Dept of AIML, RVCE Bangalore

Statistics for Data Science

Module 2: Joint Distribution, Correlation, Covariance and Independent Random Variables

1. Suppose that n people throw their hats in a box and then each picks one hat at random. Each hat can be picked by only one person and each assignment of hats to persons is equally likely. What is the expected value of X, the number of people that get back their own hat? *This does not fall under the topics I have listed above. I just thought I will include this anyways!*

Let X_i be the random variable that takes a value 1 if the i-th person selects his/her own hat & takes 0 otherwise.

Given: $P(X_i = 1) = 1/n$ & $P(X_i = 0) = 1 - 1/n$

$\therefore E[X_i] = 1 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n}$

Now we have $X = X_1 + X_2 + \dots + X_n$

$\Rightarrow E[X] = E[X_1] + E[X_2] + \dots + E[X_n] = n \cdot \frac{1}{n} = 1$ //.

2. Suppose that some customers at a restaurant take a simple survey. They are asked to rate the quality of their experience on a scale of 1 to 5, and they are asked whether or not there waiter/waitress was attentive. Let X be a random variable describing the quality of the experience, and Y a random variable which is 1 if the waiter/waitress was attentive and 0 otherwise

$p(x,y)$	x = 1	x = 2	x = 3	x = 4	x = 5	
y = 0	0.05	0.2	0.1	0.04	0.01	$= p_Y(0)$
y = 1	0.01	0.09	0.15	0.2	0.15	$= p_Y(1)$.
	$p_X(1)$	$p_X(2)$	. . .		$p_X(5)$.	

These are probability values

(a) Is the table a valid PMF?

For the table to be a valid PMF, the sum of probability values of the joint PMF = 1.

Here we see that the probability value sums to 1.

(b) Find the marginal PMF of X and Y.

$p_Y(0) = 0.4$ $\quad p_Y(1) = 0.6$. $\therefore$ Marginal PMF of Y is given by $p_Y(y) = \begin{cases} 0.4 & y = 0 \\ 0.6 & y = 1. \end{cases}$

Marginal PMF(X):

$x =$	1	2	3	4	5
$P(X = x) =$	0.06	0.29	0.25	0.24	0.16

(c) Are X and Y independent?

For X & Y to be independent, $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$ for all values of X & Y. However, we see that this does not happen. Hence X & Y are not independent.

3. Suppose X and Y are two random variables with means $\mu_X = 1$ and $\mu_Y = -3$ and variances $\sigma_X^2 = 2$ and $\sigma_Y^2 = 4$ respectively. Suppose the covariance between X and Y is known to be

$Cov(X, Y) = 1/2$, find the mean and variance of $Z = 3X + Y$.

Mean $(Z) = E[Z] = E[3X+Y] = 3E[X] + E[Y] = 3-3=0$.

$Var(Z) = Var[3X+Y]$

We know that (?) $Var[aX+bY] = a^2 Var[X] + b^2 Var[Y] + 2ab\, cov(X,Y)$

$\Rightarrow Var[Z] = Var[3X+Y] = 9 Var[X] + Var[Y] + 6 cov(X,Y)$

$= 9(2) + 4 + 6(1/2) = 25$ //.

4. Suppose X and Y are two random variables with $E(X) = E(Y) = 0$, and $VAR(X) = 1$, and suppose we know X is independent from $X + Y$. What is the covariance between X and Y?

Given X is indep of $X+Y$.

$\Rightarrow cov(X, X+Y) = 0$; To find $cov(X,Y)$.

$cov(X, X+Y) = cov(X,X) + cov(X,Y)$.

$0 = Var[X] + cov(X,Y)$

$\Rightarrow cov(X,Y) = -1$ //

5. Identify the following statement as YES or NO with proper justification.
Statement: If X and Y are independent random variables, then $VAR(3X + 2Y + 1) =$

$VAR(3X - 2Y + 3)$.

TRUE. $Var[aX \pm bY] = a^2 Var[X] + b^2 Var[Y]$

and $Var[aX + b] = a^2 Var[X]$.

Adding a constant does not affect the variance.

6. Let X, Y, and Z be random variables, where X and Y are uncorrelated. The means of the RVs are $E(X) = 1$, $E(Y) = 2$, and $E(Z) = -1$, and $E(XZ) = 5$. What is $COV(X, Y + 2Z)$?

$Cov(X, Y+2Z) = Cov(X,Y) + Cov(X,2Z)$

$= \cancel{Cov(X,Y)} + 2\,Cov(X,Z)$.

0 (Uncorrelated given)

$\Rightarrow Cov(X,Z) = E[XZ] - E[Z]E[X]$

$= 5 - (-1)(1) = \underline{6}$.

7. List out the conditions that a matrix has to satisfy for it to be deemed as a valid covariance matrix.

(a) Square matrix
(b) Real and Symmetric.
(c) Positive Semidefinite
(d) Eigenvalues are all real & non-negative

8. Which of the following matrices can be a valid covariance matrix and why? Indicate clearly which property of covariance matrix do the other matrices violate?

YES NO NO No

A. $M_1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ B. $M_2 = \begin{pmatrix} 1 & 1 \\ 3 & 2 \end{pmatrix}$ C. $M_3 = \begin{pmatrix} 5 & 0 \\ 0 & -2 \end{pmatrix}$ D. Matrix M_4 whose determinant is 15 and the eigenvalues are -3 and -5.

Fix the reasons based on Q.7.

9. You toss a fair coin 100 times. (*No no... not ordering you! Assume you do this when u really feel totally jobless!!!*. Let X denote the number of heads in the 100 tosses and Y be the number of tails in the 100 tosses. Compute the correlation coefficient $\rho_{X,Y}$. With the given information can you calculate $Var(X+Y)$? If yes, find the value. If not, what more information you need to compute $Var(X+Y)$?.

If X is the no. of heads, $Y = 100 - X$

$\therefore Var[X+Y] = Var[X + 100 - X] = Var[100] = 0$

Pl. do think before you start working out the solution.

10. Let X, Y and Z be random variables. The covariance matrix of (X, Y, Z) is given below:

$$C_{X,Y,Z} = \begin{pmatrix} Var(X) & Cov(X,Y) & Cov(X,Z) \\ Cov(X,Y) & Var(Y) & Cov(Y,Z) \\ Cov(X,Z) & Cov(Y,Z) & Var(Z) \end{pmatrix} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 4 & -1 \\ 1 & -1 & 4 \end{pmatrix}$$

Given the above information alone, is it possible to compute $Var(X+Y+Z)$? If yes, obtain the value. If your answer is a NO, then list out the information that you need to compute $Var(X+Y+Z)$.

YES. Expand $Var[X+Y+Z]$ and you get the answer.

We know $Var[X+Y] = Var[X] + Var[Y] + 2cov(X,Y)$

Similarly compute $Var[X+Y+Z] = 10$.

11. Let X, Y and Z be random variables. The covariance matrix of (X, Y, Z) is given below:

$$C_{X,Y,Z} = \begin{pmatrix} Var(X) & Cov(X,Y) & Cov(X,Z) \\ Cov(X,Y) & Var(Y) & Cov(Y,Z) \\ Cov(X,Z) & Cov(Y,Z) & Var(Z) \end{pmatrix} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 4 & -1 \\ 1 & -1 & 4 \end{pmatrix}$$

(a) Find $Cov(X, Y + Z)$.

$Cov(X, Y+Z) = Cov(X,Y) + Cov(X,Z)$
$= 0 + (-1) = -1.$

(b) Find the covariance matrix corresponding to $(X, X + Z, Y + Z)$

$Cov(X, X+Z, Y+Z).$

We know $Cov(U, V, W) = \begin{bmatrix} Var\, U & Cov(U,V) & Cov(U,W) \\ Cov(U,V) & Var\, V & Cov(W,V) \\ Cov(U,W) & Cov(V,W) & Var\, W \end{bmatrix}$

$Cov(X, X+Z, Y+Z)$
$= \begin{bmatrix} Var[X] & Cov(X, X+Z) & Cov(X, Y+Z) \\ Cov(X, X+Z) & Var(X+Z) & Cov(Y+Z, X+Z) \\ Cov(X, Y+Z) & Cov(Y+Z, X+Z) & Var(Y+Z) \end{bmatrix}$

Compute the values & fill the matrix.

12. Two fair six sided dice, a black one and a red one, are rolled, with outcomes X and Y respectively

for the black and red die respectively. Are $X + Y$ and $X - Y$ independent? Justify your choice.

X+Y & X−Y are not indep.

Assume I tell you, the sum is 12. What is the difference. Justify on those lines.

13. A chicken lays a Poisson(λ) number N of eggs. Each egg, independently, hatches a chick with probability p. Let X be the number which hatch, so $X|N \sim Bin(N, p)$. Find the correlation between N (the number of eggs) and X (the number of eggs which hatch). Your final answer should work out to a simple function of p.

NOT COVERED IN CLACS

Conditional PMF.

14. A device runs until either of two components fails, at which point the device stops running. The joint density function of the lifetimes of the two components, both measured in hours, is

$$f_{X,Y}(x, y) = \begin{cases} \frac{x + y}{8} & 0 < x < 2, 0 < y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that the device fails during the first hour of the operation?

15. Let the joint PDF of (X, Y) be

$$f_{X,Y}(x,y) = \begin{cases} \frac{x+y}{3} & \text{if} 0 < x < 2, 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Calculate the covariance $Cov(X, Y)$.

Direct Substitution in formula.

Find $f_X(x) = \int_0^1 \frac{x+y}{3} \, dy$

$f_Y(y) = \int_0^2 \frac{x+y}{3} \, dx$.

Obtain $E[X] = \int_0^1 x f_X(x) \, dx$

$E[Y] = \int_0^{2} y f_Y(y) \, dy$

$E[XY] = \int_0^2 \int_0^1 xy \left(\frac{x+y}{3}\right) dy \, dx$.

$cov(X, Y) = E[XY] - E[X]E[Y]$.