Chapter 3

Random Vectors

3.1 Joint distribution and probabilities

Let $X_1, \ldots, X_p$ $(p \geqslant 1)$ be random variables. Then the p-dimensional vector $\mathbf{X} = (X_1, \ldots, X_p)$ is called a **random vector**. The joint range space, $R_{\mathbf{X}}$, of $\mathbf{X}$ is a region of p-dimensional space $(\mathbb{R}^p)$ such that $\mathbf{X}$ must always take some value inside $R_{\mathbf{X}}$.

The **joint distribution function** of the random vector $\mathbf{X}$ is the real-valued function $F_{\mathbf{X}}$ defined by:

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leqslant x_1, X_2 \leqslant x_2, \ldots, X_p \leqslant x_p) \qquad \mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p.$$

It will sometimes be convenient to write this probability in the form $F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leqslant \mathbf{x})$, where the vector inequality is implicitly taken to apply to the vectors element-by-element.

When all the random variables $X_1, X_2, \ldots, X_p$ are discrete, then we shall refer to $\mathbf{X}$ as a discrete random vector. In this case, the joint probability mass function of $\mathbf{X}$ is

$$p_{\mathbf{X}}(\mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) \qquad \mathbf{x} \in R_{\mathbf{X}}.$$

The values of $p_{\mathbf{X}}(\mathbf{x})$ must sum to 1 when added over $R_{\mathbf{X}}$.

Example 1

Adult female subjects are to have their distance vision tested in both eyes. Each subject will be given a grade of 1, 2 or 3 for each of her eyes (where 1 is the best grade). So, the experiment on a randomly-selected subject will generate an outcome $\mathbf{X} = (X_1, X_2)$ where X_1 = grade of distance vision in right eye and X_2 = grade of distance vision in left eye. $\mathbf{X}$ is a discrete random vector with

$$R_{\mathbf{X}} = \{(x_1, x_2) : x_1 = 1, 2, 3; x_2 = 1, 2, 3\}.$$

A very large survey of Scottish women has suggested the joint probability mass function shown in Table 3.1. Find the probability that a subject gets

(a) the best possible grade for her right eye and the worst possible grade for her left eye;

(b) the best possible grade for one eye and the worst possible grade for her other eye;

(c) the same grade for both eyes;

(d) a better grade for her right eye than for her left eye.

Table 3.1: Probability mass function for eye grade of both eyes.

$p_{\mathbf{X}}(x_1, x_2)$		x_1		
		1	2	3
	1	0.20	0.03	0.02
x_2	2	0.04	0.30	0.06
	3	0.01	0.07	0.27

Solution

Definition

The p-dimensional random vector $\mathbf{X}$ is said to be a **continuous random vector** if there exists a continuous function $f_{\mathbf{X}}(\mathbf{x})$, known as the **joint probability density function**, such that, for every $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$,

$$\int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_p} f_{\mathbf{X}}(\mathbf{t})\, dt_p \cdots dt_2 dt_1 = F_{\mathbf{X}}(\mathbf{x}).$$

A valid joint probability density function must satisfy the following condition:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{t})\, dt_p \cdots dt_2 dt_1 = F_{\mathbf{X}}(\infty, \infty, \ldots, \infty) = 1.$$

Also

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^p}{\partial x_1 \partial x_2 \cdots \partial x_p} F_{\mathbf{X}}(\mathbf{x}).$$

Now, since $F_{\mathbf{X}}(\mathbf{x})$ must be non-decreasing in every dimension, so that all these partial derivatives of $F_{\mathbf{X}}$ are non-negative everywhere, then a valid joint probability density function must also satisfy the condition that:

$$f_{\mathbf{X}}(\mathbf{x}) \geqslant 0 \qquad \text{for all } \mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p.$$

Example 2

Customers at a bank each have a waiting time, X_1 minutes, until a teller becomes available to serve them and a service time, X_2 minutes, which is how long it then takes the teller to serve them. A particular bank believes that, for its customers, the random vector (X_1, X_2) has the joint probability density function:

$$f_{\mathbf{X}}(\mathbf{x}) = 2e^{-x_1} e^{-2x_2} \qquad x_1 > 0, x_2 > 0.$$

(a) Show that this is a valid joint probability density function.

(b) Find the probability that a customer's service time is greater than his or her waiting time.

Solution

Example 3

In a satellite survey of land use in part of Scotland, each photographic frame captures a $1\,\text{km}^2$ area of land. The proportions of urban, farming and forestry land in a randomly-selected frame are denoted by the random variables X_1, X_2 and X_3, respectively. The random vector $\mathbf{X} = (X_1, X_2, X_3)$ has joint range space

$$R_{\mathbf{X}} = \{(x_1, x_2, x_3) : 0 \leqslant x_1, x_2, x_3 \leqslant 1; 0 \leqslant x_1 + x_2 + x_3 \leqslant 1\}.$$

Suppose that $\mathbf{X}$ is **uniformly** distributed on its range space. This means that there is a real constant $k > 0$ such that

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} k, & \mathbf{x} \in R_{\mathbf{X}}; \\ 0, & \text{otherwise.} \end{cases}$$

(a) What value of k makes this a valid (joint) probability density function?

(b) Find the probability that less than half the area in a randomly-selected frame is urban.

(c) Find the probability that, in total, less than half the area in a randomly-selected frame is urban, farming or forestry land.

Solution

3.2 Marginal distributions and moments

Definition

Suppose that $\mathbf{X} = (X_1, \ldots, X_p)$ is a random vector. Then the marginal distribution function of X_i $(i = 1, 2, \ldots, p)$ is the function

$$F_i(x_i) = P(X_i \leqslant x_i) = F_{\mathbf{X}}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty), \qquad x_i \in \mathbb{R}.$$

When $\mathbf{X}$ is a discrete random vector, the **marginal probability mass function** of X_i $(i = 1, \ldots, p)$ consists of the probabilities:

$$p_i(x_i) = P(X_i = x_i) = \underbrace{\sum \cdots \sum}_{\text{all } x \in R_{\mathbf{X}} \text{ such that } X_i = x_i} p_{\mathbf{X}}(\mathbf{x}).$$

Thus there are $p - 1$ sums, over $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$ (i.e., missing out x_i).

Example 1 (continued)

The marginal probability mass function of X_1 (grade of vision in the right eye) is:

$$\begin{aligned}
p_1(1) &= p_{\mathbf{X}}(1,1) + p_{\mathbf{X}}(1,2) + p_{\mathbf{X}}(1,3) = 0.20 + 0.04 + 0.01 = 0.25 \\
p_1(2) &= p_{\mathbf{X}}(2,1) + p_{\mathbf{X}}(2,2) + p_{\mathbf{X}}(2,3) = 0.03 + 0.30 + 0.07 = 0.40 \\
p_1(3) &= p_{\mathbf{X}}(3,1) + p_{\mathbf{X}}(3,2) + p_{\mathbf{X}}(3,3) = 0.02 + 0.06 + 0.27 = 0.35.
\end{aligned}$$

The marginal probability mass function of X_2 (grade of vision in the left eye) is:

$$\begin{aligned}
p_2(1) &= p_{\mathbf{X}}(1,1) + p_{\mathbf{X}}(2,1) + p_{\mathbf{X}}(3,1) = 0.20 + 0.03 + 0.02 = 0.25 \\
p_2(2) &= p_{\mathbf{X}}(1,2) + p_{\mathbf{X}}(2,2) + p_{\mathbf{X}}(3,2) = 0.04 + 0.30 + 0.06 = 0.40 \\
p_2(3) &= p_{\mathbf{X}}(1,3) + p_{\mathbf{X}}(2,3) + p_{\mathbf{X}}(3,3) = 0.01 + 0.07 + 0.27 = 0.35.
\end{aligned}$$

The expected value, variance and other (marginal) moments of the individual random variables within the random vector $\mathbf{X}$ can be found from the marginal probability mass functions as usual:

$$\begin{aligned} E(X_1) = E(X_2) &= 1 \times 0.25 + 2 \times 0.40 + 3 \times 0.35 = 2.10 \\ E(X_1^2) = E(X_2^2) &= 1^2 \times 0.25 + 2^2 \times 0.40 + 3^2 \times 0.35 = 5.00 \\ \text{Var}(X_1) = \text{Var}(X_2) &= 5.00 - 2.10^2 = 5.00 - 4.41 = 0.59. \end{aligned}$$

Definition

The expected value of the random vector $\mathbf{X} = (X_1, \ldots, X_p)$ is the vector

$$E(\mathbf{X}) \equiv \big(E(X_1), \ldots, E(X_p)\big).$$

In Example 1, then,

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix} = \begin{pmatrix} 2.1 \\ 2.1 \end{pmatrix}.$$

Definition

When $\mathbf{X}$ is a continuous random vector, the **marginal probability density function** of X_i $(i = 1, \ldots, p)$ is the function

$$f_i(x_i) = \frac{d}{dx_i} F_i(x_i) \qquad x_i \in \mathbb{R}.$$

Proposition 3.1

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a continuous random vector with joint probability density function $f_{\mathbf{X}}(\mathbf{x})$. Then the marginal probability density function of X_i is

$$f_i(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \ldots, x_i, \ldots, x_p)\, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_p.$$

Proof

$$\begin{aligned} f_i(x_i) &= \frac{d}{dx_i} F_i(x_i) \\ &= \frac{d}{dx_i} F_{\mathbf{X}}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty) \\ &= \frac{d}{dx_i} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{x_i} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{t})\, dt_1 \cdots dt_p \\ &= \frac{d}{dx_i} \int_{-\infty}^{x_i} \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{t})\, dt_1 \cdots dt_{i-1} dt_{i+1} \cdots dt_p \right) dt_i \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(t_1, \ldots, t_{i-1}, x_i, t_{i+1}, \ldots, t_p)\, dt_1 \cdots dt_{i-1} dt_{i+1} \cdots dt_p \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \ldots, x_i, \ldots, x_p)\, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_p. \end{aligned}$$

Example 2 (continued)

Recall

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} 2e^{-x_1}e^{-2x_2}, & x_1 > 0, x_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

So

$$f_1(x_1) = \int_0^\infty 2e^{-x_1}e^{-2x_2}dx_2 = e^{-x_1}\int_0^\infty 2e^{-2x_2}dx_2 = e^{-x_1}\left[-e^{-2x_2}\right]_0^\infty = e^{-x_1} \qquad x_1 > 0.$$

This means that $X_1 \sim \text{Expo}(1)$. So $E(X_1) = 1$ and $\text{Var}(X_1) = 1$. For x_2,

$$f_2(x_2) = \int_0^\infty 2e^{-x_1}e^{-2x_2}dx_1 = 2e^{-2x_2}\int_0^\infty e^{-x_1}dx_1 = 2e^{-x_2}\left[-e^{-x_1}\right]_0^\infty = 2e^{-2x_2} \quad x_2 > 0.$$

This means that $X_2 \sim \text{Expo}(2)$. So $E(X_2) = \frac{1}{2}$ and $\text{Var}(X_2) = \frac{1}{4}$. So

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}.$$

Example 3 (continued)

Find the marginal probability density function and distribution function of X_1, the proportion of urban land in a randomly-selected frame.

Solution

We have

$$f_{\mathbf{X}}(\mathbf{x}) = 6 \qquad \mathbf{x} \in R_{\mathbf{X}}.$$

So

$$\begin{aligned} f_1(x_1) &= \int_0^{1-x_1}\int_0^{1-x_1-x_2} 6\,dx_3dx_2 \\ &= 6\int_0^{1-x_1}(1-x_1-x_2)\,dx_2 \\ &= 6\left[(1-x_1)x_2 - \tfrac{1}{2}x_2^2\right]_{x_2=0}^{1-x_1} \\ &= 6\left((1-x_1)^2 - \tfrac{1}{2}(1-x_1)^2\right) \\ &= 3(1-x_1)^2 \qquad 0 < x_1 < 1. \end{aligned}$$

It follows that:

$$F_1(x_1) = \int_0^{x_1} f_1(t_1)\,dt_1 = \int_0^{x_1} 3(1-t_1)^2\,dt_1 = \left[-(1-t_1)^3\right]_0^{x_1} = 1-(1-x_1)^3, \qquad 0 \leqslant x_1 \leqslant 1.$$

Also

$$E(X_1) = \int_0^1 x_1 f_1(x_1)\,dx_1 = \int_0^1 x_1 3(1-x_1)^2\,dx_1 = 3\mathbb{B}(2,3) = 3\frac{1!2!}{4!} = \frac{1}{4}.$$

The symmetry of the joint range space and the joint probability density function show us that X_2 and X_3 must both have the same marginal distribution as X_1. So, in particular,

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ E(X_3) \end{pmatrix} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}.$$

It is not always the distribution of a single variable that interests us; often we want to know about the joint (marginal) distribution of a subset of the variables in the random vector $\mathbf{X}$. Partition $\mathbf{X}$ into k sub-vectors $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$ as shown below, where $\mathbf{X}^{(j)}$ has dimension p_j, and let $\mathbf{x} \in \mathbb{R}^p$ be partitioned conformably:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \\ \vdots \\ \mathbf{X}^{(k)} \end{pmatrix} \qquad \text{and} \quad \mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(k)} \end{pmatrix}.$$

Then the **joint marginal distribution function** of $\mathbf{X}^{(j)}$ is the function

$$F^{(j)}\big(\mathbf{x}^{(j)}\big) = P\big(\mathbf{X}^{(j)} \leqslant \mathbf{x}^{(j)}\big) \qquad \mathbf{X}^{(j)} \in \mathbb{R}^{p_j},$$

where the vector inequality applies element-by-element.

When $\mathbf{X}$ is a continuous random vector, the **(marginal) probability density function** of $\mathbf{X}^{(j)}$ is the function

$$f^{(j)}\big(\mathbf{x}^{(j)}\big) = \frac{\partial^{p_j}}{\partial \mathbf{x}^{(j)}} F^{(j)}\big(\mathbf{x}^{(j)}\big) \qquad \mathbf{x}^{(j)} \in \mathbb{R}^{p_j}.$$

We can obtain this function from the joint probability density function by integrating out the random variables that are not in $\mathbf{X}^{(j)}$, in the manner suggested by Proposition 3.1.

Example 3 (continued)

Find the joint marginal probability density function of (X_1, X_2), and hence the marginal probability density function of X_1.

Solution

$$\begin{aligned} f_{12}(x_1, x_2) &= \int_0^{1-x_1-x_2} 6 \, dx_3 \\ &= 6\Big[x_3\Big]_{x_3=0}^{1-x_1-x_2} \\ &= 6(1 - x_1 - x_2) \qquad 0 \leqslant x_1, x_2 \leqslant 1; x_1 + x_2 \leqslant 1. \end{aligned}$$

We can now obtain the marginal probability density function of X_1 from the joint probability density function of (X_1, X_2):

$$f_1(x_1) = \int_0^{1-x_1} f_{12}(x_1, x_2) \, dx_2 = 6\Big[(1 - x_1)x_2 - \tfrac{1}{2}x_2^2\Big]_0^{1-x_1} = 3(1 - x_1)^2 \qquad 0 \leqslant x_1 \leqslant 1.$$

This agrees with the answer obtained previously.

3.3 Conditional distributions, moments and regression

Definition

Suppose that the discrete random vector $\mathbf{X}$ is partitioned as $\begin{pmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{pmatrix}$. Then the **conditional probability mass function** of $\mathbf{X}^{(1)}$ given that $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ is defined by

$$p_{(1)|(2)}\left(\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)}\right) = \frac{p_{\mathbf{X}}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right)}{p_{\mathbf{X}^{(2)}}\left(\mathbf{x}^{(2)}\right)},$$

for all $\mathbf{x}^{(1)}$ such that $\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) \in R_{\mathbf{X}}$.

Similarly, the **conditional probability mass function** of $\mathbf{X}^{(2)}$ given that $\mathbf{X}^{(1)} = \mathbf{x}^{(1)}$ is defined by

$$p_{(2)|(1)}\left(\mathbf{x}^{(2)} \mid \mathbf{x}^{(1)}\right) = \frac{p_{\mathbf{X}}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right)}{p_{\mathbf{X}^{(1)}}\left(\mathbf{x}^{(1)}\right)},$$

for all $\mathbf{x}^{(2)}$ such that $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \in R_{\mathbf{X}}$.

Example 1 (continued)

A woman is to have her eyes tested and two pieces of information are to be recorded: X_1 = grade of distance vision in right eye and X_2 = grade of distance vision in left eye. The vector $\mathbf{X}$ has the joint probability mass function shown below:

$p_{\mathbf{X}}(x_1, x_2)$		x_1 (right eye)		
		1	2	3
	1	0.20	0.03	0.02
x_2 (left eye)	2	0.04	0.30	0.06
	3	0.01	0.07	0.27

The conditional probability mass function of X_1 given that $X_2 = 1$ is obtained by calculating all the conditional probabilities:

$$P(X_1 = x_1 | X_2 = 1) = \frac{P(X_1 = x_1, X_2 = 1)}{P(X_2 = 1)} = \frac{p_{\mathbf{X}}(x_1, 1)}{0.25}.$$

x_1	1	2	3
$p_{1\|2}(x_1\|1)$			

In a similar way, the following conditional probability mass functions are obtained for X_1 given that $X_2 = 2$ and $X_2 = 3$:

x_1	1	2	3
$p_{1\|2}(x_1\|2)$	$\frac{4}{40} = 0.10$	$\frac{30}{40} = 0.75$	$\frac{6}{40} = 0.15$

x_1	1	2	3
$p_{1\|2}(x_1\|3)$	$\frac{1}{35}$	$\frac{7}{35}$	$\frac{27}{35}$

Conditional moments can be obtained from the conditional probability mass function in the usual way. So, in Example 1,

$$E(X_1|X_2 = 1) =$$
$$E(X_1|X_2 = 2) = 1 \times 0.10 + 2 \times 0.75 + 3 \times 0.15 = 2.05$$
$$E(X_1|X_2 = 3) = 1 \times \tfrac{1}{35} + 2 \times \tfrac{7}{35} + 3 \times \tfrac{27}{35} = \tfrac{96}{35} = 2.74$$

Definition

Suppose that the continuous random vector $\mathbf{X}$ is partitioned as $\begin{pmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{pmatrix}$. Then the **conditional probability density function** of $\mathbf{X}^{(1)}$ given that $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ is defined by

$$f_{(1)|(2)}\left(\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)}\right) = \frac{f_{\mathbf{X}}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right)}{f_{\mathbf{X}^{(2)}}\left(\mathbf{x}^{(2)}\right)},$$

for all $\mathbf{x}^{(1)}$ such that $\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) \in R_{\mathbf{X}}$.

Similarly, the **conditional probability density function** of $\mathbf{X}^{(2)}$ given that $\mathbf{X}^{(1)} = \mathbf{x}^{(1)}$ is defined by

$$f_{(2)|(1)}\left(\mathbf{x}^{(2)} \mid \mathbf{x}^{(1)}\right) = \frac{f_{\mathbf{X}}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right)}{f_{\mathbf{X}^{(1)}}\left(\mathbf{x}^{(1)}\right)},$$

for all $\mathbf{x}^{(2)}$ such that $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \in R_{\mathbf{X}}$.

Example 2 (continued)

In this example, recall that

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} 2e^{-x_1}e^{-2x_2}, & x_1 > 0, x_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We have already derived the marginal p.d.f. $f_1(x_1)$:

$$f_1(x_1) = e^{-x_1} \qquad x_1 > 0.$$

Therefore,

$$f_{2|1}(x_2|x_1) = \frac{2e^{-x_1}e^{-2x_2}}{e^{-x_1}} = 2e^{-2x_2} \qquad x_2 > 0.$$

This conditional probability density function is exactly the same as the marginal probability density function of X_2 for every possible value x_1, in this case.

Example 3 (continued)

Here, we have seen that,

$$\begin{aligned} f_{\mathbf{X}}(x_1, x_2, x_3) &= 6 && 0 \leqslant x_1, x_2, x_3 \leqslant 1; x_1 + x_2 + x_3 \leqslant 1 \\ f_{23}(x_2, x_3) &= 6(1 - x_2 - x_3) && 0 \leqslant x_2, x_3 \leqslant 1; x_2 + x_3 \leqslant 1 \\ f_3(x_3) &= 3(1 - x_3)^2 && 0 \leqslant x_3 \leqslant 1 \end{aligned}$$

The conditional expected value of $\mathbf{X}^{(1)}$, written as a function of $\mathbf{X}^{(2)}$, is known as the **regression** of $\mathbf{X}^{(1)}$ on $\mathbf{X}^{(2)}$. This need not be a linear function, i.e., the regression need not be a **linear regression**. Estimation of the form of this regression function from experimental data is discussed in the Regression Models (Level M) course.

3.4 Iterated expectation and variance

Proposition 3.2: the Laws of Iterated Expectation and Variance

For any random vector $\mathbf{X} = (X_1, X_2)$,

(a) $E(X_2) = E[E(X_2|X_1)]$

(b) $Var(X_2) = E[Var(X_2|X_1)] + Var[E(X_2|X_1)]$

(as long as the required sums or integrals converge to finite limits).

Proof

This proof is for the continuous case only; other cases are similar.

(a) Using the definition of the conditional p.d.f.

$$\begin{aligned}
\mathrm{E}(X_2|X_1 = x_1) &= \int_{x_2:(x_1,x_2)\in R_X} x_2 f_{2|1}(x_2|x_1)\,dx_2 \\
&= \int_{x_2:(x_1,x_2)\in R_X} x_2 \frac{f_{12}(x_1,x_2)}{f_1(x_1)}\,dx_2 \\
&= \frac{1}{f_1(x_1)} \int_{x_2:(x_1,x_2)\in R_X} x_2 f_{12}(x_1,x_2)\,dx_2.
\end{aligned}$$

This is a function of x_1 (but not of x_2), so we can take its expectation (with respect to x_1):

$$\begin{aligned}
\mathrm{E}[\mathrm{E}(X_2|X_1)] &= \int_{x_1} \mathrm{E}(X_2|x_1) f_1(x_1)\,dx_1 \\
&= \int_{x_1} \left[\frac{1}{f_1(x_1)} \int_{x_2:(x_1,x_2)\in R_X} x_2 f_{12}(x_1,x_2)\,dx_2 \right] f_1(x_1)\,dx_1 \\
&= \int_{x_1} \int_{x_2} x_2 f_{12}(x_1,x_2)\,dx_2 dx_1 \\
&= \mathrm{E}(X_2).
\end{aligned}$$

(b) Recall that, by definition,

$$\mathrm{Var}(X_2) = \mathrm{E}\left([X_2 - \mathrm{E}(X_2)]^2\right) \tag{3.1}$$

and analogously for the conditional variance:

$$\mathrm{Var}(X_2|X_1 = x_1) = \mathrm{E}\left([X_2 - \mathrm{E}(X_2|X_1 = x_1)]^2 \middle| X_1 = x_1\right). \tag{3.2}$$

Using (3.1),

$$\begin{aligned}
\mathrm{Var}[\mathrm{E}(X_2|X_1)] &= \mathrm{E}\left(\left[\mathrm{E}(X_2|X_1) - \mathrm{E}(\mathrm{E}[X_2|X_1])\right]^2 \right) \\
&= \mathrm{E}\left(\left[\mathrm{E}(X_2|X_1) - \mathrm{E}(X_2)\right]^2 \right) \qquad \text{using (a).}
\end{aligned} \tag{3.3}$$

Now,

$$\begin{aligned}
[X_2 - \mathrm{E}(X_2)]^2 &= \left([X_2 - \mathrm{E}(X_2|X_1)] + [\mathrm{E}(X_2|X_1) - \mathrm{E}(X_2)]\right)^2 \\
&= [X_2 - \mathrm{E}(X_2|X_1)]^2 \\
&\quad + 2[X_2 - \mathrm{E}(X_2|X_1)][\mathrm{E}(X_2|X_1) - \mathrm{E}(X_2)] \\
&\quad + [\mathrm{E}(X_2|X_1) - \mathrm{E}(X_2)]^2.
\end{aligned}$$

Taking the expectation with respect to X_2 given X_1:

$$\begin{aligned}
\mathrm{E}\big([X_2 - \mathrm{E}(X_2)]^2 \,\big|\, X_1\big) &= \mathrm{E}\big([X_2 - \mathrm{E}(X_2|X_1)]^2 \,\big|\, X_1\big) \\
&\quad + 2[\mathrm{E}(X_2|X_1) - \mathrm{E}(X_2)]\mathrm{E}\big(X_2 - \mathrm{E}(X_2|X_1)]|X_1\big) \\
&\quad + [\mathrm{E}(X_2|X_1) - \mathrm{E}(X_2)]^2 \\
&= \mathrm{Var}(X_2|X_1) + 0 + [\mathrm{E}(X_2|X_1) - \mathrm{E}(X_2)]^2,
\end{aligned} \tag{3.4}$$

using (3.2). So,

$$\begin{aligned}
\mathrm{Var}(X_2) &= \mathrm{E}\big([X_2 - \mathrm{E}(X_2)]^2\big) \\
&= \mathrm{E}\big[\mathrm{E}\big([X_2 - \mathrm{E}(X_2)]^2 \,\big|\, X_1\big)\big] \qquad \text{using (a)} \\
&= \mathrm{E}[\mathrm{Var}(X_2|X_1)] + \mathrm{Var}[\mathrm{E}(X_2|X_1)],
\end{aligned}$$

from (3.4) and then (3.3).

Example 4

Let X_1 be the number of vehicles that travel along a certain stretch of motorway in a period of 5 minutes. At peak times, $X_1 \sim \text{Poi}(\lambda)$. A proportion θ of all the vehicles that travel along this road at peak times then leave the motorway at the next interchange. Given that $X_1 = x_1$, then X_2, the number of vehicles that go off at the interchange in a period of 5 minutes, follows a $\text{Bi}(x_1, \theta)$ distribution. Find the expected value and variance of X_2.

Solution

3.5 Covariance and correlation

Definition

Suppose that $\mathbf{X}$ is a p-dimensional random vector, and let $g(\mathbf{X})$ be any real-valued function of $\mathbf{X}$. Then, if it exists, the **expected value** of $g(\mathbf{X})$ is defined to be:

$$\mathrm{E}[g(\mathbf{X})] = \begin{cases} \displaystyle\sum_{x_1} \cdots \sum_{x_p} g(x_1, \ldots, x_p) p_{\mathbf{X}}(x_1, \ldots, x_p), & \mathbf{X} \text{ discrete;} \\ \displaystyle\int_{x_1} \cdots \int_{x_p} g(x_1, \ldots, x_p) f_{\mathbf{X}}(x_1, \ldots, x_p)\, dx_p \cdots dx_1, & \mathbf{X} \text{ continuous.} \end{cases}$$

Definition

Suppose that $\mathbf{X} = (X_1, \ldots, X_p)$ is a p-dimensional random vector with finite expected value $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$. Then the **covariance** of X_i and X_j $(i \neq j)$ is defined to be

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)].$$

It is easy to show that

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j), \\ \mathrm{Cov}(X_j, X_i) &= \mathrm{Cov}(X_i, X_j), \\ \mathrm{Cov}(X_i, X_i) &= \mathrm{Var}(X_i). \end{aligned}$$

Example 1 (continued): eyes

$p_{\mathbf{X}}(x_1, x_2)$		x_1		
		1	2	3
	1	0.20	0.03	0.02
x_2	2	0.04	0.30	0.06
	3	0.01	0.07	0.27

Example 2 (continued): bank

$$\begin{aligned}
\mathrm{E}(X_1X_2) &= \int_0^\infty \int_0^\infty x_1x_2 2e^{-x_1}e^{-2x_2}\,dx_2dx_1 \\
&= \int_0^\infty x_1e^{-x_1}dx_1 \int_0^\infty 2x_2e^{-2x_2}dx_2 \\
&= \mathrm{E}(X_1)\mathrm{E}(X_2) \\
\implies \mathrm{Cov}(X_1, X_2) &= \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\mathrm{E}(X_2) = 0.
\end{aligned}$$

Example 3 (continued): land use

$$\begin{aligned}
\mathrm{E}(X_1X_2) &= \int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} x_1x_2 6\,dx_3dx_2dx_1 \\
&= 6\int_0^1 x_1 \int_0^{1-x_1} x_2(1-x_1-x_2)\,dx_2dx_1 \\
&= 6\int_0^1 x_1 \tfrac{1}{6}(1-x_1)^3\,dx_1 \\
&= \int_0^1 x_1(1-x_1)^3\,dx_1 \\
&= \mathbb{B}(2,4) = \frac{\Gamma(2)\Gamma(4)}{\Gamma(6)} = \frac{1!3!}{5!} = \frac{1}{20} \\
\implies \mathrm{Cov}(X_1, X_2) &= \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\mathrm{E}(X_2) = \frac{1}{20} - \frac{1}{4}\times\frac{1}{4} = -\frac{1}{80}.
\end{aligned}$$

The symmetry of the joint probability density function means that we can deduce that

$$\mathrm{Cov}(X_1, X_3) = \mathrm{Cov}(X_2, X_3) = -\frac{1}{80}$$

too.

- A positive value of the covariance indicates that there is a positive relationship between the two random variables; i.e., higher values of one tend to be recorded along with higher values of the other.
- A negative value of the covariance suggests a negative relationship between the two random variables; i.e., higher values of one tend to be recorded along with lower values of the other.

In Example 3, the negative covariances are to be expected given the context. The random variables are all proportions of total land use, so the more land within a $1\,\mathrm{km}^2$ region that is used for a particular purpose (e.g., urban) the less that will be available for use for other purposes (e.g., farming or forestry).

It is easier to judge the strength of the relationship between two variables if we use the correlation, a measure of association based on the covariance, rather than the covariance itself.

Definition

Suppose that $\mathbf{X} = (X_1, \ldots, X_p)$ is a p-dimensional random vector where each X_i has finite expected value μ_i and finite variance σ_i^2. Then the **correlation** of X_i and X_j $(i \neq j)$

is defined to be

$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}}.$$

Clearly, ρ_{ij} takes the same sign as $\text{Cov}(X_i, X_j)$. However, it can be shown that the correlation must take values between -1 and 1. A value of -1 or 1 means that the two random variables are exactly (negatively or positively) linearly related.

Example 3 (continued)

Definition

The **covariance matrix** (or **variance-covariance matrix**) of the p-dimensional random vector $\mathbf{X}$ is the $p \times p$ matrix

$$\text{Cov}(\mathbf{X}) = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \cdots & \text{Var}(X_p) \end{bmatrix}.$$

Definition

The **correlation matrix** of the p-dimensional random vector $\mathbf{X}$ is the $p \times p$ matrix

$$\rho(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}.$$

Proposition 3.3

Suppose that $\mathbf{X}$ is a p-dimensional random vector with finite expected value $E(\mathbf{X})$. Then, for any p-dimensional vector of constants $\mathbf{a}$ and constant a_0,

$$E(a_0 + \mathbf{a}^T\mathbf{X}) = a_0 + \mathbf{a}^T E(\mathbf{X}),$$

where $\mathbf{X}$ (and indeed $E(\mathbf{X})$) now need to be interpreted as *column* vectors. That is,

$$E(a_0 + a_1X_1 + \cdots + a_pX_p) = a_0 + a_1E(X_1) + \cdots + a_pE(X_p).$$

Proof (for the continuous case)

$$\begin{aligned}
E(a_0 + \mathbf{a}^T\mathbf{X}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (a_0 + \mathbf{a}^T\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\
&= a_0 \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{a}^T\mathbf{x} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\
&= a_0 \times 1 + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\sum_{i=1}^{p} a_i x_i \right) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\
&= a_0 + \sum_{i=1}^{p} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} a_i x_i f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\
&= a_0 + \sum_{i=1}^{p} a_i \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\
&= a_0 + \sum_{i=1}^{p} a_i E(X_i) \\
&= a_0 + \mathbf{a}^T E(\mathbf{X}).
\end{aligned}$$

Proposition 3.4

Suppose that $\mathbf{X}$ is a p-dimensional random vector with finite expected value $E(\mathbf{X})$ and finite covariance matrix $\mathrm{Cov}(\mathbf{X})$. For any $q \times p$ matrix of real constants, A, and q-dimensional vector of real constants, $\mathbf{b}$, then

(a) $E(A\mathbf{X} + \mathbf{b}) = AE(\mathbf{X}) + \mathbf{b}$,

(b) $\mathrm{Cov}(A\mathbf{X} + \mathbf{b}) = A\mathrm{Cov}(\mathbf{X})A^T$.

Proof

See Tutorial Problems.

Definition

Suppose that $\mathbf{X}$ is a random vector and partition it into k sub-vectors $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$. Then $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$ are said to be **independent** if

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{k} F_{(j)}\left(\mathbf{x}^{(j)}\right), \qquad \text{for every } \mathbf{x} \text{ in real space.}$$

An equivalent definition in the discrete case is to say that $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$ are **independent** if

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{k} p_{(j)}(\mathbf{x}^{(j)}), \qquad \text{for every } \mathbf{x} \text{ in real space.}$$

In the continuous case, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$ are **independent** if and only if

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{k} f_{(j)}(\mathbf{x}^{(j)}), \qquad \text{for every } \mathbf{x} \text{ in real space.}$$

Proposition 3.5: the Factorisation Theorem for probability density functions

Let $\mathbf{X}$ be a p-dimensional random vector with joint probability density function $f_{\mathbf{X}}(\mathbf{x})$. Suppose $\mathbf{X}$ can be partitioned as before, and that for every $\mathbf{x} \in \mathbb{R}^p$,

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{k} g_{(j)}(\mathbf{x}^{(j)}),$$

where $g_{(j)}(\mathbf{x}^{(j)})$ is a real-valued function of $\mathbf{x}^{(j)}$ alone. Then, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$ are independent random vectors and the marginal probability density function of $\mathbf{X}^{(j)}$

$$f_{(j)}(\mathbf{x}^{(j)}) \propto g_{(j)}(\mathbf{x}^{(j)}).$$

Example 2 (continued): bank

If two random variables are independent, then the covariance (and, hence, the correlation) between them must be 0. On the other hand, two random variables might have correlation zero but not be independent (see tutorial example). We say that two random variables with zero correlation are **uncorrelated**, but note that uncorrelated does not necessarily mean independent!

Summary of important results about moments of a random vector

Let $X_1, \ldots, X_p$ be random variables, and let $a_0, \ldots, a_p$ be real constants. From Proposition 3.4 (see Tutorial Examples) it follows that (assuming the moments exist):

$$\mathrm{E}(a_0 + a_1X_1 + \cdots + a_pX_p) = a_0 + \sum_{i=1}^{p} a_i\mathrm{E}(X_i) \tag{3.5}$$

$$\mathrm{Var}(a_0 + a_1X_1 + \cdots + a_pX_p) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_ia_j\mathrm{Cov}(X_i, X_j). \tag{3.6}$$

Recall that:

$$\begin{aligned}\mathrm{Cov}(X_j, X_i) &= \mathrm{Cov}(X_i, X_j),\\ \mathrm{Cov}(X_i, X_i) &= \mathrm{Var}(X_i).\end{aligned}$$

Therefore, (3.6) can be rewritten in the form

$$\mathrm{Var}(a_0 + a_1X_1 + \cdots + a_pX_p) = \sum_{i=1}^{p} a_i^2\mathrm{Var}(X_i) + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_ia_j\mathrm{Cov}(X_i, X_j).$$

In the special case when $p = 2$, this means that

$$\begin{aligned}\mathrm{Var}(a_1X_1 + a_2X_2) &= a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2) + 2a_ia_j\mathrm{Cov}(X_1, X_2)\\ \mathrm{Var}(a_1X_1 - a_2X_2) &= a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2) - 2a_ia_j\mathrm{Cov}(X_1, X_2).\end{aligned}$$

If X_1 and X_2 are independent then their covariance is 0 and so

$$\begin{aligned}\mathrm{Var}(a_1X_1 + a_2X_2) &= a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2)\\ \mathrm{Var}(a_1X_1 - a_2X_2) &= a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2).\end{aligned}$$

Now let $Y_1, \ldots, Y_m$ also be random variables, and let $b_0, \ldots, b_m$ also be real constants. Then

$$\mathrm{Cov}(a_0 + a_1X_1 + \cdots + a_pX_p, b_0 + b_1Y_1 + \cdots + b_mY_m) = \sum_{i=1}^{p}\sum_{j=1}^{m} a_ib_j\mathrm{Cov}(X_i, Y_j).$$

In particular, when $m = p$ and each $Y_i = X_i$

$$\mathrm{Cov}(a_0 + a_1X_1 + \cdots + a_pX_p, b_0 + b_1X_1 + \cdots + b_pX_p) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_ib_j\mathrm{Cov}(X_i, X_j).$$

3.6 Functions of a random vector

Suppose that we wish to find the joint probability density function of $Y_1, \ldots, Y_p$, where each Y_i is a real-valued function of the p-dimensional random vector $\mathbf{X}$, i.e., $Y_i = h_i(\mathbf{X})$. We can write:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} h_1(\mathbf{X}) \\ \vdots \\ h_p(\mathbf{X}) \end{bmatrix} \equiv \mathbf{h}(\mathbf{X}).$$

We shall assume that the vector-valued function $\mathbf{h}$ is a one-to-one transformation, i.e., for every $\mathbf{y}$ there is a *unique* $\mathbf{x}$ such that $\mathbf{h}(\mathbf{x}) = \mathbf{y}$. The **Jacobian** of this transformation is defined to be the following $p \times p$ determinant:

$$J = \det\left(\frac{\partial \mathbf{X}}{\partial \mathbf{Y}}\right) = \det\begin{pmatrix} \frac{\partial X_1}{\partial Y_1} & \cdots & \frac{\partial X_1}{\partial Y_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial X_p}{\partial Y_1} & \cdots & \frac{\partial X_p}{\partial Y_p} \end{pmatrix}.$$

Proposition 3.6

Let $\mathbf{X}$ be a continuous, p-dimensional, random vector with joint range space $R_{\mathbf{X}}$ and joint probability density function $f_{\mathbf{X}}(\mathbf{x}), \mathbf{x} \in R_{\mathbf{X}}$. Define the new p-dimensional random vector $\mathbf{Y}$ by $\mathbf{Y} = \mathbf{h}(\mathbf{X})$, where every component $h_i(\cdot)$ of $\mathbf{h}(\cdot)$ is a continuous, real-valued function. Suppose that: $\mathbf{h}(\mathbf{X})$ is a one-to-one transformation with $\mathbf{X} = h^{-1}(\mathbf{Y})$; all the partial derivatives $\partial X_i/\partial Y_j$ exist and are continuous; and the Jacobian, J, of this transformation exists and is non-zero at all points in $R_{\mathbf{X}}$. Then the joint probability density function of the random vector $\mathbf{Y}$ is:

$$f_{\mathbf{Y}}(\mathbf{y}) = |J| \times f_{\mathbf{X}}\big(\mathbf{h}^{-1}(\mathbf{y})\big), \qquad \mathbf{y} \in R_{\mathbf{Y}}.$$

Note that $|\cdot|$ here indicates the absolute value.

Proof

Follows immediately from rules for the change of variables in integral calculus.

Example 5

Suppose that X_1 and X_2 are independent random variables, with $X_i \sim \mathrm{Ga}(\alpha_i, \theta)$ $(i = 1, 2)$. Notice that these two distributions have a common scale parameter, θ. Define new random variables Y_1 and Y_2 as follows

$$Y_1 = \frac{X_1}{X_1 + X_2} \qquad \text{and} \qquad Y_2 = X_1 + X_2.$$

Since X_1 and X_2 are independent, their joint p.d.f. is

$$\begin{aligned} f_{\mathbf{X}}(x_1, x_2) &= \frac{\theta^{\alpha_1}}{\Gamma(\alpha_1)} x_1^{\alpha_1 - 1} e^{-\theta x_1} \frac{\theta^{\alpha_2}}{\Gamma(\alpha_2)} x_2^{\alpha_2 - 1} e^{-\theta x_2} \\ &= \frac{\theta^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} e^{-\theta(x_1 + x_2)}, \qquad x_1 > 0, x_2 > 0. \end{aligned}$$

The transformation is one-to-one and can be inverted to give

$$X_1 = Y_1 Y_2 \qquad \text{and} \qquad X_2 = (1 - Y_1) Y_2$$

so

$$J = \det\begin{pmatrix} \frac{\partial X_1}{\partial Y_1} & \frac{\partial X_1}{\partial Y_2} \\ \frac{\partial X_2}{\partial Y_1} & \frac{\partial X_2}{\partial Y_2} \end{pmatrix} = \det\begin{pmatrix} Y_2 & Y_1 \\ -Y_2 & 1 - Y_1 \end{pmatrix} = Y_2.$$

All the conditions for applying Proposition 3.6 are met, so Y_1 and Y_2 have joint p.d.f.

$$\begin{aligned} f_{\mathbf{Y}}(y_1, y_2) &= y_2 \frac{\theta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} (y_1 y_2)^{\alpha_1-1}[(1-y_1)y_2]^{\alpha_2-1} e^{-\theta y_2} \qquad 0 < y_1 < 1; 0 < y_2 \\ &= \frac{\theta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y_1^{\alpha_1-1}(1-y_1)^{\alpha_2-1} \times y_2^{\alpha_1+\alpha_2-1} e^{-\theta y_2} \qquad 0 < y_1 < 1; 0 < y_2. \end{aligned}$$

Since the range space of (Y_1, Y_2) is a rectangular region, and since $f_{\mathbf{Y}}(y_1, y_2)$ may be factorised into a function of Y_1 alone and a function of Y_2 alone, it follows from the Factorisation Theorem that Y_1 and Y_2 are independent. Also:

$$\begin{aligned} f_1(y_1) &\propto y_1^{\alpha_1-1}(1-y_1)^{\alpha_2-1} && 0 < y_1 < 1; \\ f_2(y_2) &\propto y_2^{\alpha_1+\alpha_2-1} e^{-\theta y_2} && 0 < y_2. \end{aligned}$$

Therefore, $Y_1 \sim \text{Be}(\alpha_1, \alpha_2)$ and $Y_2 \sim \text{Ga}(\alpha_1 + \alpha_2, \theta)$.

Suppose now that X_1 and X_2 are independent and identically distributed Exponential random variables, $\text{Expo}(\theta)$. This is the special case of Example 5 where $\alpha_1 = \alpha_2 = 1$. Then,

$$\frac{X_1}{X_1 + X_2} \text{ has the } \text{Be}(1,1) \text{ or } \text{U}(0,1) \text{ distribution,}$$

while

$$X_1 + X_2 \text{ has the } \text{Ga}(2, \theta) \text{ distribution.}$$

Example 6

Suppose that X_1, X_2 are independent random variables with $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$. Find the joint probability density function of

$$Y_1 = \frac{X_1}{\nu_1} \qquad \text{and} \qquad Y_2 = \frac{X_2/\nu_2}{X_1/\nu_1}.$$

Hence find the marginal probability density function of Y_2, and show that $Y_2 \sim \text{F}(\nu_2, \nu_1)$.

Proof

Since X_1 and X_2 are independent,

$$\begin{aligned} f_{\mathbf{X}}(x_1, x_2) = f_1(x_1) f_2(x_2) &= \frac{x_1^{\nu_1/2-1} e^{-x_1/2}}{2^{\nu_1/2}\Gamma(\nu_1/2)} \frac{x_2^{\nu_2/2-1} e^{-x_2/2}}{2^{\nu_2/2}\Gamma(\nu_2/2)} \\ &= \frac{x_1^{\nu_1/2-1} x_2^{\nu_2/2-1} e^{-(x_1+x_2)/2}}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \end{aligned}$$

for $x_1 > 0$ and $x_2 > 0$. Inverting the transformation $(X_1, X_2) \to (Y_1, Y_2)$

$$X_1 = \nu_1 Y_1 \qquad \text{and} \qquad X_2 = \nu_2 Y_1 Y_2.$$

Also $R_{\mathbf{Y}} = \{(y_1, y_2) : 0 < y_1, 0 < y_2\}$.

$$J = \det \begin{pmatrix} \dfrac{\partial X_1}{\partial Y_1} & \dfrac{\partial X_1}{\partial Y_2} \\ \dfrac{\partial X_2}{\partial Y_1} & \dfrac{\partial X_2}{\partial Y_2} \end{pmatrix} = \det \begin{pmatrix} \nu_1 & 0 \\ \nu_2 Y_2 & \nu_2 Y_1 \end{pmatrix} = \nu_1 \nu_2 Y_1.$$

So

$$\begin{aligned} f_{\mathbf{Y}}(y_1, y_2) &= |J| \times f_{\mathbf{X}}(\nu_1 y_1, \nu_2 y_1 y_2) \\ &= \nu_1 \nu_2 y_1 \frac{(\nu_1 y_1)^{\nu_1/2-1} (\nu_2 y_1 y_2)^{\nu_2/2-1} e^{-(\nu_1 y_1 + \nu_2 y_1 y_2)/2}}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \\ &= \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2} y_1^{(\nu_1+\nu_2)/2-1} y_2^{\nu_2/2-1}\, e^{-y_1(\nu_1+\nu_2 y_2)/2}}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}. \end{aligned}$$

Integrating out y_1, for $y_2 > 0$:

$$\begin{aligned} f_2(y_2) &= \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2} y_2^{\nu_2/2-1}}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \int_0^\infty y_1^{(\nu_1+\nu_2)/2-1} e^{-y_1(\nu_1+\nu_2 y_2)/2}\, dy_1 \\ &= \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2} y_2^{\nu_2/2-1}}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \frac{2^{(\nu_1+\nu_2)/2}}{(\nu_1+\nu_2 y_2)^{(\nu_1+\nu_2)/2}} \Gamma\left(\tfrac{1}{2}(\nu_1+\nu_2)\right) \\ &= \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\mathbb{B}(\nu_2/2, \nu_1/2)} \frac{y_2^{\nu_2/2-1}}{(\nu_1+\nu_2 y_2)^{(\nu_1+\nu_2)/2}}. \end{aligned}$$

This is the p.d.f. of an $F(\nu_2, \nu_1)$ distribution, so

$$Y_2 \sim F(\nu_2, \nu_1).$$

3.7 The Multinomial and Multivariate Normal distributions

The multinomial distribution and its properties

The **multinomial distribution** is the generalisation to an arbitrary number of dimensions of the Binomial distribution considered in earlier chapters. Suppose that n objects are each independently to be placed in one of $p+1$ different categories, each object having probability θ_i of being placed in the ith category ($i = 1, \ldots, p+1$). This means that $0 \leqslant \theta_i \leqslant 1$ and that $\theta_1 + \cdots + \theta_{p+1} = 1$.

Let the random variable X_i denote the total number of objects placed in the ith category ($i = 1, \ldots, p+1$). Notice that $X_1 + X_2 + \cdots + X_{p+1} = n$, since every object must be placed in one and only one of the available categories. This means that only p of the random variables need to be considered explicitly, say $X_1, \ldots, X_p$, since the value of X_{p+1} may be deduced exactly from the values of the other random variables. The random vector $\mathbf{X} = (X_1, \ldots, X_p)$ is said to follow a Multinomial distribution, often written $\mathbf{X} \sim \mathrm{Mu}(n, \theta_1, \ldots, \theta_p)$. Note the restriction $\theta_1 + \ldots + \theta_p \leqslant 1$.

$\mathbf{X}$ has joint range space

$$R_{\mathbf{X}} = \{(x_1, \ldots, x_p) : x_1, \ldots, x_p = 0, 1, \ldots, n; x_1 + \ldots + x_p \leqslant n\}.$$

$\mathbf{X}$ has joint probability mass function

$$p_{\mathbf{X}}(x_1, \ldots, x_p) = \frac{n!}{x_1! \ldots x_p!(n - x_1 - \ldots - x_p)!} \times \theta_1^{x_1} \cdots \theta_p^{x_p}(1 - \theta_1 - \ldots - \theta_p)^{n - x_1 - \ldots - x_p}, \qquad (x_1, \ldots, x_p) \in R_{\mathbf{X}}.$$

The binomial distribution is the special case of the multinomial when $p = 1$, so $\mathrm{Bi}(n, \theta)$ is the same as $\mathrm{Mu}(n, \theta)$. Marginal probability mass functions can be obtained recursively. We first find the marginal distribution of $X_1, \ldots, X_{p-1}$, by summing out X_p. Notice that $(X_1, \ldots, X_{p-1})$ has the joint range space

$$\{(x_1, \ldots, x_{p-1}) : x_1, \ldots, x_{p-1} = 0, 1, \ldots, n; x_1 + \cdots + x_{p-1} \leqslant n\}.$$

On this range space, the marginal probability mass function is

$$\begin{aligned} p_{12\cdots(p-1)}(x_1, \ldots, x_{p-1}) &= \sum_{x_p=0}^{n - x_1 - \cdots - x_{p-1}} p_{\mathbf{X}}(\mathbf{x}) \\ &= \frac{n!}{x_1! \cdots x_{p-1}!} \theta_1^{x_1} \cdots \theta_{p-1}^{x_{p-1}} \\ &\quad \times \sum_{x_p=0}^{n - x_1 - \cdots - x_{p-1}} \frac{1}{x_p!(n - x_1 - \cdots - x_p)!} \theta_p^{x_p}(1 - \theta_1 - \cdots - \theta_p)^{n - x_1 - \cdots - x_p}. \end{aligned}$$

Aiming to use the Binomial Theorem, we rewrite the constant inside the summation

as a binomial coefficient, making the required changes outside the summation:

$$\begin{aligned}
p_{12\cdots(p-1)}(x_1,\ldots,x_{p-1}) &= \frac{n!}{x_1!\cdots x_{p-1}!(n-x_1-\cdots-x_{p-1})!}\theta_1^{x_1}\cdots\theta_{p-1}^{x_{p-1}} \\
&\quad\times\sum_{x_p=0}^{n-x_1-\cdots-x_{p-1}}\frac{(n-x_1-\cdots-x_{p-1})!}{x_p!(n-x_1-\cdots-x_p)!}\theta_p^{x_p}(1-\theta_1-\cdots-\theta_p)^{n-x_1-\cdots-x_p} \\
&= \frac{n!}{x_1!\cdots x_{p-1}!(n-x_1-\cdots-x_{p-1})!}\theta_1^{x_1}\cdots\theta_{p-1}^{x_{p-1}} \\
&\quad\times\sum_{x_p=0}^{n-x_1-\cdots-x_{p-1}}\binom{n-x_1-\cdots-x_{p-1}}{x_p}\theta_p^{x_p}(1-\theta_1-\cdots-\theta_p)^{n-x_1-\cdots-x_p} \\
&= \frac{n!}{x_1!\cdots x_{p-1}!(n-x_1-\cdots-x_{p-1})!} \\
&\qquad\times\theta_1^{x_1}\cdots\theta_{p-1}^{x_{p-1}}[\theta_p+(1-\theta_1-\cdots-\theta_p)]^{n-x_1-\cdots-x_{p-1}} \\
&= \frac{n!}{x_1!\cdots x_{p-1}!(n-x_1-\cdots-x_{p-1})!} \\
&\qquad\times\theta_1^{x_1}\cdots\theta_{p-1}^{x_{p-1}}(1-\theta_1-\cdots-\theta_{p-1})^{n-x_1-\cdots-x_{p-1}}.
\end{aligned}$$

In other words, the marginal distribution of $(X_1,\ldots,X_{p-1})$ is $\mathrm{Mu}(n,\theta_1,\ldots,\theta_{p-1})$.

This can be explained intuitively as follows. We began with $p+1$ categories of objects and explicitly modelled the counts in categories $1,2,\ldots,p$; category $p+1$ may be considered as a 'miscellaneous' category which is only considered implicitly. When we restrict our attention to the marginal distribution of $X_1,\ldots,X_{p-1}$, we are implicitly combining categories p and $p+1$ into a new 'miscellaneous' category.

Since we have now established that $(X_1,\ldots,X_{p-1})$ follows a $\mathrm{Mu}(n,\theta_1,\ldots,\theta_{p-1})$ distribution, we may use the above result to infer that $(X_1,\ldots,X_{p-2})$ follows a $\mathrm{Mu}(n,\theta_1,\ldots,\theta_{p-2})$ distribution. This argument may be continued until we establish that (X_1,X_2) marginally follows a $\mathrm{Mu}(n,\theta_1,\theta_2)$ distribution and X_1 marginally follows a $\mathrm{Bi}(n,\theta_1)$ distribution.

It is now relatively easy (see Tutorial Example sheet) to show that:

$$\mathrm{E}(\mathbf{X})=\begin{bmatrix}n\theta_1\\ n\theta_2\\ \vdots\\ n\theta_p\end{bmatrix}\qquad\text{and}\qquad \mathrm{Cov}(\mathbf{X})=\begin{bmatrix}n\theta_1(1-\theta_1) & -n\theta_1\theta_2 & \cdots & -n\theta_1\theta_p\\ -n\theta_1\theta_2 & n\theta_2(1-\theta_2) & \cdots & -n\theta_2\theta_p\\ \vdots & \vdots & \ddots & \vdots\\ -n\theta_1\theta_p & -n\theta_2\theta_p & \cdots & n\theta_p(1-\theta_p)\end{bmatrix}.$$

The multivariate normal distribution and its properties

Example 7

Consider the random vector $\mathbf{X}=(X_1,\ldots,X_p)$ where $X_1,\ldots,X_p$ are independent random variables and each $X_i\sim\mathrm{N}(\mu_i,\sigma_i^2)$. Then, because of the independence of the random variables, the joint probability density function of $\mathbf{X}$ is:

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right] \\
&= (2\pi)^{-p/2}\frac{1}{\sqrt{\prod_{i=1}^{p}\sigma_i^2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{p}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right].
\end{aligned}$$

The mean vector of $\mathbf{X}$ is $E(\mathbf{X}) = (\mu_1, \ldots, \mu_p) = \boldsymbol{\mu}$ (say). The covariance matrix of $\mathbf{X}$ is $\mathrm{Cov}(\mathbf{X}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2) = \Sigma$ (say).

This means that

$$\sum_{i=1}^{p} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

and $\prod_{i=1}^{p} \sigma_i^2 = \det(\Sigma)$. So the (joint) probability density function of $\mathbf{X}$ can be written in the form:

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-p/2} [\det(\Sigma)]^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right].$$

Definition

Suppose that a p-dimensional random vector, $\mathbf{X}$, has joint probability density function

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-p/2} [\det(\Sigma)]^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right],$$

where Σ is a positive definite $p \times p$ matrix. Then $\mathbf{X}$ is said to follow a (non-singular) **Multivariate Normal (MVN) distribution**, sometimes written $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$.

We shall show later that $\boldsymbol{\mu} = E(\mathbf{X})$ and $\Sigma = \mathrm{Cov}(\mathbf{X})$. This explains the restriction of Σ to positive definite matrices in the above definition, but there is no general requirement for Σ to be diagonal, as it was in Example 7. (In fact, it is even possible to extend the class of MVN distributions to include cases where Σ is positive semi-definite but not positive definite, so that Σ is singular and the joint probability density function does not exist. In this course, we will restrict attention to cases where Σ is positive definite and, hence, non-singular.)

Proposition 3.7

The p-dimensional random vector $\mathbf{X}$ has a MVN distribution $\Leftrightarrow \mathbf{a}^T \mathbf{X}$ has a (univariate) Normal distribution, for every p-dimensional vector $\mathbf{a}$. The proof is omitted.

In particular, Proposition 3.7 tells us that, when $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ it follows that each X_i is marginally distributed as a $N(\mu_i, \sigma_{ii})$ random variable (where σ_{ii} is the ith diagonal element of Σ). However, even when every X_i is marginally normally-distributed, it does not necessarily follow that $\mathbf{X}$ has a Multivariate Normal Distribution.

Proposition 3.8

Suppose that the p-dimensional random vector $\mathbf{X}$ follows the $N_p(\boldsymbol{\mu}, \Sigma)$ distribution. If A is a $q \times p$ matrix of constants, and $\mathbf{b}$ is a q-dimensional vector of constants, then

$$A\mathbf{X} + \mathbf{b} \sim N_q(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T).$$

If $\mathbf{X}$ follows a non-singular MVN distribution, then the distribution of $A\mathbf{X} + \mathbf{b}$ is also non-singular if and only if A has rank q. (This requires in particular that $q \leqslant p$.)

Corollary

If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, then any sub-vector of $\mathbf{X}$ has a (joint) marginal distribution that is also MVN. If $\mathbf{X}$ has a non-singular distribution, then so too has any sub-vector of $\mathbf{X}$.

Proof

We can permute the elements of $\mathbf{X}$ as we please and still obtain a MVN distribution, so we shall assume (without loss of generality) that the sub-vector whose distribution we wish to find is $\mathbf{X}^{(1)}$, consisting of the first r ($1 \leqslant r \leqslant p-1$) elements of $\mathbf{X}$.

We partition $\mathbf{X}$, $\boldsymbol{\mu}$ and Σ conformably as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Let $B = [I_r, 0_{r,p-r}]$, where I_r is the $r \times r$ identity matrix and $0_{r,p-r}$ is the $r \times (p-r)$ matrix of zeros. Then B is an $r \times p$ matrix of rank r. Also,

$$B\mathbf{X} = \mathbf{X}^{(1)} \qquad \text{and} \qquad B\boldsymbol{\mu} = \boldsymbol{\mu}^{(1)} \qquad \text{and} \qquad B\Sigma B^{\mathrm{T}} = \Sigma_{11}.$$

Then, by Proposition 3.8 (with $A = B$ and $\mathbf{b} = 0$), $\mathbf{X}^{(1)} \sim N_r(\boldsymbol{\mu}^{(1)}, \Sigma_{11})$. If Σ is non-singular (i.e., positive definite), so too is $B\Sigma B^{\mathrm{T}} = \Sigma_{11}$.

Proposition 3.9

Suppose that the random vector $\mathbf{X}$ follows the $N_p(\boldsymbol{\mu}, \Sigma)$ distribution, and partition $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix},$$

where as before $\mathbf{X}^{(1)}$ consists of the first r ($1 \leqslant r \leqslant p-1$) elements of $\mathbf{X}$. Then,

- the conditional distribution of $\mathbf{X}^{(2)}$ given $\mathbf{X}^{(1)} = \mathbf{x}^{(1)}$ is

$$N_{p-r}\Big(\boldsymbol{\mu}^{(2)} + \Sigma_{12}^{\mathrm{T}}\Sigma_{11}^{-1}(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}), \Sigma_{22} - \Sigma_{12}^{\mathrm{T}}\Sigma_{11}^{-1}\Sigma_{12}\Big)$$

- the conditional distribution of $\mathbf{X}^{(1)}$ given $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ is

$$N_r\Big(\boldsymbol{\mu}^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\mathrm{T}}\Big).$$

The proof is omitted.

If $\Sigma_{12} = 0_{r,p-r}$, then the conditional distribution of $\mathbf{X}^{(2)}$ given $\mathbf{X}^{(1)} = \mathbf{x}^{(1)}$ is the same as its marginal distribution for every choice of $\mathbf{x}^{(1)}$, so $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent random vectors. It follows that, when $\mathbf{X}_1$ and $\mathbf{X}_2$ are jointly normally distributed random variables, then they are independent if and only if they are uncorrelated. (Note: this is *not* true of random variables in general.)

Example 8

Suppose that the continuous random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 & -1 \\ -1 & 9 \end{bmatrix}\right).$$

(a) Identify the marginal distributions of X_1 and X_2. Write down $E(X_1)$, $\mathrm{Var}(X_1)$, $E(X_2)$, $\mathrm{Var}(X_2)$, $\mathrm{Cov}(X_1, X_2)$ and $\rho(X_1, X_2)$.

(b) Let $Y_1 = \frac{1}{2}X_1$ and $Y_2 = X_1 + 4X_2$.

  (i) Find $\mathrm{Cov}(\mathbf{Y})$.

  (ii) What is the distribution of Y_2?

  (iii) Are Y_1 and Y_2 independent? Justify your answer.

Solution