p. 323- 325 Transformation or random variables
p. 325 Linear Transform V= X+Y W=X-Y

Understanding Probability

Chance Rules in Everyday Life

Second Edition

HENK TIJMS
Vrije University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521701723

First published 2007

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

ISBN 978-0-521-70172-3 paperback

Contents

11

Jointly distributed random variables

In experiments, one is often interested not only in individual random variables, but also in relationships between two or more random variables. For example, if the experiment is the testing of a new medicine, the researcher might be interested in cholesterol level, blood pressure, and the glucose level of a test person. Similarly, a political scientist investigating the behavior of voters might be interested in the income and level of education of a voter. There are many more examples in the physical sciences, medical sciences, and social sciences. In applications, one often wishes to make inferences about one random variable on the basis of observations of other random variables. The purpose of this chapter is to familiarize the student with the notations and the techniques relating to experiments whose outcomes are described by two or more real numbers. The discussion is restricted to the case of pairs of random variables. Extending the notations and techniques to collections of more than two random variables is straightforward.

11.1 Joint probability densities

It is helpful to discuss the joint probability mass function of two discrete random variables before discussing the concept of the joint density of two continuous random variables. In fact, Section 9.3 has dealt with the joint distribution of discrete random variables. If X and Y are two discrete random variables defined on a same sample space with probability measure P, the mass function $p(x, y)$ defined by

$$p(x, y) = P(X = x, Y = y)$$

is called the *joint probability mass function* of X and Y. As noted before, $P(X = x, Y = y)$ is the probability assigned by P to the intersection of the two sets

Table 11.1. *The joint probability mass function* $p(x, y)$.

$x \backslash y$	2	3	4	5	6	7	8	9	10	11	12	$p_X(x)$
1	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	0	0	$\frac{11}{36}$
2	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	0	$\frac{9}{36}$
3	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	$\frac{7}{36}$
4	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	$\frac{5}{36}$
5	0	0	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	0	$\frac{3}{36}$
6	0	0	0	0	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$p_Y(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	sum = 1

$A = \{\omega : X(\omega) = x\}$ and $B = \{\omega : Y(\omega) = y\}$, with ω representing an element of the sample space. The joint probability mass function uniquely determines the probability distributions $p_X(x) = P(X = x)$ and $p_Y(y) = P(Y = y)$ by

$$p_X(x) = \sum_y P(X = x, Y = y), \quad p_Y(y) = \sum_x P(X = x, Y = y).$$

These distributions are called the *marginal distributions* of X and Y.

Example 11.1 Two fair dice are rolled. Let the random variable X represent the smallest of the outcomes of the two rolls, and let Y represent the sum of the outcomes of the two rolls. What is the joint probability mass function of X and Y?

Solution. The random variables X and Y are defined on the same sample space. The sample space is the set of all 36 pairs (i, j) for $i, j = 1, \ldots, 6$, where i and j are the outcomes of the first and second dice. A probability of $\frac{1}{36}$ is assigned to each element of the sample space. In Table 11.1, we give the joint probability mass function $p(x, y) = P(X = x, Y = y)$. For example, $P(X = 2, Y = 5)$ is the probability of the intersection of the sets $A = \{(2,2), (2,3), (2,4), (2,5), (2,6), (3,2), (4,2), (5,2), (6,2)\}$ and $B = \{(1,4), (4,1), (2,3), (3,2)\}$. The set $\{(2,3), (3,2)\}$ is the intersection of these two sets and has probability $\frac{2}{36}$.

Problem 11.1 You roll a pair of dice. What is the joint probability mass function of the low and high points rolled?

Problem 11.2 Let X denote the number of hearts and Y the number of diamonds in a bridge hand. What is the joint probability mass function of X and Y?

The following example provides a good starting point for a discussion of joint probability densities.

Example 11.2 A point is picked at random inside a circular disc with radius r. Let the random variable X denote the length of the line segment between the center of the disc and the randomly picked point, and let the random variable Y denote the angle between this line segment and the horizontal axis (Y is measured in radians and so $0 \le Y < 2\pi$). What is the joint distribution of X and Y?

Solution. The two continuous random variables X and Y are defined on a common sample space. The sample space consists of all points (v, w) in the two-dimensional plane with $v^2 + w^2 \le r^2$, where the point $(0, 0)$ represents the center of the disc. The probability $P(A)$ assigned to each well-defined subset A of the sample space is taken as the area of region A divided by πr^2. The probability of the event of X taking on a value less than or equal to a and Y taking on a value less than or equal to b is denoted by $P(X \le a,\ Y \le b)$. This event occurs only if the randomly picked point falls inside the disc segment with radius a and angle b. The area of this disc segment is $\frac{b}{2\pi}\pi a^2$. Dividing this by πr^2 gives

$$P(X \le a, Y \le b) = \frac{b}{2\pi}\frac{a^2}{r^2} \qquad \text{for } 0 \le a \le r \text{ and } 0 \le b \le 2\pi.$$

We are now in a position to introduce the concept of joint density. Let X and Y be two random variables that are defined on the same sample space with probability measure P. The *joint cumulative probability distribution function* of X and Y is defined by $P(X \le x, Y \le y)$ for all x, y, where $P(X \le x, Y \le y)$ is a shorthand for $P(\{\omega : X(\omega) \le x \text{ and } Y(\omega) \le y\})$ and the symbol ω represents an element of the sample space.

Definition 11.1 *The continuous random variables X and Y are said to have a joint probability density function $f(x, y)$ if the joint cumulative probability distribution function $P(X \le a, Y \le b)$ allows for the representation*

$$P(X \le a, Y \le b) = \int_{x=-\infty}^{a}\int_{y=-\infty}^{b} f(x, y)\,dx\,dy, \qquad -\infty < a, b < \infty,$$

where the function $f(x, y)$ satisfies

$$f(x, y) \ge 0 \quad \textit{for all } x, y \quad \textit{and} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1.$$

Just as in the one-dimensional case, $f(a,b)$ allows for the interpretation

$$f(a,b)\,\Delta a\,\Delta b \approx P\left(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a, b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b\right)$$

for small positive values of Δa and Δb provided that $f(x,y)$ is continuous in the point (a,b). In other words, the probability that the random point (X,Y) falls into a small rectangle with sides of lengths Δa, Δb around the point (a,b) is approximately given by $f(a,b)\,\Delta a\,\Delta b$.

To obtain the joint probability density function $f(x,y)$ of the random variables X and Y in Example 11.2, we take the partial derivatives of $P(X \leq x, Y \leq y)$ with respect to x and y. It then follows from

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} P(X \leq x, Y \leq y)$$

that

$$f(x,y) = \begin{cases} \frac{1}{2\pi}\frac{2x}{r^2} & \text{for } 0 < x < r \text{ and } 0 < y < 2\pi, \\ 0 & \text{otherwise.} \end{cases}$$

In general, the joint probability density function is found by determining first the cumulative joint probability distribution function and taking next the partial derivatives. However, sometimes it is easier to find the joint probability density function by using its probabilistic interpretation. This is illustrated with the next example.

Example 11.3 The pointer of a spinner of radius r is spun three times. The three spins are performed independently of each other. With each spin, the pointer stops at an unpredictable point on the circle. The random variable L_i corresponds to the length of the arc from the top of the circle to the point where the pointer stops on the ith spin. The length of the arc is measured clockwise. Let $X = \min(L_1, L_2, L_3)$ and $Y = \max(L_1, L_2, L_3)$. What is the joint probability density function $f(x,y)$ of the two continuous random variables X and Y?

Solution. We can derive the joint probability density function $f(x,y)$ by using the interpretation that the probability $P(x < X \leq x + \Delta x, y < Y \leq y + \Delta y)$ is approximately equal to $f(x,y)\Delta x \Delta y$ for Δx and Δy small. The event $\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\}$ occurs only if one of the L_i takes on a value between x and $x + \Delta x$, one of the L_i a value between y and $y + \Delta y$, and the remaining L_i a value between x and y, where $0 < x < y$. There are $3 \times 2 \times 1 = 6$ ways in which L_1, L_2, L_3 can be arranged and the probability that for fixed i the random variable L_i takes on a value between a and b equals

$(b-a)/(2\pi r)$ for $0 < a < b < 2\pi r$ (explain!). Thus, by the independence of L_1, L_2, and L_3 (see the general Definition 9.2)

$$P(x < X \leq x + \Delta x, y < Y \leq y + \Delta y)$$
$$= 6\frac{(x + \Delta x - x)}{2\pi r}\frac{(y + \Delta y - y)}{2\pi r}\frac{(y - x)}{2\pi r}.$$

Hence, the joint probability density function of X and Y is given by

$$f(x, y) = \begin{cases} \frac{6(y-x)}{(2\pi r)^3} & \text{for } 0 < x < y < 2\pi r \\ 0 & \text{otherwise.} \end{cases}$$

In general, if the random variables X and Y have a joint probability density function $f(x, y)$

$$P((X, Y) \in C) = \int\int_C f(x, y)\, dx\, dy$$

for any set C of pairs of real numbers. In calculating a double integral over a nonnegative integrand, it does not matter whether we integrate over x first or over y first. This is a basic fact from calculus. The double integral can be written as a repeated one-dimensional integral. The expression for $P((X, Y) \in C)$ is very useful to determine the probability distribution function of any function $g(X, Y)$ of X and Y. To illustrate this, we derive the useful result that the sum $Z = X + Y$ has the probability density

$$f_Z(z) = \int_{-\infty}^{\infty} f(u, z - u)\, du.$$

To prove this *convolution formula*, note that

$$P(Z \leq z) = \iint_{\substack{(x,y):\\ x+y \leq z}} f(x, y)\, dx\, dy = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f(x, y)\, dx\, dy$$
$$= \int_{v=-\infty}^{z} \int_{u=-\infty}^{\infty} f(u, v - u)\, du\, dv,$$

using the change of variables $u = x$ and $v = x + y$. Next, differentiation of $P(Z \leq z)$ yields the convolution formula for $f_Z(z)$. If the random variables X and Y are nonnegative, the convolution formula reduces to

$$f_Z(z) = \int_0^z f(u, z - u)\, du \qquad \text{for } z > 0.$$

Uniform distribution over a region

Another useful result is the following. Suppose that a point (X, Y) is picked at random inside a bounded region R in the two-dimensional plane. Then, the joint probability density function $f(x, y)$ of X and Y is given by the uniform density

$$f(x, y) = \frac{1}{\text{area of region } R} \qquad \text{for } (x, y) \in R.$$

The proof is simple. For any subset $C \subseteq R$

$$P((X, Y) \in C) = \frac{\text{area of } C}{\text{area of } R},$$

being the mathematical definition of the random selection of a point inside the region R. Integral calculus tells us that area of $C = \iint_C dxdy$. Thus, for any subset $C \subseteq R$

$$P((X, Y) \in C) = \iint_C \frac{1}{\text{area of } R}\, dx\, dy,$$

showing that the random point (X, Y) has the above density $f(x, y)$.

In the following problems you are asked to apply the basic expression $P((X, Y) \in C) = \iint_C f(x, y)\, dx\, dy$ yourselves in order to find the probability density of a given function of X and Y.

Problem 11.3 A point (X, Y) is picked at random inside the triangle consisting of the points (x, y) in the plane with $x, y \geq 0$ and $x + y \leq 1$. What is the joint probability density of the point (X, Y)? Determine the probability density of each of the random variables $X + Y$ and $\max(X, Y)$.

Problem 11.4 Let X and Y be two random variables with a joint probability density

$$f(x, y) = \begin{cases} \frac{1}{(x+y)^3} & \text{for } x, y > c \\ 0 & \text{otherwise,} \end{cases}$$

for an appropriate constant c. Verify that $c = \frac{1}{4}$ and calculate the probability $P(X > a, Y > b)$ for $a, b > c$.

Problem 11.5 Independently of each other, two points are chosen at random in the interval (0, 1). What is the joint probability density of the smallest and the largest of these two random numbers? What is the probability density of the length of the middle interval of the three intervals that result from the two random points in (0,1)? What is the probability that the smallest of the three resulting intervals is larger than a?

Problem 11.6 Independently of each other, two numbers X and Y are chosen at random in the interval (0, 1). Let $Z = X/Y$ be the ratio of these two random numbers.

(a) Use the joint density of X and Y to verify that $P(Z \leq z)$ equals $\frac{1}{2}z$ for $0 < z < 1$ and equals $1 - 1/(2z)$ for $z \geq 1$.
(b) What is the probability that the first significant (nonzero) digit of Z equals 1? What about the digits $2, \ldots, 9$?
(c) What is the answer to Question (b) for the random variable $V = XY$?
(d) What is the density function of the random variable $(X/Y)U$ when U is a random number from (0, 1) that is independent of X and Y?

11.2 Marginal probability densities

If the two random variables X and Y have a joint probability density function $f(x, y)$, then each of the random variables X and Y has a probability density itself. Using the fact that $\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n)$ for any nondecreasing sequence of events A_n, it follows that

$$P(X \leq a) = \lim_{b\to\infty} P(X \leq a, Y \leq b) = \int_{-\infty}^{a} \left[\int_{-\infty}^{\infty} f(x, y)\, dy\right] dx.$$

This representation shows that X has probability density function

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \qquad -\infty < x < \infty.$$

In the same way, the random variable Y has probability density function

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx, \qquad -\infty < y < \infty.$$

The probability density functions $f_X(x)$ and $f_Y(y)$ are called the *marginal probability density functions* of X and Y. The following interpretation can be given to the marginal density $f_X(x)$ at the point $x = a$ when a is a continuity point of $f_X(x)$. For Δa small, $f_X(a)\Delta a$ gives approximately the probability that (X, Y) falls in a vertical strip in the two-dimensional plane with width Δa and around the vertical line $x = a$. A similar interpretation applies to $f_Y(b)$ for any continuity point b of $f_Y(y)$.

Example 11.4 A point (X, Y) is chosen at random inside the unit circle. What is the marginal density of X?

Solution. Denote by $C = \{(x, y) \mid x^2 + y^2 \leq 1\}$ the unit circle. The joint probability density function $f(x, y)$ of X and Y is given by $f(x, y) = 1/(\text{area of } C)$

for $(x, y) \in C$. Hence

$$f(x, y) = \begin{cases} \frac{1}{\pi} & \text{for } (x, y) \in C \\ 0 & \text{otherwise.} \end{cases}$$

Using the fact that $f(x, y)$ is equal to zero for those y satisfying $y^2 > 1 - x^2$, if follows that

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\, dy,$$

and so

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1 - x^2} & \text{for } -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Can you explain why the marginal density of X is not the uniform density on $(-1, 1)$? *Hint*: interpret $P(x < X \leq x + \Delta x)$ as the area of a vertical strip in the unit circle.

Problem 11.7 A point (X, Y) is chosen at random in the equilateral triangle having $(0, 0)$, $(1, 0)$, and $(\frac{1}{2}, \frac{1}{2}\sqrt{3})$ as corner points. Determine the marginal densities of X and Y. Before determining the function $f_X(x)$, can you explain why $f_X(x)$ must be largest at $x = \frac{1}{2}$?

A general condition for the independence of the jointly distributed random variables X and Y is stated in Definition 9.2. In terms of the marginal densities, the continuous analog of Rule 9.6 for the discrete case is:

Rule 11.1 *The jointly distributed random variables X and Y are independent if and only if*

$$f(x, y) = f_X(x) f_Y(y) \qquad \textit{for all } x, y.$$

Let us illustrate this with the random variables X and Y from Example 11.2. Then, we obtain from $f_X(x) = \int_0^{2\pi} \frac{x}{\pi r^2}\, dy$ that

$$f_X(x) = \begin{cases} \frac{2x}{r^2} & \text{for } 0 < x < r, \\ 0 & \text{otherwise.} \end{cases}$$

In the same way, we obtain from $f_Y(y) = \int_0^r \frac{x}{\pi r^2}\, dx$ that

$$f_Y(y) = \begin{cases} \frac{1}{2\pi} & \text{for } 0 < y < 2\pi, \\ 0 & \text{otherwise.} \end{cases}$$

The calculations lead to the intuitively obvious result that the angle Y has a uniform distribution on $(0, 2\pi)$. A somewhat more surprising result is that the distance X and the angle Y are independent random variables, though there is dependence between the components of the randomly picked point. The independence of X and Y follows from the observation that $f(x, y) = f_X(x)f_Y(y)$ for all x, y.

To conclude this subsection, we give a very important result for the exponential distribution.

Example 11.5 Suppose that X and Y are independent random variables, where X is exponentially distributed with expected value $1/\alpha$ and Y is exponentially distributed with expected value $1/\beta$. What is the probability distribution of $\min(X, Y)$? What is the probability that X is less than Y?

Solution. The answer to the first question is that $\min(X, Y)$ is exponentially distributed with expected value $1/(\alpha + \beta)$. It holds that

$$P(\min(X, Y) \leq z) = 1 - e^{-(\alpha+\beta)z} \quad \text{for } z \geq 0 \quad \text{and} \quad P(X < Y) = \frac{\alpha}{\alpha + \beta}.$$

The proof is simple. Noting that $P(\min(X, Y) \leq z) = 1 - P(X > z, Y > z)$, we have

$$P(\min(X, Y) \leq z) = 1 - \int_{x=z}^{\infty} \int_{y=z}^{\infty} f_X(x)f_Y(y)\,dx\,dy.$$

Also,

$$P(X < Y) = \int_{x=0}^{\infty} \int_{y=x}^{\infty} f_X(x)f_Y(y)\,dx\,dy.$$

Using the fact that $f_X(x) = \alpha e^{-\alpha x}$ and $f_Y(y) = \beta e^{-\beta y}$, it is next a matter of simple algebra to derive the results. The details are left to the reader.

Problem 11.8 The continuous random variables X and Y are nonnegative and independent. Verify that the density function of $Z = X + Y$ is given by the convolution formula

$$f_Z(z) = \int_0^z f_X(z - y)f_Y(y)dy \qquad \text{for } z \geq 0.$$

Problem 11.9 The nonnegative random variables X and Y are independent and uniformly distributed on (c, d). What is the probability density of $Z = X + Y$? What is the probability density function of $V = X^2 + Y^2$? Use the latter density to calculate the expected value of the distance of a point chosen at random inside the unit square to the center of the unit square.

11.2.1 Substitution rule

The expected value of a given function of jointly distributed random variables X and Y can be calculated by the two-dimensional substitution rule. In the continuous case, we have:

Rule 11.2 *If the random variables X and Y have a joint probability density function $f(x, y)$, then*

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

for any function $g(x, y)$ provided that the integral is well defined.

An easy consequence of Rule 11.2 is that

$$E(aX + bY) = aE(X) + bE(Y)$$

for any constants a, b provided that $E(X)$ and $E(Y)$ exist. To see this, note that

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax + by) f(x, y)\, dx\, dy \\
&\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ax f(x, y)\, dx\, dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} by f(x, y)\, dx\, dy \\
&\quad = \int_{x=-\infty}^{\infty} ax\, dx \int_{y=-\infty}^{\infty} f(x, y)\, dy + \int_{y=-\infty}^{\infty} by\, dy \int_{x=-\infty}^{\infty} f(x, y)\, dx \\
&\quad = a\int_{-\infty}^{\infty} x f_X(x)\, dx + b\int_{-\infty}^{\infty} y f_Y(y)\, dy,
\end{aligned}$$

which proves the desired result. It is left to the reader to verify from Rules 11.1 and 11.2 that

$$E(XY) = E(X)E(Y) \qquad \text{for independent } X \text{ and } Y.$$

An illustration of the substitution rule is provided by Problem 2.21: what is the expected value of the distance between two points that are chosen at random in the interval (0, 1)? To answer this question, let X and Y be two independent random variables that are uniformly distributed on (0, 1). The joint density function of X and Y is given by $f(x, y) = 1$ for all $0 < x, y < 1$. The substitution rule gives

$$\begin{aligned}E(|X-Y|) &= \int_0^1\int_0^1 |x-y|\,dx dy\\ &= \int_0^1 dx\left[\int_0^x (x-y)\,dy + \int_x^1 (y-x)\,dy\right]\\ &= \int_0^1\left[\frac{1}{2}x^2+\frac{1}{2}-\frac{1}{2}x^2-x(1-x)\right]dx = \frac{1}{3}.\end{aligned}$$

Hence, the answer to the question is $\frac{1}{3}$.

As another illustration of Rule 11.2, consider Example 11.2 again. In this example, a point is picked at random inside a circular disk with radius r and the point (0, 0) as center. What is the expected value of the rectangular distance from the randomly picked point to the center of the disk? This rectangular distance is given by $|X\cos(Y)| + |X\sin(Y)|$ (the rectangular distance from point (a, b) to (0, 0) is defined by $|a| + |b|$). For the function $g(x, y) = |x\cos(y)| + |x\sin(y)|$, we find

$$\begin{aligned}E[g(X,Y)] &= \int_0^r\int_0^{2\pi} \{x|\cos(y)| + x|\sin(y)|\}\frac{x}{\pi r^2}\,dx\,dy\\ &= \frac{1}{\pi r^2}\int_0^{2\pi}|\cos(y)|\,dy\int_0^r x^2\,dx + \frac{1}{\pi r^2}\int_0^{2\pi}|\sin(y)|\,dy\int_0^r x^2\,dx\\ &= \frac{r^3}{3\pi r^2}\left[\int_0^{2\pi}|\cos(y)|\,dy + \int_0^{2\pi}|\sin(y)|\,dy\right] = \frac{8r}{3\pi}.\end{aligned}$$

The same ideas hold in the discrete case with the probability mass function assuming the role of the density function

$$E[g(X,Y)] = \sum_x\sum_y g(x,y)p(x,y)$$

when the random variables X and Y have the joint probability mass function $p(x, y) = P(X = x, Y = y)$. •

11.3 Transformation of random variables

In statistical applications, one sometimes needs the joint density of two random variables V and W that are defined as functions of two other random variables X and Y having a joint density $f(x, y)$. Suppose that the random variables V and W are defined by $V = g(X, Y)$ and $W = h(X, Y)$ for given functions g and h. What is the joint probability density function of V and W? An answer to this question will be given under the assumption that the transformation is one-to-one. That is, it is assumed that the equations $v = g(x, y)$ and $w = h(x, y)$ can be solved uniquely to yield functions $x = a(v, w)$ and $y = b(v, w)$. Also assume that the partial derivatives of the functions $a(v, w)$ and $b(v, w)$ with respect to v and w are continuous in (v, w). Then the following transformation rule holds:

Rule 11.3 *The joint probability density function of V and W is given by*

$$f(a(v, w), b(v, w))|J(v, w)|,$$

where the Jacobian $J(v, w)$ is given by the determinant

$$\begin{vmatrix} \frac{\partial a(v,w)}{\partial v} & \frac{\partial a(v,w)}{\partial w} \\ \frac{\partial b(v,w)}{\partial v} & \frac{\partial b(v,w)}{\partial w} \end{vmatrix} = \frac{\partial a(v, w)}{\partial v}\frac{\partial b(v, w)}{\partial w} - \frac{\partial a(v, w)}{\partial w}\frac{\partial b(v, w)}{\partial v}.$$

The proof of this rule is omitted. This transformation rule looks intimidating, but is easy to use in many applications. In the next section it will be shown how Rule 11.3 can be used to devise a method for simulating from the normal distribution. However, we first give a simple illustration of Rule 11.3. Suppose that X and Y are independent $N(0, 1)$ random variables. Then, the random variables $V = X + Y$ and $W = X - Y$ are normally distributed and independent. To verify this, note that the inverse functions $a(v, w)$ and $b(v, w)$ are given by $x = \frac{v+w}{2}$ and $y = \frac{v-w}{2}$. Thus, the Jacobian $J(v, w)$ is equal to

$$\begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Since X and Y are independent $N(0, 1)$ random variables, it follows from Rule 11.1 that their joint density function is given by

$$f_{X,Y}(x, y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \times \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}, \qquad -\infty < x, y < \infty.$$

Applying Rule 11.3, we obtain that the joint density function of V and W is given by

$$f_{V,W}(v, w) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{v+w}{2})^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{v-w}{2})^2} \times \frac{1}{2}$$

$$= \frac{1}{\sqrt{2}\sqrt{2\pi}}e^{-\frac{1}{2}v^2/2} \times \frac{1}{\sqrt{2}\sqrt{2\pi}}e^{-\frac{1}{2}w^2/2}, \qquad -\infty < v, w < \infty.$$

This implies that $f_{V,W}(v, w) = f_V(v)f_W(w)$ for all v, w with the marginal density functions $f_V(v) = \frac{1}{\sqrt{2}\sqrt{2\pi}}e^{-\frac{1}{2}v^2/2}$ and $f_W(w) = \frac{1}{\sqrt{2}\sqrt{2\pi}}e^{-\frac{1}{2}w^2/2}$. Using

Rule 11.1 again, it now follows that $V = X + Y$ and $W = X - Y$ are $N(0, 2)$ distributed and independent.

11.3.1 Simulating from a normal distribution

A natural transformation of two independent standard normal random variables leads to a practically useful method for simulating random observations from the standard normal distribution. Suppose that X and Y are independent random variables each having the standard normal distribution. Using Rule 11.1, the joint probability density function of X and Y is given by

$$f(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}.$$

The random vector (X, Y) can be considered as a point in the two-dimensional plane. Let the random variable V be the distance from the point $(0, 0)$ to the point (X, Y) and let W be the angle that the line through the points $(0, 0)$ and (X, Y) makes with the horizontal axis. The random variables V and W are functions of X and Y (the function $g(x, y) = \sqrt{x^2 + y^2}$ and $h(x, y) = \arctan(y/x)$). The inverse functions $a(v, w)$ and $b(v, w)$ are very simple. By basic geometry, $x = v\cos(w)$ and $y = v\sin(w)$. We thus obtain the Jacobian

$$\begin{vmatrix} \cos(w) & -v\sin(w) \\ \sin(w) & v\cos(w) \end{vmatrix} = v\cos^2(w) + v\sin^2(w) = v,$$

using the celebrated identity $\cos^2(w) + \sin^2(w) = 1$. Hence, the joint probability density function of V and W is given by

$$f_{V,W}(v, w) = \frac{v}{2\pi} e^{-\frac{1}{2}(v^2\cos^2(w)+v^2\sin^2(w))} = \frac{v}{2\pi} e^{-\frac{1}{2}v^2}$$

for $0 < v < \infty$ and $0 < w < 2\pi$. The marginal densities of V and W are given by

$$f_V(v) = \frac{1}{2\pi}\int_0^{2\pi} v e^{-\frac{1}{2}v^2}\,dw = v e^{-\frac{1}{2}v^2}, \qquad 0 < v < \infty$$

and

$$f_W(w) = \frac{1}{2\pi}\int_0^{\infty} v e^{-\frac{1}{2}v^2}\,dv = \frac{1}{2\pi}, \qquad 0 < w < 2\pi.$$

Since $f_{V,W}(v, w) = f_V(v) f_W(w)$, we have the remarkable finding that V and W are independent random variables. The random variable V has the probability density function $v e^{-\frac{1}{2}v^2}$ for $v > 0$ and W is uniformly distributed on $(0, 2\pi)$. This result is extremely useful for simulation purposes. Using the inverse-transformation method from Section 10.3, it is a simple matter to simulate random observations from the probability distributions of V and W. If we let U_1 and U_2 denote two independent random numbers from the interval (0,1), it follows from results in Section 10.3 that random observations of V and W are given by

$$V = \sqrt{-2\ln(1 - U_1)} \quad \text{and} \quad W = 2\pi U_2.$$

Next, one obtains two random observations X and Y from the standard normal distribution by taking

$$X = V\cos(W) \quad \text{and} \quad Y = V\sin(W).$$

Theoretically, X and Y are independent of each other. However, if a pseudo-random generator is used to generate U_1 and U_2, one uses only one of two variates X and Y. It surprisingly appears that the points (X, Y) lie on a spiral in the plane when a multiplicative generator is used for the pseudo-random numbers. The explanation of this subtle dependency lies in the fact that pseudo-random numbers are not truly random. The method described above for generating normal variates is known as the Box-Muller method.

Problem 11.10 A point (V, W) is chosen inside the unit circle as follows. First, a number R is chosen at random between 0 and 1. Next, a point is chosen at random on the circumference of the circle with radius R. Use the transformation formula to find the joint density function of this point (V, W). What is the marginal density function of each of the components of the point (V, W)? Can you intuitively explain why the point (V, W) is not uniformly distributed over the unit circle?

Problem 11.11 Let (X, Y) be a point chosen at random inside the unit circle. Define V and W by $V = X\sqrt{-2\ln(Q)/Q}$ and $W = Y\sqrt{-2\ln(Q)/Q}$, where $Q = X^2 + Y^2$. Verify that the random variables V and W are independent and $N(0, 1)$ distributed. This method for generating normal variates is known as Marsaglia's polar method.

Problem 11.12 The independent random variables Z and Y have a standard normal distribution and a chi-square distribution with ν degrees of freedom. Use the transformation $V = Y$ and $W = Z/\sqrt{Y/\nu}$ to prove that the random variable $W = Z/\sqrt{Y/\nu}$ has a Student-t density with ν degrees of freedom. *Hint*: in evaluating $f_W(w)$ from $\int_0^\infty f_{V,W}(v, w)\,dv$, use the fact that the gamma density $\lambda^\alpha x^{\alpha-1}e^{-\lambda x}/\,\Gamma(\alpha)$ integrates to 1 over $(0, \infty)$.

11.4 Covariance and correlation coefficient

Let the random variables X and Y be defined on the same sample space with probability measure P. A basic rule in probability is that the expected value of the sum $X + Y$ equals the sum of the expected values of X and Y. Does a similar rule hold for the variance of the sum $X + Y$? To answer this question, we apply the definition of variance. The variance of $X + Y$ equals

$$\begin{aligned}&E[\{X + Y - E(X + Y)\}^2]\\&\quad = E[(X - E(X))^2 + 2(X - E(X))(Y - E(Y)) + (Y - E(Y))^2]\\&\quad = \text{var}(X) + 2E[(X - E(X))(Y - E(Y))] + \text{var}(Y).\end{aligned}$$

This leads to the following general definition.

Definition 11.2 *The covariance* $cov(X, Y)$ *of two random variables* X *and* Y *is defined by*

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

whenever the expectations exist.

The formula for $\text{cov}(X, Y)$ can be written in the equivalent form

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

by expanding $(X - E(X))(Y - E(Y))$ into $XY - XE(Y) - YE(X) + E(X)E(Y)$ and noting that the expectation is a linear operator. Using the fact that $E(XY) = E(X)E(Y)$ for independent random variables, the alternative formula for $\text{cov}(X, Y)$ has as direct consequence:

Rule 11.4 *If* X *and* Y *are independent random variables, then*

$$\text{cov}(X, Y) = 0.$$

However, the converse of this result is not always true. A simple example of two dependent random variables X and Y having covariance zero is given in Section 9.4. Another counterexample is provided by the random variables $X = Z$ and $Y = Z^2$, where Z has the standard normal distribution. Nevertheless, $\text{cov}(X, Y)$ is often used as a measure of the dependence of X and Y. The covariance appears over and over in practical applications (see the discussion in Section 5.2).

Using the definition of covariance and the above expression for $\text{var}(X + Y)$, we find the general rule:

Rule 11.5 *For any two random variables* X *and* Y

$$\text{var}(X + Y) = \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y).$$

If the random variables X and Y are independent, then

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

The units of $\mathrm{cov}(X,Y)$ are not the same as the units of $E(X)$ and $E(Y)$. Therefore, it is often more convenient to use the *correlation coefficient* of X and Y which is defined by

$$\rho(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}},$$

provided that $\mathrm{var}(X)>0$ and $\mathrm{var}(Y)>0$. The correlation coefficient is a dimensionless quantity with the property that

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

The reader is asked to prove this property in Problem 11.14. The random variables X and Y are said to be *uncorrelated* if $\rho(X,Y)=0$. Independent random variables are always uncorrelated, but the converse is not always true. If $\rho(X,Y)=\pm 1$, then Y is fully determined by X. In this case it can be shown that $Y=aX+b$ for constants a and b with $a\neq 0$.

The problem section of Chapter 5 contains several exercises on the covariance and correlation coefficient. Here are some more exercises.

Problem 11.13 The continuous random variables X and Y have the joint density $f(x,y)=4y^2$ for $0<x<y<1$ and $f(x,y)=0$ otherwise. What is the correlation coefficient of X and Y? Can you intuitively explain why this correlation coefficient is positive?

Problem 11.14 Verify that

$$\mathrm{var}(aX+b)=a^2\mathrm{var}(X) \quad \text{and} \quad \mathrm{cov}(aX,bY)=ab\,\mathrm{cov}(X,Y)$$

for any constants a,b. Next, evaluate the variance of the random variable $Z=Y/\sqrt{\mathrm{var}(Y)}-\rho(X,Y)X/\sqrt{\mathrm{var}(X)}$ to prove that $-1\le\rho(X,Y)\le 1$. Also, for any constants a, b, c, and d, verify that $\mathrm{cov}(aX+bY,cV+dW)$ can be worked out as $ac\,\mathrm{cov}(X,V)+ad\,\mathrm{cov}(X,W)+bc\,\mathrm{cov}(Y,V)+bd\,\mathrm{cov}(Y,W)$.

Problem 11.15 The amounts of rainfall in Amsterdam during each of the months January, February, ..., December are independent random variables with expected values of 62.1, 43.4, 58.9, 41.0, 48.3, 67.5, 65.8, 61.4, 82.1, 85.1, 89.0, and 74.9 mm and with standard deviations of 33.9, 27.8, 31.1, 24.1, 29.3, 33.8, 36.8, 32.1, 46.6, 42.4, 40.0, and 36.2 mm. What are the expected value and the standard deviation of the annual rainfall in Amsterdam? Calculate an approximate value for the probability that the total rainfall in Amsterdam next year will be larger than 1,000 mm.

Problem 11.16 Let the random variables $X_1, \ldots, X_n$ be defined on a common probability space. Prove that

$$\text{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{var}(X_i) + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{cov}(X_i, X_j).$$

Next, evaluate $\text{var}(\sum_{i=1}^{n} t_i X_i)$ in order to verify that $\sum_{i=1}^{n} \sum_{j=1}^{n} t_i t_j \sigma_{ij} \geq 0$ for all real numbers $t_1, \ldots, t_n$, where $\sigma_{ij} = \text{cov}(X_i, X_j)$. In other words, the covariance matrix $C = (\sigma_{ij})$ is positive semi-definite.

Problem 11.17 The hypergeometric distribution describes the probability mass function of the number of red balls drawn when n balls are randomly chosen from an urn containing R red and W white balls. Show that the variance of the number of red balls drawn is given by $n\frac{R}{R+W}(1 - \frac{R}{R+W})\frac{R+W-n}{R+W-1}$. *Hint:* the number of red balls drawn can be written as $X_1 + \ldots + X_R$, where X_i equals 1 if the ith red ball is selected and 0 otherwise.

Problem 11.18 What is the variance of the number of distinct birthdays within a randomly formed group of 100 persons? *Hint*: define the random variable X_i as 1 if the ith day is among the 100 birthdays, and as 0 otherwise.

Problem 11.19 You roll a pair of dice. What is the correlation coefficient of the high and low points rolled?

Problem 11.20 What is the correlation coefficient of the Cartesian coordinates of a point picked at random in the unit circle?

11.4.1 Linear predictor

Suppose that X and Y are two dependent random variables. In statistical applications, it is often the case that we can observe the random variable X but we want to know the dependent random variable Y. A basic question in statistics is: what is the best *linear* predictor of Y with respect to X? That is, for which linear function $y = \alpha + \beta x$ is

$$E[(Y - \alpha - \beta X)^2]$$

minimal? The answer to this question is

$$y = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X),$$

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X = \sqrt{\text{var}(X)}$, $\sigma_Y = \sqrt{\text{var}(Y)}$, and $\rho_{XY} = \rho(X, Y)$. The derivation is simple. Rewriting $y = \alpha + \beta x$ as $y = \mu_Y + \beta(x - \mu_X) - (\mu_Y - \alpha - \beta\mu_X)$, it follows after some algebra that $E[(Y - \alpha - \beta X)^2]$

can be evaluated as

$$\begin{aligned}
&E[\{Y-\mu_Y-\beta(X-\mu_X)+\mu_Y-\alpha-\beta\mu_X\}^2] \\
&\quad = E[\{Y-\mu_Y-\beta(X-\mu_X)\}^2]+(\mu_Y-\alpha-\beta\mu_X)^2 \\
&\quad\quad +2(\mu_Y-\alpha-\beta\mu_X)E[Y-\mu_Y-\beta(X-\mu_X)] \\
&\quad = \sigma_Y^2+\beta^2\sigma_X^2-2\beta\rho_{XY}\sigma_X\sigma_Y+(\mu_Y-\alpha-\beta\mu_X)^2.
\end{aligned}$$

In order to minimize this quadratic function in α and β, we put the partial derivatives of the function with respect to α and β equal to zero. This leads after some simple algebra to

$$\beta=\frac{\rho_{XY}\sigma_Y}{\sigma_X} \quad \text{and} \quad \alpha=\mu_Y-\frac{\rho_{XY}\sigma_Y}{\sigma_X}\mu_X.$$

For these values of α and β, we have the minimal value

$$E\left[(Y-\alpha-\beta X)^2\right]=\sigma_Y^2\left(1-\rho_{XY}^2\right).$$

This minimum is sometimes called the residual variance of Y.

The phenomenon of *regression to the mean* can be explained with the help of the best linear predictor. Think of X as the height of a 25-year-old father and think of Y as the height his newborn son will have at the age of 25 years. It is reasonable to assume that $\mu_X=\mu_Y=\mu$, $\sigma_X=\sigma_Y=\sigma$, and $\rho=\rho(X,Y)$ is positive. The best linear predictor $\hat{Y}$ of Y then satisfies $\hat{Y}-\mu=\rho(X-\mu)$ with $0<\rho<1$. If the height of the father scores above the mean, the best linear prediction is that the height of the son will score closer to the mean. Very tall fathers tend to have somewhat shorter sons and very short fathers somewhat taller ones! Regression to the mean shows up in a wide variety of places: it helps explain why great movies have often disappointing sequels, and disastrous presidents have often better successors.

Index