Alvin C. Rencher: Methods of Multivariate Analysis

CHAPTER 4

The Multivariate Normal Distribution

4.1 MULTIVARIATE NORMAL DENSITY FUNCTION

Many univariate tests and confidence intervals are based on the univariate normal distribution. Similarly, the vast majority of multivariate procedures have as their underpinning the multivariate normal distribution.

The following are some of the useful features of the multivariate normal distribution: (1) only means, variances, and covariances need be estimated in order to completely describe the distribution; (2) bivariate plots show linear trends; (3) if the variables are uncorrelated, they are independent; (4) linear functions of multivariate normal variables are also normal; (5) as in the univariate case, the convenient form of the density function lends itself to derivation of many properties and test statistics; and (6) even when the data are not multivariate normal, the multivariable normal may serve as a useful approximation, especially in inferences involving sample mean vectors, which are approximately normal by the central limit theorem (see Section 4.3.2).

Since the multivariate normal density is an extension of the univariate normal density and shares many of its features, we review the univariate normal density function in Section 4.1.1. We then describe the multivariate normal density in Sections 4.1.2–4.1.4.

4.1.1 Univariate Normal Density

If a random variable y, with mean μ and variance σ^2, is normally distributed, its density is given by

$$f(y) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} \qquad -\infty < y < \infty. \tag{4.1}$$

When y has the density (4.1), we say that y is distributed as $N(\mu, \sigma^2)$, or simply

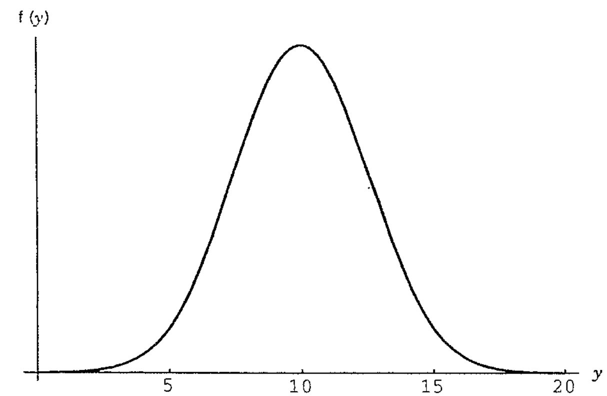


Figure 4.1 The normal density curve.

y is $N(\mu, \sigma^2)$. This function is represented by the familiar bell-shaped curve illustrated in Figure 4.1 for $\mu = 10$ and $\sigma = 2.5$.

4.1.2 Multivariate Normal Density

If $\mathbf{y}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the density is given by

$$g(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})/2}, \tag{4.2}$$

where p is the number of variables. When $\mathbf{y}$ has the density (4.2), we say that $\mathbf{y}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or simply $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The term $(y-\mu)^2/\sigma^2 = (y-\mu)(\sigma^2)^{-1}(y-\mu)$ in the exponent of the univariate normal density (4.1) measures the squared distance from y to μ in standard deviation units. Similarly, the term $(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ in the exponent of the multivariate normal density (4.2) is the squared generalized distance from $\mathbf{y}$ to $\boldsymbol{\mu}$, or the Mahalanobis distance,

$$\Delta^2 = (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}). \tag{4.3}$$

The characteristics of this distance between $\mathbf{y}$ and $\boldsymbol{\mu}$ were discussed in Section 3.12.

In the coefficient of the exponential function in (4.2), $|\boldsymbol{\Sigma}|^{1/2}$ appears as the analogue of $\sqrt{\sigma^2}$ in (4.1). In the next section, we discuss the effect of $|\boldsymbol{\Sigma}|$ on the density.

4.1.3 Generalized Population Variance

In Section 3.10, we referred to $|\mathbf{S}|$ as a generalized sample variance. Analogously, $|\boldsymbol{\Sigma}|$ is a *generalized population variance*. If σ^2 is small in the univariate normal, the y values are concentrated near the mean. Similarly, a small value of $|\boldsymbol{\Sigma}|$ in the multivariate case indicates that the $\mathbf{y}$'s are concentrated close to $\boldsymbol{\mu}$ in p-space or that there is multicollinearity among the variables. The term *multicollinearity* indicates that the variables are highly intercorrelated, in which case the effective dimensionality is less than p. (See Chapter 12 for a discussion of finding a reduced number of new dimensions that represent the data.) In the presence of multicollinearity, one or more eigenvalues of $\boldsymbol{\Sigma}$ will be near zero and $|\boldsymbol{\Sigma}|$ will be small, since $|\boldsymbol{\Sigma}|$ is the product of the eigenvalues, by (2.99).

Figure 4.2 shows, for the bivariate case, a comparison of a distribution with small $|\boldsymbol{\Sigma}|$ and a distribution with larger $|\boldsymbol{\Sigma}|$. An alternative way to portray the concentration of points in the bivariate normal distribution is with contour plots. Figure 4.3 shows contour plots for the two distributions in Figure 4.2. Each ellipse contains a different proportion of observation vectors $\mathbf{y}$. The contours in Figure 4.3 can be found by setting the density function equal to a constant and solving for $\mathbf{y}$, as illustrated in Figure 4.4. The bivariate normal density surface sliced at a constant height traces an ellipse, which contains a given proportion of the observations.

In both Figures 4.2 and 4.3, small $|\boldsymbol{\Sigma}|$ appears on the left and large $|\boldsymbol{\Sigma}|$ appears on the right. In Figure 4.3*a*, there is a larger correlation between y_1 and y_2. In Figure 4.3*b*, the variances are larger (in the natural directions). In general, for any number of variables p, a decrease in intercorrelations among the variables or an increase in the variances will lead to a larger $|\boldsymbol{\Sigma}|$.

4.1.4 Diversity of Applications of the Multivariate Normal

Nearly all the inferential procedures we discuss in this book are based on the multivariate normal distribution. We acknowledge that a major motivation for the widespread use of the multivariate normal is its mathematical tractability. From the multivariate normal assumption, a host of useful procedures can be derived. Practical alternatives to the multivariate normal are fewer than in the univariate case. Because it is not as simple to order (or rank) multivariate observation vectors as it is univariate observations, not as many nonparametric procedures are available for multivariate data.

While real data may not often be exactly multivariate normal, the multivariate normal will frequently serve as a useful approximation to the true distribution. Other reasons for our focus on the multivariate normal are the availability of tests and graphical procedures for assessing normality (see Sections 4.4 and 4.5) and the widespread use of procedures based on the multivariate normal in software packages. Fortunately, many of the procedures based on multivariate normality are robust to departures from normality.

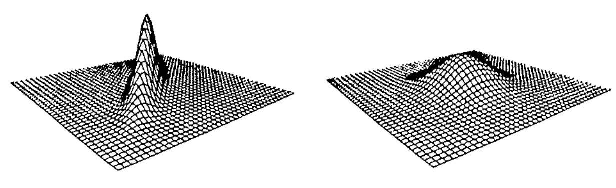


Figure 4.2 Bivariate normal densities.

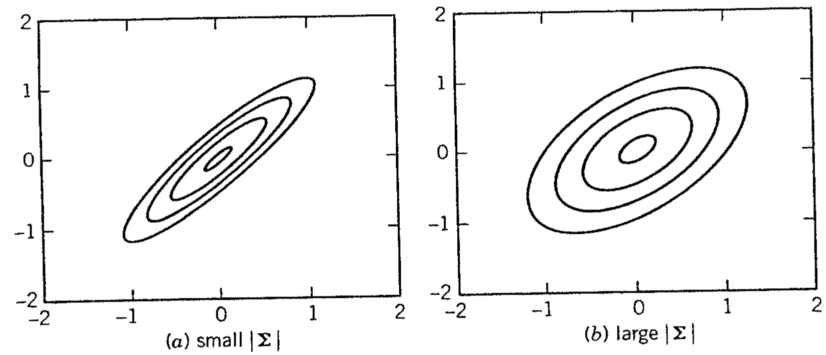


Figure 4.3 Contour plots for the distributions in Figure 4.2.

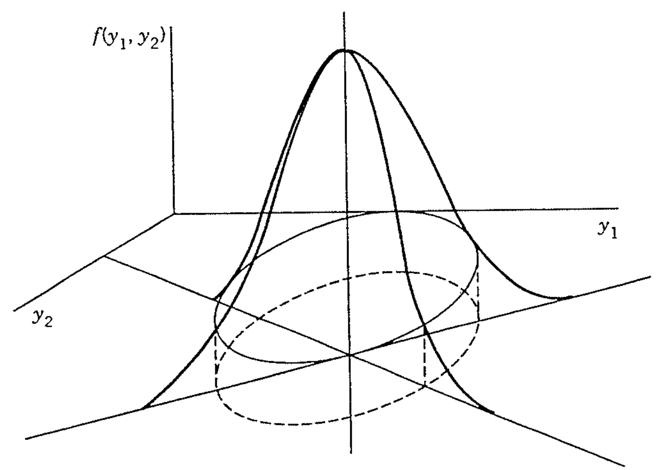


Figure 4.4 Constant density contour for bivariate normal.

4.2 PROPERTIES OF MULTIVARIATE NORMAL RANDOM VARIABLES

We list some of the properties of a random $p \times 1$ vector $\mathbf{y}$ from a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

1. Normality of linear combinations of the variables in $\mathbf{y}$

a. If $\mathbf{a}$ is a vector of constants, the linear function $\mathbf{a}'\mathbf{y} = a_1y_1 + a_2y_2 + \cdots + a_py_p$ is distributed as $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$. The mean and variance of $\mathbf{a}'\mathbf{y}$ were given previously in (3.64) and (3.65) as $E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\mu}$ and $\text{var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$ for any random vector $\mathbf{y}$. We now have the additional attribute that $\mathbf{a}'\mathbf{y}$ has a (univariate) normal distribution if $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This is a fundamental result, since we will often deal with linear combinations.

b. If $\mathbf{A}$ is a constant $q \times p$ matrix of rank q, where $q \leq p$, then $\mathbf{Ay}$ consists of q linear combinations of the variables in $\mathbf{y}$, with distribution $N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. Here, again, $E(\mathbf{Ay}) = \mathbf{A}\boldsymbol{\mu}$ and $\text{cov}(\mathbf{Ay}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$, in general, as given in (3.68) and (3.69). But we now have the additional feature that the q variables in $\mathbf{Ay}$ have a multivariate normal distribution.

2. Standardized variables

If $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a *standardized vector* $\mathbf{z}$ can be obtained in two ways:

$$\mathbf{z} = (\mathbf{T}')^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{4.4}$$

where $\boldsymbol{\Sigma} = \mathbf{T}'\mathbf{T}$ is factored using the Cholesky procedure in Section 2.7, or

$$\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{4.5}$$

where $\boldsymbol{\Sigma}^{1/2}$ is the symmetric square root matrix of $\boldsymbol{\Sigma}$ defined in (2.103) such that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$. In either (4.4) or (4.5), it follows from property 1b that $\mathbf{z}$ is distributed as $N_p(\mathbf{0}, \mathbf{I})$; that is, the z's are independently distributed as $N(0, 1)$. Thus in the multivariate case, a standardized vector of random variables has all means equal to 0, all variances equal to 1, *and* all correlations equal to 0.

3. Chi-square distribution

A *chi-square random variable* with p degrees of freedom is defined as the sum of squares of p independent standard normal random variables. Thus if $\mathbf{z}$ is the standardized vector defined in (4.4) or (4.5), then $\sum_{i=1}^{p} z_i^2 = \mathbf{z}'\mathbf{z}$ has the χ^2-distribution with p degrees of freedom, denoted χ_p^2 or $\chi^2(p)$. From either (4.4) or (4.5) we obtain $\mathbf{z}'\mathbf{z} = (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$. Hence

$$\text{If } \mathbf{y} \text{ is } N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ then } (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \text{ is } \chi_p^2. \tag{4.6}$$

4. Normality of marginal distributions

a. Any subset of the y's in $\mathbf{y}$ has a multivariate normal distribution, with mean vector consisting of the corresponding subvector of $\boldsymbol{\mu}$ and covariance matrix composed of the corresponding submatrix of $\boldsymbol{\Sigma}$. To illustrate, let $\mathbf{y}_1' = (y_1, y_2, \ldots, y_r)$ denote the subvector containing the first r elements of $\mathbf{y}$ and $\mathbf{y}_2' = (y_{r+1}, \ldots, y_p)$ consist of the remaining $p - r$ elements. Thus $\mathbf{y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ are partitioned as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{y}_1$ and $\boldsymbol{\mu}_1$ are $r \times 1$ and $\boldsymbol{\Sigma}_{11}$ is $r \times r$. Then $\mathbf{y}_1$ is distributed as $N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. Here, again, $E(\mathbf{y}_1) = \boldsymbol{\mu}_1$ and $\text{cov}(\mathbf{y}_1) = \boldsymbol{\Sigma}_{11}$ hold for any random vector partitioned in this way. But if $\mathbf{y}$ is p-variate normal, then $\mathbf{y}_1$ is r-variate normal.

b. As a special case of the above result, each y_i in $\mathbf{y}$ has the univariate normal distribution $N(\mu_i, \sigma_{ii})$. The converse of this is not true. If the density of each y_i in $\mathbf{y}$ is normal, it does not necessarily follow that $\mathbf{y}$ is multivariate normal.

In the next three properties, let the observation vector be partitioned into two subvectors denoted by $\mathbf{y}$ and $\mathbf{x}$. Or, alternatively, let $\mathbf{x}$ represent some additional variables to be considered along with those in $\mathbf{y}$. Then, as in (3.43) and (3.44),

$$E\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix} \qquad \text{cov}\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}.$$

5. Independence

a. The subvectors $\mathbf{y}$ and $\mathbf{x}$ are independent if $\boldsymbol{\Sigma}_{yx} = \mathbf{O}$.

b. Two individual variables y_i and y_j are independent if $\sigma_{ij} = 0$. Note that this is not true for many nonnormal random variables, as illustrated in Section 3.2.1.

6. Conditional distribution

If $\mathbf{y}$ and $\mathbf{x}$ are not independent, then $\boldsymbol{\Sigma}_{yx} \neq \mathbf{O}$, and the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, $f(\mathbf{y}|\mathbf{x})$, is multivariate normal with

$$E(\mathbf{y}|\mathbf{x}) = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) \tag{4.7}$$

and

$$\text{cov}(\mathbf{y}|\mathbf{x}) = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}. \tag{4.8}$$

Note that $E(\mathbf{y}|\mathbf{x})$ is a linear function of $\mathbf{x}$, while $\text{cov}(\mathbf{y}|\mathbf{x})$ does not depend on $\mathbf{x}$. The linear trend in (4.7) extends to any pair of variables. Thus to use (4.7) as a check on normality, one can examine bivariate scatter plots of all pairs of variables and look for any nonlinear trends. In (4.7), we have the justification for using the covariance or correlation to measure the relationship between two bivariate normal random variables. As noted in Section 3.2.1, the covariance and correlation are good measures of relationship only for variables with linear trends and are generally unsuitable for nonnormal random variables with a curvilinear relationship. The matrix $\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$ in (4.7) is called the *matrix of regression coefficients* because it relates $E(\mathbf{y}|\mathbf{x})$ to $\mathbf{x}$. The sample counterpart of this matrix appears in (10.37).

7. Distribution of the sum of two subvectors
If $\mathbf{y}$ and $\mathbf{x}$ are the same size (both $p \times 1$) and independent, then

$$\mathbf{y} + \mathbf{x} \quad \text{is} \quad N_p(\boldsymbol{\mu}_y + \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{yy} + \boldsymbol{\Sigma}_{xx}) \tag{4.9}$$

$$\mathbf{y} - \mathbf{x} \quad \text{is} \quad N_p(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{yy} + \boldsymbol{\Sigma}_{xx}). \tag{4.10}$$

Here, again, the mean vector and covariance matrix for $\mathbf{y} \pm \mathbf{x}$ hold in general. But if $\mathbf{y}$ and $\mathbf{x}$ are multivariate normal, then $\mathbf{y} \pm \mathbf{x}$ is multivariate normal.

To illustrate property 6, we discuss the conditional distribution for the bivariate normal. Let

$$\mathbf{u} = \begin{pmatrix} y \\ x \end{pmatrix}$$

where

$$E(\mathbf{u}) = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \qquad \text{cov}(\mathbf{u}) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}.$$

By definition $f(y|x) = g(y,x)/h(x)$, where $g(y,x)$ is the joint density of y and x and $h(x)$ is the density of x. Hence

$$g(y,x) = f(y|x)h(x)$$

and because the right side is a product, we seek a function of y and x that is independent of x and whose density can serve as $f(y|x)$. Since linear functions of y and x are normal by property 1a above, we consider $y - \beta x$ and seek the value of β so that $y - \beta x$ and x are independent.

Since $z = y - \beta x$ and x are normal and independent, $\text{cov}(x,z) = 0$. To find

$\text{cov}(x,z)$, we express x and z as functions of $\mathbf{u}$,

$$x = (0,1)\begin{pmatrix} y \\ x \end{pmatrix} = (0,1)\mathbf{u} = \mathbf{a}'\mathbf{u}$$

$$z = y - \beta x = (1, -\beta)\mathbf{u} = \mathbf{b}'\mathbf{u}.$$

Now

$$\begin{aligned} \text{cov}(x,z) &= \text{cov}(\mathbf{a}'\mathbf{u}, \mathbf{b}'\mathbf{u}) \\ &= \mathbf{a}'\boldsymbol{\Sigma}\mathbf{b} \quad [\text{by } (3.66)] \\ &= (0,1)\begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}\begin{pmatrix} 1 \\ -\beta \end{pmatrix} = (\sigma_{yx}, \sigma_x^2)\begin{pmatrix} 1 \\ -\beta \end{pmatrix} \\ &= \sigma_{yx} - \beta\sigma_x^2. \end{aligned}$$

Since $\text{cov}(x,z) = 0$, we obtain $\beta = \sigma_{yx}/\sigma_x^2$ and $y - \beta x$ becomes

$$y - \frac{\sigma_{yx}}{\sigma_x^2}x.$$

By property 1a above, the density of $y - (\sigma_{yx}/\sigma_x^2)x$ is normal with

$$E\left(y - \frac{\sigma_{yx}}{\sigma_x^2}x\right) = \mu_y - \frac{\sigma_{yx}}{\sigma_x^2}\mu_x$$

and

$$\begin{aligned} \text{var}\left(y - \frac{\sigma_{yx}}{\sigma_x^2}x\right) &= \text{var}(\mathbf{b}'\mathbf{u}) = \mathbf{b}'\boldsymbol{\Sigma}\mathbf{b} \\ &= \left(1, -\frac{\sigma_{yx}}{\sigma_x^2}\right)\begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}\begin{pmatrix} 1 \\ -\frac{\sigma_{yx}}{\sigma_x^2} \end{pmatrix} \\ &= \sigma_y^2 - \frac{\sigma_{yx}}{\sigma_x^2}. \end{aligned}$$

For a given value of x, y can be expressed as $y = \beta x + (y - \beta x)$, where βx is a fixed quantity corresponding to the given value of x and $y - \beta x$ is a random deviation. Then $f(y|x)$ is normal, with

$$\begin{aligned} E(y|x) &= \beta x + E(y - \beta x) = \beta x + \mu_y - \beta\mu_x \\ &= \mu_y + \beta(x - \mu_x) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}(x - \mu_x) \\ \text{var}(y|x) &= \sigma_y^2 - \frac{\sigma_{yx}^2}{\sigma_x^2}. \end{aligned}$$

4.3 ESTIMATION IN THE MULTIVARIATE NORMAL

4.3.1 Maximum Likelihood Estimation

When a distribution such as the multivariate normal is assumed to hold for a population, estimates of the parameters are often found by the method of *maximum likelihood*. This technique is conceptually simple: The observation vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are considered to be known and values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are sought that maximize the joint density of the y's, called the *likelihood function*. For the multivariate normal, the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{y}} \tag{4.11}$$

and

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})' \\ &= \frac{1}{n}\mathbf{W} \quad \text{say} \\ &= \frac{n-1}{n}\mathbf{S}, \end{aligned} \tag{4.12}$$

where $\mathbf{S}$ is the sample covariance matrix defined in (3.20) and (3.25). Since $\hat{\boldsymbol{\Sigma}}$ has divisor n instead of $n-1$, it is biased [see (3.31)], and we usually use $\mathbf{S}$ in place of $\hat{\boldsymbol{\Sigma}}$.

We now give a justification of $\overline{\mathbf{y}}$ as the maximum likelihood estimator of $\boldsymbol{\mu}$. Because the $\mathbf{y}_i$'s constitute a random sample, they are independent, and the joint density is the product of the densities of the y's. The likelihood function is, therefore,

$$\begin{aligned} L(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} f(\mathbf{y}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^p|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{y}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})/2} \\ &= \frac{1}{(\sqrt{2\pi})^{np}|\boldsymbol{\Sigma}|^{n/2}} e^{-\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})/2}. \end{aligned} \tag{4.13}$$

To see that $\hat{\boldsymbol{\mu}} = \overline{\mathbf{y}}$ maximizes the likelihood function, we write the exponent of (4.13) in a different form. By adding and subtracting $\overline{\mathbf{y}}$, the exponent in (4.13) becomes

$$-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}} + \overline{\mathbf{y}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \overline{\mathbf{y}} + \overline{\mathbf{y}} - \boldsymbol{\mu}).$$

When this is expanded in terms of $\mathbf{y}_i - \overline{\mathbf{y}}$ and $\overline{\mathbf{y}} - \boldsymbol{\mu}$, two of the four resulting terms vanish because $\Sigma_i(\mathbf{y}_i - \overline{\mathbf{y}}) = \mathbf{0}$, and (4.13) becomes

$$L = \frac{1}{(\sqrt{2\pi})^{np}|\boldsymbol{\Sigma}|^{n/2}} e^{-\sum_{i=1}^{n}(\mathbf{y}_i-\overline{\mathbf{y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\overline{\mathbf{y}})/2 - n(\overline{\mathbf{y}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}}-\boldsymbol{\mu})/2}. \tag{4.14}$$

Since $\boldsymbol{\Sigma}^{-1}$ is positive definite, $-n(\overline{\mathbf{y}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu})/2 \le 0$ and $0 < e^{-n(\overline{\mathbf{y}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}}-\boldsymbol{\mu})} \le 1$, with the maximum occurring when the exponent is 0. Therefore, L is maximized when $\hat{\boldsymbol{\mu}} = \overline{\mathbf{y}}$.

The maximum likelihood estimator of the population correlation matrix is the sample correlation matrix, that is,

$$\hat{\mathbf{P}}_\rho = \mathbf{R}.$$

Relationships among multinormal variables are linear, as can be seen in (4.7). Thus the estimators $\mathbf{S}$ and $\mathbf{R}$ serve well for the multivariate normal because they measure only linear relationships (see Sections 3.2.1 and 4.2). They are not as useful for some nonnormal distributions.

4.3.2 Distribution of $\overline{\mathbf{y}}$ and S

For the distribution of $\overline{\mathbf{y}} = \sum_{i=1}^{n}\mathbf{y}_i/n$, we can distinguish two cases.

(a) When $\overline{\mathbf{y}}$ is based on a random sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ from a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\overline{\mathbf{y}}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$.

(b) When $\overline{\mathbf{y}}$ is based on a random sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ from a nonnormal multivariate population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, for large n, $\overline{\mathbf{y}}$ is approximately $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$. More formally, this result is known as the *multivariate central limit theorem*: If $\overline{\mathbf{y}}$ is the mean vector of a random sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ from a population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then as $n \to \infty$, the distribution of $\sqrt{n}(\overline{\mathbf{y}} - \boldsymbol{\mu})$ approaches $N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

There are p variances in $\mathbf{S}$ and $\binom{p}{2}$ covariances, for a total of

$$p + \binom{p}{2} = p + p(p-1)/2 = p(p+1)/2$$

distinct entries. The joint distribution of these $p(p+1)/2$ distinct variables in $\mathbf{W} = (n-1)\mathbf{S} = \sum_i (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'$ is the Wishart distribution, denoted by $W_p(n-1, \boldsymbol{\Sigma})$, where $n-1$ is the degrees of freedom.

The Wishart distribution is the multivariate analogue of the χ^2-distribution, and it has similar uses. As noted in property 3 of Section 4.2, a χ^2 random variable is defined formally as the sum of squares of independent standard normal (univariate) random variables:

$$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2} \quad \text{is} \quad \chi_n^2.$$

If $\overline{y}$ is substituted for μ, then $\sum_i (y_i - \overline{y})^2/\sigma^2 = (n-1)s^2/\sigma^2$ is χ_{n-1}^2. Similarly, the formal definition of a Wishart random variable is

$$\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})' \quad \text{is} \quad W_p(n, \boldsymbol{\Sigma}), \tag{4.15}$$

where $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are independently distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. When $\overline{\mathbf{y}}$ is substituted for $\boldsymbol{\mu}$, the distribution remains Wishart with one less degree of freedom:

$$(n-1)\mathbf{S} = \sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})' \quad \text{is} \quad W_p(n-1, \boldsymbol{\Sigma}). \tag{4.16}$$

Finally, we note that when sampling from a multivariate normal distribution, $\overline{\mathbf{y}}$ and $\mathbf{S}$ are independent.

4.4 ASSESSING MULTIVARIATE NORMALITY

Many tests and graphical procedures have been suggested for evaluating whether a data set likely originated from a multivariate normal population. One possibility is to check each variable separately for univariate normality. Excellent reviews for both the univariate and multivariate cases have been given by Gnanadesikan (1977, pp. 161–195) and Seber (1984, pp. 141–155). We give a representative sample of univariate and multivariate methods in Sections 4.4.1 and 4.4.2, respectively.

4.4.1 Investigating Univariate Normality

When we have several variables, checking each for univariate normality should not be the sole approach, because (1) the variables are correlated and (2) normality of the individual variables does not guarantee joint normality. On the other hand, it is true that multivariate normality implies individual normality. Hence if even one of the separate variables is not normal, the vector is not multivariate normal. An initial check on the individual variables may therefore be useful.

A basic graphical approach for checking normality is the *Q–Q* plot that compares quantiles of a sample against the population quantiles of the univariate normal. If the points are close to a straight line, there is no indication of departure from normality. Deviation from a straight line indicates nonnormality (at least for a large sample). In fact, the type of nonlinear pattern may reveal the type of departure from normality. Some possibilities are illustrated in Figure 4.5.

Quantiles are similar to the more familiar percentiles, which are expressed in terms of percent; a test score at the 90th percentile, for example, is above 90% of the test scores and below 10% of them. Quantiles are expressed in terms of fractions or proportions. Thus the 90th percentile score becomes the 0.9 quantile score.

The sample quantiles for the *Q–Q* plot are obtained as follows. First we rank the observations $y_1, y_2, \ldots, y_n$ and denote the ordered values by $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$; thus $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$. Then the point $y_{(i)}$ is the i/n sample quantile. For example, if $n = 20$, $y_{(7)}$ is the $\frac{7}{20} = .35$ quantile, because .35 of the sample is less than or equal to $y_{(7)}$. The fraction i/n is often changed to $\left(i - \frac{1}{2}\right)/n$ as a continuity correction. If $n = 20$, $\left(i - \frac{1}{2}\right)/n$ ranges from .025 to .975 and more evenly covers the interval from 0 to 1. With this convention, $y_{(i)}$ is designated as the $\left(i - \frac{1}{2}\right)/n$ sample quantile.

The population quantiles for the *Q–Q* plot are similarly defined corresponding to $\left(i - \frac{1}{2}\right)/n$. If we denote these by $q_1, q_2, \ldots, q_n$, then q_i is the value below which a proportion $\left(i - \frac{1}{2}\right)/n$ of the observations in the population lie, that is, $\left(i - \frac{1}{2}\right)/n$ is the probability of getting an observation less than or equal to q_i. Formally, q_i can be found for the standard normal random variable y with dis-

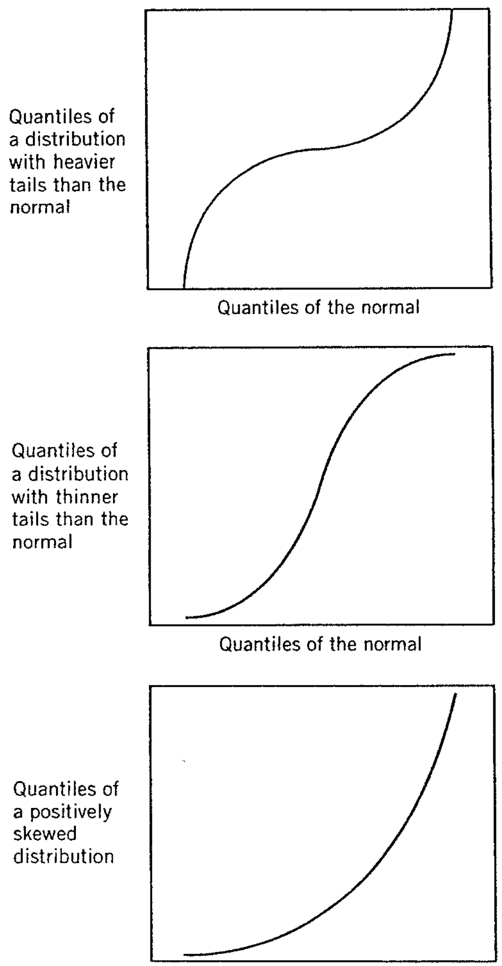


Figure 4.5 Typical Q–Q plots for nonnormal data.

tribution $N(0, 1)$ by solving

$$\Phi(q_i) = P(y < q_i) = \frac{i - \frac{1}{2}}{n}, \tag{4.17}$$

which would require numerical integration or tables of the cumulative standard normal distribution, $\Phi(x)$. Another benefit of using $\left(i - \frac{1}{2}\right)/n$ instead of i/n is that $n/n = 1$ would make $q_n = \infty$.

The population need not have the same mean and variance as the sample, since changes in mean and variance merely change the slope and intercept of the plotted line in the Q–Q plot. Therefore, we use the standard normal distribution, and the q_i values can easily be found from a table of cumulative standard normal probabilities. We then plot the pairs $(q_i, y_{(i)})$ and examine the resulting Q–Q plot for linearity.

Special graph paper, called normal probability paper, is available that eliminates the need to look up the q_i values. We need only plot $\left(i - \frac{1}{2}\right)/n$ in place of q_i, that is, plot the pairs $\left[\left(i - \frac{1}{2}\right)/n, y_{(i)}\right]$ and look for linearity as before. As an even easier alternative, most general-purpose statistical software programs now routinely provide normal probability plots of the pairs $(q_i, y_{(i)})$.

The Q–Q plots provide a good visual check on normality and are considered to be adequate for this purpose by many researchers. For those who desire a more objective procedure, several hypothesis tests are available. We give three of these that have good properties and are computationally tractable.

We discuss first a classical approach based on the following measures of skewness and kurtosis:

$$\sqrt{b_1} = \frac{\sqrt{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{3/2}} \tag{4.18}$$

and

$$b_2 = \frac{n\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{2}}. \tag{4.19}$$

These are sample estimates of the population skewness and kurtosis parameters $\sqrt{\beta_1}$ and β_2, respectively. When the population is normal, $\sqrt{\beta_1} = 0$ and $\beta_2 = 3$. If $\sqrt{\beta_1} < 0$, we have negative skewness; if $\sqrt{\beta_1} > 0$, the skewness is positive. Positive skewness is illustrated in Figure 4.6. If $\beta_2 < 3$, we have negative kurtosis, and if $\beta_2 > 3$, there is positive kurtosis. A distribution with negative kurtosis is characterized by being flatter than the normal distribution, that is, less peaked, with heavier flanks and thinner tails. A distribution with positive kurtosis has a higher peak than the normal, with an excess of values

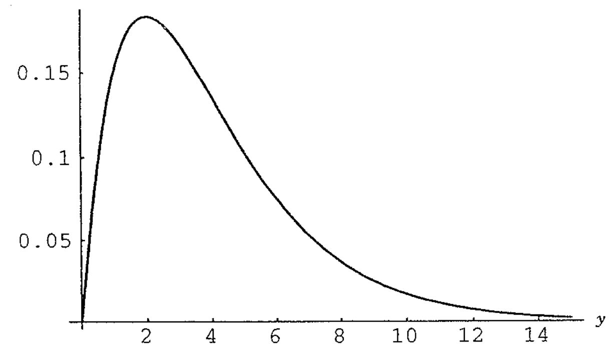


Figure 4.6 A distribution with positive skewness.

near the mean and in the tails but with thinner flanks. Positive and negative kurtosis are illustrated in Figure 4.7.

The test of normality can be carried out using the exact percentage points for $\sqrt{b_1}$ in Table A.1 for $4 \leq n \leq 25$, as given by Mulholland (1977). Alternatively, for $n \geq 8$ the function g as defined by

$$g(\sqrt{b_1}) = \delta \sinh^{-1} \frac{\sqrt{b_1}}{\lambda} \tag{4.20}$$

is approximately $N(0, 1)$, where

$$\sinh^{-1} x = \ln(x + \sqrt{x^2 + 1}). \quad (4.21))$$

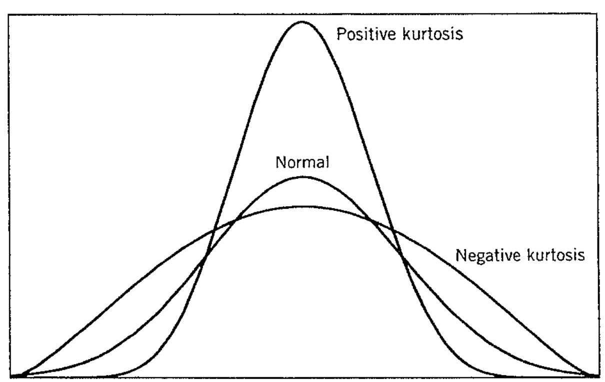


Figure 4.7 Distributions with positive and negative kurtosis compared to the normal.

Table A.2, from D'Agostino and Pearson (1973), gives values for δ and $1/\lambda$. To use b_2 as a test of normality, we can use Table A.3, obtained from D'Agostino and Tietjen (1971), which gives simulated percentiles of b_2 for selected values of n in the range $7 \leq n \leq 50$. Charts of percentiles of b_2 for $20 \leq n \leq 200$ can be found in D'Agostino and Pearson (1973).

Our second test for normality was given by D'Agostino (1971). The observations $y_1, y_2, \ldots, y_n$ are ordered as $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$, and we calculate

$$D = \frac{\sum_{i=1}^{n} \left[i - \frac{1}{2}(n+1) \right] y_{(i)}}{\sqrt{n^3 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{4.22}$$

and

$$Y = \frac{\sqrt{n}[D - (2\sqrt{\pi})^{-1}]}{.02998598}. \tag{4.23}$$

A table of percentiles for Y, given by D'Agostino (1972) for $10 \leq n \leq 250$, is provided in Table A.4.

The final test we report is by Lin and Mudholkar (1980). The test statistic is

$$z = \tanh^{-1} r = \tfrac{1}{2} \ln \frac{1+r}{1-r}, \tag{4.24}$$

where r is the sample correlation of the n pairs (y_i, x_i), $i = 1, 2, \ldots, n$, with x_i defined as

$$x_i = \frac{1}{n} \left[\sum_{j \neq i} y_j^2 - \frac{\left(\sum_{j \neq i} y_j \right)^2}{n-1} \right]^{1/3}. \tag{4.25}$$

If the y's are normal, z is approximately $N(0, 3/n)$. A more accurate upper 100α percentile is given by

$$z_\alpha = \sigma_n \left[u_\alpha + \tfrac{1}{24}(u_\alpha^3 - 3u_\alpha)\gamma_{2n} \right], \tag{4.26}$$

with

$$\sigma_n^2 = \frac{3}{n} - \frac{7.324}{n^2} + \frac{53.005}{n^3} \qquad u_\alpha = \Phi^{-1}(\alpha)$$
$$\gamma_{2n} = -\frac{11.70}{n} + \frac{55.06}{n^2},$$

where Φ is the distribution function of the $N(0, 1)$ distribution; that is, $\Phi(x)$ is the probability of an observation less than or equal to x, as in (4.17). The inverse function Φ^{-1} is essentially a quantile. For example, $u_{.05} = -1.645$ and $u_{.95} = 1.645$.

4.4.2 Investigating Multivariate Normality

Checking for multivariate normality is conceptually not as straightforward as assessing univariate normality, and consequently the state of the art is not as well developed. The complexity of this issue can be illustrated in the context of a goodness-of-fit test for normality. For a goodness-of-fit test in the univariate case, the range covered by a sample $y_1, y_2, \ldots, y_n$ is divided into several intervals, and we count how many y's fall into each interval. These observed frequencies (counts) are compared to the expected frequencies under the assumption that the sample came from a normal distribution with the same mean and variance as the sample. If the n observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are multivariate, however, the procedure is not so simple. We now have a p-dimensional region that would have to be divided into many more subregions than in the univariate case, and the expected frequencies for these subregions would be less easily obtained. With so many subregions, relatively few would contain observations; many would end up with no observations.

Thus because of the inherent "sparseness" of multivariate data, a goodness-of-fit test would be impractical. The points $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are more distant from each other in p-space than in any one of the p individual dimensions. Unless n is very large, a multivariate sample may not provide a very complete picture of the distribution from which it was taken.

As a consequence of the sparseness of the data in p-space, the tests for multivariate normality may not be very powerful. However, some check on the distribution is often desirable. Numerous procedures have been proposed for assessing multivariate normality. We discuss three of these, as well as a program containing three additional tests.

The first procedure is based on the standardized distance from each $\mathbf{y}_i$ to $\bar{\mathbf{y}}$,

$$D_i^2 = (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}). \tag{4.27}$$

Gnanadesikan and Kettenring (1972) showed that if the $\mathbf{y}_i$'s are multivariate normal, then

$$u_i = \frac{nD_i^2}{(n-1)^2} \tag{4.28}$$

has a beta distribution, which is related to the F. To obtain a Q–Q plot, the values $u_1, u_2, \ldots, u_n$ are ranked to give $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(n)}$, and we plot $(v_i, u_{(i)})$, where the quantiles v_i of the beta are given by

$$\int_0^{v_i} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{i-\alpha}{n-\alpha-\beta+1}, \tag{4.29}$$

with α and β defined as

$$\alpha = \frac{p-2}{2p} \tag{4.30}$$

and

$$\beta = \frac{n-p-2}{2(n-p-1)}. \tag{4.31}$$

A nonlinear pattern in the plot would indicate a departure from normality. A formal significance test is also available for $D_{(n)}^2 = \max_i D_i^2$. Table A.6 gives the upper 5 and 1% critical values from Barnett and Lewis (1978).

The second procedure involves scatter plots in two or three dimensions. If p is not too high, the bivariate plots of each pair of variables are often reduced in size and shown on one page, arranged to correspond to the entries in a correlation matrix. In this visual matrix, the eye readily picks out those pairs of variables that show a curved trend, outliers, or other nonnormal appearance. This plot is illustrated in connection with Example 4.5.2. The procedure is based on properties 4 and 6 of Section 4.2, from which we infer that (1) each pair of variables has a bivariate normal distribution and (2) bivariate normal variables follow a straight-line trend.

A popular option in many graphical programs is the ability to dynamically rotate a plot of three variables. While the points are rotating on the screen, a three-dimensional effect is created. The shape of the three-dimensional cloud of points is readily perceived, and we can detect various features of the data. The only drawbacks to this technique are that (1) it is a dynamic display and cannot be printed and (2) if p is very large, the number of subsets of three variables becomes unwieldy, while the number of pairs may still be tractable for plotting. These numbers are compared in Table 4.1, where $\binom{p}{2}$ and $\binom{p}{3}$ represent the number of subsets of sizes 2 and 3, respectively. Thus in many cases, the scatter plots for pairs of variables will continue to be used, even though three-dimensional plotting techniques are available.

The third procedure for assessing multivariate normality is a generalization of the univariate test based on the skewness and kurtosis measures $\sqrt{b_1}$ and b_2 as given by (4.18) and (4.19). The test is due to Mardia (1970). Let $\mathbf{y}$ and $\mathbf{x}$ be independent and identically distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then skewness and kurtosis for multivariate populations are defined by Mardia as

$$\beta_{1,p} = E[(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})]^3 \tag{4.32}$$

and

$$\beta_{2,p} = E[(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})]^2. \tag{4.33}$$

Since third-order central moments for the multivariate normal distribution are zero, $\beta_{1,p} = 0$ when $\mathbf{y}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It can also be shown that for multivariate normal $\mathbf{y}$,

$$\beta_{2,p} = p(p + 2). \tag{4.34}$$

If we define

$$g_{ij} = (\mathbf{y}_i - \bar{\mathbf{y}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_j - \bar{\mathbf{y}}), \tag{4.35}$$

where $\hat{\boldsymbol{\Sigma}} = \sum_i(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'/n$ is the maximum likelihood estimator (4.12), then sample estimates of $\beta_{1,p}$ and $\beta_{2,p}$ are given by

Table 4.1 Comparison of Number of Subsets of Sizes 2 and 3

p	$\binom{p}{2}$	$\binom{p}{3}$
6	15	20
8	28	56
10	45	120
12	66	220
15	105	455

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^3 \tag{4.36}$$

and

$$b_{2,p} = \frac{1}{n}\sum_{i} g_{ii}^2. \tag{4.37}$$

Table A.5 gives percentage points of $b_{1,p}$ and $b_{2,p}$ for $p = 2, 3, 4$, which can be used in testing for multivariate normality (Mardia 1970, 1974). For other values of p or when $n \geq 50$, the following approximate tests are available. For $b_{1,p}$

$$z_1 = \frac{(p+1)(n+1)(n+3)}{6[(n+1)(p+1)-6]} b_{1,p} \tag{4.38}$$

is approximately χ^2 with $\frac{1}{6}p(p + 1)(p + 2)$ degrees of freedom. Reject if $z_1 \geq \chi^2_{.05}$. With $b_{2,p}$, on the other hand, we wish to reject for large values (distribution too peaked) or small values (distribution too flat). For the upper 2.5% points of $b_{2,p}$ use

$$z_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \tag{4.39}$$

which is approximately $N(0, 1)$. For the lower 2.5% points we have two cases: (a) when $50 \leq n \leq 400$, use

$$z_3 = \frac{b_{2,p} - p(p+2)(n+p+1)/n}{\sqrt{8p(p+2)/(n-1)}} \tag{4.40}$$

which is approximately $N(0, 1)$; (b) when $n \geq 400$, use z_2 as given by (4.39).

Fortran programs for the above tests based on $b_{1,p}$ and $b_{2,p}$ are given by Siotani et al. (1985). They also provide programs for three additional tests for multivariate normality and several tests for univariate normality. Many of these programs are apparently unavailable elsewhere. The three multivariate tests are, briefly, as follows:

1. A test based on the third and fourth central moments

$$E[(y_i - \mu_i)(y_j - \mu_j)(y_k - \mu_k)] \tag{4.41}$$

and

$$E[(y_i - \mu_i)(y_j - \mu_j)(y_k - \mu_k)(y_l - \mu_l)]. \qquad (4.42)$$

Under normality, (4.41) is zero and (4.42) is equal to $\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$. Estimates of (4.41) and (4.42) are obtained and compared to 0 and $s_{ij}s_{kl} + s_{ik}s_{jl} + s_{il}s_{jk}$, respectively.

2. A multivariate generalization of the Shapiro–Wilk test: Define $z_i = \mathbf{c}'\mathbf{y}_i$, $i = 1, 2, \ldots, n$, where $\mathbf{c}$ is a constant vector, and

$$W(\mathbf{c}) = \frac{\sum_{i=1}^{n} a_i(z_{(i)} - \bar{z})^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}, \qquad (4.43)$$

where $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$ are the ordered values of $z_1, z_2, \ldots, z_n$ and the a_i's are coefficients tabulated in Shapiro and Wilk (1965). The hypothesis of multivariate normality is accepted if

$$\max_{\mathbf{c}}[W(\mathbf{c})] \geq k, \qquad (4.44)$$

where k corresponds to the desired significance level, α.

3. A directional normality test: This test is based on an alternative definition of the multivariate normal distribution suggested by property 1a of Section 4.2: If $\mathbf{a}'\mathbf{y}$ is $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$ for all $\mathbf{a}$, then $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. First the data vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are standardized by $\mathbf{z}_i = (\mathbf{S}^{1/2})^{-1}(\mathbf{y}_i - \bar{\mathbf{y}})$, $i = 1, 2, \ldots, n$, where $\mathbf{S}^{1/2}$ is the square root matrix given in (2.103). Then each $\mathbf{z}_i$ is multiplied by a direction vector $\mathbf{d}_\alpha$ to obtain $v_i = \mathbf{d}_\alpha'\mathbf{z}_i$ $i = 1, 2, \ldots, n$. The v's are approximately normal if the y's are multivariate normal. Several values of $\mathbf{d}_\alpha$ are used to check for normality in different directions. Various univariate normal tests can be applied to the v's.

4.5 OUTLIERS

The detection of outliers has been of concern to statisticians and other scientists for over a century. Many authors have claimed that the researcher can typically expect up to 10% of the observations to have errors in measurement or recording. Occasional stray observations from a different population than the target population are also fairly common. We do not attempt a complete summary of the vast literature covering univariate outliers, but we do review some major concepts and suggested procedures in Section 4.5.1 before moving to the multivariate case in Section 4.5.2. An alternative to detection of outliers is to use robust estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (see Rencher 1997, Section 1.10) that

are less sensitive to extreme observations than are the standard estimators $\bar{\mathbf{y}}$ and $\mathbf{S}$.

4.5.1 Outliers in Univariate Samples

Excellent surveys of the useful literature on outliers have been given by Beckman and Cook (1983), Hawkins (1980), and Barnett and Lewis (1978). We abstract a few highlights from Beckman and Cook. Many techniques have been proposed for detecting outliers in the residuals from regression, designed experiments, and so on. But we will be concerned only with simple random samples from the normal distribution. Outliers are also known as *discordant observations* or *contaminants*, which imply a discrepancy from what was expected and an origin from a nontarget population, respectively.

There are two principal approaches for dealing with outliers. The first is *identification*, which usually involves deletion of the outlier(s) but may alternatively provide important information about the model or the data. The second method involves *accommodation*, by modifying the method of analysis or the model. Robust methods, in which the influence of outliers is reduced, are the most familiar example of modification of the analysis. An example of a correction to the model is a mixture model that combines two normals with different variances, sometimes used to accommodate contaminants. For example, Marks and Rao (1978) accommodated a particular type of outlier due to patient fatigue by a mixture of two normal distributions.

In small or moderate sized univariate samples, visual methods of identifying outliers are the most frequently used. Tests are also available if a less subjective approach is desired.

Two types of *slippage* models have been proposed to account for outliers. Under the *mean slippage* model, all observations have the same variance, but one or more of the observations arise from a distribution with a different (population) mean. In the *variance slippage* model, one or more of the observations arise from a model with larger (population) variance but the same mean. Thus in the mean slippage model, the bulk of the observations arise from $N(\mu, \sigma^2)$, while the outliers originate from $N(\mu + \theta, \sigma^2)$. For the variance slippage model, the main distribution would again be $N(\mu, \sigma^2)$, with the outliers coming from $N(\mu, a\sigma^2)$ where $a > 1$. These models have led to the development of tests for rejection of outliers. We now briefly discuss some of these tests.

For a single outlier, most tests are based on the maximum studentized residual,

$$\max_i \tau_i = \max_i \left| \frac{y_i - \bar{y}}{s} \right|. \qquad (4.45)$$

If the largest or smallest observation is rejected, one could then examine the $n-1$ remaining observations for another possible outlier, and so on. This procedure is called a *consecutive test*. However, if there are two or more outliers, the less extreme ones will often make it difficult to detect the most extreme one, due to inflation of both mean and variance. This effect is called *masking*.

Ferguson (1961) showed that the maximum studentized residual (4.45) is more powerful than most other techniques for detecting intermediate or large shifts in the mean and gave the following guidelines for small shifts:

1. For outliers with small positive shifts in the mean, tests based on sample skewness are best.
2. For outliers with small shifts in the mean in either direction, tests based on the sample kurtosis are best.
3. For outliers with small positive shifts in the variance, tests based on the sample kurtosis are best.

Because of the masking problem in consecutive tests, *block tests* have been proposed for simultaneous rejection of $k > 1$ outliers. These tests work well if k is known, but in practice, it is usually not known. If the value we conjecture for k is too small, we incur the risk of failing to detect any outliers because of masking. If we set k too large, there is a high risk of rejecting more outliers than there really are, an effect known as *swamping*.

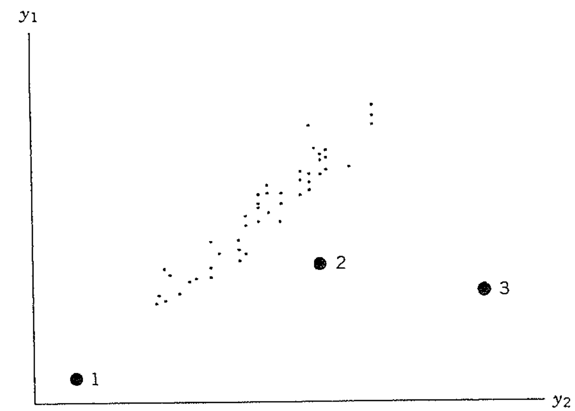
4.5.2 Outliers in Multivariate Samples

In the case of multivariate data, the problems in detecting outliers are intensified for several reasons:

1. For $p > 2$ the data cannot be readily plotted to pinpoint the outliers.
2. Multivariate data cannot be ordered as can a univariate sample, where extremes show up readily on either end.
3. An observation vector may have a large recording error in one of its components or smaller errors in several components.
4. A multivariate outlier may reflect slippage in mean, variance, or correlation. This is illustrated in Figure 4.8. Observation 1 causes a small shift in means and variances of both y_1 and y_2 but has little effect on the correlation. Observation 2 has little effect on means and variances, but it reduces the correlation somewhat. Observation 3 has a major effect on means, variances, and correlation.

Of course, as in the univariate case, one approach to outlier identification or accommodation is to use robust methods of estimation. Such methods minimize the influence of outliers in estimation or model fitting. However, an outlier sometimes furnishes valuable information, and the specific pursuit of outliers can be very worthwhile.

We present two methods of multivariate outlier identification, both of which



Figures 4.8 Bivariate sample showing three types of outliers.

turn out to be related to methods of assessing multivariate normality. (A third approach based on principal components is given in Section 12.4.) The first method, due to Wilks (1963), is designed for detection of a single outlier. Wilks' statistic is

$$w = \max_i \frac{|(n-2)\mathbf{S}_{-i}|}{|(n-1)\mathbf{S}|}, \tag{4.46}$$

where $\mathbf{S}$ is the usual sample covariance matrix and $\mathbf{S}_{-i}$ is obtained from the same sample with the ith observation deleted. It turns out that w can be expressed in terms of $D^2_{(n)} = \max_i(\mathbf{y}_i - \overline{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i - \overline{\mathbf{y}})$ as

$$w = 1 - \frac{nD^2_{(n)}}{(n-1)^2}, \tag{4.47}$$

thus basing a test for an outlier on the distances D_i^2 used in Section 4.4.2 in a graphical procedure for checking multivariate normality. Table A.6 gives the upper 5 and 1% critical values for $D^2_{(n)}$ from Barnett and Lewis (1978).

Yang and Lee (1987) provide an F-test of w as given by (4.47). Define

$$F_i = \frac{n-p-1}{p}\left[\frac{1}{1 - nD_i^2/(n-1)^2} - 1\right], \qquad i = 1, 2, \ldots, n. \tag{4.48}$$

Then the F_i are independently and identically distributed as $F_{p,n-p-1}$, and a test can be constructed in terms of $\max_i F_i$:

$$P\left(\max_i F_i > f\right) = 1 - P(\text{all } F_i \leq f) = 1 - [P(F \leq f)]^n.$$

Therefore, the test can be carried out using an F-table. Note that

$$\max_i F_i = F_{(n)} = \frac{n-p-1}{p}\left(\frac{1}{w} - 1\right), \tag{4.49}$$

where w is given in (4.47).

The second test we discuss is designed for detection of several outliers. Schwager and Margolin (1982) showed that the locally best invariant test for mean slippage is based on Mardia's (1970) sample kurtosis $b_{2,p}$ as defined by (4.35) and (4.37). Essentially this means that among all tests invariant to a class of transformations of the type $\mathbf{z} = \mathbf{Ay} + \mathbf{b}$, where $\mathbf{A}$ is nonsingular, the test using $b_{2,p}$ is most powerful for small shifts in the mean vector. This result holds if the proportion of outliers is no more than 21.13%. With some restrictions on the pattern of the outliers, the permissible fraction of outliers can go as high as $33\frac{1}{3}$%. The hypothesis is H_0: no outliers are present. This hypothesis is rejected for large values of $b_{2,p}$.

A table of critical values of $b_{2,p}$ and some approximate tests were described in Section 4.4.2 following (4.37). Here again we have a test that doubles as a check for multivariate normality and for the presence of outliers. One advantage of this test for outliers is that we do not have to specify the number of outliers and run the attendant risk of masking or swamping. Schwager and Margolin (1982) pointed out that this feature "increases the importance of performing an overall test that is sensitive to a broad range of outlier configurations. There is also empirical evidence that the kurtosis test performs well in situations of practical interest when compared with other inferential outlier procedures."

Sinha (1984) extended the result of Schwager and Margolin to cover the general case of elliptically symmetric distributions. An *elliptically symmetric distribution* is one in which $f(\mathbf{y}) = |\mathbf{\Sigma}|^{-1/2} g[(\mathbf{y}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})]$. By varying the function g, distributions with shorter or longer tails than the normal can be obtained. Of course, the critical value of $b_{2,p}$ would have to be adjusted to correspond to the distribution, but rejection for large values would be a locally best invariant test.

Table 4.2 Values of D_i^2 for the Ramus Bone Data in Table 3.8

Observation Number	D_i^2	Observation Number	D_i^2
1	0.7588	11	2.8301
2	1.2980	12	10.5718
3	1.7591	13	2.5941
4	3.8539	14	0.6594
5	0.8706	15	0.3246
6	2.8106	16	0.8321
7	4.2915	17	1.1083
8	7.9897	18	4.3633
9	11.0301	19	2.1088
10	5.3519	20	10.0931

Example 4.5.2. We use the ramus bone data set of Table 3.8 to illustrate a search for multivariate outliers, while at the same time checking for multivariate normality. An examination of each column of Table 3.8 does not reveal any apparent univariate outliers. We next calculate D_i^2 in (4.27) for each observation vector. The results are given in Table 4.2.

We see that D_9^2, D_{12}^2, and D_{20}^2 seem to stand out as possible outliers. This impression is confirmed when we compute u_i and v_i in (4.28) and (4.29) and plot them in Figure 4.9. The figure shows a distinct departure from linearity due to the three large D_i^2 values noted above and also possibly due to $D_8^2 = 7.99$. In Table A.6, the upper 5% critical value for the maximum value, $D_{(20)}^2$, is given

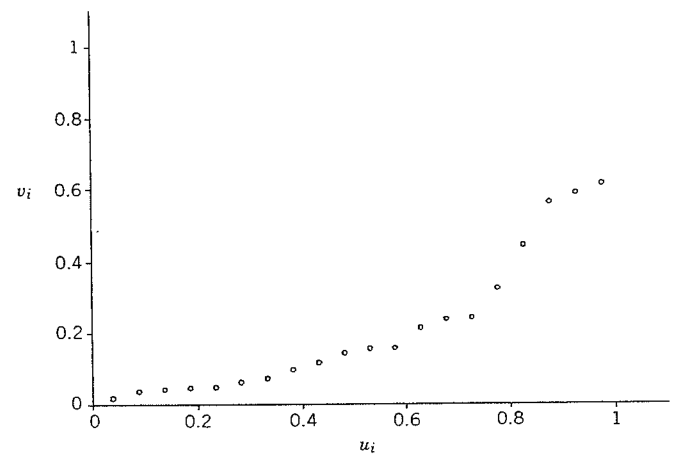


Figure 4.9 Q–Q plot of u_i and v_i for the ramus bone data of Table 3.8.

as 11.63. In our case, the largest D_i^2 is $D_9^2 = 11.03$, which does not exceed the critical value. This does not surprise us, since the test was designed to detect a single outlier, and we may have at least three.

We next calculate $b_{1,p}$ and $b_{2,p}$ as given by (4.36) and (4.37):

$$b_{1,p} = 11.338 \qquad b_{2,p} = 28.884.$$

In Table A.5, the upper .01 critical value for $b_{1,p}$ is 9.9; the upper .005 critical value for $b_{2,p}$ is 27.1. Thus both $b_{1,p}$ and $b_{2,p}$ exceed their critical values, and we have significant skewness and kurtosis, apparently caused by the three observations with large values of D_i^2.

The bivariate scatter plots are given in Figure 4.10. The three values are clearly separate from the other observations in the plot of y_1 versus y_4. In Table 3.8, the 9th, 12th, and 20th values of y_4 are not unusual, nor are the 9th, 12th, and 20th values of y_1. However, the increase from y_1 to y_4 is exceptional in each case. If these values are not due to errors in recording the data and if this sample is representative, then we appear to have a mixture of two populations. This should be taken into account in making inferences.

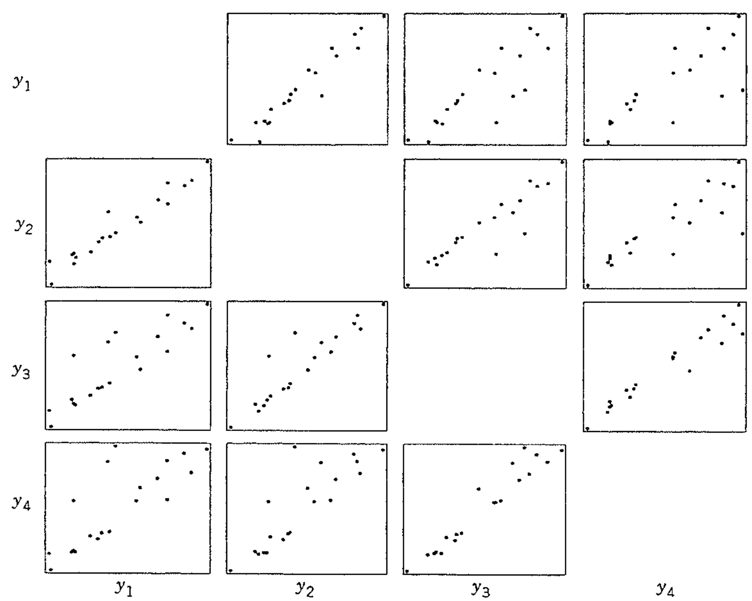


Figure 4.10 Scatter plots for the ramus bone data in Table 3.8.

PROBLEMS

4.1 Consider the two covariance matrices

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 14 & 8 & 3 \\ 8 & 5 & 2 \\ 3 & 2 & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 6 & 6 & 1 \\ 6 & 8 & 2 \\ 1 & 2 & 1 \end{pmatrix}.$$

Show that $|\boldsymbol{\Sigma}_2| > |\boldsymbol{\Sigma}_1|$ and that $\text{tr}(\boldsymbol{\Sigma}_2) < \text{tr}(\boldsymbol{\Sigma}_1)$. Thus the generalized variance of population 2 is greater than the generalized variance of population 1, even though the total variance is less. Comment on why this is true in terms of the correlations.

4.2 For $\mathbf{z} = (\mathbf{T}')^{-1}(\mathbf{y} - \boldsymbol{\mu})$ as in (4.4), show that $E(\mathbf{z}) = \mathbf{0}$ and $\text{cov}(\mathbf{z}) = \mathbf{I}$.

4.3 Show that the form of the likelihood function in (4.13) follows from the previous expression.

4.4 Show that by adding and subtracting $\overline{\mathbf{y}}$, the exponent of (4.13) has the form given in (4.14), that is,

$$\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}} + \overline{\mathbf{y}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \overline{\mathbf{y}} + \overline{\mathbf{y}} - \boldsymbol{\mu})$$
$$= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \overline{\mathbf{y}}) + \frac{n}{2}(\overline{\mathbf{y}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu}).$$

4.5 Show that $\sqrt{b_1}$ and b_2 as given in (4.18) and (4.19) are invariant to the transformation $z_i = ay_i + b$.

4.6 Show that if $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\beta_{2,p} = p(p+2)$ as in (4.34).

4.7 Show that $b_{1,p}$ and $b_{2,p}$ as given by (4.36) and (4.37) are invariant under the transformation $\mathbf{z}_i = \mathbf{A}\mathbf{y}_i + \mathbf{b}$, where $\mathbf{A}$ is nonsingular. Thus $b_{1,p}$ and $b_{2,p}$ do not depend on the units of measurement; the variables could even be standardized.

4.8 Show that $F_{(n)} = [(n-p-1)/p](1/w - 1)$ as in (4.49).

4.9 Suppose $\mathbf{y}$ is $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 1 \\ 4 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 6 & 1 & -2 \\ 1 & 13 & 4 \\ -2 & 4 & 4 \end{pmatrix}.$$

(a) Find the distribution of $z = 2y_1 - y_2 + 3y_3$.
(b) Find the joint distribution of $z_1 = y_1 + y_2 + y_3$ and $z_2 = y_1 - y_2 + 2y_3$.
(c) Find the distribution of y_2.
(d) Find the joint distribution of y_1 and y_3.
(e) Find the joint distribution of y_1, y_3, and $\frac{1}{2}(y_1 + y_2)$.

4.10 Suppose $\mathbf{y}$ is $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given in the previous problem.
(a) Find a vector $\mathbf{z}$ such that $\mathbf{z} = (\mathbf{T}')^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is $N_3(\mathbf{0}, \mathbf{I})$ as in (4.4).
(b) Find a vector $\mathbf{z}$ such that $\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is $N_3(\mathbf{0}, \mathbf{I})$ as in (4.5).
(c) What is the distribution of $(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$?

4.11 Suppose $\mathbf{y}$ is $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} -2 \\ 3 \\ -1 \\ 5 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 11 & -8 & 3 & 9 \\ -8 & 9 & -3 & -6 \\ 3 & -3 & 2 & 3 \\ 9 & -6 & 3 & 9 \end{bmatrix}.$$

(a) Find the distribution of $z = 4y_1 - 2y_2 + y_3 - 3y_4$.
(b) Find the joint distribution of $z_1 = y_1 + y_2 + y_3 + y_4$ and $z_2 = -2y_1 + 3y_2 + y_3 - 2y_4$.
(c) Find the joint distribution of $z_1 = 3y_1 + y_2 - 4y_3 - y_4$, $z_2 = -y_1 - 3y_2 + y_3 - 2y_4$, and $z_3 = 2y_1 + 2y_2 + 4y_3 - 5y_4$.
(d) What is the distribution of y_3?
(e) What is the joint distribution of y_2 and y_4?
(f) Find the joint distribution of y_1, $\frac{1}{2}(y_1 + y_2)$, $\frac{1}{3}(y_1 + y_2 + y_3)$, and $\frac{1}{4}(y_1 + y_2 + y_3 + y_4)$.

4.12 Suppose $\mathbf{y}$ is $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given in the previous problem.
(a) Find a vector $\mathbf{z}$ such that $\mathbf{z} = (\mathbf{T}')^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is $N_4(\mathbf{0}, \mathbf{I})$, as in (4.4).
(b) Find a vector $\mathbf{z}$ such that $\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is $N_4(\mathbf{0}, \mathbf{I})$, as in (4.5).
(c) What is the distribution of $(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$?

4.13 Suppose $\mathbf{y}$ is $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ -3 \\ 4 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & -3 & 0 \\ -3 & 6 & 0 \\ 0 & 0 & 5 \end{pmatrix}.$$

Which of the following random variables are independent?

(a) y_1 and y_2
(b) y_1 and y_3
(c) y_2 and y_3
(d) (y_1, y_2) and y_3
(e) (y_1, y_3) and y_2

4.14 Suppose $\mathbf{y}$ is $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu} = \begin{bmatrix} -4 \\ 2 \\ 5 \\ -1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 8 & 0 & -1 & 0 \\ 0 & 3 & 0 & 2 \\ -1 & 0 & 5 & 0 \\ 0 & 2 & 0 & 7 \end{bmatrix}.$$

Which of the following random variables are independent?

(a) y_1 and y_2
(b) y_1 and y_3
(c) y_1 and y_4
(d) y_2 and y_3
(e) y_2 and y_4
(f) y_3 and y_4
(g) (y_1, y_2) and y_3
(h) (y_1, y_2) and y_4
(i) (y_1, y_3) and y_4
(j) y_1 and (y_2, y_4)
(k) y_1 and y_2 and y_3
(l) y_1 and y_2 and y_4
(m) (y_1, y_2) and (y_3, y_4)
(n) (y_1, y_3) and (y_2, y_4)

4.15 Assume $\mathbf{y}$ and $\mathbf{x}$ are subvectors, each 2×1, where

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \quad \text{is} \quad N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\mu} = \left[\begin{array}{c} 2 \\ -1 \\ \hline 3 \\ 1 \end{array}\right] \qquad \boldsymbol{\Sigma} = \left[\begin{array}{cc|cc} 7 & 3 & -3 & 2 \\ 3 & 6 & 0 & 4 \\ \hline -3 & 0 & 5 & -2 \\ 2 & 4 & -2 & 4 \end{array}\right].$$

(a) Find $E(\mathbf{y}|\mathbf{x})$ by (4.7).
(b) Find $\text{cov}(\mathbf{y}|\mathbf{x})$ by (4.8).

4.16 Suppose $\mathbf{y}$ and $\mathbf{x}$ are subvectors, such that $\mathbf{y}$ is 2×1 and $\mathbf{x}$ is 3×1, with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ partitioned accordingly:

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ -2 \\ \hline 4 \\ -3 \\ 5 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \left[\begin{array}{cc|ccc} 14 & -8 & 15 & 0 & 3 \\ -8 & 18 & 8 & 6 & -2 \\ \hline 15 & 8 & 50 & 8 & 5 \\ 0 & 6 & 8 & 4 & 0 \\ 3 & -2 & 5 & 0 & 1 \end{array}\right].$$

Assume that $\binom{\mathbf{y}}{\mathbf{x}}$ is distributed as $N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

(a) Find $E(\mathbf{y}|\mathbf{x})$ by (4.7).

(b) Find $\text{cov}(\mathbf{y}|\mathbf{x})$ by (4.8).

4.17 Suppose that $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ is a random sample from a nonnormal multivariate population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. If n is large, what is the approximate distribution of each of the following?

(a) $\sqrt{n}(\bar{\mathbf{y}} - \boldsymbol{\mu})$

(b) $\bar{\mathbf{y}}$

4.18 For the ramus bone data treated in Example 4.5.2, check each of the four variables for univariate normality using the following techniques:

(a) Q–Q plots

(b) $\sqrt{b_1}$ and b_2 as given by (4.18) and (4.19)

(c) D'Agostino's test using D and Y given in (4.22) and (4.23)

(d) The test by Lin and Mudholkar using z defined in (4.24)

4.19 For the calcium data in Table 3.5, check for multivariate normality and outliers using the following tests:

(a) Calculate D_i^2 as in (4.27) for each observation.

(b) Compare the largest value of D_i^2 with the critical value in Table A.6.

(c) Compute u_i and v_i in (4.28) and (4.29) and plot them. Is there an indication of nonlinearity or outliers?

(d) Calculate $b_{1,p}$ and $b_{2,p}$ in (4.36) and (4.37) and compare them with critical values in Table A.5.

4.20 For the probe word data in Table 3.7, check each of the five variables for univariate normality and outliers using the following tests:

(a) Q–Q plots

(b) $\sqrt{b_1}$ and b_2 as given by (4.18) and (4.19)

(c) D'Agostino's test using D and Y given in (4.22) and (4.23)

(d) The test by Lin and Mudholkar using z defined in (4.24)

4.21 For the probe word data in Table 3.7, check for multivariate normality and outliers using the following tests:

(a) Calculate D_i^2 as in (4.27) for each observation.

(b) Compare the largest value of D_i^2 with the critical value in Table A.6.

(c) Compute u_i and v_i in (4.28) and (4.29) and plot them. Is there an indication of nonlinearity or outliers?

(d) Calculate $b_{1,p}$ and $b_{2,p}$ in (4.36) and (4.37) and compare them with critical values in Table A.5.

4.22 Six hematology variables were measured on 51 workers (Royston 1983):

y_1 = hemoglobin concentration	y_4 = lymphocyte count
y_2 = packed cell volume	y_5 = neutrophil count
y_3 = white blood cell count	y_6 = serum lead concentration

The data are given in Table 4.3. Check each of the six variables for univariate normality using the following tests.

(a) Q–Q plots

(b) $\sqrt{b_1}$ and b_2 as given by (4.18) and (4.19)

(c) D'Agostino's test using D and Y given in (4.22) and (4.23)

(d) The test by Lin and Mudholkar using z defined in (4.24)

4.23 For the hematology data in Table 4.3, check for multivariate normality using the following techniques:

(a) Calculate D_i^2 as in (4.27) for each observation.

(b) Compare the largest value of D_i^2 with the critical value in Table A.6 (extrapolate).

(c) Compute u_i and v_i in (4.28) and (4.29) and plot them. Is there an indication of nonlinearity or outliers?

(d) Calculate $b_{1,p}$ and $b_{2,p}$ in (4.36) and (4.37) and compare them with critical values in Table A.5.

Table 4.3 Hematology Data

Observation Number	y_1	y_2	y_3	y_4	y_5	y_6
1	13.4	39	4100	14	25	17
2	14.6	46	5000	15	30	20
3	13.5	42	4500	19	21	18
4	15.0	46	4600	23	16	18
5	14.6	44	5100	17	31	19
6	14.0	44	4900	20	24	19
7	16.4	49	4300	21	17	18
8	14.8	44	4400	16	26	29
9	15.2	46	4100	27	13	27
10	15.5	48	8400	34	42	36
11	15.2	47	5600	26	27	22
12	16.9	50	5100	28	17	23
13	14.8	44	4700	24	20	23
14	16.2	45	5600	26	25	19
15	14.7	43	4000	23	13	17
16	14.7	42	3400	9	22	13
17	16.5	45	5400	18	32	17
18	15.4	45	6900	28	36	24
19	15.1	45	4600	17	29	17
20	14.2	46	4200	14	25	28
21	15.9	46	5200	8	34	16
22	16.0	47	4700	25	14	18
23	17.4	50	8600	37	39	17
24	14.3	43	5500	20	31	19
25	14.8	44	4200	15	24	29
26	14.9	43	4300	9	32	17
27	15.5	45	5200	16	30	20
28	14.5	43	3900	18	18	25
29	14.4	45	6000	17	37	23
30	14.6	44	4700	23	21	27
31	15.3	45	7900	43	23	23
32	14.9	45	3400	17	15	24
33	15.8	47	6000	23	32	21
34	14.4	44	7700	31	39	23
35	14.7	46	3700	11	23	23
36	14.8	43	5200	25	19	22
37	15.4	45	6000	30	25	18
38	16.2	50	8100	32	38	18
39	15.0	45	4900	17	26	24
40	15.1	47	6000	22	33	16
41	16.0	46	4600	20	22	22
42	15.3	48	5500	20	23	23
43	14.5	41	6200	20	36	21
44	14.2	41	4900	26	20	20
45	15.0	45	7200	40	25	25
46	14.2	46	5800	22	31	22
47	14.9	45	8400	61	17	17
48	16.2	48	3100	12	15	18
49	14.5	45	4000	20	18	20
50	16.4	49	6900	35	22	24
51	14.7	44	7800	38	34	16

CHAPTER 5

Tests on One or Two Mean Vectors

5.1 MULTIVARIATE VERSUS UNIVARIATE TESTS

Hypothesis testing in a multivariate context is more complex than in a univariate setting. The number of parameters may be staggering. The p-variate normal distribution, for example, has p means, p variances, and $\binom{p}{2}$ covariances, where $\binom{p}{2}$ represents the number of pairs among the p variables. The total number of parameters is

$$p + p + \binom{p}{2} = \tfrac{1}{2}p(p + 3).$$

Each parameter corresponds to a hypothesis that could be formulated. Additionally, we might well be interested in testing hypotheses about subsets of these parameters or about functions of them. In some cases, we have the added dilemma of choosing among competing test statistics.

We first discuss the motivation for testing p variables multivariately rather than, or in addition to, univariately. There are at least four arguments for a multivariate approach to hypothesis testing:

1. The use of p univariate tests inflates the Type I error rate, α, whereas the multivariate test preserves the exact α level. For example, if we do $p = 10$ separate univariate tests at the .05 level, the probability of at least one false rejection is greater than .05. If the variables were independent (they rarely are), we would have (under H_0)

$$\begin{aligned} P(\text{at least one rejection}) &= 1 - P(\text{all 10 tests accept}) \\ &= 1 - (.95)^{10} \\ &= .40. \end{aligned}$$

The resulting overall α of .40 is not an acceptable error rate. Typically,