 - p_2 < P_0$	$Z < -Z_\alpha$
$p_1 - p_2 > P_0$	$Z > Z_\alpha$
$p_1 - p_2 \neq P_0$	$Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$

Condition: $n_1 p_1 \geq 5, n_1 q_1 \geq 5, n_2 p_2 \geq 5, n_2 q_2 \geq 5$.

Problem 1:

In a random sample of 100 men taken from village A, 60 were found to be consuming alcohol. In another sample of 200 men taken from village B, 100 were found to be consuming alcohol. Do the two villages differ significantly in respect to the proportion of men who consume alcohol?

Solution:

Let $x_1 = 60, n_1 = 100, x_2 = 100, n_2 = 200$

Same proportion $p_1 = \frac{x_1}{n_1} = \frac{60}{100} = 0.6, p_2 = \frac{x_2}{n_2} = \frac{100}{200} = 0.5$

Null hypothesis: $H_0: p_1 - p_2 = 0$

Alternative hypothesis: $H_1: p_1 - p_2 \neq 0$

Level of significance: $\alpha = 0.05$

Test statistic: under the following conditions

$n_1 p_1 = 100 \times 0.6 = 60 > 5, n_1 q_1 = 40 > 5, n_2 p_2 = 100 > 5, n_2 q_2 = 100 > 5.$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{100 \times 0.6 + 200 \times 0.5}{100 + 200} = 0.533$$

$$\boldsymbol{Z} = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.6 - 0.5}{\sqrt{0.533 \times 0.467 \times \left(\frac{1}{100} + \frac{1}{200}\right)}} = 1.6366$$

$$Z = 1.6366$$

Since $-1.96 < Z = 1.636 < 1.96$

We will fail to reject, i.e., Null hypothesis H_0 is accepted.

Problem 2:

A manufacturer of electronic equipment subjects' samples of two completing brands of transistors to an accelerated performance test. If 45 of 180 transistors of the first kind and 34 of 120 transistors of the second kind fail the test, what can conclude at the level of significance 0.05 about the difference between the corresponding sample proportions?

Solution:

Let $x_1 = 45, n_1 = 180, x_2 = 34, n_2 = 120$

Same proportion $p_1 = \frac{x_1}{n_1} = \frac{45}{180} = 0.25, p_2 = \frac{x_2}{n_2} = \frac{134}{120} = 0.283$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{45 + 34}{180 + 120} = 0.263$$

$$Q = 1 - P = 1 - 0.263 = 0.737$$

Null hypothesis: $H_0: p_1 - p_2 = 0$

Alternative hypothesis: $H_1: p_1 - p_2 \neq 0$

Level of significance: $\alpha = 0.05$

Test statistic: under the following conditions

$n_1p_1 = 180 \times 0.25 = 45 > 5, n_1q_1 = 180 \times 0.75 = 135 > 5, n_2p_2 = 120 \times 0.283 = 40 > 5, n_2q_2 = 120 \times 0.737 = 86 > 5.$

$$\boldsymbol{Z} = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.25 - 0.283}{\sqrt{0.263 \times 0.737 \times \left(\frac{1}{180} + \frac{1}{120}\right)}} = -0.647$$

$$Z = -0.647$$

Since $-1.96 < Z = -0.647 < 1.96$

We will fail to reject, i.e., Null hypothesis H_0 is accepted.

Example 3:

Random samples of 400 men and 600 women were asked whether they would like to have flyover near their residence. 200 men and 325 women were in favor of the proposal. Test the hypothesis that proportions of men and women in favor of the proposal are same at 5% level.

Solution:

Let $x_1 = 200, n_1 = 400, x_2 = 325, n_2 = 600$

Same proportion $p_1 = \frac{x_1}{n_1} = \frac{200}{400} = 0.5, p_2 = \frac{x_2}{n_2} = \frac{325}{600} = 0.541$

$$P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{200 + 325}{400 + 600} = 0.525$$

$$Q = 1 - P = 1 - 0.545 = 0.475$$

Null hypothesis: $H_0: p_1 - p_2 = 0$

Alternative hypothesis: $H_1: p_1 - p_2 \neq 0$

Level of significance: $\alpha = 0.05$

Test statistic: under the following conditions

$$\boldsymbol{Z} = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.5 - 0.541}{\sqrt{0.525 \times 0.475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.28$$

$$Z = -1.28$$

Since $-1.96 < Z = -1.28 < 1.96$

We will fail to reject, i.e., Null hypothesis H_0 is accepted.

Try yourself:

Example 4:

A cigarette manufacturing firm claims that its brand A line of cigarette outsells its brand B by 8%. If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another sample of 8% difference is valid claim.

Module 7

Hypothesis Testing-II

Small Sample test:

- If the population is normally distributed and σ is known or if σ is unknown and $n \geq 30$ then we can apply Z-test (standard Normal distribution).
- If the population is normally distributed, σ is unknown and $n < 30$, then we apply t-test (student's t distribution).

Student's t-distribution:

The probability density function of t-distribution is

$$f(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\,\Gamma\left(\frac{r}{2}\right)} \frac{1}{\left(1+\frac{t^2}{r}\right)^{\frac{r+1}{2}}}$$

where $'r'$ degrees of freedom (the number of independent values or quantities which can be assigned to statistical distribution).

Statistic for small sample test concerning one mean:

Null hypothesis: H_0: $\mu = \mu_0$

Test Statistic:

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

follows t-distribution with $n - 1$ degrees of freedom.

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

is an unbiased estimator of population standard deviation σ^2.

The relation between S and s (sample standard deviation) is

$$S = s\left(\sqrt{\frac{n}{n-1}}\right)$$

Standard error is $\frac{\sigma}{\sqrt{n}}$.

Critical region:

Level α rejection region for testing $\mu = \mu_0$ (normal population and σ unknown) one sample t-test.

Alternative hypothesis H_1	H_0 Reject Null hypothesis
$\mu < \mu_0$	$t < -t_\alpha$
$\mu > \mu_0$	$t > t_\alpha$
$\mu \neq \mu_0$	$t < -\ t_{\alpha/2}$ or t > $t_{\alpha/2}$

Where t_α *and* $t_{\alpha/2}$ are based on $'n-1'$ degrees of freedom.

Note:

The calculated value of 't' is **less than** the tabulated 't-value' for n degrees of freedom (df), accept H_0. i.e., the calculated t < tabulated t at level of significance.

Problem 1:

Scientists need to be able to detect small amounts of contaminants in the environment. As a check on current capabilities, measurements of lead content ($\mu g/L$) are taken from twelve water specimens spiked with a known concentration

2.5 2.4 2.9 2.7 2.6 2.9 2.0 2.8 2.2 2.4 2.4 2.0

Test the null hypothesis $\mu = 2.25$ against the alternative hypothesis $\mu > 2.25$ at the 0.025 level of significance.

Solution:

Null hypothesis: H_0: $\mu = 2.25$
Alternative hypothesis: H_0: $\mu > 2.25$
Level of significance: $\alpha = 0.025$
Criterion: reject H_0 if $t > 2.201$
Where 2.201 is the value of $t_{0.025}$ for $n-1 = 12-1 = 11$ degrees of freedom.

Test Statistic:

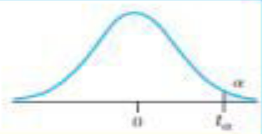
$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

From the given data $\bar{X} = 2.483, s = 0.3129, \mu = 2.25$
Then

$$t = \frac{2.483 - 2.25}{\frac{0.3129}{\sqrt{12}}} = 2.58$$

Since calculated $t = 2.58 > 2.201$, the null hypothesis must be rejected at level of significance $\alpha = 0.025$.

Table 4 Values of t_α



ν	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.00833$	$\alpha = 0.00625$	$\alpha = 0.005$	ν
1	3.078	6.314	12.706	31.821	38.204	50.923	63.657	1
2	1.886	2.920	4.303	6.965	7.650	8.860	9.925	2
3	1.638	2.353	3.182	4.541	4.857	5.392	5.841	3
4	1.533	2.132	2.776	3.747	3.961	4.315	4.604	4
5	1.476	2.015	2.571	3.365	3.534	3.810	4.032	5
6	1.440	1.943	2.447	3.143	3.288	3.521	3.707	6
7	1.415	1.895	2.365	2.998	3.128	3.335	3.499	7
8	1.397	1.860	2.306	2.896	3.016	3.206	3.355	8
9	1.383	1.833	2.262	2.821	2.934	3.111	3.250	9
10	1.372	1.812	2.228	2.764	2.870	3.038	3.169	10
11	1.363	1.796	2.201	2.718	2.820	2.891	3.106	11
12	1.356	1.782	2.179	2.681	2.780	2.934	3.055	12
13	1.350	1.771	2.160	2.650	2.746	2.896	3.012	13
14	1.345	1.761	2.145	2.624	2.718	2.864	2.977	14
15	1.341	1.753	2.131	2.602	2.694	2.837	2.947	15
16	1.337	1.746	2.120	2.583	2.673	2.813	2.921	16
17	1.333	1.740	2.110	2.567	2.655	2.793	2.898	17
18	1.330	1.734	2.101	2.552	2.639	2.775	2.878	18
19	1.328	1.729	2.093	2.539	2.625	2.759	2.861	19
20	1.325	1.725	2.086	2.528	2.613	2.744	2.845	20
21	1.323	1.721	2.080	2.518	2.602	2.732	2.831	21
22	1.321	1.717	2.074	2.508	2.591	2.720	2.819	22
23	1.319	1.714	2.069	2.500	2.582	2.710	2.807	23
24	1.318	1.711	2.064	2.492	2.574	2.700	2.797	24
25	1.316	1.708	2.060	2.485	2.566	2.692	2.787	25
26	1.315	1.706	2.056	2.479	2.559	2.684	2.779	26
27	1.314	1.703	2.052	2.473	2.553	2.676	2.771	27
28	1.313	1.701	2.048	2.467	2.547	2.669	2.763	28
29	1.311	1.699	2.045	2.462	2.541	2.663	2.756	29
inf.	1.282	1.645	1.960	2.326	2.394	2.498	2.576	inf.

Problem 2:

The height of 10 males of a given locality are found to be 70,67,62,68,61,68,70,64,64,66 inches. Is it reasonable to believe that the average height is greater than 64 inches?

Solution:

Given that $n = 10, \bar{X} = 66, s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1} = 10$

> **Null hypothesis:** $H_0: \mu = 64$
> **Alternative hypothesis:** $H_0: \mu > 64$
>
> **Level of significance:** $\alpha = 0.05$
>
> **Test Statistic:** since the population standard deviation is not known and $n < 30$, we use t-test

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{66 - 64}{\frac{\sqrt{10}}{\sqrt{10}}} = 2$$

Since calculated $t = 2 > 1.833$, the null hypothesis must be rejected at level of significance $\alpha = 0.05$.

i.e., there is no sufficient evidence to believe that the average height is greater than 64 inches.

Problem 3:

The average breaking strength of the steel rods is specified to be 18.5 thousand pounds. To test this sample of 14 rods were tested. The mean and standard deviation obtained were 17.85 and 1.955 respectively. Is the result of experiment significant?

Solution:

Given that $n = 14, \bar{x} = 17.85, \sigma = 1.955, \mu = 18.5$

> **Null hypothesis:** $H_0: \mu = 18.5$
> **Alternative hypothesis:** $H_0: \mu \neq 18.5$
>
> **Level of significance:** $\alpha = 0.05$

Test Statistic:

Since the population standard deviation is not known and $n < 30$, we use t-test

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

$$= \frac{17.85 - 18.5}{\frac{1.855}{\sqrt{14}}} = -1.311$$

Since calculated $t = -1.311 < 1.771$, the null hypothesis accepted at level of significance $\alpha = 0.05$ with 13 df.

Test of difference of means:

Null hypothesis: $H_0: \mu_1 - \mu_2 = d$

Test Statistic:

$$\boldsymbol{t} = \frac{\overline{\boldsymbol{x_1}} - \overline{\boldsymbol{x_2}}}{\sqrt{s^2\left(\frac{\boldsymbol{1}}{\boldsymbol{n_1}} + \frac{\boldsymbol{1}}{\boldsymbol{n_2}}\right)}}$$

follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

where, $s^2 = \frac{\sum(x_{1i} - \overline{x_1})^2 + \sum(x_{2i} - \overline{x_2})^2}{n_1 + n_2 - 2}$

Or

$$s^2 = \frac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2}$$

Problem 1:

Samples of two types of electric bulbs were tested for length of life and the following data were obtained

Type 1: $n_1 = 8, \overline{x_1} = 1234\ hours, s_1 = 36\ inches$

Type 2: $n_2 = 7, \overline{x_2} = 1036\ hours, s_2 = 40\ inches$

Is the difference in mean sufficient to warrant that type-1 is superior than Type 2 regarding the length of life?

Solution:

Given that

$$n_1 = 8, \overline{x_1} = 1234\ hours, s_1 = 36\ inches, n_2 = 7, \overline{x_2} = 1036\ hours, s_2 = 40\ inches$$

Null hypothesis: $H_0: \mu_1 = \mu_2$
Alternative hypothesis: $H_0: \mu_1 > \mu_2$

Level of significance: $\alpha = 0.05$

Test Statistic:

$$\boldsymbol{t} = \frac{\boldsymbol{\overline{x_1} - \overline{x_2}}}{\sqrt{s^2\left(\frac{\boldsymbol{1}}{\boldsymbol{n_1}} + \frac{\boldsymbol{1}}{\boldsymbol{n_2}}\right)}}$$

$$s^2 = \frac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2}$$

$$= \frac{8 \times 36^2 + 7 \times 40^2}{8 + 7 - 2} = 1659.07$$

$$s = 1659.07$$

$$t = \frac{(1234 - 1036)}{\sqrt{1659.07 \times \left(\frac{1}{8} + \frac{1}{7}\right)}} = \frac{198}{\sqrt{1659.07 \times 0.2678}}$$
$$= \frac{198}{\sqrt{444.39}} = 9.39$$

Follows t-distribution with 13 degrees of freedom.

Critical region:

The tabulated value of $t = 1.771$ and the calculated value of $t = 9.39$

i.e., the calculated value of $t >$ tabulated value of t

so, we reject the null hypothesis H_0 at level of significance 0.05.

Decision:

There is a statistical evidence that Type 1 is superior than Type 2.

Problem 2:

The mean height and standard deviation height of 8 randomly chosen soldiers are 166.9 and 8.29 cm respectively. The corresponding values of 6 randomly chosen sailors are 170.3 and 8.50cm respectively. Based on this data, can we conclude that soldiers are, in general, shorter than sailors?

Solution:

Given that

$$n_1 = 8, \overline{x_1} = 166.9, s_1 = 8.29, n_2 = 6, \overline{x_2} = 170.3, s_2 = 8.50$$

Null hypothesis: $H_0: \mu_1 = \mu_2$
Alternative hypothesis: $H_0: \mu_1 < \mu_2$

Level of significance: $\alpha = 0.05$

Test Statistic:

$$\boldsymbol{t} = \frac{\overline{\boldsymbol{x}_1} - \overline{\boldsymbol{x}_2}}{\sqrt{s^2\left(\frac{\mathbf{1}}{\boldsymbol{n}_1} + \frac{\mathbf{1}}{\boldsymbol{n}_2}\right)}}$$

$$s^2 = \frac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2}$$

$$= \frac{8 \times 8.29^2 + 6 \times 8.50^2}{8 + 6 - 2} = 81.940$$

$$s = 1659.07$$

$$t = \frac{(166.9 - 170.3)}{\sqrt{81.940 \times \left(\frac{1}{8} + \frac{1}{6}\right)}} = \frac{-3.4}{\sqrt{81.940 \times 0.2961}}$$

$$= \frac{-3.4}{\sqrt{24.262}} = -0.690$$

Follows t-distribution with 12 degrees of freedom.

Critical region:

The tabulated value of $t = 1.771$ and the calculated value of $t = -0.690$

i.e., the calculated value of $t <$ tabulated value of t

so, we accept the null hypothesis H_0 at level of significance 0.05.

Decision:

There is a statistical evidence that we cannot conclude that soldiers are, in general, shorter than sailors.

F-distribution:

F-distribution is used to test the equality of the variances of two populations from which two samples have been drawn.

Null hypothesis: $H_0: {\sigma_1}^2 = {\sigma_2}^2$

Test statistics:

$$F = \frac{{s_1}^2}{{s_2}^2}$$

Where, ${s_1}^2 = \frac{\sum(x_{1i}-\overline{x_1})^2}{n_1-1}$ and ${s_2}^2 = \frac{\sum(x_{2i}-\overline{x_2})^2}{n_2-1}$

Note:

- The larger among ${s_1}^2$ and ${s_2}^2$ will be the numerator.
- Here $'F'$ follows F-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom.
- The critical region value is $F_{(n_1-1, n_2-1)}$.

Level of significance α rejection region for testing $\sigma_1{}^2 = \sigma_2{}^2$:

H_1	Test statistics	Rejection H_0
${\sigma_1}^2 < {\sigma_2}^2$	$F = \frac{{s_2}^2}{{s_1}^2}$	$F > F_{\alpha,(n_1-1,n_2-1)}$
${\sigma_1}^2 > {\sigma_2}^2$	$F = \frac{{s_1}^2}{{s_2}^2}$	$F < F_{\alpha,(n_1-1,n_2-1)}$
${\sigma_1}^2 \neq {\sigma_2}^2$	$F = \frac{{s_m}^2}{{s_n}^2}$	$F \neq F_{\alpha,(n_1-1,n_2-1)}$

Problem 1:

It is desired to determine whether there is less variability in the silver plating done by company 1 than in that done by company 2. If independent random samples of size 12 of the two companies work yield $s_1 = 0.035$ mil and $s_2 = 0.062$ mil, test the null hypothesis ${\sigma_1}^2 = {\sigma_2}^2$ against the alternative hypothesis ${\sigma_1}^2 < {\sigma_2}^2$ at the 0.05 level of significance.

Solution:

Null hypothesis: H_0: ${\sigma_1}^2 = {\sigma_2}^2$

Null hypothesis: H_0: ${\sigma_1}^2 < {\sigma_2}^2$

Level of significance: $\alpha = 0.05$

Reject null hypothesis H_0, if $F > F_{\alpha,(n_1-1,n_2-1)}$ i.e., $F > F_{0.05,\ (12-1,12-1)}$

$$F > F_{0.05,\ (11,11)} = 2.85$$

Test statistics:

Here, ${s_1}^2 > {s_2}^2$

$${s_1}^2 = 0.062^2, {s_2}^2 = 0.035^2$$

$$F = \frac{{s_2}^2}{{s_1}^2}$$

$$= \frac{0.062^2}{0.035^2} = 3.14$$

Since $F = 3.14 > 2.85$, the null hypothesis must be rejected.

Problem 2:

With reference to the example dealing with the heat-producing capacity of coal from two mines

M_1: 8130 8350 8070 8390

M_2: 7950 7900 8140 7920 7840

Use the 0.01 level of significance to test whether it is reasonable to assume that the variances of the two populations sampled are equal.

Solution:

Null hypothesis: $H_0: {\sigma_1}^2 = {\sigma_2}^2$

Null hypothesis: $H_0: {\sigma_1}^2 \neq {\sigma_2}^2$

Level of significance: $\alpha = 0.01$

Reject null hypothesis H_0, if $F > F_{\alpha,(n_1-1,n_2-1)}$ i.e., $F > F_{0.01,\ (5-1,6-1)}$

$$F > F_{0.01,\ (4,5)} = 11.39$$

Test statistics:

$$where, \qquad {s_1}^2 = 15750, {s_2}^2 = 10920$$

Here, ${s_1}^2 > {s_2}^2$

$$F = \frac{{s_1}^2}{{s_2}^2}$$

$$= \frac{15750^2}{10920^2} = 1.44$$

Since $F = 1.44 < 11.4$, the null hypothesis is accepted.

Problem 3:

In one sample of 10 observations from a normal population, the sum of the squares of the deviations of the sample values from the sample mean is 102.4 and in another sample of 12 observations from another normal population, the sum of the squares of the deviations of the sample values from the sample mean is 120.5. examine whether the two normal populations have the same variance.

Solution:

Given that $n_1 = 10, n_2 = 12$

$$\sum (x - \bar{x})^2 = 102.4, \ \sum (y - \bar{y})^2 = 120.5$$

$$s_1{}^2 = \frac{\sum (x - \bar{x})^2}{n_1 - 1} = \frac{102.4}{10 - 1} = 11.37$$

$$s_2{}^2 = \frac{\sum (y - \bar{y})^2}{n_2 - 1} = \frac{120.5}{12 - 1} = 10.95$$

Null hypothesis: $H_0 : \sigma_1{}^2 = \sigma_2{}^2$

Level of significance: $\alpha = 0.05$

Reject null hypothesis H_0, if $F > F_{\alpha,(n_1-1,n_2-1)}$ i.e., $F > F_{0.01,\ (10-1,\ 12-1)}$

$$F > F_{0.05,\ (9,11)} = 2.90$$

Test statistics:

Here, $s_1{}^2 > s_2{}^2$

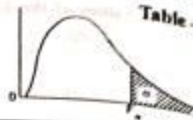
$$F = \frac{s_1{}^2}{s_2{}^2}$$

$$where, \qquad s_1{}^2 = 11.37, s_2{}^2 = 10.95$$

$$= \frac{11.37^2}{10.95^2} = 1.038$$

Since $F = 1.038 < 2.90$, the null hypothesis is accepted.

Critical Values of the F-Distribution **Table – 5**



Values of $F_{0.05}(v_1, v_2)$

v_2 \ v_1	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

(Continued) Critical Values of the F-Distribution

Values of $F_{0.05}(v_1, v_2)$

v_2 \ v_1	10	12	15	20	24	30	40	60	120	∞
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.52
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.41	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22'	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.73	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.17	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.94	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.75	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

(Continued) Critical Values of the F-Distribution

Values of $F_{0.01}(v_1, v_2)$									
	v_1								
v_2	1	2	3	4	5	6	7	8	9
1	4052	4999.5	5403	5625	5764	5859	5928	5981	6022
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	3.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

(Continued) Critical Values of the F-Distribution

Values of $F_{0.01}(\nu_1, \nu_2)$

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Chi-square distribution (or) $\chi^2 - distribution$:

The sum of k independent squared standard normal variables is Chi-square random variable with 'k'degrees of freedom i.e., $\chi^2 = Z_1^{\ 2} + Z_2^{\ 2} + Z_3^{\ 2} + \ldots + Z_k^{\ 2}$.

- The curve is non-symmetrical and skewed to the right
- The curve differs for each degrees of freedom.

Applications:

1. Hypothesis concerning one variance
2. Goodness of fit
3. Test for independence of attributes

1. Hypothesis concerning one variance:

Null hypothesis H_0: $\sigma^2 = \sigma_0^{\ 2}$

Test statistics:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^{\ 2}}$$

Where n is the sample size

s^2 is the sample variance

$\sigma_0^{\ 2}$ is the value of σ^2 given by null hypothesis.

The degrees of freedom of a χ^2 –distribution is $'n-1'$.

Critical region:

H_1	Reject H_0
$\sigma^2 < \sigma_0{}^2$	$\chi^2 < \chi^2{}_{1-\alpha}$
$\sigma^2 > \sigma_0{}^2$	$\chi^2 > \chi^2{}_{1-\alpha}$
$\sigma^2 \neq \sigma_0{}^2$	$\chi^2 < \chi^2{}_{1-\alpha/2}\ or\ \chi^2 > \chi^2{}_{1-\alpha/2}$

Problem:

A manufacturer of car batteries claims that the life of the company's batteries as approximately normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 of these batteries has a standard deviation of 1.2 years do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.

Solution:

Null hypothesis H_0: $\sigma^2 = 0.81$

Alternative hypothesis H_1: $\sigma^2 > 0.81$

Level of significance: $\alpha = 0.05$

Test statistics:

Standard deviation is $s = 1.2$, then variance is $s^2 = 1.44$

The degrees of freedom is $n - 1 = 10 - 1 = 9$

$\sigma_0{}^2 = 0.81$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0{}^2} = \frac{(10-1)1.44}{0.81} = 16$$

The tabulated value of χ^2 for 9 degrees of freedom at $\alpha = 0.05$ is 16.919.

$$\chi^2 = 16 < 16.919$$

So null hypothesis is accepted.

χ_α^2 - Critical Values of the Chi-squared Distribution with v Degrees of Freedom

v	0.30	0.25	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	29.246	30.434	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	30.319	31.528	32.912	36.741	40.113	43.194	44.140	46.963	49.645	55.476
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.993	56.893
29	32.461	33.711	35.139	39.087	42.557	45.772	46.693	49.588	52.336	58.302
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.703

2. Goodness of fit:

- Suppose we are given a set of observed frequencies obtained under some experiment and we want to test the experimental reults support a particular hypothesis or theory.
- Karl Pearson developed a test for testing the significance of discrepency between experimental (observed values) values and theoretical values (expected values) obtained under some theory or hypothesis.
- This test is known as “Chi-square test of goodness of fit”.

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

The degrees of freedom (df) for Chi-square distribution is $'n-1'$.

Note:

If the data is given in series of $'n'$ numbers, then

1. In case of Binomial distribution, $df = n - 1$
2. In case of Poisson distribution, $df = n - 2$
3. In case of Normal distribution, $df = n - 3$.

Problem 1:

The number of automobile accidents per week in a certain community are as follows: 12,8,20,2,14,10,15,6,9,4. Are these frequencies in agreement with the belief that accident conditions were the same during this 10week period.

Solution:

Expected frequency of accidents each week $= \frac{100}{10} = 10$

Null hypothesis H_0: the accident conditions were the same during the 10week period

Observed frequency (O)	**Expected frequency (E)**	$O - E$	$\frac{(O-E)^2}{E}$
12	10	2	0.4
8	10	-2	0.4
20	10	10	10
2	10	-8	6.4
14	10	4	1.6
10	10	0	0
15	10	5	2.5
6	10	-4	1.6
9	10	-1	0.1
4	10	-6	3.6
100	**100**		**26.6**

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$$

Calculated $\chi^2 = 26.6$

Here $n = 10$ observations are given, the degrees of freedom is $n - 1 = 10 - 1 = 9$

Tabulated $\chi^2 = 16.919$ at 0.05 level of significance.

Since calculated $\chi^2 >$ tabukated χ^2

Therefore, null hypothesis is rejected.

Problem 2:

A sample analysis of examination results of 500 students was made. It was found that 220 students had failed. 170 had secured a thrd class, 90 were placed in second class and 20 got a first class. Do these figures commensurate with the general examination result which is in the ratio of 4:3:2:1 for the various ctegories respectively.

Solution:

Null hypothesis H_0: the observed results commensurate with the general examination results.

Expected frequencies are in the ratio of 4:3:2:1.

Total frequency = 500.

If we divide the total frequency 500 in the ratio 4:3:2:1, we get the expected frequencies as 200,150,100,50.

Class	**Observed frequency (O)**	**Expected frequency (E)**	$O-E$	$\frac{(O-E)^2}{E}$
Failed	220	200	20	2
Third	170	150	20	2.667
Second	90	100	-10	1
First	20	50	-30	18

	500	**500**		**23.667**

$$\chi^2 = \sum\left[\frac{(\boldsymbol{O}-\boldsymbol{E})^2}{\boldsymbol{E}}\right] = 23.667$$

Calculated $\chi^2 = 23.667$

Here $n = 4$ observations are given, the degrees of freedom is $n - 1 = 4 - 1 = 3$

Tabulated $\chi^2 = 7.815$ at 0.05 level of significance.

Since calculated χ^2 >tabukated χ^2

Therefore, null hypothesis is rejected.

Problem 3:

A pair of dice are thrown 360 times and the frequency of each sum is indicated below:

Sum	2	3	4	5	6	7	8	9	10	11	12
Frequency	8	24	35	37	44	65	51	42	26	14	14

Would you say that the dice are fair on the basis of the Chi-square test at 0.05 level of significance?

Solution:

Null hypothesis $\boldsymbol{H_0}$: The dice are fair.

Alternative hypothesis $\boldsymbol{H_1}$: The dice are not fair.

Level of significance: 0.05

$n = 11$

The probabilities of getting a sum 2,3,4,5,6,7,8,9,10,11 and 12 are

X	2	3	4	5	6	7	8	9	10	11	12
$P(X)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Sum	Observed frequency (O)	Expected frequency (E) $E = 360.P(X)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
2	8	10	4	0.4
3	24	20	16	0.8
4	35	30	25	0.833
5	37	40	9	0.225
6	44	50	36	0.72
7	65	60	25	0.417
8	51	50	1	0.02
9	42	40	4	0.1
10	26	30	16	0.53
11	14	20	36	1.8
12	14	10	16	1.6
	N = **360**	**360**		**7.445**

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 7.445$$

Calculated $\chi^2 = 7.445$

Here $n = 11$ observations are given, the degrees of freedom is $n - 1 = 11 - 1 = 10$

Tabulated $\chi^2 = 19.675$ at 0.05 level of significance.

Since calculated $\chi^2 <$ tabukated χ^2

Therefore, null hypothesis is accepted.

3. Chi-square test for independence of attributes:

In general, an attribute means a quality or characteristic.

Ex: drinking, smoking, blindness, beauty, etc.

An attribute may be marked by its presence (position) or absence in a number of a given population.

Let us consider two attributes A and B. A is divided into two classes and B is divided in two classes. The various cell frequencies can be expressed in the following table known as 2x2 contingency tale.

A	a	b
B	c	d

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

The expected frequencies are given by

$E(a) = \dfrac{(a+c)(a+b)}{N}$	$E(a) = \dfrac{(b+d)(a+b)}{N}$	$a+b$
$E(a) = \dfrac{(a+c)(c+d)}{N}$	$E(a) = \dfrac{(b+d)(c+d)}{N}$	$c+d$
$a+c$	$b+d$	N (total frequency)

Note:

In this Chi-square test, we test if two attributes A and B under consideration are independent or not.

Null hypothesis H_0: Attributes are independent.

Degrees of freedom: $df = (r-1)(s-1)$

Where, $r =$ number of rows

$s =$ number of columns

Problems:

1. On the basis of information given below about the treatment of 200 patients suffering from a disease, state whether the new treatment is comparatively superior to the conventional treatment.

	Favorable	Not favorable	Total
New	60	30	90
Conventional	40	70	110

Solution:

Null hypothesis H_0: no difference between new and conventional treatment (or) new and conventional treatment are independent.

Degrees of freedom: $df = (r-1)(s-1) = (2-1)(2-1) = 1$

Where, $r =$ number of rows=2

$s =$ number of columns=2

The expected frequencies are

$\frac{(90)(100)}{200} = 45$	$\frac{(90)(100)}{200} = 45$	90
$\frac{(100)(110)}{200} = 55$	$\frac{(100)(110)}{200} = 55$	110
100	100	200

Calculation of χ^2

Observed frequency (O)	**Expected frequency (E)**	$(O-E)^2$	$\frac{(O-E)^2}{E}$
60	45	225	5
30	45	225	5
40	55	225	4.09
70	55	225	4.09
200	**200**		**18.18**

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E}\right] = 18.18$$

Calculated value of $\chi^2 = 18.18$.

Tabulated value of $\chi^2 = 3.841$ at 0.05 level of significance with 1 degrees of freedom.

Since the calculated $\chi^2 > \; tabulated\ \chi^2$.So, we reject the null hypothesis at 0.05 level of significance.

Problem 2:

Given the following contingency table for hair color and eye color. Find the value of χ^2. Is there good association between two?

		Hair color			
		Fair	**Brown**	**Black**	Total
Eye color	**Blue**	15	5	20	40
	Grey	20	10	20	50
	Brown	25	15	20	60
	Total	60	30	60	150

Solution:

Null hypothesis H_0: the two attributes, hair and eye color are independent.

Degrees of freedom: $df = (3-1)(3-1) = 4$

Where, r = number of rows=3

s = number of columns=3

The expected frequencies are

$\frac{(60)(40)}{150} = 16$	$\frac{(30)(40)}{150} = 8$	$\frac{(60)(40)}{150} = 16$	40
$\frac{(60)(50)}{150} = 20$	$\frac{(30)(50)}{150} = 10$	$\frac{(60)(50)}{150} = 20$	50
$\frac{(60)(60)}{150} = 24$	$\frac{(30)(60)}{150} = 12$	$\frac{(60)(60)}{150} = 24$	60
60	30	60	150

Calculation of χ^2

Observed frequency (O)	**Expected frequency (E)**	$(O-E)^2$	$\frac{(O-E)^2}{E}$
15	16	1	0.0625
5	8	8	1.125
20	16	16	1
20	20	0	0
10	10	0	0
20	20	0	0
25	24	1	0.042
15	12	9	0.75
20	24	16	0.665
			3.6457

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E}\right] = 3.6457$$

The tabulated value of χ^2 at 0.05 level of significance for degrees of freedom 4 is 9.488.

Since the calculated $\chi^2 < tabulated\ \chi^2$. So, we accept the null hypothesis at 0.05 level of significance.
i.e., the hair color and eye color are independent.

Design experiments:

- When comparing means across two samples, we use Z-test or t-test.
- If more than two samples are test for their means, we use ANOVA.

ANOVA:

Analysis of Variance is a hypothesis testing technique used to test the equality of two or more population means by examining the variances of samples that are taken.

Assumptions of ANOVA:

- All populations involved follow a normal distribution.
- All populations have the same variances.
- The samples are randomly selected and independent of one another or the observations are independent.

Types of ANOVA:

1. **One-way ANOVA:** Completely Randomized Design (CRD)
2. **Two-way ANOVA:** Randomized Based Design (CBD)
3. **Three-way ANOVA:** Latin Square Design (LSD)

I. Scheme for one-way classification or Completely Randomized Design (CRD):

	Observations	Mean	Sum of squares
Sample-1	$y_{11}, y_{12}, \ldots, y_{1n_1}$	$\overline{y_1}$	$\sum_{j=1}^{n_1} (y_{1j} - \overline{y_1})^2$
Sample-2	$y_{21}, y_{22}, \ldots, y_{2n_2}$	$\overline{y_2}$	$\sum_{j=1}^{n_2} (y_{2j} - \overline{y_2})^2$
. .	. .	. .	. .

.	.	.	.
Sample-i	$y_{i1}, y_{i2}, \ldots, y_{in_i}$	$\overline{y_i}$	$\sum_{j=1}^{n_i} (y_{ij} - \overline{y_i})^2$
. . . .	. . .	. . .	. . .
Sample-k	$y_{k1}, y_{k2}, \ldots, y_{kn_k}$	$\overline{y_k}$	$\sum_{j=1}^{n_k} (y_{kj} - \overline{y_k})^2$

Here, the sum of all the observations (grand total)

$$G = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

Total sample size is

$$N = \sum_{i=1}^{k} n_i$$

The overall sample mean (or grand mean) is $\bar{y} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{k} n_i} = \frac{G}{N}$

Each observation y_{ij} will be de composed as

$$y_{ij} = \bar{y} + (\overline{y_i} - \bar{y}) + (y_{ij} - \overline{y_i})$$

Taking sum of squares as measure of variation, we have to obtain

Sum of square between sample (SSB)

$$SSB = \sum_{i=1}^{k} n_i (\overline{y_i} - \bar{y})^2$$

Or we can say sum of the squared deviations of sample means from general mean (variation between sample).

Sum of squared deviation of variates from the corresponding sample means (variation within samples)

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_i})^2$$

Total variation or Total sum of squares

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2$$

Relation between all sum of squares

$$SST = SSW + SSB$$

Short cut formula:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} {y_{ij}}^2 - C$$

$$SSB = \sum_{i=1}^{k}\frac{{T_i}^2}{n_i} - C$$

Where, C is called the **correction factor** for the mean is given by

$$C = \frac{G^2}{N}, N = \sum_{i=1}^{k} T_i,\ T_i = \sum_{j=1}^{n_i} y_{ij}$$

Test statistics:

- To test the H_0that K population mean is equal, we shall compare two estimates of σ^2.

  One based on the variation **between** the sample mean.

  One based on the variation **within** the sample mean.

- Each sum of squares first converted to a mean square

$$\textbf{Mean square} = \frac{\boldsymbol{sum\ of\ squares}}{\boldsymbol{degrees\ of\ freedom}}$$

  Mean of sum of squares **between** sample

$$\textbf{MSB} = \frac{\boldsymbol{SSB}}{\boldsymbol{DF_{between}}} = \frac{\sum_{i=1}^{k} n_i\left(\overline{y_i} - \bar{y}\right)^2}{\boldsymbol{K-1}}$$

  Mean sum of squares **within** sample

$$\textbf{MSW} = \frac{\boldsymbol{SSW}}{\boldsymbol{DF_{within}}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y_i}\right)^2}{\boldsymbol{N-K}}$$

- Test statistic:

$$F = \frac{MSB}{MSW}$$

F-distribution follows $\boldsymbol{K-1}$ and $\boldsymbol{N-K}$ degrees of freedom.

ANOVA table:

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Between groups Within groups	$\boldsymbol{K-1}$ $\boldsymbol{N-K}$	SSB SSW	MSB MSW	$F = \frac{MSB}{MSW}$
Total	$\boldsymbol{N-1}$	SST		

Decision:

If $F > F_{\alpha,(N-1,N-K)}$, reject the null hypothesis H_0.

Problem 1:

Compare the means of these groups

I	II	III
1	2	2
2	4	3
5	2	4

Solution:

I	II	III
1	2	2
2	4	3
5	2	4
Total = 8	**8**	**9**

Null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$

Alternative hypothesis H_1: At least there is one difference among the means.

Level of significance:

$\alpha = 0.05$

Degrees of freedom:

$\boldsymbol{DF_{between} = K - 1 = 3 - 1 = 2}$

$\boldsymbol{DF_{within} = N - K = 9 - 3 = 6}$

$F_{\alpha,(N-1,N-K)} = F_{0.05,(2,6)} = 5.14$

sum of all the observations (grand total) $= G$

$$G = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = 1 + 2 + 5 + 2 + 4 + 2 + 2 + 3 + 4 = 25$$

Correction factor is $C = \frac{G^2}{N} = \frac{625}{9} = 69.444$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} {y_{ij}}^2 - C = (1^2 + 2^2 + 5^2 + 2^2 + 4^2 + 2^2 + 2^2 + 3^2 + 4^2) - 69.444$$
$$= 13.556$$

Sum of the squares between

$$SSB = \sum_{i=1}^{k} \frac{{T_i}^2}{n_i} - C = \left(\frac{8^2}{3} + \frac{8^2}{3} + \frac{9^2}{3}\right) - 69.444 = 69.667 - 69.444 = 0.223$$

Sum of the squares within

$$SST = SSW + SSB$$

Then, $SSW = SST - SSB = 13.556 - 0.223 = 13.333$

Mean sum of the squares

$$MSB = \frac{SSB}{DF_{between}} = \frac{0.223}{2} = 0.115$$

$$MSW = \frac{SSW}{DF_{within}} = \frac{13.333}{6} = 2.222$$

ANOVA table:

Source of variation	Degrees of freedom	Sum of Squares (SS)	Mean Squares (MS)	F
Between groups Within groups	$K-1=3-1=2$ $N-K=9-3=6$	SSB=0.223 SSW=13.333	MSB=0.115 MSW=2.222	$F=\frac{MSB}{MSW}=0.0517$
Total	$N-1=9-1=8$	SST=13.356		

Since, calculated F < tabulated F i.e., 0.0517<5.14.

So, we accept the Null hypothesis.

i.e., there is no significant difference between the means of groups.

Problem 2:

In a tin coating laboratory, the weights of 12 disks and that results are as follows:

Laboratory A	Laboratory B	Laboratory C	Laboratory D
0.25	0.18	0.19	0.23
0.27	0.28	0.25	0.30
0.22	0.21	0.27	0.28
0.30	0.23	0.24	0.28
0.27	0.25	0.18	0.24
0.28	0.20	0.26	0.34
0.32	0.27	0.28	0.20
0.24	0.19	0.24	0.18
0.31	0.24	0.25	0.24
0.26	0.22	0.20	0.28
0.21	0.29	0.21	0.22
0.28	0.16	0.19	0.21

Construct an ANOVA table.

Solution:

$\boldsymbol{K=4, N=48}$

Level of significance:

$$\alpha = 0.05$$

Degrees of freedom:

$$\boldsymbol{DF_{between} = K-1 = 4-1 = 3}$$

$$\boldsymbol{DF_{within} = N-K = 48-4 = 44}$$

Sum of all the observations (grand total) $= G$

$$G = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = 11.69$$

Correction factor is $C = \frac{G^2}{N} = \frac{11.69^2}{48} = 2.8470$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} {y_{ij}}^2 - C = (0.25^2 + 0.27^2 + \ldots + 0.21^2) - 2.8470 = 0.0809$$

Sum of the squares between

$$SSB = \sum_{i=1}^{k} \frac{{T_i}^2}{n_i} - C = \left(\frac{3.21^2}{12} + \frac{2.72^2}{12} + \frac{2.76^2}{12} + \frac{3^2}{12}\right) - 2.8470 = 0.0130$$

Sum of the squares within

$$SST = SSW + SSB$$

Then, $SSW = SST - SSB = 0.0809 - 0.0130 = 0.0679$

Mean sum of the squares

$$\boldsymbol{MSB = \frac{SSB}{DF_{between}} = \frac{0.0130}{4-1} = 0.0043}$$

$$\boldsymbol{MSW = \frac{SSW}{DF_{within}} = \frac{0.0679}{48-4} = 0.015}$$

ANOVA table:

Source of variation	Degrees of freedom	Sum of Squares (SS)	Mean Squares (MS)	F
Between groups Within groups	$\boldsymbol{K-1=4-1=3}$ $\boldsymbol{N-K=48-4=44}$	SSB=0.0130 SSW=0.0679	MSB=0.0043 MSW=0.0015	$F=\frac{MSB}{MSW}=2.87$
Total	$\boldsymbol{N-1=48-1=47}$	SST=0.0809		

$F_{\alpha,(N-1,N-K)} = F_{0.05,(3,44)} = 2.84$

Since, calculated F < tabulated F i.e., 2.87>2.84.

So, we reject the Null hypothesis.

i.e., we conclude that the laboratories are not obtaining consistent results.

Two-Way ANOVA:

Two-way ANOVA compares the means of population that are classified in two ways or the mean responses in two-factor experiments.

Randomized Block Design:

The arrangement of two-way classification.

	Blocks (columns)							
	B_1	B_2	. . .	B_j	. . .	B_r	Means	Total
Treatment 1	y_{11}	y_{12}		y_{1j}	…	y_{1r}	$\overline{y_{1.}}$	$T_{1.}$
Treatment 2	y_{21}	y_{22}	…	y_{2j}	…	y_{2r}	$\overline{y_{2.}}$	$T_{2}.$
. . .	. . .							
Treatment i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ir}	$\overline{y_{i}}.$	$T_{i}.$
.	.							

. .	. .							
Treatment C	y_{C1}	y_{C2}	…	y_{Cj}	…	y_{Cr}	$\overline{y_{C.}}$	$T_{C.}$
Means	$\overline{y_{.1}}$	$\overline{y_{.2}}$		$\overline{y_{.J}}$		$\overline{y_{.r}}$	$\overline{y_{..}}$	$T_{..}$
Total	$T_{.1}$	$T_{.2}$		$T_{.j}$		$T_{.r}$		

Where,

y_{ij} – the observation pertaining to the ith treatment and the jth block (column)

$\overline{y_{\iota}}.$ – mean of the 'r' observations for ith treatment

$\overline{y_{J}}.$ – mean of the 'C' observations for jth block

$\overline{y_{..}}$ – the grand mean of all 'rC' observations.

Note:

The dot is used for the mean is obtained by summing over the subscripts.

Model equation for randomized block design:

$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, for $i = 1,2,\dots,C$

$$j = 1,2,\dots,r$$

where,

$\mu =$ grand mean

$\alpha_i =$ mean, due to the effect of the ith treatment or between the sample

$\beta_j =$ mean, due to the effect of the jth block

$\varepsilon_{ij} =$ error, within the sample deviation.

Hypothesis for two-way ANOVA:

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_C,\ H_1: \alpha_1 \neq \alpha_2 \neq \ldots \neq \alpha_C$

$H_0: \beta_1 = \beta_2 = \ldots = \beta_r,:\ H_1: \beta_1 \neq \beta_2 \neq \ldots \neq \beta_r$

Or

$H_0: \mu_1 = \mu_2 = \ldots = \mu_C$ (columns)

$H_0: \mu_1 = \mu_2 = \ldots = \mu_r$ (rows)

Or

There is at least one mean in the columns which differ from others. Also, in rows.

Degrees of freedom:

$DF_{(column\ or\ treatments)} = C - 1$

$DF_{(row\ or\ block)} = r - 1$

$DF_{(error\ or\ within\ sample)} = Cr - 1$

Critical values:

$F_{\alpha,\ (C-1,(C-1)(r-1))}$ and $F_{\alpha,\ (r-1,(C-1)(r-1))}$

Like one-way ANOVA, we estimate for the σ^2 comparing

Variance among treatments (or between sample)

Variance among blocks, and

Measuring the experimental error or variation within samples.

Identity for analysis of two-way ANOVA classification:

$$\sum_{i=1}^{C}\sum_{j=1}^{r}\left(y_{ij}-\overline{y_{..}}\right)^2=\sum_{i=1}^{C}\sum_{j=1}^{r}\left(y_{ij}-\overline{y_{i.}}-\overline{y_{.j}}+\overline{y_{..}}\right)^2+r\sum_{i=1}^{C}\left(\overline{y_{i.}}-\overline{y_{..}}\right)^2$$
$$+C\sum_{j=1}^{r}\left(\overline{y_{.j}}-\overline{y_{..}}\right)^2$$

Each observation y_{ij} will be decomposed as

$$y_{ij}=\overline{y}_{..}+(\overline{y}_{i.}-\overline{y}_{..})+\left(y_{ij}-\overline{y}_{i.}-\overline{y_{.j}}+\overline{y}_{..}\right)$$

Sum of squares for two-way ANOVA:

Treatment sum square, $SS(Tr)=r\ \sum_{i=1}^{C}(\overline{y}_{i.}-\overline{y}_{..})^2$

Or

Short cut formula

$$\boldsymbol{SS(Tr)=\frac{\sum_{i=1}^{C}T_{i.}{}^2}{c}-\ Correction\ factor}$$

Block sum square, $SS(Bl)=C\ \sum_{j=1}^{r}\left(\overline{y_{.j}}-\overline{y}_{..}\right)^2$

Or

Short cut formula

$$\boldsymbol{SS(Bl)=\frac{\sum_{j=1}^{r}T_{.j}{}^2}{r}-\ Correction\ factor}$$

Error sum of square, $SSE=\sum_{i=1}^{C}\sum_{j=1}^{r}\left(y_{ij}-\overline{y}_{i.}-\overline{y_{.j}}+\overline{y}_{..}\right)^2$

Total sum of square, $SST=\sum_{i=1}^{C}\sum_{j=1}^{r}\left(y_{ij}-\overline{y}_{..}\right)^2$

Short cut formula

$$\boldsymbol{SST} = \sum_{i=1}^{C}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{..}\right)^2 - \boldsymbol{Correction\ factor}$$

Where, correction factor is given by $\boldsymbol{C} = \frac{\boldsymbol{T_{..}}^2}{\boldsymbol{Cr}}$

$T_{i.}$ = the sum of the r observations for the ith treatment

$T_{.j}$ = the sum of the C observations for the jth block

$T_{..}$ = the grand total of all observations

F-ratio for treatment or between sample

$$\boldsymbol{F_{Tr}} = \frac{\boldsymbol{MS(Tr)}}{\boldsymbol{MSE}} = \frac{\left(\frac{\boldsymbol{SS(Tr)}}{\boldsymbol{C-1}}\right)}{\left(\frac{\boldsymbol{SSE}}{\boldsymbol{(C-1)(r-1)}}\right)}$$

Decision: reject for H_0, if $F_{Tr} > F_{(C-1,(C-1)(r-1))}$

F-ratio for blocks

$$\boldsymbol{F_{Bl}} = \frac{\boldsymbol{MS(Bl)}}{\boldsymbol{MSE}} = \frac{\left(\frac{\boldsymbol{SS(Bl)}}{\boldsymbol{r-1}}\right)}{\left(\frac{\boldsymbol{SSE}}{\boldsymbol{(C-1)(r-1)}}\right)}$$

Decision: reject for H_0, if $F_{Bl} > F_{\alpha,\ (r-1,(C-1)(r-1))}$

Two-way ANOVA table for results

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Treatments	$\boldsymbol{r-1}$	SS(Tr)	$MS(Tr) = \frac{SS(Tr)}{r-1}$	$F_{Tr} = \frac{MS(Tr)}{MSE}$
Blocks	$\boldsymbol{C-1}$	SS(Bl)		

Error	$(C-1)(r-1)$	SSE	$MS(Bl) = \frac{SS(Bl)}{C-1}$ $MSE = \frac{SSE}{(r-1)(C-1)}$	$F_{Bl} = \frac{MS(Bl)}{MSE}$
Total	$(Cr-1)$	SST		

Problem 1:

An experiment was designed to study the performance 4 different detergents for cleaning fuel injectors. The following "cleanness" reading were obtained with specially designed equipment for 12 tanks of gas distributed over 3 different models of engines

	Engine 1	Engine 2	Engine 3	**Total**
Detergent A	45	43	51	**139**
Detergent B	47	46	52	**145**
Detergent C	48	50	55	**153**
Detergent D	42	37	49	**128**
Total	**182**	**176**	**207**	**565**

	Engine 1	Engine 2	Engine 3
Detergent A	45	43	51
Detergent B	47	46	52
Detergent C	48	50	55
Detergent D	42	37	49

Looking at the detergents as treatments and the engines as blocks, obtain the appropriate ANOVA table and test the 0.01 level of significance whether there are differences in the detergents or in the engines.

Solution:

Null hypothesis $H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$

$$\beta_1 = \beta_2 = \beta_3 = 0$$

Alternative hypothesis $H_1: \alpha_1 \neq \alpha_2 \neq \alpha_3 \neq \alpha_4 \neq 0$

$$\beta_1 \neq \beta_2 \neq \beta_3 \neq 0$$

The level of significance: $\alpha = 0.01$

$a - 1 = 4 - 1 = 3$ and $(b - 1) = 3 - 1 = 2$

$(a - 1)(b - 1) = (4 - 1)(3 - 1) = 6$

Reject H_0 for the treatment, if $F_{(tr)} > F_{0.01,\ (a-1,(a-1)(b-1))}$

$$= F_{0.01,\ (4-1,(4-1)(3-1))} = F_{0.01,\ (0.01,\ (3,6))} = 9.78$$

Reject H_0 for the block, if $F_{(Bl)} > F_{0.01,\ (b-1,(a-1)(b-1))}$

$$= F_{0.01,\ (3-1,(4-1)(3-1))} = F_{0.01,\ (0.01,\ (2,6))} = 10.92$$

Calculation:

$$a = 4, b = 3$$

$$T_{1.} = 139$$

$$T_{2.} = 145$$

$$T_{3.} = 153$$

$$T_{4.} = 128$$

And

$$T_{.1} = 182$$

$$T_{.2} = 176$$

$$T_{.3} = 207$$

$T_{..} = 565$

$\sum\sum {y_{ij}}^2 = 45^2 + 47^2 + 48^2 + 42^2 + \ldots + 49^2 = 26867$

$Correction\ factor\ is\ C = \frac{{T_{..}}^2}{ab} = \frac{565^2}{4\times3} = 26602$

$$SST = \sum_{i=1}^{C}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{..}\right)^2 - \ Correction\ factor$$

$SST = (45^2 + 47^2 + 48^2 + 42^2 + \ldots + 49^2) - 26602 = 26867 - 26602 = 265$

$SS(Tr) = \frac{\sum_{i=1}^{C} {T_{i.}}^2}{b} - \ Correction\ factor$

$= \frac{139^2 + 145^2 + 153^2 + 128^2}{3} - \ 26602 = 110.917$

$SS(Bl) = \frac{\sum_{i=1}^{r} {T_{.j}}^2}{a} - \ Correction\ factor$

$= \frac{182^2 + 176^2 + 207^2}{4} - 26602 = 135.167$

$SSE = 265 - (111 + 135) = 18.833$

Two-way ANOVA table

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Detergents	$\boldsymbol{a - 1 = 4 - 1 = 3}$	SS(Tr)=110.917	$MS(Tr) = \frac{SS(Tr)}{a-1} = \frac{110.917}{3} = 36.972$	$F_{Tr} = \frac{MS(Tr)}{MSE} = 11.78$

Engines Error	$\boldsymbol{b-1}$ $\boldsymbol{=3-1}$ $\boldsymbol{=2}$ $(a-1)(b-1)=6$	SS(Bl)=135.167 SSE=18.833	$MS(Bl)=\frac{SS(Bl)}{b-1}=\frac{135.167}{2}=67.583$ $MSE=\frac{SSE}{(a-1)(b-1)}=\frac{18.833}{6}=3.139$	$F_{Bl}=\frac{MS(Bl)}{MSE}=21.53$
Total	$(ab-1)=(12-1)=11$	SST=264.917		

Decision:

$F_{(tr)} > F_{0.01,\ (a-1,(a-1)(b-1))}$

$$F_{(tr)} > F_{0.01,\ (4-1,(4-1)(3-1))}$$

$$11.78 > 9.78$$

Reject the null hypothesis for treatment.

$$F_{(Bl)} > F_{0.01,\ (3-1,(4-1)(3-1))}$$

$$21.53 > 10.92$$

Reject the null hypothesis for Blocks.

We conclude that there are differences in the effectiveness of the 4 detergents. Also, the differences among the results obtained for the 3 engines are significant. There is an effect due to the engines, so blocking was important.

Latin Square Design (LSD) (or) Three-way ANOVA:

- In addition to rows and columns, we need to consider an extra factor known as treatments.
- Every treatment occurs only once in each row and in each column. Such a layout is known as Latin Square Design.

  Ex:

  If we are interested in studying the effects of n types of fertilizers on a yield of a certain variety of wheat, we conduct the experiment on a square field with n^2 plots of equal area and associate treatments with different fertilizers; row and column effects with variations in fertility or soil.

Procedure of LSD:

Null hypothesis: There is no significant difference in the means of columns (Groups), rows (Blocks), and treatments.

Alternative hypothesis: There is at least one mean in column which differs from others. Also, there is at least one mean in the rows which differs from others. Similarly, for treatments.

Degrees of freedom:

$\boldsymbol{DF_{rows} = n - 1}$

$\boldsymbol{DF_{columns} = n - 1}$

$\boldsymbol{DF_{treatments} = n - 1}$

$\boldsymbol{DF_{Error} = (n - 1)(n - 2)}$

Critical region:

$$F_{(n-1,(n-1)(n-2))}$$

$$G = \sum\sum x_{ij}$$

Correction factor is $C.F = \frac{G^2}{N}$

Sum of squares total:

$$SST. = \sum\sum x_{ij}{}^2 - C.F$$

sum of squares:

$$SSC = \sum \frac{C_j{}^2}{n} - CF$$

Where, C_j is the column sum of the jth column.

$$SSR = \sum \frac{R_i{}^2}{n} - CF$$

Where, R_i is the row sum of the ith row.

$$SSTr = \sum \frac{T_i{}^2}{n} - CF$$

Where, T_i is called the treatment sum of ith treatment.

$$SSE = SST - SSR - SSC - SSTr$$

ANOVA table:

Source of variation	Sum of Squares (SS)	Degrees of freedom	Mean squares (MS)	*F*
Columns	SSC	$n-1$	$\frac{SSC}{n-1}$	$F_1 = \frac{MSC}{MSE}$
Rows	SSR	$n-1$	$\frac{SSR}{n-1}$	$F_2 = \frac{MSR}{MSE}$
Treatments	SSTr	$n-1$	$\frac{SSTr}{n-1}$	$F_3 = \frac{MSTr}{MSE}$

Error	SSE	$(n-1)\,(n-2)$	$\frac{SSE}{(n-1)\,(n-2)}$	

Problem:

Analyze the variance in the Latin Square Design of yields (in Kgs) of paddy where P, Q, R, S denote the different methods of cultivation

S122	P121	R123	Q122
Q124	R123	P122	S125
P120	Q119	S120	R121
R122	S123	Q121	P122

Solution:

Null hypothesis:

There is no significance difference in the means of columns, rows and treatments (methods of cultivation).

Alternative hypothesis:

There is at least one mean in the columns which differs from others. Also, there is at least one mean in the rows which are differs from the others. Similarly, for treatments

Here, $n = 4$

Degrees of freedom:

$$\boldsymbol{DF_{rows} = n - 1 = 3}$$

$$\boldsymbol{DF_{columns} = n - 1 = 3}$$

$$\boldsymbol{DF_{treatments} = n - 1 = 3}$$

$$\boldsymbol{DF_{Error} = (n-1)(n-2) = 6}$$

Level of significance: $\boldsymbol{\alpha = 0.05}$

Critical region: $F_{(3,6)} = 4.76$

Test Statistics:

S 2	P 1	R 3	Q 2	**8**
Q 4	R 3	P 2	S 5	**14**
P 0	Q - 1	S 0	R 1	**0**
R 2	S 3	Q 1	P 2	**8**
8	**6**	**6**	**10**	**G=30**

Treatment Sum:

Sum of the treatments are $P = 5, Q = 6, R = 9, S = 10$

$G = 30, N = 16$

Correction factor is $C.F = \frac{G^2}{N} = \frac{30^2}{16} = 56.25$

$$SST = \sum\sum {x_{ij}}^2 - C.F = (2^2 + 1^2 + 2^2 + \ldots + 2^2) - 56.25 = 92 - 56.25 = 35.75$$

$$SSR = \sum \frac{{R_i}^2}{n} - CF = \frac{8^2}{4} + \frac{14^2}{4} + \frac{0^2}{4} + \frac{8^2}{4} = 24.75$$

$$SSC = \sum \frac{{RC_j}^2}{n} - CF = \frac{6^2}{4} + \frac{6^2}{4} + \frac{6^2}{4} + \frac{10^2}{4} = 2.75$$

$$SSTr = \sum \frac{{T_i}^2}{n} - CF = \frac{5^2}{4} + \frac{6^2}{4} + \frac{9^2}{4} + \frac{10^2}{4} - 10.25 = 4.25$$

$$SSE = SST - SSR - SSC - SSTr$$

$$= 35.75 - 24.75 - 2.75 - 4.25 = 4$$

Source of variation	Sum of Squares (SS)	Degrees of freedom	Mean squares (MS)	F
Columns	SSC=2.75	3	$\frac{SSC}{n-1} = 0.917$	$F_1 = \frac{MSC}{MSE} = \frac{0.917}{0.667} = 1.375$
Rows	SSR=24.75	3	$\frac{SSR}{n-1} = 8.25$	$F_2 = \frac{MSR}{MSE} = \frac{8.25}{0.667} = 12.36$
Treatments	SSTr=4.25	3	$\frac{SSTr}{n-1} = 1.417$	$F_3 = \frac{MSTr}{MSE} = \frac{1.417}{0.667} = 2.124$
Error	SSE	6	$\frac{SSE}{(n-1)\,(n-2)} = 0.667$	

Decision:

Comparing F_1, F_2, F_3 with critical region $F_{(3,6)} = 4.76$, we accept null hypothesis (columns), accept null hypothesis (treatments), reject null hypothesis (rows).