10

Covariance and correlation

In this chapter we see how the joint distribution of two or more random variables is used to compute the expectation of a combination of these random variables. We discuss the expectation and variance of a sum of random variables and introduce the notions of *covariance* and *correlation*, which express to some extent the way two random variables influence each other.

10.1 Expectation and joint distributions

China vases of various shapes are produced in the Delftware factories in the old city of Delft. One particular simple cylindrical model has height H and radius R centimeters. Due to all kinds of circumstances—the place of the vase in the oven, the fact that the vases are handmade, etc.—H and R are not constants but are random variables. The volume of a vase is equal to the random variable $V = \pi H R^2$, and one is interested in its expected value $\mathrm{E}[V]$. When f_V denotes the probability density of V, then by definition

$$\mathrm{E}[V] = \int_{-\infty}^{\infty} v f_V(v)\,\mathrm{d}v.$$

However, to obtain $\mathrm{E}[V]$, we do not necessarily need to determine f_V from the joint probability density f of H and R! Since V is a function of H and R, we can use a rule similar to the change-of-variable formula from Chapter 7:

$$\mathrm{E}[V] = \mathrm{E}\left[\pi H R^2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi h r^2 f(h,r)\,\mathrm{d}h\,\mathrm{d}r.$$

Suppose that H has a $U(25, 35)$ distribution and that R has a $U(7.5, 12.5)$ distribution. In the case that H and R are also independent, we have

$$\mathrm{E}[V] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi h r^2 f_H(h) f_R(r)\,\mathrm{d}h\,\mathrm{d}r = \int_{25}^{35}\int_{7.5}^{12.5} \pi h r^2 \cdot \frac{1}{10} \cdot \frac{1}{5}\,\mathrm{d}h\,\mathrm{d}r$$
$$= \frac{\pi}{50}\int_{25}^{35} h\,\mathrm{d}h \int_{7.5}^{12.5} r^2\,\mathrm{d}r = 9621.127\ \mathrm{cm}^3.$$

This illustrates the following general rule.

TWO-DIMENSIONAL CHANGE-OF-VARIABLE FORMULA. Let X and Y be random variables, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function.
If X and Y are *discrete* random variables with values $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, respectively, then

$$\mathrm{E}[g(X,Y)] = \sum_i \sum_j g(a_i, b_j)\mathrm{P}(X = a_i, Y = b_j)\,.$$

If X and Y are *continuous* random variables with joint probability density function f, then

$$\mathrm{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

As an example, take $g(x,y) = xy$ for discrete random variables X and Y with the joint probability distribution given in Table 10.1. The expectation of XY is computed as follows:

$$\begin{aligned}\mathrm{E}[XY] &= (0\cdot 0)\cdot 0 + (1\cdot 0)\cdot\frac{1}{4} + (2\cdot 0)\cdot 0\\ &\quad + (0\cdot 1)\cdot\frac{1}{4} + (1\cdot 1)\cdot 0 + (2\cdot 1)\cdot\frac{1}{4}\\ &\quad + (0\cdot 2)\cdot 0 + (1\cdot 2)\cdot\frac{1}{4} + (2\cdot 2)\cdot 0 = 1.\end{aligned}$$

A natural question is whether this value can also be obtained from $\mathrm{E}[X]\,\mathrm{E}[Y]$. We return to this question later in this chapter. First we address the expectation of the sum of two random variables.

Table 10.1. Joint probabilities $\mathrm{P}(X = a, Y = b)$.

	a		
b	0	1	2
0	0	1/4	0
1	1/4	0	1/4
2	0	1/4	0

QUICK EXERCISE 10.1 Compute $\mathrm{E}[X+Y]$ for the random variables with the joint distribution given in Table 10.1.

For discrete X and Y with values $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, respectively, we see that

$$\begin{aligned}
\mathrm{E}[X+Y] &= \sum_i \sum_j (a_i + b_j)\mathrm{P}(X = a_i,\, Y = b_j) \\
&= \sum_i \sum_j a_i \mathrm{P}(X = a_i,\, Y = b_j) \,+\, \sum_i \sum_j b_j \mathrm{P}(X = a_i,\, Y = b_j) \\
&= \sum_i a_i \Big(\sum_j \mathrm{P}(X = a_i,\, Y = b_j) \Big) \\
&\qquad\qquad + \sum_j b_j \Big(\sum_i \mathrm{P}(X = a_i,\, Y = b_j) \Big) \\
&= \sum_i a_i \mathrm{P}(X = a_i) \,+\, \sum_j b_j \mathrm{P}(Y = b_j) \\
&= \mathrm{E}[X] + \mathrm{E}[Y]\,.
\end{aligned}$$

A similar line of reasoning applies in case X and Y are continuous random variables. The following general rule holds.

LINEARITY OF EXPECTATIONS. For all numbers r, s, and t and random variables X and Y, one has

$$\mathrm{E}[rX + sY + t] = r\mathrm{E}[X] + s\mathrm{E}[Y] + t.$$

QUICK EXERCISE 10.2 Determine the marginal distributions for the random variables X and Y with the joint distribution given in Table 10.1, and use them to compute $\mathrm{E}[X]$ en $\mathrm{E}[Y]$. Check that $\mathrm{E}[X]+\mathrm{E}[Y]$ is equal to $\mathrm{E}[X+Y]$, which was computed in Quick exercise 10.1.

More generally, for random variables $X_1, \ldots, X_n$ and numbers $s_1, \ldots, s_n$ and t,

$$\mathrm{E}[s_1X_1 + \cdots + s_nX_n + t] = s_1\mathrm{E}[X_1] + \cdots + s_n\mathrm{E}[X_n] + t.$$

This rule is a powerful instrument. For example, it provides an easy way to compute the expectation of a random variable X with a $Bin(n,p)$ distribution. If we would use the definition of expectation, we have to compute

$$\mathrm{E}[X] = \sum_{k=0}^{n} k\mathrm{P}(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

To determine this sum is not straightforward. However, there is a simple alternative. Recall the multiple-choice example from Section 4.3. We represented

the number of correct answers out of 10 multiple-choice questions as a sum of 10 Bernoulli random variables. More generally, any random variable X with a $Bin(n,p)$ distribution can be represented as

$$X = R_1 + R_2 + \cdots + R_n,$$

where $R_1, R_2, \ldots, R_n$ are independent $Ber(p)$ random variables, i.e.,

$$R_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

Since $\mathrm{E}[R_i] = 0 \cdot (1-p) + 1 \cdot p = p$, for every $i = 1, 2, \ldots, n$, the linearity-of-expectations rule yields

$$\mathrm{E}[X] = \mathrm{E}[R_1] + \mathrm{E}[R_2] + \cdots + \mathrm{E}[R_n] = np.$$

Hence we conclude that the expectation of a $Bin(n,p)$ distribution equals np.

Remark 10.1 (More than two random variables). In both the discrete and continuous cases, the change-of-variable formula for n random variables is a straightforward generalization of the change-of-variable formula for two random variables. For instance, if $X_1, X_2, \ldots, X_n$ are continuous random variables, with joint probability density function f, and g is a function from $\mathbb{R}^n$ to $\mathbb{R}$, then

$$\mathrm{E}[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

10.2 Covariance

In the previous section we have seen that for two random variables X and Y always

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

Does such a simple relation also hold for the variance of the sum $\mathrm{Var}(X+Y)$ or for expectation of the product $\mathrm{E}[XY]$? We will investigate this in the current section.

For the variables X and Y from the example in Section 9.2 with joint probability density

$$f(x,y) = \frac{2}{75}\left(2x^2y + xy^2\right) \quad \text{for } 0 \le x \le 3 \text{ and } 1 \le y \le 2,$$

one can show that

$$\mathrm{Var}(X+Y) = \frac{939}{2000} \quad \text{and} \quad \mathrm{Var}(X) + \mathrm{Var}(Y) = \frac{989}{2500} + \frac{791}{10\,000} = \frac{4747}{10\,000}$$

(see Exercise 10.10). This shows, in contrast to the linearity-of-expectations rule, that $\mathrm{Var}(X+Y)$ is generally *not equal* to $\mathrm{Var}(X)+\mathrm{Var}(Y)$. To determine $\mathrm{Var}(X+Y)$, we exploit its definition:

$$\mathrm{Var}(X+Y) = \mathrm{E}\left[(X+Y-\mathrm{E}[X+Y])^2\right].$$

Now $X+Y-\mathrm{E}[X+Y] = (X-\mathrm{E}[X])+(Y-\mathrm{E}[Y])$, so that

$$(X+Y-\mathrm{E}[X+Y])^2 = (X-\mathrm{E}[X])^2 + (Y-\mathrm{E}[Y])^2 + 2\,(X-\mathrm{E}[X])\,(Y-\mathrm{E}[Y]).$$

Taking expectations on both sides, another application of the linearity-of-expectations rule gives

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

That is, the variance of the sum $X+Y$ equals the sum of the variances of X and Y, plus an extra term $2\mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]$. To some extent this term expresses the way X and Y influence each other.

DEFINITION. Let X and Y be two random variables. The *covariance* between X and Y is defined by

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

Loosely speaking, if the covariance of X and Y is positive, then if X has a realization larger than $\mathrm{E}[X]$, it is likely that Y will have a realization larger than $\mathrm{E}[Y]$, and the other way around. In this case we say that X and Y are *positively correlated*. In case the covariance is negative, the opposite effect occurs; X and Y are *negatively correlated*. In case $\mathrm{Cov}(X,Y)=0$ we say that X and Y are *uncorrelated*. An easy consequence of the linearity-of-expectations property (see Exercise 10.19) is the following rule.

AN ALTERNATIVE EXPRESSION FOR THE COVARIANCE. Let X and Y be two random variables, then

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

For X and Y from the example in Section 9.2, we have $\mathrm{E}[X] = 109/50$, $\mathrm{E}[Y] = 157/100$, and $\mathrm{E}[XY] = 171/50$ (see Exercise 10.10). Thus we see that X and Y are negatively correlated:

$$\mathrm{Cov}(X,Y) = \frac{171}{50} - \frac{109}{50}\cdot\frac{157}{100} = -\frac{13}{5000} < 0.$$

Moreover, this also illustrates that, in contrast to the expectation of the sum, for the expectation of the product, in general $\mathrm{E}[XY]$ is *not equal* to $\mathrm{E}[X]\,\mathrm{E}[Y]$.

Independent versus uncorrelated

Now let X and Y be two *independent* random variables. One expects that X and Y are uncorrelated: they have nothing to do with one another! This is indeed the case, for instance, if X and Y are discrete; one finds that

$$\begin{aligned}\mathrm{E}[XY] &= \sum_i \sum_j a_i b_j \mathrm{P}(X = a_i,\, Y = b_j) \\ &= \sum_i \sum_j a_i b_j \mathrm{P}(X = a_i)\, \mathrm{P}(Y = b_j) \\ &= \Big(\sum_i a_i \mathrm{P}(X = a_i) \Big) \Big(\sum_j b_j \mathrm{P}(Y = b_j) \Big) \\ &= \mathrm{E}[X]\, \mathrm{E}[Y]\,.\end{aligned}$$

A similar reasoning holds in case X and Y are continuous random variables. The alternative expression for the covariance leads to the following important observation.

INDEPENDENT VERSUS UNCORRELATED. If two random variables X and Y are independent, then X and Y are uncorrelated.

Note that the reverse is not necessarily true. If X and Y are uncorrelated, they need *not* be independent. This is illustrated in the next quick exercise.

QUICK EXERCISE 10.3 Consider the random variables X and Y with the joint distribution given in Table 10.1. Check that X and Y are dependent, but that also $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$.

From the preceding we also deduce the following rule on the variance of the sum of two random variables.

VARIANCE OF THE SUM. Let X and Y be two random variables. Then always

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)\,.$$

If X and Y are *uncorrelated*,

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

Hence, we always have that $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$, whereas $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ *only* holds for uncorrelated random variables (and hence for independent random variables!).

As with the linearity-of-expectations rule, the rule for the variance of the sum of uncorrelated random variables holds more generally. For uncorrelated random variables $X_1, X_2, \ldots, X_n$, we have

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\,.$$

This rule provides an easy way to compute the variance of a random variable with a $Bin(n,p)$ distribution. Recall the representation for a $Bin(n,p)$ random variable X:

$$X = R_1 + R_2 + \cdots + R_n.$$

Each R_i has variance

$$\begin{aligned}\mathrm{Var}(R_i) &= \mathrm{E}\left[R_i^2\right] - \left(\mathrm{E}\left[R_i\right]\right)^2 = 0^2 \cdot (1-p) + 1^2 \cdot p - \left(\mathrm{E}\left[R_i\right]\right)^2 \\ &= p - p^2 = p(1-p).\end{aligned}$$

Using the independence of the R_i, the rule for the variance of the sum yields

$$\mathrm{Var}(X) = \mathrm{Var}(R_1) + \mathrm{Var}(R_2) + \cdots + \mathrm{Var}(R_n) = np(1-p).$$

10.3 The correlation coefficient

In the previous section we saw that the covariance between random variables gives an indication of how they influence one another. A disadvantage of the covariance is the fact that it depends on the units in which the random variables are represented. For instance, suppose that the length in inches and weight in kilograms of Dutch citizens are modeled by random variables L and W. Someone prefers to represent the length in centimeters. Since 1 inch $\equiv$ 2.53 cm, one is dealing with a transformed random variable $2.53L$. The covariance between $2.53L$ and W is

$$\begin{aligned}\mathrm{Cov}(2.53L, W) &= \mathrm{E}\left[(2.53L)W\right] - \mathrm{E}\left[2.53L\right]\mathrm{E}\left[W\right] \\ &= 2.53\Big(\mathrm{E}\left[LW\right] - \mathrm{E}\left[L\right]\mathrm{E}\left[W\right]\Big) = 2.53\,\mathrm{Cov}(L,W)\,.\end{aligned}$$

That is, the covariance increases with a factor 2.53, which is somewhat disturbing since changing from inches to centimeters does not essentially alter the dependence between length and weight. This illustrates that the covariance changes under a change of units. The following rule provides the exact relationship.

COVARIANCE UNDER CHANGE OF UNITS. Let X and Y be two random variables. Then

$$\mathrm{Cov}(rX + s, tY + u) = rt\,\mathrm{Cov}(X,Y)$$

for all numbers r, s, t, and u.

See Exercise 10.14 for a derivation of this rule.

QUICK EXERCISE 10.4 For X and Y in the example in Section 9.2 (see also Section 10.2), show that $\mathrm{Cov}(-2X+7, 5Y-3) = 13/500$.

The preceding discussion indicates that the covariance $\mathrm{Cov}(X,Y)$ may not always be suitable to express the dependence between X and Y. For this reason there is a standardized version of the covariance called the correlation coefficient of X and Y.

DEFINITION. Let X and Y be two random variables. The *correlation coefficient* $\rho(X,Y)$ is defined to be 0 if $\mathrm{Var}(X) = 0$ or $\mathrm{Var}(Y) = 0$, and otherwise

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

Note that $\rho(X,Y)$ remains unaffected by a change of units, and therefore it is *dimensionless*. For instance, if X and Y are measured in kilometers, then $\mathrm{Cov}(X,Y)$, $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are in km^2, so that the dimension of $\rho(X,Y)$ is in $\mathrm{km}^2/(\sqrt{\mathrm{km}^2}\cdot\sqrt{\mathrm{km}^2})$.

For X and Y in the example in Section 9.2, recall that $\mathrm{Cov}(X,Y) = -13/5000$. We also have $\mathrm{Var}(X) = 989/2500$ and $\mathrm{Var}(Y) = 791/10\,000$ (see Exercise 10.10), so that

$$\rho(X,Y) = \frac{-\frac{13}{5000}}{\sqrt{\frac{989}{2500}\cdot\frac{791}{10\,000}}} = -0.0147.$$

QUICK EXERCISE 10.5 For X and Y in the example in Section 9.2, show that $\rho(-2X+7, 5Y-3) = 0.0147$.

The previous quick exercise illustrates the following linearity property for the correlation coefficient. For numbers r, s, t, and u fixed, $r, t \neq 0$, and random variables X and Y:

$$\rho(rX+s, tY+u) = \begin{cases} -\rho(X,Y) & \text{if } rt < 0, \\ \rho(X,Y) & \text{if } rt > 0. \end{cases}$$

Thus we see that the size of the correlation coefficient is unaffected by a change of units, but note the possibility of a change of sign.

Two random variables X and Y are "most correlated" if $X = Y$ or if $X = -Y$. As a matter of fact, in the former case $\rho(X,Y) = 1$, while in the latter case $\rho(X,Y) = -1$. In general—for nonconstant random variables X and Y—the following property holds:

$$-1 \leq \rho(X,Y) \leq 1.$$

For a formal derivation of this property, see the next remark.

Remark 10.2 (Correlations are between −1 and 1). Here we give a proof of the preceding formula. Since the variance of any random variable is nonnegative, we have that

$$\begin{aligned}
0 &\le \mathrm{Var}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}} + \frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right) \\
&= \mathrm{Var}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}}\right) + \mathrm{Var}\left(\frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right) \\
&\qquad + 2\mathrm{Cov}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}}, \frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right) \\
&= \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)} + \frac{\mathrm{Var}(Y)}{\mathrm{Var}(Y)} + \frac{2\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 2\left(1 + \rho(X,Y)\right).
\end{aligned}$$

This implies $\rho(X,Y) \ge -1$. Using the same argument but replacing X by $-X$ shows that $\rho(X,Y) \le 1$.

10.4 Solutions to the quick exercises

10.1 The expectation of $X+Y$ is computed as follows:

$$\begin{aligned}
\mathrm{E}[X+Y] &= (0+0)\cdot 0 + (1+0)\cdot\frac{1}{4} + (2+0)\cdot 0 \\
&\quad + (0+1)\cdot\frac{1}{4} + (1+1)\cdot 0 + (2+1)\cdot\frac{1}{4} \\
&\quad + (0+2)\cdot 0 + (1+2)\cdot\frac{1}{4} + (2+2)\cdot 0 = 2.
\end{aligned}$$

10.2 First complete Table 10.1 with the marginal distributions:

	a			
b	0	1	2	$\mathrm{P}(Y=b)$
0	0	1/4	0	1/4
1	1/4	0	1/4	1/2
2	0	1/4	0	1/4
$\mathrm{P}(X=a)$	1/4	1/2	1/4	1

It follows that $\mathrm{E}[X] = 0\cdot\frac{1}{4} + 1\cdot\frac{1}{2} + 2\cdot\frac{1}{4} = 1$, and similarly $\mathrm{E}[Y] = 1$. Therefore $\mathrm{E}[X] + \mathrm{E}[Y] = 2$, which is equal to $\mathrm{E}[X+Y]$ as computed in Quick exercise 10.1.

10.3 From Table 10.1, as completed in Quick exercise 10.2, we see that X and Y are dependent. For instance, $\mathrm{P}(X=0, Y=0) \neq \mathrm{P}(X=0)\,\mathrm{P}(Y=0)$. From Quick exercise 10.2 we know that $\mathrm{E}[X] = \mathrm{E}[Y] = 1$. Because we already computed $\mathrm{E}[XY] = 1$, it follows that $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$. According to the alternative expression for the covariance this means that $\mathrm{Cov}(X, Y) = 0$, i.e., X and Y are uncorrelated.

10.4 We already computed $\mathrm{Cov}(X, Y) = -13/5000$ in Section 10.2. Hence, by the linearity-of-covariance rule $\mathrm{Cov}(-2X + 7, 5Y - 3) = (-2)\cdot 5\cdot(-13/5000) = 13/500$.

10.5 From Quick exercise 10.4 we have $\mathrm{Cov}(-2X + 7, 5Y - 3) = 13/500$. Since $\mathrm{Var}(X) = 989/2500$ and $\mathrm{Var}(Y) = 791/10\,000$, by definition of the correlation coefficient and the rule for variances,

$$\begin{aligned}\rho(-2X+7, 5Y-3) &= \frac{\mathrm{Cov}(-2X+7, 5Y-3)}{\sqrt{\mathrm{Var}(-2X+7)\cdot\mathrm{Var}(5Y-3)}} \\ &= \frac{\frac{13}{500}}{\sqrt{4\mathrm{Var}(X)\cdot 25\mathrm{Var}(Y)}} = \frac{\frac{13}{500}}{\sqrt{\frac{3956}{2500}\cdot\frac{19775}{10\,000}}} = 0.0147.\end{aligned}$$

10.5 Exercises

10.1 ⊡ Consider the joint probability distribution of X and Y from Exercise 9.7, obtained from data on hair color and eye color, for which we already computed the expectations and variances of X and Y, as well as $\mathrm{E}[XY]$.

a. Compute $\mathrm{Cov}(X, Y)$. Are X and Y positively correlated, negative correlated, or uncorrelated?

b. Compute the correlation coefficient between X and Y.

10.2 ⊡ Consider the two discrete random variables X and Y with joint distribution derived in Exercise 9.2:

	a			
b	0	1	2	$\mathrm{P}(Y=b)$
-1	1/6	1/6	1/6	1/2
1	0	1/2	0	1/2
$\mathrm{P}(X=a)$	1/6	2/3	1/6	1

a. Determine $\mathrm{E}[XY]$.

b. Note that X and Y are dependent. Show that X and Y are uncorrelated.

c. Determine $\mathrm{Var}(X+Y)$.

d. Determine $\mathrm{Var}(X-Y)$.

10.3 Let U and V be the two random variables from Exercise 9.6. We have seen that U and V are dependent with joint probability distribution

	a			
b	0	1	2	$\mathrm{P}(V=b)$
0	1/4	0	1/4	1/2
1	0	1/2	0	1/2
$\mathrm{P}(U=a)$	1/4	1/2	1/4	1

Determine the covariance $\mathrm{Cov}(U,V)$ and the correlation coefficient $\rho(U,V)$.

10.4 Consider the joint probability distribution of the discrete random variables X and Y from the *Melencolia* Exercise 9.1. Compute $\mathrm{Cov}(X,Y)$.

	a			
b	1	2	3	4
1	16/136	3/136	2/136	13/136
2	5/136	10/136	11/136	8/136
3	9/136	6/136	7/136	12/136
4	4/136	15/136	14/136	1/136

10.5 ⊡ Suppose X and Y are discrete random variables taking values 0,1, and 2. The following is given about the joint and marginal distributions:

	a			
b	0	1	2	$\mathrm{P}(Y=b)$
0	8/72	...	10/72	1/3
1	12/72	9/72	...	1/2
2	...	3/72	...	...
$\mathrm{P}(X=a)$	1/3	...	...	1

a. Complete the table.

b. Compute the expectation of X and of Y and the covariance between X and Y.

c. Are X and Y independent?

10.6 ⊞ Suppose X and Y are discrete random variables taking values $c-1$, c, and $c+1$. The following is given about the joint and marginal distributions:

	a			
b	$c-1$	c	$c+1$	$\mathrm{P}(Y=b)$
$c-1$	$2/45$	$9/45$	$4/45$	$1/3$
c	$7/45$	$5/45$	$3/45$	$1/3$
$c+1$	$6/45$	$1/45$	$8/45$	$1/3$
$\mathrm{P}(X=a)$	$1/3$	$1/3$	$1/3$	1

a. Take $c=0$ and compute the expectation of X and of Y and the covariance between X and Y.

b. Show that X and Y are uncorrelated, no matter what the value of c is. *Hint:* one could compute $\mathrm{Cov}(X,Y)$, but there is a short solution using the rule on the covariance under change of units (see page 141) together with part **a**.

c. Are X and Y independent?

10.7 ⊡ Consider the joint distribution of Quick exercise 9.2 and take ε fixed between $-1/4$ and $1/4$:

	b		
a	0	1	$p_X(a)$
0	$1/4-\varepsilon$	$1/4+\varepsilon$	$1/2$
1	$1/4+\varepsilon$	$1/4-\varepsilon$	$1/2$
$p_Y(b)$	$1/2$	$1/2$	1

a. Take $\varepsilon=1/8$ and compute $\mathrm{Cov}(X,Y)$.

b. Take $\varepsilon=1/8$ and compute $\rho(X,Y)$.

c. For which values of ε is $\rho(X,Y)$ equal to -1, 0, or 1?

10.8 Let X and Y be random variables such that

$$\mathrm{E}[X]=2, \quad \mathrm{E}[Y]=3, \quad \text{and} \quad \mathrm{Var}(X)=4.$$

a. Show that $\mathrm{E}\left[X^2\right]=8$.

b. Determine the expectation of $-2X^2+Y$.

10.9 ⊞ Suppose the blood of 1000 persons has to be tested to see which ones are infected by a (rare) disease. Suppose that the probability that the test

is positive is $p = 0.001$. The obvious way to proceed is to test each person, which results in a total of 1000 tests. An alternative procedure is the following. Distribute the blood of the 1000 persons over 25 groups of size 40, and mix half of the blood of each of the 40 persons with that of the others in each group. Now test the aggregated blood sample of each group: when the test is negative *no one* in that group has the disease; when the test is positive, at least one person in the group has the disease, and one will test the other half of the blood of all 40 persons of that group separately. In total, that gives 41 tests for that group. Let X_i be the total number of tests one has to perform for the ith group using this alternative procedure.

a. Describe the probability distribution of X_i, i.e., list the possible values it takes on and the corresponding probabilities.

b. What is the expected number of tests for the ith group? What is the expected total number of tests? What do you think of this alternative procedure for blood testing?

10.10 ⊞ Consider the variables X and Y from the example in Section 9.2 with joint probability density

$$f(x,y) = \frac{2}{75}\big(2x^2y + xy^2\big) \quad \text{for } 0 \le x \le 3 \text{ and } 1 \le y \le 2$$

and marginal probability densities

$$f_X(x) = \frac{2}{225}\big(9x^2 + 7x\big) \quad \text{for } 0 \le x \le 3$$
$$f_Y(y) = \frac{1}{25}(3y^2 + 12y) \quad \text{for } 1 \le y \le 2.$$

a. Compute $\mathrm{E}[X]$, $\mathrm{E}[Y]$, and $\mathrm{E}[X+Y]$.

b. Compute $\mathrm{E}\big[X^2\big]$, $\mathrm{E}\big[Y^2\big]$, $\mathrm{E}[XY]$, and $\mathrm{E}\big[(X+Y)^2\big]$,

c. Compute $\mathrm{Var}(X+Y)$, $\mathrm{Var}(X)$, and $\mathrm{Var}(Y)$ and check that $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$.

10.11 Recall the relation between degrees Celsius and degrees Fahrenheit

$$\text{degrees Fahrenheit} = \frac{9}{5} \cdot \text{degrees Celsius} + 32.$$

Let X and Y be the average daily temperatures in degrees Celsius in Amsterdam and Antwerp. Suppose that $\mathrm{Cov}(X,Y) = 3$ and $\rho(X,Y) = 0.8$. Let T and S be the same temperatures in degrees Fahrenheit. Compute $\mathrm{Cov}(T,S)$ and $\rho(T,S)$.

10.12 Consider the independent random variables H and R from the vase example, with a $U(25,35)$ and a $U(7.5,12.5)$ distribution. Compute $\mathrm{E}[H]$ and $\mathrm{E}\big[R^2\big]$ and check that $\mathrm{E}[V] = \pi \mathrm{E}[H]\,\mathrm{E}\big[R^2\big]$.

10.13 Let X and Y be as in the triangle example in Exercise 9.15. Recall from Exercise 9.16 that X and Y represent the minimum and maximum coordinate of a point that is drawn from the unit square: $X = \min\{U, V\}$ and $Y = \max\{U, V\}$.

a. Show that $\mathrm{E}[X] = 1/3$, $\mathrm{Var}(X) = 1/18$, $\mathrm{E}[Y] = 2/3$, and $\mathrm{Var}(Y) = 1/18$. *Hint:* you might consult Exercise 8.15.

b. Check that $\mathrm{Var}(X + Y) = 1/6$, by using that U and V are independent and that $X + Y = U + V$.

c. Determine the covariance $\mathrm{Cov}(X, Y)$ using the results from **a** and **b**.

10.14 ⊞ Let X and Y be two random variables and let r, s, t, and u be arbitrary real numbers.

a. Derive from the definition that $\mathrm{Cov}(X + s, Y + u) = \mathrm{Cov}(X, Y)$.

b. Derive from the definition that $\mathrm{Cov}(rX, tY) = rt\mathrm{Cov}(X, Y)$.

c. Combine parts **a** and **b** to show $\mathrm{Cov}(rX + s, tY + u) = rt\mathrm{Cov}(X, Y)$.

10.15 In Figure 10.1 three plots are displayed. For each plot we carried out a simulation in which we generated 500 realizations of a pair of random variables (X, Y). We have chosen three different joint distributions of X and Y.

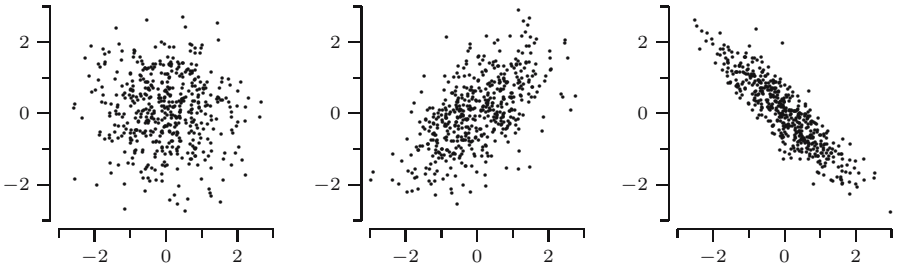

Fig. 10.1. Some scatterplots.

a. Indicate for each plot whether it corresponds to random variables X and Y that are positively correlated, negatively correlated, or uncorrelated.

b. Which plot corresponds to random variables X and Y for which $|\rho(X, Y)|$ is maximal?

10.16 ⊡ Let X and Y be random variables.

a. Express $\mathrm{Cov}(X, X + Y)$ in terms of $\mathrm{Var}(X)$ and $\mathrm{Cov}(X, Y)$.

b. Are X and $X + Y$ positively correlated, uncorrelated, or negatively correlated, or can anything happen?

c. Same question as in part **b**, but now assume that X and Y are uncorrelated.

10.17 Extending the variance of the sum rule. For mathematical convenience we first extend the sum rule to three random variables with zero expectation. Next we further extend the rule to three random variables with nonzero expectation. By the same line of reasoning we extend the rule to n random variables.

a. Let X, Y and Z be random variables with expectation 0. Show that

$$\begin{aligned}\mathrm{Var}(X+Y+Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Var}(Z) \\ + 2\mathrm{Cov}(X,Y) + 2\mathrm{Cov}(X,Z) + 2\mathrm{Cov}(Y,Z)\,.\end{aligned}$$

Hint: directly apply that for real numbers $y_1, \ldots, y_n$

$$(y_1 + \cdots + y_n)^2 = y_1^2 + \cdots + y_n^2 + 2y_1y_2 + 2y_1y_3 + \cdots + 2y_{n-1}y_n.$$

b. Now show **a** for X, Y, and Z with nonzero expectation.
Hint: you might use the rules on pages 98 and 141 about variance and covariance under a change of units.

c. Derive a general variance of the sum rule, i.e., show that if $X_1, X_2, \ldots, X_n$ are random variables, then

$$\begin{aligned}&\mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &\quad = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) \\ &\qquad + 2\mathrm{Cov}(X_1,X_2) + 2\mathrm{Cov}(X_1,X_3) + \cdots + 2\mathrm{Cov}(X_1,X_n) \\ &\qquad\qquad + 2\mathrm{Cov}(X_2,X_3) + \cdots + 2\mathrm{Cov}(X_2,X_n) \\ &\qquad\qquad\qquad \ddots \\ &\qquad\qquad\qquad + 2\mathrm{Cov}(X_{n-1},X_n)\,.\end{aligned}$$

d. Show that if the variances are all equal to σ^2 and the covariances are all equal to some constant γ, then

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = n\sigma^2 + n(n-1)\gamma.$$

10.18 ⊞ Consider a vase containing balls numbered $1, 2, \ldots, N$. We draw n balls *without replacement* from the vase. Each ball is selected with equal probability, i.e., in the first draw each ball has probability $1/N$, in the second draw each of the $N-1$ remaining balls has probability $1/(N-1)$, and so on. For $i = 1, 2, \ldots, n$, let X_i denote the number on the ball in the ith draw. From Exercise 9.18 we know that the variance of X_i equals

$$\mathrm{Var}(X_i) = \frac{1}{12}(N-1)(N+1).$$

Show that

$$\mathrm{Cov}(X_1, X_2) = -\frac{1}{12}(N+1).$$

Before you do the exercise: why do you think the covariance is negative?
Hint: use $\mathrm{Var}(X_1 + X_2 + \cdots + X_N) = 0$ (why?), and apply Exercise 10.17.

10.19 Derive the alternative expression for the covariance: $\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$.
Hint: work out $(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])$ and use linearity of expectations.

10.20 Determine $\rho\left(U, U^2\right)$ when U has a $U(0, a)$ distribution. Here a is a positive number.