Correlation & Regression

- No. of cells = $m \times n$
- No. of conditional distribution = $m+n$

  where m = no. of c^{th} class interval of x

  n = no. of class interval of y.

- No. of marginal distribution in bivariate data = 2

Positive correlation → One variable ↑ another variable ↑

One variable ↓ another variable ↓

Negative correlation → One variable ↑ another variable ↓

One variable ↓ another variable ↑

Spurious / Nonsense correlation → There is no related b/w two variable. This is due to the existence of 3rd variable

Correlation → It is expressed using r

The value of correlation ranges from -1 to 1, both inclusive

$-1 \leq r \leq 1$

Positive correlation → $0 < r < 1$

Perfect Positive → 1

Perfect Negative → $r = -1$

Negative correlation → $-1 < r < 0$

Measures of Correlation -

1. Scatter diagram → It can be applied to any type of correlation
   - linear as well as non-linear.

- The plotted points lie from lower left corner to upper right corner.

★ If it is straight line → Perfect positive

★ If it is straight line moving circle - positive correlation

- The plotted points lie from upper left corner to lower right corner

★ If it is straight line → Perfect Negative

★ If it is straight line moving circle → Negative correlation

The plotted points would be equally distributed without dupicting any particular patter
↳ No relation

- Karl Pearson's product moment correlation coffi

$$r = r_{xy} = \frac{cov(x,y)}{S_x \times S_y}$$

Q. The covariance b/w 2 variables x & y is 8.4 & their variances are 25 & 36 resp^{ly}. Calculate Karl pearson's coefficient of coorelation b/w them

$cov(x,y) = 8.4$ $\quad \sigma^2_x = 25 \quad \therefore \sigma_x = 5$

$\sigma^2_y = 36 \quad \therefore \sigma_y = 6$

$$r_{xy} = \frac{cov(x,y)}{S_x \times S_y}$$

$$= \frac{8.4}{5 \times 6} = 0.28$$

Properties of correlation coefficient

i) The coefficient of correlation is a unit-free measur.

ii) The coefficient of correlation always lies b/w -1 & 1.

iii) If two variables are related by a linear equation then correlation coefficient will always be perfect $+1$ or -1 depends on the sign of slope eqn

$y = a + bx$

$\downarrow$

slope

iv) change of origin → No impact
change of scale → No impact of value but affected by sign.

- If sign of both change of scale are same
$r_{uv} = r_{xy}$

- If sign of both change of scale are different
$r_{uv} = -r_{xy}$

Q- If $u + 5x = 6$ & $3y - 7v = 20$, the correlation coefficient b/w x & $y = 0.58$, then what could be the correlation coefficient between u & v?

$u + 5x = 6$ $\quad 3y - 7v = 20$

$u = 6 - 5x$ $\quad -7v = 20 + 3y$

$-ve$ $\quad +ve$

$r_{uv} = -r_{xy}$

$= -0.58$

Spearman's Rank correlation coefficient

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

Ex → Compute the coefficient of rank correlation between sales & advertisement expressed in thousand of rupees from the following data -

Sales : x	90	85	68	75	82	80	95	70
Advertise : y	7	6	2	3	4	5	8	1
R_x	2	3	8	6	4	5	1	7
R_y	2	3	7	6	5	4	1	8
$d = R_x - R_y$	0	0	1	0	-1	1	0	-1
d^2	0	0	1	0	1	1	0	1

$\Sigma d^2 = 4$

$$r_k = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6(4)}{8(64-1)} = 1 - \frac{24}{504}$$

$$= 1 - 0.0472$$

Coefficient of concurrent deviations $= 0.9528$

$$r_c = \pm\sqrt{\pm\frac{(2c-m)}{m}}$$

C - is the no. of concurrent deviations (same direction)
m - no. of pairs compared, $m = n-1$

Ex - Year : 1990 91 92 93 94 95 96 97
Price : 25 → 28 → 30 - 23 + 35 + 38 + 39 + 42
Demand : 35 34 35 30 29 28 26 23

$dx = + \boxed{+} \boxed{-} + + + +$

$dy = - \boxed{+} \boxed{-} - - - -$

$n = 8 \quad m = n - 1 \quad c = 2.$

$= 7$

$$\frac{2(-m}{m} = \frac{4-7}{7} = -\sqrt{-\frac{(-3)}{7}}$$

Q. For 10 pairs of observations, No. of concurrent deviations was found to be 4. Wt is value of of coefficient of concurrent deviation.

$c = 4 \quad n = 10$

$$r_c = \pm\sqrt{\frac{2c - m}{m}} = -\sqrt{\frac{8-9}{9}} \quad r_c = -\sqrt{\frac{1}{9}}$$

$$= -\frac{1}{3}$$

$cov(x,y) = 0.8$

$x - 10$	$y + 20$
-ve	+ve
-ve.	-0.8

Regression -

Estimation of Y when X is given — Y on X

Y - dependent

X - independent

$y = a + bx$

Estimated of X when Y is given — X on Y

X - dependent
Y - independent

$x = a + by$

Regression

Estimation of Y when X is given

Regression line of Y on X

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

Estimation of X when Y is given

Regression line of X on Y

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

Regression coefficient

$$b_{xy} = \frac{cov(x,y)}{var\ of\ x} \qquad b_{yx} = r \cdot \frac{SD_y}{SD_x}$$

Note - The regression coefficient remain unchanged due to a shift of origin but change due to shift of scale.

Ex - If $u = 2x + 5$ & $v = -3y - 6$ & regression coefficient of Y on X is 2.4, what is the regression coefficient of V on u?

$b_{yx} = 2.4$

$b_{xy} = b_{yx} \cdot \frac{y\,scale}{x\,scale}$

$= 2.4 \frac{(-3)}{2}$

Properties -

ii) The two lines of regression intersect at that point $(\bar{x}, \bar{y})$ mean, where x & y are the ~~variables~~ variables under consideration

iii) $r = \pm \sqrt{b_{yx} \times b_{xy}}$

Note • Product of the regression coefficient must be numerically less than unity.

• This can be applied, unlike correlation ~~for~~ for any type of relationship linear as well as curvilinear.

• The two lines of regression coincide i.e become identical when $r = -1$ or 1

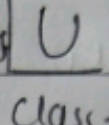
• If $r = 0$, regression lines are perpendicular to each other

Coefficient of determination / explained variance / Accounted variance $= r^2$

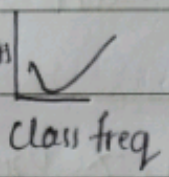
Coefficient of non-determination $= 1 - r^2$

Bell Shaped Curve- 

i) Profit, weight, height, work, mark.

U shaped curve - 

↳ minimum freq near central

L Shaped curve 

Sampling fluctuation is the variation in the value of a statistic computed from different

Sampling Distribution is the probability distribution of given statistic.

The mean of the statistic, as obtained from its sampling distribution is known as "Expectation"

The Standard deviation of the statistic is known as the "Standard Error".

- Standard Error can be regarded as a measure of the precision achieved by sampling.