5.2 Expected Values, Covariance, and Correlation

Let X and Y be jointly distributed random variables with pmf $p(x, y)$. Let $h : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be any function. Then the expected value of $h(X, Y)$, denoted by $E[h(X, Y)]$ is given by $E[h(x,y)] = \sum_{x,y} h(x,y)\, p(x,y)$

For example, $E(XY) = \sum_{x,y} xy\, p(x,y)$

Lec22

5.2.1 Covariance

When two random variables X and Y are not independent, it is frequently of interest to assess if the two random variables are linearly dependent and if so, how strongly they are related to one another.

Definition 27. The **covariance** is a measure of the strength between two random variables X and Y of linear association, which is defined as

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

Example 72. The joint and marginal pmf's for X = automobile policy deductible amount and Y = homeowner policy deductible amount are

	$p(x, y)$	y = 500	1000	5000
	100	0.3	0.05	0
x	500	0.15	0.20	0.05
	1000	0.10	0.10	0.05

Note:

$\text{cov}(X,X) = E[X^2] - E[X]^2 = \text{var}(X)$

x	100	500	1000
$p_X(x)$	0.35	0.4	0.25

y	500	1000	5000
$p_Y(y)$	0.55	0.35	0.10

Find the Cov(X, Y).

Solution.

$E[XY] = 100 \times 500 \times 0.3 + 100 \times 1000 \times 0.05 + 100 \times 5000 \times 0 + 500 \times 500 \times 0.15 + \cdots$

$= 682500$

$E[X] = 485$ $E[Y] = 1125$

$\text{cov}(X,Y) = 682,500 - 485 \times 1125 = 136,875$

5.2.2 Correlation

Note: The unit of the covariance are squared and thus it is difficult to interpret.

The correlation coefficient, ρ, divides the covariance by its maximum value to give a measure of linear strength between the values of -1 and 1.

> The correlation coefficient of X and Y, denoted by $\text{corr}(X,Y)$, $\rho_{X,Y}$, or just ρ, is defined by
>
> $$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$ ← always b/w -1 & 1 !

Example 73. (Example 72 continued) It is easily verified that

$$Var(X) = \underbrace{E[X^2]}_{353{,}500} - \underbrace{E[X]^2}_{485} = 118{,}275$$

$$Var(Y) = 1{,}721{,}875$$

$$\rho(X,Y) = \frac{136{,}875}{\sqrt{118{,}275}\sqrt{1{,}721{,}875}} = 0.303$$

Properties of covariance and correlation

1. Covariance and correlation both measure the strength of the linear association between X and Y.

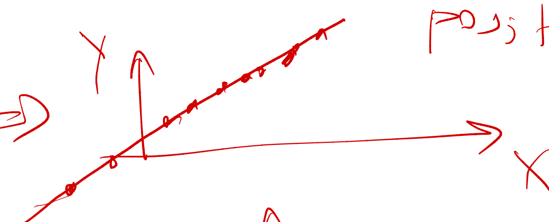
2. If the covariance and correlation are both > 0, then X and Y have a positive linear association i.e., as X increases, Y increases as well.

3. If the covariance and correlation are both < 0, then X and Y have a negative linear association, i.e., as X increases, Y decreases.

positive slope

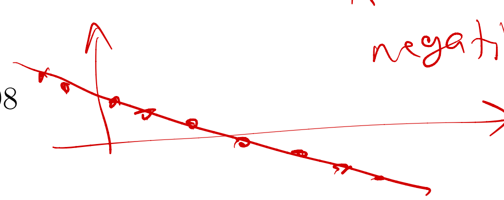
Figure:

$\rho(X,Y) = 1 \Rightarrow$



negative

$\rho(X,Y) = -1 \Rightarrow$



$\rho(X,Y)$ close to 1 :

4. If the covariance and correlation are ≈ 0, then there is ______________________ between X and Y.

5. For any two random variables X and Y, $-1 \leq \rho \leq 1$. ✓

6. If a and c are either both positive or both negative, and any b & d :

corr(X,Y) ⇒ $\rho(aX+b, cY+d) = \rho(X,Y)$

(if $ac<0$ then $\rho(aX+b, cY+d) = -\rho(X,Y)$)

7. If X and Y are independent or uncorrelated, then $\rho = 0$, but $\rho = 0$ does not imply independence. It just means that there is no **linear** association between X and Y. But it can also mean that X and Y may have a non-linear association. → →

8. $\rho = 1$ or -1 if and only if $Y = aX + b$ for some numbers a and b with $a \neq 0$.

⚠ correlation $\not\Rightarrow$ causation

5.3 Statistics and Their Distributions

Definition 28. The random variables $X_1, X_2, \cdots, X_n$ are said to form a (simple) **random sample** of size n if

1.

2.

Consider taking a random sample from a population, and compute the sample mean, $\overline{y}$, for the observations.

- Because the sample is ______________, the observations will also be ______________.

  Hence, $\overline{y}$ will also be ______________.

- Because $\overline{y}$ is random, it has a ______________________ associated with it. This distribution plays an important role in drawing conclusion about the population, This is what we called ______________________

X & Y indep. b/c of independence

$\rho(X,Y) = ?$

$$E[XY] = \sum_{x,y} xy \overbrace{p(x,y)}^{p_X(x)\,p_Y(y)}$$

$$= \sum_x \sum_y xy\, p_X(x)\, p_Y(y)$$

$$= \sum_x \Big(x\, p_X(x) \underbrace{\sum_y y\, p_Y(y)}_{E[Y]} \Big)$$

$$= E[Y] \Big(\underbrace{\sum_x x\, p_X(x)}_{E[X]} \Big)$$

$$E[XY] = E[X]E[Y]$$

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

$$\rho(X,Y) = 0$$