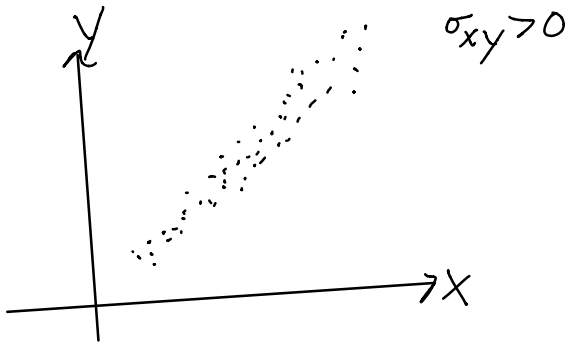
Lecture 15.

More on Covariance and Correlation

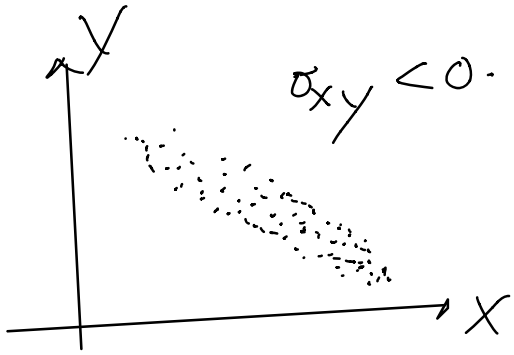
- Recall that Covariance is defined as

$$\text{cov}(X,Y) = \sigma_{XY} = E[XY] - E[X]E[Y].$$

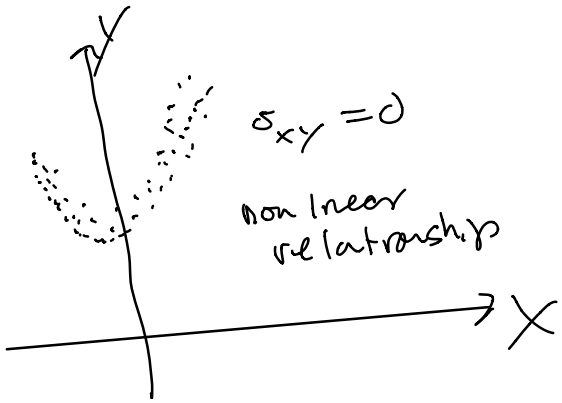
- σ_{XY} measures linear relationship between X and Y.



$\sigma_{XY} > 0$

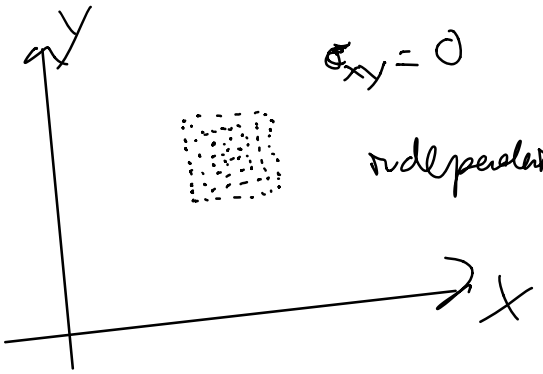


$\sigma_{XY} < 0.$



$\sigma_{XY} = 0$

non linear relationship



$\sigma_{XY} = 0$

independent

- The correlation of X and Y also measures linear relationships between X and Y.
- Correlation may be easier to interpret.

Defn: Given CRV's X, Y, jointly distributed. Then, the correlation is defined to be

$$\rho_{XY} := \frac{\mathrm{cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

(σ_x, σ_y: standard dev. of X and Y)

Exercise: Show that for all X, Y,

$$-1 \leq \rho_{XY} \leq +1.$$

- If ρ_{XY} is close to $+1$, then the possible values of X and Y tend to be near a line with positive slope with high probability.
- If $\rho_{XY} = 1$, then $Y = aX + b$ for $a > 0$.
- If $\rho_{XY} = -1$ then $Y = -aX + b$ for $a > 0$.

Linear Functions of Random Variables

Let $X_1, X_2, \ldots, X_n$ be random variables.

For constants $c_0, c_1, \ldots, c_n$,

$$Y = c_0 + c_1 X_1 + c_2 X_2 + \ldots + c_n X_n.$$

is a linear function of $X_1, \ldots, X_n$, and Y is itself a random variable.

Fact: $E[Y] = c_0 + c_1 E[X_1] + \ldots + c_n E[X_n]$.

We can verify this in the case of 2 RVs:

$$E[a + bX + cY] := \iint (a + bx + cy) f_{xy}(x,y)\,dx\,dy.$$

$$= \iint a f_{xy}(x,y)\,dx\,dy + \iint bx f_{xy}\,dx\,dy + \iint cy f_{xy}\,dx\,dy.$$

$$= a \underbrace{\iint f_{xy}\,dx\,dy}_{=1} + b \underbrace{\iint x f_{xy}\,dx\,dy}_{E[X]} + c \underbrace{\iint y f_{xy}\,dx\,dy}_{E[Y]}.$$

$$= a + bE[X] + cE[Y] \text{ as required.}$$

In general, if $Y = c_0 + c_1 X_1 + \ldots + c_n X_n$, then the variance is:

$$V(Y) = c_1^2 V(X_1) + c_2^2 V(X_2) + \cdots + c_n^2 V(X_n) + 2\sum_{i<j} c_i c_j \sigma_{X_i X_j}$$

Note: If $X_1, X_2, X_3 \ldots X_n$ are independent, then for any i, j, $\sigma_{X_i X_j} = 0$ and so

$$V(c_0 + c_1 X_1 + \cdots + c_n X_n) = \sum_{i=1}^{n} c_i^2 V(X_i).$$

As a special case: if $X_1, \ldots, X_n$ are RV's, then the average is

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

If X_i are such that $E[X_i] = \mu_i$, then

$E[\bar{X}]$ is the average of the μ_i's $\frac{\mu_1 + \cdots + \mu_n}{n}$.

For $X_1 \ldots X_n$ independent, $V(X_i) = \sigma_i^2$

$$V(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2$$

So if $\sigma_i^2 = \sigma^2 \ \forall i$, $V(\bar{X}) = \frac{\sigma^2}{n}$

Reproductive Property of the Normal Distribution

Suppose that $X_1 \dots X_n$ are independent, normal RV's with $E[X_i] = \mu_i$ and $V(X_i) = \sigma_i^2$, then

$$Y = c_0 + c_1 X_1 + \dots + c_n X_n$$

is again a normal RV with $E(Y) = c_0 + c_1\mu_1 + c_2\mu_2 + \dots + c_n\mu_n$.

and $V(Y) = \sum_{i=1}^{n} c_i^2 \sigma_i^2$.

Special case: $X_1, \dots, X_n$ are all $N(\mu, \sigma^2)$

Then $Y = X_1 + \dots + X_n = N(n\mu, n\sigma^2)$ and

$\bar{X} = \frac{X_1 + \dots + X_n}{n} = N\left(\mu, \frac{\sigma^2}{n}\right)$.

This may be viewed as a special case of the Central Limit Theorem: Let $X_1, X_2, \dots$ be a sequence of RV's, independent, with the same distribution. Then

$$\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\,\sigma^2} \xrightarrow[n \to \infty]{} N(0, 1)$$

Example: Suppose that water bottles are filled to an average of 591 ml with $\sigma = 5$ mL. Suppose the volume of bottles are independent normal RVs. Given 10 bottles, what is the prob. that the average volume is less than 581 ml?

Let X_i = the volume of the ith bottle.

So $X_i = N(591, \sigma^2 = 25)$.

Let $\bar{X} = \frac{1}{10}(X_1 + \cdots + X_{10})$.

want $P(\bar{X} \leq 585)$. Note: $\bar{X} = N\left(591, \frac{\sigma^2}{10}\right) = N\left(591, \frac{25}{10}\right)$
$= N(591, 2.5$

So $P(\bar{X} \leq 585) = P\left(Z < \frac{585 - 591}{\sqrt{2.5}}\right)$

$= \Phi(-3.79)$ [also $= 1 - P(Z < +3.79)$]

$= 0.000075.$