2
Basic Properties of Random Vectors

2.1 Cumulative Distribution Functions and Probability Density Functions

Let $\mathbf{x}=(x_1,\ldots,x_p)'$ be a random vector. By analogy with univariate theory, the *cumulative distribution function* (c.d.f.) associated with $\mathbf{x}$ is the function F defined by

$$F(\mathbf{x}^\circ)=\Pr(\mathbf{x}\leqslant\mathbf{x}^\circ)=\Pr(x_1\leqslant x_1^\circ,\ldots,x_p\leqslant x_p^\circ). \tag{2.1.1}$$

Two important cases are absolutely continuous and discrete distributions.

A random vector $\mathbf{x}$ is *absolutely continuous* if there exists a *probability density function* (p.d.f.), $f(\mathbf{x})$, such that

$$F(\mathbf{x})=\int_{-\infty}^{\mathbf{x}} f(\mathbf{u})\,d\mathbf{u}.$$

Here $d\mathbf{u}=du_1\ldots du_p$ represents the product of p differential elements and the integral sign denotes p-fold integration. Note that for any measurable set $D\subseteq R^p$,

$$P(\mathbf{x}\in D)=\int_D f(\mathbf{u})\,d\mathbf{u} \tag{2.1.2}$$

and

$$\int_{-\infty}^{\infty} f(\mathbf{u})\,d\mathbf{u}=1. \tag{2.1.3}$$

In this book we do not distinguish notationally between a random vector and its realization.

For a *discrete* random vector $\mathbf{x}$, the total probability is concentrated on

a countable (or finite) set of points $\{\mathbf{x}_j;\ j=1,2,\ldots\}$. Discrete random vectors can be handled in the above framework if we call the probability function

$$\left.\begin{aligned} f(\mathbf{x}_j)&=P(\mathbf{x}=\mathbf{x}_j), \qquad j=1,2,\ldots,\\ f(\mathbf{x})&=0, \qquad \text{otherwise,}\end{aligned}\right\} \tag{2.1.4}$$

the (*discrete*) *p.d.f.* of $\mathbf{x}$ and replace the integration in (2.1.2) by the summation

$$P(\mathbf{x}\in D)=\sum_{j:\,\mathbf{x}_j\in D} f(\mathbf{x}_j). \tag{2.1.5}$$

However, most of the emphasis in this book is directed towards absolutely continuous random vectors.

The support S of $\mathbf{x}$ is defined as the set

$$S=\{\mathbf{x}\in R^p: f(\mathbf{x})>0\}. \tag{2.1.6}$$

In examples the p.d.f. is usually defined only on S with the value zero elsewhere being assumed.

Marginal and conditional distributions Consider the partitioned vector $\mathbf{x}'=(\mathbf{x}_1',\mathbf{x}_2')$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ have k and $(p-k)$ elements, respectively $(k<p)$. The function

$$P(\mathbf{x}_1\leqslant\mathbf{x}_1^\circ)=F(x_1^\circ,\ldots,x_k^\circ,\infty,\ldots,\infty)$$

is called the *marginal cumulative distribution function* (marginal c.d.f.) of $\mathbf{x}_1$. In contrast $F(\mathbf{x})$ may then be described as the *joint c.d.f.* and $f(\mathbf{x})$ may be called the *joint p.d.f.*

Let $\mathbf{x}$ have joint p.d.f. $f(\mathbf{x})$. Then the *marginal p.d.f.* of $\mathbf{x}_1$ is given by the integral of $f(\mathbf{x})$ over $\mathbf{x}_2$; that is,

$$f_1(\mathbf{x}_1)=\int_{-\infty}^{\infty} f(\mathbf{x}_1,\mathbf{x}_2)\,d\mathbf{x}_2. \tag{2.1.7}$$

The marginal p.d.f. of $\mathbf{x}_2$, $f_2(\mathbf{x}_2)$, is defined similarly.

For a given value of $\mathbf{x}_1$, say, $\mathbf{x}_1=\mathbf{x}_1^\circ$, the *conditional p.d.f.* of $\mathbf{x}_2$ is proportional to $f(\mathbf{x}_1^\circ,\mathbf{x}_2)$, where the constant of proportionality can be calculated from the fact that this p.d.f. must integrate to one. In other words, the conditional p.d.f. of $\mathbf{x}_2$ given $\mathbf{x}_1=\mathbf{x}_1^\circ$ is

$$f(\mathbf{x}_2\mid\mathbf{x}_1=\mathbf{x}_1^\circ)=\frac{f(\mathbf{x}_1^\circ,\mathbf{x}_2)}{f_1(\mathbf{x}_1^\circ)}. \tag{2.1.8}$$

(It is assumed that $f_1(\mathbf{x}_1^\circ)$ is non-zero.) The conditional p.d.f. of $\mathbf{x}_1$ given $\mathbf{x}_2=\mathbf{x}_2^\circ$ can be defined similarly.

In general, two random variables can each have the same marginal distribution, even when their joint distributions are different. For instance, the marginal p.d.f.s of the following joint p.d.f.s,

$$f(x_1, x_2) = 1, \qquad 0 < x_1, x_2 < 1, \tag{2.1.9}$$

and (Morgenstern, 1956)

$$f(x_1, x_2) = 1 + \alpha(2x_1 - 1)(2x_2 - 1), \qquad 0 < x_1, x_2 < 1, \quad -1 \leq \alpha \leq 1, \tag{2.1.10}$$

are both uniform, although the two joint distributions are different.

Independence When the conditional p.d.f. $f(\mathbf{x}_2 \mid \mathbf{x}_1 = \mathbf{x}_1^\circ)$ is the same for all values of $\mathbf{x}_1^\circ$, then we say that $\mathbf{x}_1$ and $\mathbf{x}_2$ are *statistically independent* of each other. In such situations, $f(\mathbf{x}_2 \mid \mathbf{x}_1 = \mathbf{x}_1^\circ)$ must be $f_2(\mathbf{x}_2)$. Hence, the joint density must equal the product of the marginals, as stated in the following theorem.

Theorem 2.1.1 *If $\mathbf{x}_1$ and $\mathbf{x}_2$ are statistically independent then*

$$f(\mathbf{x}) = f_1(\mathbf{x}_1) f_2(\mathbf{x}_2).$$

Note that the variables x_1 and x_2 are independent for the p.d.f. given by (2.1.9), whereas the variables in (2.1.10) are dependent.

2.2 Population Moments

In this section we give the population analogues of the sample moments which were discussed in Section 1.4.

2.2.1 Expectation and correlation

If $\mathbf{x}$ is a random vector with p.d.f. $f(\mathbf{x})$ then the *expectation* or *mean* of a scalar-valued function $g(\mathbf{x})$ is defined as

$$E[g(\mathbf{x})] = \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}. \tag{2.2.1}$$

We assume that all necessary integrals converge, so that the expectations are finite. Expectations have the following properties:

(1) Linearity.

$$E[a_1 g_1(\mathbf{x}) + a_2 g_2(\mathbf{x})] = a_1 E[g_1(\mathbf{x})] + a_2 E[g_2(\mathbf{x})]. \tag{2.2.2}$$

(2) Partition, $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$. The expectation of a function of $\mathbf{x}_1$ may be

written in terms of the marginal distribution of $\mathbf{x}_1$ as follows:

$$E\{g(\mathbf{x}_1)\} = \int_{-\infty}^{\infty} g(\mathbf{x}_1) f(\mathbf{x})\, d\mathbf{x} = \int_{-\infty}^{\infty} g(\mathbf{x}_1) f_1(\mathbf{x}_1)\, d\mathbf{x}_1. \tag{2.2.3}$$

When f_1 is known, the second expression is useful for computation.

(3) If $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent and $g_i(\mathbf{x}_i)$ is a function of $\mathbf{x}_i$ *only* $(i = 1, 2)$, then

$$E[g_1(\mathbf{x}_1) g_2(\mathbf{x}_2)] = E[g_1(\mathbf{x}_1)] E[g_2(\mathbf{x}_2)]. \tag{2.2.4}$$

More generally, the expectation of a *matrix*-valued (or vector-valued) function of $\mathbf{x}$, $\mathbf{G}(\mathbf{x}) = (g_{ij}(\mathbf{x}))$ is defined to be the matrix

$$E[\mathbf{G}(\mathbf{x})] = (E[g_{ij}(\mathbf{x})]).$$

2.2.2 Population mean vector and covariance matrix

The vector $E(\mathbf{x}) = \boldsymbol{\mu}$ is called the *population mean vector* of $\mathbf{x}$. Thus,

$$\mu_i = \int_{-\infty}^{\infty} x_i f(\mathbf{x})\, d\mathbf{x}, \qquad i = 1, \ldots, p.$$

The population mean possesses the linearity property

$$E(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}E(\mathbf{x}) + \mathbf{b}, \tag{2.2.5}$$

where $\mathbf{A}(q \times p)$ and $\mathbf{b}(q \times 1)$ are constant.

The matrix

$$E\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\} = \boldsymbol{\Sigma} = V(\mathbf{x}) \tag{2.2.6}$$

is called the *covariance matrix* of $\mathbf{x}$ also known as the variance-covariance or dispersion matrix). For conciseness, write

$$\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.2.7}$$

to describe a random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

More generally we can define the covariance between two vectors, $\mathbf{x}(p \times 1)$ and $\mathbf{y}(q \times 1)$, by the $(p \times q)$ matrix

$$C(\mathbf{x}, \mathbf{y}) = E\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\nu})'\}, \tag{2.2.8}$$

where $\boldsymbol{\mu} = E(\mathbf{x})$, $\boldsymbol{\nu} = E(\mathbf{y})$. Notice the following simple properties of covariances. Let $V(\mathbf{x}) = \boldsymbol{\Sigma} = (\sigma_{ij})$.

(1) $\sigma_{ij} = C(x_i, x_j)$, $i \neq j$; $\sigma_{ii} = V(x_i) = \sigma_i^2$, say.

(2) $\boldsymbol{\Sigma} = E(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}'$. (2.2.9)

(3) $V(\mathbf{a}'\mathbf{x}) = \mathbf{a}' V(\mathbf{x}) \mathbf{a} = \sum a_i a_j \sigma_{ij}$, (2.2.10)

for all constant vectors $\mathbf{a}$. Since the left-hand side of (2.2.10) is always non-negative, we get the following result:

(4) $\boldsymbol{\Sigma} \geqslant 0$.

(5) $V(\mathbf{A}\mathbf{x}+\mathbf{b}) = \mathbf{A}V(\mathbf{x})\mathbf{A}'$. (2.2.11)

(6) $C(\mathbf{x}, \mathbf{x}) = V(\mathbf{x})$. (2.2.12)

(7) $C(\mathbf{x}, \mathbf{y}) = C(\mathbf{y}, \mathbf{x})'$.

(8) $C(\mathbf{x}_1+\mathbf{x}_2, \mathbf{y}) = C(\mathbf{x}_1, \mathbf{y}) + C(\mathbf{x}_2, \mathbf{y})$. (2.2.13)

(9) If $p = q$,

$$V(\mathbf{x}+\mathbf{y}) = V(\mathbf{x}) + C(\mathbf{x}, \mathbf{y}) + C(\mathbf{y}, \mathbf{x}) + V(\mathbf{y}). \tag{2.2.14}$$

(10) $C(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}) = \mathbf{A}C(\mathbf{x}, \mathbf{y})\mathbf{B}'$. (2.2.15)

(11) If x and y are independent then $C(\mathbf{x}, \mathbf{y}) = 0$. However the converse is *not* true. See Exercise 2.2.2.

Example 2.2.1 Let

$$f(x_1, x_2) = \begin{cases} x_1+x_2, & 0 \leqslant x_1, x_2 \leqslant 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2.16}$$

Then

$$\boldsymbol{\mu} = \begin{bmatrix} E(x_1) \\ E(x_2) \end{bmatrix} = \begin{bmatrix} 7/12 \\ 7/12 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 11/144 & -1/144 \\ -1/144 & 11/144 \end{bmatrix}.$$

Correlation matrix The population correlation matrix is defined in a manner similar to its sample counterpart. Let us denote the correlation coefficient between the ith and jth variables by ρ_{ij}, so that

$$\rho_{ij} = \sigma_{ij}/\sigma_i\sigma_j, \qquad i \neq j.$$

The matrix

$$\mathbf{P} = (\rho_{ij}) \tag{2.2.17}$$

with $\rho_{ii} = 1$ is called the *population correlation matrix*. Taking $\boldsymbol{\Delta} = \operatorname{diag}(\sigma_i)$, we have

$$\mathbf{P} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}.$$

The matrix $\mathbf{P} \geqslant 0$ because $\boldsymbol{\Sigma} \geqslant 0$, and $\boldsymbol{\Delta}$ is symmetric.

Generalized variance By analogy with Section 1.4.2, we may also define

the population *generalized variance* and *total variation* as $|\boldsymbol{\Sigma}|$ and $\operatorname{tr}\boldsymbol{\Sigma}$, respectively.

2.2.3 Mahalanobis space

We now turn to the population analogue of the Mahalanobis distance given by (1.6.1). If $\mathbf{x}$ and $\mathbf{y}$ are two points in space, then the *Mahalanobis distance* between $\mathbf{x}$ and $\mathbf{y}$, with metric $\boldsymbol{\Sigma}$, is the square root of

$$\Delta^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{y}). \tag{2.2.18}$$

(The subscript $\boldsymbol{\Sigma}$ may be omitted when there is no risk of confusion.) The matrix $\boldsymbol{\Sigma}$ is usually selected to be some convenient covariance matrix. Some examples are as follows.

(1) Let $\mathbf{x} \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and let $\mathbf{y} \sim (\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Then $\Delta(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ is a Mahalanobis distance between the parameters. It is invariant under transformations of the form

$$\mathbf{x} \to \mathbf{A}\mathbf{x}+\mathbf{b}, \qquad \mathbf{y} \to \mathbf{A}\mathbf{y}+\mathbf{b}, \qquad \boldsymbol{\Sigma} \to \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}',$$

where $\mathbf{A}$ is a non-singular matrix.

(2) Let $\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$, $\Delta(\mathbf{x}, \boldsymbol{\mu})$, is here a random variable.

(3) Let $\mathbf{x} \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\mathbf{y} \sim (\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. The Mahalanobis distance between $\mathbf{x}$ and $\mathbf{y}$ is $\boldsymbol{\Delta}(\mathbf{x}, \mathbf{y})$.

2.2.4 Higher moments

Following Section 1.8, a kth-order central moment for the variables $x_{i_1}, \ldots, x_{i_s}$ is

$$\mu^{(j_1,\ldots,j_s)}_{i_1,\ldots,i_s} = E\left\{\prod_{t=1}^{s}(x_{i_t} - \mu_{i_t})^{j_t}\right\},$$

where $j_1 + \ldots + j_s = k$, $j_t \neq 0$, $t = 1, \ldots, s$. Further, suitable population counterparts of the measures of multivariate skewness and kurtosis for random vector $\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are, respectively,

$$\beta_{1,p} = E\{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\}^3, \tag{2.2.19}$$

$$\beta_{2,p} = E\{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}^2, \tag{2.2.20}$$

where $\mathbf{x}$ and $\mathbf{y}$ are identically and independently distributed (see Mardia, 1970*a*). It can be seen that these measures are invariant under linear transformations.

Example 2.2.2 Consider the p.d.f. given by (2.1.10) which has uniform marginals on (0, 1). We have

$$E(x_1)=E(x_2)=\tfrac{1}{2}, \qquad \mu_{r,s}=E(x_1-\tfrac{1}{2})^r(x_2-\tfrac{1}{2})^s=\mu_r\mu_s+4\alpha\mu_{r+1}\mu_{s+1},$$

where

$$\mu_r=2^{-r-1}\{1+(-1)^r\}/(r+1),$$

which is the rth central moment for the uniform distribution on (0, 1). Let $\gamma_{rs}=\mu_{rs}/\sigma_1^r\sigma_2^s$. Then

$$\sigma_1^2=\sigma_2^2=\tfrac{1}{12},\ \rho=\tfrac{1}{3}\alpha,\ \gamma_{12}=\gamma_{03}=0, \qquad \gamma_{22}=1,\ \gamma_{13}=\tfrac{9}{5}\rho,\ \gamma_{40}=\tfrac{9}{5}.$$

Consequently, using these in the population analogue of Exercise 1.8.1, we have

$$\beta_{1,2}=0, \qquad \beta_{2,2}=4(7-13\rho^2)/\{5(1-\rho^2)^2\}.$$

2.2.5 Conditional moments

Moments of $\mathbf{x}_1\,|\,\mathbf{x}_2$ are called conditional moments. In particular, $E(\mathbf{x}_1\,|\,\mathbf{x}_2)$ and $V(\mathbf{x}_1\,|\,\mathbf{x}_2)$ are the conditional mean vector and the conditional variance-covariance matrix of $\mathbf{x}_1$ given $\mathbf{x}_2$. The *regression curve* of $\mathbf{x}_1$ on $\mathbf{x}_2$ is defined by the conditional expectation function

$$E(\mathbf{x}_1\,|\,\mathbf{x}_2),$$

defined on the support of $\mathbf{x}_2$. If this function is linear in $\mathbf{x}_2$, then the regression is called *linear*. The conditional variance function

$$V(\mathbf{x}_1\,|\,\mathbf{x}_2)$$

defines the *scedastic curve* of $\mathbf{x}_1$ on $\mathbf{x}_2$. The regression of $\mathbf{x}_1$ on $\mathbf{x}_2$ is called *homoscedastic* if $V(\mathbf{x}_1\,|\,\mathbf{x}_2)$ is a constant matrix.

Example 2.2.3 Consider the p.d.f. given by (2.2.16). The marginal density of x_2 is

$$f_2(x_2)=\int_0^1 (x_1+x_2)\,dx_1=x_2+\tfrac{1}{2}, \qquad 0<x_2<1.$$

Hence the regression curve of x_1 on x_2 is

$$E(x_1\,|\,x_2)=\int_0^1 x_1 f(x_1\,|\,x_2)\,dx_1=\int_0^1 \frac{x_1(x_1+x_2)}{x_2+\frac{1}{2}}\,dx_1=\frac{(3x_2+2)}{3(1+2x_2)}, \qquad 0<x_2<1.$$

This is a decreasing function of x_2, so the regression is not linear. Similarly,

$$E(x_1^a\,|\,x_2)=\left(\frac{1}{a+2}+\frac{1}{a+1}x_2\right)\Big/\left(\frac{1}{2}+x_2\right),$$

so that

$$V(x_1\,|\,x_2)=(1+6x_2+6x_2^2)/\{18\,(1+2x_2)^2\}, \qquad 0<x_2<1.$$

Hence the regression is not homoscedastic. ■

In general, if all the specified expectations exist, then

$$E(\mathbf{x}_1)=E\{E(\mathbf{x}_1\,|\,\mathbf{x}_2)\}. \tag{2.2.21}$$

However, note that the conditional expectations $E(\mathbf{x}_1\,|\,\mathbf{x}_2)$ may all be finite, even when $E(\mathbf{x})$ is infinite (see Exercise 2.2.6).

2.3 Characteristic Functions

Let $\mathbf{x}$ be a random p-vector. Then the characteristic function (c.f.) of $\mathbf{x}$ is defined as the function

$$\phi_{\mathbf{x}}(\mathbf{t})=E(e^{i\mathbf{t}'\mathbf{x}})=\int e^{i\mathbf{t}'\mathbf{x}}f(\mathbf{x})\,d\mathbf{x}, \qquad \mathbf{t}\in R^p. \tag{2.3.1}$$

As in the univariate case, we have the following properties.

(1) The characteristic function always exists, $\phi_{\mathbf{x}}(\mathbf{0})=1$, and $|\phi_{\mathbf{x}}(\mathbf{t})|\leqslant 1$.

(2) (Uniqueness theorem.) Two random vectors have the same c.f. if and only if they have the same distribution.

(3) (Inversion theorem.) If the c.f. $\phi_{\mathbf{x}}(t)$ is absolutely integrable, then $\mathbf{x}$ has a p.d.f. given by

$$f(\mathbf{x})=\frac{1}{(2\pi)^p}\int_{-\infty}^{\infty} e^{-i\mathbf{t}'\mathbf{x}}\phi_{\mathbf{x}}(\mathbf{t})\,d\mathbf{t}. \tag{2.3.2}$$

(4) Partition, $\mathbf{x}'=(\mathbf{x}_1',\mathbf{x}_2')$. The random vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent if and only if their joint c.f. factorizes into the product of their respective marginal c.f.s; that is if

$$\phi_{\mathbf{x}}(\mathbf{t})=\phi_{\mathbf{x}_1}(\mathbf{t}_1)\phi_{\mathbf{x}_2}(\mathbf{t}_2), \tag{2.3.3}$$

where $\mathbf{t}' = (\mathbf{t}_1', \mathbf{t}_2')$.

(5) $$E(x_1^{j_1} \dots x_p^{j_p}) = \frac{1}{i^{j_1+\dots+j_p}} \left\{ \frac{\partial^{j_1+\dots+j_p}}{\partial t_1^{j_1} \dots \partial t_p^{j_p}} \phi_{\mathbf{x}}(\mathbf{t}) \right\}_{\mathbf{t}=\mathbf{0}}$$

when this moment exists. (For a proof, differentiate both sides of (2.3.1) and put $\mathbf{t} = \mathbf{0}$.)

(6) The c.f. of the marginal distribution of $\mathbf{x}_1$ is simply $\phi_{\mathbf{x}}(\mathbf{t}_1, \mathbf{0})$.

(7) If $\mathbf{x}$ and $\mathbf{y}$ are independent random p-vectors then the c.f. of the *sum* $\mathbf{x}+\mathbf{y}$ is the *product* of the c.f.s of $\mathbf{x}$ and $\mathbf{y}$,

$$\phi_{\mathbf{x}+\mathbf{y}}(\mathbf{t}) = \phi_{\mathbf{x}}(\mathbf{t})\phi_{\mathbf{y}}(\mathbf{t}).$$

(To prove this, notice that independence implies $E(e^{i\mathbf{t}'(\mathbf{x}+\mathbf{y})}) = E(e^{i\mathbf{t}'\mathbf{x}})E(e^{i\mathbf{t}'\mathbf{y}})$.)

Example 2.3.1 Let $a(x_1, x_2)$ be a continuous positive function on an open set $S \subseteq R^2$. Let

$$f(x_1, x_2) = a(x_1, x_2)q(\theta_1, \theta_2) \exp\{x_1\theta_1 + x_2\theta_2\}, \qquad \mathbf{x} \in S,$$

be a density defined for $(\theta_1, \theta_2) \in \{(\theta_1, \theta_2) : 1/q(\theta_1, \theta_2) < \infty\}$ where

$$1/q(\theta_1, \theta_2) = \int a(x_1, x_2) \exp\{x_1\theta_1 + x_2\theta_2\}\, dx_1\, dx_2$$

is a normalization constant. Since this integral converges absolutely and uniformly over compact sets of (θ_1, θ_2), q can be extended by analytic continuation to give

$$\phi_{x_1,x_2}(t_1, t_2) = \int e^{it_1x_1 + it_2x_2} f(x_1, x_2)\, dx_1\, dx_2 = q(\theta_1, \theta_2)/q(\theta_1 + it_1, \theta_2 + it_2). \quad \blacksquare$$

We end this section with an important result.

Theorem 2.3.7 (Cramér–Wold) *The distribution of a random p-vector $\mathbf{x}$ is completely determined by the set of all one-dimensional distributions of linear combinations $\mathbf{t}'\mathbf{x}$, where $\mathbf{t} \in R^p$ ranges through all fixed p-vectors.*

Proof Let $y = \mathbf{t}'\mathbf{x}$ and let the c.f. of y be

$$\phi_y(s) = E[e^{isy}] = E[e^{is\mathbf{t}'\mathbf{x}}].$$

Clearly for $s = 1$, $\phi_y(1) = E[e^{i\mathbf{t}'\mathbf{x}}]$, which, regarded as a function of $\mathbf{t}$, is the c.f. of $\mathbf{x}$. $\blacksquare$

The Cramér–Wold theorem implies that a multivariate probability

distribution can be defined completely by specifying the distribution of all its linear combinations.

2.4 Transformations

Suppose that $f(\mathbf{x})$ is the p.d.f. of $\mathbf{x}$, and let $\mathbf{x} = \mathbf{u}(\mathbf{y})$ be a transformation from $\mathbf{y}$ to $\mathbf{x}$ which is one-to-one except possibly on sets of Lebesgue measure 0 in the supports of $\mathbf{x}$ and $\mathbf{y}$. Then the p.d.f. of $\mathbf{y}$ is

$$f\{\mathbf{u}(\mathbf{y})\}J, \tag{2.4.1}$$

where J is the Jacobian of the transformation from $\mathbf{y}$ to $\mathbf{x}$. It is defined by

$$J = \text{absolute value of } |\mathbf{J}|, \qquad \mathbf{J} = \left(\frac{\partial x_i}{\partial y_j}\right), \tag{2.4.2}$$

and we suppose that J is never zero or infinite except possibly on a set of Lebesgue measure 0. For some problems, it is easier to compute J from

$$J^{-1} = \text{absolute value of } \left|\frac{\partial y_j}{\partial x_i}\right| \tag{2.4.3}$$

using the inverse transformation $\mathbf{y} = \mathbf{u}^{-1}(\mathbf{x})$, and then substitute for $\mathbf{x}$ in terms of $\mathbf{y}$.

(1) *Linear transformation.* Let

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}, \tag{2.4.4}$$

where $\mathbf{A}$ is a non-singular matrix. Clearly, $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$. Therefore $\partial x_i/\partial y_j = a^{ij}$, and the Jacobian of the transformation $\mathbf{y}$ to $\mathbf{x}$ is

$$\text{Abs}\,|\mathbf{A}|^{-1}. \tag{2.4.5}$$

(2) *Polar transformation.* A generalization of the two-dimensional polar transformation

$$x = r\cos\theta, \qquad y = r\sin\theta, \qquad r > 0, \quad 0 \leq \theta < 2\pi,$$

to p dimensions is

$$\mathbf{x} = r\mathbf{u}(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} = (\theta_1, \dots, \theta_{p-1})', \tag{2.4.6}$$

where

$$u_i(\boldsymbol{\theta}) = \cos\theta_i \prod_{j=0}^{i-1} \sin\theta_j, \qquad \sin\theta_0 = \cos\theta_p = 1,$$

and

$$0 \leq \theta_j \leq \pi, \quad j = 1, \dots, p-2, \qquad 0 \leq \theta_{p-1} < 2\pi, \qquad r > 0.$$

Table 2.4.1 Jacobians of some transformations

Transformation $\mathbf{Y}$ to $\mathbf{X}$	Restriction	Jacobian (absolute value)
$\mathbf{X}=\mathbf{Y}^{-1}$	$\mathbf{Y}(p\times p)$ and non-singular (all elements random)	$\lvert\mathbf{Y}\rvert^{-2p}$
$\mathbf{X}=\mathbf{Y}^{-1}$	$\mathbf{Y}$ symmetric and non-singular	$\lvert\mathbf{Y}\rvert^{-p-1}$
$\mathbf{X}=\mathbf{AY}+\mathbf{B}$	$\mathbf{Y}(p\times p)$, $\mathbf{A}(p\times p)$ non-singular, $\mathbf{B}(p\times p)$	$\lvert\mathbf{A}\rvert^{p}$
$\mathbf{X}=\mathbf{AYB}$	$\mathbf{Y}(p\times q)$, $\mathbf{A}(p\times p)$, and $\mathbf{B}(q\times q)$ non-singular	$\lvert\mathbf{A}\rvert^{q}\,\lvert\mathbf{B}\rvert^{p}$
$\mathbf{X}=\mathbf{AYA}'$	$\mathbf{Y}(p\times p)$ symmetric, $\mathbf{A}(p\times p)$ non-singular	$\lvert\mathbf{A}\rvert^{p+1}$
$\mathbf{X}=\mathbf{YY}'$	$\mathbf{Y}$ lower triangular	$2^{p}\prod_{i=1}^{p} y_{ii}^{p+1-i}$

The Jacobian of the transformation from $(r, \boldsymbol{\theta})$ to $\mathbf{x}$ is

$$J = r^{p-1}\prod_{i=2}^{p-1}\sin^{p-i}\theta_{i-1}. \tag{2.4.7}$$

Note that the transformation is one to one except when $r=0$ or $\theta_i=0$ or π, for any $i=1,\ldots,p-2$.

(3) *Rosenblatt's transformation* (Rosenblatt, 1952). Suppose that $\mathbf{x}$ has p.d.f. $f(\mathbf{x})$ and denote the conditional c.d.f. of x_i given $x_1\ldots,x_{i-1}$ by $F(x_i \mid x_1,\ldots,x_{i-1})$, $i=1, 2,\ldots,p$. The Jacobian of the transformation $\mathbf{x}$ to $\mathbf{y}$, where

$$y_i = F(x_i \mid x_1,\ldots,x_{i-1}), \qquad i=1,\ldots,p, \tag{2.4.8}$$

is given by $f(x_1,\ldots,x_p)$. Hence, looking at the transformation $\mathbf{y}$ to $\mathbf{x}$, we see that $y_1,\ldots,y_p$ are independent identically distributed uniform variables on (0, 1).

Some other Jacobians useful in multivariate analysis are listed in Table 2.4.1. For their proof, see Deemer and Olkin (1951).

2.5 The Multinormal Distribution

2.5.1 Definition

In this section we introduce the most important multivariate probability distribution, namely the *multivariate normal* distribution. If we write the

p.d.f. of $N(\mu,\sigma^2)$, the univariate normal with mean μ and variance $\sigma^2>0$, as

$$f(x)=\{2\pi\sigma^2\}^{-1/2}\exp\{-\tfrac{1}{2}(x-\mu)\{\sigma^2\}^{-1}(x-\mu)\},$$

then a plausible extension to p variates is

$$f(\mathbf{x})=|2\pi\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}, \tag{2.5.1}$$

where $\boldsymbol{\Sigma}>0$. (Observe that the constant can be also written $\{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}\}^{-1}$.) Obviously, (2.5.1) is positive. It will be shown below in Theorem 2.5.1 that the total integral is unity, but first we give a formal definition.

Definition 2.5.1 *The random vector $\mathbf{x}$ is said to have a p-variate normal (or p-dimensional multinormal or multivariate normal) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ if its p.d.f. is given by* (2.5.1). *We write* $\mathbf{x}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

The quadratic form in 2.5.1 is equivalent to

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(x_i-\mu_i)(x_j-\mu_j), \qquad \boldsymbol{\Sigma}^{-1}=(\sigma^{ij}).$$

The p.d.f. may also be written in terms of correlations rather than covariances.

Theorem 2.5.1 *Let $\mathbf{x}$ have the p.d.f. given by* (2.5.1), *and let*

$$\mathbf{y}=\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}), \tag{2.5.2}$$

where $\boldsymbol{\Sigma}^{-1/2}$ is the symmetric positive-definite square root of $\boldsymbol{\Sigma}^{-1}$. Then $y_1,\ldots,y_p$ are independent $N(0,1)$ variables.

Proof From (2.5.2), we have

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\mathbf{y}'\mathbf{y}. \tag{2.5.3}$$

From (2.4.5) the Jacobian of the transformation $\mathbf{y}$ to $\mathbf{x}$ is $|\boldsymbol{\Sigma}|^{1/2}$. Hence, using (2.5.1), the p.d.f. of $\mathbf{y}$ is

$$g(\mathbf{y})=\frac{1}{(2\pi)^{p/2}}e^{-\Sigma y_i^2/2}. \quad \blacksquare$$

Note that since $g(\mathbf{y})$ integrates to 1, (2.5.1) is a density.

Corollary 2.5.1.1 *If $\mathbf{x}$ has the p.d.f. given by* (2.5.1) *then*

$$E(\mathbf{x})=\boldsymbol{\mu}, \qquad V(\mathbf{x})=\boldsymbol{\Sigma}. \tag{2.5.4}$$

Proof We have

$$E(\mathbf{y})=\mathbf{0}, \qquad V(\mathbf{y})=\mathbf{I}. \tag{2.5.5}$$

From (2.5.2),

$$\mathbf{x} = \boldsymbol{\Sigma}^{1/2}\mathbf{y} + \boldsymbol{\mu}. \tag{2.5.6}$$

Using Theorem 2.2.1, the result follows. ■

For $p = 2$, it is usual to write ρ_{12} as ρ, $-1 < \rho < 1$. In this case the p.d.f. becomes

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right\}\right],$$

where $-\infty < x_1, x_2 < \infty$.

2.5.2 Geometry

We now look at some of the above ideas geometrically. The multivariate normal distribution in p dimensions has constant density on ellipses or ellipsoids of the form

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2, \tag{2.5.7}$$

c being a constant. These ellipsoids are called the *contours* of the distribution or the "ellipsoids of equal concentrations". For $\boldsymbol{\mu} = \mathbf{0}$, these contours are centred at $\mathbf{x} = \mathbf{0}$, and when $\boldsymbol{\Sigma} = \mathbf{I}$ the contours are circles or in higher dimensions spheres or hyperspheres. Figure 2.5.1 shows a family of such contours for selected values of c for the bivariate case, and Figure 2.5.2 shows various types of contour for differing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The principal component transformation facilitates interpretation of the ellipsoids of equal concentration. Using the spectral decomposition theorem (Theorem A.6.4), write $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$, where $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_p)$ is the matrix of eigenvalues of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Gamma}$ is an orthogonal matrix whose columns are the corresponding eigenvectors. As in Section 1.5.3, define the *principal component transformation* by $\mathbf{y} = \boldsymbol{\Gamma}'(\mathbf{x}-\boldsymbol{\mu})$. In terms of $\mathbf{y}$, (2.5.7) becomes

$$\sum_{i=1}^{p} \frac{y_i^2}{\lambda_i} = c^2,$$

so that the components of $\mathbf{y}$ represent axes of the ellipsoid. This property is illustrated in Figure 2.5.1, where y_1 and y_2 represent the major and minor semi-axes of the ellipse, respectively.

2.5.3 Properties

If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the following results may be derived.

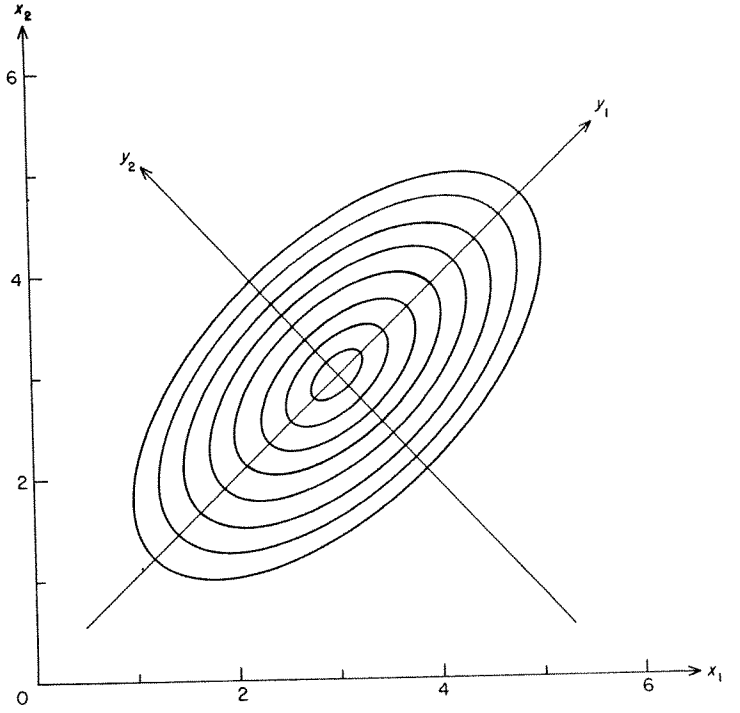

Figure 2.5.1 Ellipses of equal concentration for the bivariate normal distribution, showing the principal components y_1 *and* y_2, *where* $\boldsymbol{\mu}' = (3, 3)$, $\sigma_{11} = 3$, $\sigma_{12} = 1$, $\sigma_{22} = 3$.

Theorem 2.5.2

$$U = (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi_p^2. \tag{2.5.8}$$

Proof From (2.5.3) the left-hand side is $\sum y_i^2$, where the y_i' are independent $N(0, 1)$ by Theorem 2.5.1. Hence the result follows. ■

Using this theorem we can calculate the probability of a point $\mathbf{x}$ falling within an ellipsoid (2.5.7), from chi-square tables, since it amounts to calculating $\Pr(U \leq c^2)$.

Theorem 2.5.3 *The c.f. of* $\mathbf{x}$ *is*

$$\phi_{\mathbf{x}}(\mathbf{t}) = \exp\,(i\mathbf{t}'\boldsymbol{\mu} - \tfrac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}). \tag{2.5.9}$$

Proof Using (2.5.6), we find that

$$\phi_{\mathbf{x}}(\mathbf{t}) = E(e^{i\mathbf{t}'\mathbf{x}}) = e^{i\mathbf{t}'\boldsymbol{\mu}}E(e^{i\mathbf{u}'\mathbf{y}}), \tag{2.5.10}$$

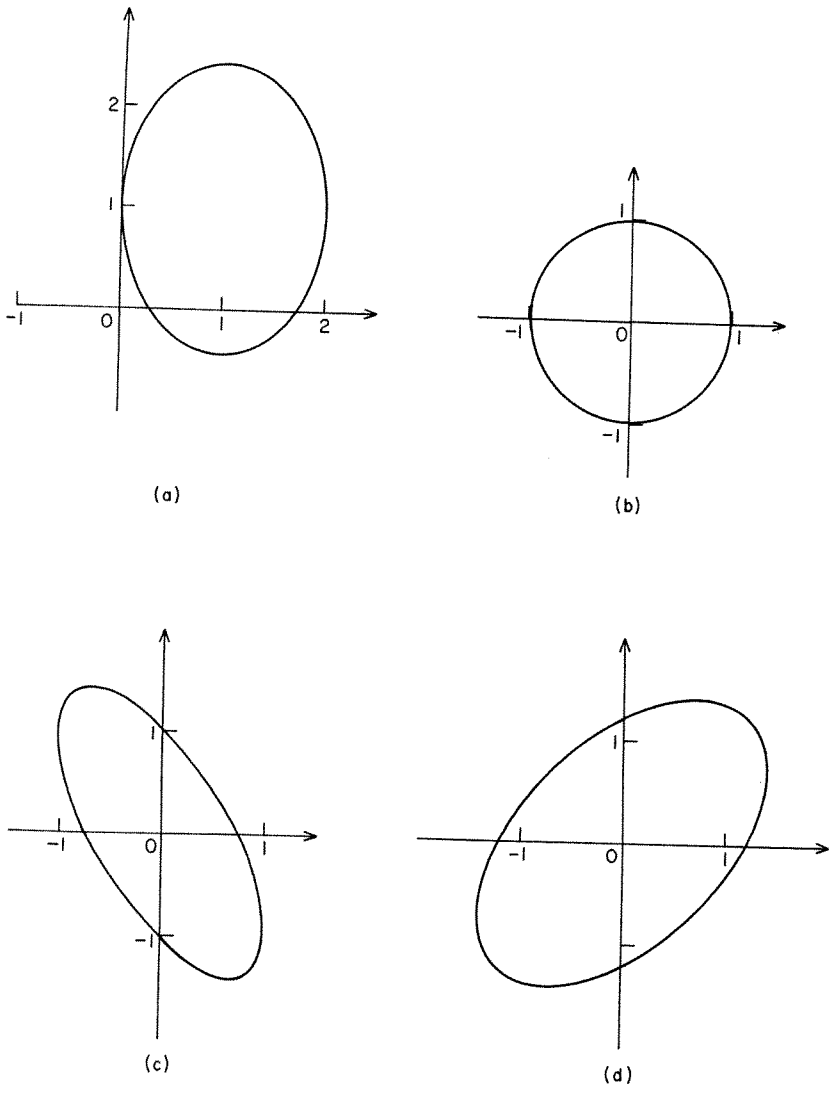

Figure 2.5.2 Ellipses of equal concentration for the bivariate normal distribution with $c = 1$.

(a) $\boldsymbol{\mu}' = (1, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. (b) $\boldsymbol{\mu}' = (0, 0)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

(c) $\boldsymbol{\mu}' = (0, 0)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$. (d) $\boldsymbol{\mu}' = (0, 0)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

where

$$\mathbf{u}' = \mathbf{t}'\boldsymbol{\Sigma}^{1/2}. \tag{2.5.11}$$

Since the y_i are independent $N(0, 1)$ from Theorem 2.5.1,

$$E(e^{i\mathbf{u}'\mathbf{y}}) = \prod_{i=1}^{p} \phi_{y_i}(u_i) = \prod_{i=1}^{p} e^{-u_i^2/2} = e^{-\mathbf{u}'\mathbf{u}/2}. \tag{2.5.12}$$

Substituting (2.5.12) and (2.5.11) in (2.5.10), we obtain the required result. ■

As an example of the use of c.f.s we prove the following result.

Theorem 2.5.4 *All non-trivial linear combinations of the elements of* $\mathbf{x}$ *are univariate normal.*

Proof Let $\mathbf{a} \neq \mathbf{0}$ be a p-vector. The c.f. of $y = \mathbf{a}'\mathbf{x}$ is

$$\phi_y(t) = \phi_{\mathbf{x}}(t\mathbf{a}) = \exp\{it\mathbf{a}'\boldsymbol{\mu} - \tfrac{1}{2}t^2\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}\},$$

which is the c.f. of a normal random variable with mean $\mathbf{a}'\boldsymbol{\mu}$ and variance $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} > 0$. ■

Theorem 2.5.5 $\beta_{1,p} = 0$, $\beta_{2,p} = p(p+2)$.

Proof Let $V = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, where $\mathbf{x}$ and $\mathbf{y}$ are i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, from (2.2.19), $\beta_{1,p} = E(V^3)$. However, the distribution of V is symmetric about $V = 0$, and therefore $E(V^3) = 0$. From (2.2.20) and (2.5.8),

$$\beta_{2,p} = E\{(\chi_p^2)^2\} = p(p+2). \quad ■$$

The multinormal distribution is explored in greater detail in Chapter 3 using a density-free approach.

*2.5.4 Singular multinormal distribution

The p.d.f. of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ involves $\boldsymbol{\Sigma}^{-1}$. However, if rank $(\boldsymbol{\Sigma}) = k < p$, we can define the (singular) density of $\mathbf{x}$ as

$$\frac{(2\pi)^{-k/2}}{(\lambda_1 \ldots \lambda_k)^{1/2}} \exp\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-}(\mathbf{x} - \boldsymbol{\mu})\}, \tag{2.5.13}$$

where

(1) $\mathbf{x}$ lies on the hyperplane $\mathbf{N}'(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0}$, where $\mathbf{N}$ is a $p \times (p-k)$ matrix such that

$$\mathbf{N}'\boldsymbol{\Sigma} = \mathbf{0}, \quad \mathbf{N}'\mathbf{N} = \mathbf{I}_{p-k}, \tag{2.5.14}$$

multinormal theory, as developed here. Again, other distributions may not have this property, e.g. linear functions of multivariate binary variables are not themselves binary.

(5) Even when the original data is not multinormal, one can often appeal to central limit theorems which prove that certain functions such as the sample mean are normal for large samples (Section 2.9).

(6) The equiprobability contours of the multinormal distribution are simple ellipses, which by a suitable change of coordinates can be made into circles (or, in the general case, hyperspheres). This geometric simplicity, together with the associated invariance properties, allows us to derive many crucial properties through intuitively appealing arguments.

In this chapter, unlike Section 2.5, we shall use a density-free approach, and try to emphasize the interrelationships between different distributions without using their actual p.d.f.s.

3.1.2 A definition by characterization

In this chapter we shall define the multinormal distribution with the help of the Cramér–Wold theorem (Theorem 2.3.7). This states that the multivariate distribution of any random p-vector $\mathbf{x}$ is completely determined by the univariate distributions of linear functions such as $\mathbf{a}'\mathbf{x}$, where $\mathbf{a}$ may be any non-random p-vector.

Definition 3.1.1 *We say that* $\mathbf{x}$ *has a* p-variate normal distribution *if and only if* $\mathbf{a}'\mathbf{x}$ *is univariate normal for all fixed* p*-vectors* $\mathbf{a}$. ■

To allow for the case $\mathbf{a}=\mathbf{0}$, we regard constants as degenerate forms of the normal distribution.

The above definition of multinormality has a useful geometric interpretation. If $\mathbf{x}$ is visualized as a random point in p-dimensional space, then linear combinations such as $\mathbf{a}'\mathbf{x}$ can be regarded as projections of $\mathbf{x}$ onto a one-dimensional subspace. Definition 3.1.1 therefore implies that the projection of $\mathbf{x}$ onto all one-dimensional subspaces has a univariate normal distribution. This geometric interpretation makes it clear that even after $\mathbf{x}$ is transformed by any arbitrary shift, rotation, or projection, it will still have the property of normality. In coordinate-dependent terms, this may be stated more precisely as follows. (In this theorem and others that follow we will assume that matrices and vectors such as $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{c}$ are non-random unless otherwise stated.)

Theorem 3.1.1 *If* $\mathbf{x}$ *has a* p*-variate normal distribution, and if* $\mathbf{y}=\mathbf{A}\mathbf{x}+\mathbf{c}$, *where* $\mathbf{A}$ *is any* $(q\times p)$ *matrix and* $\mathbf{c}$ *is any* q*-vector, then* $\mathbf{y}$ *has a* q*-variate normal distribution.* ■

Proof Let $\mathbf{b}$ be any fixed q-vector. Then $\mathbf{b}'\mathbf{y}=\mathbf{a}'\mathbf{x}+\mathbf{b}'\mathbf{c}$, where $\mathbf{a}=\mathbf{A}'\mathbf{b}$. Since $\mathbf{x}$ is multinormal, $\mathbf{a}'\mathbf{x}$ is univariate normal by Definition 3.1.1. Therefore $\mathbf{b}'\mathbf{y}$ is also univariate normal for all fixed vectors $\mathbf{b}$, and therefore $\mathbf{y}$ is multinormal by virtue of Definition 3.1.1. ■

Corollary 3.1.1.1 *Any subset of elements of a multinormal vector itself has a multinormal distribution. In particular the individual elements each have univariate normal distributions.* ■

Note that the above theorem and corollary need not assume that the covariance matrix $\mathbf{\Sigma}$ is of full rank. Therefore these results apply also to the singular multinormal distribution (Section 2.5.4). Also, the proofs do not use any intrinsic property of normality. Therefore similar results hold in principle for any other multivariate distribution defined in a similar way. That is, if we were to say that $\mathbf{x}$ has a p-variate "M" distribution whenever $\mathbf{a}'\mathbf{x}$ is univariate "M" for all fixed $\mathbf{a}$ ("M" could be "Cauchy") then results analogous to Theorem 3.1.1 and Corollary 3.1.1.1 could be derived.

However, before proceeding further, we must prove the *existence* of the multinormal distribution. This is done by showing that Definition 3.1.1 leads to the c.f. which has already been referred to in (2.5.9) and (2.5.17).

Theorem 3.1.2 *If* $\mathbf{x}$ *is multinormal with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\mathbf{\Sigma}(\mathbf{\Sigma}\geqslant 0)$, *then its c.f. is given by*

$$\phi_{\mathbf{x}}(\mathbf{t})=\exp\left(i\mathbf{t}'\boldsymbol{\mu}-\tfrac{1}{2}\mathbf{t}'\mathbf{\Sigma}\mathbf{t}\right). \tag{3.1.1}$$

Proof We follow the lines of the Cramér–Wold theorem (Theorem 2.3.7), and note that if $y=\mathbf{t}'\mathbf{x}$ then y has mean $\mathbf{t}'\boldsymbol{\mu}$ and variance $\mathbf{t}'\mathbf{\Sigma}\mathbf{t}$. Since y is univariate normal, $y\sim N(\mathbf{t}'\boldsymbol{\mu},\mathbf{t}'\mathbf{\Sigma}\mathbf{t})$. Therefore from the c.f. of the univariate normal distribution, the c.f. of y is

$$\phi_y(s)=E(\exp isy)=\exp\left(is\mathbf{t}'\boldsymbol{\mu}-\tfrac{1}{2}s^2\mathbf{t}'\mathbf{\Sigma}\mathbf{t}\right).$$

Hence the c.f. of $\mathbf{x}$ must be given by

$$\phi_{\mathbf{x}}(\mathbf{t})=E(\exp i\mathbf{t}'\mathbf{x})=E(\exp iy)=\phi_y(1)=\exp\left(i\mathbf{t}'\boldsymbol{\mu}-\tfrac{1}{2}\mathbf{t}'\mathbf{\Sigma}\mathbf{t}\right).$$

From Section 2.5, we see that (3.1.1) is indeed the c.f. of a multivariate

distribution. Hence the multinormal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ exists and its c.f. has the stated form. ■

As in Section 2.5, we may summarize the statement, "$\mathbf{x}$ is p-variate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$", by writing $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. When the dimension is clear we may omit the subscript p. We can obtain the p.d.f. when $\boldsymbol{\Sigma} > 0$ using the inversion formula (2.3.2), and it is given by (2.5.1).

Theorem 3.1.3 (a) *Two jointly multinormal vectors are independent if and only if they are uncorrelated*

(b) *For two jointly multinormal vectors, pair-wise independence of their components implies complete independence.* ■

Proof The c.f. given in Theorem 3.1.2 factorizes as required only when the corresponding submatrix of $\boldsymbol{\Sigma}$ is zero. This happens only when the vectors are uncorrelated. ■

3.2 Linear Forms

Theorem 3.1.1 proved that if $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{c}$, where $\mathbf{A}$ is any $(q \times p)$ matrix, then $\mathbf{y}$ has a q-variate normal distribution. Now from Section 2.2.2 we know that the moments of $\mathbf{y}$ are $\mathbf{A}\boldsymbol{\mu} + \mathbf{c}$ and $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$. Hence we deduce immediately the following results.

Theorem 3.2.1 *If* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{c}$, *then* $\mathbf{y} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. ■

Corollary 3.2.1.1 *If* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{\Sigma} > 0$, *then* $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I})$ *and* $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \sum y_i^2 \sim \chi_p^2$. ■

Corollary 3.2.1.2 *If* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ *and* $\mathbf{G}(q \times p)$ *is any row-orthonormal matrix, i.e. satisfying* $\mathbf{G}\mathbf{G}' = \mathbf{I}_q$, *then* $\mathbf{G}\mathbf{x} \sim N_q(\mathbf{G}\boldsymbol{\mu}, \sigma^2\mathbf{I})$. ■

Corollary 3.2.1.3 *If* $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{I})$ *and* $\mathbf{a}$ *is any non-zero* p*-vector, then* $\mathbf{a}'\mathbf{x}/\sqrt{\mathbf{a}'\mathbf{a}}$ *has the standard univariate normal distribution.* ■

Corollary 3.2.1.1 shows that any normal vector can easily be converted into standard form. It also gives an important quadratic expression which has a chi-squared distribution. From Corollary 3.2.1.2 we note that the standard multinormal distribution has a certain invariance under orthogonal transformations. Note that Corollary 3.2.1.3 also applies if $\mathbf{a}$ is a random vector independent of $\mathbf{x}$ (see Exercise 3.2.4). A further direct result of Theorem 3.1.3 is the following.

Theorem 3.2.2 *If* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then* $\mathbf{A}\mathbf{x}$ *and* $\mathbf{B}\mathbf{x}$ *are independent if and only if* $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$. ■

Corollary 3.2.2.1 *If* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ *and* $\mathbf{G}$ *is any row-orthonormal matrix, then* $\mathbf{G}\mathbf{x}$ *is independent of* $(\mathbf{I} - \mathbf{G}'\mathbf{G})\mathbf{x}$. ■

If $\mathbf{x}$ is partitioned into two subvectors, with r and s elements, respectively, then by noting two particular matrices which satisfy the conditions of Theorem 3.2.2, we may prove the following.

Theorem 3.2.3 *If* $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')' \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then* $\mathbf{x}_1$ *and* $\mathbf{x}_{2.1} = \mathbf{x}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1$ *have the following distributions and are statistically independent:*

$$\mathbf{x}_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \qquad \mathbf{x}_{2.1} \sim N_s(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1})$$

where

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1, \qquad \boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{3.2.1}$$

Proof We may write $\mathbf{x}_1 = \mathbf{A}\mathbf{x}$ where $\mathbf{A} = [\mathbf{I}, \mathbf{0}]$, and $\mathbf{x}_{2.1} = \mathbf{B}\mathbf{x}$ where $\mathbf{B} = [-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}, \mathbf{I}]$. Therefore, by Theorem 3.2.1, $\mathbf{x}_1$ and $\mathbf{x}_{2.1}$ are normal. Their moments are $\mathbf{A}\boldsymbol{\mu}$, $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$, $\mathbf{B}\boldsymbol{\mu}$, and $\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'$, which simplify to the given expressions. To prove independence note that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$, and use Theorem 3.2.2. ■

Similar results hold (using g-inverses) for the case of singular distributions. The above theorem can now be used to find the conditional distribution of $\mathbf{x}_2$ when $\mathbf{x}_1$ is known.

Theorem 3.2.4 *Using the assumptions and notation of Theorem* 3.2.3, *the conditional distribution of* $\mathbf{x}_2$ *for a given value of* $\mathbf{x}_1$ *is*

$$\mathbf{x}_2 \mid \mathbf{x}_1 \sim N_s(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22.1}).$$

Proof Since $\mathbf{x}_{2.1}$ is independent of $\mathbf{x}_1$, its conditional distribution for a given value of $\mathbf{x}_1$ is the same as its marginal distribution, which was stated in Theorem 3.2.3. Now $\mathbf{x}_2$ is simply $\mathbf{x}_{2.1}$ plus $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1$, and this term is constant when $\mathbf{x}_1$ is given. Therefore the conditional distribution of $\mathbf{x}_2 \mid \mathbf{x}_1$ is normal, and its conditional mean is

$$E[\mathbf{x}_2 \mid \mathbf{x}_1] = \boldsymbol{\mu}_{2.1} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1 = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1). \tag{3.2.2}$$

The conditional covariance matrix of $\mathbf{x}_2$ is the same as that of $\mathbf{x}_{2.1}$, namely $\boldsymbol{\Sigma}_{22.1}$. ■

If the assumption of normality is dropped from Theorem 3.2.3, then $\mathbf{x}_1$ and $\mathbf{x}_{2.1}$ still have the means and covariances stated. Instead of being independent of each other, however, all that can be said in general is that $\mathbf{x}_1$ and $\mathbf{x}_{2.1}$ are uncorrelated.

When $p=2$ and x_1 and x_2 are both scalars, then the expressions given above simplify. Putting σ_1^2, σ_2^2, and $\rho\sigma_1\sigma_2$ in place of $\mathbf{\Sigma}_{11}$, $\mathbf{\Sigma}_{22}$, and $\mathbf{\Sigma}_{12}$ we find that

$$\mathbf{\Sigma}_{22.1}=\sigma_2^2(1-\rho^2), \tag{3.2.3}$$

so that the conditional distribution of x_2 given x_1 is

$$x_2 \mid x_1 \sim N_1\{\mu_2+\rho\sigma_2\sigma_1^{-1}(x_1-\mu_1), \sigma_2^2(1-\rho^2)\}.$$

Example 3.2.1 If $\mathbf{\Sigma}$ is the equicorrelation matrix $\mathbf{\Sigma}=(1-\rho)\mathbf{I}+\rho\mathbf{1}\mathbf{1}'$, then the conditional distributions take a special form. Note that $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ are equicorrelation matrices of order $(r\times r)$ and $(s\times s)$, respectively. Also $\mathbf{\Sigma}_{12}=\rho\mathbf{1}_r\mathbf{1}_s'$ and $\mathbf{\Sigma}_{21}=\rho\mathbf{1}_s\mathbf{1}_r'$. Furthermore, we know from (A.3.2b) that

$$\mathbf{\Sigma}_{11}^{-1}=(1-\rho)^{-1}[\mathbf{I}-\alpha\mathbf{1}\mathbf{1}'], \qquad \alpha=\frac{\rho}{1+(r-1)\rho}.$$

Therefore, $\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}=\rho\mathbf{1}\mathbf{1}'\mathbf{\Sigma}_{11}^{-1}=\rho(1-\rho)^{-1}(1-\alpha r)\mathbf{1}\mathbf{1}'=\alpha\mathbf{1}\mathbf{1}'$. Substituting in (3.2.2) we find that the conditional mean is

$$E[\mathbf{x}_2 \mid \mathbf{x}_1]=\boldsymbol{\mu}_2+\alpha\mathbf{1}_r'(\mathbf{x}_1-\boldsymbol{\mu}_1)\mathbf{1}_s$$

and the conditional covariance matrix is

$$\mathbf{\Sigma}_{22.1}=(1-\rho)\mathbf{I}+\rho\mathbf{1}_s\mathbf{1}_s'-\rho\alpha\mathbf{1}_s\mathbf{1}_r'\mathbf{1}_r\mathbf{1}_s'=(1-\rho)\mathbf{I}+\rho(1-r\alpha)\mathbf{1}\mathbf{1}'.$$

Note that the conditional mean is just the original mean $\boldsymbol{\mu}_2$ with each element altered by the same amount. Moreover, this amount is proportional to $\mathbf{1}'(\mathbf{x}_1-\boldsymbol{\mu}_1)$, the sum of the deviations of the elements of $\mathbf{x}_1$ from their respective means. If $r=1$, then the conditional mean is

$$E(\mathbf{x}_2 \mid x_1)=\boldsymbol{\mu}_2+\rho(x_1-\mu_1)\mathbf{1}.$$

3.3 Transformations of Normal Data Matrices

Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be a random sample from $N(\boldsymbol{\mu},\mathbf{\Sigma})$. We call $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)'$ a data matrix from $N(\boldsymbol{\mu},\mathbf{\Sigma})$, or simply a "normal data matrix". In this section we shall consider linear functions such as $\mathbf{Y}=\mathbf{AXB}$, where $\mathbf{A}(m\times n)$ and $\mathbf{B}(p\times q)$ are fixed matrices of real numbers. The most important linear function is the sample mean $\bar{\mathbf{x}}'=n^{-1}\mathbf{1}'\mathbf{X}$, where $\mathbf{A}=n^{-1}\mathbf{1}'$ and $\mathbf{B}=\mathbf{I}_p$. The following result is immediate from Theorem 2.8.1.

Theorem 3.3.1 *If $\mathbf{X}(n\times p)$ is a data matrix from $N_p(\boldsymbol{\mu},\mathbf{\Sigma})$, and if $n\bar{\mathbf{x}}=\mathbf{X}'\mathbf{1}$, then $\bar{\mathbf{x}}$ has the $N_p(\boldsymbol{\mu},n^{-1}\mathbf{\Sigma})$ distribution.* ■

We may ask under what conditions $\mathbf{Y}=\mathbf{AXB}$ is itself a normal data matrix. Since $y_{ij}=\sum_{\alpha,\beta}a_{i\alpha}x_{\alpha\beta}b_{\beta j}$, clearly each element of $\mathbf{Y}$ is univariate normal. However, for $\mathbf{Y}$ to be a normal data matrix we also require (a) that the rows of $\mathbf{Y}$ should be independent of each other and (b) that each row should have the same distribution. The following theorem gives necessary and sufficient conditions on $\mathbf{A}$ and $\mathbf{B}$.

Theorem 3.3.2 *If $\mathbf{X}(n\times p)$ is a normal data matrix from $N_p(\boldsymbol{\mu},\mathbf{\Sigma})$ and if $\mathbf{Y}(m\times q)=\mathbf{AXB}$, then $\mathbf{Y}$ is a normal data matrix if and only if*

(a) $\mathbf{A}\mathbf{1}=\alpha\mathbf{1}$ *for some scalar* α, *or* $\mathbf{B}'\boldsymbol{\mu}=\mathbf{0}$, *and*
(b) $\mathbf{A}\mathbf{A}'=\beta\mathbf{I}$ *for some scalar* β, *or* $\mathbf{B}'\mathbf{\Sigma}\mathbf{B}=\mathbf{0}$.

When both these conditions are satisfied then $\mathbf{Y}$ is a normal data matrix from $N_q(\alpha\mathbf{B}'\boldsymbol{\mu},\beta\mathbf{B}'\mathbf{\Sigma}\mathbf{B})$. ■

Proof See Exercise 3.3.4. ■

To understand this theorem, note that post-multiplication of $\mathbf{X}$ involves adding weighted *variables*, while pre-multiplication of $\mathbf{X}$ adds weighted *objects*. Since the original objects (rows of $\mathbf{X}$) are independent, the transformed objects (rows of $\mathbf{Y}$) are also independent unless the pre-multiplication by $\mathbf{A}$ has introduced some interdependence. This clearly cannot happen when $\mathbf{A}$ is $k\mathbf{I}$, since then $\alpha=k$ and $\beta=k^2$, and both conditions of the theorem are satisfied. Similarly, all permutation matrices satisfy the conditions required on $\mathbf{A}$.

We may also investigate the correlation structure between two linear transformations of $\mathbf{X}$. Conditions for independence are stated in the following theorem.

Theorem 3.3.3 *If $\mathbf{X}$ is a data matrix from $N(\boldsymbol{\mu},\mathbf{\Sigma})$, and if $\mathbf{Y}=\mathbf{AXB}$ and $\mathbf{Z}=\mathbf{CXD}$, then the elements of $\mathbf{Y}$ are independent of the elements of $\mathbf{Z}$ if and only if either* (a) $\mathbf{B}'\mathbf{\Sigma}\mathbf{D}=\mathbf{0}$ *or* (b) $\mathbf{A}\mathbf{C}'=\mathbf{0}$. ■

Proof See Exercise 3.3.5. ■