Sec 2.4 - Correlation Coefficient.

Let X and Y have joint pdf $f(x,y)$.
Assume all expected values exist (finite).

Let $\mu_1 = E[X]$, $\sigma_1^2 = Var[X]$

Let $\mu_2 = E[Y]$, $\sigma_2^2 = Var[Y]$

def Covariance of X and Y

$$Cov(X,Y) = E[(X-\mu_1)(Y-\mu_2)]$$

$$= E[XY - \mu_2 X - \mu_1 Y + \mu_1\mu_2]$$

$$= E[XY] - \mu_2 E[X] - \mu_1 E[Y] + \mu_1\mu_2$$

$$= E[XY] - \mu_1\mu_2 - \mu_1\mu_2 + \mu_1\mu_2$$

$$= E[XY] - \mu_1\mu_2$$

OR $E[XY] - E[X]E[Y]$

def] Correlation Coefficient of X and Y

If $\sigma_1 > 0$ and $\sigma_2 > 0$ then the correlation coefficient of X and Y is

$$\rho = \frac{E[(X-\mu_1)(Y-\mu_2)]}{\sigma_1 \sigma_2}$$

$$= \frac{Cov(X,Y)}{\sigma_1 \sigma_2}$$

Note: $E[XY] = \mu_1 \mu_2 + \rho \sigma_1 \sigma_2$

$= \mu_1 \mu_2 + Cov(X,Y)$

ex] Let the RVs X and Y have joint pdf

$$f(x,y) = \begin{cases} x+y, & 0<x<1, \ 0<y<1 \\ 0, & \text{o.w.} \end{cases}$$

Compute ρ.

$$\mathbb{E}[XY] = \int_0^1 \int_0^1 xy(x+y)\,dx\,dy$$

$$= \int_0^1 \int_0^1 (x^2y + xy^2)\,dx\,dy$$

$$\vdots$$

$$= 1/3$$

$$\mu_1 = \mathbb{E}[X] = \int_0^1 \int_0^1 x(x+y)\,dx\,dy$$

$$= \int_0^1 \int_0^1 x(x+y)\,dy\,dx$$

$$\mu_2 = \mathbb{E}[Y] = 7/12$$

marginal pdf. of X

$$f_X(x) = \int_0^1 (x+y)\,dy$$

$$\mathbb{E}[X] = \int_0^1 x\left(\int_0^1 (x+y)\,dy\right)dx$$

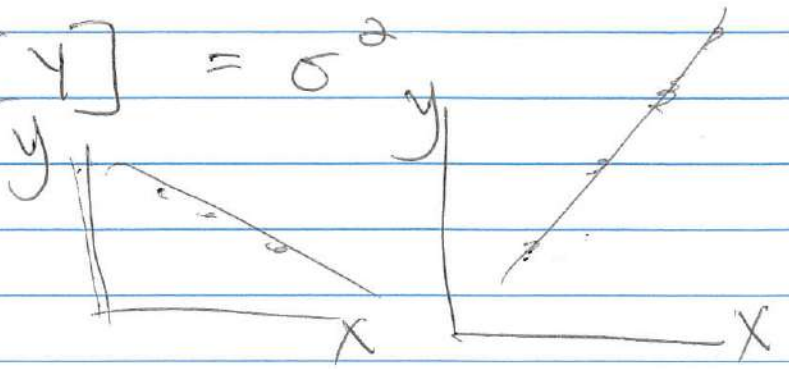
$$\mathbb{E}[X^2] = \int_0^1 \int_0^1 x^2(x+y)\,dx\,dy = \frac{5}{12}$$

$$= \mathbb{E}[Y^2].$$

$$\text{Var}[X] = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

$$= \text{Var}[Y] = \sigma^2$$

$$\sigma = \sqrt{\frac{11}{144}}$$



$$\text{Cov}(X,Y) = E[XY] - \mu_1 \mu_2$$

$$= \frac{1}{3} - \left(\frac{7}{12}\right)\left(\frac{7}{12}\right) = -\frac{1}{144}$$

Correlation Coefficient of X and Y.

$$\rho = \frac{-\frac{1}{144}}{\sqrt{\frac{11}{144}}\sqrt{\frac{11}{144}}} = -\frac{1}{11}$$

Properties

(1) $-1 \leq \rho \leq 1$

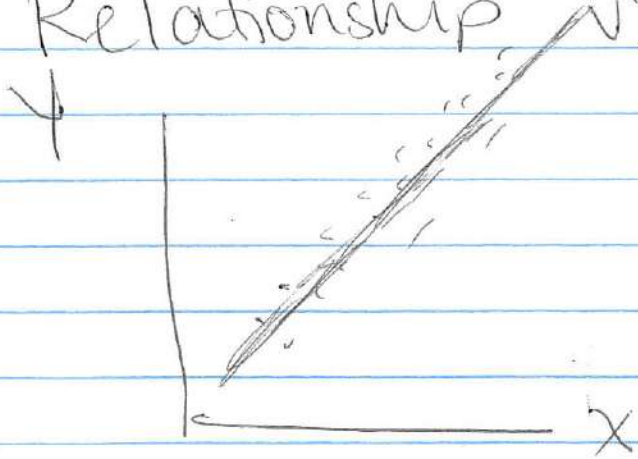
(2) measure intensity of the concentration of the probability for X and Y about a line

(3) If $\rho = 1$, then we can make a line $y = a + bx$, $b > 0$ such that it contains all of the distribution of X and Y.

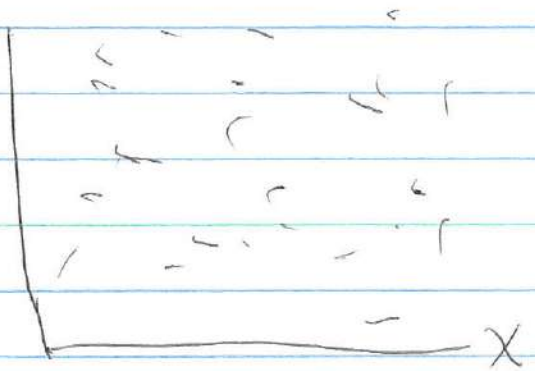
(4) If $\rho = -1$, then we can make a line $y = a + bx$, $b < 0$ such that it contains all of the probability of the dist. of X and Y

(5) If X and Y are independent (sec 2.5), then $\rho = 0$.

Relationship via data.

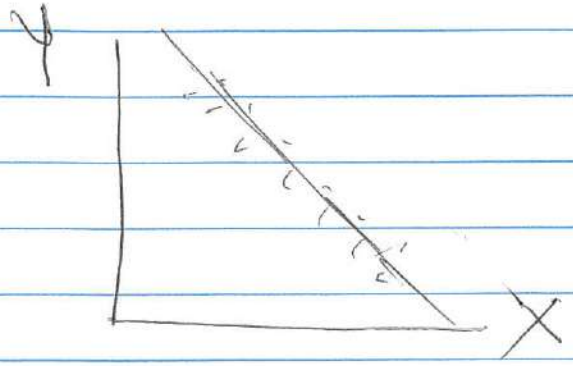


← r close to 1 strong relationship.



vague cloud.

r close or = 0

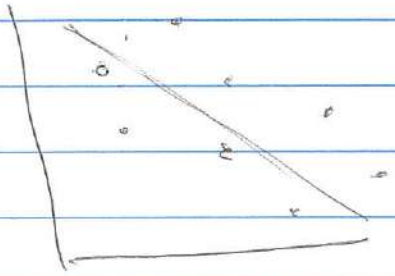


r close to −1

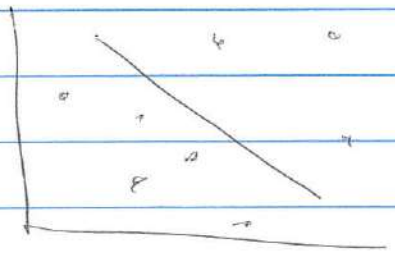
strong relationship

$-1 \leq r \leq -0.7$ strong neg.

$-0.7 \leq r \leq -0.3$ moderate neg.



$-0.3 \leq r \leq -0.1$ weak neg.



$-0.1 \leq r \leq 0.1$ No rel.

$0.1 \leq r \leq 0.3$ weak pos.

$0.3 \leq r \leq 0.7$ moderate pos

$0.7 \leq r \leq 1$ strong pos.

Thm 1.

Suppose (X,Y) have joint dist. $f(x,y)$

Let $\mu_1 = \mathbb{E}[X]$

$\mu_2 = \mathbb{E}[Y]$

$\sigma_1^2 = \text{Var}[X] > 0$ and finite

$\sigma_2^2 = \text{Var}[Y] > 0$ and finite.

Let ρ be the correlation coeff. between X and Y.

If $\mathbb{E}[Y|X]$ is linear in X

then $\mathbb{E}[Y|X] = \mu_2 + \rho \cdot \frac{\sigma_2}{\sigma_1}(X-\mu_1)$

and $\mathbb{E}[\text{Var}(Y|X)] = \sigma_2^2(1-\rho^2)$

If $\mathbb{E}[Y|X=x] = a + bx$

slope $= b = \rho \frac{\sigma_2}{\sigma_1}$ and

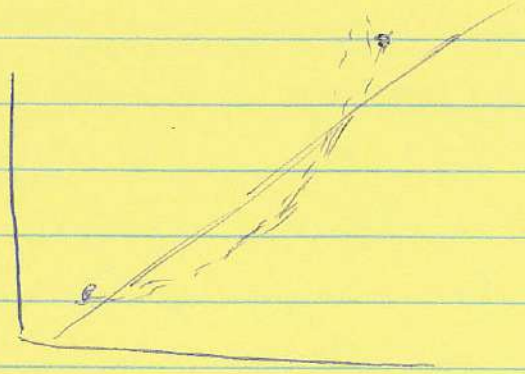
y-int $a = \mu_2 - \rho \frac{\sigma_2}{\sigma_1}\mu_1$

$$\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(X - \mu_1)$$

$$= \mu_2 + \rho \frac{\sigma_2}{\sigma_1} X - \rho \frac{\sigma_2}{\sigma_1} \mu_1$$

$$= \left[\mu_2 - \rho \frac{\sigma_2}{\sigma_1} \mu_1\right] + \left[\rho \frac{\sigma_2}{\sigma_1}\right] X$$

a $\quad + \quad b \quad X$



For data:

$$\hat{y} = a + bx$$

$b = r \cdot \frac{s_y}{s_x}$

$\hat{y}$: predicted value

$a = \bar{y} - r \cdot \frac{s_y}{s_x} \cdot \bar{x}$

$\bar{y}$: mean of y from data.

s_y, s_x: std dev.

$\bar{x}$: mean of x from data