Covariance and Correlation

Definition of covariance: Covariance of X and Y is

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

We can also denote $Cov(X,Y) = \sigma_{X,Y}$.

Two special cases: $Cov(X,X) = Var(X)$ and $Cov(X,c) = 0$.

Definition of correlation: Correlation of X and Y is

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

By the definition we see $\rho(X,X) = 1$, and $\rho(X,-X) = -1$.

We can show $|\rho| \leq 1$.

A list of properties of Covariance

(1). $Cov(X, Y) = E(XY) - EX \cdot EY$

(2). $Cov(X, Y) = Cov(Y, X)$

(3). $Cov(aX, bY) = abCov(X, Y)$

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

Remarks:

(a) We often use property (1) to compute $Cov(X, Y)$.

(b) Property (2) says the covariance operation is symmetric about X and Y.

(c) By (3) we know $Cov(X, -Y) = -Cov(X, Y)$.

(d) By (4) and (2) we know

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

Ex1

If X and Y are independent, then

$$Cov(X, Y) = 0$$

and

$$\rho_{X,Y} = 0.$$

Proof: $Cov(X, Y) = E(XY) - EX \cdot EY$. By independence, we know $E(XY) = EX \cdot EY$, so $Cov(X, Y) = 0$.

By definition of correlation,

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = 0.$$

Therefore,

if $Cov(X, Y) \neq 0$, then X and Y are not independent.

Ex2

Suppose X and Z are independent. $Var(X) = 1$ and $Var(Z) = 0.01$. $Y = X + Z$. Compute the correlation between X and Y.

Solution: Use the definition.

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

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

$$\begin{aligned} Var(Y) &= Var(X+Z) \\ &= Var(X) + Var(Z) \\ &\quad \textit{by independence assumption} \\ &= 1.01. \end{aligned}$$

So

$$\rho_{X,Y} = \frac{1}{\sqrt{1.01}} = 0.995.$$

A useful variance formula

$$\begin{aligned} & Var(aX+bY) \\ = \; & a^2Var(X)+2abCov(X,Y)+b^2Var(Y). \end{aligned}$$

Proof:

$$\begin{aligned} & Var(aX+bY) \\ = \; & Cov(aX+bY, aX+bY) \\ = \; & Cov(aX, aX+bY)+Cov(bY, aX+bY) \\ = \; & Cov(aX, aX)+Cov(aX, bY) \\ & +Cov(bY, aX)+Cov(bY, bY) \\ = \; & a^2Cov(X,X)+2abCov(X,Y)+b^2Cov(Y,Y) \\ = \; & a^2Var(X)+2abCov(X,Y)+b^2Var(Y). \end{aligned}$$

A more general formula If $X_1, X_2, \dots, X_n$ be n random variables, $a_1, a_2, \dots, a_n$ are constants. Then

$$\begin{aligned} & Var(a_1X_1+\cdots+a_nX_n) \\ = \; & \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum_{i<j}^{n} a_ia_jCov(X_i, X_j). \end{aligned}$$

The proof is left as an optional Hw problem.

Ex3 If $Var(X) = 1$, $Var(Y) = 2$ and $Cov(X,Y) = -1$. If $U = 3X-2Y$, $V = X+2Y$. Find $Var(U)$, $Var(V)$ and $Cov(U,V)$.

$$\begin{aligned}
Var(U) &= Var(3X-2Y) \\
&= 3^2 Var(X) + 2 \cdot 3 \cdot (-2) Cov(X,Y) \\
&\quad + (-2)^2 Var(Y) \\
&= 9 \cdot 1 + 2 \cdot 3 \cdot (-2) \cdot (-1) + 4 \cdot 2 = 29.
\end{aligned}$$

$$\begin{aligned}
Var(V) &= Var(X+2Y) \\
&= 1^2 Var(X) + 2 \cdot 1 \cdot 2 Cov(X,Y) \\
&\quad + (2)^2 Var(Y) \\
&= 1 - 4 + 8 = 5.
\end{aligned}$$

$$\begin{aligned}
Cov(U,V) &= Cov(3X-2Y, X+2Y) \\
&= Cov(3X-2Y, X) + Cov(3X-2Y, 2Y) \\
&= Cov(3X, X) + Cov(-2Y, X) \\
&\quad + Cov(3X, 2Y) + Cov(-2Y, 2Y) \\
&= 3Cov(X,X) + (-2)Cov(Y,X) \\
&\quad + 3 \cdot 2 Cov(X,Y) + (-2) \cdot 2 Cov(Y,Y) \\
&= 3Var(X) + 4Cov(X,Y) - 4Var(Y) \\
&= 3 - 4 - 4(2) = -9.
\end{aligned}$$

Proof of $|\rho| \leq 1$

We want to show for any X, Y, $|\rho(X,Y)| \leq 1$.

Let $Z_t = X - tY$ where t is a real number. Then we have

$$Var(Z_t) = Var(X) - 2tCov(X,Y) + t^2 Var(Y)$$

Let $g(t) = Var(Z_t)$ as a function of t. Consider $t = t_0 = \frac{Cov(X,Y)}{Var(Y)}$. We find

$$g(t_0) = Var(X) - \frac{[Cov(X,Y)]^2}{Var(Y)}$$

Since $g(t) = Var(Z_t) \geq 0$ for all t, we must have

$$Var(X) - \frac{[Cov(X,Y)]^2}{Var(Y)} \geq 0$$

which means

$$1 \geq \frac{[Cov(X,Y)]^2}{Var(X)Var(Y)}.$$

Thus by the definition of $\rho(X,Y)$ we see

$$1 \geq \rho^2(X,Y).$$

Markov Inequality

Let X be a positive random variable and $E[X] < \infty$. Then for every positive real number a, we have

$$\Pr(X > a) \leq \frac{E[X]}{a}.$$

Proof: We note that

$$Y = X - aI(X > a) \geq 0$$

Why? because if $X \leq a$ then $Y = X - 0 = X > 0$; and if $X \geq a$, then $Y = X - a \geq 0$. Since Y is a non-negative random variable, by the definition of expectation, its mean is greater or equal to zero, so $E[Y] \geq 0$.

$$E[Y] = E[X - aI(X > a)] = E[X] - aE[I(X > a)]$$

Thus we end up with

$$E[X] \geq aE[I(X > a)] = a \Pr(X > a).$$

We used the fact $E[I(X > a)] = \Pr(X > a)$.

Chebyshev Inequality

Let X be a random variable with mean μ and variance σ^2. Then for any $a > 0$ we have

$$\Pr(|X-\mu| > a) \leq \frac{Var(X)}{a^2}.$$

Proof: We use Markov inequality. Observe the following fact

$$\Pr(|X-\mu| > a) = \Pr((X-\mu)^2 > a^2)$$

By Markov inequality, we know

$$\Pr((X-\mu)^2 > a^2) \leq \frac{E[(X-\mu)^2]}{a^2}$$

where we consider $(X-\mu)^2$ is the X in Markov inequality. Note that

$$E[(X-\mu)^2] = Var(X)$$

Hence we end up with

$$\Pr(|X-\mu| > a) \leq \frac{Var(X)}{a^2}.$$

Let $a = k\sigma$ in Chebyshev inequality, we then have

$$\Pr(|X - \mu| > k\sigma) \leq \frac{Var(X)}{k^2\sigma^2} = \frac{1}{k^2}$$

suppose $k = 3$, then

$$\Pr(|X - \mu| > 3\sigma) \leq \frac{1}{9} = 0.11.$$

Or we can conclude that

$$\Pr(\mu - 3\sigma < X < \mu + 3\sigma) \geq 1 - \frac{1}{9} = 0.89.$$

With 89% chance the value of X will fall into the interval $(\mu - 3\sigma, \mu + 3\sigma)$.

An important application of Chebyshev inequality

Let $X_1, \ldots, X_n$ be iid random variables. We assume they have a common mean μ and variance σ^2. Let

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Then for $a > 0$,

$$\Pr(|\bar{X} - \mu| > a) \leq \frac{\sigma^2}{na^2}.$$

Why? Use Chebyshev inequality on $\bar{X}$. By the iid assumption,

$$Var(\bar{X}) = \frac{1}{n} Var(X_1) = \frac{1}{n}\sigma^2$$

Chebyshev inequality says

$$\Pr(|\bar{X} - \mu| > a) \leq \frac{Var(\bar{X})}{a^2} = \frac{\sigma^2}{na^2}.$$

The Law of Large Numbers

Because for any $a > 0$ (no matter how small it is),

$$\Pr(|\bar{X} - \mu| > a) \leq \frac{\sigma^2}{na^2},$$

we can conclude

$$\lim_{n \to \infty} \Pr(|\bar{X} - \mu| > a) = 0.$$

The above equation tells us the sample mean converges to the true mean in probability. This is called <u>the weak law of large numbers.</u>

There is also <u>the strong law of large numbers</u>. The mathmatical expression is

$$\Pr(\lim_{n \to \infty} \bar{X} = \mu) = 1.$$