Lecture 9: Transformations; Joint Distributions

Relevant textbook passages:

Pitman [5]: Chapter 5; Section 6.4–6.5

Larsen–Marx [4]: Sections 3.7, 3.8, 3.9, 3.11

9.1 Density of a function of a random variable; aka change of variable

Pitman [5]: Section 4.4, pp. 302–309

If X is a random variable with cumulative distribution function F_X and density $f_X = F'_X$, and g is a (Borel) function, then

$$Y = g(X)$$

is a random variable. The cumulative distribution function F_Y of Y is given by

$$F_Y(y) = P(Y \leqslant y) = P(g(X) \leqslant y).$$

The density f_Y is then given by

$$f_Y(y) = \frac{d}{dy} P(g(X) \leqslant y),$$

provided the derivative exists.

There are two special cases of particular interest. If $g'(x) > 0$ for all x, so that g is strictly increasing, then g has an inverse, and

$$g(X) \leqslant y \iff X \leqslant g^{-1}(y),$$

so

$$F_Y(y) = F_X\big(g^{-1}(y)\big)$$

is the cumulative distribution function of Y. The density f_Y of Y is found by differentiating this.

Similarly, if $g'(x) < 0$ for all x, then g is strictly decreasing, and

$$Y \leqslant y \iff g(X) \leqslant y \iff X \geqslant g^{-1}(y),$$

and if F is continuous, this is just

$$F_Y(y) = 1 - F_X\big(g^{-1}(y)\big),$$

and we may differentiate that (with respect to y, to get the density.)

Start with the case $g' > 0$. By the Inverse Function Theorem [2, Theorem 6.7, p. 252], then

$$\frac{d}{dy} g^{-1}(y) = \frac{1}{g'\big(g^{-1}(y)\big)}$$

So in this case we have

$$f_Y(y) = \frac{d}{dy}F_X\big(g^{-1}(y)\big) = F'_X\big(g^{-1}(y)\big)\frac{d}{dy}g^{-1}(y) = \frac{f_X\big(g^{-1}(y)\big)}{g'\big(g^{-1}(y)\big)}.$$

Or letting $y = g(x)$,

$$f_Y\big(g(x)\big) = f_X(x)/g'(x).$$

When $g' < 0$, then g is decreasing and we must differentiate $1 - F_g$ to find the density of Y. In this case

$$f_Y\big(g(x)\big) = -f_X(x)/g'(x).$$

So to sum up:

9.1.1 Theorem (Density of a monotone function of X) *Let X is a random variable with cumulative distribution function F_X and density $f_X = F'_X$ and let*

$$Y = g(X).$$

If g is everywhere differentiable and either $g'(x) > 0$ for all x in the range of X, or $g'(x) < 0$ for all x in the range of X, then Y has a density f_Y given by

$$f_Y(y) = \frac{f_X\big(g^{-1}(y)\big)}{\big|g'\big(g^{-1}(y)\big)\big|}$$

9.1.2 Example (Change of scale and location) We say that the random variable Y is a change of **location and scale** of the random variable X if

$$Y = aX + b, \qquad \text{or equivalently,} \qquad X = \frac{Y-b}{a},$$

where $a > 0$. If $X \sim F_X$ with density f_X, then

$$F_Y(y) = P(Y \leqslant y) = P(aX + b \leqslant y) = P\left(X \leqslant \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right),$$

so

$$f_Y(y) = \frac{1}{a}f_X\left(\frac{y-b}{a}\right).$$

□

9.1.3 Example Let $X \sim U(0,1)$, and let $Y = 2X - 1$. Then by Example 9.1.2 we have

$$f_X(x) = \begin{cases} 1, & 0 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{so} \quad f_Y(y) = \begin{cases} 1/2, & -1 \leqslant y \leqslant 1 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, $Y \sim U[-1,1]$. □

9.1.4 Example Let Z have the standard normal density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2},$$

and let $Y = \sigma Z + \mu$. By Example 9.1.2 we have

$$f_Y(y) = f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)},$$

□

9.1.5 Example Let $X \sim U[0,1]$, (so $f(x) = 1$ for $0 \leqslant x \leqslant 1$). Let

$$g(x) = x^a,$$

so

$$g'(x) = ax^{a-1}.$$

Now

$$g^{-1}(y) = y^{1/a},$$

and as x ranges over the unit interval, so does y. To apply the inverse function theorem use

$$\frac{d}{dy}g^{-1}(y) = \frac{1}{g'\big(g^{-1}(y)\big)} = \frac{1}{a\big(g^{-1}(y)\big)^{a-1}} = \frac{1}{a\big(y^{1/a}\big)^{a-1}}.$$

So the density of $g(X) = X^a$ is given by $f\big(g^{-1}(y)\big)/\frac{d}{dy}g^{-1}(y)$ or

$$f_g(y) = \begin{cases} \frac{1}{a}y^{(a-1)/a} & (0 \leqslant y \leqslant 1) \\ 0 & \text{otherwise.} \end{cases}$$

□

Even if g is not strictly increasing or decreasing, if we can find a nice expression for F_Y, we may still be able to find the density of Y.

9.1.6 Example (A non-monotonic transformation) Let X have a Uniform$[-1,1]$ distribution, and $Y = X^2$. Then $0 \leqslant Y \leqslant 1$ and

$$F_Y(y) = P\big(X^2 \leqslant y\big) = P(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) = \sqrt{y} \qquad (0 \leqslant y \leqslant 1),$$

and $F_Y(y) = 0$ for $y \leqslant 0$ and $F_Y(y) = 1$ for $y \geqslant 1$. Then

$$f_Y(y) = \frac{d}{dy}\sqrt{y} = \frac{1}{2\sqrt{y}} \qquad (0 \leqslant y \leqslant 1),$$

and $f_Y(y) = 0$ for $y \leqslant 0$ or $y \geqslant 1$. □

9.2 Random vectors and joint distributions

Recall that a **random variable** X is a real-valued function on the sample space $(\Omega, \mathcal{F}, P)$, where P is a probability measure on Ω; and that it induces a probability measure P_X on $\boldsymbol{R}$, called the **distribution** of X, given by

$$P_X(I) = P(X \in I) = P\big(\{\omega \in \Omega : X(\omega) \in I\}\big),$$

for every interval in $\boldsymbol{R}$. The distribution is enough to calculate the expectation of any (Borel) function of X.

Now suppose I have more than one random variable on the same sample space. Then I can consider the **random vector** (X, Y) or $\boldsymbol{X} = (X_1, \ldots, X_n)$.

- A random vector $\boldsymbol{X}$ defines a probability $P_{\boldsymbol{X}}$ on $\boldsymbol{R}^{\mathrm{n}}$, called the distribution of $\boldsymbol{X}$ via:

$$P_{\boldsymbol{X}}(I_1 \times \cdots \times I_n) = P(\boldsymbol{X} \in I_1 \times \cdots \times I_n) = P\left\{\omega \in \Omega : X_1(\omega) \in I_1, \ldots, X_n(\omega) \in I_n\right\},$$

where each I_i is an interval. This distribution is also called the **joint distribution** of $X_1, \ldots, X_n$.

- We can use this to define a **joint cumulative distribution function**, denoted $F_{\boldsymbol{X}}$, by

$$F_{\boldsymbol{X}}(x_1, \dots, x_n) = P(X_i \leqslant x_i, \text{ for all } i = 1, \dots, n)$$

- Given a joint cumulative distribution function we can recover the joint distribution. For instance, Suppose (X, Y) has joint cumulative distribution function F. The probability of the rectangle $(a_1, b_1] \times (a_2, b_2]$ is given by

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, b_1).$$

To see this, consult Figure 9.1. The rectangle is the event

$$((X, Y) \leqq (b_1, b_2)) \setminus \big[((X, Y) \leqq (b_1, b_2)) \cup ((X, Y) \leqq (b_1, b_2))\big].$$

The probability is computed by using the inclusion/exclusion principle to compute the probability of the union and subtracting it from $P(X, Y) \leqq (b_1, b_2))$. There are higher dimensional

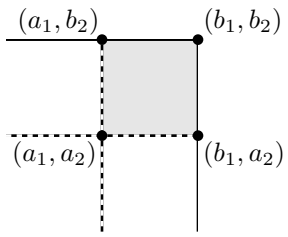

Figure 9.1. Joint cumulative distribution function and rectangles.

versions, but the expressions are complicated (see, e.g., [1, pp. 394–395]). In this class we shall mostly deal with joint densities.

- **N.B.** While the subscript $_{X,Y}$ or $_{\boldsymbol{X}}$ is often used to identify a joint distribution or cumulative distribution function, it is also frequently omitted. You are supposed to figure out the domain of the function by inspecting its arguments.

9.2.1 Example Let $S = \{SS, SF, FS, FF\}$ and let P be the probability measure on S defined by

$$P(SS) = \frac{7}{12},\ P(SF) = \frac{3}{12},\ P(FS) = \frac{1}{12},\ P(FF) = \frac{1}{12}.$$

Define the random variables X and Y by

$$X(SS) = 1,\ X(SF) = 1,\ X(FS) = 0,\ X(FF) = 0,$$
$$Y(SS) = 1,\ Y(SF) = 0,\ Y(FS) = 1,\ Y(FF) = 0.$$

That is, X and Y indicate Success or Failure on two different experiments, but the experiments are not necessarily independent.

Then

$$P_X(1) = \frac{10}{12},\ P_X(0) = \frac{2}{12},$$
$$P_Y(1) = \frac{8}{12},\ P_Y(0) = \frac{4}{12},$$
$$P_{X,Y}(1,1) = \frac{7}{12},\ P_{X,Y}(1,0) = \frac{3}{12},\ P_{X,Y}(0,1) = \frac{1}{12},\ P_{X,Y}(0,0) = \frac{1}{12}.$$

□

9.3 Joint PMFs

A random vector $\boldsymbol{X}$ on a probability space $(\Omega, \mathcal{F}, P)$ is **discrete**, if you can enumerate its range.

Pitman [5]: Section 3.1; also p. 348
Larsen–Marx [4]: Section 3.7

When $X_1, \dots, X_n$ are discrete, the **joint probability mass function** of the random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ is usually denoted $p_{\boldsymbol{X}}$, and is given by

$$p_{\boldsymbol{X}}(x_1, x_2, \dots, x_n) = P(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } \cdots \text{ and } X_n = x_n).$$

If X and Y are independent random variables, then $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.
For a function g of X and Y we have

$$\boldsymbol{E}\, g(X, Y) = \sum_x \sum_y g(x, y) p_{X,Y}(x, y).$$

9.4 Joint densities

Pitman [5]: Chapter 5.1
Larsen–Marx [4]: Section 3.7

Let X and Y be random variables on a probability space $(\Omega, \mathcal{F}, P)$. The random vector (X, Y) has a **joint density** $f_{X,Y}(x, y)$ if for every rectangle $I_1 \times I_2 \subset \boldsymbol{R}^2$,

$$P((X, Y) \in I_1 \times I_2) = \int_{I_1} \int_{I_1} f_{X,Y}(x, y)\, dx\, dy.$$

If X and Y are **independent**, then $f_{X,Y}(x, y) = f_X(x) f_Y(y)$.
For example,

$$P(X \geqslant Y) = \int_{-\infty}^{\infty} \int_{y}^{\infty} f_{X,Y}(x, y)\, dx\, dy.$$

For a function g of X and Y we have

$$\boldsymbol{E}\, g(X, Y) = \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

Again, the subscript $_{X,Y}$ or $_{\boldsymbol{X}}$ on a density is frequently omitted.

9.4.1 Theorem *If the joint cumulative distribution function $F\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}$ is differentiable, then the joint density is given by*

$$f(x_1, \dots, x_n) = \frac{\partial^n F(x_1, \dots, x_n)}{\partial x_1 \dots, \partial x_n}.$$

9.5 Recovering marginal distributions from joint distributions

So now we have random variables X and Y, and the random vector (X, Y). They have distributions P_X, P_Y, and $P_{X,Y}$. How are they related? The **marginal distribution** of X is just the distribution P_X of X alone. We can recover its probability mass function from the joint probability mass function $p_{X,Y}$ as follows.

In the discrete case:

$$p_X(x) = P(X = x) = \sum_y p_{X,Y}(x, y)$$

Likewise

$$p_Y(y) = P(Y = y) = \sum_y p_{X,Y}(x, y)$$

If X and Y are independent random variables, then $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

Larsen–Marx [4]: p. 169

For the density case, the **marginal density** of X, denoted f_X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy,$$

and the marginal density f_Y of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

The recovery of a marginal density of X from a joint density of X and Y is sometimes described as **"integrating out"** out y.

We just showed that if we know the joint distribution of two random variables, we can recover their marginal distributions. Can we go the other way? That is, if I have the marginal distributions of random variables X and Y, can I uniquely recover the joint distribution of X and Y? The answer is No, as the this simple example shows:

9.5.1 Example Suppose X and Y are both Bernoulli random variables with

$$P(X = 1) = P(Y = 1) = 0.5.$$

Here are three different joint distributions that give rise to these marginals:

$Y=1$	0.25	0.25
$Y=0$	0.25	0.25
	$X=0$	$X=1$

,

$Y=1$	0.20	0.30
$Y=0$	0.30	0.20
	$X=0$	$X=1$

$Y=1$	0.50	0.00
$Y=0$	0.00	0.50
	$X=0$	$X=1$

You can see that there are plenty of other joint distributions that work. □

9.6 The expectation of a sum

I already asserted that the expectation of a sum of random variables was the sum of their expectations, and proved it for the case of discrete random variables. If the random vector (X, Y) has a joint density, then it is straightforward to show that $\boldsymbol{E}(X + Y) = \boldsymbol{E}\,X + \boldsymbol{E}\,Y$. Since $x + y$ is a (Borel) function of the vector (x, y), we have (Section 9.4) that

$$\begin{aligned}
\boldsymbol{E}(X+Y) &= \iint (x+y) f_{X,Y}(x,y)\,dx\,dy \\
&= \iint x f_{X,Y}(x,y)\,dy\,dx + \iint y f_{X,Y}(x,y)\,dx\,dy \\
&= \int x \left(\int f_{X,Y}(x,y)\,dy \right) dx + \int y \left(\int f_{X,Y}(x,y)\,dx \right) dy \\
&= \int x f_X(x)\,dx + \int y f_Y(y)\,dy \\
&= \boldsymbol{E}\,X + \boldsymbol{E}\,Y.
\end{aligned}$$

9.7 The distribution of a sum

We already know to calculate the expectation of a sum of random variables—since **expectation is a positive linear operator**, the expectation of a sum is the sum of the expectations.

We are now in a position to describe the *distribution* of the sum of two random variables. Let $Z = X + Y$.

Pitman [5]: p. 147

Larsen–Marx [4]: p. 178ff.

Discrete case:

$$P(Z = z) = \sum_{(x,y):x+y=z} P(x, y) = \sum_{\text{all } x} p_{X,Y}(x, z - x)$$

9.7.1 Density of a sum

If (X, Y) has joint density $f_{X,Y}(x, y)$, what is the density of $X + Y$?

To find the density of a sum, we first find its cumulative distribution function. Now $X+Y \leqslant t$ if and only if $X \leqslant t - Y$, so

$$P(X + Y \leqslant t) = \iint_{\{(x,y):x\leqslant t-y\}} f_{X,Y}(x, y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{t-y} f_{X,Y}(x, y)\,dx\,dy.$$

I have written the limits of integration as $-\infty$ and ∞, but the density may well be zero in much of this region, so it helps to pay attention the nonzero region.

9.7.1 Example Let X and Y be independent Uniform$[0, 1]$ random variables. Their joint density is 1 on the square $[0, 1] \times [0, 1]$. The probability that their sum is $\leqslant t$ is just the area of the square lying below the line $x + y = t$. For $0 \leqslant t \leqslant 1$, this is a triangle with area t^2. For $1 \leqslant t \leqslant 2$, the region is more complicated, but by symmetry it is easy to see it's area is $1 - (1 - t)^2$. So

$$F_{X+Y}(t) = \begin{cases} 0, & t \leqslant 0 \\ t^2/2, & 0 \leqslant t \leqslant 1 \\ 1 - (2 - t)^2/2, & 1 \leqslant 2 \\ 1 & t \geqslant 2. \end{cases}$$

□

Recalling that the density is the derivative of the cdf, so to find the density we need only differentiate the cumulative distribution function.

9.7.2 Example (continued) The derivative of F_{X+Y} for the example above is

$$\frac{d}{dt}F_{X+Y}(t) = \begin{cases} 0, & t \leqslant 0 \text{ or } t \geqslant 2 \\ t, & 0 \leqslant t \leqslant 1 \\ 2 - t & 1 \leqslant t \leqslant 2. \end{cases}$$

□

More generally the derivative of the cumulative distribution function is given by

Pitman [5]: pp. 372–373

$$\begin{aligned} f_{X+Y}(t) = \frac{d}{dt}P(X + Y \leqslant t) &= \frac{d}{dt}\iint_{\{(x,y):x\leqslant t-y\}} f_{X,Y}(x, y)\,dx\,dy \\ &= \frac{d}{dt}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{t-y} f_{X,Y}(x, y)\,dx\right)dy \\ &= \int_{-\infty}^{\infty}\frac{d}{dt}\left(\int_{-\infty}^{t-y} f_{X,Y}(x, y)\,dx\right)dy \\ &= \int_{-\infty}^{\infty} f_{X,Y}(t - y, y)\,dy. \end{aligned}$$

So if X and Y are independent, we get the **convolution**

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(t-y) f_Y(y)\, dy.$$

9.8 ⋆ Expectation of a random vector

Since random vectors are just vector-valued functions on a sample space S, we can add them and multiply them just like any other functions. For example, the sum of random vectors $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is given by

$$(\boldsymbol{X} + \boldsymbol{Y})(\omega) = \boldsymbol{X} + \boldsymbol{Y}(\omega) = \big(X_1(\omega), \ldots, X_n(\omega)\big) + \big(Y_1(\omega), \ldots, Y_n(\omega)\big).$$

Thus the set of random vectors is a vector space. In fact, the subset of random vectors whose components have a finite expectation is also a vector subspace of the vector space of all random vectors.

If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random vector, and each X_i has expectation $\boldsymbol{E}\, X_i$, the expectation of $\boldsymbol{X}$ is defined to be

$$\boldsymbol{E}\, \boldsymbol{X} = (\boldsymbol{E}\, X_1, \ldots, \boldsymbol{E}\, X_n).$$

- Expectation is a **linear operator** on the space of random vectors. This means that

$$\boldsymbol{E}(a\boldsymbol{X} + b\boldsymbol{Y}) = a\, \boldsymbol{E}\, \boldsymbol{X} + b\, \boldsymbol{E}\, \boldsymbol{Y}.$$

- Expectation is a **positive operator** on the space of random vectors. For vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$, define $\boldsymbol{x} \geqq 0$ if $x_i \geqslant 0$ for each $i = 1, \ldots, n$. Then

$$\boldsymbol{X} \geqq 0 \implies \boldsymbol{E}\, \boldsymbol{X} \geqq 0.$$

9.9 Covariance

Pitman [5]: § 6.4, p. 430

When X and Y are independent, we proved

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

More generally however, since **expectation is a positive linear operator**,

$$\begin{aligned}
\boldsymbol{Var}(X+Y) &= \boldsymbol{E}\big((X+Y) - \boldsymbol{E}(X+Y)\big)^2 \\
&= \boldsymbol{E}\big((X - \boldsymbol{E}\, X) + (Y - \boldsymbol{E}\, Y)\big)^2 \\
&= \boldsymbol{E}\big((X - \boldsymbol{E}\, X)^2 + 2(X - \boldsymbol{E}\, X)(Y - \boldsymbol{E}\, Y) + (Y - \boldsymbol{E}\, Y)^2 \\
&= \boldsymbol{Var}(X) + \boldsymbol{Var}(Y) + 2\, \boldsymbol{E}(X - \boldsymbol{E}\, X)(Y - \boldsymbol{E}\, Y).
\end{aligned}$$

9.9.1 Definition *The* ***covariance*** *of X and Y is defined to be*

$$\boldsymbol{Cov}(X, Y) = \boldsymbol{E}(X - \boldsymbol{E}\, X)(Y - \boldsymbol{E}\, Y). \tag{1}$$

In general

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

There is another way to write the covariance:

$$\boldsymbol{Cov}(X,Y) = \boldsymbol{E}(XY) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y). \tag{2}$$

Proof: Since **expectation is a positive linear operator**,

$$\begin{aligned}\boldsymbol{Cov}(X,Y) &= \boldsymbol{E}\big((X - \boldsymbol{E}(X))\big(Y - \boldsymbol{E}(Y)\big)\big)\\ &= \boldsymbol{E}\big(XY - X\,\boldsymbol{E}(Y) - Y\,\boldsymbol{E}(X) + \boldsymbol{E}(X)\,\boldsymbol{E}(Y)\big)\\ &= \boldsymbol{E}(XY) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y) + \boldsymbol{E}(X)\,\boldsymbol{E}(Y)\\ &= \boldsymbol{E}(XY) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y).\end{aligned}$$

∎

9.9.2 Remark It follows that for any random variable X,

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

If X and Y are independent, then $\boldsymbol{Cov}(X,Y) = 0$.

The converse is not true.

9.9.3 Example (Covariance = 0, but variables are not independent) (Cf. Feller [3, p. 236])

Let X be a random variable that assumes the values ± 1 and ± 2, each with probability $1/4$. ($\boldsymbol{E}\,X = 0$)

Define $Y = X^2$, and let $\bar{Y} = \boldsymbol{E}\,Y$ $(= 2.5)$. Then

$$\begin{aligned}\boldsymbol{Cov}(X,Y) &= \boldsymbol{E}\big(X(Y - \bar{Y})\big)\\ &= \big(1(1-\bar{Y})\big)\,\frac{1}{4} + \big((-1)(1-\bar{Y})\big)\,\frac{1}{4} + \big(2(4-\bar{Y})\big)\,\frac{1}{4} + \big((-2)(4-\bar{Y})\big)\,\frac{1}{4}\\ &= 0.\end{aligned}$$

But X and Y are not independent:

$$P(X = 1 \;\&\; Y = 1) = P(X = 1) = 1/2,$$

but $P(X = 1) = 1/4$ and $P(Y = 1) = 1/2$, so

$$P(X = 1) \cdot P(Y = 1) = 1/8.$$

□

9.9.4 Example (Covariance = 0, but variables are not independent) Let U, V be independent and identically distributed random variables with $\boldsymbol{E}\,U = \boldsymbol{E}\,V = 0$. Define

$$X = U + V, \qquad Y = U - V.$$

Since $\boldsymbol{E}\,X = \boldsymbol{E}\,Y = 0$,

$$\boldsymbol{Cov}(X,Y) = \boldsymbol{E}(XY) = \boldsymbol{E}\big((U+V)(U-V)\big) = \boldsymbol{E}(U^2 - V^2) = \boldsymbol{E}\,U^2 - \boldsymbol{E}\,V^2 = 0$$

since U and V have the same distribution.

But are X and Y independent?

If U and V are integer-valued, then X and Y are also integer-valued, but more importantly they have the same parity. That is, X is odd if and only if Y is odd. (This is a handy fact for KenKen solvers.)

So let U and V be independent and assume the values ± 1 and ± 2, each with probability 1/4. ($\boldsymbol{E}\,U = \boldsymbol{E}\,V = 0$.) Then

$$P(X \text{ is odd}) = P(X \text{ is even}) = P(Y \text{ is odd}) = P(Y \text{ is even}) = \frac{1}{2},$$

but

$$P(X \text{ is even and } Y \text{ is odd}) = 0 \neq \frac{1}{4} = P(X \text{ is even})P(Y \text{ is odd}),$$

so X and Y are not independent. □

Pitman [5]: p. 432

9.9.5 Remark The product $(X - \boldsymbol{E}\,X)(Y - \boldsymbol{E}\,Y)$ is positive at outcomes ω where $X(\omega)$ and $Y(\omega)$ are either both above or both below their means, and negative when one is above and the other below. So one very loose interpretation of positive covariance is that the random variables are probably both above average or below average rather than not. Of course this is just a tendency.

9.9.6 Example (The effects of covariance) For mean zero random variables that have a positive covariance, the joint density tends to concentrate on the diagonal. Figure 9.2 shows the joint density of two standard normals with various covariances. Figure 9.4 shows random samples from these distributions.

□

9.10 ⋆ A covariance menagerie

Recall that for independent random variables X and Y, $\boldsymbol{Var}(X + Y) = \boldsymbol{Var}\,X + \boldsymbol{Var}\,Y$, and $\boldsymbol{Cov}(XY) = 0$. For any random variable X with finite variance, $\boldsymbol{Var}\,X = \boldsymbol{E}(X^2) - (\boldsymbol{E}\,X)^2$, so $\boldsymbol{E}(X^2) = \boldsymbol{Var}\,X + (\boldsymbol{E}\,X)^2$. Also, if $\boldsymbol{E}\,X = 0$, then $\boldsymbol{Cov}(XY) = \boldsymbol{E}(XY)$ (Why?).

9.10.1 Theorem (A Covariance Menagerie) *Let $X_1, \dots, X_n$ be independent and identically distributed random variables with common mean μ and variance σ^2. Define*

$$S = \sum_{i=1}^{n} X_i, \quad \text{and} \quad \bar{X} = S/n,$$

and let

$$D_i = X_i - \bar{X}, \quad (i = 1, \dots, n)$$

be the deviation of X_i from $\bar{X}$.

Then

1. $\boldsymbol{E}(X_i X_j) = (\boldsymbol{E}\,X_i)(\boldsymbol{E}\,X_j) = \mu^2$, *for $i \neq j$ (by independence).*
2. $\boldsymbol{E}(X_i^2) = \sigma^2 + \mu^2$.
3. $\boldsymbol{E}(X_i S) = \sum_{\substack{j=1 \\ j \neq i}}^{n} \boldsymbol{E}(X_i X_j) + \boldsymbol{E}(X_i^2) = \boldsymbol{E}(X_i^2) = \sigma^2 + n\mu^2$.
4. $\boldsymbol{E}(X_i \bar{X}) = (\sigma^2/n) + \mu^2$.
5. $\boldsymbol{E}(S) = n\mu$.
6. $\boldsymbol{Var}(S) = n\sigma^2$.

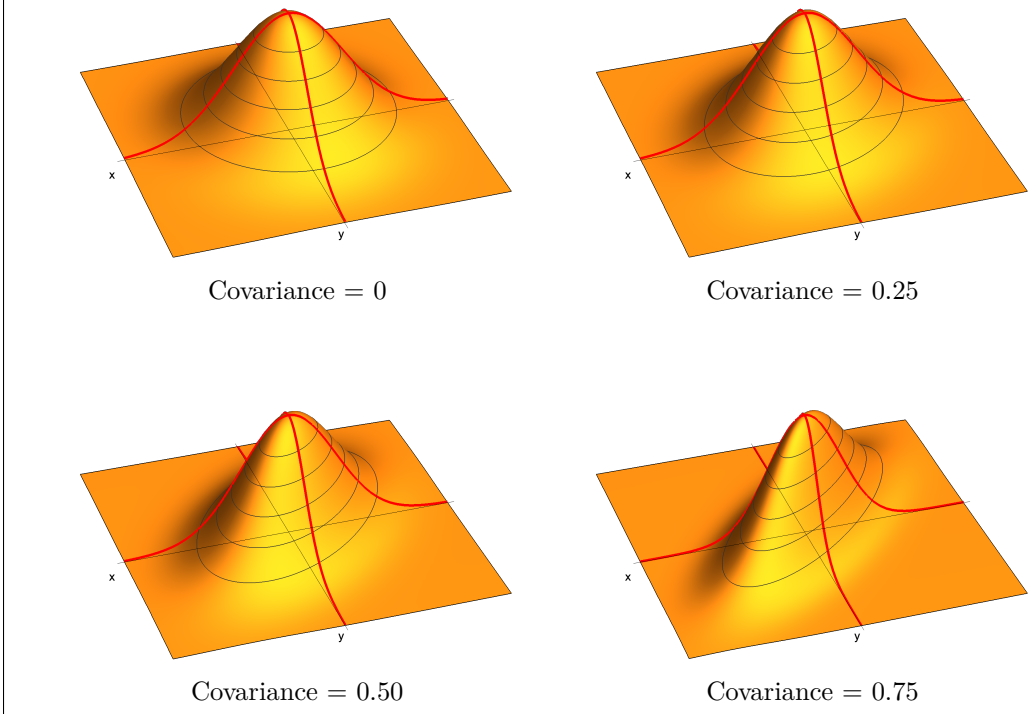

Figure 9.2. Joint density of standard normals, as covariance changes.

7. $\boldsymbol{E}(S^2) = n\sigma^2 + n^2\mu^2$.

8. $\boldsymbol{E}(\bar{X}) = \mu$.

9. $\boldsymbol{Var}(\bar{X}) = \sigma^2/n$.

10. $\boldsymbol{E}(\bar{X}^2) = (\sigma^2/n) + \mu^2$.

11. $\boldsymbol{E}(D_i) = 0, \qquad i = 1, \dots, n.$

12. $\boldsymbol{Var}(D_i) = \boldsymbol{E}(D_i^2) = (n-1)\sigma^2/n$:

$$\begin{aligned}\boldsymbol{Var}(D_i) &= \boldsymbol{E}(X_i - \bar{X})^2 \\ &= \boldsymbol{E}(X_i^2) - 2\,\boldsymbol{E}(X_i\bar{X}) + \boldsymbol{E}(\bar{X}^2) \\ &= (\sigma^2 + \mu^2) - 2\big((\sigma^2/n) + \mu^2\big) + \big((\sigma^2/n) + \mu^2\big) = \left(1 - \frac{1}{n}\right)\sigma^2.\end{aligned}$$

13. $\boldsymbol{Cov}(D_i, D_j) = \boldsymbol{E}(D_iD_j) = -\sigma^2/n$:

$$\begin{aligned}\boldsymbol{E}(D_iD_j) &= \boldsymbol{E}\big((X_i - \bar{X})(X_j - \bar{X})\big) \\ &= \boldsymbol{E}(X_iX_j) - \boldsymbol{E}(X_i\bar{X}) - \boldsymbol{E}(X_j\bar{X}) + \boldsymbol{E}(\bar{X}^2) \\ &= \mu^2 - \big[(\sigma^2/n) + \mu^2\big] - \big[(\sigma^2/n) + \mu^2\big] + \big[(\sigma^2/n) + \mu^2\big] = -\sigma^2/n.\end{aligned}$$

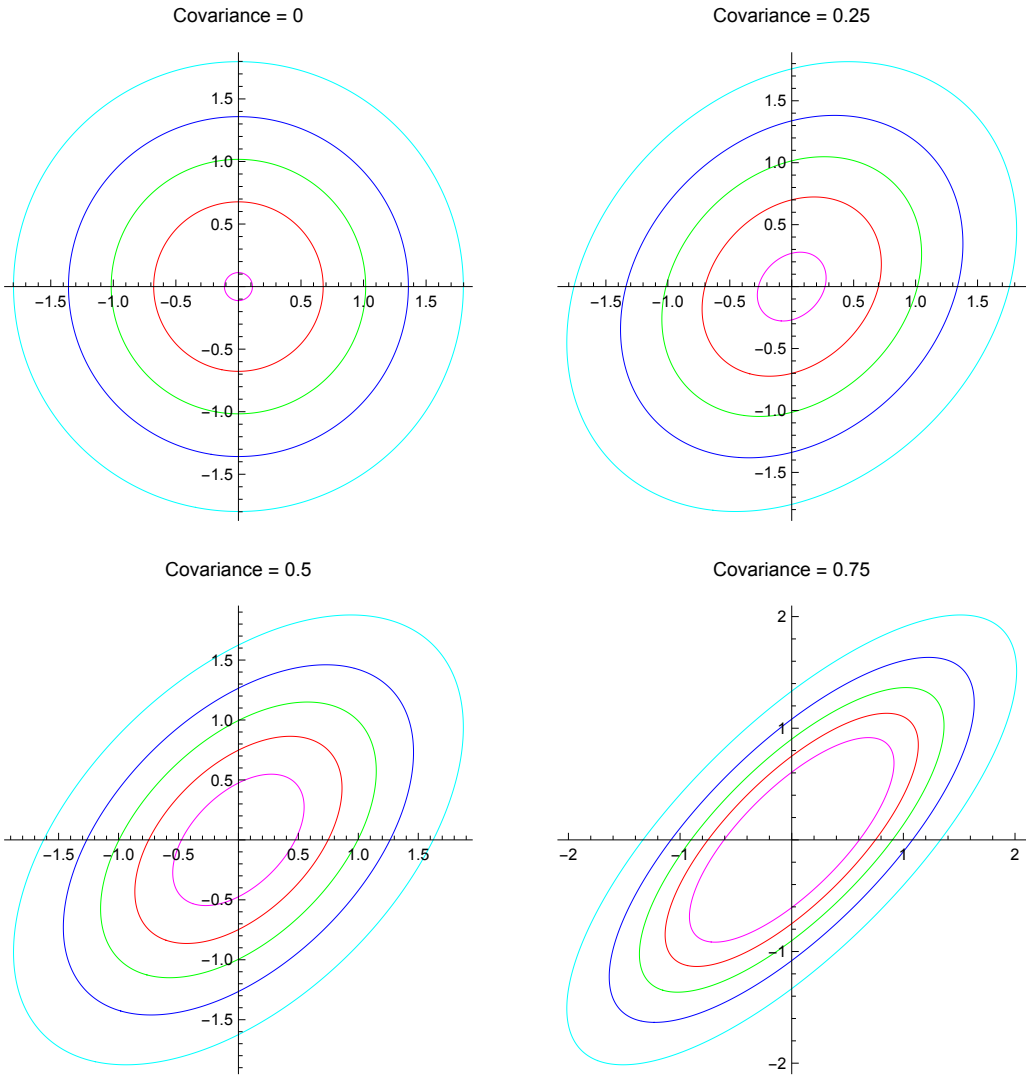

Figure 9.3. Contours of the joint density of standard normals, as covariance changes.

Note that this means that deviations from the mean are negatively correlated. This makes sense, because if one variate is bigger than the mean, another must be smaller to offset the difference.

14. $\boldsymbol{Cov}(D_i, S) = \boldsymbol{E}(D_i S) = 0.$

$$\begin{aligned}\boldsymbol{E}(D_i S) = \boldsymbol{E}\big(\big(X_i - (S/n)\big)S\big) &= \boldsymbol{E}(X_i S) - \boldsymbol{E}(S^2/n)\\ &= (\sigma^2 + n\mu^2) - (n\sigma^2 + n^2\mu^2)/n = 0.\end{aligned}$$

15. $\boldsymbol{Cov}(D_i, \bar{X}) = \boldsymbol{E}(D_i \bar{X}) = \boldsymbol{E}(D_i S)/n = 0.$

The proof of each is a straightforward plug-and-chug calculation. The only reason for writing this as a theorem is to be able to refer to it easily.

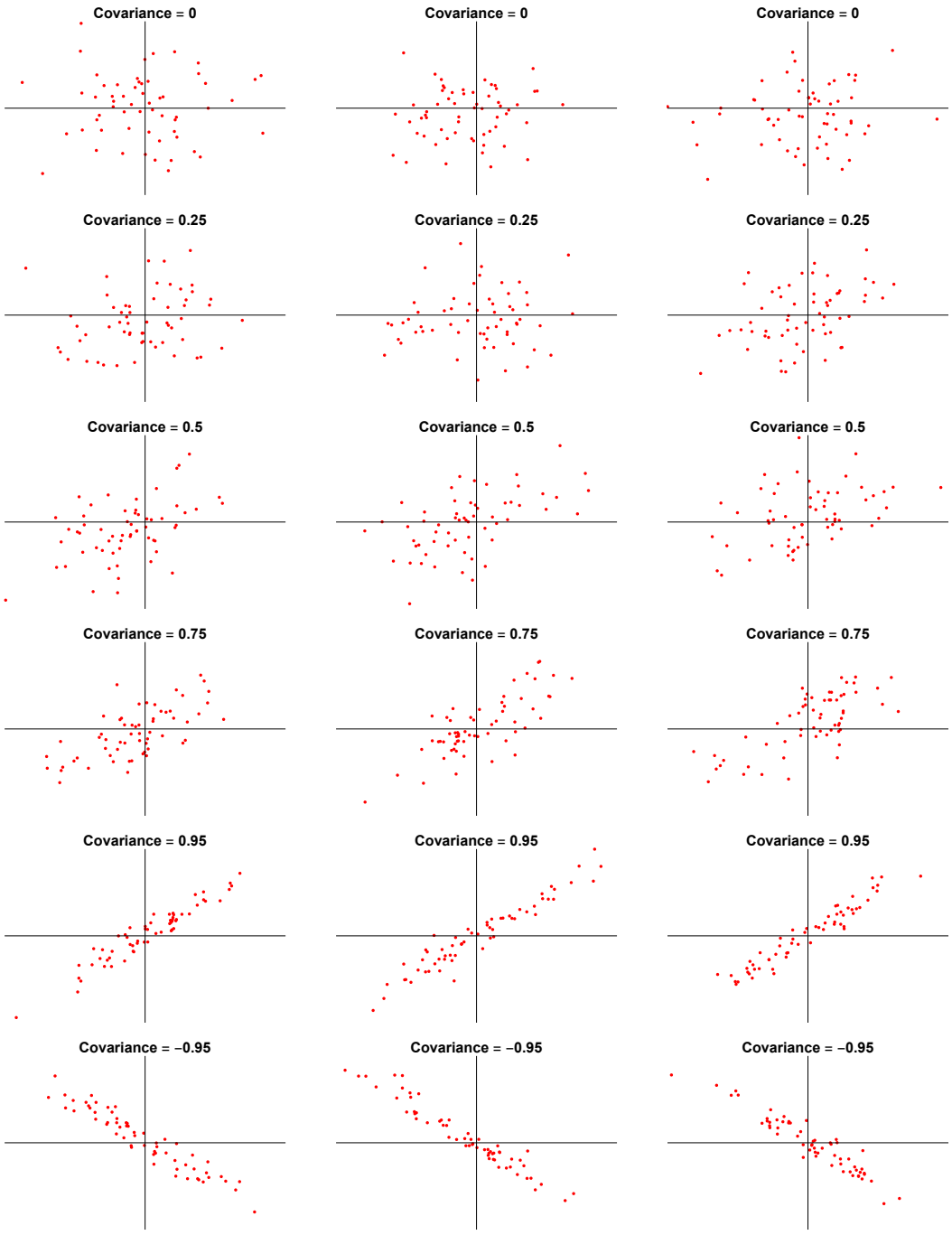

Figure 9.4. Random samples from standard normals, as covariance changes.

9.11 ⋆ Covariance matrix of a random vector

In general, we define the **covariance matrix** of a random vector by

$$\boldsymbol{Cov}\,\boldsymbol{X} = \begin{bmatrix} \vdots \\ \cdots\ \boldsymbol{E}(X_i - \boldsymbol{E}\,X_i)(X_j - \boldsymbol{E}\,X_j)\ \cdots \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \cdots\ \boldsymbol{Cov}(X_i, X_j)\ \cdots \\ \vdots \end{bmatrix}$$

9.12 ⋆ Variance of a linear combination of random variables

9.12.1 Proposition *Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random vector with covariance matrix*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \vdots \\ \cdots\ \boldsymbol{Cov}(X_i, X_i)\ \cdots \\ \vdots \end{bmatrix}$$

and let $\boldsymbol{a} = (a_1, \dots, a_n)$. The random variable

$$Z = \boldsymbol{a} \cdot \boldsymbol{X} = \boldsymbol{a}'\boldsymbol{X} = \sum_{i=1}^{n} a_i X_i$$

has variance given by

$$\boldsymbol{Var}\,Z = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{Cov}(X_i, X_j) a_i a_j,$$

where $\boldsymbol{a}$ is treated as a column vector, and $\boldsymbol{a}'$ is its transpose, a row vector.

Proof: This just uses the fact that **expectation is a positive linear operator**. Since adding constants don't change variance, we may subtract means and assume that each $\boldsymbol{E}\,X_i = 0$. Then $\boldsymbol{Cov}(X_i, X_j) = \boldsymbol{E}(X_i X_j)$. Then Z has mean 0, so

$$\begin{aligned} \boldsymbol{Var}\,Z = \boldsymbol{E}\,Z^2 &= \boldsymbol{E}\left(\sum_{i=1} a_i X_i\right)\left(\sum_{j=1} a_j X_j\right) \\ &= \boldsymbol{E}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j a_i a_j = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{E}(X_i X_j) a_i a_j \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{Cov}(X_i, X_j) a_i a_j. \end{aligned}$$

∎

Since $\boldsymbol{Var}\,Z \geqslant 0$ and $\boldsymbol{a}$ is arbitrary, we see that $\boldsymbol{Cov}\,\boldsymbol{X}$ is a positive semidefinite matrix.

9.13 ⋆ An inner product for random variables

Since random variables are just functions on the probability space $(\Omega, \mathcal{F}, P)$, the set of random variables is a **vector space** under the usual operations of addition of functions and multiplication by scalars. The collection of random variables that have finite variance is a linear subspace, often denoted $L_2(P)$, and it has a natural inner product.

9.13.1 Fact (Inner product on the space $L_2(P)$ of random variables) *Let $L_2(P)$ denote the linear space of random variables that have finite variance. Then*

$$(X, Y) = \boldsymbol{E}\, XY$$

is a real inner product on $L_2(P)$.

The proof of this straightforward, and is essentially the same as the proof that the Euclidean inner product on $\boldsymbol{R}^{\mathrm{m}}$ is an inner product.

The next result is just the Cauchy–Schwartz Inequality for this inner product, but I've written out a self-contained proof for you.

9.13.2 Cauchy–Schwartz Inequality

$$\boldsymbol{E}(XY)^2 \leqslant (\boldsymbol{E}\, X^2)(\boldsymbol{E}\, Y^2), \tag{3}$$

with equality only if X and Y are linearly dependent, that is, only if there exist a, b not both zero such that $aX + bY = 0$ a.s..

Proof: If either X or Y is zero a.s., then we have equality, so assume X, Y are nonzero. Define the quadratic polynomial $Q\colon \boldsymbol{R} \to \boldsymbol{R}$ by

$$Q(\lambda) = \boldsymbol{E}\big((\lambda X + Y)^2\big) \geqslant 0.$$

Since **expectation is a positive linear operator**,

$$Q(\lambda) = \boldsymbol{E}(X^2)\lambda^2 + 2\,\boldsymbol{E}(XY)\lambda + \boldsymbol{E}(Y^2).$$

Since this is always $\geqslant 0$, the discriminant of the quadratic polynomial $Q(\lambda)$ is nonpositive,[1] that is, $4\,\boldsymbol{E}(XY)^2 - 4\,\boldsymbol{E}(X^2)\,\boldsymbol{E}(Y^2) \leqslant 0$, or $\boldsymbol{E}(XY)^2 \leqslant \boldsymbol{E}(X^2)\,\boldsymbol{E}(Y^2)$. Equality in (3) can occur only if the discriminant is zero, in which case Q has a real root. That is, there is some λ for which $Q(\lambda) = \boldsymbol{E}\big((\lambda X + Y)^2\big) = 0$. But this implies that $\lambda X + Y = 0$ (almost surely). ∎

9.13.3 Corollary

$$\big|\boldsymbol{Cov}(X, Y)\big|^2 \leqslant \boldsymbol{Var}X\ \boldsymbol{Var}Y. \tag{4}$$

Proof: Apply the Cauchy–Schwartz inequality to the random variables, $X - \boldsymbol{E}\, X$ and $Y - \boldsymbol{E}\, Y$, and then take square roots. ∎

9.14 Covariance is bilinear

Since **expectation is a positive linear operator**, it is routine to show that

$$\boldsymbol{Cov}(aX + bY, cZ + dW) = ac\,\boldsymbol{Cov}(X, Z) + bc\,\boldsymbol{Cov}(Y, Z) + ad\,\boldsymbol{Cov}(X, W) + bd\,\boldsymbol{Cov}(Y, W).$$

[1] In case you have forgotten how you derived the quadratic formula in Algebra I, rewrite the polynomial as

$$Q(z) = \alpha x^2 + \beta z + \gamma = \tfrac{1}{\alpha}\big(\alpha z + \tfrac{\beta}{2}\big)^2 - (\beta^2 - 4\alpha\gamma)/4\alpha,$$

and note that the only way to guarantee that $Q(z) \geqslant 0$ for all z is to have $\alpha > 0$ and $\beta^2 - 4\alpha\gamma \leqslant 0$.

9.15 Correlation

9.15.1 Definition *The **correlation** between X and Y is defined to be*

$$\operatorname{Corr}(X,Y) = \frac{\boldsymbol{Cov}(X,Y)}{(\operatorname{SD} X)(\operatorname{SD} Y)}$$

It is also equal to

$$\operatorname{Corr}(X,Y) = \boldsymbol{Cov}(X^*,Y^*) = \boldsymbol{E}(X^*Y^*),$$

where X^* and Y^* are the standardization of X and Y.

Let X have mean μ_X and standard deviation σ_X, and ditto for Y. Recall that

$$X^* = \frac{X-\mu_X}{\sigma_X}$$

has mean 0 and std. dev. 1. Thus by the alternate formula for covariance

$$\boldsymbol{Cov}(X^*,Y^*) = \boldsymbol{E}(X^*Y^*) - \underset{=0}{\boldsymbol{E}(X^*)}\,\underset{=0}{\boldsymbol{E}(Y^*)}\,.$$

Now

$$\begin{aligned}\boldsymbol{E}(X^*Y^*) &= \boldsymbol{E}\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\\ &= \frac{\boldsymbol{E}(XY)-\boldsymbol{E}(X)\,\boldsymbol{E}(Y)}{\sigma_X\sigma_Y}\\ &= \operatorname{Corr}(X,Y)\end{aligned}$$

Pitman [5]: p. 433

Corollary 9.13.3 (the Cauchy–Schwartz Inequality) implies:

$$-1 \leqslant \operatorname{Corr}(X,Y) \leqslant 1$$

If the correlation between X and Y is zero, then the random variables $X - \boldsymbol{E}\,X$ and $Y - \boldsymbol{E}\,Y$ are orthogonal in our inner product.

9.16 Linear transformations of random vectors

Let

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

be a random vector. Define

$$\boldsymbol{\mu} = \boldsymbol{E}\,\boldsymbol{X} = \begin{bmatrix} \boldsymbol{E}\,X_1 \\ \vdots \\ \boldsymbol{E}\,X_n \end{bmatrix}.$$

Define the **covariance matrix** of $\boldsymbol{X}$ by

$$\boldsymbol{Var}\,\boldsymbol{X} = \big[\boldsymbol{Cov}\,X_iX_j\big] = \big[\boldsymbol{E}(X_i - \boldsymbol{\mu}_i)(X_j - \boldsymbol{\mu}_j)\big] = \big[\sigma_{ij}\big] = \boldsymbol{E}\left((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\right).$$

Let

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

be an $m \times n$ matrix of constants, and let

$$\boldsymbol{Y} = \boldsymbol{AX}$$

Then, since **expectation is a positive linear operator**,

$$\boldsymbol{E}\,\boldsymbol{Y} = \boldsymbol{A\mu}.$$

Moreover

$$\boldsymbol{Var}\,\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{Var}\,\boldsymbol{X})\boldsymbol{A}'$$

since $\boldsymbol{Var}\,\boldsymbol{Y} = \boldsymbol{E}\left((\boldsymbol{AX} - \boldsymbol{A\mu})(\boldsymbol{AX} - \boldsymbol{A\mu})'\right) = \boldsymbol{E}\left(\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{A}'\right) = \boldsymbol{A}(\boldsymbol{Var}\,\boldsymbol{X})\boldsymbol{A}'$.

The covariance matrix $\boldsymbol{\Sigma}$ of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is always positive semidefinite, since for any vector $\boldsymbol{w}$ of weights, $\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$ is the variance of the random variable $\boldsymbol{w}'\boldsymbol{Y}$, and variances are always nonnegative.

Bibliography

[1] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis*, 3d. ed. Berlin: Springer–Verlag.

[2] T. M. Apostol. 1967. *Calculus, Volume I: One-variable calculus with an introduction to linear algebra*, 2d. ed. New York: John Wiley & Sons.

[3] W. Feller. 1968. *An introduction to probability theory and its applications*, 3d. ed., volume 1. New York: Wiley.

[4] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[5] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.