Unit- III

Curve fitting and Sampling Theory

by Dr. Mukta Garg

Curve fitting :

To find a relationship between the set of paired observations x and y (say) we plot their corresponding values on the graph taking one of the variables along the x-axis and other along the y-axis, i.e. $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

The resulting diagram showing a collection of dots is called a scatter diagram. A smooth curve that approximates the above set of points is known as the approximating curve.

Several equations of different types can be obtained to express the given data approximately. But the problem is to find the equation of the curve of 'best fit' which may be most suitable for predicting the unknown values. The process of finding such an equation of 'best fit' is known as <u>curve-fitting</u>.

Principle of least squares

The principle of least squares provides an elegant procedure for fitting a unique curve to a given data.

Let the curve

$$y = a_1 + a_2 x + a_3 x^2 + \cdots + a_m x^{m-1}$$

be fitted to the set of n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Now we have to determine the constants $a_1, a_2, \ldots$ such that it represents the curve of best fit.

<u>Case 1</u>: If $n = m$, then substituting the values (x_i, y_i) in ①, we get n equation from which a unique set of n constants can be determined.

<u>Case 2</u>: If $n > m$, then no. of equations are more than no. of constants & we may not get unique constants. So we apply principle of least squares.

At $x = x_i$, compute ~~observed~~ expected (or calculated) value

$$\eta_i = a_1 + a_2 x_i + a_3 x_i^2 + \cdots + a_m x_i^{m-1}$$

and the observed (or experimental) value be y_i. Calculate error at $x = x_i$, say $e_1, e_2, \ldots, e_n$.

$\therefore$ let $\boxed{E = e_1^2 + e_2^2 + \cdots + e_n^2}$.

Then the curve of best fit is that for which E is minimum. This is known as the <u>principle of least squares</u>.

Method of least squares

Suppose it is required to fit the curve

$$y = a + bx + cx^2$$

to a given set of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_5, y_5)$. For any x_i, the observed value is y_i and the expected value is $\eta_i = a + bx_i + cx_i^2$ so that the error $e_i = y_i - \eta_i$

$\therefore$ the sum of the squares of these errors is

$$E = e_1^2 + e_2^2 + \ldots + e_5^2$$

$$= [y_1 - (a + bx_1 + cx_1^2)]^2 + \ldots + [y_5 - (a + bx_5 + cx_5^2)]$$

$\hookrightarrow$ (a function of a, b and c)

For E to be minimum, we have

$$\frac{\partial E}{\partial a} = 0$$

$$\Rightarrow -2[y_1 - (a + bx_1 + cx_1^2)] - \ldots - 2[y_5 - (a + bx_5 + cx_5^2)] = 0 \quad \text{— (2)}$$

$$\frac{\partial E}{\partial b} = 0$$

$$\Rightarrow -2x_1[y_1 - (a + bx_1 + cx_1^2)] - \ldots - 2x_5[y_5 - (a + bx_5 + cx_5^2)] = 0 \quad \text{— (3)}$$

$$\frac{\partial E}{\partial c} = 0$$

$$\Rightarrow -2x_1^2[y_1 - (a + bx_1 + cx_1^2)] - \ldots - 2x_5^2[y_5 - (a + bx_5 + cx_5^2)] = 0 \quad \text{— (4)}$$

$$(2) \Rightarrow y_1 + y_2 + y_3 + y_4 + y_5 = 5a + b(x_1 + x_2 + \ldots + x_5) + c(x_1^2 + x_2^2 + \ldots + x_5^2)$$

$$\Rightarrow \boxed{\Sigma y_i = 5a + b\Sigma x_i + c\Sigma x_i^2}$$

Similarly,

③ $\Rightarrow$ $$\boxed{\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3}$$

and

④ $\Rightarrow$ $$\boxed{\Sigma x_i^2 y_i = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4}$$

These equations can be solved as simultaneous equations in a, b, c. The values of these constants when substituted in ① give the desired curve of best fit.

Procedure

a) To fit the straight line: $y = a + bx$

(i) Substitute the observed set of n values in this equation.

(ii) Form normal equations for each constant

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

b) To fit the parabola: $y = a + bx + cx^2$

(i) Form the normal eq'ns

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

(ii) Solve these as simultaneous eq'ns for a, b, c.

(iii) Substitute these values in $y = a + bx + cx^2$, which is the required parabola of best fit.

In general, the curve $y = a + bx + cx^2 + \dots + kx^{m-1}$ can be fitted to a given data by writing m normal equations.

Q1 Fit a straight line trend by the method of least squares to the following data:

Soln:

Year (x)	u $x-1994$	y Milk consumption	v $y-112$	u^2	uv
1990	-4	102.3	-9.7	16	38.8
1991	-3	101.9	-10.1	9	30.3
1992	-2	105.8	-6.2	4	12.4
1993	-1	112.0	0	1	0
1994	0	114.8	2.8	0	0
1995	1	118.7	6.7	1	6.7
1996	2	124.5	12.5	4	25.0
1997	3	102.9	-9.1	9	-27.3
	-4		-13.1	44	85.9

[Remark: The calculations get simplified when the central value of x is zero. It is therefore, advisable to make the central value zero if values are large]

Let the straight line to be fitted be

$$v = a + bu,$$

where a & b are to be determined using

$$\Sigma v = na + b\Sigma u$$

$$\Sigma uv = a\Sigma u + b\Sigma u^2$$

Substituting the values of Σv, Σu etc,

$$-13.1 = 8a - 4b$$

$$85.9 = -4a + 44b$$

$$\Rightarrow \quad b = 1.89, \quad a = -0.69$$

$\therefore$ equation of straight line is

$$v = -0.69 + 1.89u$$

$$\Rightarrow \boxed{y - 112 = -0.69 + 1.89(x - 1994)}$$

Q2 Fit a straight line $y = a + bx$ to the following data by the method of least squa

x :	0	1	3	6	8
y :	1	3	2	5	4

x	y	x^2	xy
0	1	0	0
1	3	1	3
3	2	9	6
6	5	36	30
8	4	64	32
18	15	110	71

normal equations are

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

$$\Rightarrow \quad 15 = 5a + 18b$$
$$71 = 18a + 110b$$

$$\Rightarrow \quad a = 1.6, \quad b = 0.38$$

$\therefore$ equation of straight line is

$$\boxed{y = 1.6 + 0.38x}$$

Q3 Fit a second degree parabola following data:

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

Soln:

x	u $(x-5)$	y	v $(y-8)$	u^2	u^3	u^4	uv	u^2v
1	-4	2	-6	16	-64	256	24	-96
2	-3	6	-2	9	-27	81	6	-18
3	-2	7	-1	4	-8	16	2	-4
4	-1	8	0	1	-1	1	0	0
5	0	10	2	0	0	0	0	0
6	1	11	3	1	1	1	3	3
7	2	11	3	4	8	16	6	12
8	3	10	2	9	27	81	6	18
9	4	9	1	16	64	256	4	16
	0		2	60	0	708	51	-69

Let the parabola of second degree to be fitted be

$$v = a + bu + cu^2$$

The normal equations for this are:

$$\Sigma v = na + b\Sigma u + c\Sigma u^2$$

$$\Sigma vu = a\Sigma u + b\Sigma u^2 + c\Sigma u^3$$

$$\Sigma vu^2 = a\Sigma u^2 + b\Sigma u^3 + c\Sigma u^4$$

substituting the values of Σu, Σv etc, in these equations:

$$2 = 9a + 60c$$

$$51 = 60b$$

$$\Rightarrow b = 0.85$$

$$-69 = 60a + 708c$$

$$\Rightarrow -23 = 20a + 236c$$

$\therefore$ solving these equations, we get

$$a = 2.002$$

$$b = 0.85$$

$$c = -0.267$$

$\therefore$ the equation of parabola is

$$v = 2.002 + 0.85(x-5) - 0.267(x-5)^2$$

$$\Rightarrow \boxed{y = -0.923 + (3.52)x - (0.267)x^2}$$

Practice Problems :

1) Fit a straight line to the following data:

Year x:	1961	1971	1981	1991	2001
Production y:	8	10	12	10	16

(in thousand tons)

and find the expected production in 2006.

[Ans: 15.2 thousand tons]

2) The following table gives the results of the measurements of train resistance; V is the velocity in miles per hour. R is the resistance in pounds per ton:

V:	20	40	60	80	100	120
R:	5.5	9.1	14.9	22.8	33.3	46.0

If R is related to V by the relation $R = a + bV + cV^2$, find a, b and c.

[Ans: $R = 3.48 - 0.002V + 0.0029V^2$]

Correlation

Whenever two variables x and y are so related that an increase in the one is accompanied by an increase or decrease in the other, then the variables are said to be correlated.

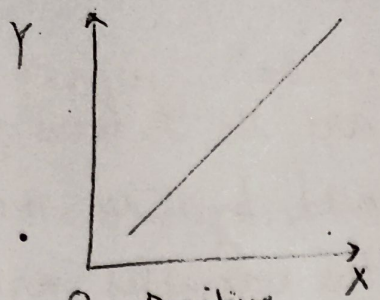
For example, the hike of price of a certain comodity varies with its demand in the market.

If an increase in one variable corresponds to an increase in the other, the correlation is said to be positive. If increase in one corresponds to the decrease in the other the correlation is said to be negative. If there is no relationship between the two variables, they are said to be independent.

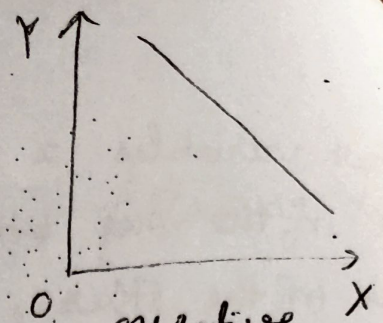
* If two variables vary in such a way that their ratio is always constant, then the correlation is said to be perfect.

Dot-diagram

Plot corresponding values of two variables, taking one on x-axis & the other along y-axis, we get a collection of dots. This collection of dots is called a dot diagram or a scatter diagram.



Positive correlation



negative correlation

Karl Pearson's coefficient of correlation

Let x & y be two random variables.
$X = x - \bar{x}$ and $Y = y - \bar{y}$, that is, X & Y are the deviations measured from their respective means.

$$\text{Cov}(x,y) = P = \frac{\Sigma XY}{n},$$ where x & y can take n values

and the coefficient of correlation

$$\boxed{r = \frac{\text{cov}(x,y)}{\sqrt{\text{Var}(x)}\ \sqrt{\text{Var}(y)}}}$$

$$= \frac{P}{\sigma_x \sigma_y},$$ σ_x and σ_y being the standard deviation of these series.

$$= \boxed{\frac{\Sigma XY}{\sqrt{(\Sigma X^2)(\Sigma Y^2)}}},$$ because

$$\sigma^2 = \frac{1}{n} \Sigma (x-\bar{x})^2$$

$$= \frac{\Sigma(x^2 - 2\bar{x}x + \bar{x}}{n}$$

$$= \frac{\Sigma x^2}{n} - 2\bar{x}\frac{\Sigma x}{n} + \bar{x}$$

$$= \frac{\Sigma x^2}{n} - 2\bar{x}^2 + \bar{x}^2$$

$$= \Sigma x^2 - \bar{x}^2$$

$$= \boxed{\frac{\Sigma XY}{n\,\sigma_x \sigma_y}}$$

* Coefficient of correlation lies in the interval $[-1, 1]$

Ten students got the following percentage of marks in Economics and Statistics.

Roll no.	1	2	3	4	5	6	7	8	9	10
Marks in Eco	78	36	98	25	75	82	90	62	65	39
Marks in stats	84	51	91	60	68	62	86	58	53	47

Calculate the coefficient of correlation.

Sol'n: Let the marks of two subjects be denoted by x & y resp. Then

$$\bar{x} = \frac{650}{10} = 65 \quad \text{and} \quad \bar{y} = \frac{660}{10} = 66.$$

x	y	$X = x - 65$	$Y = y - 66$	X^2	Y^2	XY
78	84	13	18	169	324	234
36	51	-29	-15	841	225	435
98	91	33	25	1089	625	825
25	60	-40	-6	1600	36	240
75	68	10	2	100	4	20
82	62	17	-4	289	16	-68
90	86	25	20	625	400	500
62	58	-3	-8	9	64	24
65	53	0	-13	0	169	0
39	47	-26	-19	676	361	494
650	660	0	0	5398	2224	2704

$$\therefore r = \frac{\Sigma XY}{\sqrt{\Sigma X^2}\sqrt{\Sigma Y^2}} = \frac{2704}{\sqrt{5398} \times \sqrt{2224}}$$

$$= \frac{2704}{73.4 \times 47.1} = \boxed{0.78}$$

Spearman's rank correlation co

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be ranks of n individuals corresponding characteristics and d be the difference of ranks between paired items in two series. Then rank coefficient of correlation is given by

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

Q1. Compute spearman's rank correlation coefficient for the following data:

Person	A	B	C	D	E	F	G	H	I	J
rank in stats	9	10	6	5	7	2	4	8	1	3
rank in income	1	2	3	4	5	6	7	8	9	10

Sol'n:

Person	rank in statistics	rank in income	$d = R_1 - R_2$	d^2
A.	9	1	8	64
B	10	2	8	64
C	6	3	3	9
D	5	4	1	1
E	7	5	2	4
F	2	6	-4	16
G	4	7	-3	9
H	8	8	0	0
I	1	9	-8	64
J	3	10	-7	49
				$\Sigma d^2 = 28$

rank correlation coefficient,

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

$$= 1 - \frac{6 \times 280}{10(100-1)}$$

$$= 1 - 1.697$$

$$= \boxed{-0.697}$$

Q2 Establish the formula

$$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

where r is the correlation coefficient between x & y.

Sol'n: we know that $\sigma_x^2 = \frac{\Sigma(x-\bar{x})^2}{n}$

$$\sigma_{x-y}^2 = \frac{\Sigma[(x-y)-\overline{(x-y)}]^2}{n}$$

$\overline{x-y}$ = mean of $(x-y)$ series

= mean of x − mean of y

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

$$\sigma_{x-y}^2 = \frac{\Sigma[(x-y)-(\bar{x}-\bar{y})]^2}{n}$$

$$= \frac{\Sigma[(x-\bar{x})-(y-\bar{y})]^2}{n}$$

$$= \frac{\Sigma(x-\bar{x})^2}{n} + \frac{\Sigma(y-\bar{y})^2}{n} - \frac{2\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

$$= \sigma_x^2 + \sigma_y^2 - \frac{2\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

Since, $r = \dfrac{\Sigma XY}{n\sigma_x \sigma_y}$ (corre coffi similarly,

$$= \frac{\Sigma (x-\bar{x})(y-\bar{y})}{n\,\sigma_x \sigma_y}$$

Also,

$$\therefore \frac{\Sigma (x-\bar{x})(y-\bar{y})}{n} = r\,\sigma_x \sigma_y$$

$$\therefore \boxed{\sigma^2_{x-y} = \sigma_x^2 + \sigma_y^2 - 2r\,\sigma_x \sigma_y}$$

Q3. If X and Y are uncorrelated random variables, find the coefficient of correlation between X+Y and X−Y.

Soln Let $x = X - \bar{X}$ & $y = Y - \bar{Y}$

since X & Y are not correlated (i.e. independent

$\therefore \text{cov}(X, Y) = 0$

$\Rightarrow \boxed{\Sigma xy = 0}$

Let $u = X+Y$, $v = X-Y$

Then $r = \dfrac{\Sigma (u-\bar{u})(v-\bar{v})}{n\,\sigma_u \sigma_v}$

$$\sigma_u^2 = \frac{\Sigma(u-\bar{u})^2}{n} = \frac{1}{n}\Sigma[(X-\bar{X}) + (Y-\bar{Y})]^2$$

$$= \frac{1}{n}\Sigma(x+y)^2$$

$$= \frac{1}{n}[\Sigma x^2 + \Sigma y^2 + 2\Sigma xy]$$

$$= \sigma_x^2 + \sigma_y^2 \quad (\because \Sigma xy = 0)$$

Similarly, $\sigma_v^2 = \sigma_x^2 + \sigma_y^2$

Also, $\Sigma(u-\bar{u})(v-\bar{v}) = \Sigma(X-\bar{X}+Y-\bar{Y})[(X-\bar{X})-(Y-\bar{Y})]$

$$= \Sigma(x+y)(x-y)$$

$$= \Sigma x^2 - \Sigma y^2$$

$$= n\sigma_x^2 - n\sigma_y^2$$

$$\therefore r = \frac{\Sigma(u-\bar{u})(v-\bar{v})}{n\,\sigma_u\,\sigma_v}$$

$$= \frac{n(\sigma_x^2-\sigma_y^2)}{\sqrt{n(\sigma_x^2+\sigma_y^2)}\,\sqrt{n(\sigma_x^2+\sigma_y^2)}}$$

$$= \boxed{\frac{\sigma_x^2-\sigma_y^2}{\sigma_x^2+\sigma_y^2}}$$

<u>Practice Problems</u>:

1) x and y are two random variables with the same standard deviation and correlation coefficient r. Show that the coefficient of correlation between x and $x+y$ is $\sqrt{\frac{1+r}{2}}$.

2.) Find the coefficient of correlation between x and y from the table of their values.

x :	1	3	4	6	8	9	11	14
y :	1	2	4	4	5	7	8	9

Line of regression

If the dot diagram indicates some relationship between two variables x and y, then the dots of the dot diagram will be concentrated round a curve. This curve is called curve of regression. When the curve is a straight line, it is called a line of regression, that is, a line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

Equations of the line of regression

Let x and y be two discrete random variables and $\bar{x}$, $\bar{y}$ denote the means of x series, y series respectively.

The line of regression passes through $(\bar{x}, \bar{y})$

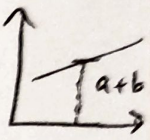
And the line of regression of y on x is given by

$$\boxed{y-\bar{y} = r\,\frac{\sigma_y}{\sigma_x}\,(x-\bar{x})}$$

[r = coeff. of correlation]

with slope $= r\,\dfrac{\sigma_y}{\sigma_x}$.

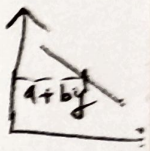
[general form $y = a+bx$



Also the line of regression of x on y is given by

$$\boxed{x-\bar{x} = r\,\frac{\sigma_x}{\sigma_y}\,(y-\bar{y})}$$

[general form $x=a+by$



$x = a+b$

with slope $= r\,\dfrac{\sigma_x}{\sigma_y}$

→ note that $\boxed{b_{yx} = r\,\dfrac{\sigma_y}{\sigma_x}}$ and $\boxed{b_{xy} = r\,\dfrac{\sigma_x}{\sigma_y}}$

are known as the <u>coefficients of regression</u>.

Moreover,

$$b_{yx}\,b_{xy} = \left(r\,\frac{\sigma_y}{\sigma_x}\right)\left(r\,\frac{\sigma_x}{\sigma_y}\right)$$

$= r^2$, <u>where r is the coefficient of correlation</u>.

Q1 Find the correlation coefficient be[illegible] x and y, when the lines of regress[illegible] are: $2x - 9y + 6 = 0$ and $x - 2y + 1 =$

Sol'n: Suppose that line of regression of x on y is

$$2x - 9y + 6 = 0$$

& that of y on x is $x - 2y + 1 = 0$

Now, $2x - 9y + 6 = 0 \Rightarrow x = \frac{9}{2}y - 3 \Rightarrow b_{xy} = \frac{9}{2}$

and $x - 2y + 1 = 0 \Rightarrow y = \frac{1}{2}x + \frac{1}{2} \Rightarrow b_{yx} = \frac{1}{2}$

$\therefore$ coeff of correlation, $r = \sqrt{b_{xy} \cdot b_{yx}}$

$$= \cancel{\sqrt{\frac{2 \times 2}{9}} = \frac{2}{3}}$$

$$= \sqrt{\frac{9}{2} \cdot \frac{1}{2}} = \frac{3}{2} > 1,$$

which is not possible

because coefficient of correlation, $r \leq 1$.

$\therefore$ regression line of x on y is $x - 2y + 1 = 0$

$\Rightarrow b_{xy} = 2$

and regression line of y on x is $2x - 9y + 6 = 0$

$\Rightarrow b_{yx} = \frac{2}{9}$.

$\therefore r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{2 \times \frac{2}{9}} = \frac{2}{3} < 1$

Hence the correlation coefficient b/w x & y is $\boxed{\frac{2}{3}}$.

Q2 Find the regression line of y on x for the following data

x:	1	3	4	6	8	9	11	14
y:	1	2	4	4	5	7	8	9

Estimate the value of y, when $x = 10$. Also find its coefficient of regression.

Soln:

x	y	xy	x^2
1	1	1	1
3	2	6	9
4	4	16	16
6	4	24	36
8	5	40	64
9	7	63	81
11	8	88	121
14	9	126	196
56	40	364	524

Let $y = a + bx$ be the line of regression of y on x, where a & b are determined by normal equations:

$$\Sigma y = na + b\,\Sigma x$$

$$\Sigma xy = a\,\Sigma x + b\,\Sigma x^2$$

$\Rightarrow$ $40 = 8a + 56b$

and $364 = 56a + 524b$

$\Rightarrow$ $a = \frac{6}{11}$, $b = \frac{7}{11}$

$\therefore$ the eq^n of line of regression is

$$\boxed{y = \frac{6}{11} + \frac{7}{11}x}$$

and coeffient of regression, $\boxed{b_{yx} = \frac{7}{11}}$

Moreover, at $x = 10$, $\boxed{y = \frac{76}{11}}$.

Q3 The regression equations calculated from a given set of observations for two random variables are

$$x = -0.4y + 6.4$$

and $y = -0.6x + 4.6$

Calculate $\bar{x}, \bar{y}$ and r.

Soln: The regression equations are

$$x = -0.4y + 6.4$$
$$y = -0.6x + 4.6$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = -0.4 \quad \text{—(1)}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = -0.6 \quad \text{—(2)}$$

$\Rightarrow r^2 = 0.24$

$\Rightarrow r = \pm 0.49$

Since σ_x & σ_y are always positive

$\therefore$ r is negative (from (1) & (2))

$\therefore \boxed{r = -0.49}$

To find, $\bar{x}$ & $\bar{y}$, since these line pass through $(\bar{x}, \bar{y})$, we have

$$\bar{x} = -0.4\bar{y} + 6.4$$
$$\bar{y} = -0.6\bar{x} + 4.6$$

Solving these equations we get $\boxed{\bar{x} = 6}$ and $\boxed{\bar{y} = 1}$

Practice Problems

1.) Prove that arithmetic mean of the coeff. of regression is greater than the coeffi. of correlation.

2.) Find the correlation coefficient and the equations of regression lines from the following data:

x	1	2	3	4	5
y	2	5	3	8	7

[Ans: $r = 0.81$, $x = 0.5y + 0.5$, $y = 1.3x + 1.1$]

3.) The following regression equations and variances are obtained from a correlation table:

$20x - 9y - 107 = 0$, $4x - 5y + 33 = 0$, variance of $x = 9$.

Find (i) mean values of x and y,
(ii) the standard deviation of y.

[Ans: $\bar{x} = 13$, $\bar{y} = 17$, $\sigma_y = 4$]

Sampling Theory

Many a-times one is interested in drawing valid conclusions about a large group of individuals or objects. Instead of examining the entire group one may think of examining a small part of it. The group of individuals under study is called <u>population</u>. A part selected from the population is called a <u>sample</u>. The process of selection of a sample is called <u>sampling</u>.

A Random sample is one in which each member of population has an equal chance of being included in it.

There are ${}^{N}C_{n}$ different samples of size n that can be picked up from a population of size N.

Parameters and Statistics

The statistical constants of the population such as mean (μ), standard deviation (σ) are called parameters.

The mean ($\bar{x}$), standard deviation (s) of a sample are known as statistics.

Aims of sampling

The population parameters are not known gener
The sample characteristics are utilised to appro
determine population parameters. Thus static a
an estimate of the parameter. The logic of the
sampling theory is the logic of induction. In
induction, we pass from a particular (sample)
to general (population). This type of generalization
here is known as statistical inference.

Sampling distribution

From a population a number of samples are
drawn of equal size n. Find out the mean
of each sample. The means of samples are not
equal. The means with their respective frequen
are grouped. The frequency distribution so
formed is known as sampling distribution
of mean. Similarly, we can have sampling
distribution of standard deviation.

Standard error (S.E.)

The standard deviation of a sampling distributio
of a statistic is known as standard error.

Sampling distribution of means from infinite population

Let the population be infinitely large and having a population mean μ and a population variance of σ^2. Let X be a random variable denoting the measurement of the charateristic, then $E(X) = \mu$ and $Var(X) = \sigma^2$.

The sample mean, $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$, where $x_1, x_2, \dots, x_n$ are independent v'bles from the infinitly large population.

$$\therefore E(\bar{x}) = E\left[\frac{x_1 + x_2 + \dots + x_n}{n}\right]$$

$$= \frac{1}{n}[E(x_1) + \dots + E(x_n)]$$

$$= \frac{1}{n}[n\mu]$$

$$= \boxed{\mu}$$

$$Var(\bar{x}) = Var\left[\frac{x_1 + x_2 + \dots + x_n}{n}\right]$$

$$= \frac{1}{n^2}[Var(x_1) + \dots + Var(x_n)]$$

$$= \frac{n\sigma^2}{n^2}$$

$$= \boxed{\frac{\sigma^2}{n}}$$

The expected value of the sample mean is same as population mean. The variance of sample mean is the variance of the popu. divided by the sample size.

The standard deviation $\left(\frac{\sigma}{\sqrt{n}}\right)$ of $\bar{x}$ is called standard error of the mean.

It is denoted by $\sigma_{\bar{x}}$.

Sampling from Normal Population

If $x \sim N(\mu, \sigma^2)$, then $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

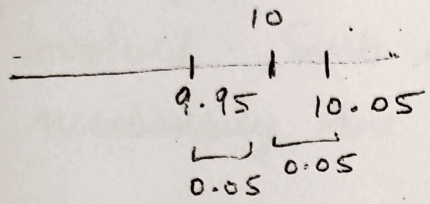
Ques The diameter of a component produced on a semi-automatic machine is known to be distributed normally with a mean of 10 mm and a standard deviation of 0.1 mm. If we pick up a random sample of size 5, what is the probability that the sample mean will be between 9.95 and 10.05 mm?

Soln: Let x be a random variable representing the diameter of one component picked up at random.

Here $x \sim N(10, 0.01)$

$\therefore \bar{x} \sim N\left(10, \frac{0.01}{5}\right)$ $\left[\bar{x} = N\left(\bar{x}, \frac{\sigma^2}{n}\right)\right]$

$Pr\{9.95 \le \bar{x} \le 10.05\} = 2 \times Pr\{10 \le \bar{x} \le 10.05\}$



$\left(\text{Put } z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$

$$= 2 \times Pr\left\{\frac{10-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{10.05-\mu}{\frac{\sigma}{\sqrt{n}}}\right\}$$

$$= 2 \times Pr\left\{0 \le z \le \frac{10.05-10}{\frac{0.1}{\sqrt{5}}}\right\}$$

$$= 2 \times Pr\{0 \le z \le 1.12\}$$

$$= 2 \times 0.3686$$

$$= \boxed{0.7372}$$

Sampling distribution of the variance

We use a sample statistic called sample variance to estimate the population variance. The sample variance is usually denoted by s^2 and is given by

$$\boxed{s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Hypothesis : Very often it is required to make decisions about populations on the basis of sample information. Such decisions are called statistical decisions.

In attempting to reach decisions, it is often necessary to make assumptions about the population involved. Such assumptions, which are not necessarily true, are called statistical hypothesis.

Null hypothesis : The hypothesis which is tested for possible rejection under the assumption that it is true is called null hypothesis. It is denoted by H_0.

Tests of significance : Procedures which enable us to decide, on the basis of sample information, whether to accept or reject hypothesis or to determine whether observed sampling results differ significantly from expected results, are called tests of significance.

Level of significance : The probability level below which we reject the hypothesis is called the level of significance.

Alternative hypothesis: Any hypothesis which complementary to the null hypothesis is called an alternative hypothesis, usually denoted H_1.

For example, if we want to test the null hypothesis that the population has a specified mean μ_0, then

$H_0 : \mu = \mu_0$ and H_1 could be

(i) $H_1 : \mu \neq \mu_0$ (that if $\mu > \mu_0$ or $\mu < \mu_0$)

(ii) $H_1 : \mu > \mu_0$

(iii) $H_1 : \mu < \mu_0$

The alternative hypothesis in (i) is known as a <u>two tailed alternative</u> and the alternatives in (ii) and (iii) are known as <u>right-tailed</u> and <u>left-tailed</u> alternatives respectively.

Q It is believed that the precision (as measured by the variance) of an instrument is no more than 0.16. Write down the null and alternative hypothesis for testing this belief.

Soln:

$H_0 : \sigma^2 = 0.16$

$H_1 : \sigma^2 > 0.16$.

called by

The t- Distribution (for small sample)

The students distribution is used to test the significance of

(i) The mean of a small sample.

(ii) The difference between the means of two small samples or to compare two small samples.

(iii) The correlation coefficient.

Let x_i $(i=1,2,\ldots,n)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then student's t is defined by the statistic

$$\boxed{t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}}$$, where

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean and

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

with $(n-1)$ degree of freedom.

Degree of freedom (d.f)

The no. of independent variates which make up the statistic is known as the degree of freedom (d.f.)

Q1 Ten individuals are chosen at random from population and their heights are found to be in inches 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. Do the suggestion that the mean height of under is 65. For 9 degree of freedom t at 5% level significance = 2.262.

Remark: Working rule:

To calculate significance of sample mean at 5% level.

Calculate $t = \frac{\bar{x} - \mu}{s} \sqrt{n}$ and compare it to the value of t with (n-1) degrees of freedom at 5% level, obtained from the table. Let this tabulated value of t by t_0.

If $t < t_0$, then we accept the hypothesis.

If $t > t_0$, we reject the hypothesis and the sample is not drawn from the population.

$H_0 : \mu = 65$ inches

x	$x - \bar{x}$ $(x-67)$	$(x-\bar{x})^2$ $(x-67)^2$
63	−4	16
63	−4	16
64	−3	9
65	−2	4
66	−1	1
69	2	4
69	2	4
70	3	9
70	3	9
71	4	16
$\Sigma x = 670$		$\Sigma(x-\bar{x})^2$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{670}{10} = 67$$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{88}{9}} = 3.13$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{67-65}{\frac{3.13}{\sqrt{10}}}$$

$$= \frac{2\sqrt{10}}{3.13} = 2.02$$

Since $t_0 = 2.262$

and $t < t_0$, that is, calculated value of t (2.02) is less than the table value of t (2.262). $\therefore$ The hypothesis is accepted and the mean height of universe is 65 inches.

Q2 A random sample of size 16 values from a normal population showed a mean of 53 and a sum of squares of deviation from the mean equals to 150. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95% and 99% confidence limits of the mean of the population.

$\gamma = 15$, $\alpha = 0.05$, $t = 2.131$

degree of freedom

$\alpha = 0.01$, $t = 2.947$

level of significance

Soln:

$\mu = 56$, $n = 16$, $\bar{x} = 53$, $\Sigma(x-\bar{x})^2 = 150$

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{150}{15} = 10$$

$$s = \pm\sqrt{10}$$

$$\therefore t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{53-56}{\frac{\pm\sqrt{10}}{\sqrt{16}}} = \frac{-3\times 4}{\pm\sqrt{10}}$$

$$= 3.79$$

$$\therefore t = 3.79$$

Since $3.79 > t_{0.05} = 2.131$

and $3.79 > 2.947 = t_{0.01}$,

the sample cannot be regarded as taken from the population.

Testing for difference between means of two small samples

Let the mean and variance of the first and second population be μ_1 and σ_1^2 and μ_2 & σ_2^2 respectively. Let $\bar{x}_1, \bar{x}_2$ be the mean of small sample of size n_1, n_2 from first and second population respectively.

Then
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
with $(n_1 + n_2 - 2)$ d

Ques 1: Two independent samples of 8 and 7 items respectively had the following values of the variables (weight in ounces):

(x_1) Sample 1:	9	11	13	11	15	9	12	14
(x_2) Sample 2:	10	12	10	14	9	8	10	

Is the difference between the means of the sample significant?

[Given for $\underset{\text{d.f.}}{\nu} = 13$, $t_{0.05} = 2.16$]

~~Soln~~ ... [Do yourself]

[Hint: $\mu_1 = \mu_2$ $\therefore$ $\mu_1 - \mu_2 = 0$

$\bar{x}_1 = 11.75$, $\bar{x}_2 = 10.43$.

calculate $s = \sqrt{\dfrac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(y_1-\bar{y}_1)^2}{n_1+n_2-2}}$

use test statistic, $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$]

(same as next question)

Ques2: Samples of two types of electric light bulbs were tested for length of life and following data were obtained

	Type I	Type II
sample no.	$n_1 = 8$	$n_2 = 7$
sample means	$\bar{x}_1 = 1234$ hrs	$\bar{x}_2 = 1036$ hrs
sample S.D.'s	$s_1 = 36$ hrs	$s_2 = 40$ hrs

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life?

soln: null hypothesis, $H_0 : \mu_1 = \mu_2$

alternative hypothesis, $H_1 : \mu_1 > \mu_2$

(type I is superior to type II)

test statistic,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \text{ where}$$

$$s^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2\right]$$

$$= \frac{1}{n_1+n_2-2}\left[n_1 s_1^2 + n_2 s_2^2\right]$$

$$= \frac{1}{13}\left[8 \times (36)^2 + 7 \times (40)^2\right]$$

$$= 1659.08$$

$$\therefore t = \frac{198}{\sqrt{1659.08\left(\frac{1}{8}+\frac{1}{7}\right)}} = \boxed{9.39}$$

$t_{0.05} = 2.16$ for 13 d.f. $\rightarrow t_0$

and $t > t_0$

$\therefore$ H_0 is rejected and hence type I is definitely superior to type II.

<u>Practice problem</u> :

<u>Ques</u> An experiment was conducted on nine individuals. The experiment showed that due to smoking, the pulse rate increased in the following order:

5, 3, 4, −1, 2, −3, 4, 3, 1.

Can you maintain that smoking leads to ~~the~~ an increase in the pulse rate?

(t for 8 d.f. at 5% level of significance = 2.31).

The Chi-Square distribution

Chi-square is a measure of actual divergence of the observed and expected frequencies. If f_o is the observed frequency and f_e is the expected frequency of a class interval, then χ^2 is defined as

$$\boxed{\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}}$$

with $(n-1)$ d.f., where n is the no. of frequencies.

Independence of Attributes: Let us consider two attributes A and B, A divided into r classes $A_1, A_2, ---, A_r$ and B divided into s classes $B_1, B_2, ---, B_s$. The various cell frequencies can be expressed in table known as contingency table. In an $r \times s$ contingency table, the no. of degrees of freedom is:

$$\boxed{d.f. = (r-1)(s-1)}$$

Q1 From the following table, showing the number of plants having certain characters, test the hypothesis that the flower colour is independent of flatness of leaf

	flat leaves	curled leaves	Total
White flowers	99	36	135
red flowers	20	5	25
Total	119	41	160

Soln: <u>null hypothesis</u>: The flower colour is dependent of flatness of leaf.

The following table shows the theoretical (expected) frequencies.

	flat leaves	curled leaves	Total
white flowers	$\frac{135 \times 119}{160} = 100$	$\frac{135 \times 41}{160} = 35$	135
red flowers	$\frac{25 \times 119}{160} = 19$	$\frac{25 \times 41}{160} = 6$	25
Total	119	41	160

$$\chi^2 = \Sigma \frac{(fo - fe)^2}{fe}$$

$$= \frac{(90-100)^2}{100} + \frac{(36-35)^2}{35} + \frac{(20-19)^2}{19} + \frac{(5-6)^2}{6}$$

$$= \frac{1}{100} + \frac{1}{35} + \frac{1}{19} + \frac{1}{6} = 0.2579$$

degree of freedom (d.f.) $= (r-1)(c-1)$
$= (2-1)(2-1)$
$= 1$

$\chi_0^2 = 0.0158$ at 0.1 level of significance with 1 degree of freedom.

Since $\chi^2 > \chi_0^2$, we reject the hypothesis and the flower colour is independent of flatness of leaf at the 0.1 level of significance.

Q2 For a 2×2 table

a	b
c	d

$f_o(a)$

prove that chi-square test of ~~significance~~ independence gives

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)},$$ where

$N = a+b+c+d$.

Soln: Table for expected frequencies

$f_e(b)$

$\frac{(a+b)(a+c)}{N}$	$\frac{(a+b)(b+d)}{N}$	$a+b$
$\frac{(a+c)(c+d)}{N}$	$\frac{(b+d)(c+d)}{N}$	$c+d$
$a+c$	$b+d$	N

$f_e(a)$ $f_e(d)$ $f_e(c)$

$$\therefore \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Consider, a - $\frac{(a+b)(a+c)}{N} = \frac{a(a+b+c+d) - (a^2+ac+bc)}{N}$

$$= \frac{ad-bc}{N}$$

Similarly $\rightarrow f_o^{(a)} - f_e^{(a)}$

b - $\frac{(a+b)(b+d)}{N} = -\frac{(ad-bc)}{N}$

c - $\frac{(a+c)(c+d)}{N} = -\frac{(ad-bc)}{N}$

d - $\frac{(b+d)(c+d)}{N} = \frac{ad-bc}{N}$

$\therefore$

$$\chi^2 = \frac{(ad-bc)^2}{N^2}\left[\frac{1}{f_e(a)} + \frac{1}{f_e(b)} + \frac{1}{f_e(c)} + \frac{1}{f_e(d)}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\frac{1}{(a+b)(a+c)} + \frac{1}{(a+b)(b+d)} + \frac{1}{(a+c)(c+d)} + \frac{1}{(b+d)(c+d)}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\frac{(a+b+c+d)^2}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\therefore \boxed{\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}}$$

Q3 The following table gives the number of air craft accidents that occurs during various days of the week. Find whether the accidents are uniformly distributed over the week.

Days	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
no. of accidents	14	16	8	12	11	9	14

The values of chi-square significant at 5, 6, 7 are respectively 11.07, 12.59, 14.07 at the 5% level of significance.

Soln: null hypothesis: The accidents are uniformly distributed over the week.

Expected frequencies of accidents $= \frac{14+16+8+12+11+9+14}{7}$

$$= \frac{84}{7}$$

$$= \boxed{12}$$

$\therefore$ table for expected frequencies

Days	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri
no. of accidents	12	12	12	12	12	12

$$\therefore \chi^2 = \frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12}$$

$$= \frac{1}{12}[4+16+16+0+1+9+4]$$

$$= \frac{50}{12} = \boxed{4.17}$$

degree of freedom (d.f.) $= n-1 = 7-1 = \boxed{6}$

$\chi^2_{0.05}$ for 6 d.f. $= 12.59 > \chi^2$

$\therefore$ null hypothesis is accepted.

Hence, the accidents are uniformly distributed over the week.

Fri Sep 12

F- Distribution

It is used to test

(i) whether two independent samples $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ have been drawn from the normal population with the same variance σ^2.

(ii) whether the two independent estimates of the population variance are homogeneous or not.

Under null hypothesis

$$H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2 \quad \text{(population variances are equal)}$$

F statistic is given by

$$F = \frac{S_X^2}{S_Y^2}, \text{ where}$$

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

and $$S_Y^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_j - \bar{y})^2$$

It has degree of freedom (ν_1, ν_2) where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

Q: Two random samples are:

sample	size	sum of squares of deviation from the mean
1	10	90
2	12	108

Test whether the samples come from the same normal population at 5% level of significance.

[Given: $F_{0.05}(9,11) = 2.90$, $F_{0.05}(11,9) = 3.10$]

Soln: null hypothesis: The two samples have been drawn from the same normal population.

$n_1 = 10$, $\Sigma(x_i - \bar{x})^2 = 90$

$n_2 = 12$, $\Sigma(y_i - \bar{y})^2 = 108$

$$s_1^2 = \frac{1}{n_1 - 1}\Sigma(x_i - \bar{x})^2 = \frac{90}{10} = \boxed{10}$$

$$s_2^2 = \frac{1}{n_2 - 1}\Sigma(y_i - \bar{y})^2 = \frac{108}{11} = \boxed{9.82}$$

$$F = \frac{s_1^2}{s_2^2} = \frac{10}{9.82} = 1.018$$

$F_{0.05}(9,11) = 2.90$

Since

$F < F_{0.05}(9, 11)$, we accept the hypothesis.

Q2 Two random samples from two normal population are given below:

sample 1	16	26	27	23	24	22
sample 2:	33	42	35	32	28	31

Do the estimates of population variances differ significantly?

d.f.	(5,5)	(5,6)	(6,5)
5% value of F	5.05	4.39	4.95

x_i	$x_i - \bar{x}$ $x_i - 23$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$ $y_i - 33.5$	$(y_i - \bar{y})^2$
16	-7	49	33	-0.5	0.25
26	3	9	42	8.5	72.25
27	4	16	35	1.5	2.25
23	0	0	32	-1.5	2.25
24	1	1	28	-5.5	30.25
22	-1	1	31	-2.5	6.25
		76			113.50

$$s_x^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{76}{5} = 15.2$$

$$s_y^2 = \frac{\Sigma(y-\bar{y})^2}{n-1} = \frac{113.5}{5} = 22.7$$

$$F = \frac{s_y^2}{s_x^2} = \frac{22.7}{15.2} = 1.4934$$

[<u>remark</u> : Greater of the two variances s_x^2, s_y^2 is to be taken in the numerator]

Tabulated value of $F_{0.05} = 5.05$

Since $F < F_{0.05}$, we accept the hypothesis. The difference is not significant.

<u>Fisher's Z- Distribution</u>

It is the statistical distribution of half the logarithm of an F-distribution variate, that is, on substitution $z = \frac{1}{2}\log_e F$ or $F = e^{2z}$ in the F-distribution, we obtain Fisher's z-distribution: It is of the form

$$y = y_0 \cdot e^{r_1 z}(r_1 e^{2z} + r_2)$$, where

$r_1 = n_1 - 1$, $r_2 = n_2 - 1$ are d.f. of two samples of size n_1 & n_2.

Practice problems:

Q1 Two random samples drawn from two normal populations are:

Sample I	20	16	26	27	23	22	18	29	25	19
Sample II	27	33	42	35	32	34	38	28	41	23 30

Obtain the estimates of the variances of the populations & test whether the population have the same variance. (Given $F_{0.05} = 3.11$ for 11 and 9 d.f.

[Ans: $F = 2.368$, we accept the hypo]

Q2 The diameters of two random samples, each of size 10 of bulbs produced by two machines have standard deviations $S_1 = 0.01$ and $S_2 = 0.015$. Assuming that the diameters have independent distributions, test the hypothesis that, the two machines are equally good by testing.

[Ans: $F = 1.5$, yes hypothesis is correct.