Random Variable Measures

Mean, Variance and Covariance

X = Number of weekly work-outs of a person,

Y = Number of strokes

What is the probability that if a person works-out for 5 days in a week his likeliness of having a stroke comes down to zero?

P(X = 5, Y = 0) = ?

Instructors

Mousum Dutta
Chief Data Scientist, Spotle.ai
IIT Kharagpur

Guest Instructor

Rimjhim Ray
Head of Product, Spotle.ai
MBA, SP Jain

Expected Value

The Expected Value E(X) of a random variable X, denoting the outcome of a certain experiment, is the average value of the outcomes over a large number of repetitions.

Consider the example:

A coin is flipped. You get Rs 10 when the toss results in a head, give Rs 10 if the outcome is a tail. Your earning can look like:

{10, -10, 10,10,10,-10,-10,10,-10,-10…N}

The expected income obtained by averaging the income over a sufficiently large N, is Zero.

Expected Value – Discrete Random Variable

The Expected Value E(X) of a discrete random variable is the probability-weighted average of X, or it is the average of values that X takes multiplied by the associated probability.

In the coin example in the previous slide:

$$E(X) = (-10*1/2 + 10*1/2)/\ 1$$

$$= 0$$

Note since sums of probabilities is always 1, the denominator is always 1.

Therefore, the formula for the expected value of a discrete random variable is given by:

$$E(X) = \sum x_i P(X = x_i) = \sum x_i p_i$$

Exercise – Expected Value

Little Massy is a fantastic footballer. In any match, the probability of him scoring one goal is 0.1, 2 goals is 0.1, 3 goals is 0.3, 4 goals is 0.3, 5 goals is 0.2.

How many goals are Messy expected to score in a series of 10 matches?

Exercise – Expected Value

To solve the problem, plot the number of goals and probability:

Number of goals (X)	Probability P(X)
1	0.1
2	0.1
3	0.3
4	0.3
5	0.2

$$E(X) = 1 \times 0.1 + 2 \times 0.1 + 3 \times 0.3 + 4 \times 0.3 + 5 \times 0.2 = 3.4$$

Expected goals in a series of 10 matches

$$= 10 \times E(X)$$

$$= 34$$

Expected Value – Continuous Random Variable

For a continuous random variable, X, with probability density function, f(x), the Expected Value is given by:

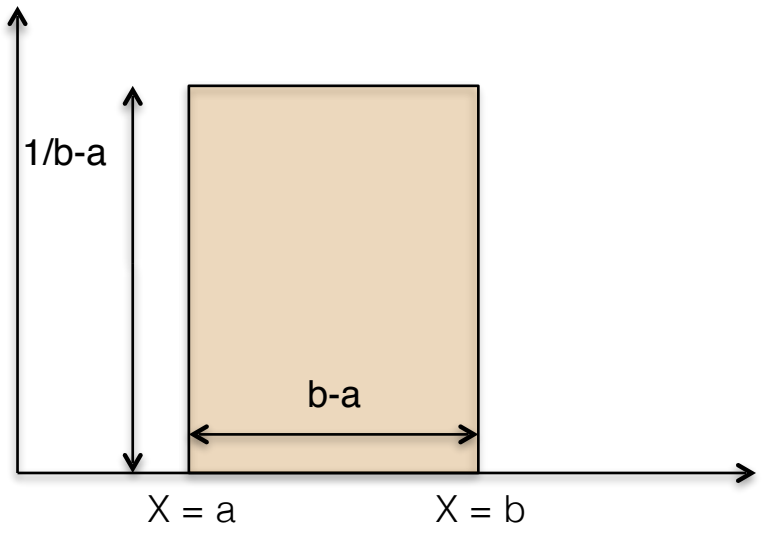
$$E(X) = \int_{x_{min}}^{x_{max}} xf(x)dx$$

Consider a random variable X, with $f(x) = 3x^2$ where X can take values between 0 and 1:

$$E(X) = \int_0^1 x.3x^2dx = 3/4$$

Expected Value – Uniform Random Distribution

For a continuous random variable, **X**, that varies between a and b with equal probability, $P(X = x_i) = 1/(b-a)$ for $a <= x_i <= b$.



$E(X) = \int_a^b (1/(b-a)).xdx$

Recall $\int xdx = x^2/2$

Therefore integrating $\int(1/(b-a)).xdx$, we get $(1/(b-a)).((b^2/2) – (a^2/2))$

$E(X) = (a+b)/2$

Variance

The Variance of a random variable X, indicates how dispersed the random variable is around the mean. Lower variance indicates X converges around the mean while a higher value of variance shows X is more dispersed.

Standard Deviation of the random variable is given by the square root of the variance.

Formula:

$$\text{Var}(X) = \sum(x_i - E(X))^2 p_i$$

$$= \sum x_i^2 p_i - 2E(X)(\sum x_i p_i) + E(X)^2 \sum p_i$$

$$= E(X^2) - (E(X))^2$$

Variance – Illustrative Problem

Consider the coin flip example: A coin is flipped. You get Rs 10 when the toss results in a head, give Rs 10 if the outcome is a tail. Your earning can look like:

{10, -10, 10, 10, 10, -10, -10, 10, -10, -10…N}

Here E(X) = 0, as calculated earlier

Variance:

$$\text{Var}(X) = \sum(x_i - E(X))^2 p_i$$

$$= (10 - 0)^2.1/2 + (-10 - 0)^2.1/2$$

$$= 100$$

Standard Deviation = 10

Joint Probability Distribution

Consider 2 variables **X and Y**:

X denoting the number of weekly work-outs of a person, Y denoting the number of strokes he is likely to suffer over his life-time. We will be interested in finding out the probability of events involving both variables, for e.g. P(X=0, Y=2) or P(X=5, Y=0).

In this case, we construct the **joint distribution of X and Y** to compute the probability of events involving both variables to show, for example, the efficacy of exercise in reducing heart diseases.

Joint Probability Distribution – Discrete Random Variable

The joint probability mass function of 2 discrete random variables X and Y, is given by p(xi,yj) giving probability of $X=x_i$ and $Y=y_j$ occurring simultaneously.

Let X and Y be outcomes of 2 distinct dice rolls. Then the joint probability distribution of X and Y is given by the following table, where cell(i,j) = $p(x_i,y_j)$:

Y \ X	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Important Properties: $0 <= p(x_i,y_j) <=1$ and $\sum \sum p(x_i,y_j) = 1$

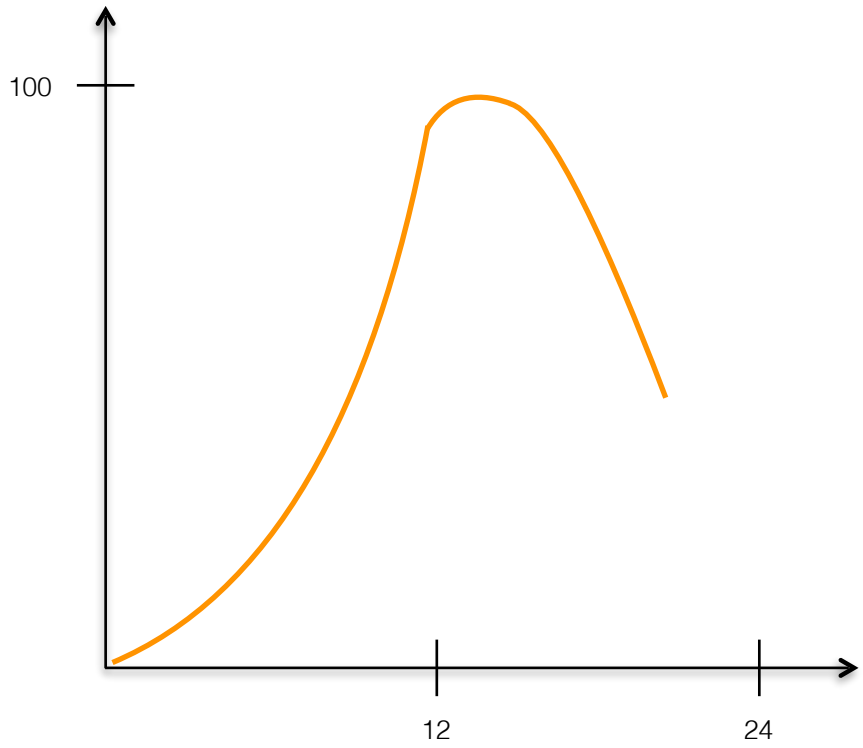
Joint Probability Distribution – Continuous Random Variable

X and Y are 2 continuous random variables, X denotes the time spent daily to study for an exam and Y denotes the marks obtained in the exam.

The probability that X and Y falls in the intervals dx and dy simultaneously, is given by f(x,y) dxdy where f(x,y) is the joint probability density function of X and Y.

Given that X belongs to [0,24] and Y belongs to [0,100] the following properties hold:

$f(x,y) >= 0$ and $\int_0^{100}\int_0^{24} f(x,y)dxdy = 1$



Joint Cumulative Distributive Function

The joint CDF of 2 jointly distributed random variables belonging to [a,b][c,] simply gives the probability of X <= xi and Y<= yj at any point (xi,yj) in the given sample space. In general:

$$F(x,y) = P(X<=x_i, Y<=y_j)$$

Case I: X and Y are discrete random variables with pdf = $p(x_i,y_j)$. The joint CDF is: given by:

$$F(x_i, y_j) = \sum_{x<=x_i,\ y<=y_j} \sum p(x,y)$$

Case II: X and Y are continuous random variables with pdf = f(x,y)

$$F(xi, yj) = \int^{y_j}\int^{x_i} f(x,y)dxdy = 1$$

Marginal Probability Distribution

Given 2 jointly distributed random variables X any Y, we may just want to find the probability distribution function of one variable say X. The pmf (or pdf) of X without Y or Y without X is called the marginal PMF or PDF.

Let X and Y be outcomes of 2 distinct dice rolls. The marginal probability of P(X=3) and P(Y=5) are illustrated in the table below:

Y \ X	1	2	3	4	5	6	
1	1/36	1/36	1/36	1/36	1/36	1/36	
2	1/36	1/36	1/36	1/36	1/36	1/36	
3	1/36	1/36	1/36	1/36	1/36	1/36	
4	1/36	1/36	1/36	1/36	1/36	1/36	
5	1/36	1/36	1/36	1/36	1/36	1/36	P(Y=5) =1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	
		P(X=3)=1/6					

Covariance

The covariance of 2 random variables X and Y measures the joint variability of X and Y. It is a measure of how changes in one variable cause changes in a second variable.

A positive covariance means the variables move together in same direction. For example the random variables H and W, denoting the Height and Weight of a person.

A negative covariance means the variables move together in same direction. For example the random variables T and S, denoting the time spent online and S denoting the attention span of a person.

Covariance

Covariance measures the total variation of two random variables from their expected values.

$\text{Cov}(X,Y) = \sum(x_i - E(X))(y_j - E(Y))/n = E(XY) - E(X)E(Y)$

Covariance shows whether 2 variables are likely to move in the same direction or in opposite direction. It does not show the strength of the relationship. The strength of the relationship is given by the correlation. Correlation indicates a scaled measure of covariance.

Correlation = COV(X,Y)/ STDEV(X)*STDEV(Y)

Exercise – Covariance and Correlation

Let X denote the average number of children per household in a neighbourhood. Let Y denote child mortality rate (per 100) in the neighbourhood. The values for X and Y for 10 neighbourhoods are given below:

	1	2	3	4	5	6	7	8	9	10
X	2	1	3	1	4	5	6	1	2	3
Y	9	9	15	14	20	21	25	7	5	12

How are X and Y related – positive covariance, negative covariance, independent?
Calculate COV(X,Y) and the correlation coefficient.