4.3 Sampling from a Multivariate Normal Distribution and Maximum Likelihood Estimation

We discussed sampling and selecting random samples briefly in Chapter 3. In this section, we shall be concerned with samples from a multivariate normal population—in particular, with the sampling distribution of $\bar{\mathbf{X}}$ and $\mathbf{S}$.

The Multivariate Normal Likelihood

Let us assume that the $p \times 1$ vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ represent a random sample from a multivariate normal population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Since $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are mutually independent and each has distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the joint density function of all the observations is the product of the marginal normal densities:

$$\left\{\begin{array}{c}\text{Joint density} \\ \text{of } \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\end{array}\right\} = \prod_{j=1}^{n}\left\{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j-\boldsymbol{\mu})/2}\right\}$$

$$= \frac{1}{(2\pi)^{np/2}}\frac{1}{|\boldsymbol{\Sigma}|^{n/2}} e^{-\sum_{j=1}^{n}(\mathbf{x}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j-\boldsymbol{\mu})/2} \tag{4-11}$$

When the numerical values of the observations become available, they may be substituted for the $\mathbf{x}_j$ in Equation (4-11). The resulting expression, now considered as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the fixed set of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, is called the *likelihood*.

Many good statistical procedures employ values for the population parameters that "best" explain the observed data. One meaning of *best* is to select the parameter values that *maximize* the joint density evaluated at the observations. This technique is called *maximum likelihood estimation*, and the maximizing parameter values are called *maximum likelihood estimates*.

At this point, we shall consider maximum likelihood estimation of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for a multivariate normal population. To do so, we take the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ as fixed and consider the joint density of Equation (4-11) evaluated at these values. The result is the likelihood function. In order to simplify matters, we rewrite the likelihood function in another form. We shall need some additional properties for the trace of a square matrix. (The trace of a matrix is the sum of its diagonal elements, and the properties of the trace are discussed in Definition 2A.28 and Result 2A.12.)

Result 4.9. Let $\mathbf{A}$ be a $k \times k$ symmetric matrix and $\mathbf{x}$ be a $k \times 1$ vector. Then

(a) $\mathbf{x}'\mathbf{A}\mathbf{x} = \operatorname{tr}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \operatorname{tr}(\mathbf{A}\mathbf{x}\mathbf{x}')$

(b) $\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{k} \lambda_i$, where the λ_i are the eigenvalues of $\mathbf{A}$.

Proof. For Part a, we note that $\mathbf{x}'\mathbf{A}\mathbf{x}$ is a scalar, so $\mathbf{x}'\mathbf{A}\mathbf{x} = \operatorname{tr}(\mathbf{x}'\mathbf{A}\mathbf{x})$. We pointed out in Result 2A.12 that $\operatorname{tr}(\mathbf{BC}) = \operatorname{tr}(\mathbf{CB})$ for any two matrices $\mathbf{B}$ and $\mathbf{C}$ of dimensions $m \times k$ and $k \times m$, respectively. This follows because $\mathbf{BC}$ has $\sum_{j=1}^{k} b_{ij}c_{ji}$ as its ith diagonal element, so $\operatorname{tr}(\mathbf{BC}) = \sum_{i=1}^{m}\left(\sum_{j=1}^{k} b_{ij}c_{ji}\right)$. Similarly, the jth diagonal element of $\mathbf{CB}$ is $\sum_{i=1}^{m} c_{ji}b_{ij}$, so $\operatorname{tr}(\mathbf{CB}) = \sum_{j=1}^{k}\left(\sum_{i=1}^{m} c_{ji}b_{ij}\right) = \sum_{i=1}^{m}\left(\sum_{j=1}^{k} b_{ij}c_{ji}\right) = \operatorname{tr}(\mathbf{BC})$. Let $\mathbf{x}'$ be the matrix $\mathbf{B}$ with $m = 1$, and let $\mathbf{A}\mathbf{x}$ play the role of the matrix $\mathbf{C}$. Then $\operatorname{tr}(\mathbf{x}'(\mathbf{A}\mathbf{x})) = \operatorname{tr}((\mathbf{A}\mathbf{x})\mathbf{x}')$, and the result follows.

Part b is proved by using the spectral decomposition of (2-20) to write $\mathbf{A} = \mathbf{P}'\boldsymbol{\Lambda}\mathbf{P}$, where $\mathbf{P}\mathbf{P}' = \mathbf{I}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix with entries $\lambda_1, \lambda_2, \ldots, \lambda_k$. Therefore, $\operatorname{tr}(\mathbf{A}) = \operatorname{tr}(\mathbf{P}'\boldsymbol{\Lambda}\mathbf{P}) = \operatorname{tr}(\boldsymbol{\Lambda}\mathbf{P}\mathbf{P}') = \operatorname{tr}(\boldsymbol{\Lambda}) = \lambda_1 + \lambda_2 + \cdots + \lambda_k$. ■

Now the exponent in the joint density in (4-11) can be simplified. By Result 4.9(a),

$$(\mathbf{x}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) = \operatorname{tr}[(\mathbf{x}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})]$$
$$= \operatorname{tr}[\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})'] \tag{4-12}$$

Next,

$$\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) = \sum_{j=1}^{n}\operatorname{tr}[(\mathbf{x}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})]$$
$$= \sum_{j=1}^{n}\operatorname{tr}[\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})']$$
$$= \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})'\right)\right] \tag{4-13}$$

since the trace of a sum of matrices is equal to the sum of the traces of the matrices, according to Result 2A.12(b). We can add and subtract $\bar{\mathbf{x}} = (1/n)\sum_{j=1}^{n}\mathbf{x}_j$ in each term $(\mathbf{x}_j - \boldsymbol{\mu})$ in $\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})'$ to give

$$\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu})(\mathbf{x}_j - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu})'$$
$$= \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' + \sum_{j=1}^{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'$$
$$= \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})' \tag{4-14}$$

because the cross-product terms, $\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}} - \boldsymbol{\mu})'$ and $\sum_{j=1}^{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\mathbf{x}_j - \bar{\mathbf{x}})'$, are both matrices of zeros. (See Exercise 4.15.) Consequently, using Equations (4-13) and (4-14), we can write the joint density of a random sample from a multivariate normal population as

$$\left\{\begin{array}{c}\text{Joint density of} \\ \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\end{array}\right\} = (2\pi)^{-np/2}|\boldsymbol{\Sigma}|^{-n/2}$$
$$\times \exp\left\{-\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\right)\right]\Big/2\right\} \tag{4-15}$$

Substituting the observed values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ into the joint density yields the likelihood function. We shall denote this function by $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, to stress the fact that it is a function of the (unknown) population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Thus, when the vectors $\mathbf{x}_j$ contain the specific numbers actually observed, we have

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'+n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right)\right]/2} \tag{4-16}$$

It will be convenient in later sections of this book to express the exponent in the likelihood function (4-16) in different ways. In particular, we shall make use of the identity

$$\begin{aligned}
&\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'+n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right)\right] \\
&= \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\right)\right] + n\,\mathrm{tr}[\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'] \\
&= \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\right)\right] + n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})
\end{aligned} \tag{4-17}$$

Maximum Likelihood Estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

The next result will eventually allow us to obtain the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Result 4.10. Given a $p \times p$ symmetric positive definite matrix $\mathbf{B}$ and a scalar $b > 0$, it follows that

$$\frac{1}{|\boldsymbol{\Sigma}|^b} e^{-\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B})/2} \le \frac{1}{|\mathbf{B}|^b}(2b)^{pb}e^{-bp}$$

for all positive definite $\underset{(p\times p)}{\boldsymbol{\Sigma}}$, with equality holding only for $\boldsymbol{\Sigma} = (1/2b)\mathbf{B}$.

Proof. Let $\mathbf{B}^{1/2}$ be the symmetric square root of $\mathbf{B}$ [see Equation (2-22)], so $\mathbf{B}^{1/2}\mathbf{B}^{1/2} = \mathbf{B}$, $\mathbf{B}^{1/2}\mathbf{B}^{-1/2} = \mathbf{I}$, and $\mathbf{B}^{-1/2}\mathbf{B}^{-1/2} = \mathbf{B}^{-1}$. Then $\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B}) = \mathrm{tr}[(\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2})\mathbf{B}^{1/2}] = \mathrm{tr}[\mathbf{B}^{1/2}(\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2})]$. Let η be an eigenvalue of $\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}$. This matrix is positive definite because $\mathbf{y}'\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}\mathbf{y} = (\mathbf{B}^{1/2}\mathbf{y})'\boldsymbol{\Sigma}^{-1}(\mathbf{B}^{1/2}\mathbf{y}) > 0$ if $\mathbf{B}^{1/2}\mathbf{y} \neq \mathbf{0}$ or, equivalently, $\mathbf{y} \neq \mathbf{0}$. Thus, the eigenvalues η_i of $\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}$ are positive by Exercise 2.17. Result 4.9(b) then gives

$$\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B}) = \mathrm{tr}(\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}) = \sum_{i=1}^{p}\eta_i$$

and $|\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}| = \prod_{i=1}^{p}\eta_i$ by Exercise 2.12. From the properties of determinants in Result 2A.11, we can write

$$\begin{aligned}
|\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}| &= |\mathbf{B}^{1/2}||\boldsymbol{\Sigma}^{-1}||\mathbf{B}^{1/2}| = |\boldsymbol{\Sigma}^{-1}||\mathbf{B}^{1/2}||\mathbf{B}^{1/2}| \\
&= |\boldsymbol{\Sigma}^{-1}||\mathbf{B}| = \frac{1}{|\boldsymbol{\Sigma}|}|\mathbf{B}|
\end{aligned}$$

or

$$\frac{1}{|\boldsymbol{\Sigma}|} = \frac{|\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}|}{|\mathbf{B}|} = \frac{\prod_{i=1}^{p}\eta_i}{|\mathbf{B}|}$$

Combining the results for the trace and the determinant yields

$$\frac{1}{|\boldsymbol{\Sigma}|^b}e^{-\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\mathbf{B}]/2} = \frac{\left(\prod_{i=1}^{p}\eta_i\right)^b}{|\mathbf{B}|^b}e^{-\sum_{i=1}^{p}\eta_i/2} = \frac{1}{|\mathbf{B}|^b}\prod_{i=1}^{p}\eta_i^b e^{-\eta_i/2}$$

But the function $\eta^b e^{-\eta/2}$ has a maximum, with respect to η, of $(2b)^b e^{-b}$, occurring at $\eta = 2b$. The choice $\eta_i = 2b$, for each i, therefore gives

$$\frac{1}{|\boldsymbol{\Sigma}|^b}e^{-\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B})/2} \le \frac{1}{|\mathbf{B}|^b}(2b)^{pb}e^{-bp}$$

The upper bound is uniquely attained when $\boldsymbol{\Sigma} = (1/2b)\mathbf{B}$, since, for this choice,

$$\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2} = \mathbf{B}^{1/2}(2b)\mathbf{B}^{-1}\mathbf{B}^{1/2} = (2b)\underset{(p\times p)}{\mathbf{I}}$$

and

$$\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\mathbf{B}] = \mathrm{tr}[\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}] = \mathrm{tr}[(2b)\mathbf{I}] = 2bp$$

Moreover,

$$\frac{1}{|\boldsymbol{\Sigma}|} = \frac{|\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}|}{|\mathbf{B}|} = \frac{|(2b)\mathbf{I}|}{|\mathbf{B}|} = \frac{(2b)^p}{|\mathbf{B}|}$$

Straightforward substitution for $\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\mathbf{B}]$ and $1/|\boldsymbol{\Sigma}|^b$ yields the bound asserted. ■

The maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are those values—denoted by $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$—that maximize the function $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in (4-16). The estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ will depend on the observed values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ through the summary statistics $\bar{\mathbf{x}}$ and $\mathbf{S}$.

Result 4.11. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a normal population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Then

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_j-\bar{\mathbf{X}})(\mathbf{X}_j-\bar{\mathbf{X}})' = \frac{(n-1)}{n}\mathbf{S}$$

are the *maximum likelihood estimators* of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. Their observed values, $\bar{\mathbf{x}}$ and $(1/n)\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'$, are called the *maximum likelihood estimates* of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Proof. The exponent in the likelihood function [see Equation (4-16)], apart from the multiplicative factor $-\frac{1}{2}$, is [see (4-17)]

$$\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\right)\right] + n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})$$