

25. Suppose that in Exercise 24, corresponding to any fixed point  $(p_1, p_2, p_3) \in$  this system of three mass points, and show that the position of  $\mathbf{x}$  defines a one-to-one correspondence between  $\mathcal{P}$  and  $\mathcal{C}$ . [For any point  $\mathbf{x} \in \mathcal{C}$ , the corresponding values  $(p_1, p_2, p_3)$  are called the *barycentric coordinates* of  $\mathbf{x}$ .]

26. Show that the correspondence between  $\mathcal{P}$  and  $\mathcal{C}$  defined in Exercise 25 is the same as the correspondence defined in Exercise 24.

27. Suppose that each of  $k$  statisticians has his own prior distribution for a certain parameter  $W$ , and let  $\xi_i$  be the g.p.d.f. which statistician  $i$  assigns to  $W$  ( $i = 1, \dots, k$ ). Suppose also that an executive forms his opinion about  $W$  from the opinions of the  $k$  statisticians and that he assigns to  $W$  the g.p.d.f.  $\xi^*$  defined, at each point  $w \in \Omega$ , as follows:

$$\xi^*(w) = \alpha_1 \xi_1(w) + \dots + \alpha_k \xi_k(w).$$

Here  $\alpha_1, \dots, \alpha_k$  are weights such that  $\alpha_i \geq 0$  ( $i = 1, \dots, k$ ) and  $\alpha_1 + \dots + \alpha_k = 1$ . The value of  $\alpha_i$  reflects the relative weight that the executive gives to the opinion of statistician  $i$ . Suppose further that the  $k$  statisticians and the executive observe together the value of a random variable  $X$  whose conditional g.p.d.f. when  $W = w$  is  $f(\cdot | w)$ . Show that the posterior g.p.d.f. of the executive will again be a linear combination of the posterior g.p.d.f.'s of the  $k$  statisticians, with new weights  $\beta_1, \dots, \beta_k$  which will depend on the observed value of  $X$ . Also, discuss the conditions under which the weight  $\beta_1$  in the posterior g.p.d.f. will be greater than the weight  $\alpha_1$  in the prior g.p.d.f.

## CHAPTER 9

# conjugate prior distributions

### 9.1 SUFFICIENT STATISTICS

Consider a statistical problem in which a large amount of experimental data has been collected. The treatment of the data is often simplified if the statistician computes a few numerical values, or statistics, and considers these values as summaries of the relevant information in the data. In some problems, a statistical analysis that is based on these few summary values can be just as effective as any analysis that could be based on all the observed values. In this chapter we shall consider problems for which fully informative summaries of this type are available. Such summaries are known as *sufficient statistics*.

Suppose that  $W$  is a parameter which takes values in the space  $\Omega$ . Also, suppose that  $X$  is a random variable, or random vector, which takes values in the sample space  $S$ . We shall let  $f(\cdot | w)$  denote the conditional g.p.d.f. of  $X$  when  $W = w$  ( $w \in \Omega$ ). It is assumed that the observed value of  $X$  will be available for making inferences and decisions relating to the parameter  $W$ . In this context, any function  $T$  of the observation  $X$ , whether or not  $T$  is a real-valued function, is called a *statistic*.

Loosely speaking, a statistic  $T$  is called a sufficient statistic if, for any prior distribution of  $W$ , its posterior distribution depends on the

observed value of  $X$  only through  $T(X)$ . More formally, for any prior g.p.d.f.  $\xi$  of  $W$  and any observed value  $x \in S$ , let  $\xi(\cdot | x)$  denote the posterior g.p.d.f. of  $W$ . For simplicity, it will be assumed in this section that for every value of  $x \in S$  and every prior g.p.d.f.  $\xi$ , the posterior g.p.d.f.  $\xi(\cdot | x)$  exists and is specified by Bayes' theorem. Then it is said that a statistic  $T$  is a *sufficient statistic* for the family of g.p.d.f.'s  $\{f(\cdot | w), w \in \Omega\}$  if  $\xi(\cdot | x_1) = \xi(\cdot | x_2)$  for any prior g.p.d.f.  $\xi$  and any two points  $x_1 \in S$  and  $x_2 \in S$  such that  $T(x_1) = T(x_2)$ . There is a good reason for saying that a statistic which has this property is sufficient. In order to be able to compute the posterior distribution of  $W$  from any prior distribution, the statistician needs only the value of  $T(X)$ . He does not need the value of  $X$  itself, which may be a vector of high dimension. It should be emphasized here that the values of any g.p.d.f. can be arbitrarily changed on any set of points having probability 0. Hence, when we say that the g.p.d.f.'s  $\xi(\cdot | x)$  in the preceding definition or the g.p.d.f.'s  $f(\cdot | w)$  in the following theorem have certain properties, we mean that there are versions of these g.p.d.f.'s which have those properties.

The following theorem, which is known as the *factorization criterion*, provides an easy way of recognizing sufficient statistics.

**Theorem 1** *A statistic  $T$  is sufficient for a family of g.p.d.f.'s  $\{f(\cdot | w), w \in \Omega\}$  if, and only if,  $f(x|w)$  can be factored as follows for all values  $x \in S$  and  $w \in \Omega$ :*

$$f(x|w) = u(x)v[T(x), w]. \quad (1)$$

Here, the function  $v$  is positive and does not depend on  $w$  and the function  $u$  is nonnegative and depends on  $x$  only through  $T(x)$ .

*Proof* Suppose first that the factorization indicated in Eq. (1) is correct. Then, for any prior g.p.d.f.  $\xi$  of  $W$  and any points  $w \in \Omega$  and  $x \in S$ , the posterior g.p.d.f. of  $W$  is

$$\xi(w|x) = \frac{v[T(x), w]\xi(w)}{\int_{\Omega} v[T(x), w']\xi(w') dw'}. \quad (2)$$

Since the right side of Eq. (2) depends on the observed value  $x$  only through the value  $T(x)$ , it follows that  $T$  is a sufficient statistic.

Conversely, suppose that  $T$  is a sufficient statistic. Let  $\xi$  be any prior g.p.d.f. of  $W$  such that  $\xi(w) > 0$  at every point  $w \in \Omega$ . The posterior g.p.d.f.  $\xi(w|x)$  is specified at any points  $w \in \Omega$  and  $x \in S$  as follows:

$$\xi(w|x) = \frac{f(x|w)\xi(w)}{\int_{\Omega} f(x|w')\xi(w') dw'}. \quad (3)$$

Since  $T$  is a sufficient statistic, then  $\xi(w|x) = r[T(x), w]$ , where the function  $r$  involves only  $T(x)$  and  $w$ . Hence, it follows from Eq. (3) that

$$f(x|w) = \left[ \int_{\Omega} f(x|w')\xi(w') dv(w') \right] \frac{r[T(x), w]}{\xi(w)}. \quad (4)$$

Equation (4) exhibits a factorization of the form indicated in Eq. (1). ■

Throughout the remainder of this chapter and in many other parts of this book, we shall need to consider random variables  $X_1, \dots, X_n$  whose joint distribution depends on the value  $w$  of some parameter  $W$  in the following way: For any given value  $w$  of  $W$  ( $w \in \Omega$ ), the variables  $X_1, \dots, X_n$  form a random sample from a specified distribution whose g.p.d.f. is  $f(\cdot | w)$ . Hence, the conditional joint g.p.d.f.  $f_n(\cdot | w)$  of  $X_1, \dots, X_n$ , when  $W = w$ , is specified by the equation

$$f_n(x_1, \dots, x_n|w) = f(x_1|w) \cdots f(x_n|w). \quad (5)$$

When the distribution of  $X_1, \dots, X_n$  satisfies these conditions, we shall say that  $X_1, \dots, X_n$  is a *random sample from the specified distribution with an unknown value of the parameter  $W$* . We shall now present three examples of sufficient statistics which will also illustrate this terminology.

**EXAMPLE 1** Suppose that  $X_1, \dots, X_n$  is a random sample from a Bernoulli distribution with an unknown value of the parameter  $W$ . In other words, the only possible values of each random variable  $X_i$  are 0 and 1, and for any given value  $w$  of  $W$  such that  $0 < w < 1$ , the joint p.f.  $f_n(\cdot | w)$  of  $X_1, \dots, X_n$  is specified by the equation

$$f_n(x_1, \dots, x_n|w) = w^y(1-w)^{n-y}. \quad (6)$$

Here  $y = \sum_{i=1}^n x_i$ . Hence, the joint p.f. given in Eq. (6) depends on the values of the random variables  $X_1, \dots, X_n$  only through their sum.

Accordingly, let  $T$  be the statistic defined by the equation

$$T(X_1, \dots, X_n) = \sum_{i=1}^n X_i. \quad (7)$$

It follows that  $T$  is a sufficient statistic for the family of p.f.'s specified by Eq. (6) for  $0 < w < 1$ .

**EXAMPLE 2** Suppose that  $X_1, \dots, X_n$  is a random sample from a normal distribution with an unknown value of the mean and an unknown value of the variance. Then, for any given values  $\mu$  and  $\sigma^2$  of the mean and variance such that  $-\infty < \mu < \infty$  and  $\sigma^2 > 0$ , the conditional joint

p.d.f.  $f_n(\cdot | \mu, \sigma^2)$  of  $X_1, \dots, X_n$  is specified by the equation

$$f_n(x_1, \dots, x_n | \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right]. \quad (8)$$

If we let  $\bar{x} = (1/n)\sum_{i=1}^n x_i$ , then

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2. \quad (9)$$

Therefore, the p.d.f. defined by Eq. (8) depends on the values of the observations  $X_1, \dots, X_n$  only through the two values  $\bar{x}$  and  $\sum_{i=1}^n (x_i - \bar{x})^2$ .

Accordingly, let  $\mathbf{T}$  be the two-dimensional vector statistic defined by the equation

$$\mathbf{T}(X_1, \dots, X_n) = \{ \bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2 \}. \quad (10)$$

It follows that  $\mathbf{T}$  is a sufficient statistic for the family of p.d.f.'s specified by Eq. (8). It is sometimes said that the two components of the vector  $\mathbf{T}(X_1, \dots, X_n)$  are *jointly sufficient statistics*.

**EXAMPLE 3** Suppose that  $X_1, \dots, X_n$  is a random sample from the uniform distribution on the interval  $(0, W)$ , where the value of the parameter  $W$  is unknown. For any given value  $w$  of  $W$  such that  $w > 0$ , the conditional p.d.f.  $f(\cdot | w)$  of any single observation  $X_i$  is specified by the equation

$$f(x|w) = \begin{cases} \frac{1}{w} & \text{for } x < w, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

The sample space  $S$  of each observation is assumed to be the set of positive numbers. It follows that the conditional joint p.d.f.  $f_n(\cdot | w)$  of  $X_1, \dots, X_n$ , when  $W = w$  ( $w > 0$ ), is as follows:

$$f_n(x_1, \dots, x_n | w) = \begin{cases} \frac{1}{w^n} & \text{for } x_i < w \ (i = 1, \dots, n), \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

This p.d.f. can be rewritten in the form

$$f_n(x_1, \dots, x_n | w) = \begin{cases} \frac{1}{w^n} & \text{for } \max \{x_1, \dots, x_n\} < w, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

This function depends on the values of the observations  $X_1, \dots, X_n$  only through the value of  $\max \{x_1, \dots, x_n\}$ .

Accordingly, let  $T$  be the statistic defined by the equation

$$T(X_1, \dots, X_n) = \max \{X_1, \dots, X_n\}. \quad (14)$$

It follows that  $T$  is a sufficient statistic for the family of p.d.f.'s specified by Eq. (12).

Other examples of sufficient statistics are given in Exercises 1 to 9 at the end of this chapter.

### Further Remarks and References

The concept of a sufficient statistic was introduced by Fisher (1922). A rigorous measure-theoretic presentation of this concept is given by Halmos and Savage (1949). Sufficient statistics have also been studied by Lehmann and Scheffé (1950), Bahadur (1954), and Dynkin (1961) and are discussed in the books by Savage (1954), Lehmann (1959), and Raiffa and Schlaifer (1961).

## 9.2 CONJUGATE FAMILIES OF DISTRIBUTIONS

In each of the three examples presented in Sec. 9.1 and in each of the examples presented in Exercises 1 to 7, there is a sufficient statistic which can be represented by one or two real-valued functions of the observations  $X_1, \dots, X_n$  in a random sample, and the dimension of this sufficient statistic remains fixed regardless of the size  $n$  of the sample. The planning and analysis of an experiment become much easier for the statistician if it can be assumed that the observations are to be drawn from a family of distributions for which there is a sufficient statistic of fixed dimension. One important simplification results from the fact that under these conditions there must exist a standard family of distributions of the parameter  $W$  which has the following property: If the prior distribution of  $W$  belongs to this family, then for any sample size  $n$  and any values of the observations in the sample, the posterior distribution of  $W$  must also belong to the same family. A family of distributions with this property is said to be *closed under sampling*. The family is also called a *conjugate family of distributions* because of the special relationship which must exist between this family of distributions of the parameter and the family of distributions of the observations.

Therefore, when there exists a sufficient statistic of fixed dimension, it is possible for the statistician to handle only prior and posterior distributions which belong to a relatively small conjugate family. In order for this simplification to be of much use, however, the conjugate family of distributions of  $W$  must still be rich enough to permit the statistician to

find within the family, in a wide variety of situations, a distribution which will adequately represent his prior distribution of  $W$ .

As an example, suppose that  $X_1, \dots, X_n$  is a random sample from the Bernoulli distribution with an unknown value of the parameter  $W$ . Suppose also that the prior distribution of  $W$  is a beta distribution with specified values of the parameters  $\alpha$  and  $\beta$  such that  $\alpha > 0$  and  $\beta > 0$ . Then the prior p.d.f.  $\xi$  of  $W$  is

$$\xi(w) \propto w^{\alpha-1}(1-w)^{\beta-1} \quad \text{for } 0 < w < 1. \quad (1)$$

Here we are utilizing the proportionality symbol  $\propto$  to indicate that the function  $\xi$  is basically as given on the right side of the relation (1) but that there may be a factor which does not involve  $w$ .

In general, if the prior g.p.d.f. of  $W$  is  $\xi$  and the conditional g.p.d.f. of  $X$  when  $W = w$  is  $f(\cdot | w)$ , then the posterior g.p.d.f.  $\xi(\cdot | x)$  of  $W$  when  $X = x$  is

$$\xi(w|x) \propto \xi(w)f(x|w) \quad \text{for } w \in \Omega. \quad (2)$$

The use of the proportionality symbol in the relation (2) is justified since the conditional g.p.d.f. of  $W$  is equal to the right side of (2) divided by the factor  $\int_{\Omega} f(x|w')\xi(w') dw'$ , which does not involve  $w$ .

We shall now return to our example. The conditional joint p.f.  $f_n(\cdot | w)$  of  $X_1, \dots, X_n$  when  $W = w$  is given by Eq. (6) of Sec. 9.1. Therefore, from the relations (1) and (2), the posterior p.d.f.  $\xi(\cdot | x_1, \dots, x_n)$  of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is

$$\xi(w|x_1, \dots, x_n) \propto w^{\alpha+\nu-1}(1-w)^{\beta+n-\nu-1}. \quad (3)$$

Here  $y = \sum_{i=1}^n x_i$ . It can be seen from the relation (3) that the posterior distribution of  $W$  is a beta distribution with parameters  $\alpha + y$  and  $\beta + n - y$ . We have proved the following theorem.

**Theorem 1** Suppose that  $X_1, \dots, X_n$  is a random sample from a Bernoulli distribution with an unknown value of the parameter  $W$ . Suppose also that the prior distribution of  $W$  is a beta distribution with parameters  $\alpha$  and  $\beta$  such that  $\alpha > 0$  and  $\beta > 0$ . Then the posterior distribution of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is a beta distribution with parameters  $\alpha + y$  and  $\beta + n - y$ , where  $y = \sum_{i=1}^n x_i$ .

In other words, the family of beta distributions is a conjugate family for samples from a Bernoulli distribution. It should be emphasized here that the term "parameter" is now doing double duty. The random variable  $W$ , whose value is not known by the statistician, is a parameter whose possible values  $w$  index the family of g.p.d.f.'s  $f(\cdot | w)$  of each obser-

vation  $X$  which is available. The parameters  $\alpha$  and  $\beta$ , whose values are specified, index the conjugate family of p.d.f.'s of  $W$ .

The construction of conjugate families of distributions will be studied more systematically in the next section.

### Further Remarks and References

A family of g.p.d.f.'s  $\{f(\cdot | w), w \in \Omega\}$ , each of which is defined on a given sample space  $S$ , is said to be an *exponential family* if the g.p.d.f.'s are of the following form for any points  $x \in S$  and  $w \in \Omega$ :

$$f(x|w) = a(w)b(x) \exp \left[ \sum_{i=1}^k g_i(w)h_i(x) \right]. \quad (4)$$

Consider an exponential family of this type and suppose that  $X_1, \dots, X_n$  are random variables, or random vectors, all of whose values lie in the sample space  $S$  and whose joint g.p.d.f.  $f_n(\cdot | w)$ , for any point  $w \in \Omega$ , is specified by the equation

$$f_n(x_1, \dots, x_n|w) = \prod_{j=1}^n f(x_j|w). \quad (5)$$

Let  $\mathbf{T}$  be the  $k$ -dimensional vector defined by the equation

$$\mathbf{T}(X_1, \dots, X_n) = \left\{ \sum_{j=1}^n h_1(X_j), \dots, \sum_{j=1}^n h_k(X_j) \right\}. \quad (6)$$

Then the statistic  $\mathbf{T}$  is a sufficient statistic of fixed dimension  $k$  for each sample size  $n$ .

Darmonis (1935), Koopman (1936), and Pitman (1936) have shown that among families of distributions which satisfy certain regularity conditions, a sufficient statistic of fixed dimension will exist only for exponential families [see also Fraser (1963)]. Almost all the examples which we have considered in this chapter involve exponential families (see Exercise 11). A family of uniform distributions is not an exponential family, but there is a sufficient statistic of fixed dimension for such a family. However, one of the regularity conditions which is not satisfied by a family of uniform distributions is the condition that the set of points  $x$  such that  $f(x|w) > 0$  must be the same set for every value  $w \in \Omega$ .

### 9.3 CONSTRUCTION OF THE CONJUGATE FAMILY

Now consider again the example summarized in Theorem 1 of Sec. 9.2. For any positive constants  $\alpha$  and  $\beta$ , let  $g(\cdot | \alpha, \beta)$  denote the p.d.f. of a

beta distribution with parameters  $\alpha$  and  $\beta$ . In that example, the family of beta distributions was found to be the appropriate conjugate family because of the following two properties:

First, consider any observed values  $x_1, \dots, x_n$  of the variables  $X_1, \dots, X_n$ . The conditional joint p.f.  $f_n(x_1, \dots, x_n|w)$  of  $X_1, \dots, X_n$  is specified by Eq. (6) of Sec. 9.1. If this function is regarded as a function of  $w$ , then it follows from the definition of the p.d.f. of a beta distribution that

$$f_n(x_1, \dots, x_n|w) \propto g(w|y + 1, n - y + 1). \tag{1}$$

Therefore, for any observed values  $x_1, \dots, x_n$ , the function  $f_n(x_1, \dots, x_n|w)$  is proportional to the p.d.f. of a beta distribution.

Second, if  $g(\cdot | \alpha_1, \beta_1)$  and  $g(\cdot | \alpha_2, \beta_2)$  are the p.d.f.'s of any two beta distributions, then there is another p.d.f.  $g(\cdot | \alpha_3, \beta_3)$  such that for  $0 < w < 1$ ,

$$g(w|\alpha_3, \beta_3) \propto g(w|\alpha_1, \beta_1)g(w|\alpha_2, \beta_2). \tag{2}$$

In fact, it follows from the definition of the p.d.f. of a beta distribution that

$$g(w|\alpha_3, \beta_3) \propto w^{\alpha_1+\alpha_2-2}(1-w)^{\beta_1+\beta_2-2}. \tag{3}$$

Since the right side of the relation (3) is proportional to the p.d.f. of a beta distribution with parameters  $\alpha_1 + \alpha_2 - 1$  and  $\beta_1 + \beta_2 - 1$ , we obtain the equations

$$\alpha_3 = \alpha_1 + \alpha_2 - 1 \quad \text{and} \quad \beta_3 = \beta_1 + \beta_2 - 1. \tag{4}$$

A family of p.d.f.'s which satisfies the relation (2) is said to be *closed under multiplication*.

If the prior distribution of  $W$  is taken to be a beta distribution with p.d.f.  $g(\cdot | \alpha, \beta)$ , then the posterior p.d.f.  $\xi(\cdot | x_1, \dots, x_n)$  of  $W$  will satisfy the relation

$$\xi(w|x_1, \dots, x_n) \propto f_n(x_1, \dots, x_n|w)g(w|\alpha, \beta). \tag{5}$$

It now follows from the relations (1) to (4) that the posterior p.d.f. of  $W$  must be that of a beta distribution with parameters  $\alpha + y$  and  $\beta + n - y$ . This result is the one that was presented in Theorem 1 of Sec. 9.2.

This development suggests a method for determining a conjugate family of distributions in any problem for which there exists a sufficient statistic of fixed dimension. The statistician need only determine a family of p.d.f.'s of the parameter  $W$  such that (1) for any sample size  $n$  and any observed values  $x_1, \dots, x_n$ , the conditional joint g.p.d.f.  $f_n(x_1, \dots, x_n|w)$ , regarded as a function of  $w$ , is proportional to one of the p.d.f.'s in the family, and (2) the family is closed under multiplication.

We shall now show that whenever the family of g.p.d.f.'s  $\{f_n(\cdot | w)\}$ ,  $w \in \Omega$  has a sufficient statistic  $T_n(X_1, \dots, X_n)$  of fixed dimension  $k$  ( $k \geq 1$ ) for every sample size  $n$ , there must exist a simple conjugate family of this type. It follows from Theorem 1 of Sec. 9.1 that for each value of  $n$ , there is a function  $v_n$  such that

$$f_n(x_1, \dots, x_n|w) \propto v_n[T_n(x_1, \dots, x_n), w]. \tag{6}$$

If we let  $T_n(x_1, \dots, x_n) = t$  and assume that  $\int_{\Omega} v_n(t, w) dw < \infty$ , then there exists a p.d.f.  $g(\cdot | t, n)$  on the space  $\Omega$  such that

$$g(w|t, n) \propto v_n(t, w). \tag{7}$$

Consider the family of p.d.f.'s  $g(\cdot | t, n)$  for all possible sample sizes  $n$  and all possible values  $t$  of the statistic  $T_n(X_1, \dots, X_n)$ . It follows from the relations (6) and (7) that  $f_n(x_1, \dots, x_n|w)$  must be proportional to one of the p.d.f.'s in this family. Furthermore, as will now be demonstrated, this family of p.d.f.'s is closed under multiplication.

Consider any two p.d.f.'s  $g(\cdot | s, m)$  and  $g(\cdot | t, n)$  which belong to the family. Then there must exist observed values  $x_1, \dots, x_m$  and  $y_1, \dots, y_n$  of samples of size  $m$  and size  $n$  such that  $T_m(x_1, \dots, x_m) = s$  and  $T_n(y_1, \dots, y_n) = t$ . If these observations are combined, they form a sample of size  $m + n$  and their g.p.d.f.'s satisfy the equation

$$f_{m+n}(x_1, \dots, x_m, y_1, \dots, y_n|w) = f_m(x_1, \dots, x_m|w)f_n(y_1, \dots, y_n|w). \tag{8}$$

If we let  $u = T_{m+n}(x_1, \dots, x_m, y_1, \dots, y_n)$ , then it follows from the relations (6) to (8) that

$$g(w|u, m + n) \propto g(w|s, m)g(w|t, n). \tag{9}$$

Therefore, the family is closed under multiplication and it satisfies the properties of a conjugate family.

It is often convenient to choose a slightly larger family of p.d.f.'s than the one we have just constructed for the conjugate family. For example, it can be seen from the relation (1) that for samples from a Bernoulli distribution, the function  $f_n(x_1, \dots, x_n|w)$  must always be proportional to the p.d.f. of a beta distribution for which both parameters  $\alpha$  and  $\beta$  are positive integers. It follows that this subfamily of the family of beta distributions will be closed under multiplication, and it could have been chosen as the conjugate family. However, we found it convenient to choose the whole family of beta distributions as the conjugate family and to verify that it also was closed under multiplication.

In the subsequent sections of this chapter we shall present several theorems which identify conjugate families for samples from various

distributions. Each of these theorems was discovered by writing down the function  $f_n(x_1, \dots, x_n|w)$  and then recognizing it as being proportional to a p.d.f. which belongs to one of the standard families of distributions described in Chaps. 4 and 5. When this family has been identified and the theorem has been formulated, its proof consists only of the verification that the family is closed under sampling.

When the function  $f_n(x_1, \dots, x_n|w)$  is regarded as a function of  $w$ , for given values  $x_1, \dots, x_n$  of the observations, it is called the *likelihood function*. Thus, the likelihood function will be of fundamental importance in our development of conjugate families of distributions.

#### Further Remarks and References

The concept of a conjugate family of distributions was formalized by Raiffa and Schlaifer (1961), who also studied in detail many of the families which will be presented here. Other theories of statistical inference in which the likelihood function is of fundamental importance are discussed by Fisher (1956), Barnard (1949, 1962, 1967), Birnbaum (1962), Barnard, Jenkins, and Winsten (1962), and Stein (1962b). These theories are related to, but distinct from, the Bayesian approach which is being presented in this book.

### 9.4 CONJUGATE FAMILIES FOR SAMPLES FROM VARIOUS STANDARD DISTRIBUTIONS

In this section we shall present conjugate families of distributions for samples from Poisson, negative binomial, and exponential distributions.

**Theorem 1** Suppose that  $X_1, \dots, X_n$  is a random sample from a Poisson distribution with an unknown value of the mean  $W$ . Suppose also that the prior distribution of  $W$  is a gamma distribution with parameters  $\alpha$  and  $\beta$  such that  $\alpha > 0$  and  $\beta > 0$ . Then the posterior distribution of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is a gamma distribution with parameters  $\alpha + \sum_{i=1}^n x_i$  and  $\beta + n$ .

*Proof* Let  $f_n(x_1, \dots, x_n|w)$  denote the value of the likelihood function when  $W = w$  and  $X_i = x_i$  ( $i = 1, \dots, n$ ), and let  $\xi$  denote the prior p.d.f. of  $W$ . If  $y = \sum_{i=1}^n x_i$ , then it follows from the hypotheses of the theorem that for  $w > 0$ ,

$$f_n(x_1, \dots, x_n|w) \propto w^y e^{-nw} \quad (1)$$

and

$$\xi(w) \propto w^{\alpha-1} e^{-\beta w}. \quad (2)$$

If  $\xi(\cdot|x_1, \dots, x_n)$  denotes the posterior p.d.f. of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ), then

$$\xi(w|x_1, \dots, x_n) \propto f_n(x_1, \dots, x_n|w)\xi(w). \quad (3)$$

It now follows from the relations (1) to (3) that

$$\xi(w|x_1, \dots, x_n) \propto w^{\alpha+y-1} e^{-(\beta+n)w}. \quad (4)$$

It can be seen from the relation (4) that the posterior p.d.f. of  $W$  is that of a gamma distribution with parameters  $\alpha + y$  and  $\beta + n$ . ■

The coefficient of variation, as defined by the ratio (5) of Sec. 3.6, is commonly used as a measure of the dispersion of the distribution of a positive random variable. It follows from Eq. (4) of Sec. 4.8 and from Theorem 1 that the coefficient of variation of the posterior distribution of  $W$  is  $(\alpha + \sum_{i=1}^n x_i)^{-1}$ .

Let  $\epsilon$  be a fixed positive number, and suppose that observations are to be drawn from the Poisson distribution until the coefficient of variation of the posterior distribution of  $W$  is not greater than  $\epsilon$ . Then sampling must be continued until the inequality  $\alpha + \sum_{i=1}^n x_i \geq 1/\epsilon^2$  has been established.

The next theorem describes a conjugate family of distributions for a sample from a negative binomial distribution. The proof of this theorem serves as Exercise 17 at the end of this chapter.

**Theorem 2** Suppose that  $X_1, \dots, X_n$  is a random sample from a negative binomial distribution with parameters  $r$  and  $W$ , where  $r$  has a specified value ( $r > 0$ ) and the value of  $W$  is unknown. Suppose also that the prior distribution of  $W$  is a beta distribution with parameters  $\alpha$  and  $\beta$  such that  $\alpha > 0$  and  $\beta > 0$ . Then the posterior distribution of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is a beta distribution with parameters  $\alpha + rn$  and  $\beta + \sum_{i=1}^n x_i$ .

Theorem 2, together with Theorem 1 of Sec. 9.2, provides an opportunity to illustrate, for samples from a Bernoulli distribution, an important property which was discussed briefly in Sec. 8.12. Suppose that each item in a large population of manufactured items can be classified as either defective or nondefective. Suppose also that the proportion  $W$  of defective items in the population is unknown but that  $W$  has a specified prior distribution. Suppose further that a sample of items is to be selected from the population and inspected. Consider the following four sampling methods, any one of which might be used in obtaining the

sample:

1. A random sample of  $n$  items is selected from the population, where  $n$  is a fixed positive integer.
2. Items are selected at random from the population one at a time until exactly  $y$  defective items have been obtained, where  $y$  is a fixed positive integer.
3. Items are selected at random from the population one at a time until the inspector is called away to another problem.
4. Items are selected at random from the population until the inspector feels that he has accumulated sufficient information about  $W$ .

For any one of these four methods, let  $\mathbf{x}$  denote the vector of all observed values which were found during the sampling process, and let  $g(\mathbf{x}|w)$  denote the value of the likelihood function when  $W = w$ . Furthermore, let  $n$  denote the total number of items which were inspected, and let  $y$  denote the number of these items which were defective. Then, regardless of which method of sampling was used, the following basic relation must hold:

$$g(\mathbf{x}|w) \propto w^y (1-w)^{n-y}. \quad (5)$$

It follows that the posterior distribution of  $W$  will be the same regardless of which method of sampling was used. In other words, the posterior distribution of  $W$  will depend only on the total number  $n$  of items which were inspected and the number  $y$  which were defective, and not on the sampling method which led to these results.

The next theorem describes a conjugate family of distributions for a sample from an exponential distribution. The proof of this theorem serves as Exercise 20.

**Theorem 3** Suppose that  $X_1, \dots, X_n$  is a random sample from an exponential distribution with an unknown value of the parameter  $W$ . Suppose also that the prior distribution of  $W$  is a gamma distribution with parameters  $\alpha$  and  $\beta$  such that  $\alpha > 0$  and  $\beta > 0$ . Then the posterior distribution of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is a gamma distribution with parameters  $\alpha + n$  and  $\beta + \sum_{i=1}^n x_i$ .

### 9.5 CONJUGATE FAMILIES FOR SAMPLES FROM A NORMAL DISTRIBUTION

We shall begin by considering a normal distribution for which the value of the precision or, equivalently, the value of the variance is specified.

**Theorem 1** Suppose that  $X_1, \dots, X_n$  is a random sample from a normal distribution with an unknown value of the mean  $W$  and a specified value of the precision  $r$  ( $r > 0$ ). Suppose also that the prior distribution of  $W$  is a normal distribution with mean  $\mu$  and precision  $\tau$  such that  $-\infty < \mu < \infty$  and  $\tau > 0$ . Then the posterior distribution of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is a normal distribution with mean  $\mu'$  and precision  $\tau + nr$ , where

$$\mu' = \frac{\tau\mu + nr\bar{x}}{\tau + nr}. \quad (1)$$

*Proof* For  $-\infty < w < \infty$ , the likelihood function  $f_n(x_1, \dots, x_n|w)$  satisfies the following relation:

$$f_n(x_1, \dots, x_n|w) \propto \exp \left[ -\frac{r}{2} \sum_{i=1}^n (x_i - w)^2 \right]. \quad (2)$$

However,

$$\sum_{i=1}^n (x_i - w)^2 = n(w - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x})^2. \quad (3)$$

Since the final term in Eq. (3) does not involve  $w$ , we may rewrite relation (2) as follows:

$$f_n(x_1, \dots, x_n|w) \propto \exp \left[ -\frac{nr}{2} (w - \bar{x})^2 \right]. \quad (4)$$

The prior p.d.f.  $\xi$  of  $W$  satisfies the relation

$$\xi(w) \propto \exp \left[ -\frac{\tau}{2} (w - \mu)^2 \right], \quad (5)$$

and the posterior p.d.f.  $\xi(\cdot | x_1, \dots, x_n)$  of  $W$  will be proportional to the product of the functions specified by the relations (4) and (5). However, it can be shown that

$$\tau(w - \mu)^2 + nr(w - \bar{x})^2 = (\tau + nr)(w - \mu')^2 + \frac{nr\tau(\bar{x} - \mu)^2}{\tau + nr}. \quad (6)$$

Since the final term in Eq. (6) does not involve  $w$ , it can be included in the proportionality factor, and we obtain the relation

$$\xi(w|x_1, \dots, x_n) \propto \exp \left[ -\frac{\tau + nr}{2} (w - \mu')^2 \right]. \quad (7)$$

Here  $\mu'$  is specified by Eq. (1). It follows from the relation (7) that the posterior distribution of  $W$  is a normal distribution with mean  $\mu'$  and precision  $\tau + nr$ . ■

Theorem 1 reveals the advantages of expressing our results in terms of the precision rather than the variance. The mean  $\mu'$  of the posterior

distribution of  $W$  can be written in the following form:

$$\mu' = \frac{n\tau}{\tau + n\tau} \bar{x} + \frac{\tau}{\tau + n\tau} \mu. \quad (8)$$

It is seen that  $\mu'$  is a weighted average of  $\bar{x}$  and  $\mu$ , where  $\bar{x}$  is the value of the sample mean and  $\mu$  is the mean of the prior distribution of  $W$ . Therefore, we may conveniently regard the mean of the posterior distribution as a weighted average of an estimate of  $W$  formed from the sample and an estimate of  $W$  formed from the prior distribution. The weights of  $\bar{x}$  and  $\mu$  in this average are proportional to  $n\tau$  and  $\tau$ , where  $n\tau$  is the precision of the conditional distribution of the sample mean for any given value of  $W$  and  $\tau$  is the precision of the prior distribution of  $W$ . The larger the sample size  $n$  and the higher the precision  $\tau$  of each observation, the greater will be the weight that is given to  $\bar{x}$ .

The form of the precision of the posterior distribution of  $W$  is particularly simple. The precision increases by the amount  $\tau$  with each observation that is taken, regardless of the observed values. Therefore, as the number of observations increases, the distribution of  $W$  becomes more concentrated around its mean. Moreover, the concentration must increase in a fixed, predetermined way, while the values of the mean will depend on the observed values.

In the next theorem, we shall consider a normal distribution for which the value of the mean is specified but the value of the precision is unknown. The proof of this theorem serves as Exercise 24.

**Theorem 2** Suppose that  $X_1, \dots, X_n$  is a random sample from a normal distribution with a specified value of the mean  $m$  ( $-\infty < m < \infty$ ) and an unknown value of the precision  $W$ . Suppose also that the prior distribution of  $W$  is a gamma distribution with parameters  $\alpha$  and  $\beta$  such that  $\alpha > 0$  and  $\beta > 0$ . Then the posterior distribution of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is a gamma distribution with parameters  $\alpha + (n/2)$  and  $\beta'$ , where

$$\beta' = \beta + \frac{1}{2} \sum_{i=1}^n (x_i - m)^2. \quad (9)$$

For a gamma distribution with parameters  $\alpha$  and  $\beta$ , the coefficient of variation is  $\alpha^{-1}$ . Therefore, it follows from Theorem 2 that the coefficient of variation of the posterior distribution of  $W$  must decrease in a fixed, predetermined way as the sample size  $n$  increases.

## 9.6 SAMPLING FROM A NORMAL DISTRIBUTION WITH UNKNOWN MEAN AND UNKNOWN PRECISION

We shall now consider the important problem of sampling from a normal distribution for which both the mean and the precision are unknown.

A conjugate family for this problem must be a family of bivariate distributions.

**Theorem 1** Suppose that  $X_1, \dots, X_n$  is a random sample from a normal distribution with an unknown value of the mean  $M$  and an unknown value of the precision  $R$ . Suppose also that the prior joint distribution of  $M$  and  $R$  is as follows: The conditional distribution of  $M$  when  $R = r$  ( $r > 0$ ) is a normal distribution with mean  $\mu$  and precision  $\tau$  such that  $-\infty < \mu < \infty$  and  $\tau > 0$ , and the marginal distribution of  $R$  is a gamma distribution with parameters  $\alpha$  and  $\beta$  such that  $\alpha > 0$  and  $\beta > 0$ . Then the posterior joint distribution of  $M$  and  $R$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is as follows: The conditional distribution of  $M$  when  $R = r$  is a normal distribution with mean  $\mu'$  and precision  $(\tau + n)r$ , where

$$\mu' = \frac{\tau\mu + n\bar{x}}{\tau + n}, \quad (1)$$

and the marginal distribution of  $R$  is a gamma distribution with parameters  $\alpha + (n/2)$  and  $\beta'$ , where

$$\beta' = \beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{\tau n(\bar{x} - \mu)^2}{2(\tau + n)}. \quad (2)$$

*Proof* For  $-\infty < m < \infty$  and  $r > 0$ , let  $f_n(x_1, \dots, x_n | m, r)$  denote the value of the likelihood function when  $M = m$ ,  $R = r$ , and  $X_i = x_i$  ( $i = 1, \dots, n$ ), and let  $\xi$  denote the prior p.d.f. of  $M$  and  $R$ . Then

$$f_n(x_1, \dots, x_n | m, r) \propto r^{n/2} \exp \left[ -\frac{r}{2} \sum_{i=1}^n (x_i - m)^2 \right] \quad (3)$$

and

$$\xi(m, r) \propto r^{\frac{1}{2}} e^{-(r/2)(m-\mu)^2} r^{\alpha-1} e^{-\beta r}. \quad (4)$$

The posterior p.d.f.  $\xi(\cdot | x_1, \dots, x_n)$  of  $M$  and  $R$  will be proportional to the product of the right sides of the relations (3) and (4). It follows from Eqs. (3) and (6) of Sec. 9.5 that this p.d.f. can be specified by the relation

$$\xi(m, r | x_1, \dots, x_n) \propto \left\{ r^{\frac{1}{2}} \exp \left[ -\frac{(\tau + n)r}{2} (m - \mu')^2 \right] \right\}^{(\tau + n/2 - 1) e^{-\beta r}}. \quad (5)$$

Here,  $\mu'$  is defined by Eq. (1) and  $\beta'$  by Eq. (2).

The function inside the braces in relation (5), when regarded as a function of  $m$ , must be proportional to the conditional p.d.f. of  $M$  when  $R$  is known since the variable  $m$  does not appear inside the set of parentheses on the right. However, for each fixed value of  $r$ , the function inside the braces is proportional to the p.d.f. of a normal distribution for which the mean and the precision are as given in the statement of the

suddenly told that the value of  $R$  is  $r$ , where  $r$  is a large number, then because of this knowledge about  $R$ , he can now also specify the value of  $M$  with high precision. On the other hand, if the statistician is suddenly told that the value of  $R$  is  $r'$ , where  $r'$  is a small number, then the statistician's distribution for  $M$  will still have a very large variance. Again, this is not an important deficiency of the conjugate family, for suppose that the statistician learns the value of even a single observation drawn at random from the normal distribution. If the precision  $R$  of the distribution is large, then the observation provides very precise information about the value of the mean  $M$ ; whereas if the precision  $R$  is small, then the observation provides little information about the value of  $M$ .

**A Numerical Example**

Consider a normal distribution in which the values of the mean  $M$  and the precision  $R$  are unknown, and suppose that a statistician wishes to select a normal-gamma distribution from the conjugate family to represent the prior distribution of  $M$  and  $R$ . If he specifies that  $E(M) = 2$ ,  $\text{Var}(M) = 5$ ,  $E(R) = 3$ , and  $\text{Var}(R) = 3$ , what values should be selected for the parameters  $\mu$ ,  $\tau$ ,  $\alpha$ , and  $\beta$  of the prior distribution? Since  $R$  has a gamma distribution with parameters  $\alpha$  and  $\beta$ , the values  $\alpha = 3$  and  $\beta = 1$  can be found from Eq. (4) of Sec. 4.8. Furthermore,  $\mu = 2$  since  $E(M) = \mu$ . Finally, since  $M$  has a  $t$  distribution with  $2\alpha$  degrees of freedom and precision  $\alpha r/\beta$ , it follows from Eq. (4) of Sec. 4.12 that

$$\text{Var}(M) = \frac{\beta}{\tau(\alpha - 1)} \tag{9}$$

Hence,  $\tau = 0.1$ . This value completes the specification of the prior distribution.

Now suppose that a random sample of 10 observations is taken from the given normal distribution and it is found that for these 10 values,  $\bar{x} = 4.20$  and  $\Sigma_{i=1}^{10} (x_i - \bar{x})^2 = 5.40$ . Then, by Theorem 1, the parameters  $\mu'$ ,  $\tau'$ ,  $\alpha'$ , and  $\beta'$  of the posterior distribution of  $M$  and  $R$  are  $\mu' = 4.18$ ,  $\tau' = 10.1$ ,  $\alpha' = 8$ , and  $\beta' = 3.94$ . It follows from these values that the means and variances of  $M$  and  $R$  are now  $E(M) = 4.18$ ,  $\text{Var}(M) = 0.056$ ,  $E(R) = 2.03$ , and  $\text{Var}(R) = 0.515$ .

Next, suppose that 10 more observations are now taken from the given normal distribution and it is found that for these 10 values,  $\bar{x} = 4.48$  and  $\Sigma_{i=1}^{10} (x_i - \bar{x})^2 = 5.82$ . By considering the posterior distribution which we just found as the prior distribution for these new observations, the parameters  $\mu''$ ,  $\tau''$ ,  $\alpha''$ , and  $\beta''$  of the new posterior distribution of  $M$  and  $R$  become  $\mu'' = 4.33$ ,  $\tau'' = 20.1$ ,  $\alpha'' = 13$ , and  $\beta'' = 7.08$ . There-

theorem. It now follows that the function inside the parentheses on the right must be proportional to the marginal p.d.f. of  $R$ . Therefore, the marginal distribution of  $R$  is a gamma distribution for which the parameters are as given in the statement of the theorem. ■

When the joint p.d.f.  $\xi$  of  $M$  and  $R$  is a normal-gamma p.d.f., as specified by the relation (4), the conditional distribution of  $M$  for any given value  $R = r$  will be normal but the marginal distribution of  $M$  will not be normal. The marginal p.d.f.  $\xi_M$  of  $M$  is defined by the equation

$$\xi_M(m) = \int_0^\infty \xi(m, r) dr \quad \text{for } -\infty < m < \infty. \tag{6}$$

Therefore, if we make use of the proportionality symbol and drop all factors which do not involve  $m$ , it follows from the relation (4) that  $\xi_M$  has the form

$$\xi_M(m) \propto \left[ \beta + \frac{\tau}{2} (m - \mu)^2 \right]^{-\alpha-\frac{1}{2}} \tag{7}$$

or

$$\xi_M(m) \propto \left[ 1 + \frac{1}{2\alpha} \frac{\alpha\tau(m - \mu)^2}{\beta} \right]^{-(2\alpha+1)/2}. \tag{8}$$

From a comparison of the function given in the relation (8) with the p.d.f. of the  $t$  distribution specified by Eq. (5) of Sec. 4.12, it is seen that the marginal distribution of  $M$  is a  $t$  distribution with  $2\alpha$  degrees of freedom, location parameter  $\mu$ , and precision  $\alpha r/\beta$ . The posterior marginal distribution of  $M$  is obtained by replacing  $\mu$ ,  $\tau$ ,  $\alpha$ , and  $\beta$  by their posterior values as given in Theorem 1. Hence, although the number of degrees of freedom  $2\alpha + n$  of the posterior  $t$  distribution will not depend on the observed values  $x_1, \dots, x_n$ , both the location parameter and the precision of the posterior distribution will depend on these values.

An interesting feature of the conjugate family of joint distributions of  $M$  and  $R$  specified by Theorem 1 is the following: For any normal-gamma distribution in this family, the variables  $M$  and  $R$  are dependent. There is no joint distribution in the family such that  $M$  has a normal distribution,  $R$  has a gamma distribution, and  $M$  and  $R$  are independent. This is not an important deficiency of the family. In fact, even if the prior distribution of  $M$  and  $R$  specified that these variables were independent, their posterior distribution after the value of a single observation had been noted would specify that they were dependent.

Another interesting feature is that for each distribution in this conjugate family, the precision of the conditional distribution of  $M$  when  $R = r$  must be proportional to  $r$ . In other words, the prior distribution of  $M$  and  $R$  must have the following property: If the statistician is

fore, the means and variances of  $M$  and  $R$  now have the values  $E(M) = 4.33$ ,  $\text{Var}(M) = 0.029$ ,  $E(R) = 1.84$ , and  $\text{Var}(R) = 0.260$ .

Since  $M$  now has a  $t$  distribution with 26 degrees of freedom, location parameter 4.33, and precision 36.9, it follows from tables of the  $t$  distribution that

$$\text{Pr}\{-2.056 \leq (36.9)^{1/2}(M - 4.33) \leq 2.056\} = 0.95. \quad (10)$$

Upon simplification, Eq. (10) reduces to the equation

$$\text{Pr}\{3.99 \leq M \leq 4.67\} = 0.95. \quad (11)$$

### 9.7 SAMPLING FROM A UNIFORM DISTRIBUTION

In this section we shall describe conjugate families of distributions for samples from a uniform distribution such that either the value of one end point or the values of both end points are unknown.

**Theorem 1** Suppose that  $X_1, \dots, X_n$  is a random sample from a uniform distribution on the interval  $(0, W)$ , where the value of  $W$  is unknown. Suppose also that the prior distribution of  $W$  is a Pareto distribution with parameters  $w_0$  and  $\alpha$  such that  $w_0 > 0$  and  $\alpha > 0$ . Then the posterior distribution of  $W$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is a Pareto distribution with parameters  $w'_0$  and  $\alpha + n$ , where

$$w'_0 = \max\{w_0, x_1, \dots, x_n\}. \quad (1)$$

*Proof* For  $w > w_0$ , the prior p.d.f.  $\xi$  of  $W$  is of the following form:

$$\xi(w) \propto \frac{1}{w^{\alpha+1}}. \quad (2)$$

Furthermore,  $\xi(w) = 0$  for  $w \leq w_0$ . The likelihood function  $f_n(x_1, \dots, x_n|w)$  is specified by Eq. (13) of Sec. 9.1.

It follows from these relations that the posterior p.d.f.  $\xi(w|x_1, \dots, x_n)$  of  $W$  will be positive only for values of  $w$  such that  $w > w_0$  and  $w > \max\{x_1, \dots, x_n\}$ . Therefore,  $\xi(w|x_1, \dots, x_n) > 0$  only if  $w > w'_0$  where  $w'_0$  is defined by Eq. (1). Furthermore, for  $w > w'_0$ ,

$$\xi(w|x_1, \dots, x_n) \propto \frac{1}{w^{\alpha+n+1}}. \quad (3)$$

It is seen from the relation (3) that the posterior distribution of  $W$  must be a Pareto distribution whose parameters are as specified in the statement of the theorem. ■

In the next theorem, we shall consider a uniform distribution in which both end points are unknown. The proof of this theorem serves as Exercise 32.

**Theorem 2** Suppose that  $X_1, \dots, X_n$  is a random sample from a uniform distribution on the interval  $(W_1, W_2)$ , where the values of  $W_1$  and  $W_2$  are unknown. Suppose also that the prior joint distribution of  $W_1$  and  $W_2$  is a bilateral bivariate Pareto distribution with parameters  $r_1, r_2$ , and  $\alpha$  such that  $r_1 < r_2$  and  $\alpha > 0$ . Then the posterior joint distribution of  $W_1$  and  $W_2$  when  $X_i = x_i$  ( $i = 1, \dots, n$ ) is a bilateral bivariate Pareto distribution with parameters  $r'_1, r'_2$ , and  $\alpha + n$ , where

$$r'_1 = \min\{r_1, x_1, \dots, x_n\} \quad \text{and} \quad r'_2 = \max\{r_2, x_1, \dots, x_n\}. \quad (4)$$

Let us now consider a numerical example which illustrates the use of Theorem 2. Suppose that if a certain type of atomic particle is passed through a container of water, the horizontal deflection of the particle, as measured in appropriate units, has a uniform distribution on the interval  $(W_1, W_2)$ , where the values of  $W_1$  and  $W_2$  are unknown. Suppose also that it is known that  $W_1 < -0.4$  and  $W_2 > 0.1$ . If bounds of this type cannot be specified in advance by the statistician, then they can certainly be specified after the horizontal deflections  $y_1$  and  $y_2$  of two particles have been observed. In fact, since  $y_1$  and  $y_2$  must lie in the interval  $(W_1, W_2)$ , then, if we assume that  $y_1 < y_2$ , it follows that  $W_1 < y_1$  and  $W_2 > y_2$ .

Suppose further that the statistician wishes to select a bilateral bivariate Pareto distribution to represent the prior distribution of  $W_1$  and  $W_2$  and that he expects the length  $W_2 - W_1$  of the interval  $(W_1, W_2)$  to be about 2.5 units. What values of the parameters  $r_1, r_2$ , and  $\alpha$  of the prior distribution should be selected? It follows from the above bounds on  $W_1$  and  $W_2$  that  $r_1 = -0.4$  and  $r_2 = 0.1$ . Furthermore, by Eq. (2) of Sec. 5.7, if  $\alpha > 1$ , then

$$E(W_2 - W_1) = \frac{(\alpha + 1)(r_2 - r_1)}{\alpha - 1}. \quad (5)$$

If it is assumed that  $E(W_2 - W_1) = 2.5$ , then  $\alpha = 1.5$ . This value completes the specification of the prior distribution.

Suppose now that the deflections of five particles are observed and found to have the values  $-0.27, -0.45, -0.36, -0.12$ , and  $0.47$ . Since the minimum of these five values is  $-0.45$  and the maximum is  $0.47$ , it follows from Theorem 2 that the values of the parameters  $r'_1, r'_2$  and  $\alpha'$  of the posterior distribution of  $W$  are  $r'_1 = -0.45, r'_2 = 0.47$ , and  $\alpha' = 6.5$ . Therefore, it is now known that  $W_1 < -0.45$  and  $W_2 > 0.47$ . Furthermore, by Eq. (5), the expected length  $E(W_2 - W_1)$  is now 1.25.

Next, suppose that the deflections of five more particles are found to have the values  $-0.39, -0.07, 0.43, 0.01$ , and  $-0.14$ . The parameters  $r''_1, r''_2$ , and  $\alpha''$  of the new posterior distribution of  $W_1$  and  $W_2$  are  $r''_1 = r'_1$ ,

$r'_2 = r'_2$ , and  $\alpha'' = 11.5$ . Therefore, by Eq. (5), the expected length  $E(W_2 - W_1)$  is now 1.10.

Since the interval (-0.45, 0.47), whose length is 0.92, must be wholly contained in the interval  $(W_1, W_2)$  and since the expected length of the interval  $(W_1, W_2)$  is itself 1.10, the statistician now has relatively precise information about the values of  $W_1$  and  $W_2$ . In fact, by Eq. (3) of Sec. 5.7, it follows that the variances of  $W_1$  and  $W_2$  now have the common value 0.0093.

### 9.3 A CONJUGATE FAMILY FOR MULTINOMIAL OBSERVATIONS

In the next theorem, it is shown that the family of Dirichlet distributions is a conjugate family for observations which have a multinomial distribution.

**Theorem 1** Suppose that the random vector  $\mathbf{X} = (X_1, \dots, X_k)'$  has a multinomial distribution with parameters  $n$  and  $\mathbf{W} = (W_1, \dots, W_k)'$ , where  $n$  is a specified positive integer and the values of the components of the vector  $\mathbf{W}$  are unknown. Suppose also that the prior distribution of  $\mathbf{W}$  is a Dirichlet distribution with parametric vector  $\alpha = (\alpha_1, \dots, \alpha_k)'$  such that  $\alpha_i > 0$  ( $i = 1, \dots, k$ ). Then the posterior distribution of  $\mathbf{W}$  when  $X_i = x_i$  ( $i = 1, \dots, k$ ) is a Dirichlet distribution with parametric vector  $\alpha^* = (\alpha_1 + x_1, \dots, \alpha_k + x_k)'$ .

*Proof* Let  $\Omega$  denote the set of points  $\mathbf{w} = (w_1, \dots, w_k)'$  such that  $w_i > 0$  ( $i = 1, \dots, k$ ) and  $w_1 + \dots + w_k = 1$ . Then for any given value  $\mathbf{w}$  of  $\mathbf{W}$  such that  $\mathbf{w} \in \Omega$ , the likelihood function  $f(x_1, \dots, x_k | \mathbf{w})$  satisfies the following relation:

$$f(x_1, \dots, x_k | \mathbf{w}) \propto \prod_{i=1}^k w_i^{x_i}. \quad (1)$$

Furthermore, for  $\mathbf{w} \in \Omega$ , the prior p.d.f.  $\xi$  of  $\mathbf{W}$  satisfies the relation

$$\xi(\mathbf{w}) \propto \prod_{i=1}^k w_i^{\alpha_i - 1}. \quad (2)$$

Therefore, for  $\mathbf{w} \in \Omega$ , the posterior p.d.f.  $\xi(\cdot | x_1, \dots, x_k)$  of  $\mathbf{W}$  must satisfy the relation

$$\xi(\mathbf{w} | x_1, \dots, x_k) \propto \prod_{i=1}^k w_i^{\alpha_i + x_i - 1}. \quad (3)$$

The function in the relation (3) is proportional to the p.d.f. of a Dirichlet distribution whose parametric vector is as specified in the statement of the theorem. ■

As an example, suppose that in a large shipment of manufactured items, there are items of  $k$  different types. For  $i = 1, \dots, k$ , let  $W_i$  denote the proportion of items which are of type  $i$ , and assume that the prior distribution of the vector  $\mathbf{W} = (W_1, \dots, W_k)'$  is a Dirichlet distribution with parametric vector  $\alpha = (\alpha_1, \dots, \alpha_k)'$ . If items are selected at random from the shipment, one at a time, then it follows from Theorem 1 that the posterior distribution of  $\mathbf{W}$  at each stage will be a Dirichlet distribution, and for  $i = 1, \dots, k$ , the  $i$ th component of the parametric vector  $\alpha$  will be increased by 1 unit each time an item of type  $i$  is selected.

### 9.9 CONJUGATE FAMILIES FOR SAMPLES FROM A MULTIVARIATE NORMAL DISTRIBUTION

In the remainder of this chapter we shall consider problems in which samples are taken from a nonsingular,  $k$ -dimensional multivariate normal distribution ( $k \geq 1$ ). In each problem, the mean vector of the distribution must be a  $k$ -dimensional vector, i.e., a point in  $R^k$ , and the precision matrix of the distribution must be a symmetric  $k \times k$  positive definite matrix. Any observation  $\mathbf{X}$  from this distribution will be a  $k$ -dimensional random vector whose value  $\mathbf{x}$  will lie in the space  $R^k$ . The results to be derived will be generalizations of those obtained in Secs. 9.5 and 9.6, in which we dealt with samples from a univariate normal distribution. We shall begin by discussing the problem of sampling from a distribution for which the precision matrix  $\mathbf{r}$  is specified. If  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are the values of a sample of observations  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , we shall, as usual, let  $\bar{\mathbf{x}}$  denote the sample mean vector as defined by the equation

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i. \quad (1)$$

**Theorem 1** Suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  is a random sample from a multivariate normal distribution with an unknown value of the mean vector  $\mathbf{M}$  and a specified precision matrix  $\mathbf{r}$ . Suppose also that the prior distribution of  $\mathbf{M}$  is a multivariate normal distribution with mean vector  $\boldsymbol{\mu}$  and precision matrix  $\boldsymbol{\tau}$  such that  $\boldsymbol{\mu} \in R^k$  and  $\boldsymbol{\tau}$  is a symmetric positive definite matrix. Then the posterior distribution of  $\mathbf{M}$  when  $\mathbf{X}_i = \mathbf{x}_i$  ( $i = 1, \dots, n$ ) is a multivariate normal distribution with mean vector  $\boldsymbol{\mu}^*$  and precision matrix  $\boldsymbol{\tau} + n\mathbf{r}$ , where

$$\boldsymbol{\mu}^* = (\boldsymbol{\tau} + n\mathbf{r})^{-1}(\boldsymbol{\tau}\boldsymbol{\mu} + n\bar{\mathbf{x}}). \quad (2)$$

*Proof* For  $\mathbf{M} = \mathbf{m}$  and  $\mathbf{X}_i = \mathbf{x}_i$  ( $i = 1, \dots, n$ ), the likelihood function

$f_n(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{m})$  satisfies the following relation:

$$f_n(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{m}) \propto \exp \left[ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})' \mathbf{r} (\mathbf{x}_i - \mathbf{m}) \right]. \quad (3)$$

However,

$$\begin{aligned} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})' \mathbf{r} (\mathbf{x}_i - \mathbf{m}) \\ = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{r} (\mathbf{x}_i - \bar{\mathbf{x}}) + n(\mathbf{m} - \bar{\mathbf{x}})' \mathbf{r} (\mathbf{m} - \bar{\mathbf{x}}). \end{aligned} \quad (4)$$

Therefore, the relation (3) can be rewritten in the following form:

$$f_n(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{m}) \propto \exp \left[ -\frac{1}{2} (\mathbf{m} - \bar{\mathbf{x}})' (nr) (\mathbf{m} - \bar{\mathbf{x}}) \right]. \quad (5)$$

The prior p.d.f.  $\xi$  of  $\mathbf{M}$  satisfies the relation

$$\xi(\mathbf{m}) \propto \exp \left[ -\frac{1}{2} (\mathbf{m} - \mathbf{y})' \boldsymbol{\tau} (\mathbf{m} - \mathbf{y}) \right]. \quad (6)$$

The posterior p.d.f.  $\xi(\cdot | \mathbf{x}_1, \dots, \mathbf{x}_n)$  of  $\mathbf{M}$  is proportional to the product of the functions specified in the relations (5) and (6). However, it can be verified that

$$\begin{aligned} (\mathbf{m} - \mathbf{y})' \boldsymbol{\tau} (\mathbf{m} - \mathbf{y}) + (\mathbf{m} - \bar{\mathbf{x}})' (nr) (\mathbf{m} - \bar{\mathbf{x}}) \\ = (\mathbf{m} - \mathbf{y}^*)' (\boldsymbol{\tau} + nr) (\mathbf{m} - \mathbf{y}^*) \\ + (\text{terms which do not involve } \mathbf{m}). \end{aligned} \quad (7)$$

Since the terms in Eq. (7) which do not involve  $\mathbf{m}$  can be absorbed in the proportionality factor, we obtain the following relation:

$$\xi(\mathbf{m} | \mathbf{x}_1, \dots, \mathbf{x}_n) \propto \exp \left[ -\frac{1}{2} (\mathbf{m} - \mathbf{y}^*)' (\boldsymbol{\tau} + nr) (\mathbf{m} - \mathbf{y}^*) \right]. \quad (8)$$

The p.d.f. specified by the relation (8) is that of a multivariate normal distribution for which the mean vector and the precision matrix are as specified in the statement of the theorem. ■

The analogy between the multivariate results of this theorem and the univariate results of Theorem 1 of Sec. 9.5 is evident, and the discussion which was given following that theorem is relevant here.

Now suppose that we are sampling from a multivariate normal distribution with a specified mean vector but an unknown precision matrix.

**Theorem 2** Suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  is a random sample from a multivariate normal distribution with a specified mean vector  $\mathbf{m}$  and an unknown value of the precision matrix  $\mathbf{R}$ . Suppose also that the prior distribution of  $\mathbf{R}$  is a Wishart distribution with  $\alpha$  degrees of freedom and precision matrix  $\boldsymbol{\tau}$  such that  $\alpha > k - 1$  and  $\boldsymbol{\tau}$  is a symmetric positive definite matrix. Then the posterior distribution of  $\mathbf{R}$  when  $\mathbf{X}_i = \mathbf{x}_i$  ( $i = 1, \dots, n$ ) is a Wishart

distribution with  $\alpha + n$  degrees of freedom and precision matrix  $\boldsymbol{\tau}^*$ , where

$$\boldsymbol{\tau}^* = \boldsymbol{\tau} + \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})'. \quad (9)$$

*Proof* The likelihood function  $f_n(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{r})$  satisfies the following relation:

$$f_n(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{r}) \propto |\mathbf{r}|^{n/2} \exp \left[ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})' \mathbf{r} (\mathbf{x}_i - \mathbf{m}) \right]. \quad (10)$$

The exponent of  $e$  in relation (10) is a real number which may be regarded as a  $1 \times 1$  matrix. Therefore, by Eqs. (1) and (2) of Sec. 3.5, we obtain the following relation:

$$\begin{aligned} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})' \mathbf{r} (\mathbf{x}_i - \mathbf{m}) &= \text{tr} \left[ \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})' \right] \\ &= \text{tr} \left\{ \left[ \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})' \right] \mathbf{r} \right\}. \end{aligned} \quad (11)$$

Furthermore, the p.d.f.  $\xi$  of  $\mathbf{R}$  satisfies the relation

$$\xi(\mathbf{r}) \propto |\mathbf{r}|^{(c-k-1)/2} \exp \left[ -\frac{1}{2} \text{tr} (\boldsymbol{\tau} \mathbf{r}) \right]. \quad (12)$$

Since the posterior p.d.f. of  $\mathbf{R}$  is proportional to the product of the functions given in the relations (10) and (12), it follows from Eq. (11) that the posterior distribution of  $\mathbf{R}$  must be a Wishart distribution as specified in the statement of the theorem. ■

This theorem is a straightforward multivariate generalization of Theorem 2 of Sec. 9.5. The analogy between the two theorems is slightly obscured, however, by the fact that a one-dimensional Wishart distribution with  $\alpha$  degrees of freedom and precision matrix  $\beta$ , where  $\beta$  is simply a positive number, is the same as a gamma distribution with parameters  $\alpha/2$  and  $\beta/2$ . Although a slightly different definition of the parameters of one of these distributions would have removed this discrepancy, the definitions which we are using here are the traditional ones.

It should be noted that the Wishart p.d.f. is well defined by the relation (12) even when the number of degrees of freedom  $\alpha$  is not an integer, provided that  $\alpha > k - 1$ . However, if  $\alpha$  is chosen to be an integer in the prior distribution of  $\mathbf{R}$ , then the number of degrees of freedom of the posterior distribution will also be an integer.

## 9.10 MULTIVARIATE NORMAL DISTRIBUTIONS WITH UNKNOWN MEAN VECTOR AND UNKNOWN PRECISION MATRIX

We shall now extend Theorem 1 of Sec. 9.6 by considering the problem of sampling from a multivariate normal distribution for which both the mean

vector and the precision matrix are unknown. If  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are the values of a sample of observations  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , we shall let  $\mathbf{s}$  denote the symmetric  $k \times k$  nonnegative definite matrix which is defined by the equation

$$\mathbf{s} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'. \quad (1)$$

**Theorem 1** Suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  is a random sample from a multivariate normal distribution with an unknown value of the mean vector  $\mathbf{M}$  and an unknown value of the precision matrix  $\mathbf{R}$ . Suppose also that the prior joint distribution of  $\mathbf{M}$  and  $\mathbf{R}$  is as follows: The conditional distribution of  $\mathbf{M}$  when  $\mathbf{R} = \mathbf{r}$  is a multivariate normal distribution with mean vector  $\mathbf{y}$  and precision matrix  $\nu\mathbf{r}$  such that  $\nu \in R^k$  and  $\nu > 0$ , and the marginal distribution of  $\mathbf{R}$  is a Wishart distribution with  $\alpha$  degrees of freedom and precision matrix  $\tau$  such that  $\alpha > k - 1$  and  $\tau$  is a symmetric positive definite matrix. Then the posterior joint distribution of  $\mathbf{M}$  and  $\mathbf{R}$  when  $\mathbf{X}_i = \mathbf{x}_i$  ( $i = 1, \dots, n$ ) is as follows: The conditional distribution of  $\mathbf{M}$  when  $\mathbf{R} = \mathbf{r}$  is a multivariate normal distribution with mean vector  $\mathbf{y}^*$  and precision matrix  $(\nu + n)\mathbf{r}$ , where

$$\mathbf{y}^* = \frac{\nu\mathbf{y} + n\bar{\mathbf{x}}}{\nu + n}, \quad (2)$$

and the marginal distribution of  $\mathbf{R}$  is a Wishart distribution with  $\alpha + n$  degrees of freedom and precision matrix  $\tau^*$ , where

$$\tau^* = \tau + \mathbf{s} + \frac{\nu n}{\nu + n}(\mathbf{y} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{x}})'. \quad (3)$$

*Proof* When  $\mathbf{M} = \mathbf{m}$ ,  $\mathbf{R} = \mathbf{r}$ , and  $\mathbf{X}_i = \mathbf{x}_i$  ( $i = 1, \dots, n$ ), the likelihood function  $f_n(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{m}, \mathbf{r})$  is proportional to the function specified by the right side of relation (10) of Sec. 9.9. Furthermore, it follows from Eqs. (4) and (11) of Sec. 9.9 and from Eq. (1) that the sum which appears in the exponent of this function can be written in the form

$$\sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})'\mathbf{r}(\mathbf{x}_i - \mathbf{m}) = n(\mathbf{m} - \bar{\mathbf{x}})'\mathbf{r}(\mathbf{m} - \bar{\mathbf{x}}) + \text{tr}(\mathbf{s}\mathbf{r}). \quad (4)$$

The prior joint p.d.f.  $\xi$  of  $\mathbf{M}$  and  $\mathbf{R}$  satisfies the following relation:

$$\xi(\mathbf{m}, \mathbf{r}) \propto |\mathbf{r}|^{\frac{\alpha}{2}} \exp \left[ -\frac{\nu}{2}(\mathbf{m} - \mathbf{y})'\mathbf{r}(\mathbf{m} - \mathbf{y}) \right] \times |\mathbf{r}|^{(\alpha-k-1)/2} \exp \left[ -\frac{1}{2} \text{tr}(\tau\mathbf{r}) \right]. \quad (5)$$

It can be verified that

$$\begin{aligned} \nu(\mathbf{m} - \mathbf{y})'\mathbf{r}(\mathbf{m} - \mathbf{y}) + n(\mathbf{m} - \bar{\mathbf{x}})'\mathbf{r}(\mathbf{m} - \bar{\mathbf{x}}) \\ = (\nu + n)(\mathbf{m} - \mathbf{y}^*)'\mathbf{r}(\mathbf{m} - \mathbf{y}^*) + \frac{\nu n}{\nu + n}(\mathbf{y} - \bar{\mathbf{x}})'\mathbf{r}(\mathbf{y} - \bar{\mathbf{x}}). \end{aligned} \quad (6)$$

Furthermore, the final term in Eq. (6) can be rewritten as follows:

$$\frac{\nu n}{\nu + n}(\mathbf{y} - \bar{\mathbf{x}})'\mathbf{r}(\mathbf{y} - \bar{\mathbf{x}}) = \text{tr} \left[ \frac{\nu n}{\nu + n}(\mathbf{y} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{x}})'\mathbf{r} \right]. \quad (7)$$

From relation (10) of Sec. 9.9 and from (4) to (7), we can obtain the following relation for the posterior joint p.d.f.  $\xi(\mathbf{m}, \mathbf{r} | \mathbf{x}_1, \dots, \mathbf{x}_n)$  of  $\mathbf{M}$  and  $\mathbf{R}$ :

$$\begin{aligned} \xi(\mathbf{m}, \mathbf{r} | \mathbf{x}_1, \dots, \mathbf{x}_n) \propto \left\{ |\mathbf{r}|^{\frac{\alpha}{2}} \exp \left[ -\frac{\nu + n}{2}(\mathbf{m} - \mathbf{y}^*)'\mathbf{r}(\mathbf{m} - \mathbf{y}^*) \right] \right\} \\ \times \left\{ |\mathbf{r}|^{(\alpha+n-k-1)/2} \exp \left[ -\frac{1}{2} \text{tr}(\tau^*\mathbf{r}) \right] \right\}. \end{aligned} \quad (8)$$

The function inside the first set of braces in the relation (8), when regarded as a function of  $\mathbf{m}$ , must be proportional to the conditional p.d.f. of  $\mathbf{M}$  when  $\mathbf{R} = \mathbf{r}$ , since the variable  $\mathbf{m}$  does not appear inside the second set of braces. This function is proportional to the p.d.f. of a multivariate normal distribution for which the mean vector and the precision matrix are as specified in the statement of the theorem. It now follows that the function inside the second set of braces in the relation (8) is proportional to the marginal p.d.f. of  $\mathbf{R}$ , and it is proportional to the p.d.f. of the Wishart distribution specified in the statement of the theorem. ■

### 9.11 THE MARGINAL DISTRIBUTION OF THE MEAN VECTOR

We shall now find the marginal distribution of  $\mathbf{M}$  when the joint distribution of  $\mathbf{M}$  and  $\mathbf{R}$  is a multivariate normal-Wishart distribution of the form specified in Theorem 1 of Sec. 9.10. If the analogy with the univariate results given in Theorem 1 of Sec. 9.6 can be extended far enough, then the discussion following that theorem would lead us to conclude that the marginal distribution of  $\mathbf{M}$  here will be a multivariate  $t$  distribution. Such a conclusion is correct, as we shall now show.

Suppose, as in Theorem 1 of Sec. 9.10, that the conditional distribution of  $\mathbf{M}$  when  $\mathbf{R} = \mathbf{r}$  is a multivariate normal distribution with mean vector  $\mathbf{y}$  and precision matrix  $\nu\mathbf{r}$  and that the marginal distribution of  $\mathbf{R}$  is a Wishart distribution with  $\alpha$  degrees of freedom and precision matrix  $\tau$ . Then the joint p.d.f.  $\xi$  of  $\mathbf{M}$  and  $\mathbf{R}$ , as specified by the relation (5) of Sec. 9.10, can be rewritten as follows:

$$\xi(\mathbf{m}, \mathbf{r}) \propto |\mathbf{r}|^{(\alpha-k)/2} \exp \left[ -\frac{1}{2} \text{tr} \{[\tau + \nu(\mathbf{m} - \mathbf{y})(\mathbf{m} - \mathbf{y})']\mathbf{r}\} \right]. \quad (1)$$

The marginal p.d.f.  $\xi_M$  of  $\mathbf{M}$  is obtained by integrating the  $k(k+1)/2$  distinct variables in the symmetric matrix  $\mathbf{r}$  over the set of all values such that  $\mathbf{r}$  is positive definite. For any positive integer  $n$  ( $n \geq k$ ) and any positive definite matrix  $\mathbf{T}$ , the p.d.f. of the Wishart distribution specified by Eq. (11) of Sec. 5.5 must integrate to unity over this set. Hence, by

integrating the function on the right side of relation (1) over this set, we obtain the relation

$$\xi_{\mathbf{M}}(\mathbf{m}) \propto |\boldsymbol{\tau} + \nu(\mathbf{m} - \mathbf{y})|^{-(\alpha+\nu)/2}. \quad (2)$$

A standard result in the theory of determinants (see Exercise 40) may be stated as follows: If  $\mathbf{A}$  is any  $k \times k$  nonsingular matrix and  $\mathbf{v}$  is any  $k$ -dimensional column vector, then

$$|\mathbf{A} + \mathbf{v}\mathbf{v}'| = |\mathbf{A}|(1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{v}). \quad (3)$$

From the relations (2) and (3), we can obtain the following result:

$$\xi_{\mathbf{M}}(\mathbf{m}) \propto [1 + \nu(\mathbf{m} - \mathbf{y})'\boldsymbol{\tau}^{-1}(\mathbf{m} - \mathbf{y})]^{-(\alpha+\nu)/2}. \quad (4)$$

When the function on the right side of (4) is rewritten so as to bring it into the form of the p.d.f. of a multivariate  $t$  distribution specified by Eq. (9) of Sec. 5.6, it can be seen that  $\mathbf{M}$  has a multivariate  $t$  distribution with  $\alpha - k + 1$  degrees of freedom, location vector  $\mathbf{y}$ , and precision matrix  $\nu(\alpha - k + 1)\boldsymbol{\tau}^{-1}$ . This  $t$  distribution is the marginal distribution of  $\mathbf{M}$  under the prior joint distribution of  $\mathbf{M}$  and  $\mathbf{R}$  specified in Theorem 1 of Sec. 9.10. The posterior marginal distribution of  $\mathbf{M}$  is obtained by replacing  $\mathbf{y}$ ,  $\nu$ ,  $\alpha$ , and  $\boldsymbol{\tau}$  in the  $t$  distribution by their posterior values as given in that same theorem. Hence, the number of degrees of freedom  $\alpha + \nu - k + 1$  of the posterior distribution does not depend on the observed values  $\mathbf{x}_1, \dots, \mathbf{x}_n$  in the sample. However, the location vector of the posterior distribution depends on the value of the sample mean vector  $\bar{\mathbf{x}}$ , and the precision matrix of the posterior distribution depends on both the vector  $\bar{\mathbf{x}}$  and the matrix  $\mathbf{s}$  formed from the sample.

### 9.12 THE DISTRIBUTION OF A CORRELATION

Suppose that  $X_1$  and  $X_2$  are two random variables whose joint distribution is a bivariate normal distribution with precision matrix  $\mathbf{R}$ . Let the elements of the  $2 \times 2$  matrix  $\mathbf{R}$  be defined by the equation

$$\mathbf{R} = \begin{pmatrix} R_{11} & R_{12} \\ R_{12} & R_{22} \end{pmatrix}. \quad (1)$$

Since the precision matrix  $\mathbf{R}$  is the inverse of the covariance matrix of  $X_1$  and  $X_2$ , it follows that the correlation  $P$  of  $X_1$  and  $X_2$  is specified by the equation

$$P = \frac{-R_{12}}{(R_{11}R_{22})^{1/2}}. \quad (2)$$

Suppose now that the value of the precision matrix  $\mathbf{R}$  is unknown and that in accordance with the results developed in the preceding sections, it is assumed that the distribution of  $\mathbf{R}$  is a Wishart distribution with  $\alpha$  degrees of freedom and precision matrix  $\boldsymbol{\tau}$  such that  $\alpha > 1$  and  $\boldsymbol{\tau}$  is a symmetric, positive definite matrix whose elements are specified by the equation

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{12} & \tau_{22} \end{bmatrix}. \quad (3)$$

In this section, we shall derive the distribution of the correlation  $P$ . It will be shown that, in general, this distribution is quite complicated, that it is not one of the standard distributions which have been discussed in this book, and that the p.d.f. of  $P$  involves a certain integral which cannot be evaluated directly. However, if  $\tau_{12} = 0$  in the distribution of  $\mathbf{R}$ , then these complications are not present and the p.d.f. of  $P$  has a simple form.

Let  $c$  be the constant defined by Eq. (5) of Sec. 5.5 with  $n = \alpha$  and  $k = 2$ . The joint p.d.f.  $f(\tau_{11}, \tau_{22}, \tau_{12})$  of the three random variables  $R_{11}$ ,  $R_{22}$ , and  $R_{12}$  is positive for any values  $\tau_{11}, \tau_{22}$ , and  $\tau_{12}$  such that  $\tau_{11} > 0$ ,  $\tau_{22} > 0$ , and  $\tau_{11}\tau_{22} - \tau_{12}^2 > 0$ . For such values,

$$f(\tau_{11}, \tau_{22}, \tau_{12}) = c|\boldsymbol{\tau}|^{\alpha/2}(\tau_{11}\tau_{22} - \tau_{12}^2)^{(\alpha-3)/2} \times \exp \left[ -\frac{1}{2}(\tau_{11}\tau_{11} + \tau_{22}\tau_{22} + 2\tau_{12}\tau_{12}) \right]. \quad (4)$$

Let  $Y$  be a random variable defined as follows:

$$Y = \left( \frac{\tau_{22}R_{22}}{\tau_{11}R_{11}} \right)^{1/2}. \quad (5)$$

We shall now compute the joint p.d.f. of the three random variables  $R_{11}$ ,  $Y$ , and  $P$ . The joint p.d.f.  $g(\tau_{11}, y, \rho)$  of these variables will be positive for any values  $\tau_{11}, y$ , and  $\rho$  such that  $\tau_{11} > 0$ ,  $y > 0$ , and  $|\rho| < 1$ . The original variables  $\tau_{22}$  and  $\tau_{12}$  can be expressed in terms of the variables  $\tau_{11}, y$ , and  $\rho$  as follows:

$$\tau_{22} = \frac{\tau_{11}}{\tau_{22}} \tau_{11}y^2 \quad \text{and} \quad \tau_{12} = - \left( \frac{\tau_{11}}{\tau_{22}} \right)^{1/2} \tau_{11}y\rho. \quad (6)$$

Therefore, when the transformation from the three variables  $\tau_{11}, y$ , and  $\rho$  to the three variables  $\tau_{11}, \tau_{22}$ , and  $\tau_{12}$  is considered, the Jacobian  $J$  of this transformation will be the determinant of a  $3 \times 3$  matrix each of whose elements on one side of the main diagonal is 0. Therefore, the value of  $J$  is specified by the equation

$$J = \frac{\partial \tau_{11}}{\partial \tau_{11}} \cdot \frac{\partial \tau_{22}}{\partial y} \cdot \frac{\partial \tau_{12}}{\partial \rho} = -2 \left( \frac{\tau_{11}}{\tau_{22}} \right)^{3/2} \tau_{11}^2 y^2. \quad (7)$$

By changing the variables in Eq. (4) to  $r_{11}$ ,  $y$ , and  $\rho$  and multiplying the result by  $|J|$  as specified by Eq. (7), we obtain the following equation:

$$g(r_{11}, y, \rho) = 2c \left( \frac{\tau_{11}}{\tau_{22}} |\epsilon| \right)^{\alpha/2} (\tau_{11}y)^{\alpha-1} (1 - \rho^2)^{(\alpha-3)/2} \\ \times \exp \left\{ -\frac{\tau_{11}}{2} \left[ \tau_{11}(1 + y^2) - 2\tau_{12} \left( \frac{\tau_{11}}{\tau_{22}} \right) y\rho \right] \right\}. \quad (8)$$

The marginal joint p.d.f.  $h(y, \rho)$  of  $Y$  and  $P$  can be obtained by integrating the p.d.f. in Eq. (8) over all positive values of  $r_{11}$ . Let

$$c' = 2^{\alpha+1} c \Gamma(\alpha) \left( \frac{|\epsilon|}{\tau_{11}\tau_{22}} \right)^{\alpha/2}. \quad (9)$$

Then the result of this integration is

$$h(y, \rho) = \frac{c'(1 - \rho^2)^{(\alpha-3)/2}}{y|y + y^{-1} - 2\tau_{12}(\tau_{11}\tau_{22})^{-1/2}\rho|^\alpha}. \quad (10)$$

Finally, for any value  $\rho$  such that  $|\rho| < 1$ , the marginal p.d.f.  $h_P(\rho)$  of  $P$  is defined by the equation

$$h_P(\rho) = \int_0^\infty h(y, \rho) dy. \quad (11)$$

However, in general, this integration cannot be carried out in closed form.

Suppose, however, that  $\tau_{12} = 0$  in the precision matrix of the original Wishart distribution of  $\mathbf{R}$ . Then it can be seen from Eq. (10) that the joint p.d.f.  $h(y, \rho)$  can be factored into the product of a function of  $y$  and a function of  $\rho$ . Therefore, the random variables  $Y$  and  $P$  are independent, and the p.d.f.  $h_P(\rho)$  of  $P$  must have the following simple form:

$$h_P(\rho) = c''(1 - \rho^2)^{(\alpha-3)/2}. \quad (12)$$

Here,  $c''$  is a constant which makes the integral of the p.d.f. (12) equal to unity.

The distribution of the correlation has also been studied by Lindley (1965), sec. 8.2.

### 9.13 PRECISION MATRICES HAVING AN UNKNOWN FACTOR

In this section we shall consider the problem of sampling from a multivariate normal distribution for which the value of the mean vector  $\mathbf{M}$  is unknown and the precision matrix is the product of a specified matrix and an unknown positive number  $W$ . In other words, the precision matrix is of the form  $W\mathbf{r}$ , where  $\mathbf{r}$  is a specified symmetric  $k \times k$  positive definite matrix. A commonly occurring situation of this type involves sampling from a multivariate normal distribution for which it is known that the components are independent and have the same variance

but where the value of this variance is unknown. In such a case,  $W^{-1}$  is the unknown common variance and the precision matrix is of the form  $W\mathbf{I}$ , where  $\mathbf{I}$  is the identity matrix. The proof of the next theorem is to be developed as Exercise 44.

**Theorem 1** Suppose that  $X_1, \dots, X_n$  is a random sample from a multivariate normal distribution for which the mean vector  $\mathbf{M}$  has an unknown value and the precision matrix is of the form  $W\mathbf{r}$ , where  $\mathbf{r}$  is a specified positive definite matrix and the value of  $W$  is unknown. Suppose also that the prior joint distribution of  $\mathbf{M}$  and  $W$  is as follows: The conditional distribution of  $\mathbf{M}$  when  $W = w$  is a multivariate normal distribution with mean vector  $\mathbf{y}$  and precision matrix  $w\boldsymbol{\tau}$  such that  $\mathbf{y} \in R^k$  and  $\boldsymbol{\tau}$  is a symmetric  $k \times k$  positive definite matrix, and the marginal distribution of  $W$  is a gamma distribution with parameters  $\alpha$  and  $\beta$  such that  $\alpha > 0$  and  $\beta > 0$ . Then the posterior joint distribution of  $\mathbf{M}$  and  $W$  when  $\mathbf{X}_i = \mathbf{x}_i$  ( $i = 1, \dots, n$ ) is as follows: The conditional distribution of  $\mathbf{M}$  when  $W = w$  is a multivariate normal distribution with mean vector  $\mathbf{y}^*$  and precision matrix  $w(\boldsymbol{\tau} + n\boldsymbol{\tau})$ , where

$$\mathbf{y}^* = (\boldsymbol{\tau} + n\boldsymbol{\tau})^{-1}(n\mathbf{y} + n\bar{\mathbf{x}}), \quad (1)$$

and the marginal distribution of  $W$  is a gamma distribution with parameters  $\alpha + (nk/2)$  and  $\beta^*$ , where

$$\beta^* = \beta + \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})' \boldsymbol{\tau} (\mathbf{x}_i - \bar{\mathbf{x}}) + \frac{1}{2} (\mathbf{y}^* - \mathbf{y})' \boldsymbol{\tau} (\mathbf{y}^* - \mathbf{y}). \quad (2)$$

It can be shown (see Exercise 45) that when the joint distribution of  $\mathbf{M}$  and  $W$  is a multivariate normal-gamma distribution as specified in Theorem 1, the marginal distribution of the mean vector  $\mathbf{M}$  is a multivariate  $t$  distribution with  $2\alpha$  degrees of freedom, location vector  $\mathbf{y}$ , and precision matrix  $(\alpha/\beta)\boldsymbol{\tau}$ .

### Further Remarks and References

Posterior distributions for the parameters of multivariate normal distributions have been studied by Ando and Kaufman (1965), Geisser (1964; 1965a, b; 1966), and Geisser and Cornfield (1963). Probability models related to the one presented in Theorem 1 will be used in Chap. 11 for studying some problems of regression and analysis of variance.

### EXERCISES

1. Suppose that  $X_1, \dots, X_n$  is a random sample from a Poisson distribution with an unknown value of the mean. For any given value  $\lambda$  of the mean such that

$\lambda > 0$ , let  $f_n(\cdot | \lambda)$  denote the conditional joint p.f. of  $X_1, \dots, X_n$ . Let  $T$  be the statistic defined by the equation

$$T(X_1, \dots, X_n) = \sum_{i=1}^n X_i.$$

Show that  $T$  is a sufficient statistic for the family of p.f.'s  $f_n(\cdot | \lambda)$ .

2. Suppose that  $X_1, \dots, X_n$  is a random sample from a normal distribution with an unknown value of the mean and a specified value  $\sigma^2 > 0$  of the variance. For any given value  $\mu$  of the mean such that  $-\infty < \mu < \infty$ , let  $f_n(\cdot | \mu)$  denote the conditional joint p.d.f. of  $X_1, \dots, X_n$ . If the statistic  $T$  is defined as in Exercise 1, show that  $T$  is a sufficient statistic for the family of p.d.f.'s  $f_n(\cdot | \mu)$ .

3. As in Example 2 of Sec. 9.1, suppose that  $X_1, \dots, X_n$  is a random sample from a normal distribution with an unknown value of the mean and an unknown value of the variance. Let  $V$  be the two-dimensional vector defined by the equation

$$V(X_1, \dots, X_n) = \left\{ \sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right\}.$$

Show that  $V$  is a sufficient statistic for the family of p.d.f.'s  $f_n(\cdot | \mu, \sigma^2)$  specified by Eq. (8) of Sec. 9.1.

4. Suppose that  $X_1, \dots, X_n$  is a random sample from a gamma distribution with parameters  $\alpha$  and  $W$ , where the value of  $\alpha$  is specified ( $\alpha > 0$ ) and the value of the parameter  $W$  is unknown. For any given value  $\beta$  of the parameter  $W$  such that  $\beta > 0$ , let  $f_n(\cdot | \beta)$  denote the conditional joint p.d.f. of  $X_1, \dots, X_n$ . If the statistic  $T$  is as defined in Exercise 1, show that  $T$  is a sufficient statistic for the family of p.d.f.'s  $f_n(\cdot | \beta)$ .

5. Suppose that  $X_1, \dots, X_n$  is a random sample from a gamma distribution with parameters  $W_1$  and  $W_2$ , both of whose values are unknown. For any given values  $\alpha$  and  $\beta$  of the parameters  $W_1$  and  $W_2$  such that  $\alpha > 0$  and  $\beta > 0$ , let  $f_n(\cdot | \alpha, \beta)$  denote the conditional joint p.d.f. of  $X_1, \dots, X_n$ . Let  $T$  be the two-dimensional vector defined by the equation

$$T(X_1, \dots, X_n) = \left\{ \sum_{i=1}^n X_i, \sum_{i=1}^n \log X_i \right\}.$$

Show that  $T$  is a sufficient statistic for the family of p.d.f.'s  $f_n(\cdot | \alpha, \beta)$ .

6. Suppose that  $X_1, \dots, X_n$  is a random sample from a uniform distribution on the interval  $(W_1, W_2)$ , where the values of the parameters  $W_1$  and  $W_2$  are unknown. For any given values  $w_1$  and  $w_2$  of  $W_1$  and  $W_2$  such that  $-\infty < w_1 < w_2 < \infty$ , let  $f_n(\cdot | w_1, w_2)$  denote the conditional joint p.d.f. of  $X_1, \dots, X_n$ . Let  $T$  be the two-dimensional vector defined by the equation

$$T(X_1, \dots, X_n) = \{\min(X_1, \dots, X_n), \max(X_1, \dots, X_n)\}.$$

Show that  $T$  is a sufficient statistic for the family of p.d.f.'s  $f_n(\cdot | w_1, w_2)$ .

7. Suppose that  $X_1, \dots, X_n$  is a random sample from a uniform distribution on the interval  $(W - 1, W + 1)$ , where the value of the parameter  $W$  is unknown. For any given value  $w$  of  $W$  such that  $-\infty < w < \infty$ , let  $f_n(\cdot | w)$  denote the conditional joint p.d.f. of  $X_1, \dots, X_n$ . If the two-dimensional statistic  $T$  is as defined in Exercise 6, show that  $T$  is a sufficient statistic for the family of p.d.f.'s  $f_n(\cdot | w)$ .

8. Suppose that  $X_1, \dots, X_n$  is a random sample of vectors from a  $k$ -dimensional multivariate normal distribution with an unknown value of the mean vector

## EXERCISES

$W$  and a specified nonsingular covariance matrix  $\Sigma$ . For any given value  $w$  of  $W$  such that  $w \in P^k$ , let  $f_n(\cdot | w)$  denote the conditional joint p.d.f. of the random vectors  $X_1, \dots, X_n$ . Let

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Show that  $\bar{X}$  is a sufficient statistic for the family of p.d.f.'s  $f_n(\cdot | w)$ .

9. Suppose that  $X_1, \dots, X_n$  is a random sample from a  $k$ -dimensional multivariate normal distribution with an unknown value of the mean vector and an unknown value of the covariance matrix. For any given value  $\mu$  of the mean vector and any given value  $\Sigma$  of the covariance matrix such that  $\mu \in P^k$  and  $\Sigma$  is a symmetric  $k \times k$  positive definite matrix, let  $f_n(\cdot | \mu, \Sigma)$  denote the conditional joint p.d.f. of the random vectors  $X_1, \dots, X_n$ . Let the statistic  $V$  be specified by the equation

$$V(X_1, \dots, X_n) = \left\{ \bar{X}, \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' \right\}.$$

Show that  $V$  is a sufficient statistic for the family of p.d.f.'s  $f_n(\cdot | \mu, \Sigma)$ .

10. Let  $\{f(\cdot | w), w \in \Omega\}$  be a family of g.p.d.f.'s each of which is defined on the sample space  $S$ . Let  $T_1$  and  $T_2$  be two statistics which have the following property: For any two points  $x_1 \in S$  and  $x_2 \in S$ ,  $T_1(x_1) = T_1(x_2)$  if, and only if,  $T_2(x_1) = T_2(x_2)$ . Show that  $T_1$  is a sufficient statistic for the specified family of g.p.d.f.'s if, and only if,  $T_2$  is a sufficient statistic for that family.

11. Show that each of the following families of distributions is an exponential family:

- The family of Bernoulli distributions with an unknown value of the parameter
- The family of Poisson distributions with an unknown value of the mean
- The family of normal distributions with an unknown value of the mean and a specified value of the variance
- The family of normal distributions with an unknown value of the mean and an unknown value of the variance
- The family of gamma distributions with unknown values of the parameters  $\alpha$  and  $\beta$
- The family of  $k$ -dimensional multivariate normal distributions with an unknown value of the mean vector and an unknown value of the covariance matrix

12. Show that if the mean of a beta distribution is  $\mu$  and the variance is  $\sigma^2$ , then  $\sigma^2 < \mu(1 - \mu)$ .

13. Let  $\mu$  and  $\sigma^2$  be numbers such that  $0 < \mu < 1$ ,  $\sigma^2 > 0$ , and  $\sigma^2 < \mu(1 - \mu)$ . Also, let  $\alpha$  and  $\beta$  be the parameters of the unique beta distribution which has mean  $\mu$  and variance  $\sigma^2$ . Show that

$$\alpha = \mu \left[ \frac{\mu(1 - \mu)}{\sigma^2} - 1 \right] \quad \text{and} \quad \beta = (1 - \mu) \left[ \frac{\mu(1 - \mu)}{\sigma^2} - 1 \right].$$

14. Let  $W$  denote the unknown probability that a certain machine will produce a defective item, and suppose that the prior distribution of  $W$  is a uniform distribution on the interval  $(0, 1)$ . Suppose that after the items in a random sample produced by the machine have been inspected, the posterior distribution of  $W$  is a beta distribution with parameters  $\alpha = 7$  and  $\beta = 95$ . Show that 100 items were inspected and that six of them were defective.

15. Show that for any given positive numbers  $\mu$  and  $\sigma^2$ , there is a unique gamma distribution which has mean  $\mu$  and variance  $\sigma^2$ .
16. Suppose that when magnetic recording tape is manufactured by a certain process, the mean number  $W$  of defects on a 1,200-ft roll of tape is unknown, and suppose that the prior distribution of  $W$  is a gamma distribution whose mean is 2 and whose variance is 1. Suppose also that the number of defects on any roll of tape when  $W = w$  has a Poisson distribution with mean  $w$ . Suppose further that after a random sample of rolls of tape has been taken and the number of defects on each roll has been counted, the mean of the posterior distribution of  $W$  is 1.6 and the variance is 0.16. Show that eight rolls of tape were included in the random sample and that the average number of defects per roll in the sample was 1.5. Prove Theorem 2 of Sec. 9.4.
17. An unknown proportion  $W$  of the items produced by a certain machine are defective. Suppose that the prior distribution of  $W$  is a beta distribution with parameters  $\alpha = 1$  and  $\beta = 99$ . Suppose also that items produced by the machine are selected at random and observed one at a time until exactly five defective items have been found. If, when sampling terminates, the mean of the posterior distribution of  $W$  is 0.02, show that 195 nondefective items were observed during the sampling process.
19. Suppose that in a large population of voters, the proportion  $W$  who belong to the Liberal Party is unknown, and suppose that the prior distribution of  $W$  is a beta distribution with parameters  $\alpha = 1$  and  $\beta = 10$ .
- (a) If, in a random sample of 1,000 voters, it is found that 123 belong to the Liberal Party, what is the posterior distribution of  $W$ ?
- (b) Suppose that instead of taking a random sample as in part a, voters are selected one at a time until exactly 123 have been found who belong to the Liberal Party. Suppose that a total of 1,000 voters had to be selected in order to accomplish this. What is the posterior distribution of  $W$ ?
20. Prove Theorem 3 of Sec. 9.4.
21. The length of life of a lamp manufactured by a certain process has an exponential distribution with an unknown value of the parameter  $W$ . Suppose that the prior distribution of  $W$  is a gamma distribution for which the coefficient of variation is 0.5. A random sample of lamps is to be tested, and the length of life of each of the lamps is to be noted. If the coefficient of variation of the posterior distribution of  $W$  must be reduced to the value 0.1, show that 96 lamps should be tested.
22. Extend Theorem 3 of Sec. 9.4 so that it covers the case of a random sample from a gamma distribution with parameters  $\alpha$  and  $W$ , where the value of  $\alpha$  is specified ( $\alpha > 0$ ) and the value of  $W$  is unknown.
23. Consider a normal distribution for which the value of the mean  $W$  is unknown and the variance is 4, and suppose that the prior distribution of  $W$  is a normal distribution whose variance is 9. How large a random sample must be taken from the given normal distribution in order to be able to specify an interval having a length of 1 unit such that the probability that  $W$  lies in this interval is at least 0.95? (Answer:  $n = 62$ .)
24. Prove Theorem 2 of Sec. 9.5.
25. Suppose that the value of the precision  $W$  of a normal distribution is unknown, and suppose that the distribution of  $W$  is a gamma distribution with parameters  $\alpha$  and  $\beta$ . Let  $V$  denote the variance of the given normal distribution.
- (a) Find the p.d.f. of  $V$ .
- (b) Show that if  $\alpha > 1$ ,  $E(V) = \beta/(\alpha - 1)$ .
- (c) Show that if  $\alpha > 2$ ,  $\text{Var}(V) = \beta^2/[(\alpha - 1)^2(\alpha - 2)]$ .

## EXERCISES

26. Suppose that a random sample is to be taken from a normal distribution with a specified value of the mean and an unknown value of the precision  $W$ . Suppose also that the prior distribution of  $W$  is a gamma distribution and that the coefficient of variation of the posterior distribution of  $W$  must be reduced to the value 0.1. Show that this requirement will be satisfied, regardless of the value of the coefficient of variation of the prior distribution, if a sample of size  $n = 200$  is taken.
27. Consider a normal distribution with an unknown value of the mean  $M$  and an unknown value of the precision  $R$ , and suppose that the prior joint distribution of  $M$  and  $R$  is as specified in Theorem 1 of Sec. 9.6. Find the conditional distribution of  $R$  when  $M = m$ .
28. Consider the conditions specified in Exercise 27. Suppose that the coefficient of variation of the prior distribution of  $R$  has the value 0.5. How large a random sample must be taken from the given normal distribution in order that the coefficient of variation of the posterior distribution of  $R$  will be reduced to the value 0.1? (Answer:  $n = 192$ .)
29. Consider the conditions specified in Exercise 27. Suppose that under the conditional posterior distribution of  $M$  when  $R = 3$ , the variance of  $M$  must be reduced to the value 0.01. Show that this requirement will be satisfied, regardless of the values of the parameters of the prior distribution, if a random sample of size  $n = 34$  is taken from the given normal distribution.
30. The length of time for which a certain man must wait each morning for a bus taking him to work is uniformly distributed on the interval  $(0, W)$ , where the value of  $W$  is unknown and the prior distribution of  $W$  is a Pareto distribution with parameters  $w_0 > 0$  and  $\alpha = 1$ . On how many mornings must the man observe his waiting time before he will be able to specify an interval having a length of 0.01 unit such that the probability that the unknown value of  $\log W$  lies in this interval is at least 0.95? (Answer:  $n = 299$ .)
31. Consider the prior distribution of  $W$  specified in Exercise 30. On how many mornings must the man observe his waiting time in order that the coefficient of variation of the posterior distribution of  $W$  will be reduced to the value 0.01? (Answer:  $n = 101$ .)
32. Prove Theorem 2 of Sec. 9.7.
33. Consider a uniform distribution on the interval  $(W_1, W_2)$ , where the values of  $W_1$  and  $W_2$  are unknown, and suppose that the prior joint distribution of  $W_1$  and  $W_2$  is a bilateral bivariate Pareto distribution with parameters  $r_1 > 0$ ,  $r_2 > 0$ , and  $\alpha = 2$ . How large a random sample must be taken from the uniform distribution in order that the coefficient of variation of the posterior distribution of the random variable  $W_2 - W_1$  will be reduced to the value 0.01? (Answer:  $n = 140$ .)
34. Suppose that a box contains  $N$  balls, of which an unknown number  $W$  are red and the rest are blue. Suppose also that the prior distribution of  $W$  is a hypergeometric distribution with parameters  $A$ ,  $B$ , and  $N$ , where  $A$  and  $B$  are positive integers such that  $A + B \geq N$ .
- (a) Now suppose that although the exact value of  $W$  is unknown, the statistician knows that  $r \leq W \leq s$ , where  $r$  and  $s$  are integers such that  $0 < r \leq s < N$ . Show that there are unique values of  $A$  and  $B$  such that, under the prior hypergeometric distribution,
- $$\Pr\{r \leq W \leq s\} = 1, \quad \Pr\{W = r\} > 0, \quad \text{and} \quad \Pr\{W = s\} > 0.$$
- (b) Next, suppose that the statistician knows only that  $0 \leq W \leq N$ . Show that any prior hypergeometric distribution such that  $A \geq N$  and  $B \geq N$  will assign positive probability to each integer  $0, 1, 2, \dots, N$ .

- (c) Suppose that one ball is selected from the box at random. What is the probability, under the prior distribution, that it will be red?
- (d) Suppose that  $n$  balls ( $1 \leq n < N$ ) are selected at random from the box without replacement and that  $x$  of these balls are red. Show that the posterior distribution of the number of red balls among the  $N - n$  balls remaining in the box is a hypergeometric distribution with parameters  $A - x, B - (n - x)$ , and  $N - n$ .
35. Suppose that there are  $k$  different types of items in a very large population and let  $W_i$  be the unknown proportion of the population that includes items of type  $i$  ( $i = 1, 2, \dots, k$ ). Suppose also that the prior distribution of  $\mathbf{W} = (W_1, \dots, W_k)'$  is a Dirichlet distribution with parametric vector  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)'$  such that  $\alpha_1 + \dots + \alpha_k = 6$ . How large a random sample of items must be taken in order to be sure that no matter what the values of the individual components of the vector  $\boldsymbol{\alpha}$  are and no matter what the observed outcomes are, the posterior variance of each proportion  $W_i$  ( $i = 1, \dots, k$ ) will be at most 0.005? (Answer:  $n = 43$ .)

36. Consider a bivariate normal distribution with an unknown mean vector  $\mathbf{M} = (M_1, M_2)'$  and a precision matrix  $\mathbf{r}$  which is known to be

$$\mathbf{r} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{4}{3} \end{bmatrix}.$$

Suppose that the prior distribution of  $\mathbf{M}$  is a bivariate normal distribution for which the precision matrix is

$$\boldsymbol{\tau} = \begin{bmatrix} 1 & -1 \\ -1 & 6 \end{bmatrix}.$$

How large a random sample must be taken in order that the variance of the posterior distribution of the random variable  $M_1 - M_2$  will be reduced to the value 0.01? (Answer:  $n = 297$ .)

37. Suppose that a random  $k \times k$  symmetric, positive definite matrix  $\mathbf{R}$  has a Wishart distribution with  $\alpha$  degrees of freedom ( $\alpha > k - 1$ ) and precision matrix  $\boldsymbol{\tau}$ . Show that the coefficient of variation of the determinant  $|\mathbf{R}|$  is

$$\left[ \frac{k(2\alpha - k + 3)}{(\alpha - k + 2)(\alpha - k + 1)} \right]^{\frac{1}{2}}.$$

Hint: See Exercise 14 of Chap. 5.

38. Consider a bivariate normal distribution with a specified mean vector and an unknown precision matrix  $\mathbf{R}$ . Suppose that the prior distribution of  $\mathbf{R}$  is a Wishart distribution with 3 degrees of freedom and precision matrix  $\boldsymbol{\tau}$ . How large a random sample must be taken in order that the coefficient of variation of the posterior distribution of the determinant  $|\mathbf{R}|$  will be reduced to the value 0.1? (Answer:  $n = 399$ .)

39. Consider again the conditions described in Exercise 38, and suppose that the elements of the  $2 \times 2$  matrix  $\mathbf{R}$  are

$$\mathbf{R} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}.$$

How large a random sample must be taken in order that the coefficient of variation of the posterior distribution of the random variable  $R_{11}$  will be reduced to the value 0.1? (Answer:  $n = 197$ .)

40. Prove that Eq. (3) of Sec. 9.11 is correct.

41. Consider a multivariate normal distribution with an unknown value of the mean vector  $\mathbf{M}$  and an unknown value of the precision matrix  $\mathbf{R}$ , and suppose that the prior joint distribution of  $\mathbf{M}$  and  $\mathbf{R}$  is a multivariate normal-Wishart distribution as specified in Theorem 1 of Sec. 9.10. What is the conditional distribution of  $\mathbf{R}$  when  $\mathbf{M} = \mathbf{m}$ ?

42. Consider a bivariate normal distribution with an unknown value of the mean vector  $\mathbf{M}$  and an unknown value of the precision matrix  $\mathbf{R}$ . Suppose that the prior joint distribution of  $\mathbf{M}$  and  $\mathbf{R}$  is a bivariate normal-Wishart distribution as specified in Theorem 1 of Sec. 9.10, and suppose that  $\alpha = 3$  in this prior distribution. How large a random sample must be taken in order that the coefficient of variation of the posterior distribution of the determinant  $|\mathbf{R}|$  will be reduced to the value 0.1? Hint: Note that the answer in this exercise must be the same as the answer in Exercise 38.

43. Consider the conditions specified in Exercise 42, and suppose again that  $\alpha = 3$  in the prior joint distribution of  $\mathbf{M}$  and  $\mathbf{R}$ . Suppose that a random sample is taken, and let  $\mathbf{y}^*$  be the location vector and  $\mathbf{T}^*$  the precision matrix of the posterior bivariate  $t$  distribution of the two-dimensional vector  $\mathbf{M}$ . Determine the size of a random sample that must be taken in order for the posterior distribution of  $\mathbf{M}$  to satisfy the following equation:

$$\Pr\{(\mathbf{M} - \mathbf{y}^*)' \mathbf{T}^* (\mathbf{M} - \mathbf{y}^*) \leq 10\} \geq 0.95.$$

Hint: See expression (16) of Sec. 5.6. (Answer:  $n = 5$ .)

44. Prove Theorem 1 of Sec. 9.13.

45. Prove that if the joint distribution of the random vector  $\mathbf{M}$  and the random variable  $W$  is a multivariate normal-gamma distribution as specified in Theorem 1 of Sec. 9.13, then the marginal distribution of  $\mathbf{M}$  is a multivariate  $t$  distribution with  $2\alpha$  degrees of freedom, location vector  $\mathbf{y}$ , and precision matrix  $(\alpha/\beta)\boldsymbol{\tau}$ .

46. Consider five different normal distributions with unknown means  $M_1, \dots, M_5$  and with a common unknown precision  $W$ . Suppose that the prior joint distribution of  $M_1, \dots, M_5$  and  $W$  is as follows: For any given value  $W = w$ , the random variables  $M_1, \dots, M_5$  are independent and have the same normal distribution with mean 3 and precision  $2w$ . Furthermore, the marginal distribution of  $W$  is a gamma distribution with parameters  $\alpha = 10$  and  $\beta = 5$ . Suppose also that a random sample of eight observations is taken from each of the five normal distributions. Let  $x_i$  denote the value of the  $j$ th observation from the  $i$ th distribution, and for  $i = 1, \dots, 5$ , let  $\bar{x}_i$  denote the average of the eight observations from the  $i$ th distribution. Finally, let the values  $\mu_1, \dots, \mu_5$  and  $c$  be defined as follows:

$$\mu_i = \frac{3 + 4\bar{x}_i}{5} \quad i = 1, \dots, 5,$$

and

$$c = \frac{2.37}{600} \left[ 50 + 5 \sum_{i=1}^5 \sum_{j=1}^8 (x_{ij} - \bar{x}_i)^2 + 8 \sum_{i=1}^5 (\bar{x}_i - 3)^2 \right].$$

Show that under the posterior joint distribution of  $M_1, \dots, M_5$ ,

$$\Pr \left[ \sum_{i=1}^5 (M_i - \mu_i)^2 \leq c \right] = 0.95.$$

Hint: See expression (16) of Sec. 5.6.

47. Suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  is a random sample of symmetric  $k \times k$  matrices from a Wishart distribution with  $m$  degrees of freedom ( $m > k - 1$ ) and an unknown value of the precision matrix  $\mathbf{R}$ . Suppose also that the prior distribution of  $\mathbf{R}$  is a Wishart distribution with  $\alpha$  degrees of freedom and precision matrix  $\boldsymbol{\tau}$  such that  $\alpha > k - 1$  and  $\boldsymbol{\tau}$  is a  $k \times k$  positive definite matrix. Show that the posterior distribution of  $\mathbf{R}$  when  $\mathbf{X}_i = \mathbf{x}_i$  ( $i = 1, \dots, n$ ) is a Wishart distribution with  $\alpha + mn$  degrees of freedom and precision matrix  $\boldsymbol{\tau} + \sum_{i=1}^n \mathbf{x}_i$ .