CHAPTER 8

More About Estimation

8.1 Bayesian Estimation

In Chapter 6 we introduced point and interval estimation for various parameters. In Chapter 7 we observed how such inferences should be based upon sufficient statistics for the parameters if they exist. In this chapter we introduce other concepts related to estimation and begin this by considering *Bayesian estimates*, which are also based upon sufficient statistics if the latter exist.

In introducing the interesting and sometimes controversial Bayesian method of estimation, the student should constantly keep in mind that making statistical inferences from the data does not strictly follow a mathematical approach. Clearly, up to now, we have had to construct models before we have been able to make such inferences. These models are subjective, and the resulting inference depends greatly on the model selected. For illustration, two statisticians could very well select different models for exactly the same situation and make different inferences with exactly the same data. Most statisticians would use some type of model diagnostics to see if

the models seem to be reasonable ones, but we must still recognize that there can be differences among statisticians' inferences.

We shall now describe the Bayesian approach to the problem of estimation. This approach takes into account any prior knowledge of the experiment that the statistician has and it is one application of a principle of statistical inference that may be called *Bayesian statistics*. Consider a random variable X that has a distribution of probability that depends upon the symbol θ, where θ is an element of a well-defined set Ω. For example, if the symbol θ is the mean of a normal distribution, Ω may be the real line. We have previously looked upon θ as being some constant, although an unknown constant. Let us now introduce a random variable Θ that has a distribution of probability over the set Ω; and, just as we look upon x as a possible value of the random variable X, we now look upon θ as a possible value of the random variable Θ. Thus the distribution of X depends upon θ, an experimental value of the random variable Θ. We shall denote the p.d.f. of Θ by $h(\theta)$ and we take $h(\theta) = 0$ when θ is not an element of Ω. Moreover, we now denote the p.d.f. of X by $f(x|\theta)$ since we think of it as a conditional p.d.f. of X, given $\Theta = \theta$.

Say $X_1, X_2, \ldots, X_n$ is a random sample from this conditional distribution of X. Thus we can write the joint conditional p.d.f. of $X_1, X_2, \ldots, X_n$, given $\Theta = \theta$, as

$$f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta).$$

Thus the joint p.d.f. of $X_1, X_2, \ldots, X_n$ and Θ is

$$g(x_1, x_2, \ldots, x_n, \theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)h(\theta).$$

If Θ is a random variable of the continuous type, the joint marginal p.d.f. of $X_1, X_2, \ldots, X_n$ is given by

$$g_1(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n, \theta)\, d\theta.$$

If Θ is a random variable of the discrete type, integration would be replaced by summation. In either case the conditional p.d.f. of Θ, given $X_1 = x_1, \ldots, X_n = x_n$, is

$$k(\theta|x_1, x_2, \ldots, x_n) = \frac{g(x_1, x_2, \ldots, x_n, \theta)}{g_1(x_1, x_2, \ldots, x_n)} = \frac{f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)h(\theta)}{g_1(x_1, x_2, \ldots, x_n)}.$$

This relationship is another form of Bayes' formula.

Example 1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with mean θ, where θ is the observed value of a random variable Θ having a gamma distribution with known parameters α and β. Thus

$$g(x_1, \ldots, x_n, \theta) = \left[\frac{\theta^{x_1}e^{-\theta}}{x_1!} \cdots \frac{\theta^{x_n}e^{-\theta}}{x_n!}\right]\left[\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}\right],$$

provided that $x_i = 0, 1, 2, 3, \ldots$, $i = 1, 2, \ldots, n$ and $0 < \theta < \infty$, and is equal to zero elsewhere. Then

$$g_1(x_1, \ldots, x_n) = \int_0^{\infty} \frac{\theta^{\Sigma x_i + \alpha - 1}e^{-(n+1/\beta)\theta}}{x_1! \cdots x_n!\,\Gamma(\alpha)\beta^{\alpha}}\,d\theta$$

$$= \frac{\Gamma\left(\sum_1^n x_i + \alpha\right)}{x_1! \cdots x_n!\,\Gamma(\alpha)\beta^{\alpha}(n + 1/\beta)^{\Sigma x_i + \alpha}}.$$

Finally, the conditional p.d.f. of Θ, given $X_1 = x_1, \ldots, X_n = x_n$, is

$$k(\theta|x_1, \ldots, x_n) = \frac{g(x_1, \ldots, x_n, \theta)}{g_1(x_1, \ldots, x_n)}$$

$$= \frac{\theta^{\Sigma x_i + \alpha - 1}e^{-\theta/[\beta/(n\beta+1)]}}{\Gamma\left(\sum x_i + \alpha\right)[\beta/(n\beta + 1)]^{\Sigma x_i + \alpha}},$$

provided that $0 < \theta < \infty$, and is equal to zero elsewhere. This conditional p.d.f. is one of the gamma type with parameters $\alpha^* = \Sigma x_i + \alpha$ and $\beta^* = \beta/(n\beta + 1)$.

In Example 1 it is extremely convenient to notice that it is not really necessary to determine $g_1(x_1, \ldots, x_n)$ to find $k(\theta|x_1, \ldots, x_n)$. If we divide

$$f(x_1|\theta)f(x_2|\theta) \cdots f(x_n|\theta)h(\theta)$$

by $g_1(x_1, \ldots, x_n)$, we must get the product of a factor, which depends upon $x_1, \ldots, x_n$ but does *not* depend upon θ, say $c(x_1, \ldots, x_n)$, and

$$\theta^{\Sigma x_i + \alpha - 1}e^{-\theta/[\beta/(n\beta+1)]}.$$

That is,

$$k(\theta|x_1, \ldots, x_n) = c(x_1, \ldots, x_n)\theta^{\Sigma x_i + \alpha - 1}e^{-\theta/[\beta/(n\beta+1)]},$$

provided that $0 < \theta < \infty$ and $x_i = 0, 1, 2, \ldots$, $i = 1, 2, \ldots, n$. However, $c(x_1, \ldots, x_n)$ must be that "constant" needed to make $k(\theta|x_1, \ldots, x_n)$ a p.d.f., namely

$$c(x_1, \ldots, x_n) = \frac{1}{\Gamma\left(\sum x_i + \alpha\right)[\beta/(n\beta + 1)]^{\Sigma x_i + \alpha}}.$$

Accordingly, Bayesian statisticians frequently write that $k(\theta|x_1, \ldots, x_n)$ is proportional to

$$g(x_1, x_2, \ldots, x_n, \theta);$$

that is,

$$k(\theta|x_1, \ldots, x_n) \propto f(x_1|\theta) \cdots f(x_n|\theta)h(\theta).$$

Note that in the right-hand member of this expression all factors involving constants and $x_1, \ldots, x_n$ alone (not θ) can be dropped. For illustration, in solving the problem presented in Example 1, the Bayesian statistician would simply write

$$k(\theta|x_1, \ldots, x_n) \propto \theta^{\Sigma x_i} e^{-n\theta} \theta^{\alpha - 1} e^{-\theta/\beta}$$

or, equivalently,

$$k(\theta|x_1, \ldots, x_n) \propto \theta^{\Sigma x_i + \alpha - 1} e^{-\theta/[\beta/(n\beta + 1)]},$$

$0 < \theta < \infty$ and is equal to zero elsewhere. Clearly, $k(\theta|x_1, \ldots, x_n)$ must be a gamma p.d.f. with parameters $\alpha^* = \Sigma x_i + \alpha$ and $\beta^* = \beta/(n\beta + 1)$.

There is another observation that can be made at this point. Suppose that there exists a sufficient statistic $Y = u(X_1, \ldots, X_n)$ for the parameter so that

$$f(x_1|\theta) \cdots f(x_n|\theta) = g[u(x_1, \ldots, x_n)|\theta]H(x_1, \ldots, x_n),$$

where now $g(y|\theta)$ is the p.d.f. of Y, given $\Theta = \theta$. Then we note that

$$k(\theta|x_1, \ldots, x_n) \propto g[u(x_1, \ldots, x_n)|\theta]h(\theta)$$

because the factor $H(x_1, \ldots, x_n)$ that does not depend upon θ can be dropped. Thus, if a sufficient statistic Y for the parameter exists, we can begin with the p.d.f. of Y if we wish and write

$$k(\theta|y) \propto g(y|\theta)h(\theta),$$

where now $k(\theta|y)$ is the conditional p.d.f. of Θ, given the sufficient statistic $Y = y$. The following discussion assumes that a sufficient statistic Y does exist; but more generally, we could replace Y by $X_1, X_2, \ldots, X_n$ in what follows. Also, we now use $g_1(y)$ to be the marginal p.d.f. of Y; that is, in the continuous case,

$$g_1(y) = \int_{-\infty}^{\infty} g(y|\theta)h(\theta)\, d\theta.$$

In Bayesian statistics, the p.d.f. $h(\theta)$ is called the *prior p.d.f.* of Θ, and the conditional p.d.f. $k(\theta|y)$ is called the *posterior p.d.f.* of Θ. This is because $h(\theta)$ is the p.d.f. of Θ prior to the observation of Y, whereas $k(\theta|y)$ is the p.d.f. of Θ after the observation of Y has been made. In many instances, $h(\theta)$ is not known; yet the choice of $h(\theta)$ affects the p.d.f. $k(\theta|y)$. In these instances the statistician takes into account all prior knowledge of the experiment and *assigns* the prior p.d.f. $h(\theta)$. This, of course, injects the problem of *personal* or *subjective probability* (see the Remark, Section 1.1).

Suppose that we want a point estimate of θ. From the Bayesian viewpoint, this really amounts to selecting a decision function δ, so that $\delta(y)$ is a predicted value of θ (an experimental value of the random variable Θ) when both the computed value y and the conditional p.d.f. $k(\theta|y)$ are known. Now, in general, how would we predict an experimental value of any random variable, say W, if we want our prediction to be "reasonably close" to the value to be observed? Many statisticians would predict the mean, $E(W)$, of the distribution of W; others would predict a median (perhaps unique) of the distribution of W; some would predict a mode (perhaps unique) of the distribution of W; and some would have other predictions. However, it seems desirable that the choice of the decision function should depend upon the loss function $\mathscr{L}[\theta, \delta(y)]$. One way in which this dependence upon the loss function can be reflected is to select the decision function δ in such a way that the conditional expectation of the loss is a minimum. A *Bayes' solution* is a decision function δ that minimizes

$$E\{\mathscr{L}[\Theta, \delta(y)] | Y = y\} = \int_{-\infty}^{\infty} \mathscr{L}[\theta, \delta(y)]k(\theta|y)\, d\theta,$$

if Θ is a random variable of the continuous type. The usual modification of the right-hand member of this equation is made for random variables of the discrete type. If, for example, the loss function is given by $\mathscr{L}[\theta, \delta(y)] = [\theta - \delta(y)]^2$, the Bayes' solution is given by $\delta(y) = E(\Theta|y)$, the mean of the conditional distribution of Θ, given $Y = y$. This follows from the fact that $E[(W - b)^2]$, if it exists, is a minimum when $b = E(W)$. If the loss function is given by $\mathscr{L}[\theta, \delta(y)] = |\theta - \delta(y)|$, then a median of the conditional distribution of Θ, given $Y = y$, is the Bayes' solution. This follows from the fact

that $E(|W - b|)$, if it exists, is a minimum when b is equal to any median of the distribution of W.

The conditional expectation of the loss, given $Y = y$, defines a random variable that is a function of the statistic Y. The expected value of that function of Y, in the notation of this section, is given by

$$\int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \mathscr{L}[\theta, \delta(y)] k(\theta|y)\, d\theta \right\} g_1(y)\, dy$$

$$= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \mathscr{L}[\theta, \delta(y)] g(y|\theta)\, dy \right\} h(\theta)\, d\theta,$$

in the continuous case. The integral within the braces in the latter expression is, for every given $\theta \in \Omega$, the risk function $R(\theta, \delta)$; accordingly, the latter expression is the mean value of the risk, or the expected risk. Because a Bayes' solution minimizes

$$\int_{-\infty}^{\infty} \mathscr{L}[\theta, \delta(y)] k(\theta|y)\, d\theta$$

for every y for which $g_1(y) > 0$, it is evident that a Bayes' solution $\delta(y)$ minimizes this mean value of the risk. We now give an illustrative example.

Example 2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $b(1, \theta)$, $0 < \theta < 1$. We seek a decision function δ that is a Bayes' solution. The sufficient statistic $Y = \sum_1^n X_i$, and Y is $b(n, \theta)$. That is, the conditional p.d.f. of Y, given $\Theta = \theta$, is

$$g(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}, \qquad y = 0, 1, \ldots, n,$$

$$= 0 \qquad \text{elsewhere.}$$

We take the prior p.d.f. of the random variable Θ to be

$$h(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1}, \qquad 0 < \theta < 1,$$

$$= 0 \qquad \text{elsewhere.}$$

where α and β are assigned positive constants. Thus the conditional p.d.f. of Θ, given $Y = y$, is, at points of positive probability density,

$$k(\theta|y) \propto \theta^y (1 - \theta)^{n-y} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1}, \qquad 0 < \theta < 1.$$

That is,

$$k(\theta|y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+y)\Gamma(n+\beta-y)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}, \qquad 0<\theta<1,$$

and $y = 0, 1, \ldots, n$. We take the loss function to be $\mathscr{L}[\theta, \delta(y)] = [\theta - \delta(y)]^2$. Because Y is a random variable of the discrete type, whereas Θ is of the continuous type, we have for the expected risk,

$$\int_0^1 \left\{\sum_{y=0}^{n}[\theta-\delta(y)]^2\binom{n}{y}\theta^y(1-\theta)^{n-y}\right\}h(\theta)\,d\theta$$

$$= \sum_{y=0}^{n}\left\{\int_0^1 [\theta-\delta(y)]^2 k(\theta|y)\,d\theta\right\}g_1(y).$$

The Bayes' solution $\delta(y)$ is the mean of the conditional distribution of Θ, given $Y = y$. Thus

$$\begin{aligned}\delta(y) &= \int_0^1 \theta k(\theta|y)\,d\theta \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+y)\Gamma(n+\beta-y)}\int_0^1 \theta^{\alpha+y}(1-\theta)^{\beta+n-y-1}\,d\theta \\ &= \frac{\alpha+y}{\alpha+\beta+n}.\end{aligned}$$

This decision function $\delta(y)$ minimizes

$$\int_0^1 [\theta-\delta(y)]^2 k(\theta|y)\,d\theta$$

for $y = 0, 1, \ldots, n$ and, accordingly, it minimizes the expected risk. It is very instructive to note that this Bayes' solution can be written as

$$\delta(y) = \left(\frac{n}{\alpha+\beta+n}\right)\frac{y}{n} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta}$$

which is a weighted average of the maximum likelihood estimate y/n of θ and the mean $\alpha/(\alpha+\beta)$ of the prior p.d.f. of the parameter. Moreover, the respective weights are $n/(\alpha+\beta+n)$ and $(\alpha+\beta)/(\alpha+\beta+n)$. Thus we see that α and β should be selected so that not only is $\alpha/(\alpha+\beta)$ the desired prior mean, but the sum $\alpha+\beta$ indicates the worth of the prior opinion, relative to a sample of size n. That is, if we want our prior opinion to have as much weight as a sample size of 20, we would take $\alpha+\beta = 20$. So if our prior mean is $\frac{3}{4}$; we have that α and β are selected so that $\alpha = 15$ and $\beta = 5$.

Example 3. Suppose that $Y = \bar{X}$, the sufficient statistic, is the mean of a random sample of size n that arises from the normal distribution $N(\theta, \sigma^2)$, where σ^2 is known. Then $g(y|\theta)$ is $N(\theta, \sigma^2/n)$. Further suppose that we

are able to assign prior knowledge to θ through a prior p.d.f. $h(\theta)$ that is $N(\theta_0, \sigma_0^2)$. Then we have that

$$k(\theta|y) \propto \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(y-\theta)^2}{2(\sigma^2/n)} - \frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right].$$

If we eliminate all constant factors (including factors involving y only), we have

$$k(\theta|y) \propto \exp\left[-\frac{(\sigma_0^2 + \sigma^2/n)\theta^2 - 2(y\sigma_0^2 + \theta_0\sigma^2/n)\theta}{2(\sigma^2/n)\sigma_0^2}\right].$$

This can be simplified, by completing the square, to read (after eliminating factors not involving θ)

$$k(\theta|y) \propto \exp\left[-\frac{\left(\theta - \dfrac{y\sigma_0^2 + \theta_0\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\right)^2}{\dfrac{2(\sigma^2/n)\sigma_0^2}{(\sigma_0^2 + \sigma^2/n)}}\right].$$

That is, the posterior p.d.f. of the parameter is obviously normal with mean

$$\frac{y\sigma_0^2 + \theta_0\sigma^2/n}{\sigma_0^2 + \sigma^2/n} = \left(\frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\right)y + \left(\frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\right)\theta_0$$

and variance $(\sigma^2/n)\sigma_0^2/(\sigma_0^2 + \sigma^2/n)$. If the square-error loss function is used, this posterior mean is the Bayes' solution. Again, note that it is a weighted average of the maximum likelihood estimate $y = \bar{x}$ and the prior mean θ_0. Observe here and in Example 2 that the Bayes' solution gets closer to the maximum likelihood estimate as n increases. Thus the Bayesian procedures permit the decision maker to enter his or her prior opinions into the solution in a very formal way such that the influences of these prior notions will be less and less as n increases.

In Bayesian statistics all the information is contained in the posterior p.d.f. $k(\theta|y)$. In Examples 2 and 3 we found Bayesian point estimates using the square-error loss function. It should be noted that if $\mathscr{L}[\delta(y), \theta] = |\delta(y) - \theta|$, the absolute value of the error, then the Bayes' solution would be the median of the posterior distribution of the parameter, which is given by $k(\theta|y)$. Hence the Bayes' solution changes, *as it should*, with different loss functions.

If an interval estimate of θ is desired, we can now find two functions $u(y)$ and $v(y)$ so that the conditional probability

$$\Pr[u(y) < \Theta < v(y)|Y = y] = \int_{u(y)}^{v(y)} k(\theta|y)\, d\theta,$$

is large, say 0.95. The experimental values of $X_1, X_2, \ldots, X_n$, say

$x_1, x_2, \ldots, x_n$, provide us with an experimental value of Y, say y. Then the interval $u(y)$ to $v(y)$ is an interval estimate of θ in the sense that the conditional probability of Θ belonging to that interval is equal to 0.95. For illustration, in Example 3 where the posterior p.d.f. of the parameter was normal, the interval, whose end points are found by taking the mean of that distribution and adding and subtracting 1.96 of its standard deviation,

$$\frac{y\sigma_0^2 + \theta_0\sigma^2/n}{\sigma_0^2 + \sigma^2/n} \pm 1.96\sqrt{\frac{(\sigma^2/n)\sigma_0^2}{\sigma_0^2 + \sigma^2/n}}$$

serves as an interval estimate for θ with posterior probability of 0.95.

EXERCISES

8.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that is $b(1, \theta)$. Let the prior p.d.f. of Θ be a beta one with parameters α and β. Show that the posterior p.d.f. $k(\theta|x_1, x_2, \ldots, x_n)$ is exactly the same as $k(\theta|y)$ given in Example 2.

8.2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $N(\theta, \sigma^2)$, $-\infty < \theta < \infty$, where σ^2 is a given positive number. Let $Y = \bar{X}$, the mean of the random sample. Take the loss function to be $\mathscr{L}[\theta, \delta(y)] = |\theta - \delta(y)|$. If θ is an observed value of the random variable Θ that is $N(\mu, \tau^2)$, where $\tau^2 > 0$ and μ are known numbers, find the Bayes' solution $\delta(y)$ for a point estimate of θ.

8.3. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution with mean θ, $0 < \theta < \infty$. Let $Y = \sum_1^n X_i$ and take the loss function to be $\mathscr{L}[\theta, \delta(y)] = [\theta - \delta(y)]^2$. Let θ be an observed value of the random variable Θ. If Θ has the p.d.f. $h(\theta) = \theta^{\alpha-1}e^{-\theta/\beta}/\Gamma(\alpha)\beta^\alpha$, $0 < \theta < \infty$, zero elsewhere, where $\alpha > 0$, $\beta > 0$ are known numbers, find the Bayes' solution $\delta(y)$ for a point estimate of θ.

8.4. Let Y_n be the nth order statistic of a random sample of size n from a distribution with p.d.f. $f(x|\theta) = 1/\theta$, $0 < x < \theta$, zero elsewhere. Take the loss function to be $\mathscr{L}[\theta, \delta(y_n)] = [\theta - \delta(y_n)]^2$. Let θ be an observed value of the random variable Θ, which has p.d.f. $h(\theta) = \beta\alpha^\beta/\theta^{\beta+1}$, $\alpha < \theta < \infty$, zero elsewhere, with $\alpha > 0$, $\beta > 0$. Find the Bayes' solution $\delta(y_n)$ for a point estimate of θ.

8.5. Let Y_1 and Y_2 be statistics that have a trinomial distribution with parameters n, θ_1, and θ_2. Here θ_1 and θ_2 are observed values of the random variables Θ_1 and Θ_2, which have a Dirichlet distribution with known

parameters α_1, α_2, and α_3 (see Example 1, Section 4.5). Show that the conditional distribution of Θ_1 and Θ_2 is Dirichlet and determine the conditional means $E(\Theta_1|y_1, y_2)$ and $E(\Theta_2|y_1, y_2)$.

8.6. Let X be $N(0, 1/\theta)$. Assume that the unknown θ is a value of a random variable Θ which has a gamma distribution with parameters $\alpha = r/2$ and $\beta = 2/r$, where r is a positive integer. Show that X has a marginal t-distribution with r degrees of freedom. This procedure is called *compounding*, and it may be used by a Bayesian statistician as a way of first presenting the t-distribution, as well as other distributions.

8.7. Let X have a Poisson distribution with parameter θ. Assume that the unknown θ is a value of a random variable Θ that has a gamma distribution with parameters $\alpha = r$ and $\beta = (1 - p)/p$, where r is a positive integer and $0 < p < 1$. Show, by the procedure of compounding, that X has a marginal distribution which is negative binomial, a distribution that was introduced earlier (Section 3.1) under very different assumptions.

8.8. In Example 2 let $n = 30$, $\alpha = 10$, and $\beta = 5$ so that $\delta(y) = (10 + y)/45$ is the Bayes' estimate of θ.

(a) If Y has the binomial distribution $b(30, \theta)$, compute the risk $E\{[\theta - \delta(Y)]^2\}$.

(b) Determine those values of θ for which the risk of part (a) is less than $\theta(1 - \theta)/30$, the risk associated with the maximum likelihood estimator Y/n of θ.

8.9. Let Y_4 be the largest order statistic of a sample of size $n = 4$ from a distribution with uniform p.d.f. $f(x; \theta) = 1/\theta$, $0 < x < \theta$, zero elsewhere. If the prior p.d.f. of the parameter is $g(\theta) = 2/\theta^3$, $1 < \theta < \infty$, zero elsewhere, find the Bayesian estimator $\delta(Y_4)$ of θ, based upon the sufficient statistic Y_4, using the loss function $|\delta(y_4) - \theta|$.

8.10. Consider a random sample $X_1, X_2, \ldots, X_n$ from the *Weibull distribution* with p.d.f. $f(x; \theta, \tau) = \theta\tau x^{\tau-1}e^{-\theta x^\tau}$, $0 < x < \infty$, where $0 < \theta$, $0 < \tau$, zero elsewhere.

(a) If τ is known, find the m.l.e. of θ.

(b) If the parameter θ has a prior gamma p.d.f. $g(\theta)$ with parameters α and $\beta^* = 1/\beta$, show that the compound distribution is a *Burr type* with p.d.f. $h(x) = \alpha\tau\beta^\alpha x^{\tau-1}/(x^\tau + \beta)^{\alpha+1}$, $0 < x < \infty$, zero elsewhere.

(c) If, in the Burr distribution, τ and β are known, find the m.l.e. of α based on a random sample of size n.

8.2 Fisher Information and the Rao–Cramér Inequality

Let X be a random variable with p.d.f. $f(x; \theta)$, $\theta \in \Omega$, where the parameter space Ω is an interval. We consider only special cases,

sometimes called *regular cases*, of probability density functions as we wish to differentiate under an integral (summation) sign. In particular, this means that the parameter θ does not appear in endpoints of the interval in which $f(x; \theta) > 0$.

With these assumptions, we have (in the continuous case, but the discrete case can be handled in a similar manner) that

$$\int_{-\infty}^{\infty} f(x; \theta)\, dx = 1$$

and, by taking the derivative with respect to θ,

$$\int_{-\infty}^{\infty} \frac{\partial f(x; \theta)}{\partial \theta}\, dx = 0. \tag{1}$$

The latter expression can be rewritten as

$$\int_{-\infty}^{\infty} \frac{\dfrac{\partial f(x; \theta)}{\partial \theta}}{f(x; \theta)} f(x; \theta)\, dx = 0$$

or, equivalently,

$$\int_{-\infty}^{\infty} \frac{\partial \ln f(x; \theta)}{\partial \theta} f(x; \theta)\, dx = 0.$$

If we differentiate again, it follows that

$$\int_{-\infty}^{\infty} \left[\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} f(x; \theta) + \frac{\partial \ln f(x; \theta)}{\partial \theta} \frac{\partial f(x; \theta)}{\partial \theta} \right] dx = 0. \tag{2}$$

We rewrite the second term of the left-hand member of this equation as

$$\int_{-\infty}^{\infty} \frac{\partial \ln f(x; \theta)}{\partial \theta} \frac{\dfrac{\partial f(x; \theta)}{\partial \theta}}{f(x; \theta)} f(x; \theta)\, dx = \int_{-\infty}^{\infty} \left[\frac{\partial \ln f(x; \theta)}{\partial \theta} \right]^2 f(x; \theta)\, dx.$$

This is called *Fisher information* and is denoted by $I(\theta)$. That is,

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{\partial \ln f(x; \theta)}{\partial \theta} \right]^2 f(x; \theta)\, dx;$$

but, from Equation (2), we see that $I(\theta)$ can be computed from

$$I(\theta) = -\int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} f(x; \theta)\, dx.$$

Sometimes, one expression is easier to compute than the other, but often we prefer the second expression.

Remark. Note that the information is the weighted mean of either

$$\left[\frac{\partial \ln f(x;\theta)}{\partial\theta}\right]^2 \quad \text{or} \quad -\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2},$$

where the weights are given by the p.d.f. $f(x;\theta)$. That is, the greater these derivatives on the average, the more information that we get about θ. Clearly, if they were equal to zero [so that θ would not be in $\ln f(x;\theta)$], there would be zero information about θ. As we study more and more statistics, we learn to recognize that the function

$$\frac{\partial \ln f(x;\theta)}{\partial\theta}$$

is a very important one. For example, it played a major role in finding the m.l.e. $\hat{\theta}$ by solving

$$\sum_{i=1}^{n}\frac{\partial \ln f(x_i;\theta)}{\partial\theta}=0$$

for θ.

Example 1. Let X be $N(\theta,\sigma^2)$, where $-\infty<\theta<\infty$ and σ^2 is known. Then

$$f(x;\theta)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right], \qquad -\infty<x<\infty,$$

where $-\infty<\theta<\infty$, and

$$\ln f(x;\theta)=-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{(x-\theta)^2}{2\sigma^2}.$$

Thus

$$\frac{\partial \ln f(x;\theta)}{\partial\theta}=\frac{x-\theta}{\sigma^2}$$

and

$$\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}=\frac{-1}{\sigma^2}.$$

Clearly, $E[(X-\theta)^2/\sigma^4]=-E[-1/\sigma^2]=1/\sigma^2$. That is, in this case, it does not matter much which way we compute $I(\theta)$, as

$$I(\theta)=E\left\{\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2\right\} \quad \text{or} \quad I(\theta)=-E\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right],$$

because each is very easy. Of course, the information is greater with smaller values of σ^2.

Example 2. Let X be binomial $b(1, \theta)$. Thus

$$\ln f(x; \theta) = x \ln \theta + (1 - x) \ln (1 - \theta),$$

$$\frac{\partial \ln f(x; \theta)}{\partial \theta} = \frac{x}{\theta} - \frac{1 - x}{1 - \theta},$$

and

$$\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{1 - x}{(1 - \theta)^2}.$$

Clearly,

$$I(\theta) = -E\left[\frac{-X}{\theta^2} - \frac{1 - X}{(1 - \theta)^2}\right]$$

$$= \frac{\theta}{\theta^2} + \frac{1 - \theta}{(1 - \theta)^2} = \frac{1}{\theta} + \frac{1}{1 - \theta} = \frac{1}{\theta(1 - \theta)},$$

which is larger for θ values close to zero or 1.

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution having p.d.f. $f(x; \theta)$. Thus the likelihood function (the joint p.d.f. of $X_1, X_2, \ldots, X_n$) is

$$L(\theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta).$$

Of course,

$$\ln L(\theta) = \ln f(x_1; \theta) + \ln f(x_2; \theta) + \cdots + \ln f(x_n; \theta)$$

and

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\partial \ln f(x_1; \theta)}{\partial \theta} + \frac{\partial \ln f(x_2; \theta)}{\partial \theta} + \cdots + \frac{\partial \ln f(x_n; \theta)}{\partial \theta}. \quad (3)$$

It seems reasonable to define the *Fisher information* in the random sample as

$$I_n(\theta) = E\left\{\left[\frac{\partial \ln L(\theta)}{\partial \theta}\right]^2\right\}.$$

Note if we square Equation (3), we obtain cross-product terms like

$$2E\left[\frac{\partial \ln f(X_i; \theta)}{\partial \theta} \frac{\partial \ln f(X_j; \theta)}{\partial \theta}\right], \qquad i \neq j,$$

which from the independence of X_i and X_j equals

$$2E\left[\frac{\partial \ln f(X_i;\theta)}{\partial\theta}\right]E\left[\frac{\partial \ln f(X_j;\theta)}{\partial\theta}\right]=0.$$

The fact that this product equals zero follows immediately from Equation (1). Hence we have the result that

$$I_n(\theta)=\sum_{i=1}^{n}E\left\{\left[\frac{\partial \ln f(X_i;\theta)}{\partial\theta}\right]^2\right\}.$$

However, each term of this summation equals $I(\theta)$, and hence

$$I_n(\theta)=nI(\theta).$$

That is, the Fisher information in a random sample of size n is n times the Fisher information in one observation. So, in the two examples of this section, the Fisher information in a random sample of size n is n/σ^2 in Example 1 and $n/[\theta(1-\theta)]$ in Example 2.

We can now prove a very important inequality involving the variance of an estimator, say $Y=u(X_1,X_2,\ldots,X_n)$, of θ, which can be biased. Suppose that

$$E(Y)=E[u(X_1,X_2,\ldots,X_n)]=k(\theta).$$

That is, in the continuous case,

$$k(\theta)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}u(x_1,\ldots,x_n)f(x_1;\theta)\cdots f(x_n;\theta)\,dx_1\cdots dx_n;$$

$$\begin{aligned}k'(\theta)&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}u(x_1,x_2,\ldots,x_n)\left[\sum_{1}^{n}\frac{1}{f(x_i;\theta)}\frac{\partial f(x_i;\theta)}{\partial\theta}\right]\\&\qquad\times f(x_1;\theta)\cdots f(x_n;\theta)\,dx_1\cdots dx_n\\&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}u(x_1,x_2,\ldots,x_n)\left[\sum_{1}^{n}\frac{\partial \ln f(x_i;\theta)}{\partial\theta}\right]\\&\qquad\times f(x_1;\theta)\cdots f(x_n;\theta)\,dx_1\cdots dx_n.\end{aligned}\tag{4}$$

Define the random variable Z by $Z=\sum_{1}^{n}[\partial \ln f(X_i;\theta)/\partial\theta]$. In accordance with Equation (1) we have $E(Z)=\sum_{1}^{n}E[\partial \ln f(X_i;\theta)/\partial\theta]=0$. Moreover, Z is the sum of n independent random variables each with mean zero and consequently with variance $E\{[\partial \ln f(X;\theta)/\partial\theta]^2\}$. Hence the variance of Z is the sum of the n variances,

$$\sigma_Z^2=nE\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right]=I_n(\theta)=nI(\theta).$$

Because $Y = u(X_1, \ldots, X_n)$ and $Z = \sum_1^n [\partial \ln f(X_i; \theta)/\partial\theta]$, Equation (4) shows that $E(YZ) = k'(\theta)$. Recall that

$$E(YZ) = E(Y)E(Z) + \rho\sigma_Y\sigma_Z,$$

where ρ is the correlation coefficient of Y and Z. Since $E(Y) = k(\theta)$ and $E(Z) = 0$, we have

$$k'(\theta) = k(\theta) \cdot 0 + \rho\sigma_Y\sigma_Z \qquad \text{or} \qquad \rho = \frac{k'(\theta)}{\sigma_Y\sigma_Z}.$$

Now $\rho^2 \leq 1$. Hence

$$\frac{[k'(\theta)]^2}{\sigma_Y^2\sigma_Z^2} \leq 1 \qquad \text{or} \qquad \frac{[k'(\theta)]^2}{\sigma_Z^2} \leq \sigma_Y^2.$$

If we replace σ_Z^2 by its value, we have

$$\sigma_Y^2 \geq \frac{[k'(\theta)]^2}{nE\left[\left(\dfrac{\partial \ln f(X; \theta)}{\partial\theta}\right)^2\right]} = \frac{[k'(\theta)]^2}{nI(\theta)}.$$

This inequality is known as the *Rao–Cramér inequality*.

If $Y = u(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of θ, so that $k(\theta) = \theta$, then the Rao–Cramér inequality becomes, since $k'(\theta) = 1$,

$$\sigma_Y^2 \geq \frac{1}{nI(\theta)}.$$

Note that in Examples 1 and 2 of this section $1/nI(\theta)$ equals σ^2/n and $\theta(1 - \theta)/n$, respectively. In each case, the unbiased estimator, $\bar{X}$, of θ, which is based upon the sufficient statistic for θ, has a variance that is equal to this Rao–Cramér lower bound of $1/nI(\theta)$.

We now make the following definitions.

Definition 1. Let Y be an unbiased estimator of a parameter θ in such a case of point estimation. The statistic Y is called an *efficient estimator* of θ if and only if the variance of Y attains the Rao–Cramér lower bound.

Definition 2. In cases in which we can differentiate with respect to a parameter under an integral or summation symbol, the ratio of the Rao–Cramér lower bound to the actual variance of any unbiased estimation of a parameter is called the *efficiency* of that statistic.

Example 3. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson

distribution that has the mean $\theta > 0$. It is known that $\bar{X}$ is an m.l.e. of θ; we shall show that it is also an efficient estimator of θ. We have

$$\frac{\partial \ln f(x;\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}(x \ln \theta - \theta - \ln x!)$$
$$= \frac{x}{\theta} - 1 = \frac{x-\theta}{\theta}.$$

Accordingly,

$$E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial \theta}\right)^2\right] = \frac{E(X-\theta)^2}{\theta^2} = \frac{\sigma^2}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}.$$

The Rao–Cramér lower bound in this case is $1/[n(1/\theta)] = \theta/n$. But θ/n is the variance of $\bar{X}$. Hence $\bar{X}$ is an efficient estimator of θ.

Example 4. Let S^2 denote the variance of a random sample of size $n > 1$ from a distribution that is $N(\mu, \theta)$, $0 < \theta < \infty$, where μ is known. We know that $E[nS^2/(n-1)] = \theta$. What is the efficiency of the estimator $nS^2/(n-1)$? We have

$$\ln f(x;\theta) = -\frac{(x-\mu)^2}{2\theta} - \frac{\ln(2\pi\theta)}{2},$$

$$\frac{\partial \ln f(x;\theta)}{\partial \theta} = \frac{(x-\mu)^2}{2\theta^2} - \frac{1}{2\theta},$$

and

$$\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2} = -\frac{(x-\mu)^2}{\theta^3} + \frac{1}{2\theta^2}.$$

Accordingly,

$$-E\left[\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right] = \frac{\theta}{\theta^3} - \frac{1}{2\theta^2} = \frac{1}{2\theta^2}.$$

Thus the Rao–Cramér lower bound is $2\theta^2/n$. Now nS^2/θ is $\chi^2(n-1)$, so the variance of nS^2/θ is $2(n-1)$. Accordingly, the variance of $nS^2/(n-1)$ is $2(n-1)[\theta^2/(n-1)^2] = 2\theta^2/(n-1)$. Thus the efficiency of the estimator $nS^2/(n-1)$ is $(n-1)/n$. With μ known, what is the efficient estimator of the variance?

Example 5. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size $n > 2$ from a distribution with p.d.f.

$$f(x;\theta) = \theta x^{\theta-1} = \exp(\theta \ln x - \ln x + \ln \theta), \qquad 0 < x < 1,$$
$$= 0 \qquad \text{elsewhere.}$$

It is easy to verify that the Rao–Cramér lower bound is θ^2/n. Let

$Y_i = -\ln X_i$. We shall indicate that each Y_i has a gamma distribution. The associated transform $y_i = -\ln x_i$, with inverse $x_i = e^{-y_i}$, is one-to-one and the transformation maps the space $\{x_i: 0 < x_i < 1\}$ onto the space $\{y_i: 0 < y_i < \infty\}$. We have $|J| = e^{-y_i}$. Thus Y_i has a gamma distribution with $\alpha = 1$ and $\beta = 1/\theta$. Let $Z = -\sum_1^n \ln X_i$. Then Z has a gamma distribution with $\alpha = n$ and $\beta = 1/\theta$. Accordingly, we have $E(Z) = \alpha\beta = n/\theta$. This suggests that we compute the expectation of $1/Z$ to see if we can find an unbiased estimator of θ. A simple integration shows that $E(1/Z) = \theta/(n-1)$. Hence $(n-1)/Z$ is an unbiased estimator of θ. With $n > 2$, the variance of $(n-1)/Z$ exists and is found to be $\theta^2/(n-2)$, so that the efficiency of $(n-1)/Z$ is $(n-2)/n$. This efficiency tends to 1 as n increases. In such an instance, the estimator is said to be *asymptotically efficient*.

The concept of joint efficient estimators of several parameters has been developed along with the associated concept of joint efficiency of several estimators. But limitations of space prevent their inclusion in this book.

EXERCISES

8.11. Prove that $\bar{X}$, the mean of a random sample of size n from a distribution that is $N(\theta, \sigma^2)$, $-\infty < \theta < \infty$, is, for every known $\sigma^2 > 0$, an efficient estimator of θ.

8.12. Show that the mean $\bar{X}$ of a random sample of size n from a distribution which is $b(1, \theta)$, $0 < \theta < 1$, is an efficient estimator of θ.

8.13. Given $f(x; \theta) = 1/\theta$, $0 < x < \theta$, zero elsewhere, with $\theta > 0$, formally compute the reciprocal of

$$nE\left\{\left[\frac{\partial \ln f(X; \theta)}{\partial \theta}\right]^2\right\}.$$

Compare this with the variance of $(n+1)Y_n/n$, where Y_n is the largest item of a random sample of size n from this distribution. Comment.

8.14. Given the p.d.f.

$$f(x; \theta) = \frac{1}{\pi[1 + (x-\theta)^2]}, \qquad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Show that the Rao–Cramér lower bound is $2/n$, where n is the size of a random sample from this Cauchy distribution.

8.15. Let X have a gamma distribution with $\alpha = 4$ and $\beta = \theta > 0$.
(a) Find the Fisher information $I(\theta)$.

(b) If $X_1, X_2, \ldots, X_n$ is a random sample from this distribution, show that the m.l.e. of θ is an efficient estimator of θ.

8.16. Let X be $N(0, \theta)$, $0 < \theta < \infty$.
(a) Find the Fisher information $I(\theta)$.
(b) If $X_1, X_2, \ldots, X_n$ is a random sample from this distribution, show that the m.l.e. of θ is an efficient estimator of θ.

8.3 Limiting Distributions of Maximum Likelihood Estimators

We use the notation and assumptions of Section 8.2 as much as possible here. In particular, $f(x; \theta)$ is the p.d.f., $I(\theta)$ is the Fisher information, and the likelihood function is

$$L(\theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta).$$

Also, we can differentiate under the integral (summation) sign, so that

$$Z = \frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(X_i; \theta)}{\partial \theta}$$

has mean zero and variance $nI(\theta)$. In addition, we want to be able to find the maximum likelihood estimator $\hat{\theta}$ by solving

$$\frac{\partial[\ln L(\theta)]}{\partial \theta} = 0.$$

That is,

$$\frac{\partial[\ln L(\hat{\theta})]}{\partial \theta} = 0,$$

where now, with $\hat{\theta}$ in this expression, $L(\hat{\theta}) = f(X_1; \hat{\theta}) \cdots f(X_n; \hat{\theta})$. We can approximate the left-hand member of this latter equation by a linear function found from the first two terms of a Taylor's series expanded about θ, namely

$$\frac{\partial[\ln L(\theta)]}{\partial \theta} + (\hat{\theta} - \theta) \frac{\partial^2[\ln L(\theta)]}{\partial \theta^2} \approx 0,$$

when $L(\theta) = f(X_1; \theta) f(X_2; \theta) \cdots f(X_n; \theta)$.

Obviously, this approximation is good enough only if $\hat{\theta}$ is close to θ, and an adequate mathematical proof involves certain regularity

conditions, all of which we have not given here. But a heuristic argument can be made by solving for $\hat{\theta} - \theta$ to obtain

$$\hat{\theta} - \theta = \frac{\dfrac{\partial[\ln L(\theta)]}{\partial\theta}}{-\dfrac{\partial^2[\ln L(\theta)]}{\partial\theta^2}} = \frac{Z}{-\dfrac{\partial^2[\ln L(\theta)]}{\partial\theta^2}}.$$

Let us rewrite this equation as

$$\frac{\hat{\theta} - \theta}{\sqrt{\dfrac{1}{nI(\theta)}}} = \frac{Z/\sqrt{nI(\theta)}}{-\dfrac{1}{n}\dfrac{\partial^2[\ln L(\theta)]}{\partial\theta^2}\Big/I(\theta)}. \tag{1}$$

Since Z is the sum of the i.i.d. random variables

$$\frac{\partial \ln f(X_i; \theta)}{\partial\theta}, \qquad i = 1, 2, \ldots, n,$$

each with mean zero and variance $I(\theta)$, the numerator of the right-hand member of Equation (1) is limiting $N(0, 1)$ by the central limit theorem. Moreover, the mean

$$\frac{1}{n}\sum_{i=1}^{n}\frac{-\partial^2 \ln f(X_i; \theta)}{\partial\theta^2}$$

converges in probability to its expected value, namely $I(\theta)$. So the denominator of the right-hand member of Equation (1) converges in probability 1. Thus, by Slutsky's theorem given in Section 5.5, the right-hand member of Equation (1) is limiting $N(0, 1)$. Hence the left-hand member also has this limiting standard normal distribution. That means that we can say that $\hat{\theta}$ has an approximate normal distribution with mean θ and variance $1/nI(\theta)$.

The preceding result means that in a regular case of estimation and in some limiting sense, the m.l.e. $\hat{\theta}$ is unbiased and its variance achieves the Rao–Cramér lower bound. That is, the m.l.e. $\hat{\theta}$ is asymptotically efficient.

Example 1. In Exercise 8.14 we examined the Rao–Cramér lower bound of the variance of an unbiased estimator of θ, the median of a certain Cauchy distribution. We now know that the m.l.e. $\hat{\theta}$ of θ has an approximate normal distribution with mean θ and variance equal to the lower bound of $2/n$. Hence, once we compute $\hat{\theta}$, we can say, for illustration, that $\hat{\theta} \pm 1.96\sqrt{2/n}$ provides an approximate 95 percent confidence interval for θ.

To determine $\hat{\theta}$, there are many numerical methods that can be used. In the Cauchy case, one of the easiest is given by the following:

$$0 = \frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}.$$

In the denominator of the right-hand member, we use a preliminary estimate of θ that is not influenced too much by extreme observations. For illustration, the sample median, say $\hat{\theta}_0$, is very good one while the sample mean $\bar{x}$ would be a poor choice. This provides weights

$$w_{i1} = \frac{2}{1 + (x_i - \hat{\theta}_0)^2}, \qquad i = 1, 2, \ldots, n,$$

so that we can solve

$$0 = \sum_{i=1}^{n} (w_{i1})(x_i - \theta) \quad \text{to get} \quad \hat{\theta}_1 = \frac{\sum w_{i1} x_i}{\sum w_{i1}}.$$

Now $\hat{\theta}_1$ can be used to obtain new weights and $\hat{\theta}_2$:

$$w_{i2} = \frac{2}{1 + (x_i - \hat{\theta}_1)^2}, \qquad \hat{\theta}_2 = \frac{\sum w_{i2} x_i}{\sum w_{i2}}.$$

This iterative process can be continued until adequate convergence is obtained; that is, at some step k, $\hat{\theta}_k$ will be close enough to $\hat{\theta}$ to be used as the m.l.e.

Example 2. Suppose that the random sample arises from a distribution with p.d.f.

$$f(x; \theta) = \theta x^{\theta - 1}, \qquad 0 < x < 1, \quad \theta \in \Omega = \{\theta : 0 < \theta < \infty\},$$

zero elsewhere. We have

$$\ln f(x; \theta) = \ln \theta + (\theta - 1) \ln x,$$

$$\frac{\partial \ln f(x; \theta)}{\partial \theta} = \frac{1}{\theta} + \ln x,$$

and

$$\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} = -\frac{1}{\theta^2}$$

Since $E(-1/\theta^2) = -1/\theta^2$, the lower bound of the variance of every unbiased estimator of θ is θ^2/n. Moreover, the maximum likelihood estimator $\hat{\theta} = -n/\ln \prod_{i=1}^{n} X_i$ has an approximate normal distribution with mean θ and variance θ^2/n. Thus, in a limiting sense, $\hat{\theta}$ is the unbiased minimum variance estimator of θ; that is, $\hat{\theta}$ is asymptotically efficient.

Example 3. The m.l.e. for θ in

$$f(x;\theta)=\frac{\theta^x e^{-\theta}}{x!}, \qquad x=0,1,2,\ldots, \quad \theta\in\Omega=\{\theta:0<\theta<\infty\},$$

is $\hat{\theta}=\bar{X}$, the mean of a random sample. Now

$$\ln f(x;\theta)=x\ln\theta-\theta-\ln x!$$

and

$$\frac{\partial[\ln f(x;\theta)]}{\partial\theta}=\frac{x}{\theta}-1 \qquad \text{and} \qquad \frac{\partial^2[\ln f(x;\theta)]}{\partial\theta^2}=-\frac{x}{\theta^2}.$$

Thus

$$-E\left(-\frac{X}{\theta^2}\right)=\frac{\theta}{\theta^2}=\frac{1}{\theta}$$

and $\hat{\theta}=\bar{X}$ has an approximate normal distribution with mean θ and standard deviation $\sqrt{\theta/n}$. That is, $Y=(\bar{X}-\theta)/\sqrt{\theta/n}$ has a limiting standard normal distribution. The problem in practice is how best to estimate the standard deviation in the denominator of Y. Clearly, we might use $\bar{X}$ for θ there, but does that create too much dependence between the numerator and denominator? If so, this requires a very large sample size for $(\bar{X}-\theta)/\sqrt{\bar{X}/n}$ to have an approximate normal distribution. It might be better to approximate $I(\theta)$ by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial[\ln f(x_i;\hat{\theta})]}{\partial\hat{\theta}}\right\}^2=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\bar{x}}-1\right)^2=\frac{s^2}{\bar{x}^2}.$$

Thus $nI(\theta)$ is approximated by $ns^2/\bar{x}^2$ and we can say that

$$\frac{\sqrt{n}(\bar{X}-\theta)}{\bar{X}/S}$$

is approximately $N(0,1)$. We do not know exactly which of these two solutions, or others like simply using $s/\sqrt{n}$ in the denominator, is best. Fortunately, however, if the Poisson model is correct, usually

$$\sqrt{\frac{\bar{x}}{n}}\approx\frac{\bar{x}}{\sqrt{n}\,s}\approx\frac{s}{\sqrt{n}}.$$

If this is not true, we should check the Poisson assumption, which requires, among other things, that $\mu=\sigma^2$. Hence, for illustration, either

$$\bar{x}\pm 1.96\sqrt{\frac{\bar{x}}{n}} \quad \text{or} \quad \bar{x}\pm\frac{1.96\bar{x}}{\sqrt{n}\,s} \quad \text{or} \quad \bar{x}\pm\frac{1.96s}{\sqrt{n}}$$

serves as an approximate 95 percent confidence interval for θ. In situations like this, we recommend that a person try all three because they should be in substantial agreement. If not, check the Poisson assumption.

The fact that the m.l.e. $\hat{\theta}$ has an approximate normal distribution with mean θ and variance $1/nI(\theta)$ suggests that $\hat{\theta}$ (really a sequence $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_n, \ldots$) converges in probability to θ. Of course, $\hat{\theta}_n$ can be biased; say $E(\hat{\theta}_n - \theta) = b_n(\theta)$, where $b_n(\theta)$ is the bias. However, $b_n(\theta)$ equals zero in the limit. Moreover, if we assume that the variances exist and

$$\lim_{n\to\infty} [\text{var}\,(\hat{\theta}_n)] = \lim_{n\to\infty} \left[\frac{1}{nI(\theta)}\right],$$

then the limit of the variances is obviously zero. Hence, from Chebyshev's inequality, we have

$$\Pr\,[|\hat{\theta}_n - \theta| \geq \epsilon] \leq \frac{E[(\hat{\theta}_n - \theta)^2]}{\epsilon^2}.$$

However,

$$\lim_{n\to\infty} E[(\hat{\theta}_n - \theta)^2] = \lim_{n\to\infty} [b_n^2(\theta) + \text{var}\,(\hat{\theta}_n)] = 0$$

and thus

$$\lim_{n\to\infty} \Pr\,[|\hat{\theta}_n - \theta| \geq \epsilon] = 0$$

for each fixed $\epsilon > 0$. Any estimator, not just maximum likelihood estimators, that enjoys this property is said to be a *consistent estimator* of θ. As illustrations, we note that all the unbiased estimators based upon the complete sufficient statistics in Chapter 7 and all the estimators in Sections 8.1 and 8.2 are consistent ones.

We close this section by considering the extension of these limiting distributions to maximum likelihood estimators of two or more parameters. For convenience, we restrict ourselves to the regular case involving two parameters, but the extension to more than two is obvious once the reader understands multivariate normal distributions (Section 4.10).

Suppose that the random sample $X_1, X_2, \ldots, X_n$ arises from a distribution with p.d.f. $f(x; \theta_1, \theta_2)$, $(\theta_1, \theta_2) \in \Omega$, in which regularity conditions exist. Without describing these conditions in any detail, let us simply say that the space of X where $f(x; \theta_1, \theta_2) > 0$ does not

involve θ_1 and θ_2, and we are able to differentiate under the integral (summation) signs. The information matrix of the sample is equal to

$$\mathbf{I}_n = n \times$$

$$\begin{bmatrix} E\left\{\left[\frac{\partial \ln f(X;\theta_1,\theta_2)}{\partial\theta_1}\right]^2\right\}, & E\left\{\frac{\partial \ln f(X;\theta_1,\theta_2)}{\partial\theta_1}\frac{\partial \ln f(X;\theta_1,\theta_2)}{\partial\theta_2}\right\}, \\ E\left\{\frac{\partial \ln f(X;\theta_1,\theta_2)}{\partial\theta_1}\frac{\partial \ln f(X;\theta_1,\theta_2)}{\partial\theta_2}\right\} & E\left\{\left[\frac{\partial \ln f(X;\theta_1,\theta_2)}{\partial\theta_2}\right]^2\right\} \end{bmatrix}$$

$$= -n \begin{bmatrix} E\left[\frac{\partial^2 \ln f(X;\theta_1,\theta_2)}{\partial\theta_1^2}\right] & E\left[\frac{\partial^2 \ln f(X;\theta_1,\theta_2)}{\partial\theta_1\,\partial\theta_2}\right] \\ E\left[\frac{\partial^2 \ln f(X;\theta_1,\theta_2)}{\partial\theta_1\,\partial\theta_2}\right] & E\left[\frac{\partial^2 \ln f(X;\theta_1,\theta_2)}{\partial\theta_2^2}\right] \end{bmatrix}.$$

One can immediately see the similarity of this to the one-parameter case.

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are maximum likelihood estimators of θ_1 and θ_2, then $\hat{\theta}_1$ and $\hat{\theta}_2$ have an approximate bivariate normal distribution with means θ_1 and θ_2 and variance–covariance matrix $\mathbf{I}_n^{-1}$. That is, the approximate variances and covariances are found, respectively, in the matrix

$$\mathbf{I}_n^{-1} \approx \begin{pmatrix} \text{var}\,(\hat{\theta}_1) & \text{cov}\,(\hat{\theta}_1, \hat{\theta}_2) \\ \text{cov}\,(\hat{\theta}_1, \hat{\theta}_2) & \text{var}\,(\hat{\theta}_2) \end{pmatrix}.$$

An illustration will help us understand this result that has simply been given to the reader to accept without any mathematical derivation.

Example 4. Let the random sample $X_1, X_2, \ldots, X_n$ arise from $N(\theta_1, \theta_2)$. Then

$$\ln f(x;\theta_1,\theta_2) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{(x-\theta_1)^2}{2\theta_2},$$

$$\frac{\partial \ln f(x;\theta_1,\theta_2)}{\partial\theta_1} = \frac{x-\theta_1}{\theta_2},$$

$$\frac{\partial \ln f(x;\theta_1,\theta_2)}{\partial\theta_2} = -\frac{1}{2\theta_2} + \frac{(x-\theta_1)^2}{2\theta_2^2},$$

$$\frac{\partial^2 \ln f(x;\theta_1,\theta_2)}{\partial\theta_1^2} = \frac{-1}{\theta_2},$$

$$\frac{\partial^2 \ln f(x;\theta_1,\theta_2)}{\partial\theta_1\,\partial\theta_2} = \frac{-(x-\theta_1)}{\theta_2^2},$$

$$\frac{\partial^2 \ln f(x;\theta_1,\theta_2)}{\partial\theta_2^2} = \frac{1}{2\theta_2^2} - \frac{(x-\theta_1)^2}{\theta_2^3}.$$

If we take the expected value of these three second partial derivatives and multiply by $-n$, we obtain the information matrix of the sample, namely,

$$\mathbf{I}_n = \begin{bmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{bmatrix}.$$

Hence the approximate variance–covariance matrix of the maximum likelihood estimators $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$ is

$$\mathbf{I}_n^{-1} = \begin{bmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n} \end{bmatrix}.$$

It is not surprising that the covariance equals zero as we know that $\bar{X}$ and S^2 are independent. In addition, we know that

$$\text{var}\,(\bar{X}) = \frac{\theta_2}{n}$$

and

$$\text{var}\,(S^2) = \text{var}\left[\left(\frac{\theta_2}{n}\right)\left(\frac{nS^2}{\theta_2}\right)\right] = \frac{\theta_2^2}{n^2}\,\text{var}\left(\frac{nS^2}{\theta_2}\right) = \frac{2(n-1)\theta_2^2}{n^2}$$

since nS^2/θ_2 is $\chi^2(n-1)$. While $\text{var}\,(S^2) \neq 2\theta_2^2/n$, it is true that

$$\frac{2\theta_2^2}{n} \approx \frac{2(n-1)\theta_2^2}{n^2}$$

for large n.

EXERCISES

8.17. Let $X_1, X_2, \ldots, X_n$ be a random sample from each of the following distributions. In each case, find the m.l.e. $\hat{\theta}$, $\text{var}\,(\hat{\theta})$, $1/nI(\theta)$, where $I(\theta)$ is the Fisher information of a single observation X, and compare $\text{var}\,(\hat{\theta})$ and $1/nI(\theta)$.

(a) $b(1,\theta)$, $0 \leq \theta \leq 1$.

(b) $N(\theta, 1)$, $-\infty < \theta < \infty$.
(c) $N(0, \theta)$, $0 < \theta < \infty$.
(d) Gamma ($\alpha = 5, \beta = \theta$), $0 < \theta < \infty$.

8.18. Referring to Exercise 8.17 and using the fact that $\hat{\theta}$ has an approximate $N[\theta, 1/nI(\theta)]$, in each case construct an approximate 95 percent confidence interval for θ.

8.19. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution with unknown means θ_1 and θ_2 and with known variances and correlation coefficient, σ_1^2, σ_2^2, and ρ, respectively. Find the maximum likelihood estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ_1 and θ_2 and their approximate variance–covariance matrix. In this case, does the latter provide the exact variances and covariance?

8.20. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution with means equal to zero and variances θ_1 and θ_2, respectively, and known correlation coefficient ρ. Find the maximum likelihood estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ_1 and θ_2 and their approximate variance–covariance matrix.

8.4 Robust *M*-Estimation

In Example 1 of Section 8.3 we found the m.l.e. of the center θ of the Cauchy distribution with p.d.f.

$$f(x; \theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \qquad -\infty < x < \infty,$$

where $-\infty < \theta < \infty$. The logarithm of the likelihood function of a random sample $X_1, X_2, \ldots, X_n$ from this distribution is

$$\ln L(\theta) = -n \ln \pi - \sum_{i=1}^{n} \ln [1 + (x_i - \theta)^2].$$

To maximize, we differentiated $\ln L(\theta)$ to obtain

$$\frac{d \ln L(\theta)}{d\theta} = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2} = 0.$$

The solution of this equation cannot be found in closed form, but the equation can be solved by some iterative process. There, to do this, we used the weight function

$$w(x - \hat{\theta}_0) = \frac{2}{1 + (x - \hat{\theta}_0)^2},$$

where $\hat{\theta}_0$ is some preliminary estimator of θ, like the sample median. Note that values of x for which $|x - \hat{\theta}_0|$ is relatively large do not have much weight. That is, in finding the maximum likelihood estimator of θ, the outlying values are downweighted greatly.

The generalization of this special case is described as follows. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with a p.d.f. of the form $f(x - \theta)$, where θ is a location parameter such that $-\infty < \theta < \infty$. Thus

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i - \theta) = -\sum_{i=1}^{n} \rho(x_i - \theta),$$

where $\rho(x) = -\ln f(x)$, and

$$\frac{d \ln L(\theta)}{d\theta} = -\sum_{i=1}^{n} \frac{f'(x_i - \theta)}{f(x_i - \theta)} = \sum_{i=1}^{n} \Psi(x_i - \theta),$$

where $\rho'(x) = \Psi(x)$. For the Cauchy distribution, we have that these functions are

$$\rho(x) = \ln \pi + \ln (1 + x^2),$$

and

$$\Psi(x) = \frac{2x}{1 + x^2}.$$

In addition, we define a weight function as

$$w(x) = \frac{\Psi(x)}{x},$$

which equals $2/(1 + x^2)$ in the Cauchy case.

To appreciate how outlying observations are handled in estimating a center θ of different models progressing from a fairly light-tailed distribution like the normal to a very heavy-tailed distribution like the Cauchy, it is an easy exercise (Exercise 8.21) to show that standard normal distribution, with p.d.f. $\varphi(x)$, has

$$\rho(x) = \frac{1}{2} \ln 2\pi + \frac{x^2}{2}, \qquad \Psi(x) = x, \quad w(x) = 1.$$

That is, in estimating the center θ in $\varphi(x - \theta)$ each value of x has the weight 1 to yield the estimator $\hat{\theta} = \bar{X}$.

Also, the double exponential distribution, with p.d.f.

$$f(x) = \frac{1}{2} e^{-|x|}, \qquad -\infty < x < \infty,$$

has, provided that $x \neq 0$,

$$\rho(x) = \ln 2 + |x|, \qquad \Psi(x) = \text{sign}\,(x), \quad w(x) = \frac{\text{sign}\,(x)}{x} = \frac{1}{|x|}.$$

Here $\hat{\theta} = \text{median}\,(X_i)$ because in solving

$$\sum_{i=1}^{n} \Psi(x_i - \theta) = \sum_{i=1}^{n} \text{sign}\,(x_i - \theta) = 0$$

we need as many positive values of $x_i - \theta$ as negative values. The weights in the double exponential case are of the order $1/|x - \theta|$, while those in the Cauchy case are $2/[1 + (x - \theta)^2]$. That is, in estimating the center, outliers are downweighted more severely in a Cauchy situation, as the tails of the distribution are heavier than those of the double exponential distribution. On the other hand, extreme values from the double exponential distribution are downweighted more than those under normal assumptions in arriving at an estimate of the center θ.

Thus we suspect that the m.l.e. associated with one of these three distributions would not necessarily be a good estimator in another situation. This is true; for example, $\bar{X}$ is a very poor estimator of the median of a Cauchy distribution, as the variance of $\bar{X}$ does not even exist if the sample arises from a Cauchy distribution. Intuitively, $\bar{X}$ is not a good estimator with the Cauchy distribution, because the very small or very large values (outliers) that can arise from that distribution influence the mean $\bar{X}$ of the sample too much.

An estimator that is fairly good (small variance, say) for a wide variety of distributions (not necessarily the best for any one of them) is called a *robust estimator*. Also estimators associated with the solution of the equation

$$\sum_{i=1}^{n} \Psi(x_i - \theta) = 0$$

are frequently called *robust M-estimators* (denoted by $\hat{\theta}$) because they can be thought of as *m*aximum likelihood estimators. So in finding a robust M-estimator we must select a Ψ function which will provide an estimator that is good for each distribution in the collection under consideration. For certain theoretical reasons that we cannot explain at this level, Huber suggested a Ψ function that is a combination of

those associated with the normal and double exponential distributions,

$$\begin{aligned}\Psi(x) &= -k, \qquad x < -k \\ &= x, \qquad -k \le x \le k, \\ &= k, \qquad k < x,\end{aligned}$$

with weight $w(x) = 1$, $|x| \le k$, and $k/|x|$, provided that $k < |x|$. In Exercise 8.23 the reader is asked to find the p.d.f. $f(x)$ so that the M-estimator associated with this Ψ function is the m.l.e. of the location parameter θ in the p.d.f. $f(x - \theta)$.

With Huber's Ψ function, another problem arises. Note that if we double (for illustration) each $X_1, X_2, \ldots, X_n$, estimators such as $\bar{X}$ and median (X_i) also double. This is not at all true with the solution of the equation

$$\sum_{i=1}^{n} \Psi(x_i - \theta) = 0,$$

where the Ψ function is that of Huber. One way to avoid this difficulty is to solve another, but similar, equation instead,

$$\sum_{i=1}^{n} \Psi\left(\frac{x_i - \theta}{d}\right) = 0, \tag{1}$$

where d is a robust estimate of the scale. A popular d to use is

$$d = \frac{\text{median } |x_i - \text{median } (x_i)|}{0.6745}.$$

The divisor 0.6745 is inserted in the definition of d because then d is a consistent estimate of σ and thus is about equal to σ, if the sample arises from a normal distribution. That is, σ can be approximated by d under normal assumptions.

That scheme of selecting d also provides us with a clue for selecting k. For if the sample actually arises from a normal distribution, we would want most of the values $x_1, x_2, \ldots, x_n$ to satisfy the inequality

$$\left|\frac{x_i - \theta}{d}\right| \le k$$

because then

$$\Psi\left(\frac{x_i - \theta}{d}\right) = \frac{x_i - \theta}{d}.$$

That is, for illustration, if *all* the values satisfy this inequality, then Equation (1) becomes

$$\sum_{i=1}^{n} \Psi\left(\frac{x_i - \theta}{d}\right) = \sum_{i=1}^{n} \frac{x_i - \theta}{d} = 0.$$

This has the solution $\bar{x}$, which of course is most desirable with normal distributions. Since d approximates σ, popular values of k to use are 1.5 and 2.0, because with those selections most normal variables would satisfy the desired inequality.

Again an iterative process must usually be used to solve Equation (1). One such scheme, Newton's method, is described. Let $\hat{\theta}_0$ be a first estimate of θ, such as $\hat{\theta}_0 = \text{median}\,(x_i)$. Approximate the left-hand member of Equation (1) by the first two terms of Taylor's expansion about $\hat{\theta}_0$ to obtain

$$\sum_{i=1}^{n} \Psi\left(\frac{x_i - \hat{\theta}_0}{d}\right) + (\theta - \hat{\theta}_0) \sum_{i=1}^{n} \Psi'\left(\frac{x_i - \hat{\theta}_0}{d}\right)\left(-\frac{1}{d}\right) = 0,$$

approximately. The solution of this provides a second estimate of θ,

$$\hat{\theta}_1 = \hat{\theta}_0 + \frac{d \sum_{i=1}^{n} \Psi\left(\frac{x_i - \hat{\theta}_0}{d}\right)}{\sum_{i=1}^{n} \Psi'\left(\frac{x_i - \hat{\theta}_0}{d}\right)},$$

which is called the one-step M-estimate of θ. If we use $\hat{\theta}_1$ in place of $\hat{\theta}_0$, we obtain $\hat{\theta}_2$, the two-step M-estimate of θ. This process can continue to obtain any desired degree of accuracy. With Huber's Ψ function, the denominator of the second term,

$$\sum_{i=1}^{n} \Psi'\left(\frac{x_i - \hat{\theta}_0}{d}\right),$$

is particularly easy to compute because $\Psi'(x) = 1$, $-k \le x \le k$, and zero elsewhere. Thus that denominator simply counts the number of $x_1, x_2, \ldots, x_n$ such that $|x_i - \hat{\theta}_0|/d \le k$.

Say that the scale parameter σ is known (here σ is not necessarily the standard deviation for it does not exist for a distribution like the Cauchy). Two terms of Taylor's expansion of

$$\sum_{i=1}^{n} \Psi\left(\frac{X_i - \hat{\theta}}{\sigma}\right) = 0$$

about θ provides the approximation

$$\sum_{i=1}^{n} \Psi\left(\frac{X_i - \theta}{\sigma}\right) + (\hat{\theta} - \theta) \sum_{i=1}^{n} \Psi'\left(\frac{X_i - \theta}{\sigma}\right)\left(-\frac{1}{\sigma}\right) = 0.$$

This can be rewritten

$$\hat{\theta} - \theta = \frac{\sigma \sum \Psi\left(\frac{X_i - \theta}{\sigma}\right)}{\sum \Psi'\left(\frac{X_i - \theta}{\sigma}\right)}. \tag{2}$$

For the asymmetric Ψ functions that we have considered

$$E\left[\Psi\left(\frac{X - \theta}{\sigma}\right)\right] = 0,$$

provided that X has a symmetric distribution about θ. Clearly,

$$\operatorname{var}\left[\Psi\left(\frac{X - \theta}{\sigma}\right)\right] = E\left[\Psi^2\left(\frac{X - \theta}{\sigma}\right)\right].$$

Thus Equation (2) can be rewritten as

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma \sqrt{E\left[\Psi^2\left(\frac{X - \theta}{\sigma}\right)\right] \Big/ \left\{E\left[\Psi'\left(\frac{X - \theta}{\sigma}\right)\right]\right\}^2}}$$

$$= \frac{\sum \Psi\left(\frac{X_i - \theta}{\sigma}\right) \Big/ \sqrt{nE\left[\Psi^2\left(\frac{X - \theta}{\sigma}\right)\right]}}{\sum \Psi'\left(\frac{X_i - \theta}{\sigma}\right) \Big/ nE\left[\Psi'\left(\frac{X_i - \theta}{\sigma}\right)\right]}. \tag{3}$$

Clearly, by the central limit theorem, the numerator of the right-hand member of Equation (3) has a limiting standardized normal distribution, while the denominator converges in probability to 1. Thus the left-hand member has a limiting distribution that is $N(0, 1)$. In application we must approximate the denominator of the left-hand member. So we say that the robust M-estimator $\hat{\theta}$ has an approximate normal distribution with mean θ and variance

$$v = \frac{\left(\frac{d^2}{n}\right) \frac{1}{n} \sum_{i=1}^{n} \Psi^2\left(\frac{x_i - \hat{\theta}_k}{d}\right)}{\left\{\frac{1}{n} \sum_{i=1}^{n} \Psi'\left(\frac{x_i - \hat{\theta}_k}{d}\right)\right\}^2},$$

where $\hat{\theta}_k$ is the (last) k-step estimator of θ. Of course, $\hat{\theta}$ is approximated by $\hat{\theta}_k$; and an approximate 95 percent confidence interval for θ is given by $\hat{\theta}_k - 1.96\sqrt{v}$ to $\hat{\theta}_k + 1.96\sqrt{v}$.

EXERCISES

8.21. Verify that the functions $\rho(x)$, $\Psi(x)$, and $w(x)$ given in the text for the normal and double exponential distributions are correct.

8.22. Compute the one-step M-estimate $\hat{\theta}_1$ using Huber's Ψ with $k = 1.5$ if $n = 7$ and the seven observations are 2.1, 5.2, 2.3, 1.4, 2.2, 2.3, and 1.6. Here take $\hat{\theta}_0 = 2.2$, the median of the sample. Compare $\hat{\theta}_1$ with $\bar{x}$.

8.23. Let the p.d.f. $f(x)$ be such that the M-estimator associated with Huber's Ψ function is a maximum likelihood estimator of the location parameter in $f(x - \theta)$. Show that $f(x)$ is of the form $ce^{-\rho_1(x)}$, where $\rho_1(x) = x^2/2$, $|x| \leq k$ and $\rho_1(x) = k|x| - k^2/2$, $k < |x|$.

8.24. Plot the Ψ functions associated with the normal, double exponential, and Cauchy distributions in addition to that of Huber. Why is the M-estimator associated with the Ψ function of the Cauchy distribution called a *redescending* M-estimator?

8.25. Use the data in Exercise 8.22 to find the one-step redescending M-estimator $\hat{\theta}_1$ associated with $\Psi(x) = \sin(x/1.5)$, $|x| \leq 1.5\pi$, zero elsewhere. This was first proposed by D. F. Andrews. Compare this to $\bar{x}$ and the one-step M-estimator of Exercise 8.22. [It should be noted that there is no p.d.f. $f(x)$ that could be associated with this $\Psi(x)$ because $\Psi(x) = 0$ if $|x| > 1.5\pi$.]

ADDITIONAL EXERCISES

8.26. Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma distribution with $\alpha = 2$ and $\beta = 1/\theta$, $0 < \theta < \infty$.

(a) Find the m.l.e., $\hat{\theta}$, of θ. Is $\hat{\theta}$ unbiased?

(b) What is the approximating distribution of $\hat{\theta}$?

(c) If the prior distribution of the parameter is exponential with mean 2, determine the Bayes' estimator associated with a square-error loss function.

8.27. If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with p.d.f. $f(x; \theta) = 3\theta^3(x + \theta)^{-4}$, $0 < x < \infty$, zero elsewhere, where $0 < \theta$, show that $Y = 2\bar{X}$ is an unbiased estimator of θ and determine its efficiency.

8.28. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f. $f(x; \theta) = \dfrac{\theta}{(1 + x)^{\theta + 1}}$, $0 < x < \infty$, zero elsewhere, where $0 < \theta$.

(a) Find the m.l.e., $\hat{\theta}$, of θ and argue that it is a complete sufficient statistic for θ. Is $\hat{\theta}$ unbiased?
(b) If $\hat{\theta}$ is adjusted so that it is an unbiased estimator of θ, what is a lower bound for the variance of this unbiased estimator?

8.29. If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\theta, 1)$, find a lower bound for the variance of an estimator of $k(\theta) = \theta^2$. Determine an unbiased minimum variance estimator of θ^2 and then compute its efficiency.

8.30. Suppose that we want to estimate the middle, θ, of a symmetric distribution using a robust estimator because we believe that the tails of this distribution are much thicker than those of a normal distribution. A t-distribution with 3 degrees of freedom with center at θ (not at zero) is such a distribution, so we decide to use the m.l.e., $\hat{\theta}$, associated with that distribution as our robust estimator. Evaluate $\hat{\theta}$ for the five observations: 10.1, 20.7, 11.3, 12.5, 6.0. Here we assume that the spread parameter is equal to 1.

8.31. Consider the normal distribution $N(0, \theta)$. With a random sample $X_1, X_2, \ldots, X_n$ we want to estimate the standard deviation $\sqrt{\theta}$. Find the constant c so that $Y = c \sum_{i=1}^{n} |X_i|$ is an unbiased estimator of $\sqrt{\theta}$ and determine its efficiency.

CHAPTER 9

Theory of Statistical Tests

9.1 Certain Best Tests

In Chapter 6 we introduced many concepts associated with tests of statistical hypotheses. In this chapter we consider some methods of constructing good statistical tests, beginning with testing a simple hypothesis H_0 against a simple alternative hypothesis H_1. Thus, in all instances, the parameter space is a set that consists of exactly two points. Under this restriction, we shall do three things:

1. Define a best test for testing H_0 against H_1.
2. Prove a theorem that provides a method of determining a best test.
3. Give two examples.

Before we define a best test, one important observation should be made. Certainly, a test specifies a critical region; but it can also be said that a choice of a critical region defines a test. For instance, if one is given the critical region $C = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 \geq 1\}$, the test is determined: Three random variables X_1, X_2, X_3 are to be considered; if the observed values are x_1, x_2, x_3, accept H_0 if