Chapter: Covariance and Correlation

Content Developers: Vaishali Kapoor & Rakhi Arora

College / University: Rajdhani College (University of Delhi)

Table of Contents

Learning outcomes

After you have read this chapter, you should be able to:-

1. Define Covariance.
2. Calculate the covariance for the discrete and Continuous Random Variables.
3. Consider the special cases of covariance.
4. Compute correlation.

Introduction

So far we have studied Joint Probability mass / Density function and expected value of some function of random variables. It is also of interested, at times, to know whether two random variables have some sort of relationship or not. For example, someone may be interested in knowing whether marks obtained by a student is positively affected by number of hours denoted to studying by that student or is negatively affected by hours denoted to watching T.V. If X is marks obtained and Y is number of hours daily spent on studying, then one is interested in knowing whether X and Y are related. If Yes, positively or negatively (the answer we expect is positive) and if it do affect, how strongly are they related.

For answering the above question, we need to learn statistical techniques of covariance and correlation. This chapter aids is understanding these and solving through these techniques. First section of this chapter covers covariance for discrete and continuous random variables and various theorems and its proofs and corollaries are covered. Second section of this chapter focuses on measuring strength of relationship between X and Y called correlation.

1. Covariance

When X and Y are two random variables and are not independent then covariance between two random variables X and Y is

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

where μ_X is mean of variable X and

μ_Y is mean of variable Y.

$(X - \mu_X)$ is deviation of X variable from its mean value

$(Y-\mu_Y)$ is deviation of Y variable from its mean value.

So covariance is expected value of deviations of X and Y from its respective mean values.

If suppose, X and Y are positively related to each other, then this means that when X attain large value then corresponding Y value also tend to be larger and small values of X correspond to small values of Y. Then most of the probability mass or density will be associated with $(X - \mu_X)$ and $(Y - \mu_Y)$, either both positive or both negative so the product tends to be positive. Thus for strong positive relationship $Cov.(X,Y)$ should be positive. For if there exists strong negative relationship, signs of $(X - \mu_X)$ and $(Y - \mu_Y)$ would be opposite, yielding a negative $Cov.(X,Y)$. If they are not related at all, then positive product values would tend to be cancelled out with negative product values, yielding $Cov.(X,Y)$ near zero.

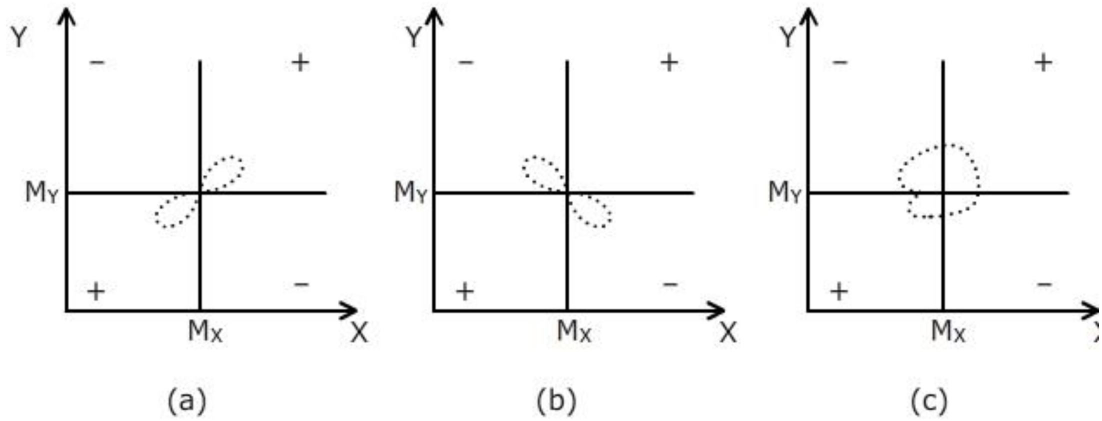


Figure 1: Different Possible relationships between random variables X and Y.

In the above figure, '+' and '–' signs are marked ± show areas where if X and Y values are plotted shows 'positive' or 'negative' values respectively. In panel (a) X and Y have positive relationship, in panel (b) X and Y have negative relationship and in panel (c) X and Y have no relationship so covariance would be positive in first case (a) negative in case (b) and around zero in case (c).

b) Covariance for Discrete Random Variables

If two discrete random variables X and Y are not independent then covariance between X and Y is given by

$$Cov.(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

$$= \sum_{all\,x}\sum_{all\,y}(x-\mu_X)(y-\mu_Y)\rho_{XY}(x,y)$$

$$= \left[\sum_{all\,x}\sum_{all\,y} xy\,\rho_{XY}(x,y)\right] - \mu_X\mu_Y$$

Example 1 : The joint PMF for X = automobile policy deductible and Y = homeowner policy deductible amount is given in following table:

$X\downarrow$ \ $Y\rightarrow$	**0**	**100**	***200***	$P_X(x)$
100	0.20	0.10	0.20	0.5
250	0.05	0.15	0.30	0.5
$P_Y(y)$	0.25	0.25	0.5	1

$$\mu_X = E(X) = \sum xP_X(x)$$

$$= 0.5\times100 + 0.5\times250$$

$$= 175$$

$$\mu_Y = E(Y) = \sum yP_Y(y)$$

$$= 0.25\times0 + 0.25\times100 + 0.5\times200$$

$$= 125$$

$$Cov(X,Y) = \sum_{all\,x}\sum_{all\,y}(x-175)(y-125)P_{XY}(x,y)$$

$$= (100-175)(0-125)(0.20)+(250-175)(0-125)(0.05)$$
$$+(100-175)(100-125)(0.10)+(250-175)(100-125)(0.15)$$
$$+(100-175)(200-125)(0.20)+(250-175)(200-125)(0.30)$$
$$= 1875$$

Positive value of covariance suggests positive relationship between automobile policy deductible amount and homeowner policy automobile amount.

(b) An alternative formula (a shortcut)

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= E[XY - \mu_X X - Y\mu_Y + \mu_X \mu_Y]$$
$$= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X \mu_Y$$
$$= E(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y$$
$$= E(XY) - \mu_X \mu_Y$$

c) Covariance for continuous random variables

For continuous random variables X and Y that are not independent, covariance between X and Y is given by

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y)\, f_{XY}(x,y)dxdy$$
$$= \left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{XY}(x,y)\right] - \mu_X \mu_Y$$
$$= E(XY) - \mu_X \mu_Y$$

Example 2: If joint PDF for X and Y is given by

$$f_{XY}(x,y) = \begin{cases} 24xy & 0 \le x \le 1,\ 0 \le y \le 1,\ x+y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and marginal PDF is $f_X(x) = \begin{cases} 12x(1-x)^2 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$

with $f_Y(g)$ obtained by replacing X by Y in $f_X(x)$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{XY}(x,y)dxdy$$

$$= \int_0^1\int_0^{1-x} xy\, 24xy\, dydx$$

$$= \int_0^1\left[\int_0^{1-x} 24x^2y^2\, dy\right]dx$$

$$= \int_0^1 \frac{24}{3}x^2\, y^3\Big|_0^{1-x} dx$$

$$= \int_0^1 8x^2(1-x)^3 dx$$

$$= 8\int_0^1 x^2 - x^5 + 3x^4 - 3x^3$$

$$= 8\left[\frac{x^3}{3} - \frac{x^6}{6} + \frac{3x^5}{5} - \frac{3x^4}{4}\Big|_0^1\right]$$

$$= 8 \times \frac{20x^3 - 10x^6 + 36 - 45}{60}\Bigg|_0^1 = \frac{2}{15}$$

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x\, f_X(x)dx$$

$$= \int_0^1 x.12x(1-x)^2 dx$$

$$= \int_0^1 (12x^2 + 12x^4 - 24x^3)dx$$

$$= 12\frac{x^3}{3} + 12\frac{x^5}{5} - 24\frac{x^4}{4}\Bigg|_0^1$$

$$= 4 + \frac{12}{5} - 6$$

$$= -2 + \frac{12}{5}$$

$$= \frac{2}{5}$$

Also, $\mu_Y = \frac{2}{5}$ since marginal PDF is same.

$$Cov(X, Y) = E(X, Y) - \mu_X \mu_Y$$

$$= \frac{2}{15} - \frac{2}{5} \times \frac{2}{5}$$

$$= \frac{10 - 12}{75} = -\frac{2}{75}$$

This minus signs shows negative relationship between *X* and *Y* (since $x + y \leq 1$, more of *X* would mean less of *Y*).

d) Special Cases

(i) Till now we assumed that *X* and *Y* are not independent what if they are ? Then we expect that there should not be any relationship or covariance between *X* and *Y* be zero.

$$Cov(X,Y) = E(X,Y) - E(X)E(Y)$$

$$= E(X)E(Y) - E(X)E(Y) \quad [\because X \text{ and } Y \text{ are independent}]$$

$$= 0$$

So as covariance value is observed if *X* and *Y* are independent.

But above result must be used with a caution as reverse of it is not true. Consider the sample space = {(–2, 4), (–1, 1), (0, 0), (1, 1), (2, 4)}, where each point is equally likely. Random variable *X* is first component of sample and Y be the second

$$P(X=1, Y=1)=\frac{1}{5}$$

$$P(X=1)=\frac{1}{5}$$

$$P(Y=1)=\frac{2}{5}$$

So $\quad P(X=1, Y=1) \neq P(X=1)$ and $P(Y=1)$

So X and Y are dependent.

$$E(XY)=\frac{1}{5}[(-8)+(-1)+0+1+8]=0$$

$$E(X)=\frac{1}{5}[(-2)+(-1)+0+1+2]=0$$

$$E(Y)=\frac{1}{5}[4+1+10+1+4]=2$$

$$Cov(X,Y)=0-2\times 0=0$$

Therefore if $Cov(X,Y)$ is zero it can't be concluded that X and Y are independent (it may or may not be)

(ii) $\quad Cov(X,X)$

What if one wish to calculate covariance between X and X?

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

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

$$=Var(X)$$

Covariance (X, X) is the same thing as Variance of X which gives quantitative measure of how much spread true is in the distribution or population of x values

c) Suppose X and Y are random variables and a and b are constants then

$$Var(aX+bY)=a^2Var(X)+b^2Var(Y)+2abCov(X,Y)$$

Proof to it is as follows:

$$Var(aX+bY)=E[(aX+bY)^2]-(a\mu_X+b\mu_Y)^2$$

$$=E[a^2X^2+b^2Y^2+2abXY]-[a^2\mu_X^2+b^2\mu_Y^2+2ab\mu_X\mu_Y]$$

$$=a^2E(X^2)+b^2E(Y^2)+2abE(XY)-a^2\mu_X^2+b^2\mu_Y^2$$

$$=2ab\mu_X\mu_Y$$

$$=a^2[E(X^2)-\mu_X^2]+b^2[E(Y^2)-\mu_Y^2]+2ab[E(XY)-\mu_X\mu_Y]$$

$$=a^2V(x)+b^2V(Y)+2ab\,\mathrm{cov}(X,Y)$$

If X and Y are independent then $\mathrm{cov}(X,Y)=0$.

Suppose there are n variables $X_1,X_2,...,X_n$, then

$$\mathrm{var}\left(\sum_{i=1}^{n} a_iX_i\right)=\sum_{i=1}^{n} a_i^2\,\mathrm{var}(X_i)a_j+2\sum_{i<j} a_ja_j\,\mathrm{cov}(X_i,\ X_j)$$

If $X_1,...,X_n$ are independent random variables and all $a_i's$ are equal to 1 then

$$\mathrm{var}(X_1+...+X_n)=V(X_1)+V(X_2)+...+V(X_n)$$

e) Scaling of Variables

In the two examples solved in this chapter, covariance in first example was 1875 and is second it was $-\frac{2}{75}$

Could we conclude that X and Y variables in first example have strong (positive) relationship and a weaker (negative) relationship emerges in second example? The Answer to it is No!

To see this, lets scale our random variables by a and lets check covariance for two new variables aX and aY and compare it with covariance of X and Y.

$$\text{cov}(aX,\, aY) = E(aXaY) - E(aX)E(aY)$$

$$= E(a^2 XY) - aE(X)aE(Y)$$

$$= a^2[E(XY) - E(X)E(Y)]$$

$$= a^2 \,\text{cov}(X,\, Y)$$

If $\text{cov}(X, Y)$ is 1800 and $a = 10$ then $\text{cov}(aX, aY)$ is 1,80,000. $Cov(X, Y)$ is positive and so is $\text{cov}(aX, aY)$; but $\text{cov}(aX, aY)$ > $\text{cov}(X, Y)$ for | a | > 1. Thus it can be concluded that covariance accurately tells the direction of relationship between X and Y - positive or negative but does not tell the strength of relationship as it is affected by scaling of variables.

2. Correlation

As discussed in the last section that covariance suffers from the defect that scaling of variable alters the value of covariance and then it does not serve as a measure of strength of relationship. So a better measure is studied called correlation coefficient.

The covariance of X and Y necessarily reflects the units of both random variables, which can make it difficult to interpret. The measure of strength of relationship should be dimensionless measure of dependency so that one *xy* relationship can be compared to another. Dividing $\text{cov}(X, Y)$ by $\sigma_X\, \sigma_Y$ accomplishes this task. Also, this scales the quotient to be a number between -1 and 1.

For two random variables, X and Y, correlation coefficient of X and Y is given by

$$\rho(X,Y)=\frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}=\text{cov}(X^*,Y^*)$$

where $X^*=(X-\mu_X)/\sigma_X$

$$Y^*=(Y-\mu_Y)/\sigma_Y$$

since $\text{cov}(X^*,Y^*)=E\left[\left(\frac{X-\mu_X}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]-E\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)$

$$=\frac{1}{\sigma_X\sigma_Y}E[(X-\mu_X)(Y-\mu_Y)]-\frac{1}{\sigma_X\sigma_Y}E(X-\mu_X)E(Y-\mu_Y)$$

$$=\frac{1}{\sigma_X\sigma_Y}\text{cov}(X,Y)-0$$

Since $E(X-\mu_X)=E(X)-\frac{n\mu_X}{n}=\mu_X-\mu_X=0$

$$E(Y-\mu_Y)=E(X)-\frac{n\mu_Y}{n}=\mu_Y-\mu_Y=0$$

X^* and Y^* are called standardised variables. So correlation coefficient is covariance between these standardized variables.

Some Useful Results for $\rho(x,y)$

a) $|\rho(X,Y)|\le 1$

$$Var(X^*\pm Y^*)\ge 0$$

$$Var(X^*\pm Y^*)=Var(X^*)\pm 2\,\text{cov}(X^*,Y^*)+Var(Y^*)$$

$$=\text{var}\left(\frac{X-\mu_X}{\sigma_X}\right)\pm 2\,\text{cov}(X^*,Y^*)+\text{var}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)$$

$$=\frac{1}{\sigma_X^2}[\text{var}(X)-0]\pm 2\,\text{cov}(X^*,Y^*)+\frac{1}{\sigma_Y^2}[\text{var}(Y)-0]$$

[Since variance of constant is zero]

$$= \frac{\text{var}(X)}{\text{var}(X)} \pm 2\rho(X,Y) + \frac{\text{var}(Y)}{\text{var}(Y)}$$

$$= 1 \pm 2\rho(X,Y) + 1$$

$$\text{var}(X^* \pm Y^*) = 2[1 \pm \rho(X,Y)] \geq 0$$

$\Rightarrow \quad 1 \pm \rho(X,Y) \geq 0$

$\Rightarrow \quad |\rho(X,Y)| \leq 1$

b) $|\rho(X,Y)| = 1$ iff $Y = aX + b$ for some constants a and b.

Suppose $\rho(X,Y) = 1$ then $\text{var}(X^* - Y^*) = 0$. A random variable with zero variance is constant. So it readily follows that Y is a linear function of X i.e. $X^* - Y^* = \alpha$ (say)

$X^* = Y^* + \alpha$ (linear relationship)

Since it is iff only first part of statement is proved. For second part,

Let $\quad Y = aX + b$

then $\quad E(Y) = aE(X) + b$ and $V(Y) = a^2V(Y)$

$$\sigma_Y = a\sigma_X$$

$$\rho(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \, \sigma_Y}$$

Putting $Y = aX + b$, $E(Y) = aE(X)$ and $\sigma_Y = a\sigma_X$

$$\rho(X,Y) = \frac{E(X(aX+b)) - E(X)a(E(X)+b)}{\sigma_X a.\sigma_X}$$

$$= \frac{E(aX^2 + bX) - a(E(X))^2 - bE(X)}{a\sigma_X^2}$$

$$= \frac{aE(X^2) - a(E(X))^2 + bE(X) - bE(X)}{a\sigma_X^2}$$

$$\rho(X,Y) = \frac{a(\sigma_X^2)}{a.\sigma_X^2} = 1$$

Hence second part of statement is also proved. Similar results can be obtained for $\rho(X,Y) = -1$.

c) For some constants *a* and *c* either both negative or both positive,

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

$$\rho(aX+b,\ cY+d) = \frac{\text{cov}(aX+b,\ cY+d)}{\sqrt{V(aX+b)V(cY+d)}}$$

$$= \frac{a.c\,\text{cov}(X,Y)}{\sqrt{a^2V(X)c^2V(Y)}}$$

$$= \frac{a.c\,\text{cov}(X,Y)}{|\,a.c\,|\,\sigma_X\sigma_Y}$$

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

This amounts to saying that scaling variables up or down in the same direction does not affect the correlation coefficient.

d) Correlation coefficient measures only existence of linear relationship between random variables *X* and *Y*.

ρ = 0 implies that $\text{cov}(X,Y)$ would be zero.

A zero value of covariance as discussed earlier, need not imply that *X* and *Y* are independent and so is for ρ; $\rho = 0$ does not imply that there is no relationship between *X* and *Y*.

For if *X* and *Y* are linearly related as $Y = aX + b$ then $|\rho(X,Y)| = 1$. $\rho(X,Y)$ shows strongest possible linear relationship between *X* and *Y* $|\rho| < 1$ indicates that relationship is not completely linear, but there could be a very strong non linear relationship. So, $\rho = 0$ implies that *X* and Y are (linearly) uncorrelated but there could be high dependence between X and Y given by some non-linear relationship.

Example 3: Let X and Y be discrete random variables with joint pmf

$$P(x,y)=\left\{\frac{1}{4} \qquad \text{for } (x,y)=(-4,1),(4,-1)\ (2,2),\ (-2,\ -2)\right\}$$

$$\mu_X = \frac{1}{4}.(-4)+\frac{1}{4}.(4)+\frac{1}{4}(2)+\frac{1}{4}(-2)$$

$$= 0$$

$$\mu_Y = \frac{1}{4}(1)+\frac{1}{4}(-1)+\frac{1}{4}(2)+\frac{1}{4}(-2)$$

$$= 0$$

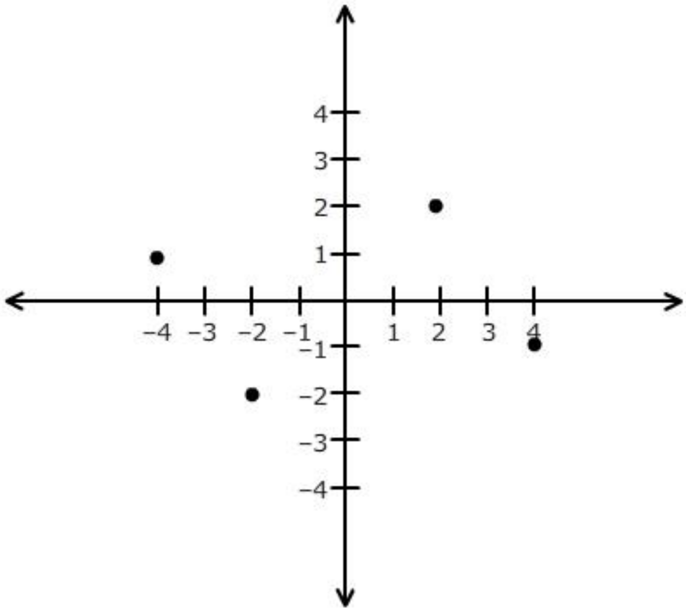
$$E(XY)=\frac{1}{4}(-4)+\frac{1}{4}(-4)+\frac{1}{4}(4)+\frac{1}{4}(4)$$

$$= 0$$

$$\text{cov}(X,Y)=E(XY)-\mu_X\mu_Y=0$$

so $\rho_{XY}=0$

Plotting all pairs of X and Y on the graph shows that two variables are dependent but $\rho(X,Y)=0$ represents absence of any linear relationship as evident in following graph



Example 4: Continuing from Example 1, we know that

$$\text{cov}(X,Y) = 1875,\ E(X) = 175\ \ \&\ \ E(Y) = 125$$

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

$$= [0.5\times(100)^2 + 0.5\times(250)^2] - (175)^2$$

$$= 36250 - (175)^2$$

$$= 5625$$

$$\Rightarrow \qquad \sigma_X = \sqrt{5625} = 75$$

$$\sigma_Y^2 = E(Y^2) - (E(Y))^2$$

$$= [0 + 0.25(100)^2 + 0.5(200)^2](125)^2$$

$$= 6875$$

$$\sigma_Y = \sqrt{6875} = 82.92$$

$$\rho(X,Y) = \frac{1875}{75\times 82.92} = 0.301$$

Example 5:

Risk on Securities

In the last chapter, we calculated expected returns of the two securities but we did not mention anything of risk. Risk is measured by standard deviation. Standard deviation is an estimate of the likely divergence of an actual return from expected return. So standard deviation is useful measure of risk as it weighs the deviation with possible probability of that outcome.

Let's revisit the old example of securities A & B. Following are the returns in different states.

State	1	2	3	4	5	***Total***
Returns on Security A, R_A	10%	12	8	14	19	
Returns on Security B,R_B	20%	25	33	27	22	
Probability	.10	.25	.35	.20	.10	1
R_A-E(R_A)	-1.5	0.5	-3.5	3.5	8.5	
R_B-E(R_B)	-7.4	-2.4	5.6	-0.4	-5.4	
(R_A-E(R_A)*(R_A-E(R_B))	11.1	-1.2	-19.6	-1.4	-45.9	

Where, E (R_A) = 11.5 % & E(R_B)=27.4(as calculated earlier)

V(R_A) :Variance of returns on Security A = $\Sigma P_i(R_A-E(R_A))^2$

= $(.10X(-1.5)^2) + (.25X(0.5)^2)+(.35X(-3.5)^2)+(.20X(3.5)^2)+(.10X(8.5)^2)$

= 0.225+0.0625+4.2875+2.45+7.225=14.25

Standard deviation of Returns on security A=√14.25=**3.775%**

Similarly,

V (R_B): Variance of Return on Security B$=\Sigma P_i(R_B-E(R_B))^2$

$= (.10 \times (-7.4)^2) + .35 \times (2.4)^2 + .35 \times (5.6)^2 + .20 \times (-0.4)^2 + .10 \times (-5.4)^2$

$= 5.476+1.44+10.976+0.032+2.916=20.84$

Standard deviation of returns on security B = $\sqrt{20.84}$=**4.56%**

Though the expected returns on security B are higher but also risk is higher on security B measured by standard deviation.

Portfolio

Let us study the same portfolio of A & B with same weights of 0.25& 0.75 respectively. ***V(Rp): Variance of Portfolio***$=V(0.25R_A+0.25R_B)$

$(0.25)^2 V(R_A) + (0.75)^2 V(R_B) + 2(0.25)(0.75)\,Cov(R_A R_B)$

Risk on portfolio is least compared to standard deviation of either $Cov(R_A R_B)$ $= =\Sigma P_i[(R_B-E(R_B))*(R_A-E(R_A))]$

$= (0.1 \times 11.1)+(0.25 \times (-1.2))+.(35 \times (-19.6))+(.20 \times (-1.4))+(.10 \times (-5.4))$

$= 0.11+ (-0.3) + (-6.86)+(-0.28)+(-0.54)$

$= -7.87$

$V(R_P) = [(0.25)^2 \times 14.25]+(0.75)^2 \times 20.84]+[2 \times (0.75) \times (0.25) \times (-7.87)]$

$=0.8906+11.7225+ (-2.95125)$

$= 9.66$

Standard deviation of Returns on Portfolio = 3.10% of security A or B . It is due to the reason that the two securities are negatively related shown by negative value of covariance). Hence the portfolio is diversified.

Correlation between returns of A & B is given by =

$$\frac{Cov(RA,RB)}{\sigma A \sigma B} = \frac{-7.87}{3.775*4.56} = -0.457$$

Exercises:

Q.1 Suppose that two dice are thrown. Let x be the number showing on the first die and let y be the larger of the two numbers showing. Find cov (X,Y).

Q.2 Show that Cov(ax+b,cy+d) = ac cov (x,y) for any constants a,b,c,& d.

Q.3 Let x & y be random variables with

fXY(x,y) ={1, -y<x<y, 0<y<1

0, elsewhere

Show that cov(x,y) = 0 but that x & y are dependent.

Q.4 Suppose that $f_{XY}(x,y)= a^2 e^{-a(x+y)}$, 0≤ x, 0 ≤y. Find Var(X+Y).

References:

1. Jay L. Devore, Probability and Statistics for Engineers, Cengage Learning, 2010.
2. William G. Cochran, Sampling Techniques, John Wiley, 2007.
3. Richard J. Larsen and Morris. L. Marx, An Introduction to Mathematical Statistics and its Applications, Prentice Hall, 2011.

.