7/3/13

4. Correlation

In a bivariate data if the change in one variable create, any change in the other variable, the two variables are said to correlated. It is of 2 types.

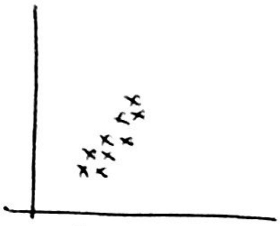
(i) **Positive correlation**: If the change in two variables is in the same direction, it is called as '+ve correlation.

Ex: $X\uparrow, Y\uparrow$ or $X\downarrow, Y\downarrow$. (Income & Expenditure

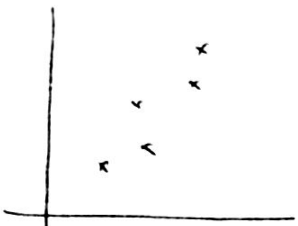
Negative correlation: If the change is in opposite directions, then it is '-ve correlation.

Ex: $X\uparrow, Y\downarrow$; $X\downarrow, Y\uparrow$ (Pressure & Volume)

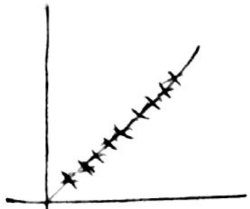
Scattered diagrams: Used to get an idea about correlation.



high '+ve correlation.



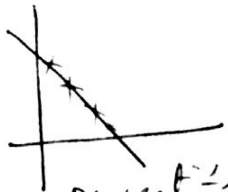
poor '+ve



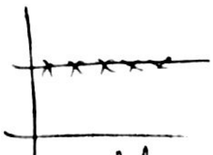
Perfect positive

high '-ve correlation

poor '-ve



perfect '-ve



null correlation

Karl-Pearson's correlation coefficient:

This is a numerical measurement for the linear relationship of x and y. It is denoted by r and is given as

$$\boxed{r = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}}$$

$$\frac{E\left[(X-E(X))(Y-E(Y))\right]}{\sqrt{E(X-E(X))^2}\sqrt{E(Y-E(Y))^2}}$$

$$= \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}\sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}}$$

Note:

$$\boxed{-1 \leq r \leq +1}$$

Find the correlation coefficient.

X	Y	XY	X^2	Y^2
80	44	3520	6400	1936
90	32	2880	8100	1024
65	75	4875	4225	5625

To find the correlation coefficient, we calculate the foll. table.

$\Sigma x^2 = 18725$ $\qquad$ $\Sigma xy = 11275$

$\Sigma y^2 = 8585$

$\bar{x} = 78.33$

$\bar{y} = 50.33$

$$Cov(x,y) = \frac{1}{n}\Sigma xy - \bar{x}\bar{y}$$

$$= \frac{1}{3}(11275) - (78.33)(50.33) = -184.02$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2} = 10.30$$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2} = 18.13$$

$\therefore$ correlation coeff $= \underline{\underline{-0.99}}$

The two variables are negatively correlated and correlation is high.

correlation is high if (>0.5).

Properties of r

* Correlation coeff. is independent of change of origin and scale.

Let x, y are two variables for which the correlation coefficient is $r_{x,y}$

Consider $U = \frac{x-a}{h}$, $V = \frac{y-b}{k}$ where a, b, h, k are constants

we can write $x = a + Uh \Rightarrow E(x) = a + hE(U)$

$$\Rightarrow x - E(x) = \cancel{a} + Uh - \cancel{a} - hE(U)$$
$$= h[U - E(U)].$$

Similarly $y - E(y) = k[V - E(V)]$.

We know that

$$r_{xy} = \frac{E[(x - E(x))(y - E(y))]}{\sqrt{E(x - E(x))^2}\sqrt{E(y - E(y))^2}}$$

$$= \frac{E[h(U - E(U))\, k(V - E(V))]}{\sqrt{E(h(U - E(U)))^2}\sqrt{E(k(V - E(V)))^2}}$$

$$= \frac{\cancel{h}\cancel{k}\, E[(U - E(U))(V - E(V))]}{\cancel{h}\sqrt{E(U - E(U))^2}\ \cancel{k}\sqrt{E(V - E(V))^2}}$$

$$= \frac{\mathrm{cov}(U, V)}{\sigma_U \sigma_V} = r_{UV}.$$

* correlation coefficient lies b/n -1 and $+1$

$$\boxed{-1 \leq r \leq 1}$$

$$r_{xy} = \frac{E[(x - E(x))(y - E(y))]}{\sqrt{E(x - E(x))^2}\sqrt{E(y - E(y))^2}}$$

Let $(X - E(X)) = U$, $Y - E(Y) = V$

$$r_{XY} = \frac{E(UV)}{\sqrt{E(U)^2}\sqrt{E(V)^2}}$$

$$r^2_{XY} = \frac{(E(UV))^2}{E(U)^2 E(V)^2}$$

from Cauchy-Schwartz inequality. we have

$$(E(XY))^2 \le E(X^2)\,E(Y^2).$$

$$r^2_{XY} \le 1.$$

$$\longrightarrow \boxed{-1 \le r \le 1}$$

* For independent random variable, correlation is 0 but the converse is not true.

$$r_{XY} = \frac{E\left[(X-E(X))(Y-E(Y))\right]}{\sqrt{E(X-E(X))^2}\sqrt{E(Y-E(Y))^2}}$$

$$E\left[(X-E(X))(Y-E(Y))\right]$$

$$= E\left[(XY) - XE(Y) - E(X)Y + E(X)E(Y)\right]$$

$$= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)$$

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

$$= 0 \quad [\text{if } X \,\&\, Y \text{ are independent}]$$

$$r = 0.$$

Though $r = 0$, here $Y = X^2$ ie; there may exist other than linear relationship b/n X and Y.

Spearman's Rank Correlation coefficient:

It is useful to measure the correlation b/w 2 qualitative variables. It is denoted by ρ given as

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)} \qquad \boxed{-1 \leq \rho \leq 1}$$

$$\rho = 1 - \frac{6\left(\Sigma d_i^2 + \frac{m(m^2-1)}{12}\right)}{n(n^2-1)}$$

<u>Regression</u>: It is useful to estimate the value of one variable for a given value of other variable.

<u>Regression of Y on X</u>: It is useful to estimate of Y, for a given value of X, and is given as

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

b_{YX} is regression coefficient, $b_{YX} = \frac{\text{cov}(X,Y)}{V(X)}$

$$b_{YX} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\frac{1}{n}\Sigma x^2 - \bar{x}^2} = r\frac{\sigma_y}{\sigma_x}$$

Regression line of X on Y is useful to estimate the value of x for a given value of Y and is given as

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

$$b_{XY} = \frac{\text{cov}(X,Y)}{V(Y)} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\frac{1}{n}\Sigma y^2 - \bar{y}^2} = \frac{r\sigma_x}{\sigma_y}$$

$$= \frac{\text{cov}(x,y)}{\sigma_y^2}$$

14/8/13 **Properties of Regression coefficients**

1. Correlation coefficient is the geometric mean of regression coefficients.

$$b_{XY} = \frac{\mathrm{Cov}(X,Y)}{V(Y)} \qquad b_{YX} = \frac{\mathrm{Cov}(X,Y)}{V(X)}$$

$$GM = \sqrt{b_{XY} b_{YX}} = \sqrt{\frac{\mathrm{Cov}(X,Y)}{V(Y)} \cdot \frac{\mathrm{Cov}(X,Y)}{V(X)}}$$

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

2. If one of the regression coeff. is greater than unity, the other is less than unity.

let $b_{XY} > 1$

wkt $r = \sqrt{b_{XY} b_{YX}} \Rightarrow r^2 = b_{XY} b_{YX} \leq 1$

$$\Rightarrow b_{XY} \leq \frac{1}{b_{YX}} < 1.$$

3. The arithmetic mean of regression coefficient is greater than that of correlation coefficient

Assume that $\dfrac{b_{XY} + b_{YX}}{2} > r.$

$$b_{XY} + b_{YX} > 2r$$

$$\cancel{r}\frac{\sigma_X}{\sigma_Y} + \cancel{r}\frac{\sigma_Y}{\sigma_X} > 2\cancel{r} \Rightarrow \frac{\sigma^2_X + \sigma^2_Y}{\sigma_X \sigma_Y} > 2.$$

$$\sigma^2_X + \sigma^2_Y - 2\sigma_X\sigma_Y > 0$$

$$\Rightarrow (\sigma_X - \sigma_Y)^2 > 0.$$

A squared quantity is always > 0 $\Rightarrow$ Assumption is true

4. Regression coefficients are independent of change of origin but not change of scale.

$$U = \frac{X-a}{h}, \quad V = \frac{Y-b}{k}$$

wkt $b_{XY} = \frac{cov(X,Y)}{V(Y)} = \frac{E\left((X-E(X))(Y-E(Y))\right)}{E\left(Y-E(Y)\right)^2}$

$X = a + Uh \quad Y = kV + b$

$E(X) = a + hE(U)$

$\Rightarrow X - E(X) = h\left(U - E(U)\right)$

|||ly $Y - E(Y) = k\left(V - E(V)\right)$

$$b_{XY} = \frac{E\left[h(U-E(U))\,k(V-E(V))\right]}{E\left(k(V-E(V))\right)^2}$$

$$= \frac{h\cancel{k}\,E\left((U-E(U))(V-E(V))\right)}{k^{\cancel{2}}\,E\left(V-E(V)\right)^2}$$

$$= \frac{h}{k}\,b_{UV}$$

5. Angle b/n 2 regression lines.

$$\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2}$$

$(y - \bar{y}) = \overset{m_1}{b_{YX}}(x - \bar{x})$

$(x - \bar{x}) = b_{XY}(y - \bar{y}) \Rightarrow (y - \bar{y}) = \overset{m_2}{\frac{1}{b_{XY}}}(x - \bar{x})$

$$\tan\theta = \frac{b_{YX} - \frac{1}{b_{XY}}}{\ }$$

$$\frac{\dfrac{r\sigma_y}{\sigma_x} - \dfrac{1}{r\dfrac{\sigma_x}{\sigma_y}}}{1 + \dfrac{\cancel{r}\dfrac{\sigma_y}{\sigma_x}}{\cancel{r}\dfrac{\sigma_x}{\sigma_y}}} = \frac{r^2 - 1}{\dfrac{\sigma_x}{\sigma_y}\left[1 + \dfrac{\sigma_y^2}{\sigma_x^2}\right]}\left(\frac{1}{r}\right)$$

$$\tan\theta = \left(\frac{r^2-1}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}$$

$$\theta = \tan^{-1}\left[\left(\frac{r^2-1}{r}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right].$$

Case 1: If $r = \pm 1$, $\Rightarrow \theta = \tan^{-1}(0) = 0$

$\therefore$ the two regression lines completely coincides with each other

Case 2: If $r = 0 \Rightarrow \theta = \pi/2$ then the two regression lines are $\perp$ to each other.

Small Samples.

(i) t-test for testing the significance of single mean.

(i) <u>Null hypothesis</u>: The population mean is equal to given value.

$$H_0 : \mu = \mu_0 .$$

(ii) <u>Alternative hypothesis</u>: The population mean is not equal to given value.

(iii) Fix the level of significance, α.

(iv) Under H_0, compute the test statistic.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x} = \frac{1}{n}\left(\Sigma x_i\right)$ = sample mean.

s^2 = sample mean sum of squares.

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

(v) ν = degrees of freedom = $n-1$

If t_{cal} is in acceptance region, accept H_0.

Or else reject H_0.

(ii) t test for 2 pop^n means differences.

(i) <u>Null hypothesis</u>: two population means are equal (No significant difference)

$$H_0 : \mu_1 = \mu_2$$

(ii) <u>Alternative hypothesis</u>: The two population means are not equal.

$$H_1 : \mu_1 \neq \mu_2$$

(iii) Fix the level of significance, α.

(iv) Under H_0, compute t,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right)}} \sim t_{n_1+n_2-2}$$

where $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}$

$$= \frac{\Sigma (x_i - \bar{x})^2 + \Sigma (y_i - \bar{y})^2}{n_1 + n_2 - 2}$$

$\bar{x}_1 \rightarrow$ first sample mean.

$n_1 \rightarrow$ first sample size.

$\bar{x}_2 \rightarrow$ second sample mean.

$n_2 \rightarrow$ second sample size.

ν = degrees of freedom = $n_1 + n_2 - 2$.

If t_{cal} is in accepted region, accept H_0 otherwise reject H_0.

(iii) Paired t-test or t-test for dependent samples.

The test statistic is $t = \bar{d} / (S/\sqrt{n}) \sim t_{n-1}$ where

$$d_i = x_i - y_i \Rightarrow \bar{d} = \frac{1}{n} \Sigma d_i$$

$$S^2 = \frac{1}{n-1} \Sigma (d_i - \bar{d})^2.$$

(iv) t test for correlation coefficient

Null hypothesis: variables are null correlated.

$$H_0 : \rho = 0.$$

Alternative hypothesis: Variables are correlated

$$H_1 : \rho \neq 0.$$

Under H_0, test statistic is

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$$

If t_{cal} is in acceptance region, accept H_0. otherwise reject H_0.

F-test for two population variances.

(i) <u>Null hypothesis</u>: The two population variances are equal $H_0 : \sigma_1^2 = \sigma_2^2$

(ii) <u>Alternative hypothesis</u>: The two population variances are not equal.

(iii) Fix the level of significance (α).

(iv) Compute the test statistic,

$$F = \frac{S_1^2}{S_2^2} \sim F_{(\underset{(v_1)}{n_1-1},\ \underset{(v_2)}{n_2-1})} \quad \text{if } S_1^2 > S_2^2$$

$$= \frac{S_2^2}{S_1^2} \sim F_{(\underset{(v_1)}{n_2-1},\ \underset{(v_2)}{n_1-1})} \quad \text{if } S_2^2 > S_1^2$$

where $S_1^2 = \frac{1}{n_1-1} \Sigma (x_i - \bar{x})^2$

$$S_2^2 = \frac{1}{n_2-1} \Sigma (y_i - \bar{y})^2$$

(v) If F_{cal} is in acceptance region, accept H_0 otherwise reject H_0.

Note: 1) F must be always greater than unity.

2) For a 2 tail test, let α = either 10% or 20%

for left tail critical value, $F_\alpha(v_1, v_2) = \frac{1}{F_{1-\alpha}(v_2, v_1)}$

Anova - I way.

The observations of the data are different with respect to one factor i.e; either row or column.

Ex:

					R_i
X	40	50	80	60	230
Y	45	50	80		175
Z	45	55	85		185
					590 → G.T (Grand Total).

$$\text{Correction factor} = \frac{G^2}{\text{no. of obs}} = \frac{(590)^2}{10}$$

$$= 34810.$$

Sum of squares according to rows

$$(SSR) = \sum \frac{R_i^2}{n_i} - c.f.$$

$$= \frac{(230)^2}{4} + \frac{(175)^2}{3} + \frac{(185)^2}{3} - 34810$$

$$= 31.667.$$

Total sum of squares (TSS)

$$= \sum\sum y_{ij}^2 - C.f$$

$$= 2490.$$

Anova 1-way Table.

	Source of Variance	Sum of Squares (s.s)	degrees of freedom (d.f)	Mean sum of squares (s.s/d.f)	F_{cal}	$F_{tab.}$
1.	Rows.	31.66	3 − 1 = 2	15.83.	22.182	19.4
2.	Error.	2458.34	9 − 2 = 7.	351.143		
	Total.	2490	10 − 1 = 9	276.68		

Rows − Total = error.

degrees of freedom = d.f = no. of rows − 1.

$$F_{cal} = \frac{351.14.}{15.83} \sim F_{(7,2)} = \underline{\underline{22.182}}$$

Ans

Anova 2 way table

	S	R	W	R_i
A	8	4	3	15
B	5	6	7	18
C	9	2	4	15
D	5	8	6	19
C_j	27	20	20	67 → G.T

For testing the quality of saleman A, B, C, D.

H_{01}: The sales of all saleman are equal.

H_{11}: They are not equal

H_{02}: For the quality of seasons (Summer, Rainy, Winter).

H_{12}: All the three seasons are not equal.

Same process of Anova I way table is continued where as a new row of 'column' is added to the table.