# Correlation & Regression

## # correlation

→ Degree of association between two variable

→ correlation can be negative positive or zero

→ $-1 \leq r \leq 1$

→ for Direction Relation b/w $x$ & $y$

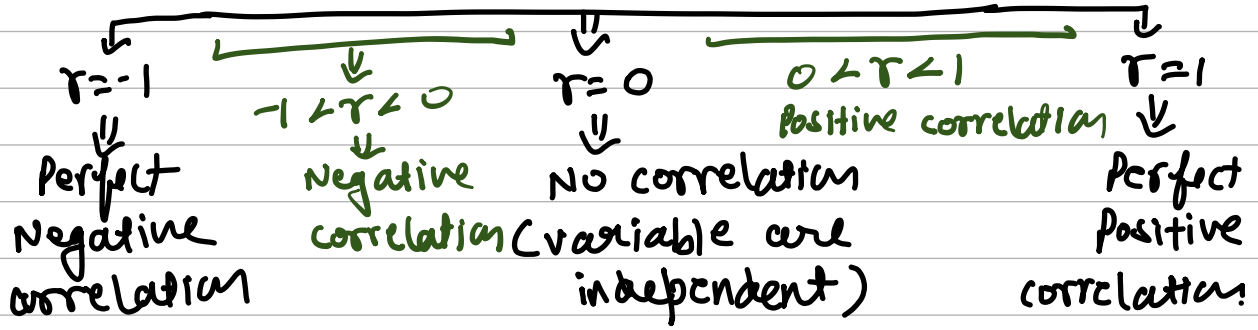
$x \uparrow y \uparrow$ or $x \downarrow y \downarrow$

$r$ is positive

→ for inverse Relation b/w $x$ & $y$

$x \downarrow y \uparrow$ or $x \uparrow y \downarrow$

$r$ is negative

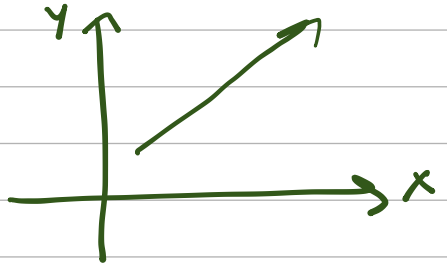


Methods of calculating correlation

- Graphical method
  - Scattered Diagram method
- non Graphical method
  - Karl Pearson method
  - Spearman method
  - concurrent Deviation

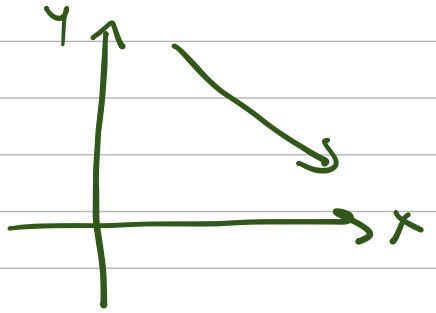
# Scattered Diagram method

Plot the values of two variable on graph



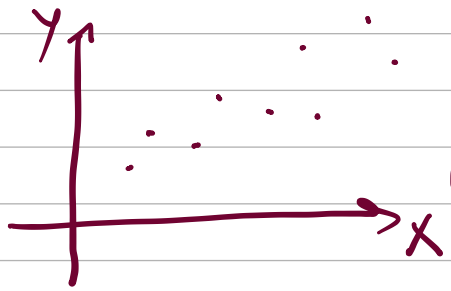

for straight line
(lower left to upper right)

$r = 1$

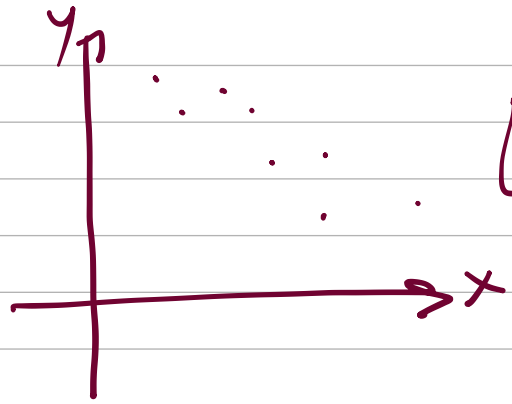



for straight line
(upper left to lower right)

$r = -1$

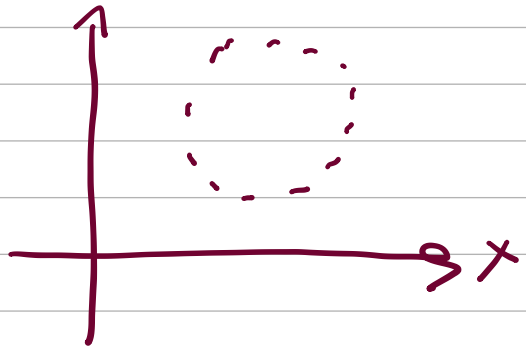



$0 < r < 1$ (No straight line)

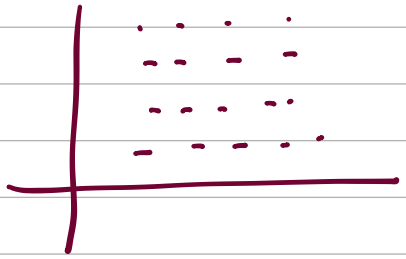



$-1 < \gamma < 0$

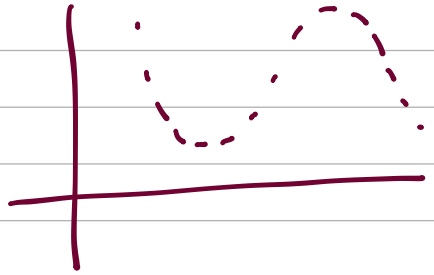
No straight Line




$\gamma = 0$



$\gamma = 0$



$\gamma = 0$

→ Exact magnitude can not be calculated in scattered Diagram method

# Karl Pearson's coefficient correlation

$$r = \frac{cov.(x,y)}{\sigma_x \sigma_y}$$

or

$$r = \frac{\Sigma (x_i - \bar{x}).(y_i - \bar{y})}{N \sigma_x \sigma_y}$$

$$r = \frac{\Sigma (x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \times \Sigma (y_i - \bar{y})^2}}$$

or

$$r = \frac{\Sigma xy - \frac{\Sigma x \times \Sigma y}{N}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{N}} \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{N}}}$$

→ covariance $= cov(x,y) = \frac{\Sigma (x_i - \bar{x}).(y_i - \bar{y})}{N}$

→ $cov(x,y)$ can be any Real number (Negative, Positive & zero)

→ $cov(x,y)$ does not change with origin but it changes with change of scale

→ Correlation does not change with change of origin

i.e. $u_i = x_i - A$ & $v_i = y_i - B$

then
$$r = \frac{\Sigma uv - \frac{\Sigma u \times \Sigma v}{N}}{\sqrt{\Sigma u^2 - \frac{(\Sigma u)^2}{N}}\sqrt{\Sigma v^2 - \frac{(\Sigma v)^2}{N}}}$$

→ Correlation does not change with change of scale (provided scale is positive)

→ when scale is negative, the sign of correlation may change but magnitude does not change

e.g. $r(x,y) = 0.5$

If $u_i = 2x_i$ & $v_i = 4y_i$

then $r(u_i, v_i) = (+)(+)\,0.5 = 0.5$

e.g. $r(x,y) = 0.6$

If $u_i = -2x_i$ & $v_i = 3y_i$

then $r(u,v) = (-)(+)\,0.6 = -0.6$

e.g. If $r(x,y) = 0.7$

& $u_i = -2x_i$ & $v_i = -4y_i$

then $r(u,v) = (-)(-)\,0.7 = (+)\,0.7$

# Spearman's Rank correlation

→ This method used for qualitative characters & level of agreements & disagreements b/w opinions of Judges

→ when no numbers repeat

$$r = 1 - \frac{6 \Sigma D^2}{N^3 - N}$$

when some numbers repeat

→

$$r = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{N^3 - N}$$

→ $\Sigma D = 0$

| X | Y | $R_1$ | $R_2$ | $D = R_1 - R_2$ | $D^2$ |
|---|---|---|---|---|---|
| | | | | | |
| | | | | | |
| | | | | 0 | |

# Concurrent Deviation method

Concurrent Deviation

⇓

when x & y both increase or both decrease

$$r = \pm\sqrt{\pm\left(\frac{2c-m}{m}\right)}$$

where $c$ = Total no of concurrent Deviations

$m = n - 1$

| | x | y | | Concurrent Deviation |
|---|---|---|---|---|
| + | 10 | 12 | + | yes (+) |
| + | 12 | 15 | - | No (-) |
| - | 15 | 14 | + | No (-) |
| + | 14 | 15 | + | yes (+) |
| - | 16 | 18 | + | No (-) |
| | 12 | 20 | | |

$c = 2$

$m = 6 - 1 = 5$

$$r = \sqrt{\frac{2(2)-5}{5}}$$

$$r = \sqrt{\frac{-1}{5}}$$

$$r = -\sqrt{\frac{1}{5}}$$

→ when $\frac{2c-m}{m}$ is negative then (-) sign will be taken out from $\sqrt{}$ sign

# Bivariate Frequency Distribution Table

| maths \ stats | 0-5 | 5-10 | 10-15 | 15-20 | Total |
|---|---|---|---|---|---|
| 0-10 | 2 | 5 | 1 | 0 | 8 |
| 10-20 | 3 | 2 | 3 | 4 | 12 |
| 20-30 | 4 | 1 | 2 | 6 | 13 |
| Total | 9 | 8 | 6 | 10 | 33 |

Frequency Distribution of marks of maths

| marks | No of students |
|---|---|
| 0-10 | 8 |
| 10-20 | 12 |
| 20-30 | 13 |
| | 33 |

This is called marginal distribution

Frequency Distribution of marks in maths when score in stats is '5-10'

| marks | No of students |
|---|---|
| 0-10 | 5 |
| 10-20 | 2 |
| 20-30 | 1 |
| | 8 |

This is called conditional distribution

# For Bivariate Distribution table of "m x n"

Total cells = m x n

Total marginal Distribution = 2

Total conditional Distribution = m + n

# coefficient of Determination

$$= r^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

# coefficient of Non Determination

$$= 1 - r^2$$

# Regressions

→ Establishing mathematical relation b/w two variables (Independent & Dependent)

→ Prediction of Dependent variable

→ for Linear Regression Least Square method is used

There two Linear Regression Lines

| | when y depends on x | when x depends on y |
|---|---|---|
| Standard form | $y = a + bx$ | $x = a + by$ |
| How to find = ? | use formula $y - \bar{y} = b_{yx}(x - \bar{x})$ | use formula $x - \bar{x} = b_{xy}(y - \bar{y})$ |
| slope of line | slope = $b_{yx}$ | slope = $1/b_{xy}$ |

$b_{yx}$ & $b_{xy}$ are know as Regression coefficients

# Calculation of Regression coefficients

$$b_{yx} = \frac{cov(x,y)}{\sigma_x^2}$$

$$b_{yx} = \frac{\Sigma(x_i - \bar{x})\cdot(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$b_{yx} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{N}}{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

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

$$b_{xy} = \frac{\Sigma(x_i - \bar{x})\cdot(y_i - \bar{y})}{\Sigma(y_i - \bar{y})^2}$$

$$b_{xy} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{N}}{\Sigma y^2 - \frac{(\Sigma y)^2}{N}}$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

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

→ 'r' will be positive if both $b_{yx}$ & $b_{xy}$ are positive

→ 'r' will be negative if both $b_{yx}$ & $b_{xy}$ are negative.

→ $b_{yx} \times b_{xy} \leq 1$

# $r \leq \frac{b_{yx} + b_{xy}}{2}$

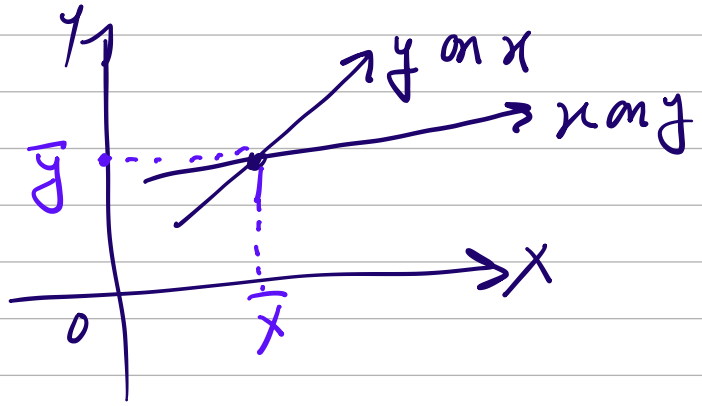
# If Regression Line y on x is: $ax+by+C=0$

then $b_{yx} = -\frac{a}{b}$

# If Regression Line x on y is $Ax+By+C=0$

then $b_{xy} = -\frac{B}{A}$

# Two Regression lines intersect each other at $(\bar{x}, \bar{y})$




# Regression coefficients does not change with change of origin

$u_i = x_i - A$ & $v_i = y_i - B$

then $b_{yx} = b_{uv}$ & $b_{xy} = b_{uv}$

# Regression coefficients changes with change of scale

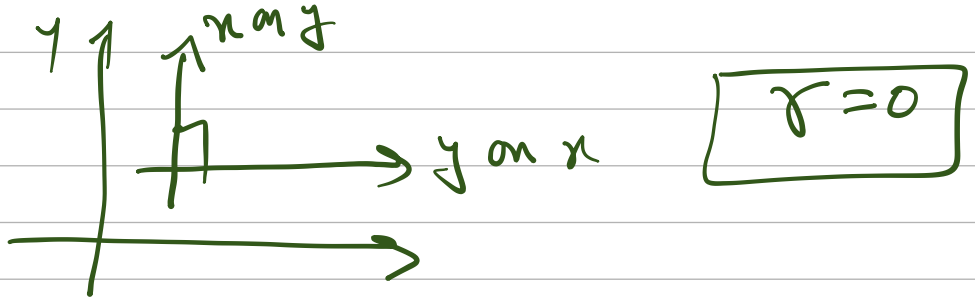
$u_i = a x_i + b$ & $v_i = c y_i + d$

So $b_{vu} = \frac{c}{a} \times b_{yx}$

or

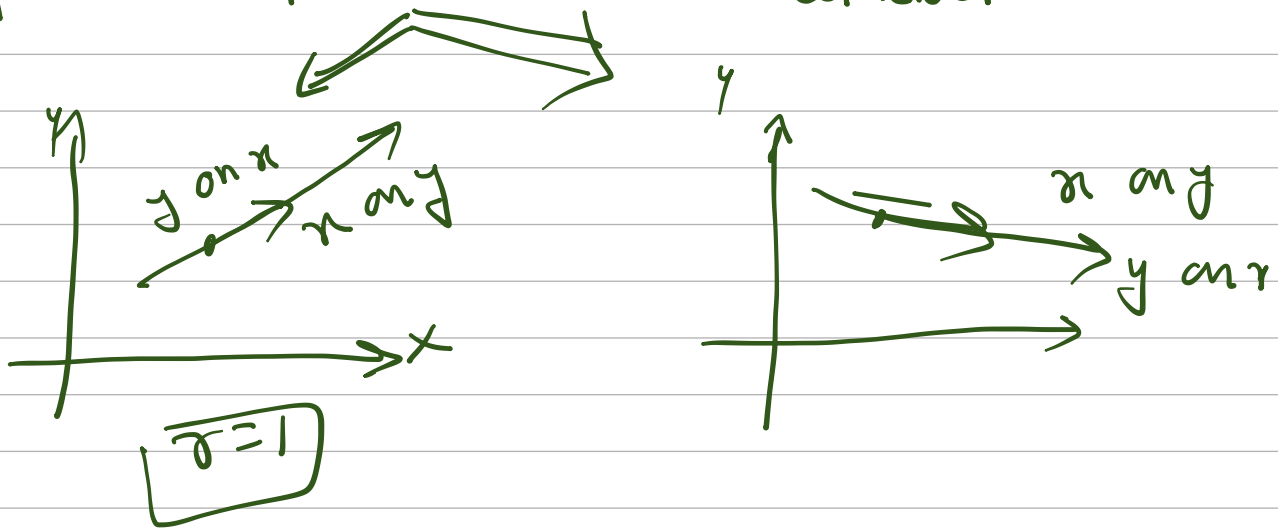
$b_{vu} = \frac{\text{scale of } y}{\text{scale of } x} \times b_{yx}$

# when two lines are perpendicular




$r = 0$

# when two lines are coincident




$r = 1$