Math 170E: Winter 2023

Lecture 19, Mon 27th Feb

The correlation coefficient

Midterm 2

- This Friday, 12-12:50pm MS4000A (here).
- Content: Focus on lectures $10 \le X \le 18$

Last Friday

- 2 sided cheat sheet, simple calculator.
- Practice on Canvas
- No lecture Friday
- My OH (this week): Wed 5:30-6:30, Thu 4:30-5:30

Last time:

- Let X, Y be a pair of discrete random variables taking values in sets $S_X, S_Y \subseteq \mathbb{R}$

- Let $S = S_X \times S_Y$ and let X, Y have joint PMF $p_{X,Y}(x, y)$

- If $g : S \to \mathbb{R}$, we define

$$\mathbb{E}[g(X, Y)] = \sum_{(x,y)\in S} g(x, y) p_{X,Y}(x, y)$$

$\to p_{X,Y}(x,y) = p_X(x) p_Y(y)$

- If X, Y are independent and $g : S_X \to \mathbb{R}$, $h : S_Y \to \mathbb{R}$, then

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\}$

- We have the Cauchy-Schwarz inequality:

$$\boxed{|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}}$$

Today:

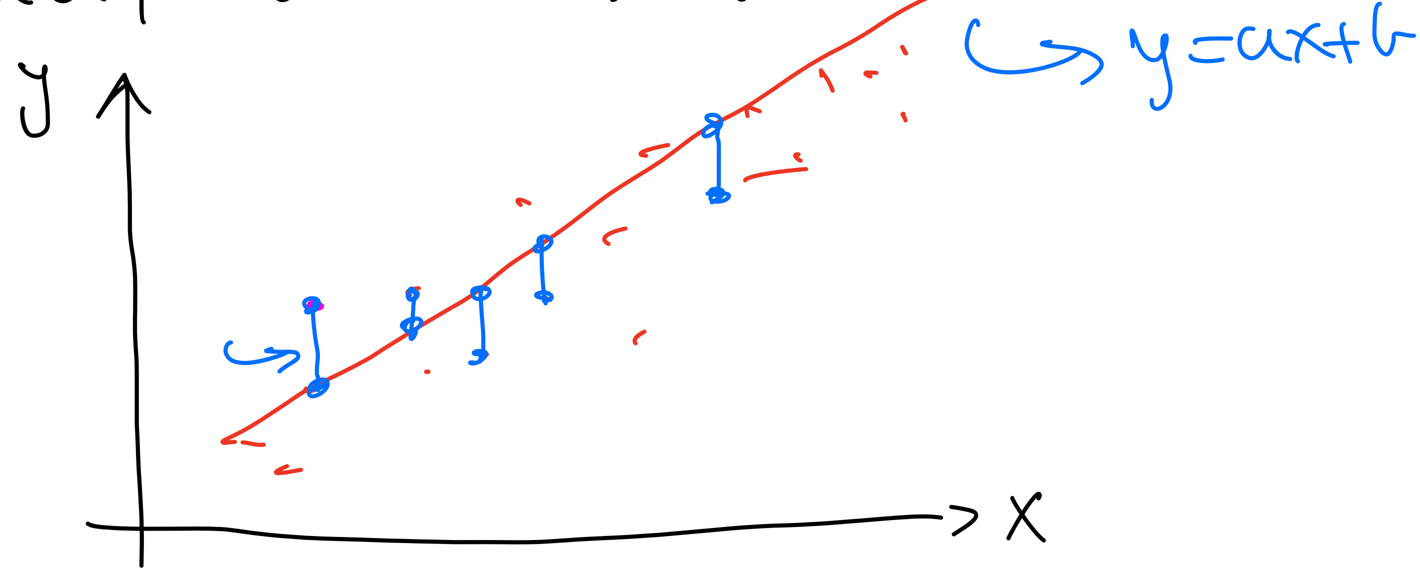
We'll discuss today:

- the definition of the *covariance* and *correlation coefficient* of two random variables
- how to compute the *covariance* and *correlation coefficient* of two random variables
- the 'least-squares' line of best fit

Following problem:

Do an experiment & measured two variables (x, y)

↳ Gives data points (x_j, y_j), $j = 1, \dots, N$.



↳ Plot indicates that there is a relationship b/w x & y. The simplest relationship is a linear one:

$$y = ax + b$$

for constants a, b. find some "good" a, b.

Ask for a minimal choice of (a,b).

$\hookrightarrow$ We ask for (a,b) such that they minimise the sum of the vertical distances.

If a pair (a_N, b_N) minimises the above distance the we say that the line

$$y = a_N x + b_N$$

is a line of best fit.

How to find (a_N, b_N)?

Minimise over $(a,b) \in \mathbb{R}^2$, the function:

$$d(a,b) = \sum_{i=1}^{N} (y_i - a x_i - b)^2$$

$\Leftrightarrow \nabla d(a,b) = (0,0).$

Solve for (a,b):

$$\begin{cases} a_N = \dfrac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)}{\boxed{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2}} \\ b_N = \frac{1}{N}\sum_{i=1}^{N} y_i - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) a_N \end{cases}$$

$\to$ by Cauchy-Schwarz inequality > 0

So sign of a_N purely depends on the sign of

$$\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)$$

If we think of x's generated by a r.v. X
y's ———————— Y

then as $N \to +\infty$, should converge to: (law of large numbers statement).

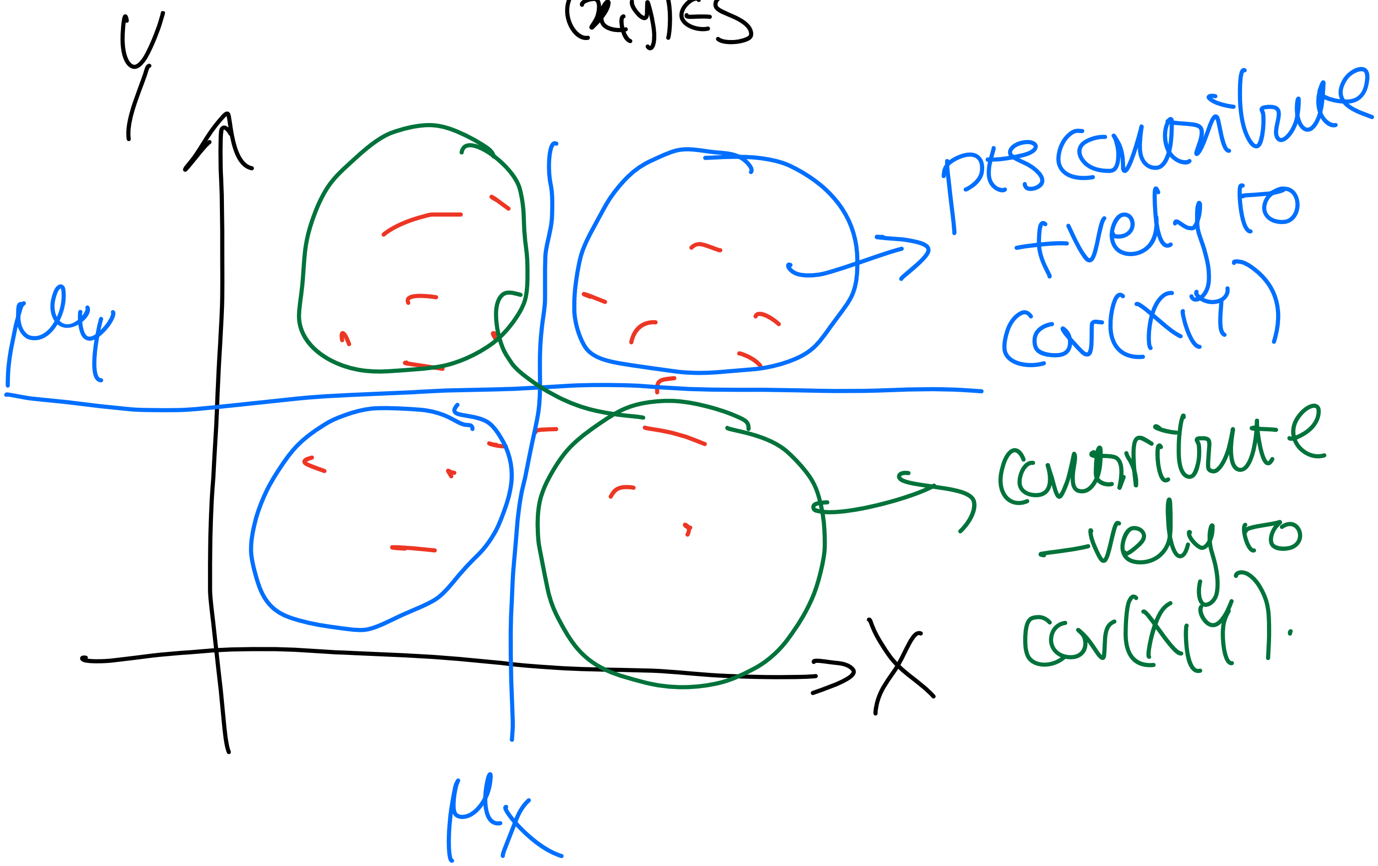
$$\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

$$\left|\sum_{n=1}^{N} a_n b_n\right| \leq \left(\sum_{n=1}^{N} a_n^2\right)^{1/2} \left(\sum_{n=1}^{N} b_n^2\right)^{1/2} \ldots$$

Definition 4.7: Let X, Y be a pair of discrete random variables taking values. We define the covariance of X, Y to be

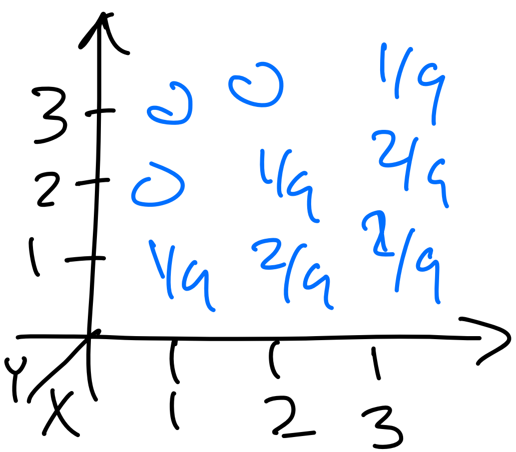
$$\mathrm{cov}(X,Y) = \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$= \sum_{(x,y)\in S} (x-\mu_X)(y-\mu_Y)\, p_{X,Y}(x,y)$$



Example 7:

- You choose two numbers at random from the set $\{1, 2, 3\}$
- Let X be the larger and Y be the smaller of these two numbers
- What is $\mathrm{cov}(X, Y)$?



Last time

$$\hookrightarrow E[XY] = 4$$

$$E[X] = 22/9$$

$$E[Y] = 14/9$$

$$\hookrightarrow \mathrm{cov}(X, Y) = 4 - \frac{22}{9} \times \frac{14}{9} = \frac{16}{81}.$$

Proposition 4.8: If X is a random variable, then

$$\text{cov}(X, X) = \text{var}(X).$$

Proof:

$$\text{cov}(X, X) = \mathbb{E}[(X-\mu_X)(X-\mu_X)]$$

$$= \mathbb{E}[(X-\mu_X)^2] = \text{var}(X).$$

Proposition 4.9: If X, Y are *independent*, then

$$\operatorname{cov}(X, Y) = 0$$

Proof:

If X, Y indep, $E[XY] = \mu_X \mu_Y$.

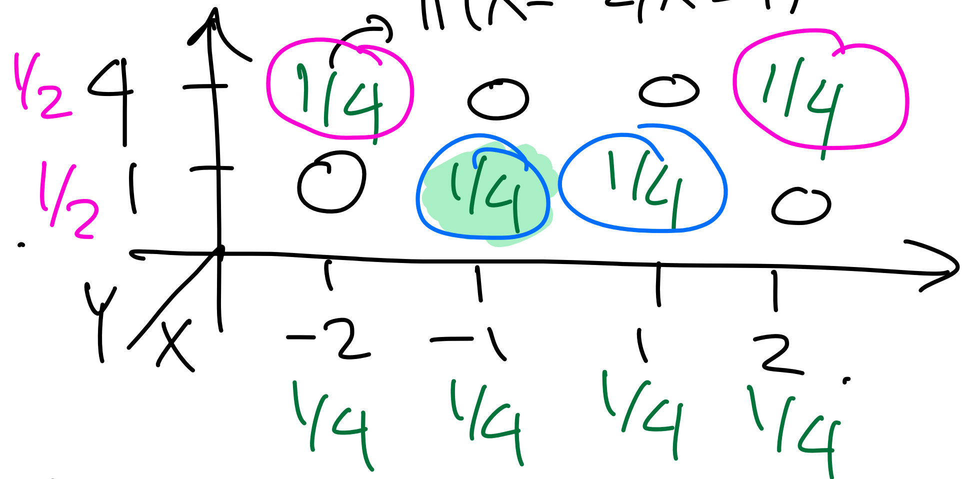
If indep $\Longrightarrow \operatorname{cov}(X, Y) = 0$

$\not\Longleftarrow$

Example 8:

- Let X be a discrete r.v. which is uniform on $\{-2, -1, 1, 2\}$ and set $Y = X^2$
- What is $\text{cov}(X, Y)$?
- Are X and Y independent?

$$\mathbb{P}(X=-2, X^2=4) = \mathbb{P}(X=-2)$$



$$p_{X,Y}(x,y) = \mathbb{P}(X=x, X^2=y).$$

Now

$$\mathbb{E}[XY] = \sum xy\, p_{X,Y}(x,y) = 0$$

$$\mathbb{E}[X] = 0 \Rightarrow \text{cov}(X,Y) = 0.$$

But X & Y are not independent!

$$p_{X,Y}(-1,1) = 1/4 \neq \frac{1}{2} \times \frac{1}{4} = p_X(-1)\,p_Y(1).$$

Proposition 4.10: Let X, Y be (discrete) random variables and $a, b \in \mathbb{R}$. Then,

$$\operatorname{cov}(aX, bY) = ab \operatorname{cov}(X, Y)$$

Proof:

$$\operatorname{cov}(aX, bY) = \mathbb{E}[abXY] - \mathbb{E}[aX]\mathbb{E}[bY]$$

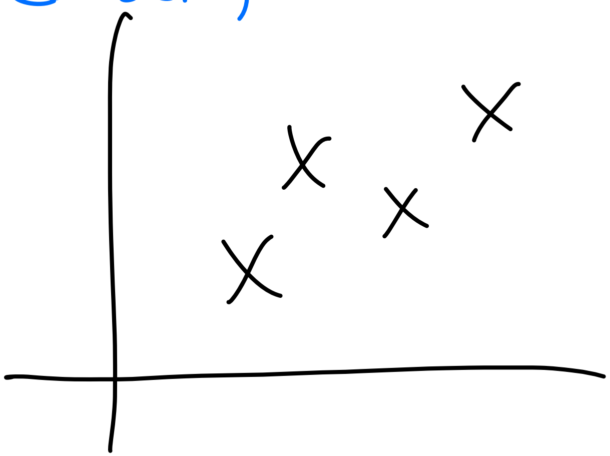
$$= ab \operatorname{cov}(X, Y).$$

Definition 4.11:

- Let X, Y be a pair of (discrete) random variables
- We define the correlation coefficient of X, Y to be

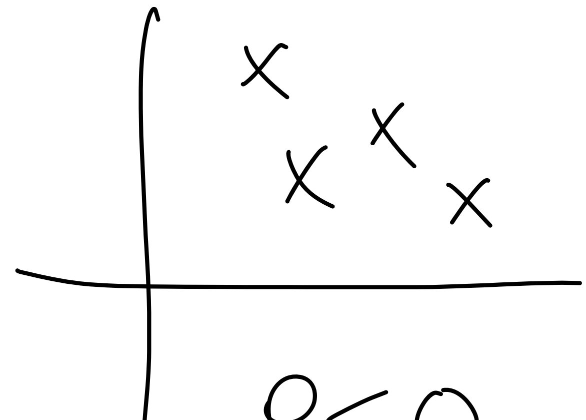
(Pearson)

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$



$\rho > 0$

+vely correlated



$\rho < 0$

−vely correlated.

$$\rho(aX, bY) = \frac{\text{cov}(aX, bY)}{\sqrt{\text{var}(aX)\,\text{var}(bY)}} = \frac{ab\,\text{cov}(X,Y)}{\sqrt{a^2b^2}\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{ab}{\sqrt{a^2b^2}}\rho(X,Y)$$

$$= \frac{ab}{|ab|}\rho(X,Y).$$

$$\text{var}(aX) = a^2\,\text{var}(X)$$

Proposition 4.12:
If X, Y are (discrete) random variables, then

[If X, Y are indep, then $\rho(X,Y)=0$]

$$-1 \leq \rho(X, Y) \leq 1$$

In particular, $|\rho(X, Y)| = 1$ **if and only if** then $Y = aX + b$ for some $a, b \in \mathbb{R}$.

Proof:

$$|\mathrm{cov}(X,Y)| = |\mathbb{E}[\underbrace{(X-\mu_X)}_{W}\underbrace{(Y-\mu_Y)}_{Z}]|$$

Cauchy-Schwarz
$$\leq \sqrt{\mathbb{E}[(X-\mu_X)^2]\,\mathbb{E}[(Y-\mu_Y)^2]}$$

$$= \sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}$$

$$|\rho(X,Y)| = \frac{|\mathrm{cov}(X,Y)|}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} \leq 1.$$