5 Joint Probability Distributions and Random Samples

5.2 Expected Values, Covariance, and Correlation

Expected Values, Covariance, and Correlation

Any function $h(X)$ of a single rv X is itself a random variable.

However, to compute $E[h(X)]$, it is not necessary to obtain the probability distribution of $h(X)$; instead, $E[h(X)]$ is computed as a weighted average of $h(x)$ values, where the weight function is the pmf $p(x)$ or pdf $f(x)$ of X.

A similar result holds for a function $h(X, Y)$ of two jointly distributed random variables.

3

Expected Values, Covariance, and Correlation

Proposition

Let X and Y be jointly distributed rv's with pmf $p(x, y)$ or pdf $f(x, y)$ according to whether the variables are discrete or continuous.

Then the expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X, Y)}$, is given by

$$E[h(X, Y)] = \begin{cases} \sum_x \sum_y h(x, y) \cdot p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y)\, dx\, dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

4

Example 13

Five friends have purchased tickets to a certain concert. If the tickets are for seats 1–5 in a particular row and the tickets are randomly distributed among the five, what is the expected number of seats separating any particular two of the five?

Let X and Y denote the seat numbers of the first and second individuals, respectively. Possible (X, Y) pairs are $\{(1, 2), (1, 3), \ldots, (5, 4)\}$, and the joint pmf of (X, Y) is

$$p(x, y) = \begin{cases} \frac{1}{20} & x = 1, \ldots, 5;\ y = 1, \ldots, 5;\ x \neq y \\ 0 & \text{otherwise} \end{cases}$$

Example 13

cont'd

The number of seats separating the two individuals is $h(X, Y) = |X - Y| - 1$.

The accompanying table gives $h(x, y)$ for each possible (x, y) pair.

$h(x, y)$		x = 1	2	3	4	5
	1	—	0	1	2	3
	2	0	—	0	1	2
y	3	1	0	—	0	1
	4	2	1	0	—	0
	5	3	2	1	0	—

Example 13

cont'd

Thus

$$E[h(X, Y)] = \sum_{(x, y)}\sum h(x, y) \cdot p(x, y)$$

$$= \sum_{x=1}^{5}\sum_{\substack{y=1 \\ x \neq y}}^{5} (|x - y| - 1) \cdot \frac{1}{20}$$

$$= 1$$

Covariance

Covariance

When two random variables X and Y are not independent, it is frequently of interest to assess how strongly they are related to one another.

Definition

The **covariance** between two rv's X and Y is

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

$$= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) & X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x, y)\, dx\, dy & X, Y \text{ continuous} \end{cases}$$

9

Covariance

That is, since $X - \mu_X$ and $Y - \mu_Y$ are the deviations of the two variables from their respective mean values, the covariance is the expected product of deviations. Note that $\text{Cov}(X, X) = E[(X - \mu_X)^2] = V(X)$.

The rationale for the definition is as follows.

Suppose X and Y have a strong positive relationship to one another, by which we mean that large values of X tend to occur with large values of Y and small values of X with small values of Y.

10

Covariance

Then most of the probability mass or density will be associated with $(x - \mu_X)$ and $(y - \mu_Y)$, either both positive (both X and Y above their respective means) or both negative, so the product $(x - \mu_X)(y - \mu_Y)$ will tend to be positive.

Thus for a strong positive relationship, $\text{Cov}(X, Y)$ should be quite positive.

For a strong negative relationship, the signs of $(x - \mu_X)$ and $(y - \mu_Y)$ will tend to be opposite, yielding a negative product.

11

Covariance

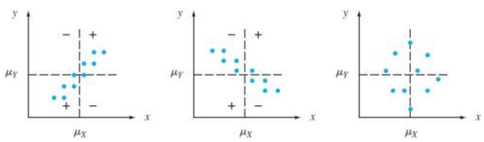
Thus for a strong negative relationship, $\text{Cov}(X, Y)$ should be quite negative.

If X and Y are not strongly related, positive and negative products will tend to cancel one another, yielding a covariance near 0.

12

Covariance

Figure 5.4 illustrates the different possibilities. The covariance depends on *both* the set of possible pairs and the probabilities. In Figure 5.4, the probabilities could be changed without altering the set of possible pairs, and this could drastically change the value of Cov(X, Y).



$p(x, y)$ = 1/10 for each of ten pairs corresponding to indicated points:

(a) positive covariance; (b) negative covariance; (c) covariance near zero

Figure 5.4

Example 15

The joint and marginal pmf's for

X = automobile policy deductible amount and

Y = homeowner policy deductible amount in Example 5.1 were

$p(x, y)$		y = 0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

x	100	250
$p_X(x)$	.5	.5

y	0	100	200
$p_Y(y)$	.25	.25	.5

from which $\mu_X = \Sigma x p_X(x) = 175$ and $\mu_Y = 125$.

Example 15

cont'd

Therefore,

$$\text{Cov}(X, Y) = \sum_{(x,y)}\sum (x - 175)(y - 125)p(x, y)$$

$$= (100 - 175)(0 - 125)(.20) + \ldots + (250 - 175)(200 - 125)(.30)$$

$$= 1875$$

Covariance

The following shortcut formula for Cov(X, Y) simplifies the computations.

Proposition

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

According to this formula, no intermediate subtractions are necessary; only at the end of the computation is $\mu_X \cdot \mu_Y$ subtracted from $E(XY)$. The proof involves expanding $(X - \mu_X)(Y - \mu_Y)$ and then taking the expected value of each term separately.

Correlation

Correlation

Definition

The **correlation coefficient** of X and Y, denoted by Corr(X, Y), $\rho_{X,Y}$, or just ρ, is defined by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Example 17

It is easily verified that in the insurance scenario of Example 15, $E(X^2) = 36{,}250$,

$\sigma_X^2 = 36{,}250 - (175)^2 = 5625$,

$\sigma_X = 75$, $E(Y^2) = 22{,}500$,

$\sigma_Y^2 = 6875$, and $\sigma_Y = 82.92$.

This gives

$$\rho = \frac{1875}{(75)(82.92)} = .301$$

Correlation

The following proposition shows that ρ remedies the defect of Cov(X, Y) and also suggests how to recognize the existence of a strong (linear) relationship.

Proposition

1. If a and c are either both positive or both negative,

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

2. For any two rv's X and Y, $-1 \le \text{Corr}(X, Y) \le 1$.

Correlation

If we think of $p(x, y)$ or $f(x, y)$ as prescribing a mathematical model for how the two numerical variables X and Y are distributed in some population (height and weight, verbal SAT score and quantitative SAT score, etc.), then ρ is a population characteristic or parameter that measures how strongly X and Y are related in the population.

We will consider taking a sample of pairs $(x_1, y_1), \ldots, (x_n, y_n)$ from the population.

The sample correlation coefficient r will then be defined and used to make inferences about ρ.

21

Correlation

The correlation coefficient ρ is actually not a completely general measure of the strength of a relationship.

Proposition

1. If X and Y are independent, then $\rho = 0$, but $\rho = 0$ does not imply independence.

2. $\rho = 1$ or -1 iff $Y = aX + b$ for some numbers a and b with $a \neq 0$.

22

Correlation

This proposition says that ρ is a measure of the degree of **linear** relationship between X and Y, and only when the two variables are perfectly related in a linear manner will ρ be as positive or negative as it can be.

A ρ less than 1 in absolute value indicates only that the relationship is not completely linear, but there may still be a very strong nonlinear relation.

23

Correlation

Also, $\rho = 0$ does not imply that X and Y are independent, but only that there is a complete absence of a linear relationship. When $\rho = 0$, X and Y are said to be **uncorrelated.**

Two variables could be uncorrelated yet highly dependent because there is a strong nonlinear relationship, so be careful not to conclude too much from knowing that $\rho = 0$.

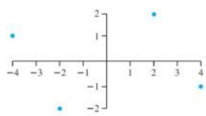
24

Example 18

Let X and Y be discrete rv's with joint pmf

$$p(x, y) = \begin{cases} \frac{1}{4} & (x, y) = (-4, 1), (4, -1), (2, 2), (-2, -2) \\ 0 & \text{otherwise} \end{cases}$$

The points that receive positive probability mass are identified on the (x, y) coordinate system in Figure 5.5.



The population of pairs for Example 18

Figure 5.5

Example 18

cont'd

It is evident from the figure that the value of X is completely determined by the value of Y and vice versa, so the two variables are completely dependent. However, by symmetry $\mu_X = \mu_Y = 0$ and

$$E(XY) = (-4)\frac{1}{4} + (-4)\frac{1}{4} + (4)\frac{1}{4} + (4)\frac{1}{4} = 0$$

The covariance is then $\text{Cov}(X,Y) = E(XY) - \mu_X \cdot \mu_Y = 0$ and thus $\rho_{X,Y} = 0$. Although there is perfect dependence, there is also complete absence of any linear relationship!

Correlation

A value of ρ near 1 does not necessarily imply that increasing the value of X *causes* Y to increase. It implies only that large X values are *associated* with large Y values.

For example, in the population of children, vocabulary size and number of cavities are quite positively correlated, but it is certainly not true that cavities cause vocabulary to grow.

Instead, the values of both these variables tend to increase as the value of age, a third variable, increases.

Correlation

For children of a fixed age, there is probably a low correlation between number of cavities and vocabulary size.

In summary, association (a high correlation) is not the same as causation.