Lesione 7

Recap:

Definition: covariance, measures association between X & Y

$$Cov(X,Y) = E[(X-EX)\cdot(Y-EY)] = E[X\cdot Y] - EX\cdot EY$$

positive >>0 if X forces Y to increase, negative <<0, if X increasing causes Y to decrease.

The linear correlation coefficient:

$$\rho_{x,y} = cov\left(\frac{X-EX}{\sqrt{Var X}}, \frac{Y-EY}{\sqrt{Var Y}}\right) = \frac{cov(X,Y)}{\sqrt{Var X \cdot Var Y}} \Rightarrow -1 \le \rho_{x,y} \le 1$$

covariance of standarized forms of X and Y

$\rho = 1$ or $\rho = -1$ if there is a linear relationship

$X+Y$

$E[X+Y] = EX + EY$

$Var[X+Y] = Var(X) + Var(Y) + 2\,Cov(X,Y)$

Proprieties of Cov

1.

2. $Cov(X,X) = Var(X)$ and $cov(X,Y) = cov(Y,X)$ covariance is symmetric

3. $Cov(aX+b, Y) = a\,cov(X,Y)$, $\forall a, b \in \mathbb{R}$

4) $Cov(X+Y, Z) = cov(X,Z) + cov(Y,Z)$

5) Combination of 3 and 4

if $cov(X,Y) = 0$, X and Y are said to be uncorrelated or not correlated

Independent Random Variables

Let $X_1, X_2, \dots, X_n$ be n random variables.

They are independent if jointly

marginal distribution.

$$P(X_1\in E_1, X_2\in E_2, \dots, X_n\in E_n) = P(X_1\in E_1)\cdot P(X_2\in E_2)\cdot \dots \cdot P(X_n\in E_n)$$

$\forall E_1, E_2, \dots, E_n \in F$

X and Y are independent $\Longleftrightarrow$

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B) \quad \forall A, B \in \mathbb{R}$$

Propretics:

1) X, Y are independent $\Longleftrightarrow$

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y) \quad \forall x,y \quad \text{for } X, Y \text{ discrete r.v}$$

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \quad \forall x,y \quad \text{for } X, Y \text{ are abs. cont. r.v.}$$

i.e If the joint distribution function can be factorized as the product of the marginal densities.

2) If X, Y are independent, then $E(X \cdot Y) = E(X) \cdot E(Y)$, so $\text{cov}(X,Y) = 0$

(In general) The opposite does NOT hold:
if $\text{cov}(X,Y) = 0 \not\Rightarrow X, Y$ are necessarily independent.

If X, Y are independent $\Rightarrow$

$$\text{Var}(X,Y) = \text{Var}(X) + \text{Var}(Y) + 2\cancelto{0}{\text{Cov}(X,Y)} = \text{Var}(X) + \text{Var}(Y)$$

Exercises:

We have a couple of discrete random variable with joint distribution

$$\begin{array}{c|c|c|c}
X \backslash Y & 0 & 1 & p_X \\ \hline
-1 & 0 & 1/4 & 1/4 \\ \hline
0 & 1/2 & 0 & 1/2 \\ \hline
1 & 0 & 1/4 & 1/4 \\ \hline
p_Y & 1/2 & 1/2 & 1 \checkmark
\end{array}$$

↳ Sum of distribution

a) EX, EY

$Y \sim Be(p = 1/2) \rightarrow$ i.e. 1 with $P(\ ) = 1/2$ and 0 with $P(\ ) = \frac{1}{2}$

$\Rightarrow EY = 1/2$

$EX = 0$ since it's simmetric with respect to 0.

b) $Cov(X,Y) = E[X \cdot Y] - \overset{0}{\cancel{EX}} \cdot EY$

$= \cancel{-1 \cdot 0 \cdot 0} + -1 \cdot 1 \cdot 1/4 + \cancel{0 \cdot 0 \cdot 1/2} + \cancel{0 \cdot 1 \cdot 0} + \cancel{1 \cdot 0 \cdot 0} + 1 \cdot 1 \cdot \frac{1}{4}$

$= 0 \Rightarrow X$ and Y are uncorrelated.

c) Are X and Y independent?

They are not independent because there are values of x and y such that $p_{X,Y}(x,y) \neq p_X(x) \cdot p_Y(y)$

$0 = p(X=-1, Y=0) \neq \frac{1}{2} \cdot \frac{1}{4} = P(X=-1) \cdot P(Y=0)$

d) $P(Y = X^2) = P(X=-1, Y=1) + P(X=0, Y=0) + P(X=1, Y=1)$

we just do through possible values of X, find the associated Y and find the product of the probabilities

$\stackrel{!}{=} \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1 \Rightarrow Y = X^2$ almost surely.

$\Rightarrow$ X and Y are not independent.

Exercise 2

Table of joint distribution of couple of discrete random variables.

X \ Y	2,5	4	5,5	p_X
1	0,1	0,2	0,1	.4
2	0,15	0,1	0,05	.3
3	0,05	0,1	0,15	.3
p_Y	.3	.4	.3	1

a) $cov(X,Y)$

b) Are X and Y independent?

$E(Y) = 4$

$E(X) = 1.9$

$$E[X \cdot Y] = 1 \cdot 2{,}5 \cdot 0{,}1 + 1 \cdot 4 \cdot 0{,}2 + \ldots + 3 \cdot 5{,}5 \cdot 0{,}15 = \sum_{(x,y) \in S} x \cdot y \cdot p_{X,Y}(x,y)$$

$= 7{,}75$

$\Rightarrow Cov(X,Y) = E[X \cdot Y] - EX \cdot EY = 7{,}75 - 4 \cdot 1{,}9 = 0{,}15 \neq 0$

$Cov(X,Y) \underset{\neq 0}{\Rightarrow}$ they are not independent, since the $cov(X,Y) \neq 0$, if they were independent $\Rightarrow cov(X,Y) = 0$

sum of

Distributions of Independent Random variables

Linear combinations of random variables

Theorem

1. Let $X_1, X_2, \ldots, X_n$ be indeput random variable with $X_i \sim N(\mu_i, \sigma_i^2)$

Then: linear reprecibility (property of gaussian)

Let $Y = c_1 X_1 + c_2 X_2 + \ldots + c_n X_n \sim N(\mu_Y, \sigma_Y^2)$,

where: $\mu_Y = c_1\mu_1 + \ldots + c_n \mu_n$

The linear combination gaussians is also gaussian

$\sigma_Y^2 = c_1^2 \sigma_1^2 + \ldots + c_n \sigma_n^2$

Assuming all $c_i = 1$, $\sum_i c_i$, the reproducibily is lost unless

2. $X_1, X_2, \ldots$ independent, $X_i \sim Poi(\lambda_i)$, then

$Y = X_1 + X_2 + \ldots + X_n \sim Poi(\lambda_1 + \ldots + \lambda_n)$

Reproducibility of Poisson

similarly 3. $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Be(p)$, then $Y = X_1 + X_2 + \ldots + X_n \sim Bin(n,p)$

iid = independent and identically distributed

$$\begin{cases} 1, & p = p \\ 0, & p = 1-p \end{cases}$$

since Bin is the number of successes over n Bernoulli trials.

Example

$X_1, X_2, \ldots, X_n$ independent r.v $X_i \sim N(2,3)\ \forall i$

$\Rightarrow X_1, X_2, \ldots, X_{10} \overset{iid}{\sim} N(2,3)$

a) Compute probability that at least one of the range variables is negative

It's easier to use the complement which is:
"All the r.v.s are ≥ 0" $= A^c$

$$P(A) = 1 - P(A^c) = 1 - P(X_1 \geq 0, X_2 \geq 0, \dots, X_{10} \geq 0)$$

since independent

$$= 1 - P(X_1 \geq 0) \cdot P(X_2 \geq 0) \cdot P(X_3 \geq 0) \cdot \dots \cdot P_{10}(X_{10} \geq 0)$$

since iid $\longrightarrow$

$$\overset{!}{=} 1 - \left[P(X_1 \geq 10)\right]^{10}$$

$$P(X_i \geq 0) = 1 - P(X_1 \leq 0) = 1 - P\left(\frac{X_1 - EX_1}{\sqrt{\operatorname{Var} X_1}} \leq \frac{0-2}{\sqrt{3}}\right) = 1 - \phi\left(\frac{-2}{\sqrt{3}}\right) \overset{!}{=} \phi\left(\frac{2}{\sqrt{3}}\right) \overset{!}{\approx} \phi(1{,}15) = 0{,}87493$$

$X_i \sim N(2,3)$

$$= 1 - \left[0{,}87493\right]^{10} = 0{,}73713$$

b) Computer the probability that exactly 2 r.v among $\{X_1, \dots, X_{10})$ are > 1

We can introduce 10 new r.v such that:

$$Y_i = \begin{cases} 1 \text{ if } X_i > 1 \\ 0 \quad \text{otherwise} \end{cases}$$

Since X_i independent, Y_i will also be.

$$Y_i \overset{iid}{\sim} Be(p)$$

$$p = P(X_i > 1) = 1 - \phi\left(\frac{1-2}{\sqrt{3}}\right) = \phi\left(\frac{1}{\sqrt{3}}\right) \simeq \phi(0,58) = 0,71804$$

$$Y = Y_1 + Y_2 + \dots + Y_{10} \sim Bin(10, p)$$

$$P(Y=2) = \binom{10}{2} p^2 (1-p)^8 \simeq 0,00090$$

Using sequence of iid Bernoulli variables we have solved the problem quickly.

Example 2:

X_i = n° of phone calls received by a call center in the hour i

$i = 1,2,3,4$

$$X_i \overset{iid}{\sim} Poi(5), \ i = 1,2,3,4$$

a) n° of phone calls in the first 4 hours overall.

$$= X_1 + X_2 + X_3 + X_4 \sim Poi(4 \times 5) = Poi(20)$$

b) Probability number of calls in first 4 hours is larger than 20.

$$P(X_1 + X_2 + X_3 + X_4 > 20) = 1 - P(X_1 + X_2 + X_3 + X_4 \leq 20)$$

$$= 1 - \sum_{j=0}^{20} \frac{e^{-\lambda} \lambda^j}{j!} = 0,44091$$

Definition

A random sample is a vector (X_1, X_2, X_n) of independent and identically distributed random variable, i.e.

$$X_1, X_2, \dots, X_n \overset{iid}{\sim} f \text{ or } X_i \overset{iid}{\sim} f, \ i = 1,2,\dots,n$$

Here f represents the density of the common distribution

Alternative notation $X_i \overset{iid}{\sim} F$, $i = 1, 2, n$, where F is the cumm. distribution function.

This is a mathematical notation to formalize that the random variables represent the variable of interest n subjects

We consider the sum of iid finite sequence or r.v.s

$$S_n = X_1 + X_2 + \dots + X_n$$

Let $\mu := \mathbb{E}(X_1) = \mathbb{E}(X_i)$ and $\sigma^2 : Var(X_1) = Var(X_i)$

Therefore:

$$\mathbb{E}[S_n] = \mathbb{E}[X_1 + X_2 \dots + X_n] = \mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]$$

Definition Sample mean

Let $(X_1, X_2, \dots, X_n)$ be a random sample of size n

$$\bar{x}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{S_n}{n}$$

Let us compute the mean and variance of S_n

$$\mathbb{E}[\bar{X}_n] = \frac{\mathbb{E}[S_n]}{n} = \frac{n\mu}{\mu} = \mu = \mathbb{E}(X_1)$$

$Var[$

Sample mean interpretation.

Do we in general know the distribution of S_n or $\bar{X}_n$

Only in the cases that we saw before, where where all variables are iid.

We can only describe the distribution of S_n or $\bar{X}_n$ only for large n.

↳ This will be the central limit theorem.