# UCS 2H27 - Business Intelligence

## Unit 3

## Understanding Probability & Statistics

### Assignment Numericals

1) A research investigator collected data on savings and investment from 16 households. Savings showed a mean of Rs 6565.0 and a variance of Rs 250.00. As against this, mean investment was found at Rs 4525.00 & variance as Rs 520.00. If the coefficient of correlation between savings and investments is 0.67, Find the most appropriate value of savings against an investment of Rs 9000 and that of investment against a savings of Rs 5600

**Ans**

Savings: $\bar{X} = Rs. 6565$

Investment: $\bar{Y} = Rs\ 4525.00$

$\sigma^2(X) = 250.0$

$\sigma^2(Y) = 520$

$r = 0.67$

$Y = bX + a$

$b = r * \left(\frac{\sigma(Y)}{\sigma(X)}\right)$

$a = \bar{Y} - b\bar{X}$

$b = 0.67 \left(\frac{\sqrt{520}}{\sqrt{250}}\right)$ $\cancel{= 0.9663}$ $= 0.9663$

$= 1.4422 \times 0.67$

$= 0.9662$

$$a = \overline{Y} - b\overline{X}$$ $\overline{X} = 4525$

$$= 4525 - (0.9662)(6565)$$

$$= 4525 - (6343.103)$$

$a \simeq -1818.103$

$$\boxed{Y = -1818.103 + 0.9662X}$$

when investment is 9000

$$9000 = -1818.103 + 0.9662X$$

$$X = \frac{9000 + 1818.103}{0.9662}$$

when savings is 5600

$$Y = -1818.103 + 0.9662(5600)$$

(Q) Out of two lines of regression given by $x+2y-5=0$ and $2x+3y-8=0$, which one is the regression line of X and Y. Use the equations to find the means of X and Y. If the variance of X is 12, find the variance of Y.

Ans

$$x + 2y - 5 = 0 \rightarrow (1)$$

$$2x + 3y - 8 = 0 \rightarrow (2)$$

Assume that (1) is the regression line of y on x.

$$2y = 5 - x$$

$$y = \frac{-x}{2} + \frac{5}{2}$$

i.e $\boxed{b_{yx} = -1/2}$

Assume that 2 is the regression line of x on y

$$2x = 8 - 3y$$

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

$$\boxed{b_{xy} = -3/2}$$

$$r^2 = b_{xy} \cdot b_{yx} = \frac{-1}{2} \times \frac{-3}{2} = \frac{+3}{4} < 1$$

our assumption is correct

$$r^2 = 0.75$$

$$r = \sqrt{0.75} = \pm 0.866$$

both $b_{yx}$ and $b_{xy}$ are -ve

$$\Rightarrow \boxed{r = -0.866}$$

Variance of $X = 12$

Variance of $Y = ?$

$$b_{yx} = r \cdot \frac{\sigma_Y}{\sigma_X}$$

$$\cancel{-}0.5 = \cancel{-}0.866 \times \left(\frac{\sigma_Y}{\sqrt{12}}\right)$$

$$\sigma_Y = 2$$

variance $\boxed{\sigma^2_Y = 4}$

Finding the mean

$\bar{x} + 2\bar{y} = 5 \quad \times 2$

$2\bar{x} + 3\bar{y} = 8$

$2\bar{x} + 4\bar{y} = 10$

$2\bar{x} + 3\bar{y} = 8$

$\boxed{\bar{y} = 2}$

$\bar{x} + 4 = 5$

$\boxed{\bar{x} = 1}$

(3) In a partially destroyed laboratory on the analysis of correlation data, only the following results are legible.

variance of $x = 9$

regression equations are $8x - 10y + 66 = 0$ and

$40x - 18y = 214$

A. Mean $\quad 8x - 10y + 66 = 0 \quad$ — (1)

$40x - 18y = 214 \quad$ — (2)

Solving (1), (2)

$$\boxed{\begin{aligned}\bar{x} &= 13\\ \bar{y} &= 17\end{aligned}}$$

B. Correlation Coefficient $\quad$ Let $8x - 10y + 66 = 0$ be the regression line of y on x.

$10y = 8x + 66$

$y = \frac{8}{10}x + \frac{66}{10}$

$\boxed{b_{yx} = 0.8}$

Let $40x - 18y = 214$ be the regression line of $x$ on $y$.

$$40x = 18y + 214$$

$$x = \frac{18}{40}y + \frac{214}{40}$$

$$b_{xy} = \frac{9}{20}$$

correlation coefficient : $r^2 = 0.8 \times \frac{9}{20} = \frac{\cancel{8}^{4}}{10} \times \frac{9}{\cancel{20}_{10}} = \frac{36}{100} = 0.36$

$\boxed{r = \pm 0.6}$

both $b_{yx}$ and $b_{xy}$ are positive $\boxed{r = +0.6}$

$\sigma^2 x = 9 \quad \sigma x = 3$

$$b_{yx} = r \cdot \frac{\sigma y}{\sigma x}$$

$$\frac{\cancel{8}^{4}}{\cancel{10}} = \frac{\cancel{6}^{2}}{\cancel{10}} \times \frac{\sigma y}{\cancel{3}_{1}}$$

$\boxed{\sigma y = 4}$

(4) For the following data, find the equations of the regression lines.

| | Mark in Maths | Mark in English |
|---|---|---|
| Mean | 62.5 | 39 |
| SD | 9.5 | 10 |

Coefficient of Correlation between marks in Math & English $= 0.60$

(a) Estimate the marks in English when marks in Math is 70.

(b) Estimate the marks in Math corresponding to 54 marks in English.

$X \rightarrow$ Mathematics $\qquad Y \rightarrow$ English

$\bar{X} \Rightarrow 62.5 \qquad \bar{Y} = 39$

$\sigma x = 9.5 \qquad \sigma y = 10$

$r = 0.60$

<u>equations of regression lines</u>

$\boxed{X \text{ on } Y}$

$$b_{xy} = r \cdot \frac{\sigma x}{\sigma y}$$

$$= 0.6 \left( \frac{9.5}{10} \right)$$

$$= 0.57$$

$$\bar{X} = b_{xy} \cdot \bar{Y} + c$$

$$62.5 = 0.57(39) + c$$

$$\underline{c = 40.27}$$

The line is: $\boxed{X = 0.57Y + 40.27}$

$\boxed{Y \text{ on } X}$

$$b_{yx} = \frac{r \cdot \sigma y}{\sigma x}$$

$$= 0.6 \left( \frac{10}{9.5} \right)$$

$$= 0.63$$

$$\bar{Y} = b_{yx} \bar{X} + c$$

$$39 = 0.63(62.5) + c$$

$$c = -0.375$$

$\boxed{Y = 0.63X - 0.375}$

(8) Given the following data:

| | X | Y |
|---|---|---|
| Mean | 36 | 85 |
| std | 11 | 8 |

correlation coeff between X and Y = 0.6

Find the 2 regression equations

Estimate X when Y = 75

$\boxed{\text{X on Y}}$

$$b_{xy} = r \cdot \frac{\sigma x}{\sigma y}$$

$$= 0.6 \left(\frac{11}{8}\right)$$

$$= 0.825$$

$$\bar{X} = b_{xy} \bar{Y} + c$$

$$36 = (0.825)(85) + c$$

$\boxed{\text{Y on X}}$

$$b_{yx} = r \cdot \frac{\sigma y}{\sigma x}$$

$$= 0.6 \left(\frac{8}{11}\right)$$

$$= 0.436$$

$$\bar{Y} = (0.436)(36) + c$$

(9) Suppose that for a certain population, we can predict log earnings from log height as follows:

→ A person who is 66 inches tall is predicted to have earnings of \$30,000.

→ Every increase of 1% in height corresponds to a predicted increase of 0.8% in earnings

→ The earnings of approx 95% of people fall within a factor of 1.1 of predicted values.

Give the eqn. of the regression line and the residual standard deviation of the regression.

Ans

$y = mx + c$

$c = y - mx$

$c = \log(30{,}000) - \left(\frac{0.008}{0.01}\right) \log 66$

$= 6.957$

$\log(\text{earnings}) = 6.957 + \left(\frac{0.008}{0.01}\right) * \log(\text{height})$

$sd = 0.1 \times \frac{0.5}{0.95}$

# Bayes Theorem

$$P(B_i|A) = \frac{P(B_i) \times P(A|B_i)}{\sum_{i=1}^{n} P(B_i)\,P(A|B_i)}$$

3 machines $M_1$, $M_2$ and $M_3$ produce identical items. Of their respective outputs 5%, 4%, 3% of items are faulty. On a certain day $M_1$ has produced 25%, $M_2$ has produced 30% and $M_3$ the rest. An item selected at random is found to be faulty. What are the chances it was produced by the machine with the highest op

$P(M_1) = 0.25$

$P(M_2) = 0.35$

$P(M_3) = 0.45$

$P(F|M_1) = 0.05$

$P(F|M_2) = 0.04$

$P(F|M_3) = 0.03$

$$P(M_3|F) = \frac{P(F|M_3)\,P(M_3)}{P(M_3)\,P(F|M_3) + P(M_2)\,P(F|M_2) + P(M_1)\,P(F|M_1)}$$

# Correlation Coefficient

Ex) Compute the coefficient of correlation between x and y with the following data.

| | | | | | | |
|---|---|---|---|---|---|---|
| x | 1 | 3 | 5 | 7 | 8 | 10 |
| y | 8 | 12 | 15 | 17 | 18 | 20 |

| x | y | $x_i^2$ | $y_i^2$ | $x_iy_i$ |
|---|---|---|---|---|
| 1 | 8 | 1 | 64 | 8 |
| 3 | 12 | 9 | 144 | 36 |
| 5 | 15 | 25 | 225 | 75 |
| 7 | 17 | 49 | 289 | 119 |
| 8 | 18 | 64 | 324 | 144 |
| 10 | 20 | 100 | 400 | 200 |

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

$$= \frac{6.582 - 34 \times 90}{\sqrt{(6 \times 248 - (34)^2)(6(1446) - (90)^2)}}$$

$$r_{xy} = \underline{0.9879}$$

$$1 - \frac{6\sum (x_i - y_i)^2}{n(n^2-1)}$$

Ex2 Compute the correlations between X & Y

X : 65 67 66 71 67 70 68 69

Y : 67 68 68 70 64 67 72 70

change origin to approx mean

| X | Y | u | v | $u^2$ | $v^2$ | uv |
|---|---|---|---|---|---|---|
| 65 | 67 | -3 | -1 | 9 | 1 | 3 |
| 67 | 68 | -1 | 0 | 1 | 0 | 0 |
| 66 | 68 | -2 | 0 | 4 | 0 | 0 |
| 71 | 70 | 3 | 2 | 9 | 4 | 6 |
| 67 | 64 | -1 | -4 | 1 | 16 | 4 |
| 70 | 67 | 2 | -1 | 4 | 1 | -2 |
| 68 | 72 | 0 | 4 | 0 | 16 | 0 |
| 69 | 70 | 1 | 2 | 1 | 4 | 2 |

$$r = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}}$$