Covariance and Correlation

We have seen how to *summarize a data-based relative frequency distribution* by measures of location and spread, such as the sample mean and sample variance. Likewise, we have seen how to summarize *probability distribution* of a random variable X by similar measures of location and spread, the mean and variance parameters. Now we ask,

> For a pair of random variables X and Y having *joint* probability distribution
> $$p(x, y) = P(X = x, Y = y)$$
> how might we summarize the distribution?

For *location* features of a joint distribution, we simply use the means μ_X and μ_Y of the corresponding *marginal* distributions for X and Y. Likewise, for *spread* features we use σ_X^2 and σ_Y^2. For *joint* distributions, however, we can go further and explore a further type of feature: the manner in which X and Y are *interrelated* or manifest *dependence*. For example, consider the joint distribution given by the following table for $p(x, y)$:

		Y	
		0	1
X	0	.7	.1
	1	.1	.1

We see that there is indeed some dependence here: if a pair (X, Y) is selected at random according to this distribution, the probability that $(X, Y) = (0, 0)$ is selected is .70, whereas the *product* of the probabilities that $X = 0$ and $Y = 0$ is $.8 \times .8 = .64 \neq .70$. So the events $X = 0$ and $Y = 0$ are *dependent events*. But we can go further, asking: How might we characterize the *extent* or *quantity* of this "dependence" feature?

There are in fact a variety of possible ways to formulate a suitable measure of dependence. We shall consider here one very useful approach: "covariance."

Covariance

One way that X and Y can exhibit dependence is to "vary together" – i.e., the distribution $p(x, y)$ might attach *relatively high probability* to pairs (x, y) for which the deviation of x above its mean, $x - \mu_X$, and the deviation of y above its mean, $y - \mu_Y$, are either *both positive* or *both negative* and relatively large in magnitude. Thus, for example, the information that a pair (x, y) had an x with positive deviation $x - \mu_X$ would suggest that, unless something unusual had occurred, the y of the given pair also had a positive deviation above its mean. A natural *numerical measure* which takes account of this type of information is the sum of terms

(1) $$\sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) .$$

For the kind of dependence just described, this sum would tend to be dominated by large *positive* terms.

Another way that X and Y could exhibit dependence is to "vary oppositely," in which case pairs (x, y) such that one of $x - \mu_X$ and $y - \mu_Y$ is positive and the other negative would receive relatively high probability. In this case the sum (1) would tend to be dominated by *negative* terms.

Consequently, the sum (1) tends to indicate something about the *kind of dependence* between X and Y. We call it the *covariance* of X and Y and use the following notation and representations:

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) = E\left[(X - \mu_X)(Y - \mu_Y)\right] .$$

It is easily checked that an equivalent formula for computing the covariance is:

$$\sigma_{XY} = E(XY) - E(X)E(Y).$$

For the example of $p(x, y)$ considered above, we find:

$$\mu_X = \mu_Y = .2, \ E(XY) = .1, \ \sigma_{XY} = .06 .$$

Does this indicate a strong relationship between X and Y? What kind of relationship? How can we tell whether a particular value for σ_{XY} is meaningfully large or not? We'll return to these questions below, but for now let us explore some other aspects of covariance.

Independence of X and Y implies Covariance $= 0$. This important fact is seen (for the discrete case) as follows. If X and Y are *independent*, then their joint distribution factors into the product of marginals. Using this, we have

$$\begin{aligned}
\sigma_{XY} &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) \\
&= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p_X(x)p_Y(y) \\
&= \left[\sum_x (x - \mu_X)p_X(x)\right]\left[\sum_y (y - \mu_Y)p_Y(y)\right] \\
&= [\mu_X - \mu_X][\mu_Y - \mu_Y] \\
&= 0 \cdot 0 = 0 .
\end{aligned}$$

Of course, we should want any reasonable measure of dependence to reduce to the value 0 in the case of an absence of dependence. □

For some measures of dependence which have been proposed in the literature, the *converse* holds as well: if the measure has value 0, then the variables are independent. However, for the covariance measure, this converse is *not* true.

EXAMPLE. Consider the joint probability distribution

		Y		
		-1	0	1
X	0	0	$1/3$	0
	1	$1/3$	0	$1/3$

Note that for this distribution we have a *very strong relationship* between X and Y: ignoring pairs (x, y) with 0 probability, we have

$$X = Y^2 .$$

On the other hand,

$$\sigma_{XY} = E(XY) - E(X)E(Y) = E(Y^3) - E(X) \cdot 0 = E(Y^3) = 0.$$

Thus the covariance measure *fails to detect the dependence structure.* □

Preliminary Conclusions about Covariance:

a) Independence of X and Y implies $\sigma_{XY} = 0$;
b) $\sigma_{XY} = 0$ does not necessarily indicate independence;
c) $\sigma_{XY} \neq 0$ indicates some kind of dependence is present (what kind?).

In order to be able to interpret the meaning of a nonzero covariance value, we ask

How big can σ_{XY} be in value?

It turns out that the covariance always lies between two limits which may be expressed in terms of the variances of X and Y:

$$-\sigma_X \sigma_Y \leq \sigma_{XY} \leq +\sigma_X \sigma_Y .$$

(This follows from the *Cauchy-Schwarz inequality*: for *any* two r.v.'s W and Z, $|E(WZ)| \leq [E(W^2)]^{1/2}[E(Z^2)]^{1/2}$, with equality if and only if W and Z are proportional.)

Moreover, the covariance can attain one of these limits *only* in the case that $X - \mu_X$ and $Y - \mu_Y$ are *proportional*: i.e., for some constant c,

$$X - \mu_X = c(Y - \mu_Y) ,$$

i.e.,

$$Y = \frac{1}{c} X + (\mu_Y - \frac{\mu_X}{c}) ,$$

i.e., X and Y satisfy a *linear relationship*, $Y = aX + b$ for some choice of a and b.

Interpretation of covariance. The above considerations lead to an interpretation of covariance: σ_{XY} measures the *degree of linear relationship* between X and Y. (This is why $\sigma_{XY} = 0$ for the example in which $X = Y^2$ with X symmetric about 0. This is a *purely quadratic* relationship, quite *nonlinear*.) □

Thus we see that covariance measures a particular kind of dependence, *the degree of linear relationship*. We can assess the strength of a covariance measure by comparing its magnitude with the largest possible value, $\sigma_X\sigma_Y$. If σ_{XY} attains this magnitude, then the variables X and Y have a *purely linear* relationship. If $\sigma_{XY} = 0$, however, then either the relationship between X and Y can be assumed to be of some *nonlinear* type, or else the variables are *independent.* If σ_{XY} lies between these values in magnitude, then we conclude that X and Y have a relationship which is a mixture of linear and other components.

A somewhat undesirable aspect of the covariance measure is that its value changes if we transform the variables involved to other units. For example, if the variables are X and Y and we transform to new variables

$$X^* = cX, \quad Y^* = dY,$$

note that

$$\begin{aligned} \mathrm{Cov}(X^*, Y^*) &= E\left[(X^* - E(X^*))(Y^* - E(Y^*))\right] \\ &= E\left[c(X - E(X))d(Y - E(Y))\right] \\ &= cd\ \mathrm{Cov}(X, Y)\ . \end{aligned}$$

Thus, counter to our intuition that dependence relationships should not be altered by simply rescaling the variables, the covariance measure indeed is affected.

It is preferable to have a dependence measure which is not sensitive to irrelevant details such as units of measurement. Next we see how to convert to an equivalent measure which is "dimensionless" in this sense.

Correlation

Recall that the upper and lower limits for the possible values which the covariance can take are given in terms of the variances, which also change with rescaling of variables. Consequently, in order to decide whether a particular value for the covariance is "big" or "small," we need to assess it *relative to the variances of the two variables.* One way to do this is to *divide the covariance by the product of the standard deviations of the variables*, producing the quantity

$$\rho_{XY} = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}\ ,$$

which we call the *correlation of X and Y.* It is easily seen (*check*) that this measure does *not* change when we rescale the variables X and Y to $X^* = cX$, $Y^* = dY$

as considered above. Also, the upper and lower limits we saw for the covariance translate to the following limits for the value of the correlation:

$$-1 \leq \rho_{XY} \leq 1 \ .$$

Thus we can judge the *strength of the dependence* by how close the correlation measure comes to either of its extreme values +1 or −1 (and away from the value 0 that suggests an absence of dependence). However, keep in mind that — like the covariance — the correlation is especially sensitive to a certain kind of dependence, namely *linear dependence*, and can have a small value (near 0) even when there is strong dependence of other kinds. (We saw in the previous lecture an example when the covariance — and hence the correlation — is 0 even when there is strong dependence, but of a nonlinear kind.)

Let us now illustrate the use of the correlation parameter to judge the degree of (linear) dependence in the following probability distribution which we considered above:

		Y = 0	Y = 1
X	0	.7	.1
	1	.1	.1

We had calculated its covariance to be $\sigma_{XY} = .06$, but at that time we did not assess how meaningful this value is. Now we do so, by converting to correlation. First finding the variances $\sigma_X^2 = .16 = \sigma_Y^2$, we then compute

$$\rho_{XY} = \frac{.06}{\sqrt{.16}\sqrt{.16}} = .36 \ ,$$

which we might interpret as "moderate" (not close to 0, but not close to 1, either). Moreover, note that a *positive* value of correlation is obtained in this example, indicating that deviations of X from its mean tend to be associated with deviations of Y from its mean in the *same* direction.

In what cases does correlation attain the value +1? −1? Suppose that Y is related to X *exactly as a linear transformation of* X: $Y = a + bX$, for some choice of a, b. Then we have the following analysis:

$$E(XY) = E[X(a + bX)] = aE(X) + bE(X^2)$$

and so

$$\begin{aligned} \sigma_{XY} &= E(XY) - E(X)E(Y) = [aE(X) + bE(X^2)] - E(X)E(a + bX) \\ &= \cdots \\ &= b\sigma_X^2 \end{aligned}$$

and thus, finally,

$$\rho_{XY} = \frac{b\sigma_X^2}{\sigma_X \sigma_Y} = \frac{b}{|b|} \ ,$$

which we see takes value $+1$ if $b > 0$ and value -1 if $b < 0$. It can be shown, also, that the values $+1$ and -1 are attainable *only* in the case of *exactly linear* relationships. This is the basis for characterizing correlation as a measure of the "degree of linear relationship" between X and Y. Despite this intuitive appeal of the correlation measure, note that it really doesn't leave us with a precise meaning in the case of a value intermediate between 0 and ± 1. intermediate values. $\square$

Despite any shortcomings the correlation measure may have, it has very wide application. This is based on

- its fairly successful *intuitive appeal* as a measure of dependence,
- the central importance of confirming (or disproving) *linear* relationships,
- its application to calculating *variances of linear combinations of r.v.'s*,
- its application in analyzing and interpreting *regression models.*

As a final illustration, let us consider the special case that Y is defined to be just X itself, i.e., $Y = X$. This is a special case of the *linear* transformation with $a = 0$ and $b = 1$, so we have: $\rho = 1$, i.e.,

$$\mathrm{Corr}(X, X) = 1 \; .$$

Likewise, we can talk of the "covariance of a r.v. X with itself," and we readily find that it reduces to the variance of X:

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) \; .$$

(To see this immediately, just go back and check the definition of covariance.)