5

The Distribution of Two Random Variables

Often we may be interested in the joint behavior of two (or more) random variables defined on the same probability space.

5.1 JOINT DISTRIBUTIONS

Example 5.1 Given the chance experiment of tossing 3 fair coins, consider the two random variables X = "no. of heads" and Y = "no. of changes in the sequence of outcomes". The sample space of this experiment and the values taken by these two random variables are shown in Table 9. It is clear from the table that

$$\Pr(\text{“}X = 1 \text{ and } Y = 1\text{”}) = \Pr(HTT, TTH) = 1/4.$$

This probability is called the *joint probability* that $X = 1$ and $Y = 1$. □

More generally, given two discrete random variables X and Y, let

$$p(x, y) = \Pr(X = x, Y = y) = \Pr(\text{“}X = x \text{ and } Y = y\text{”}).$$

The collection of $p(x, y)$ for all possible values x and y of X and Y defines the *joint probability distribution* of X and Y. The joint probability $p(x, y)$, viewed as a function of x and y, is called the *joint frequency function* of X and Y. A joint frequency function must satisfy:

1. $0 \leq p(x, y) \leq 1$ for any pair (x, y),
2. $\sum_x \sum_y p(x, y) = 1$.

We say that $p(x, y)$ is a *valid* joint frequency function if it satisfies these two properties.

Example 5.2 The joint probability distribution of X and Y in Example 5.1 may be summarized by the following table:

	0	1	2
0	1/8	0	0
1	0	1/4	1/8
2	0	1/4	1/8
3	1/8	0	0

Table 9 Sample space of the experiment consisting of tossing 3 fair coins and values taken by the two random variables X = "no. of heads" and Y = "no. of changes in the sequence of outcomes".

points in the sample space	x	y
TTT	0	0
TTH	1	1
THT	1	2
HTT	1	1
HHT	2	1
HTH	2	2
THH	2	1
HHH	3	0

where the rows refer to the values of X and the columns to the values of Y. This represents a valid probability assignment because $0 \leq p(x, y) \leq 1$ for all pair (x, y) and $\sum_x \sum_y p(x, y) = 1$.

A three-dimensional bar graph of the joint probability distribution of X and Y is shown in Figure 13. □

If X and Y are continuous random variables, then their *joint density function* is any nonnegative function $f(x, y)$ such that

$$F(a, b) = \Pr(X \leq a, Y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y)\, dx\, dy.$$

5.2 MARGINAL DISTRIBUTIONS

Definition 5.1 Given two discrete random variables X and Y, the probability $p_X(x) = \Pr(X = x)$, viewed as a function of x, is called the *marginal frequency function* of X. Similarly, the probability $p_Y(y) = \Pr(Y = y)$, viewed as function of y, is called the *marginal frequency function* of Y. □

What is the relationship between the marginal frequency functions $p_X(x)$ and $p_Y(y)$ and the joint frequency function $p(x, y)$?

Example 5.3 In the case of Example 5.1, we have

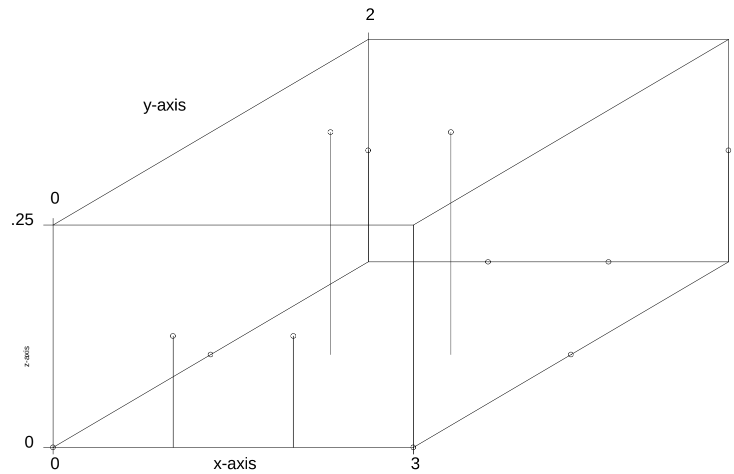
$$\begin{aligned} p_X(2) &= \Pr(X = 2, Y = 0) + \Pr(X = 2, Y = 1) + \Pr(X = 2, Y = 2) \\ &= p(2, 0) + p(2, 1) + p(2, 2) = \frac{3}{8}. \end{aligned}$$

□

In general, we have

$$p_X(x) = \sum_y p(x, y), \qquad p_Y(y) = \sum_x p(x, y).$$

Figure 13 Three-dimensional bar graph of the joint probability distribution of X and Y in Example 5.1.



Example 5.4 The joint probability table in Example 5.1 may be modified as follows in order to display the marginal frequencies of X and Y:

	0	1	2	$p_X(x)$
0	1/8	0	0	1/8
1	0	1/4	1/8	3/8
2	0	1/4	1/8	3/8
3	1/8	0	0	1/8
$p_Y(y)$	1/4	1/2	1/4	1.0

Notice that the marginal probabilities of X are obtained by adding along the rows of the table, whereas the marginal probabilities of Y are obtained by adding along the columns. □

If the joint distribution of Y and X is continuous, the *marginal density functions* of X and Y are defined as

$$f_X(x) = \int f(x,y)\,dy, \qquad f_Y(y) = \int f(x,y)\,dx,$$

respectively, where $f(x,y)$ is the joint density function of X and Y.

Although we can always go fron the joint distribution of two random variables to their marginal distributions, the converse is not true in general. In other words, knowledge of the marginal distributions of two random variables is not generally enough to recover their joint distribution. As we shall see, however, there is one important case in which this is possible.

5.3 CONDITIONAL DISTRIBUTIONS

Definition 5.2 Given two discrete random variables X and Y, the conditional probability of the event $\{Y = y\}$ given the event $\{X = x\}$ is called the *conditional probability of* $Y = y$ *given* $X = x$ and denoted by

$$p(y \mid x) = \Pr(Y = y \mid X = x).$$

□

To find $p(y \mid x)$, just put $A = \{Y = y\}$ and $B = \{X = x\}$ in the definition of conditional probability. Provided that $p_X(x) = \Pr(X = x) > 0$, we get

$$p(y \mid x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} = \frac{p(x, y)}{p_X(x)}.$$

Notice that, if we now consider two possible values y and y' of Y, their odds conditional on $X = x$ are

$$\frac{p(y \mid x)}{p(y' \mid x)} = \frac{p(x, y)}{p(x, y')}.$$

Example 5.5 In the case of Example 5.1 we have

$$p(1 \mid 1) = \frac{p(1, 1)}{p_X(1)} = \frac{1/4}{3/8} = 2/3,$$
$$p(2 \mid 1) = \frac{p(2, 1)}{p_X(1)} = \frac{1/8}{3/8} = 1/3.$$

and

$$\frac{p(1 \mid 1)}{p(2 \mid 1)} = \frac{p(1, 1)}{p(2, 1)} = 2.$$

□

The collection of $p(y \mid x)$ for all possible values of Y defines the *conditional probability distribution of* Y *given* $X = x$. The conditional probability $p(y \mid x)$, viewed as a function of y for x fixed, is called the *conditional probability function of* Y *given* $X = x$. A conditional frequency function must satisfy:

1. $0 \le p(y \mid x) \le 1$ for all y and any x,
2. $\sum_y \sum_y p(y \mid x) = 1$ for any x.

The conditional distribution of X given $Y = y$ and the conditional frequency function of X given $Y = y$ are similarly defined. To avoid ambiguities, we may sometimes write the conditional distribution of Y given $X = x$ as $p_{Y|X}(y \mid x)$ and the conditional distribution of X given $Y = y$ as $p_{X|Y}(x \mid y)$. Notice that the conditional frequency function of Y given $X = x$ changes as x changes. Similarly, the conditional frequency function of X given $Y = y$ changes as y changes.

Example 5.6 In the case of Example 5.1, there are 3 conditional frequency functions for X, one for each of the possible values of X:

	0	1	2	3
$p(x \mid 0)$	1/2	0	0	1/2
$p(x \mid 1)$	0	1/2	1/2	0
$p(x \mid 2)$	0	1/2	1/2	0

depending on whether $Y = 0, 1$ or 2. □

If the joint distribution of Y and X is continuous, the *conditional density function* of Y given $X = x$ is defined as

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)},$$

provided that $f_X(x) > 0$.

5.4 MEANS AND VARIANCES OF CONDITIONAL DISTRIBUTIONS

Since the conditional distribution of Y given $X = x$ is a well defined probability distribution, its mean and variance can be computed in the standard way.

Consider first the case when both X and Y are discrete random variables. The conditional mean of Y given $X = x$ may then be computed by averaging the possible values of Y using $p(y \mid x)$ as probability weights. Thus, the conditional mean Y given $X = x$ is

$$\mu(x) = \mathrm{E}(Y \mid X = x) = \sum_y y\, p(y \mid x),$$

which, is sometimes called the *regression function* of Y given X. The conditional variance of Y given $X = x$ is

$$\begin{aligned} \mathrm{Var}(Y \mid X = x) &= \sum_y [y - \mu(x)]^2\, p(y \mid x), \\ &= \sum_y y^2 p(y \mid x) - [\mu(x)]^2 \\ &= \mathrm{E}(Y^2 \mid X = x) - [\mu(x)]^2. \end{aligned}$$

The conditional mean and variance of X given $Y = y$ are similarly defined.

Example 5.7 In the case of Example 5.1, we have

$$\mathrm{E}(X \mid Y = 0) = 1.5$$
$$\mathrm{E}(X \mid Y = 1) = 1.5$$

In this example, the two conditional distributions have the same mean. However

$$\begin{aligned}\mathrm{Var}(X \mid Y = 0) &= 9/4\\ \mathrm{Var}(X \mid Y = 1) &= 1/4\end{aligned}$$

Thus, the distribution of X conditional on $Y = 0$ is more spread out than the conditional one given $Y = 1$. □

If the joint distribution of Y and X is continuous, then the conditional mean of Y given $X = x$ is defined as

$$\mu(x) = \mathrm{E}(Y \mid X = x) = \int y\, f(y \mid x)\, dy,$$

and the conditional variance of Y given $X = x$ is defined as

$$\mathrm{Var}(Y \mid X = x) = \int [y - \mu(x)]^2\, f(y \mid x)\, dy \int y^2 f(y \mid x)\, dy - [\mu(x)]^2.$$

There exists an important relationship between the conditional mean of Y and the unconditional or marginal mean $\mathrm{E}(Y)$. This relationship is easier to establish when both X and Y are discrete random variables. Taking the mean of the possible values of $\mu(x)$ gives

$$\begin{aligned}\mathrm{E}[\mu(X)] = \sum_x \mu(x) p_X(x) &= \sum_x \left[\sum_y y\, p(y \mid x)\right] p_X(x)\\ &= \sum_y y \left[\sum_x p(y \mid x)\, p_X(x)\right] = \sum_y y\, p_Y(y).\end{aligned}$$

Thus,

$$\mathrm{E}[\mu(X)] = \mathrm{E}(Y),$$

that is, the mean of the conditional mean is equal to the unconditional mean. This relationship is known as the *Law of Iterated Expectations*. By a similar argument, one can show that

$$\mathrm{Var}(Y) = \mathrm{E}[\sigma^2(X)] + \mathrm{Var}[\mu(X)],$$

where $\sigma^2(x) = \mathrm{Var}(Y \mid X = x)$. This relationship is known as the *Law of Total Variance.*

5.5 INDEPENDENCE

Recall that two events A and B are independent if

$$\Pr(B \mid A) = \Pr(B)$$

or, equivalently,

$$\Pr(A \cap B) = \Pr(A)\ \Pr(B).$$

Therefore, given two discrete random variables X and Y, the two events $\{X = x\}$ and $\{Y = y\}$ are independent if

$$p_{Y|X}(y \mid x) = p_Y(y) \tag{5.1}$$

or, equivalently,

$$p(x, y) = p_X(x)\, p_Y(y). \tag{5.2}$$

If (5.1) or (5.2) are true for all pairs (x, y), then we say that X and Y are *independent random variables*. It is clear from (5.2) that if X and Y are independent, then their joint distribution can be reconstructed by knowledge of their marginal distributions.

Example 5.8 In the case of Example 5.1 we have

$$p_{X|Y}(0 \mid 0) = \frac{1}{2} \neq \frac{1}{8} = p_X(0).$$

Thus X and Y cannot be independent. □

If the joint distribution of X and Y is continuous, then X and Y are independent if

$$f_{Y|X}(y \mid x) = f_Y(y)$$

or, equivalently,

$$f(x, y) = f_X(x)\, f_Y(y).$$

5.6 FUNCTIONS OF TWO RANDOM VARIABLES

Example 5.9 Consider the experiment of tossing 3 fair coins. Given the random variables X = "no. of heads" and Y = "no. of changes in the sequence of outcomes", define the new random variable $W = 2X - Y$. It is easy to see that:

w	$p_W(w)$	$w\, p_W(w)$
0	1/4	0
1	1/4	1/4
2	1/8	1/4
3	1/4	3/4
6	1/8	3/4

Therefore

$$\mu_W = \mathrm{E}(W) = \sum_w w\, p_W(w) = 2.0.$$

The same result can also be obtained directly from the joint distribution of X and Y in Table 9 by using the formula

$$\mu_W = \sum_x \sum_y (2x - y)\, p(x, y).$$

□

In general, if X and Y are discrete random variables and $W = g(X,Y)$, then

$$\mu_W = \mathrm{E}(W) = \sum_x \sum_y g(x,y)\, p(x,y).$$

Thus, we do not need to tabulate $p_W(w)$ and we can just work with the joint distribution of X and Y.

If the joint distribution of X and Y is continuous, then

$$\mu_W = \mathrm{E}(W) = \int \int g(x,y)\, f(x,y)\, dx\, dy.$$

5.6.1 LINEAR COMBINATIONS OF TWO RANDOM VARIABLES

We now consider the important special case when the transformation $g(X,Y)$ is of the form

$$g(X,Y) = a + bX + cY, \tag{5.3}$$

where a, b and c are arbitrary constants. Such a transformation will be called *linear* although, strictly speaking, the term *affine* should be used and the term linear reserved for the special case when $a = 0$.

Examples of (5.3) include the sum $g(X,Y) = X+Y$ of two random variables, where $a = 0$ and $b = c = 1$, and the difference $g(X,Y) = X - Y$ of two random variables, where $a = 0$ and $b = -c = 1$.

If $W = g(X,Y)$ is a linear transformation of X and Y, then the mean of W is easily obtained from the mean of X and Y as

$$\mu_W = a + b\mu_X + c\mu_Y.$$

where $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$. We give the proof for the discrete case.

Proof.

$$\begin{aligned}
\mathrm{E}(W) &= \sum_x \sum_y (a + bx + cy)\, p(x,y) \\
&= a \sum_x \sum_y p(x,y) + b \sum_x \sum_y x\, p(x,y) + c \sum_x \sum_y y\, p(x,y) \\
&= a + b \sum_x x \left[\sum_y p(x,y)\right] + c \sum_y y \left[\sum_x p(x,y)\right] \\
&= a + b \sum_x x\, p_X(x) + c \sum_y y\, p_Y(y) \\
&= a + b\mu_X + c\mu_Y,
\end{aligned}$$

where we used the fact that $\sum_x \sum_y p(x,y) = 1$, $\sum_y p(x,y) = p_X(x)$ and $\sum_x p(x,y) = p_Y(y)$. □

As a special case of the above relationship we obtain

$$\mathrm{E}(X+Y) = \mu_X + \mu_Y, \qquad \mathrm{E}(X-Y) = \mu_X - \mu_Y.$$

In order to determine what is the variance of W, we need to introduce first the concept of covariance.

5.6.2 COVARIANCE

We are often interested in studying whether and how two random variables X and Y "vary together". Recall that a measure of how variable is X alone is the variance

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)^2] = \mathrm{E}(X^2) - \mu_X^2.$$

To measure the "co-variation" of X and Y we consider the following measure, called the *covariance* between X and Y,

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

where $\mu_Y = \mathrm{E}(Y)$. Clearly $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. If X and Y are discrete random variables, then

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)\, p(x, y) \\ &= \sum_x \sum_y xy\, p(x, y) - \mu_X \mu_Y.\end{aligned}$$

If the joint distribution of X and Y is continuous, then

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \int\int (x - \mu_X)(y - \mu_Y)\, f(x, y)\, dx\, dy \\ &= \int\int xy\, f(x, y)\, dx\, dy - \mu_X \mu_Y.\end{aligned}$$

Does the proposed measure make sense? Suppose that the joint distribution of two discrete random variables X and Y is as in Figure 14. Each point in the graph correspond to a pair of possible value of X and Y. Assuming that all points are equally probable, each of them receives probability 1/10. It is then clear from the graph that high probability weight is assigned to values of X and Y such that, either $x - \mu_X > 0$ and $y - \mu_Y > 0$, or $x - \mu_X < 0$ and $y - \mu_Y < 0$. Since deviations from the mean has the same sign with high probability, we conclude that $\mathrm{Cov}(X, Y) > 0$ in this case.

Suppose now that the joint distribution of X and Y is as in Figure 15. Assume again that all points are equally probable. In this case, high probability weight is assigned to values of X and Y such that, either $x - \mu_X > 0$ and $y - \mu_Y < 0$, or $x - \mu_X < 0$ and $y - \mu_Y > 0$. Since deviations from the mean has opposite sign with high probability, we conclude that $\mathrm{Cov}(X, Y) < 0$ in this case.

In some cases, positive and negative deviations cancel out so that $\mathrm{Cov}(X, Y) = 0$. In this case, we say that X and Y are *uncorrelated*.

Example 5.10 In the case of Example 5.1 we obtain

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)\, p(x, y) \\ &= (-1.5)(-1)(1/8) + (3 - 1.5)(-1)(1/8) + \\ &\quad + (1 - 1.5)(1 - 1)(1/4) + (2 - 1.5)(1 - 1)(1/4) + \\ &\quad + (1 - 1.5)(2 - 1)(1/8) + (2 - 1.5)(2 - 1)(1/8) \\ &= (1.5 - 1.5 - .5 + .5)(1/8) = 0.\end{aligned}$$

Figure 14 Scatterplot of positively correlated random variables.

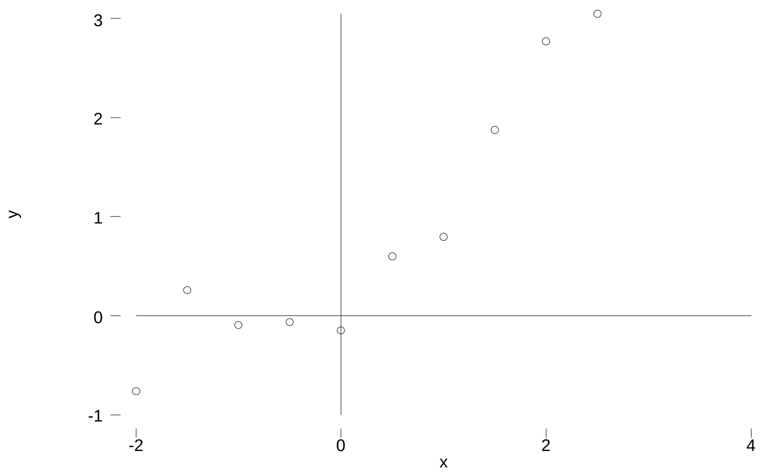
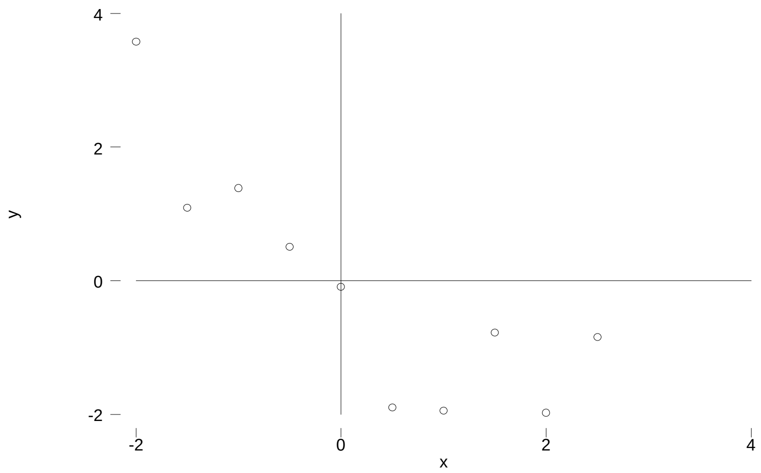


Figure 15 Scatterplot of negatively correlated random variables.



Thus, X and Y are uncorrelated. However, as we have already seen, they are not independent. □

What happens if X and Y are independent? In this case

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \sum_x \sum_y xy\, p(x,y) - \mu_X \mu_Y \\
&= \sum_x \sum_y xy\, p_X(x)\, p_Y(y) - \mu_X \mu_Y \\
&= \left[\sum_x x\, p_X(x)\right] \left[\sum_y y\, p_Y(y)\right] - \mu_X \mu_Y \\
&= \mu_X \mu_Y - \mu_X \mu_Y = 0.
\end{aligned}$$

We therefore conclude that if X and Y are independent, then they are also uncorrelated. The converse, however, is not true as the previous example demonstrates.

We are now in the position to be able to derive the following formula for the variance of the linear transformation $W = a + bX + cY$

$$\mathrm{Var}(W) = b^2\,\mathrm{Var}(X) + c^2\,\mathrm{Var}(Y) + 2bc\,\mathrm{Cov}(X,Y). \tag{5.4}$$

Proof.

$$\begin{aligned}
\mathrm{Var}(W) &= \mathrm{E}[(W - \mu_W)^2] \\
&= \mathrm{E}[(b(X-\mu_X) + c(Y-\mu_Y))^2] \\
&= \mathrm{E}[b^2(X-\mu_X)^2 + c^2(Y-\mu_Y)^2 + 2bc(X-\mu_X)(Y-\mu_Y)] \\
&= b^2\,\mathrm{E}[(X-\mu_X)^2] + c^2\,\mathrm{E}[(Y-\mu_Y)^2] + 2bc\,\mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= b^2\,\mathrm{Var}(X) + c^2\,\mathrm{Var}(Y) + 2bc\,\mathrm{Cov}(X,Y).
\end{aligned}$$

□

If $\mathrm{Cov}(X,Y) = 0$, that is, X and Y are uncorrelated, then (5.4) becomes

$$\mathrm{Var}(W) = b^2\,\mathrm{Var}(X) + c^2\,\mathrm{Var}(Y).$$

In particular, if X and Y are uncorrelated, then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

5.6.3 CORRELATION

One problem with the covariance as a measure of the degree of association between two random variables is that it depends on the units in which the two random variables are measured.

Example 5.11 Replace X by $W = 100X$ in the definition of covariance. Then $\mathrm{E}(W) = 100\,\mathrm{E}(X)$ and therefore

$$\begin{aligned}
\mathrm{Cov}(W,Y) &= \mathrm{E}(WY) - \mathrm{E}(W)\,\mathrm{E}(Y) \\
&= 100\,\mathrm{E}(XY) - 100\,\mathrm{E}(X)\,\mathrm{E}(Y) = 100\,\mathrm{Cov}(X,Y).
\end{aligned}$$

□

To eliminate the undesirable effect of the scale of measurement, we can "standardize" $\mathrm{Cov}(X,Y)$ dividing by the product of the standard deviations σ_X and σ_Y of X and Y, provided that both are positive. The resulting number is called the *correlation* between X and Y

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \mathrm{E}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right].$$

Example 5.12 Consider again Example 5.11. Since $W = 100X$, we have that $\sigma_W = 100\sigma_X$. Therefore

$$\mathrm{Corr}(W,Y) = \frac{\mathrm{Cov}(W,Y)}{\sigma_W\,\sigma_Y} = \frac{100\,\mathrm{Cov}(X,Y)}{100\sigma_X\,\sigma_Y} = \mathrm{Corr}(X,Y).$$

□

Cleary, $\mathrm{Corr}(X,Y) = 0$ if and only if $\mathrm{Cov}(X,Y) = 0$. Moreover, one can prove that

$$-1 \le \mathrm{Corr}(X,Y) \le 1.$$

To better understand what is measured by correlations, suppose that it is known that Y is exactly a linear transformation of the random variable X, that is $Y = a+bX$, where $b \neq 0$. Then $\mu_Y = a + b\mu_X$ and therefore

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}[(X-\mu_X)(Y-\mu_Y)]\\ &= \mathrm{E}[(X-\mu_X)(a+bX-a-b\mu_X)]\\ &= b\,\mathrm{E}[(X-\mu_X)(X-\mu_X)] = b\,\mathrm{Var}(X).\end{aligned}$$

Moreover, since $\sigma_Y = |b|\sigma_X$, we also have

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{b\sigma_X^2}{|b|\sigma_X^2} = \begin{cases}1, & \text{if } b>0,\\ -1, & \text{if } b<0.\end{cases}$$

Conversely, it can be shown that if $|\,\mathrm{Corr}(X,Y)| = 1$, then there exists an exact linear relationship between X and Y. If the relationship between X and Y is not linear, but $|\,\mathrm{Corr}(X,Y)|$ is near one, then we can conclude that the relationship between X and Y could well be approximated by a linear one.

Notice that if there exists an exact relationship between X and Y, but one that is not linear, then $|\,\mathrm{Corr}(X,Y)| < 1$. Thus, correlation is only a measure of the strength of a linear relationship and therefore it may not be able to detect nonlinear relationships, even when they are exact.

Example 5.13 Suppose that X and Y are related by the following exact relationship

$$X^2 + Y^2 = 1,$$

that is, each (X,Y) pair lies on a circle centered at the origin with radius equal to one. It can be shown that in this case $\mathrm{Corr}(X,Y) = 0$. □