Chapter 3

MEASURES OF PERFORMANCE, NOTIONS OF OPTIMALITY, AND OPTIMAL PROCEDURES

3.1 INTRODUCTION

Here we develop the theme of Section 1.3, which is how to appraise and select among decision procedures. In Sections 3.2 and 3.3 we show how the important Bayes and minimax criteria can in principle be implemented. However, actual implementation is limited. Our examples are primarily estimation of a real parameter. In Section 3.4, we study, in the context of estimation, the relation of the two major decision theoretic principles to the nondecision theoretic principle of maximum likelihood and the somewhat out of favor principle of unbiasedness. We also discuss other desiderata that strongly compete with decision theoretic optimality, in particular computational simplicity and robustness. We return to these themes in Chapter 6, after similarly discussing testing and confidence bounds, in Chapter 4 and developing in Chapters 5 and 6 the asymptotic tools needed to say something about the multiparameter case.

3.2 BAYES PROCEDURES

Recall from Section 1.3 that if we specify a parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, action space $\mathcal{A}$, loss function $l(\theta, a)$, then for data $X \sim P_\theta$ and any decision procedure δ randomized or not we can define its risk function, $R(\cdot, \delta) : \Theta \to R^+$ by

$$R(\theta, \delta) = E_\theta l(\theta, \delta(X)).$$

We think of $R(\cdot, \delta)$ as measuring a priori the performance of δ for this model. Strict comparison of δ_1 and δ_2 on the basis of the risks alone is not well defined unless $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all θ or vice versa. However, by introducing a Bayes prior density (say) π for θ comparison becomes unambiguous by considering the scalar Bayes risk,

$$r(\pi, \delta) \equiv ER(\boldsymbol{\theta}, \delta) = El(\boldsymbol{\theta}, \delta(X)), \tag{3.2.1}$$

where $(\boldsymbol{\theta}, X)$ is given the joint distribution specified by (1.2.3). Recall also that we can define

$$R(\pi) = \inf\{r(\pi, \delta) : \delta \in \mathcal{D}\} \tag{3.2.2}$$

the *minimum Bayes risk* of the problem, and that in Section 1.3 we showed how in an example, we could identify the *Bayes rules* δ^*_π such that

$$r(\pi, \delta^*_\pi) = R(\pi). \tag{3.2.3}$$

In this section we shall show systematically how to construct Bayes rules. This exercise is interesting and important even if we do not view π as reflecting an implicitly believed in prior distribution on $\boldsymbol{\theta}$. After all, if π is a density and $\Theta \subset R$

$$r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)d\theta \tag{3.2.4}$$

and π may express that we care more about the values of the risk in some rather than other regions of Θ. For testing problems the hypothesis is often treated as more important than the alternative. We may have vague prior notions such as "$|\theta| \geq 5$ is physically implausible" if, for instance, θ denotes mean height of people in meters. If π is then thought of as a weight function roughly reflecting our knowledge, it is plausible that δ^*_π if computable will behave reasonably even if our knowledge is only roughly right. Clearly, $\pi(\theta) \equiv c$ plays a special role ("equal weight") though (Problem 3.2.4) the parametrization plays a crucial role here. It is in fact clear that prior and loss function cannot be separated out clearly either. Thus, considering $l_1(\theta, a)$ and $\pi_1(\theta)$ is equivalent to considering $l_2(\theta, a) = \pi_1(\theta)l_1(\theta, a)$ and $\pi_2(\theta) \equiv 1$. Issues such as these and many others are taken up in the fundamental treatises on Bayesian statistics such as Jeffreys (1948) and Savage (1954) and are reviewed in the modern works of Berger (1985) and Bernardo and Smith (1994). We don't pursue them further except in Problem 3.2.5, and instead turn to construction of Bayes procedures.

We first consider the problem of estimating $q(\theta)$ with quadratic loss, $l(\theta, a) = (q(\theta) - a)^2$, using a nonrandomized decision rule δ. Suppose $\boldsymbol{\theta}$ is a random variable (or vector) with (prior) frequency function or density $\pi(\theta)$. Our problem is to find the function δ of $\mathbf{X}$ that minimizes $r(\pi, \delta) = E(q(\boldsymbol{\theta}) - \delta(\mathbf{X}))^2$. This is just the problem of finding the best mean squared prediction error (MSPE) predictor of $q(\theta)$ given $\mathbf{X}$ (see Remark 1.4.5). Using our results on MSPE prediction, we find that either $r(\pi, \delta) = \infty$ for all δ or the Bayes rule δ^* is given by

$$\delta^*(\mathbf{X}) = E[q(\boldsymbol{\theta}) \mid \mathbf{X}]. \tag{3.2.5}$$

This procedure is called the *Bayes estimate for squared error loss.*

In view of formula (1.2.8) for the posterior density and frequency functions, we can give the Bayes estimate a more explicit form. In the continuous case with θ real valued and prior density π,

$$\delta^*(\mathbf{x}) = \frac{\int_{-\infty}^{\infty} q(\theta)p(x \mid \theta)\pi(\theta)d\theta}{\int_{-\infty}^{\infty} p(x \mid \theta)\pi(\theta)d\theta}. \tag{3.2.6}$$

In the discrete case, as usual, we just need to replace the integrals by sums. Here is an example.

Example 3.2.1. *Bayes Estimates for the Mean of a Normal Distribution with a Normal Prior.* Suppose that we want to estimate the mean θ of a normal distribution with known variance σ^2 on the basis of a sample $X_1, \ldots, X_n$. If we choose the conjugate prior $\mathcal{N}(\eta_0, \tau^2)$ as in Example 1.6.12, we obtain the posterior distribution

$$\mathcal{N}\left(\eta_0\left(\frac{\sigma^2}{n\tau^2+\sigma^2}\right)+\bar{x}\left(\frac{n\tau^2}{n\tau^2+\sigma^2}\right),\ \frac{\sigma^2}{n}\left(1+\frac{\sigma^2}{n\tau^2}\right)^{-1}\right).$$

The Bayes estimate is just the mean of the posterior distribution

$$\delta^*(\mathbf{X}) = \eta_0\left[\frac{1/\tau^2}{n/\sigma^2+1/\tau^2}\right] + \bar{X}\left[\frac{n/\sigma^2}{n/\sigma^2+1/\tau^2}\right]. \tag{3.2.7}$$

Its Bayes risk (the MSPE of the predictor) is just

$$\begin{aligned} r(\pi,\delta^*) &= E(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid\mathbf{X}))^2 = E[E((\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid\mathbf{X}))^2\mid\mathbf{X})] \\ &= E\left[\frac{\sigma^2}{n}\Big/\left(1+\frac{\sigma^2}{n\tau^2}\right)\right] = \frac{1}{n/\sigma^2+1/\tau^2}. \end{aligned}$$

No finite choice of η_0 and τ^2 will lead to $\bar{X}$ as a Bayes estimate. But $\bar{X}$ is the limit of such estimates as prior knowledge becomes "vague" ($\tau \to \infty$ with η_0 fixed). In fact, $\bar{X}$ is the estimate that (3.2.6) yields, if we substitute the prior "density" $\pi(\theta) \equiv 1$ (Problem 3.2.1). Such priors with $\int \pi(\theta) = \infty$ or $\sum \pi(\theta) = \infty$ are called *improper*. The resulting Bayes procedures are also called improper.

Formula (3.2.7) reveals the Bayes estimate in the proper case to be a weighted average

$$w\eta_0 + (1-w)\bar{X}$$

of the estimate to be used when there are no observations, that is, η_0, and $\bar{X}$ with weights inversely proportional to the Bayes risks of these two estimates. Because the Bayes risk of $\bar{X}$, σ^2/n, tends to 0 as $n \to \infty$, the Bayes estimate corresponding to the prior density $\mathcal{N}(\eta_0, \tau^2)$ differs little from $\bar{X}$ for n large. In fact, $\bar{X}$ is approximately a Bayes estimate for any one of these prior distributions in the sense that $[r(\pi,\bar{X}) - r(\pi,\delta^*)]/r(\pi,\delta^*) \to 0$ as $n \to \infty$. For more on this, see Section 5.5. □

We now turn to the problem of finding Bayes rules for general action spaces $\mathcal{A}$ and loss functions l. To begin with we consider only nonrandomized rules. If we look at the proof of Theorem 1.4.1, we see that the key idea is to consider what we should do given $\mathbf{X} = \mathbf{x}$. Thus, $E(Y \mid \mathbf{X})$ is the best predictor because $E(Y \mid \mathbf{X} = \mathbf{x})$ minimizes the conditional MSPE $E((Y-a)^2 \mid \mathbf{X} = \mathbf{x})$ as a function of the action a. Applying the same idea in the general Bayes decision problem, we form the *posterior risk*

$$r(a \mid \mathbf{x}) = E(l(\boldsymbol{\theta}, a) \mid \mathbf{X} = \mathbf{x}).$$

This quantity $r(a \mid \mathbf{x})$ is what we expect to lose, if $\mathbf{X} = \mathbf{x}$ and we use action a. Intuitively, we should, for each $\mathbf{x}$, take that action $a = \delta^*(\mathbf{x})$ that makes $r(a \mid \mathbf{x})$ as small as possible. This action need not exist nor be unique if it does exist. However,

Proposition 3.2.1. *Suppose that there exists a function $\delta^*(\mathbf{x})$ such that*

$$r(\delta^*(\mathbf{x}) \mid \mathbf{x}) = \inf\{r(a \mid \mathbf{x}) : a \in \mathcal{A}\}. \tag{3.2.8}$$

Then δ^* is a Bayes rule.

Proof. As in the proof of Theorem 1.4.1, we obtain for any δ

$$r(\pi, \delta) = E[l(\boldsymbol{\theta}, \delta(\mathbf{X}))] = E[E(l(\boldsymbol{\theta}, \delta(\mathbf{X})) \mid \mathbf{X})]. \tag{3.2.9}$$

But, by (3.2.8),

$$E[l(\theta, \delta(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}] = r(\delta(\mathbf{x}) \mid \mathbf{x}) \geq r(\delta^*(\mathbf{x}) \mid \mathbf{x}) = E[l(\theta, \delta^*(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}].$$

Therefore,

$$E[l(\theta, \delta(\mathbf{X})) \mid \mathbf{X}] \geq E[l(\theta, \delta^*(\mathbf{X})) \mid \mathbf{X}],$$

and the result follows from (3.2.9). □

As a first illustration, consider the oil-drilling example (Example 1.3.5) with prior $\pi(\theta_1) = 0.2$, $\pi(\theta_2) = 0.8$. Suppose we observe $x = 0$. Then the posterior distribution of $\boldsymbol{\theta}$ is by (1.2.8)

$$\pi(\theta_1 \mid X = 0) = \frac{1}{9},\ \pi(\theta_2 \mid X = 0) = \frac{8}{9}.$$

Thus, the posterior risks of the actions a_1, a_2, and a_3 are

$$\begin{aligned} r(a_1 \mid 0) &= \frac{1}{9}l(\theta_1, a_1) + \frac{8}{9}l(\theta_2, a_1) = 10.67 \\ r(a_2 \mid 0) &= 2, \qquad\qquad r(a_3 \mid 0) = 5.89. \end{aligned}$$

Therefore, a_2 has the smallest posterior risk and, if δ^* is the Bayes rule,

$$\delta^*(0) = a_2.$$

Similarly,

$$r(a_1 \mid 1) = 8.35,\ r(a_2 \mid 1) = 3.74,\ r(a_3 \mid 1) = 5.70$$

and we conclude that

$$\delta^*(1) = a_2.$$

Therefore, $\delta^* = \delta_5$ as we found previously. The great advantage of our new approach is that it enables us to compute the Bayes procedure without undertaking the usually impossible calculation of the Bayes risks of all competing procedures.

More generally consider the following class of situations.

Example 3.2.2. *Bayes Procedures When Θ and $\mathcal{A}$ Are Finite.* Let $\Theta = \{\theta_0, \ldots, \theta_p\}$, $\mathcal{A} = \{a_0, \ldots, a_q\}$, let $\omega_{ij} \geq 0$ be given constants, and let the loss incurred when θ_i is true and action a_j is taken be given by

$$l(\theta_i, a_j) = \omega_{ij}.$$

Let $\pi(\theta)$ be a prior distribution assigning mass π_i to θ_i, so that $\pi_i \geq 0$, $i = 0, \ldots, p$, and $\sum_{i=0}^{p} \pi_i = 1$. Suppose, moreover, that $\mathbf{X}$ has density or frequency function $p(\mathbf{x} \mid \theta)$ for each θ. Then, by (1.2.8), the posterior probabilities are

$$P[\boldsymbol{\theta} = \theta_i \mid \mathbf{X} = \mathbf{x}] = \frac{\pi_i p(\mathbf{x} \mid \theta_i)}{\Sigma_j \pi_j p(\mathbf{x} \mid \theta_j)}$$

and, thus,

$$r(a_j \mid \mathbf{x}) = \frac{\Sigma_i \omega_{ij} \pi_i p(\mathbf{x} \mid \theta_i)}{\Sigma_i \pi_i p(\mathbf{x} \mid \theta_i)}. \tag{3.2.10}$$

The optimal action $\delta^*(\mathbf{x})$ has

$$r(\delta^*(\mathbf{x}) \mid \mathbf{x}) = \min_{0 \leq j \leq q} r(a_j \mid \mathbf{x}).$$

Here are two interesting specializations.

(a) *Classification:* Suppose that $p = q$, we identify a_j with θ_j, $j = 0, \ldots, p$, and let

$$\begin{aligned} \omega_{ij} &= 1, \quad i \neq j \\ \omega_{ii} &= 0. \end{aligned}$$

This can be thought of as the *classification problem* in which we have $p + 1$ known disjoint populations and a new individual $\mathbf{X}$ comes along who is to be classified in one of these categories. In this case,

$$r(\theta_i \mid \mathbf{x}) = P[\boldsymbol{\theta} \neq \theta_i \mid \mathbf{X} = \mathbf{x}]$$

and minimizing $r(\theta_i \mid \mathbf{x})$ is equivalent to the reasonable procedure of maximizing the posterior probability,

$$P[\boldsymbol{\theta} = \theta_i \mid \mathbf{X} = \mathbf{x}] = \frac{\pi_i p(\mathbf{x} \mid \theta_i)}{\Sigma_j \pi_j p(\mathbf{x} \mid \theta_j)}.$$

(b) *Testing:* Suppose $p = q = 1$, $\pi_0 = \pi$, $\pi_1 = 1 - \pi$, $0 < \pi < 1$, a_0 corresponds to deciding $\theta = \theta_0$ and a_1 to deciding $\theta = \theta_1$. This is a special case of the testing formulation of Section 1.3 with $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. The Bayes rule is then to

$$\begin{aligned} &\text{decide } \theta = \theta_1 \text{ if } (1 - \pi) p(\mathbf{x} \mid \theta_1) > \pi p(\mathbf{x} \mid \theta_0) \\ &\text{decide } \theta = \theta_0 \text{ if } (1 - \pi) p(\mathbf{x} \mid \theta_1) < \pi p(\mathbf{x} \mid \theta_0) \end{aligned}$$

and decide either a_0 or a_1 if equality occurs. See Sections 1.3 and 4.2 on the option of randomizing between a_0 and a_1 if equality occurs. As we let π vary between zero and one, we obtain what is called the class of *Neyman–Pearson tests*, which provides the solution to the problem of minimizing P (type II error) given P (type I error) $\leq \alpha$. This is treated further in Chapter 4. □

To complete our illustration of the utility of Proposition 3.2.1, we exhibit in "closed form" the Bayes procedure for an estimation problem when the loss is not quadratic.

Example 3.2.3. *Bayes Estimation of the Probability of Success in n Bernoulli Trials.* Suppose that we wish to estimate θ using $X_1, \ldots, X_n$, the indicators of n Bernoulli trials with probability of success θ. We shall consider the loss function l given by

$$l(\theta, a) = \frac{(\theta - a)^2}{\theta(1-\theta)},\ 0 < \theta < 1,\ a \text{ real}. \tag{3.2.11}$$

This close relative of quadratic loss gives more weight to parameter values close to zero and one. Thus, for θ close to zero, this $l(\theta, a)$ is close to the relative squared error $(\theta - a)^2/\theta$. It makes $\bar{X}$ have constant risk, a property we shall find important in the next section. The analysis can also be applied to other loss functions. See Problem 3.2.5.

By sufficiency we need only consider the number of successes, S. Suppose now that we have a prior distribution. Then, if all terms on the right-hand side are finite,

$$\begin{aligned} r(a \mid k) &= E\left\{ \frac{(\boldsymbol{\theta} - a)^2}{\boldsymbol{\theta}(1-\boldsymbol{\theta})} \middle| S = k \right\} = E\left\{ \frac{\boldsymbol{\theta}}{(1-\boldsymbol{\theta})} \middle| S = k \right\} \\ &- 2aE\left\{ \frac{1}{(1-\boldsymbol{\theta})} \middle| S = k \right\} + a^2 E\left\{ \frac{1}{\boldsymbol{\theta}(1-\boldsymbol{\theta})} \middle| S = k \right\}. \end{aligned} \tag{3.2.12}$$

Minimizing this parabola in a, we find our Bayes procedure is given by

$$\delta^*(k) = \frac{E(1/(1-\boldsymbol{\theta}) \mid S = k)}{E(1/\boldsymbol{\theta}(1-\boldsymbol{\theta}) \mid S = k)} \tag{3.2.13}$$

provided the denominator is not zero. For convenience let us now take as prior density the density $b_{r,s}(\theta)$ of the beta distribution $\beta(r, s)$. In Example 1.2.1 we showed that this leads to a $\beta(k + r, n + s - k)$ posterior distribution for $\boldsymbol{\theta}$ if $S = k$. If $1 \leq k \leq n - 1$ and $n \geq 2$, then all quantities in (3.2.12) and (3.2.13) are finite, and

$$\begin{aligned} \delta^*(k) &= \frac{\int_0^1 (1/(1-\theta)) b_{k+r,n-k+s}(\theta)\, d\theta}{\int_0^1 (1/\theta(1-\theta)) b_{k+r,n-k+s}(\theta)\, d\theta} \\ &= \frac{B(k + r, n - k + s - 1)}{B(k + r - 1, n - k + s - 1)} = \frac{k + r - 1}{n + s + r - 2}, \end{aligned} \tag{3.2.14}$$

where we are using the notation B.2.11 of Appendix B. If $k = 0$, it is easy to see that $a = 0$ is the only a that makes $r(a \mid k) < \infty$. Thus, $\delta^*(0) = 0$. Similarly, we get $\delta^*(n) = 1$. If we assume a uniform prior density, $(r = s = 1)$, we see that the Bayes procedure is the usual estimate, $\bar{X}$. This is *not* the case for quadratic loss (see Problem 3.2.2). □

"Real" computation of Bayes procedures

The closed forms of (3.2.6) and (3.2.10) make the computation of (3.2.8) appear straightforward. Unfortunately, this is far from true in general. Suppose, as is typically the case, that $\theta = (\theta_1, \ldots, \theta_p)$ has a hierarchically defined prior density,

$$\pi(\theta_1, \theta_2, \ldots, \theta_p) = \pi_1(\theta_1)\pi_2(\theta_2 \mid \theta_1) \ldots \pi_p(\theta_p \mid \theta_{p-1}). \tag{3.2.15}$$

Here is an example.

Example 3.2.4. The *random effects model* we shall study in Volume II has

$$X_{ij} = \mu + \Delta_i + \epsilon_{ij},\ 1 \le i \le I,\ 1 \le j \le J \tag{3.2.16}$$

where the ϵ_{ij} are i.i.d. $\mathcal{N}(0, \sigma_e^2)$ and μ and the vector $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_I)$ are independent of $\{\epsilon_{ij} : 1 \le i \le I,\ 1 \le j \le J\}$ with $\Delta_1, \ldots, \Delta_I$ i.i.d. $\mathcal{N}(0, \sigma_\Delta^2)$, $1 \le j \le J$, $\mu \sim \mathcal{N}(\mu_0, \sigma_\mu^2)$. Here the X_{ij} can be thought of as measurements on individual i and Δ_i is an "individual" effect. If we now put a prior distribution on $(\mu, \sigma_e^2, \sigma_\Delta^2)$ making them independent, we have a Bayesian model in the usual form. But it is more fruitful to think of this model as parametrized by $\theta = (\mu, \sigma_e^2, \sigma_\Delta^2, \Delta_1, \ldots, \Delta_I)$ with the $X_{ij} \mid \theta$ independent $\mathcal{N}(\mu + \Delta_i, \sigma_e^2)$. Then $p(\mathbf{x} \mid \theta) = \prod_{i,j} \varphi_{\sigma_e}(x_{ij} - \mu - \Delta_i)$ and

$$\pi(\theta) = \pi_1(\mu)\pi_2(\sigma_e^2)\pi_3(\sigma_\Delta^2) \prod_{i=1}^{I} \varphi_{\sigma_\Delta}(\Delta_i) \tag{3.2.17}$$

where φ_σ denotes the $\mathcal{N}(0, \sigma^2)$ density.

In such a context a loss function frequently will single out some single coordinate θ_s (e.g., Δ_1 in 3.2.17) and to compute $r(a \mid \mathbf{x})$ we will need the posterior distribution of $\Delta_1 \mid \mathbf{x}$. But this is obtainable from the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{X} = \mathbf{x}$ only by integrating out θ_j, $j \ne s$, and if p is large this is intractable. In recent years so-called Markov Chain Monte Carlo (MCMC) techniques have made this problem more tractable and the use of Bayesian methods has spread. We return to the topic in Volume II. □

Linear Bayes estimates

When the problem of computing $r(\pi, \delta)$ and δ_π is daunting, an alternative is to consider a class $\widetilde{\mathcal{D}}$ of procedures for which $r(\pi, \delta)$ is easy to compute and then to look for $\widetilde{\delta}_\pi \in \widetilde{\mathcal{D}}$ that minimizes $r(\pi, \delta)$ for $\delta \in \widetilde{\mathcal{D}}$. An example is *linear Bayes* estimates where, in the case of squared error loss $[q(\theta) - a]^2$, the problem is equivalent to minimizing the mean squared prediction error among functions of the form $b_0 + \sum_{j=1}^d b_j X_j$. If in (1.4.14) we identify $q(\boldsymbol{\theta})$ with Y and $\mathbf{X}$ with $\mathbf{Z}$, the solution is

$$\widetilde{\delta}(\mathbf{X}) = Eq(\boldsymbol{\theta}) + [\mathbf{X} - E(\mathbf{X})]^T \boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is as defined in Section 1.4. For example, if in the model (3.2.16), (3.2.17) we set $q(\boldsymbol{\theta}) = \Delta_1$, we can find the linear Bayes estimate of Δ_1 by using 1.4.6 and Problem 1.4.21. We find from (1.4.14) that the best linear Bayes estimator of Δ_1 is

$$\delta_L(\mathbf{X}) = E(\Delta_1) + (\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\beta} \tag{3.2.18}$$

where $E(\Delta_1) = 0$, $\mathbf{X} = (X_{11}, \ldots, X_{1J})^T$, $\boldsymbol{\mu} = E(\mathbf{X})$ and $\boldsymbol{\beta} = \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\Delta_1}$. For the given model

$$E(X_{1j}) = EE(X_{1j} \mid \theta) = E(\mu + \Delta_1) = E(\mu)$$

$$\text{Var}(X_{1j}) = E\,\text{Var}(X_{1j} \mid \theta) + \text{Var}\,E(X_{1j} \mid \theta) = E(\sigma_\epsilon^2) + \sigma_\mu^2 + \sigma_{\Delta_1}^2$$

$$\begin{aligned}\text{Cov}(X_{1j}, X_{1k}) &= E\,\text{Cov}(X_{1j}, X_{1k} \mid \theta) + \text{Cov}(E(X_{1j} \mid \theta), E(X_{1k} \mid \theta)) \\ &= 0 + \text{Cov}(\mu + \Delta_1, \mu + \Delta_1) = \sigma_\mu^2 + \sigma_{\Delta_1}^2\end{aligned}$$

$$\text{Cov}(\Delta_1, X_{1j}) = E\,\text{Cov}(X_{1j}, \Delta_1 \mid \theta) + \text{Cov}(E(X_{1j} \mid \theta), E(\Delta_1 \mid \theta)) = 0 + \sigma_{\Delta_1}^2 = \sigma_{\Delta_1}^2.$$

From these calculations we find $\boldsymbol{\beta}$ and $\delta_L(\mathbf{X})$. We leave the details to Problem 3.2.10. Linear Bayes procedures are useful in actuarial science, for example, Bühlmann (1970) and Norberg (1986).

Bayes estimation, maximum likelihood, and equivariance

As we have noted earlier, the maximum likelihood estimate can be thought of as the mode of the Bayes posterior density when the prior density is (the usually improper) prior $\pi(\theta) \equiv c$. When modes and means coincide for the improper prior (as in the Gaussian case), the MLE is an improper Bayes estimate. In general, computing means is harder than modes and that again accounts in part for the popularity of maximum likelihood.

An important property of the MLE is equivariance: An estimating method M producing the estimate $\widehat{\theta}_M$ is said to be *equivariant* with respect to reparametrization if for every one-to-one function h from Θ to $\Omega = h(\Theta)$, the estimate of $\omega \equiv h(\theta)$ is $\widehat{\omega}_M = h(\widehat{\theta}_M)$; that is, $\widehat{(h(\theta))}_M = h(\widehat{\theta}_M)$. In Problem 2.2.16 we show that the MLE procedure is equivariant. If we consider squared error loss, then the Bayes procedure $\widehat{\theta}_B = E(\boldsymbol{\theta} \mid X)$ is not equivariant for nonlinear transformations because

$$E(h(\boldsymbol{\theta}) \mid X) \neq h(E(\boldsymbol{\theta} \mid X))$$

for nonlinear h (e.g., Problem 3.2.3).

The source of the lack of equivariance of the Bayes risk and procedure for squared error loss is evident from (3.2.9): In the discrete case the conditional Bayes risk is

$$r_\Theta(a \mid x) = \sum_{\theta \in \Theta} [\theta - a]^2 \pi(\theta \mid x). \tag{3.2.19}$$

If we set $\boldsymbol{\omega} = h(\boldsymbol{\theta})$ for h one-to-one onto $\Omega = h(\Theta)$, then $\boldsymbol{\omega}$ has prior $\lambda(\omega) \equiv \pi(h^{-1}(\omega))$ and in the ω parametrization, the posterior Bayes risk is

$$\begin{aligned}r_\Omega(a \mid x) &= \sum_{\omega \in \Omega} [\omega - a]^2 \lambda(\omega \mid x) \\ &= \sum_{\theta \in \Theta} [h(\theta) - a]^2 \pi(\theta \mid x).\end{aligned} \tag{3.2.20}$$

Thus, the Bayes procedure for squared error loss is not equivariant because squared error loss is not equivariant and, thus, $r_\Omega(a \mid x) \neq r_\Theta(h^{-1}(a) \mid x)$.

Loss functions of the form $l(\theta, a) = Q(P_\theta, P_a)$ are necessarily equivariant. The Kullback–Leibler divergence $K(\theta, a)$, $\theta, a \in \Theta$, is an example of such a loss function. It satisfies $K_\Omega(\omega, a) = K_\Theta(\theta, h^{-1}(a))$, thus, with this loss function,

$$r_\Omega(a \mid \mathbf{x}) = r_\Theta(h^{-1}(a) \mid \mathbf{x}).$$

See Problem 2.2.38. In the discrete case using K means that the importance of a loss is measured in probability units, with a similar interpretation in the continuous case (see (A.7.10)). In the $\mathcal{N}(\theta, \sigma_0^2)$ case the KL (Kullback–Leibler) loss $K(\theta, a)$ is $\sigma^{-2}\frac{1}{2}n(a-\theta)^2$ (Problem 2.2.37), that is, equivalent to squared error loss. In canonical exponential families

$$K(\boldsymbol{\eta}, \mathbf{a}) = \sum_{j=1}^{k} [\eta_j - a_j]\dot{A}(\boldsymbol{\eta}) + A(\boldsymbol{\eta}) - A(\mathbf{a}).$$

Moreover, if we can find the KL loss Bayes estimate $\widehat{\boldsymbol{\eta}}_{BKL}$ of the canonical parameter $\boldsymbol{\eta}$ and if $\boldsymbol{\eta} \equiv \mathbf{c}(\boldsymbol{\theta}) : \Theta \to \mathcal{E}$ is one-to-one, then the KL loss Bayes estimate of $\boldsymbol{\theta}$ in the general exponential family is $\widehat{\boldsymbol{\theta}}_{BKL} = c^{-1}(\widehat{\boldsymbol{\eta}}_{BKL})$.

For instance, in Example 3.2.1 where μ is the mean of a normal distribution and the prior is normal, we found the squared error Bayes estimate $\widehat{\mu}_B = w\eta_0 + (1-w)\bar{X}$, where η_0 is the prior mean and w is a weight. Because the KL loss is equivalent to squared error for the canonical parameter μ, then if $\omega = h(\mu)$, $\widehat{\omega}_{BKL} = h(\widehat{\mu}_{BKL})$, where $\widehat{\mu}_{BKL} = w\eta_0 + (1-w)\bar{X}$.

Bayes procedures based on the Kullback–Leibler divergence loss function are important for their applications to model selection and their connection to "minimum description (message) length" procedures. See Rissanen (1987) and Wallace and Freeman (1987). More recent reviews are Shibata (1997), Dowe, Baxter, Oliver, and Wallace (1998), and Hansen and Yu (2001). We will return to this in Volume II.

Bayes methods and doing reasonable things

There is a school of Bayesian statisticians (Berger, 1985; DeGroot, 1969; Lindley, 1965; Savage, 1954) who argue on normative grounds that a decision theoretic framework and rational behavior force individuals to use only Bayes procedures appropriate to their personal prior π. This is not a view we espouse because we view a model as an imperfect approximation to imperfect knowledge. However, given that we view a model and loss structure as an adequate approximation, it is good to know that generating procedures on the basis of Bayes priors viewed as weighting functions is a reasonable thing to do. This is the conclusion of the discussion at the end of Section 1.3. It may be shown quite generally as we consider all possible priors that the class $\mathcal{D}_0$ of Bayes procedures and their limits is *complete* in the sense that for any $\delta \in \mathcal{D}$ there is a $\delta_0 \in \mathcal{D}_0$ such that $R(\theta, \delta_0) \leq R(\theta, \delta)$ for all θ.

Summary. We show how Bayes procedures can be obtained for certain problems by computing posterior risk. In particular, we present Bayes procedures for the important cases of classification and testing statistical hypotheses. We also show that for more complex problems, the computation of Bayes procedures require sophisticated statistical numerical techniques or approximations obtained by restricting the class of procedures.

3.3 MINIMAX PROCEDURES

In Section 1.3 on the decision theoretic framework we introduced minimax procedures as ones corresponding to a worst-case analysis; the true θ is one that is as "hard" as possible. That is, δ_1 is better than δ_2 from a minimax point of view if $\sup_\theta R(\theta, \delta_1) < \sup_\theta R(\theta, \delta_2)$ and δ^* is said to be *minimax* if

$$\sup_\theta R(\theta, \delta^*) = \inf_\delta \sup_\theta R(\theta, \delta).$$

Here θ and δ are taken to range over Θ and $\mathcal{D} = \{$all possible decision procedures (not randomized)$\}$ while $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. It is fruitful to consider proper subclasses of $\mathcal{D}$ and subsets of $\mathcal{P}$, but we postpone this discussion.

The nature of this criterion and its relation to Bayesian optimality is clarified by considering a so-called zero sum game played by two players N (Nature) and S (the statistician). The statistician has at his or her disposal the set $\mathcal{D}$ or the set $\overline{\mathcal{D}}$ of all randomized decision procedures whereas Nature has at her disposal all prior distributions π on Θ. For the basic game, S picks δ without N's knowledge, N picks π without S's knowledge and then all is revealed and S pays N

$$r(\pi, \delta) = \int R(\theta, \delta) d\pi(\theta)$$

where the notation $\int R(\theta, \delta) d\pi(\theta)$ stands for $\int R(\theta, \delta)\pi(\theta)d\theta$ in the continuous case and $\sum R(\theta_j, \delta)\pi(\theta_j)$ in the discrete case.

S tries to minimize his or her loss, N to maximize her gain. For simplicity, we assume in the general discussion that follows that all sup's and inf's are assumed. There are two related partial information games that are important.

I: N is told the choice δ of S before picking π and S knows the rules of the game. Then N naturally picks π_δ such that

$$r(\pi_\delta, \delta) = \sup_\pi r(\pi, \delta), \tag{3.3.1}$$

that is, π_δ is *least favorable against* δ. Knowing the rules of the game S naturally picks δ^* such that

$$r(\pi_{\delta^*}, \delta^*) = \sup_\pi r(\pi, \delta^*) = \inf_\delta \sup_\pi r(\pi, \delta). \tag{3.3.2}$$

We claim that δ^* is minimax. To see this we note first that,

$$r(\pi, \delta) = \int R(\theta, \delta) d\pi(\theta) \leq \sup_\theta R(\theta, \delta)$$

for all π, δ. On the other hand, if $R(\theta_\delta, \delta) = \sup_\theta R(\theta, \delta)$, then if π_δ is point mass at θ_δ, $r(\pi_\delta, \delta) = R(\theta_\delta, \delta)$ and we conclude that

$$\sup_\pi r(\pi, \delta) = \sup_\theta R(\theta, \delta) \tag{3.3.3}$$

and our claim follows.

II: S is told the choice π of N before picking δ and N knows the rules of the game. Then S naturally picks δ_π such that

$$r(\pi, \delta_\pi) = \inf_\delta r(\pi, \delta).$$

That is, δ_π is a Bayes procedure for π. Then N should pick π^* such that

$$r(\pi^*, \delta_{\pi^*}) = \sup_\pi r(\pi, \delta_\pi) = \sup_\pi \inf_\delta r(\pi, \delta). \tag{3.3.4}$$

For obvious reasons, π^* is called a *least favorable* (to S) *prior* distribution. As we shall see by example, although the right-hand sides of (3.3.2) and (3.3.4) are always defined, least favorable priors and/or minimax procedures may not exist and, if they exist, may not be unique.

The key link between the search for minimax procedures in the basic game and games I and II is the von Neumann minimax theorem of game theory, which we state in our language.

Theorem 3.3.1. *(von Neumann). If both Θ and $\mathcal{D}$ are finite, then:*

$$(a) \qquad \underline{v} \equiv \sup_\pi \inf_\delta r(\pi, \delta),\ \bar{v} \equiv \inf_\delta \sup_\pi r(\pi, \delta)$$

are both assumed by (say) π^* (least favorable), δ^* minimax, respectively. Further,

$$\underline{v} = r(\pi^*, \delta^*) = \bar{v} \tag{3.3.5}$$

and, hence, $\delta^* = \delta_{\pi^*}$, $\pi^* = \pi_{\delta^*}$.

$\underline{v}$ and $\bar{v}$ are called the *lower* and *upper values* of the basic game. When $\underline{v} = \bar{v} = v$ (say), v is called the *value* of the game.

Remark 3.3.1. Note (Problem 3.3.3) that von Neumann's theorem applies to classification and testing when $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ (Example 3.2.2) but is too restrictive in its assumption for the great majority of inference problems. A generalization due to Wald and Karlin—see Karlin (1959)—states that the conclusions of the theorem remain valid if Θ and $\mathcal{D}$ are compact subsets of Euclidean spaces. There are more far-reaching generalizations but, as we shall see later, without some form of compactness of Θ and/or $\mathcal{D}$, although equality of $\underline{v}$ and $\bar{v}$ holds quite generally, existence of least favorable priors and/or minimax procedures may fail.

The main practical import of minimax theorems is, in fact, contained in a converse and its extension that we now give. Remarkably these hold without essentially any restrictions on Θ and D and are easy to prove.

Proposition 3.3.1. *Suppose δ^{**}, π^{**} can be found such that*

$$\delta^{**} = \delta_{\pi^{**}},\ \pi^{**} = \pi_{\delta^{**}} \tag{3.3.6}$$

that is, δ^{**} is Bayes against π^{**} and π^{**} is least favorable against δ^{**}. Then $\underline{v} = \bar{v} = r(\pi^{**}, \delta^{**})$. That is, π^{**} is least favorable and δ^{**} is minimax.

To utilize this result we need a characterization of π_δ. This is given by

Proposition 3.3.2. *π_δ is least favorable against δ iff*

$$\pi_\delta\{\theta : R(\theta, \delta) = \sup_{\theta'} R(\theta', \delta)\} = 1. \tag{3.3.7}$$

That is, π_δ assigns probability only to points θ at which the function $R(\cdot, \delta)$ is maximal.

Thus, combining Propositions 3.3.1 and 3.3.2 we have a simple criterion, "A Bayes rule with constant risk is minimax."

Note that π_δ may not be unique. In particular, if $R(\theta, \delta) \equiv$ constant, the rule has constant risk, then all π are least favorable.

We now prove Propositions 3.3.1 and 3.3.2.

Proof of Proposition 3.3.1. Note first that we always have

$$\underline{v} \leq \bar{v} \tag{3.3.8}$$

because, trivially,

$$\inf_\delta r(\pi, \delta) \leq r(\pi, \delta') \tag{3.3.9}$$

for all π, δ'. Hence,

$$\underline{v} = \sup_\pi \inf_\delta r(\pi, \delta) \leq \sup_\pi r(\pi, \delta') \tag{3.3.10}$$

for all δ' and $\underline{v} \leq \inf_{\delta'} \sup_\pi r(\pi, \delta') = \bar{v}$. On the other hand, by hypothesis,

$$\underline{v} \geq \inf_\delta r(\pi^{**}, \delta) = r(\pi^{**}, \delta^{**}) = \sup_\pi r(\pi, \delta^{**}) \geq \bar{v}. \tag{3.3.11}$$

Combining (3.3.8) and (3.3.11) we conclude that

$$\underline{v} = \inf_\delta r(\pi^{**}, \delta) = r(\pi^{**}, \delta^{**}) = \sup_\pi r(\pi, \delta^{**}) = \bar{v} \tag{3.3.12}$$

as advertised. □

Proof of Proposition 3.3.2. π is least favorable for δ iff

$$E_\pi R(\boldsymbol{\theta}, \delta) = \int R(\theta, \delta) d\pi(\theta) = \sup_\pi r(\pi, \delta). \tag{3.3.13}$$

But by (3.3.3),

$$\sup_\pi r(\pi, \delta) = \sup_\theta R(\theta, \delta). \tag{3.3.14}$$

Because $E_\pi R(\boldsymbol{\theta}, \delta) = \sup_\theta R(\theta, \delta)$, (3.3.13) is possible iff (3.3.7) holds. □

Putting the two propositions together we have the following.

Theorem 3.3.2. *Suppose δ^* has $\sup_\theta R(\theta, \delta^*) = r < \infty$. If there exists a prior π^* such that δ^* is Bayes for π^* and $\pi^*\{\theta : R(\theta, \delta^*) = r\} = 1$, then δ^* is minimax.*

Example 3.3.1. *Minimax Estimation in the Binomial Case.* Suppose S has a $\mathcal{B}(n, \theta)$ distribution and $\bar{X} = S/n$, as in Example 3.2.3. Let $l(\theta, a) = (\theta - a)^2/\theta(1-\theta)$, $0 < \theta < 1$. For this loss function,

$$R(\theta, \bar{X}) = \frac{E(\bar{X} - \theta)^2}{\theta(1-\theta)} = \frac{\theta(1-\theta)}{n\theta(1-\theta)} = \frac{1}{n},$$

and $\bar{X}$ does have constant risk. Moreover, we have seen in Example 3.2.3 that $\bar{X}$ is Bayes, when $\boldsymbol{\theta}$ is $\mathcal{U}(0, 1)$. By Theorem 3.3.2 we conclude that $\bar{X}$ is minimax and, by Proposition 3.3.2, the uniform distribution least favorable.

For the usual quadratic loss neither of these assertions holds. The minimax estimate is

$$\delta^*(S) = \frac{S + \frac{1}{2}\sqrt{n}}{n + \sqrt{n}} = \frac{\sqrt{n}}{\sqrt{n}+1}\bar{X} + \frac{1}{\sqrt{n}+1} \cdot \frac{1}{2}.$$

This estimate does have constant risk and is Bayes against a $\beta(\sqrt{n}/2, \sqrt{n}/2)$ prior (Problem 3.3.4). This is an example of a situation in which the minimax principle leads us to an unsatisfactory estimate. For quadratic loss, the limit as $n \to \infty$ of the ratio of the risks of δ^* and $\bar{X}$ is > 1 for every $\theta \neq \frac{1}{2}$. At $\theta = \frac{1}{2}$ the ratio tends to 1. Details are left to Problem 3.3.4. □

Example 3.3.2. *Minimax Testing. Satellite Communications.* A test to see whether a communications satellite is in working order is run as follows. A very strong signal is beamed from Earth. The satellite responds by sending a signal of intensity $v > 0$ for n seconds or, if it is not working, does not answer. Because of the general "noise" level in space the signals received on Earth vary randomly whether the satellite is sending or not. The mean voltage per second of the signal for each of the n seconds is recorded. Denote the mean voltage of the signal received through the ith second less expected mean voltage due to noise by X_i. We assume that the X_i are independently and identically distributed as $\mathcal{N}(\mu, \sigma^2)$ where $\mu = v$, if the satellite functions, and 0 otherwise. The variance σ^2 of the "noise" is assumed known. Our problem is to decide whether "$\mu = 0$" or "$\mu = v$." We view this as a decision problem with $0 - 1$ loss. If the number of transmissions is fixed, the minimax rule minimizes the maximum probability of error (see (1.3.6)). What is this risk?

A natural first step is to use the characterization of Bayes tests given in the preceding section. If we assign probability π to 0 and $1 - \pi$ to v, use $0 - 1$ loss, and set $L(\mathbf{x}, 0, v) = p(\mathbf{x} \mid v)/p(\mathbf{x} \mid 0)$, then the Bayes test decides $\mu = v$ if

$$L(\mathbf{x}, 0, v) = \exp\left\{\frac{v}{\sigma^2}\Sigma x_i - \frac{nv^2}{2\sigma^2}\right\} \geq \frac{\pi}{1-\pi}$$

and decides $\mu = 0$ if

$$L(\mathbf{x}, 0, v) < \frac{\pi}{1-\pi}.$$

This test is equivalent to deciding $\mu = v$ (Problem 3.3.1) if, and only if,

$$T = \frac{1}{\sigma\sqrt{n}}\Sigma x_i \geq t,$$

where,

$$t = \frac{\sigma}{v\sqrt{n}}\left[\log\frac{\pi}{1-\pi} + \frac{nv^2}{2\sigma^2}\right].$$

If we call this test δ_π,

$$\begin{aligned} R(0,\delta_\pi) &= 1 - \Phi(t) = \Phi(-t) \\ R(v,\delta_\pi) &= \Phi\left(t - \tfrac{v\sqrt{n}}{\sigma}\right). \end{aligned}$$

To get a minimax test we must have $R(0,\delta_\pi) = R(v,\delta_\pi)$, which is equivalent to

$$-t = t - \frac{v\sqrt{n}}{\sigma}$$

or

$$t = \frac{v\sqrt{n}}{2\sigma}.$$

Because this value of t corresponds to $\pi = \frac{1}{2}$, the intuitive test, which decides $\mu = \nu$ if and only if $T \geq \frac{1}{2}[E_0(T) + E_v(T)]$, is indeed minimax. □

If Θ is not bounded, minimax rules are often not Bayes rules but instead can be obtained as limits of Bayes rules. To deal with such situations we need an extension of Theorem 3.3.2.

Theorem 3.3.3. *Let δ^* be a rule such that $\sup_\theta R(\theta,\delta^*) = r < \infty$, and let $\{\pi_k\}$ denote a sequence of prior distributions. Let $r_k = \inf_\delta r(\pi_k,\delta)$, where $r(\pi_k,\delta)$ denotes the Bayes risk wrt π_k. If*

$$r_k \to r \text{ as } k \to \infty, \tag{3.3.15}$$

then δ^* is minimax.

Proof. By assumption

$$\sup_\theta R(\theta,\delta^*) = r_k + o(1)$$

where $o(1) \to 0$ as $k \to \infty$. For any competitor δ

$$\sup_\theta R(\theta,\delta) \geq E_{\pi_k}(R(\boldsymbol{\theta},\delta)) \geq r_k = \sup_\theta R(\theta,\delta^*) - o(1). \tag{3.3.16}$$

If we let $k \to \infty$ the left-hand side of (3.3.16) is unchanged, whereas the right tends to $\sup_\theta R(\theta,\delta^*)$. □

To apply this result, we need to find a sequence of priors π_k such that $\inf_\delta r(\pi_k, \delta) \to \sup_\theta R(\theta, \delta^*)$. Here are two examples.

Example 3.3.3. *Normal Mean.* We now show that $\bar{X}$ is minimax in Example 3.2.1. Identify π_k with the $\mathcal{N}(\eta_0, \tau^2)$ prior where $k = \tau^2$. We know that $R(\theta, \bar{X}) = \sigma^2/n$, whereas the Bayes risk of the Bayes rule of Example 3.2.1 is

$$\inf_\delta r(\pi_k, \delta) = \frac{\tau^2}{(\sigma^2/n) + \tau^2}\frac{\sigma^2}{n} = \frac{\sigma^2}{n} - \frac{1}{(\sigma^2/n) + \tau^2}\frac{\sigma^2}{n}.$$

Because $(\sigma^2/n)/((\sigma^2/n)+\tau^2) \to 0$ as $\tau^2 \to \infty$, we can conclude that $\bar{X}$ is minimax. □

Example 3.3.4. *Minimax Estimation in a Nonparametric Setting (after Lehmann).* Suppose $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}$

$$\mathcal{F} = \{F : \mathrm{Var}_F(X_1) \leq M\}.$$

Then $\bar{X}$ is minimax for estimating $\theta(F) \equiv E_F(X_1)$ with quadratic loss. This can be viewed as an extension of Example 3.3.3. Let π_k be a prior distribution on $\mathcal{F}$ constructed as follows:[(1)]

(i) $\pi_k\{F : \mathrm{Var}_F(X_1) \neq M\} = 0$.

(ii) $\pi_k\{F : F \neq \mathcal{N}(\mu, M) \text{ for some } \mu\} = 0$.

(iii) F is chosen by first choosing $\mu = \theta(F)$ from a $\mathcal{N}(0, k)$ distribution and then taking $F = \mathcal{N}(\theta(F), M)$.

The Bayes risk is now the same as in Example 3.3.3 with $\sigma^2 = M$. Because, evidently,

$$\max_{\mathcal{F}} R(F, \bar{X}) = \max_{\mathcal{F}} \frac{\mathrm{Var}_F(X_1)}{n} = \frac{M}{n},$$

Theorem 3.3.3 applies and the result follows. □

Remark 3.3.2. If δ^* has constant risk and is Bayes with respect to some prior π, then $\inf_\delta r(\pi, \delta) = \sup_\theta R(\theta, \delta^*)$ is satisfied and δ^* is minimax. See Problem 3.3.4 for an example.

Minimax procedures and symmetry

As we have seen, minimax procedures have constant risk or at least constant risk on the "most difficult" θ. There is a deep connection between symmetries of the model and the structure of such procedures developed by Hunt and Stein, Lehmann, and others, which is discussed in detail in Chapter 9 of Lehmann (1986) and Chapter 5 of Lehmann and Casella (1998), for instance. We shall discuss this approach somewhat in Volume II but refer to Lehmann (1986) and Lehmann and Casella (1998) for further reading.

Summary. We introduce the minimax principle in the context of the theory of games. Using this framework we connect minimaxity and Bayes methods and develop sufficient conditions for a procedure to be minimax and apply them in several important examples.

More specifically, we show how finding minimax procedures can be viewed as solving a *game* between a statistician S and nature N in which S selects a decision rule δ and N selects a prior π. The *lower (upper) value* $\underline{v}(\bar{v})$ of the game is the supremum (infimum) over priors (decision rules) of the infimum (supremum) over decision rules (priors) of the Bayes risk. A prior for which the Bayes risk of the Bayes procedure equals the lower value of the game is called *least favorable.* When $\underline{v} = \bar{v}$, the game is said to have a *value* v. Von Neumann's Theorem states that if Θ and $\mathcal{D}$ are both finite, then the game of S versus N has a value v, there is a least favorable prior π^* and a minimax rule δ^* such that δ^* is the Bayes rule for π^* and π^* maximizes the Bayes risk of δ^* over all priors. Moreover, v equals the Bayes risk of the Bayes rule δ^* for the prior π^*. We show that Bayes rules with constant risk, or more generally with constant risk over the support of some prior, are minimax. This result is extended to rules that are limits of Bayes rules with constant risk and we use it to show that $\bar{x}$ is a minimax rule for squared error loss in the $\mathcal{N}(\theta, \sigma_0^2)$ model.

3.4 UNBIASED ESTIMATION AND RISK INEQUALITIES

3.4.1 Unbiased Estimation, Survey Sampling

In the previous two sections we have considered two decision theoretic optimality principles, Bayes and minimaxity, for which it is possible to characterize and, in many cases, compute procedures (in particular estimates) that are best in the class of all procedures, $\mathcal{D}$, according to these criteria. An alternative approach is to specify a proper subclass of procedures, $\mathcal{D}_0 \subset \mathcal{D}$, on other grounds, computational ease, symmetry, and so on, and then see if within the $\mathcal{D}_0$ we can find $\delta^* \in \mathcal{D}_0$ that is best according to the "gold standard," $R(\theta, \delta) \geq R(\theta, \delta^*)$ for all θ, all $\delta \in \mathcal{D}_0$. Obviously, we can also take this point of view with humbler aims, for example, looking for the procedure $\delta_\pi^* \in \mathcal{D}_0$ that minimizes the Bayes risk with respect to a prior π among all $\delta \in \mathcal{D}_0$. This approach has early on been applied to parametric families $\mathcal{D}_0$. When $\mathcal{D}_0$ is the class of linear procedures and l is quadratic loss, the solution is given in Section 3.2.

In the non-Bayesian framework, if $\mathbf{Y}$ is postulated as following a linear regression model with $E(Y) = \mathbf{z}^T\boldsymbol{\beta}$ as in Section 2.2.1, then in estimating a linear function of the β_j it is natural to consider the computationally simple class of linear estimates, $S(\mathbf{Y}) = \sum_{i=1}^n d_i Y_i$. This approach coupled with the principle of unbiasedness we now introduce leads to the famous Gauss–Markov theorem proved in Section 6.6.

We introduced, in Section 1.3, the notion of bias of an estimate $\delta(X)$ of a parameter $q(\theta)$ in a model $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$ as

$$\text{Bias}_\theta(\delta) \equiv E_\theta \delta(X) - q(\theta).$$

An estimate such that $\text{Bias}_\theta(\delta) \equiv 0$ is called *unbiased.* This notion has intuitive appeal,

ruling out, for instance, estimates that ignore the data, such as $\delta(X) \equiv q(\theta_0)$, which can't be beat for $\theta = \theta_0$ but can obviously be arbitrarily terrible. The most famous unbiased estimates are the familiar estimates of μ and σ^2 when $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ given by (see Example 1.3.3 and Problem 1.3.8)

$$\widehat{\mu} = \bar{X} \tag{3.4.1}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{3.4.2}$$

Because for unbiased estimates mean square error and variance coincide we call an unbiased estimate $\delta^*(X)$ of $q(\theta)$ that has minimum MSE among all unbiased estimates for all θ, UMVU (uniformly minimum variance unbiased). As we shall see shortly for $\bar{X}$ and in Volume 2 for s^2, these are both UMVU.

Unbiased estimates play a particularly important role in survey sampling.

Example 3.4.1. *Unbiased Estimates in Survey Sampling.* Suppose we wish to sample from a finite population, for instance, a census unit, to determine the average value of a variable (say) monthly family income during a time between two censuses and suppose that we have available a list of families in the unit with family incomes at the last census. Write $x_1, \ldots, x_N$ for the unknown current family incomes and correspondingly $u_1, \ldots, u_N$ for the known last census incomes. We ignore difficulties such as families moving. We let $X_1, \ldots, X_n$ denote the incomes of a sample of n families drawn at random without replacement. This leads to the model with $\mathbf{x} = (x_1, \ldots, x_N)$ as parameter

$$\begin{aligned} P_{\mathbf{x}}[X_1 = a_1, \ldots, X_n = a_n] &= \frac{1}{\binom{N}{n}} \text{ if } \{a_1, \ldots, a_n\} \subset \{x_1, \ldots, x_N\} \\ &= 0 \text{ otherwise.} \end{aligned} \tag{3.4.3}$$

We want to estimate the parameter $\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$. It is easy to see that the natural estimate $\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased (Problem 3.4.14) and has

$$MSE(\bar{X}) = \mathrm{Var}_{\mathbf{x}}(\bar{X}) = \frac{\sigma_{\mathbf{x}}^2}{n}\left(1 - \frac{n-1}{N-1}\right) \tag{3.4.4}$$

where

$$\sigma_{\mathbf{x}}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2. \tag{3.4.5}$$

This method of sampling does not use the information contained in $u_1, \ldots, u_N$. One way to do this, reflecting the probable correlation between $(u_1, \ldots, u_N)$ and $(x_1, \ldots, x_N)$, is to estimate by a regression estimate

$$\widehat{\bar{X}}_R \equiv \bar{X} - b(\bar{U} - \bar{u}) \tag{3.4.6}$$

where b is a prespecified positive constant, U_i is the last census income corresponding to X_i, and $\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i$, $\bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i$. Clearly for each b, $\widehat{\bar{X}}_R$ is also unbiased. If the correlation of U_i and X_i is positive and $b < 2\text{Cov}(\bar{U}, \bar{X})/\text{Var}(\bar{U})$, this will be a better estimate than $\bar{X}$ and the best choice of b is $b_{\text{opt}} \equiv \text{cov}(\bar{U}, \bar{X})/\text{Var}(\bar{U})$ (Problem 3.4.19). The value of b_{opt} is unknown but can be estimated by

$$\widehat{b}_{\text{opt}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(U_i - \bar{U})}{\frac{1}{N}\sum_{j=1}^{N}(U_j - \bar{U})^2}.$$

The resulting estimate is no longer unbiased but behaves well for large samples—see Problem 5.3.11.

An alternative approach to using the u_j is to not sample all units with the same probability. Specifically let $0 \leq \pi_1, \ldots, \pi_N \leq 1$ with $\sum_{j=1}^{N} \pi_j = n$. For each unit $1, \ldots, N$ toss a coin with probability π_j of landing heads and select x_j if the coin lands heads. The result is a sample $S = \{X_1, \ldots, X_M\}$ of random size M such that $E(M) = n$ (Problem 3.4.15). If the π_j are not all equal, $\bar{X}$ is not unbiased but the following estimate known as the *Horvitz–Thompson estimate* is:

$$\widehat{\bar{x}}_{HT} \equiv \frac{1}{N}\sum_{i=1}^{M} \frac{X_i}{\pi_{J_i}} \tag{3.4.7}$$

where J_i is defined by $X_i = x_{J_i}$. To see this write

$$\widehat{\bar{x}}_{HT} = \frac{1}{N}\sum_{j=1}^{N} \frac{x_j}{\pi_j} 1(x_j \in S).$$

Because $\pi_j = P[x_j \in S]$ by construction unbiasedness follows. A natural choice of π_j is $\frac{u_j}{\bar{u}}n$. This makes it more likely for big incomes to be included and is intuitively desirable. It is possible to avoid the undesirable random sample size of these schemes and yet have specified π_j. The Horvitz–Thompson estimate then stays unbiased. Further discussion of this and other sampling schemes and comparisons of estimates are left to the problems. □

Discussion. Unbiasedness is also used in stratified sampling theory (see Problem 1.3.4). However, outside of sampling, the unbiasedness principle has largely fallen out of favor for a number of reasons.

(i) Typically unbiased estimates do not exist—see Bickel and Lehmann (1969) and Problem 3.4.18, for instance.

(ii) Bayes estimates are necessarily biased—see Problem 3.4.20—and minimax estimates often are.

(iii) Unbiased estimates do not obey the attractive equivariance property. If $\widetilde{\theta}$ is unbiased for θ, $q(\widetilde{\theta})$ is biased for $q(\widetilde{\theta})$ unless q is linear. They necessarily in general differ from maximum likelihood estimates except in an important special case we develop later.

Nevertheless, as we shall see in Chapters 5 and 6, good estimates in large samples are approximately unbiased. We expect that $|\text{Bias}_\theta(\widehat{\theta}_n)|/\text{Var}_\theta^{\frac{1}{2}}(\widehat{\theta}_n) \to 0$ or equivalently $\text{Var}_\theta(\widehat{\theta}_n)/MSE_\theta(\widehat{\theta}_n) \to 1$ as $n \to \infty$. In particular we shall show that maximum likelihood estimates are approximately unbiased and approximately best among all estimates. The arguments will be based on asymptotic versions of the important inequalities in the next subsection.

Finally, unbiased estimates are still in favor when it comes to estimating residual variances. For instance, in the linear regression model $\mathbf{Y} = \mathbf{Z}_D\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ of Section 2.2, the variance $\sigma^2 = \text{Var}(\boldsymbol{\varepsilon}_i)$ is estimated by the unbiased estimate $S^2 = \widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}}/(n-p)$ where $\widehat{\boldsymbol{\varepsilon}} = (\mathbf{Y} - \mathbf{Z}_D\widehat{\boldsymbol{\beta}})$, $\widehat{\boldsymbol{\beta}}$ is the least squares estimate, and p is the number of coefficients in $\boldsymbol{\beta}$. This preference of S^2 over the MLE $\widehat{\sigma}^2 = \widehat{\boldsymbol{\varepsilon}}^T\boldsymbol{\varepsilon}/n$ is in accord with optimal behavior when both the number of observations and number of parameters are large. See Problem 3.4.9.

3.4.2 The Information Inequality

The one-parameter case

We will develop a lower bound for the variance of a statistic, which can be used to show that an estimate is UMVU. The lower bound is interesting in its own right, has some decision theoretic applications, and appears in the asymptotic optimality theory of Section 5.4.

We suppose throughout that we have a regular parametric model and further that Θ is an open subset of the line. From this point on we will suppose $p(x,\theta)$ is a density. The discussion and results for the discrete case are essentially identical and will be referred to in the future by the same numbers as the ones associated with the continuous-case theorems given later. We make two regularity assumptions on the family $\{P_\theta : \theta \in \Theta\}$.

(I) The set $A = \{x : p(x,\theta) > 0\}$ does not depend on θ. For all $x \in A$, $\theta \in \Theta$, $\partial/\partial\theta \log p(x,\theta)$ exists and is finite.

(II) If T is any statistic such that $E_\theta(|T|) < \infty$ for all $\theta \in \Theta$, then the operations of integration and differentiation by θ can be interchanged in $\int T(x)p(x,\theta)dx$. That is, for integration over R^q,

$$\frac{\partial}{\partial\theta}\left[\int T(x)p(x,\theta)dx\right] = \int T(x)\frac{\partial}{\partial\theta}p(x,\theta)dx \tag{3.4.8}$$

whenever the right-hand side of (3.4.8) is finite.

Note that in particular (3.4.8) is assumed to hold if $T(x) = 1$ for all x, and we can interchange differentiation and integration in $\int p(x,\theta)dx$.

Assumption II is practically useless as written. What is needed are simple sufficient conditions on $p(x,\theta)$ for II to hold. Some classical conditions may be found in Apostol (1974), p. 167. Simpler assumptions can be formulated using Lebesgue integration theory. For instance, suppose I holds. Then II holds provided that for all T such that $E_\theta(|T|) < \infty$

for all θ, the integrals

$$\int T(x)\left[\frac{\partial}{\partial\theta}p(x,\theta)\right]dx \text{ and } \int\left|T(x)\left[\frac{\partial}{\partial\theta}p(x,\theta)\right]\right|dx$$

are continuous functions[3] of θ. It is not hard to check (using Laplace transform theory) that a one-parameter exponential family quite generally satisfies Assumptions I and II.

Proposition 3.4.1. *If $p(x,\theta)=h(x)\exp\{\eta(\theta)T(x)-B(\theta)\}$ is an exponential family and $\eta(\theta)$ has a nonvanishing continuous derivative on Θ, then I and II hold.*

For instance, suppose $X_1,\ldots,X_n$ is a sample from a $\mathcal{N}(\theta,\sigma^2)$ population, where σ^2 is known. Then (see Table 1.6.1) $\eta(\theta)=\theta/\sigma^2$ and I and II are satisfied. Similarly, I and II are satisfied for samples from gamma and beta distributions with one parameter fixed.

If I holds it is possible to define an important characteristic of the family $\{P_\theta\}$, the *Fisher information number*, which is denoted by $I(\theta)$ and given by

$$I(\theta)=E_\theta\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right)^2=\int\left(\frac{\partial}{\partial\theta}\log p(x,\theta)\right)^2 p(x,\theta)dx. \tag{3.4.9}$$

Note that $0\le I(\theta)\le\infty$.

Lemma 3.4.1. *Suppose that I and II hold and that*

$$E\left|\frac{\partial}{\partial\theta}\log p(X,\theta)\right|<\infty.$$

Then

$$E_\theta\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right)=0 \tag{3.4.10}$$

and, thus,

$$I(\theta)=\text{Var}\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right). \tag{3.4.11}$$

Proof.

$$\begin{aligned}E_\theta\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right) &= \int\left\{\left[\frac{\partial}{\partial\theta}p(x,\theta)\right]\Big/\,p(x,\theta)\right\}p(x,\theta)dx\\ &= \int\frac{\partial}{\partial\theta}p(x,\theta)dx=\frac{\partial}{\partial\theta}\int p(x,\theta)dx=0.\end{aligned}$$

□

Example 3.4.2. Suppose $X_1,\ldots,X_n$ is a sample from a Poisson $\mathcal{P}(\theta)$ population. Then

$$\frac{\partial}{\partial\theta}\log p(\mathbf{x},\theta)=\frac{\sum_{i=1}^n x_i}{\theta}-n \text{ and } I(\theta)=\text{Var}\left(\frac{\sum_{i=1}^n X_i}{\theta}\right)=\frac{1}{\theta^2}n\theta=\frac{n}{\theta}.$$

□

Here is the main result of this section.

Theorem 3.4.1. (Information Inequality). *Let $T(X)$ be any statistic such that $Var_\theta(T(X)) < \infty$ for all θ. Denote $E_\theta(T(X))$ by $\psi(\theta)$. Suppose that I and II hold and $0 < I(\theta) < \infty$. Then for all θ, $\psi(\theta)$ is differentiable and*

$$\mathrm{Var}_\theta(T(X)) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}. \tag{3.4.12}$$

Proof. Using I and II we obtain,

$$\psi'(\theta) = \int T(x)\frac{\partial}{\partial\theta}p(x,\theta)dx = \int T(x)\left(\frac{\partial}{\partial\theta}\log p(x,\theta)\right)p(x,\theta)dx. \tag{3.4.13}$$

By (A.11.14) and Lemma 3.4.1,

$$\psi'(\theta) = \mathrm{Cov}\left(\frac{\partial}{\partial\theta}\log p(X,\theta), T(X)\right). \tag{3.4.14}$$

Now let us apply the correlation (Cauchy–Schwarz) inequality (A.11.16) to the random variables $\partial/\partial\theta \log p(X,\theta)$ and $T(X)$. We get

$$|\psi'(\theta)| \leq \sqrt{\mathrm{Var}(T(X))\mathrm{Var}\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right)}. \tag{3.4.15}$$

The theorem follows because, by Lemma 3.4.1, $\mathrm{Var}\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right) = I(\theta)$. □

The lower bound given in the information inequality depends on $T(X)$ through $\psi(\theta)$. If we consider the class of unbiased estimates of $q(\theta) = \theta$, we obtain a universal lower bound given by the following.

Corollary 3.4.1. *Suppose the conditions of Theorem 3.4.1 hold and T is an unbiased estimate of θ. Then*

$$\mathrm{Var}_\theta(T(X)) \geq \frac{1}{I(\theta)}. \tag{3.4.16}$$

The number $1/I(\theta)$ is often referred to as the *information* or *Cramér–Rao lower bound* for the variance of an unbiased estimate of $\psi(\theta)$.

Here's another important special case.

Proposition 3.4.2. *Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ is a sample from a population with density $f(x,\theta)$, $\theta \in \Theta$, and that the conditions of Theorem 3.4.1 hold. Let $I_1(\theta) = E\left(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\right)^2$, then*

$$I(\theta) = nI_1(\theta) \text{ and } \mathrm{Var}_\theta(T(\mathbf{X})) \geq \frac{[\psi'(\theta)]^2}{nI_1(\theta)}, \tag{3.4.17}$$

Proof. This is a consequence of Lemma 3.4.1 and

$$I(\theta) = \text{Var}\left[\frac{\partial}{\partial\theta}\log p(\mathbf{X},\theta)\right] = \text{Var}\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i,\theta)\right]$$
$$= \sum_{i=1}^{n}\text{Var}\left[\frac{\partial}{\partial\theta}\log f(X_i,\theta)\right] = nI_1(\theta).$$

□

$I_1(\theta)$ is often referred to as the information contained in one observation. We have just shown that the information $I(\theta)$ in a sample of size n is $nI_1(\theta)$.

Next we note how we can apply the information inequality to the problem of unbiased estimation. If the family $\{P_\theta\}$ satisfies I and II and if there exists an unbiased estimate T^* of $\psi(\theta)$ such that $\text{Var}_\theta[T^*(X)] = [\psi'(\theta)]^2/I(\theta)$ for all $\theta \in \Theta$, then T^* is UMVU as an estimate of ψ.

Example 3.4.2. *(Continued).* For a sample from a $\mathcal{P}(\theta)$ distribution, the MLE is $\widehat{\theta} = \bar{X}$. Because $\bar{X}$ is unbiased and $\text{Var}(\bar{X}) = \theta/n$, $\bar{X}$ is UMVU.

Example 3.4.3. Suppose $X_1, \ldots, X_n$ is a sample from a normal distribution with unknown mean θ and known variance σ^2. As we previously remarked, the conditions of the information inequality are satisfied. By Corollary 3.4.1 we see that the conclusion that $\bar{X}$ is UMVU follows if

$$\text{Var}(\bar{X}) = \frac{1}{nI_1(\theta)}. \tag{3.4.18}$$

Now $\text{Var}(\bar{X}) = \sigma^2/n$, whereas if φ denotes the $\mathcal{N}(0,1)$ density, then

$$I_1(\theta) = E\left[\frac{\partial}{\partial\theta}\log\left\{\frac{1}{\sigma}\varphi\left(\frac{X_1-\theta}{\sigma}\right)\right\}\right]^2 = E\left(\frac{(X_1-\theta)}{\sigma^2}\right)^2 = \frac{1}{\sigma^2},$$

and (3.4.18) follows. Note that because $\bar{X}$ is UMVU whatever may be σ^2, we have in fact proved that $\bar{X}$ is UMVU even if σ^2 is unknown. □

We can similarly show (Problem 3.4.1) that if $X_1, \ldots, X_n$ are the indicators of n Bernoulli trials with probability of success θ, then $\bar{X}$ is a UMVU estimate of θ. These are situations in which $\mathbf{X}$ follows a one-parameter exponential family. This is no accident.

Theorem 3.4.2. *Suppose that the family $\{P_\theta : \theta \in \Theta\}$ satisfies assumptions I and II and there exists an unbiased estimate T^* of $\psi(\theta)$, which achieves the lower bound of Theorem 3.4.1 for every θ. Then $\{P_\theta\}$ is a one-parameter exponential family with density or frequency function of the form*

$$p(x,\theta) = h(x)\exp[\eta(\theta)T^*(x) - B(\theta)]. \tag{3.4.19}$$

Conversely, if $\{P_\theta\}$ is a one-parameter exponential family of the form (1.6.1) with natural sufficient statistic $T(X)$ and $\eta(\theta)$ has a continuous nonvanishing derivative on Θ, then $T(X)$ achieves the information inequality bound and is a UMVU estimate of $E_\theta(T(X))$.

Proof. We start with the first assertion. Our argument is essentially that of Wijsman (1973). By (3.4.14) and the conditions for equality in the correlation inequality (A.11.16) we know that T^* achieves the lower bound for all θ if, and only if, there exist functions $a_1(\theta)$ and $a_2(\theta)$ such that

$$\frac{\partial}{\partial\theta}\log p(X,\theta) = a_1(\theta)T^*(X) + a_2(\theta) \tag{3.4.20}$$

with P_θ probability 1 for each θ. From this equality of random variables we shall show that $P_\theta[X \in A^*] = 1$ for all θ where

$$A^* = \left\{x : \frac{\partial}{\partial\theta}\log p(x,\theta) = a_1(\theta)T^*(x) + a_2(\theta) \text{ for all } \theta \in \Theta\right\}. \tag{3.4.21}$$

Upon integrating both sides of (3.4.20) with respect to θ we get (3.4.19).

The passage from (3.4.20) to (3.4.19) is highly technical. However, it is necessary. Here is the argument. If A_θ denotes the set of x for which (3.4.20) hold, then (3.4.20) guarantees $P_\theta(A_\theta) = 1$ and assumption I guarantees $P_{\theta'}(A_\theta) = 1$ for all θ' (Problem 3.4.6). Let $\theta_1, \theta_2, \ldots$ be a denumerable dense subset of Θ. Note that if $A^{**} = \cap_m A_{\theta_m}$, $P_{\theta'}(A^{**}) = 1$ for all θ'. Suppose without loss of generality that $T(x_1) \neq T(x_2)$ for $x_1, x_2 \in A^{**}$. By solving for a_1, a_2 in

$$\frac{\partial}{\partial\theta}\log p(x_j,\theta) = a_1(\theta)T^*(x_j) + a_2(\theta) \tag{3.4.22}$$

for $j = 1, 2$, we see that a_1, a_2 are linear combinations of $\partial\log p(x_j,\theta)/d\theta$, $j = 1, 2$ and, hence, continuous in θ. But now if x is such that

$$\frac{\partial}{\partial\theta}\log p(x,\theta) = a_1(\theta)T^*(x) + a_2(\theta) \tag{3.4.23}$$

for all $\theta_1, \theta_2, \ldots$ and both sides are continuous in θ, then (3.4.23) must hold for all θ. Thus, $A^{**} = A^*$ and the result follows.

Conversely in the exponential family case (1.6.1) we assume without loss of generality (Problem 3.4.3) that we have the canonical case with $\eta(\theta) = \theta$ and $B(\theta) = A(\theta) = \log\int h(x)\exp\{\theta T(x)\}dx$. Then

$$\frac{\partial}{\partial\theta}\log p(X,\theta) = T(X) - A'(\theta) \tag{3.4.24}$$

so that

$$I(\theta) = \mathrm{Var}_\theta(T(X) - A'(\theta)) = \mathrm{Var}_\theta T(X) = A''(\theta). \tag{3.4.25}$$

But $\psi(\theta) = A'(\theta)$ and, thus, the information bound is $[A''(\theta)]^2/A''(\theta) = A''(\theta) = \mathrm{Var}_\theta(T(X))$ so that $T(X)$ achieves the information bound as an estimate of $E_\theta T(X)$. □

Example 3.4.4. In the Hardy-Weinberg model of Examples 2.1.4 and 2.2.6,

$$\begin{aligned} p(\mathbf{x},\theta) &= 2^{n_2}\exp\{(2n_1+n_2)\log\theta + (2n_3+n_3)\log(1-\theta)\} \\ &= 2^{n_2}\exp\{(2n_1+n_2)[\log\theta - \log(1-\theta)] + 2n\log(1-\theta)\} \end{aligned}$$

where we have used the identity $(2n_1 + n_2) + (2n_3 + n_2) = 2n$. Because this is an exponential family, Theorem 3.4.2 implies that $T = (2N_1 + N_2)/2n$ is UMVU for estimating $E(T) = (2n)^{-1}[2n\theta^2 + 2n\theta(1-\theta)] = \theta$.

This T coincides with the MLE $\widehat{\theta}$ of Example 2.2.6. The variance of $\widehat{\theta}$ can be computed directly using the moments of the multinomial distribution of (N_1, N_2, N_3), or by transforming $p(\mathbf{x}, \theta)$ to canonical form by setting $\eta = \log[\theta/(1-\theta)]$ and then using Theorem 1.6.2. A third method would be to use $\mathrm{Var}(\widehat{\theta}) = 1/I(\theta)$ and formula (3.4.25). We find (Problem 3.4.7) $\mathrm{Var}(\widehat{\theta}) = \theta(1-\theta)/2n$. □

Note that by differentiating (3.4.24), we have

$$\frac{\partial^2}{\partial\theta^2}\log p(X,\theta) = -A''(\theta).$$

By (3.4.25) we obtain

$$I(\theta) = -E_\theta \frac{\partial^2}{\partial\theta^2}\log p(X,\theta). \tag{3.4.26}$$

It turns out that this identity also holds outside exponential families:

Proposition 3.4.3. *Suppose $p(\cdot,\theta)$ satisfies in addition to I and II: $p(\cdot,\theta)$ is twice differentiable and interchange between integration and differentiation is permitted. Then* (3.4.26) *holds.*

Proof. We need only check that

$$\frac{\partial^2}{\partial\theta^2}\log p(x,\theta) = \frac{1}{p(x,\theta)}\frac{\partial^2}{\partial\theta^2}p(x,\theta) - \left(\frac{\partial}{\partial\theta}\log p(x,\theta)\right)^2 \tag{3.4.27}$$

and integrate both sides with respect to $p(x,\theta)$. □

Example 3.4.2. *(Continued).* For a sample from a $\mathcal{P}(\theta)$ distribution

$$E_\theta\left(-\frac{\partial^2}{\partial\theta^2}\log p(\mathbf{X},\theta)\right) = \theta^{-2}E\left(\sum_{i=1}^n X_i\right) = \frac{n}{\theta},$$

which equals $I(\theta)$. □

Discussion. It often happens, for instance, in the $\mathcal{U}(0,\theta)$ example, that I and II fail to hold, although UMVU estimates exist. See Volume II. Even worse, as Theorem 3.4.2 suggests, in many situations, assumptions I and II are satisfied and UMVU estimates of $\psi(\theta)$ exist, but the variance of the best estimate is not equal to the bound $[\psi'(\theta)]^2/I(\theta)$. Sharpenings of the information inequality are available but don't help in general.

Extensions to models in which θ is multidimensional are considered next.

The multiparameter case

We will extend the information lower bound to the case of several parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. In particular, we will find a lower bound on the variance of an estimator

$\widehat{\theta}_1 = T$ of θ_1 when the parameters $\theta_2, \ldots, \theta_d$ are unknown. We assume that Θ is an open subset of R^d and that $\{p(x, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is a *regular parametric model* with conditions I and II satisfied when differentiation is with respect θ_j, $j = 1, \ldots, d$. Let $p(x, \boldsymbol{\theta})$ denote the density or frequency function of X where $X \in \mathcal{X} \subset R^q$.

The *(Fisher) information matrix* is defined as

$$I_{d\times d}(\boldsymbol{\theta}) = (I_{jk}(\boldsymbol{\theta}))_{1\le j\le d, 1\le k\le d}, \tag{3.4.28}$$

where

$$I_{jk}(\boldsymbol{\theta}) = E\left(\frac{\partial}{\partial\theta_j}\log p(X,\boldsymbol{\theta})\frac{\partial}{\partial\theta_k}\log p(X,\boldsymbol{\theta})\right). \tag{3.4.29}$$

Proposition 3.4.4. *Under the conditions in the opening paragraph,*

(a)

$$E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta_j}\log p(X,\boldsymbol{\theta})\right) = 0 \tag{3.4.30}$$

$$I_{jk}(\boldsymbol{\theta}) = Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta_j}\log p(X,\boldsymbol{\theta}), \frac{\partial}{\partial\theta_k}\log p(X,\boldsymbol{\theta})\right). \tag{3.4.31}$$

That is,

$$E_{\boldsymbol{\theta}}(\nabla_{\boldsymbol{\theta}}\log p(X,\boldsymbol{\theta})) = \mathbf{0},$$

and

$$I(\boldsymbol{\theta}) = Var(\nabla_{\boldsymbol{\theta}}\log p(X,\boldsymbol{\theta})).$$

(b) *If $X_1, \ldots, X_n$ are i.i.d. as X, then $\mathbf{X} = (X_1, \ldots, X_n)^T$ has information matrix $nI_1(\boldsymbol{\theta})$ where I_1 is the information matrix of X.*

(c) *If, in addition, $p(\cdot, \theta)$ is twice differentiable and double integration and differentiation under the integral sign can be interchanged,*

$$I(\boldsymbol{\theta}) = -\left\|E_{\boldsymbol{\theta}}\left(\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p(X,\boldsymbol{\theta})\right)\right\|, \ 1\le j\le d, \ 1\le k\le d. \tag{3.4.32}$$

Proof. The arguments follow the $d = 1$ case and are left to the problems.

Example 3.4.5. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2)$. Then

$$\log p(x,\boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$I_{11}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\mu^2}\log p(x,\boldsymbol{\theta})\right] = E[\sigma^{-2}] = \sigma^{-2}$$

$$I_{12}(\boldsymbol{\theta}) = -E\left[\frac{\partial}{\partial\sigma^2}\frac{\partial}{\partial\mu}\log p(x,\boldsymbol{\theta})\right] = -\sigma^{-4}E(x-\mu) = 0 = I_{21}(\boldsymbol{\theta})$$

$$I_{22}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{(\partial\sigma^2)^2}\log p(x,\boldsymbol{\theta})\right] = \sigma^{-4}/2.$$

Thus, in this case

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-4}/2 \end{pmatrix}. \tag{3.4.33}$$

□

Example 3.4.6. *Canonical d-Parameter Exponential Family.* Suppose

$$p(x, \boldsymbol{\theta}) = \exp\{\sum_{j=1}^{d} T_j(x)\theta_j - A(\boldsymbol{\theta})\}h(x) \tag{3.4.34}$$

$\boldsymbol{\theta} \in \Theta$ open. The conditions I, II are easily checked and because

$$\nabla_{\boldsymbol{\theta}} \log p(x, \boldsymbol{\theta}) = \mathbf{T}(X) - \dot{A}(\boldsymbol{\theta}),$$

then

$$I(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}} \mathbf{T}(X).$$

By (3.4.30) and Corollary 1.6.1,

$$I(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}} \mathbf{T}(X) = \ddot{A}(\boldsymbol{\theta}). \tag{3.4.35}$$

□

Next suppose $\widehat{\theta}_1 = T$ is an estimate of θ_1 with $\theta_2, \ldots, \theta_d$ assumed unknown. Let $\psi(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} T(X)$ and let $\dot{\psi}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta})$ be the $d \times 1$ vector of partial derivatives. Then

Theorem 3.4.3. *Assume the conditions of the opening paragraph hold and suppose that the matrix $I(\boldsymbol{\theta})$ is nonsingular. Then for all $\boldsymbol{\theta}$, $\dot{\psi}(\boldsymbol{\theta})$ exists and*

$$\mathrm{Var}_{\boldsymbol{\theta}}(T(X)) \geq [\dot{\psi}(\boldsymbol{\theta})]^T I^{-1}(\boldsymbol{\theta})\dot{\psi}(\boldsymbol{\theta}) \,. \tag{3.4.36}$$

Proof. We will use the prediction inequality $\mathrm{Var}(Y) \geq \mathrm{Var}(\mu_L(\mathbf{Z}))$, where $\mu_L(\mathbf{Z})$ denotes the optimal MSPE linear predictor of Y; that is,

$$\mu_L(\mathbf{Z}) = \mu_Y + (\mathbf{Z} - \mu_{\mathbf{Z}})^T \textstyle\sum_{\mathbf{ZZ}}^{-1} \sum_{\mathbf{Z}Y}. \tag{3.4.37}$$

Now set $Y = T(X)$, $\mathbf{Z} = \nabla_{\boldsymbol{\theta}} \log p(X, \boldsymbol{\theta})$. Then, by (B.5.3),

$$\mathrm{Var}_{\boldsymbol{\theta}}(T(X)) \geq \textstyle\sum_{\mathbf{Z}Y}^{T} I^{-1}(\boldsymbol{\theta}) \sum_{\mathbf{Z}Y}, \tag{3.4.38}$$

where $\sum_{\mathbf{Z}Y} = E_{\boldsymbol{\theta}}(T\nabla_{\boldsymbol{\theta}} \log p(X, \boldsymbol{\theta})) = \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(T(X))$ and the last equality follows from the argument in (3.4.13). □

Here are some consequences of this result.

Example 3.4.6. *(continued). UMVU Estimates in Canonical Exponential Families.* Suppose the conditions of Example 3.4.6 hold. We claim that each of $T_j(X)$ is a UMVU

estimate of $E_{\boldsymbol{\theta}}T_j(X)$. This is a different claim than $T_j(X)$ is UMVU for $E_{\boldsymbol{\theta}}T_j(X)$ if θ_i, $i \neq j$, are known. To see our claim note that in our case

$$\psi(\boldsymbol{\theta}) = \frac{\partial A(\boldsymbol{\theta})}{\partial \theta_1},\ \dot{\psi}(\boldsymbol{\theta}) = \left(\frac{\partial^2 A}{\partial \theta_1^2}, \ldots, \frac{\partial^2 A}{\partial \theta_1 \partial \theta_d}\right) \tag{3.4.39}$$

where, without loss of generality, we let $j = 1$. We have already computed in Proposition 3.4.4

$$I^{-1}(\boldsymbol{\theta}) = \left(\frac{\partial^2 A}{\partial \theta_i \partial \theta_j}\right)^{-1}_{d\times d}. \tag{3.4.40}$$

We claim that in this case

$$\dot{\psi}^T(\boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta}) \dot{\psi}(\boldsymbol{\theta}) = \frac{\partial^2 A}{\partial \theta_1^2} \tag{3.4.41}$$

because $\dot{\psi}^T(\boldsymbol{\theta})$ is the first row of $I(\boldsymbol{\theta})$ and, hence, $\dot{\psi}^T(\boldsymbol{\theta})I^{-1}(\boldsymbol{\theta}) = (1, 0, \ldots, 0)$. But $\frac{\partial^2}{\partial \theta_1^2}A(\boldsymbol{\theta})$ is just $\mathrm{Var}_{\boldsymbol{\theta}}T_1(X)$.

Example 3.4.7. *Multinomial Trials.* In the multinomial Example 1.6.6 with $X_1, \ldots, X_n$ i.i.d. as X and $\lambda_j = P(X = j)$, $j = 1, \ldots, k$, we transformed the multinomial model $\mathcal{M}(n, \lambda_1, \ldots, \lambda_k)$ to the canonical form

$$p(\mathbf{x}, \boldsymbol{\theta}) = \exp\{\mathbf{T}^T(\mathbf{x})\boldsymbol{\theta} - A(\boldsymbol{\theta})\}$$

where $\mathbf{T}^T(\mathbf{x}) = (T_1(\mathbf{x}), \ldots, T_{k-1}(\mathbf{x}))$,

$$T_j(\mathbf{X}) = \sum_{j=1}^{n} 1[X_i = j],\ \mathbf{X} = (X_1, \ldots, X_n)^T,\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})^T,$$

$\theta_j = \log(\lambda_j/\lambda_k)$, $j = 1, \ldots, k-1$, and

$$A(\boldsymbol{\theta}) = n \log\left(1 + \sum_{j=1}^{k-1} e^{\theta_j}\right).$$

Note that

$$\frac{\partial}{\partial \theta_j}A(\boldsymbol{\theta}) = \frac{n e^{\theta_j}}{1 + \sum_{l=1}^{k-1} e^{\theta_l}} = n\lambda_j = nE(T_j(\mathbf{X}))$$

$$\frac{\partial^2}{\partial \theta_j^2}A(\boldsymbol{\theta}) = \frac{n e^{\theta_j}\left(1 + \sum_{l=1}^{k-1} e^{\theta_l} - e^{\theta_j}\right)}{\left(1 + \sum_{l=1}^{k-1} e^{\theta_l}\right)^2} = n\lambda_j(1 - \lambda_j) = \mathrm{Var}(T_j(\mathbf{X})).$$

Thus, by Theorem 3.4.3, the lower bound on the variance of an unbiased estimator of $\psi_j(\theta) = E(n^{-1}T_j(\mathbf{X})) = \lambda_j$ is $\lambda_j(1 - \lambda_j)/n$. But because N_j/n is unbiased and has variance $\lambda_j(1 - \lambda_j)/n$, then N_j/n is UMVU for λ_j. □

Example 3.4.8. *The Normal Case.* If $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ then $\bar{X}$ is UMVU for μ and $\frac{1}{n}\sum X_i^2$ is UMVU for $\mu^2 + \sigma^2$. But it does not follow that $\frac{1}{n-1}\sum(X_i - \bar{X})^2$ is UMVU for σ^2. These and other examples and the implications of Theorem 3.4.3 are explored in the problems. □

Here is an important extension of Theorem 3.4.3 whose proof is left to Problem 3.4.21.

Theorem 3.4.4. *Suppose that the conditions of Theorem* 3.4.3 *hold and*

$$\mathbf{T}(X) = (T_1(X), \ldots, T_d(X))^T$$

is a d-dimensional statistic. Let

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\mathbf{T}(X))_{d\times 1} = (\psi_1(\boldsymbol{\theta}), \ldots, \psi_d(\boldsymbol{\theta}))^T$$

and $\dot{\boldsymbol{\psi}}(\boldsymbol{\theta}) = \left(\frac{\partial \psi_i}{\partial \theta_j}(\boldsymbol{\theta})\right)_{d\times d}$. *Then*

$$Var_{\boldsymbol{\theta}}\mathbf{T}(X) \geq \dot{\boldsymbol{\psi}}(\boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta}) \dot{\boldsymbol{\psi}}^T(\boldsymbol{\theta}) \tag{3.4.42}$$

where $A \geq B$ means $\mathbf{a}^T(A - B)\mathbf{a} \geq 0$ for all $\mathbf{a}_{d\times 1}$.

Note that both sides of (3.4.42) are $d \times d$ matrices. Also note that

$$\widehat{\boldsymbol{\theta}} \text{ unbiased } \Rightarrow \text{Var}_{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}} \geq I^{-1}(\boldsymbol{\theta}).$$

In Chapters 5 and 6 we show that in smoothly parametrized models, reasonable estimates are asymptotically unbiased. We establish analogues of the information inequality and use them to show that under suitable conditions the MLE is asymptotically optimal.

Summary. We study the important application of the unbiasedness principle in survey sampling. We derive the information inequality in one-parameter models and show how it can be used to establish that in a canonical exponential family, $T(\mathbf{X})$ is the UMVU estimate of its expectation. Using inequalities from prediction theory, we show how the information inequality can be extended to the multiparameter case. Asymptotic analogues of these inequalities are sharp and lead to the notion and construction of efficient estimates.

3.5 NONDECISION THEORETIC CRITERIA

In practice, even if the loss function and model are well specified, features other than the risk function are also of importance in selection of a procedure. The three principal issues we discuss are the speed and numerical stability of the method of computation used to obtain the procedure, interpretability of the procedure, and robustness to model departures.

3.5.1 Computation

Speed of computation and numerical stability issues have been discussed briefly in Section 2.4. They are dealt with extensively in books on numerical analysis such as Dahlquist,

Björk, and Anderson (1974). We discuss some of the issues and the subtleties that arise in the context of some of our examples in estimation theory.

Closed form versus iteratively computed estimates

At one level closed form is clearly preferable. For instance, a method of moments estimate of (λ, p) in Example 2.3.2 is given by

$$\widehat{\lambda} = \frac{\bar{X}}{\widehat{\sigma}^2},\ \widehat{p} = \frac{\bar{X}^2}{\widehat{\sigma}^2}$$

where $\widehat{\sigma}^2$ is the empirical variance (Problem 2.2.11). It is clearly easier to compute than the MLE. Of course, with ever faster computers a difference at this level is irrelevant. But it reappears when the data sets are big and the number of parameters large.

On the other hand, consider the Gaussian linear model of Example 2.1.1. Then least squares estimates are given in closed form by equation (2.2.10). The closed form here is deceptive because inversion of a $d \times d$ matrix takes on the order of d^3 operations when done in the usual way and can be numerically unstable. It is in fact faster and better to solve equation (2.1.9) by, say, Gaussian elimination for the particular $\mathbf{Z}_D^T\mathbf{Y}$.

Faster versus slower algorithms

Consider estimation of the MLE $\widehat{\boldsymbol{\theta}}$ in a general canonical exponential family as in Section 2.3. It may be shown that, in the algorithm we discuss in Section 2.4, if we seek to take enough steps J so that $|\widehat{\boldsymbol{\theta}}^{(J)} - \theta| \leq \varepsilon < 1$ then J is of the order of $\log \frac{1}{\varepsilon}$ (Problem 3.5.1). On the other hand, at least if started close enough to $\widehat{\boldsymbol{\theta}}$, the Newton–Raphson method in which the jth iterate, $\widehat{\boldsymbol{\theta}}^{(j)} = \widehat{\boldsymbol{\theta}}^{(j-1)} - \ddot{A}^{-1}(\widehat{\boldsymbol{\theta}}^{(j-1)}(T(X) - \dot{A}(\widehat{\theta}^{(j-1)}))$, takes on the order of $\log\log \frac{1}{\varepsilon}$ steps (Problem 3.5.2). The improvement in speed may however be spurious since $\ddot{A}^{-1}$ is costly to compute if d is large—though the same trick as in computing least squares estimates can be used.

The interplay between estimated variance and computation

As we have seen in special cases in Examples 3.4.3 and 3.4.4, estimates of parameters based on samples of size n have standard deviations of order $n^{-1/2}$. It follows that striving for numerical accuracy of order smaller than $n^{-1/2}$ is wasteful. Unfortunately it is hard to translate statements about orders into specific prescriptions without assuming at least bounds on the constants involved.

3.5.2 Interpretability

Suppose that in the normal $\mathcal{N}(\mu, \sigma^2)$ Example 2.1.5 we are interested in the parameter μ/σ. This parameter, the signal-to-noise ratio, for this population of measurements has a clear interpretation. Its maximum likelihood estimate $\bar{X}/\widehat{\sigma}$ continues to have the same intuitive interpretation as an estimate of μ/σ even if the data are a sample from a distribution with

mean μ and variance σ^2 other than the normal. On the other hand, suppose we initially postulate a model in which the data are a sample from a gamma, $\mathcal{G}(p, \lambda)$, distribution. Then $E(X)/\sqrt{\text{Var}(X)} = (p/\lambda)(p/\lambda^2)^{-1/2} = p^{1/2}$. We can now use the MLE $\widehat{p}^{1/2}$, which as we shall see later (Section 5.4) is for n large a more precise estimate than $\bar{X}/\widehat{\sigma}$ if this model is correct. However, the form of this estimate is complex and if the model is incorrect it no longer is an appropriate estimate of $E(X)/[\text{Var}(X)]^{1/2}$. We return to this in Section 5.5.

3.5.3 Robustness

Finally, we turn to robustness.

This is an issue easy to point to in practice but remarkably difficult to formalize appropriately. The idea of robustness is that we want estimation (or testing) procedures to perform reasonably even when the model assumptions under which they were designed to perform excellently are not exactly satisfied. However, what reasonable means is connected to the choice of the parameter we are estimating (or testing hypotheses about). We consider three situations.

(a) The problem dictates the parameter. For instance, the Hardy–Weinberg parameter θ has a clear biological interpretation and is *the* parameter for the experiment described in Example 2.1.4. Similarly, economists often work with median housing prices, that is, the parameter ν that has half of the population prices on either side (formally, ν is any value such that $P(X \leq \nu) \geq \frac{1}{2}$, $P(X \geq \nu) \geq \frac{1}{2}$). Alternatively, they may be interested in total consumption of a commodity such as coffee, say $\theta = N\mu$, where N is the population size and μ is the expected consumption of a randomly drawn individual.

(b) We imagine that the random variable X^* produced by the random experiment we are interested in has a distribution that follows a "true" parametric model with an interpretable parameter θ, but we do not necessarily observe X^*. The actual observation X is X^* contaminated with "gross errors"—see the following discussion. But θ is still the target in which we are interested.

(c) We have a qualitative idea of what the parameter is, but there are several parameters that satisfy this qualitative notion. This idea has been developed by Bickel and Lehmann (1975a, 1975b, 1976) and Doksum (1975), among others. For instance, we may be interested in the center of a population, and both the mean μ and median ν qualify. See Problem 3.5.13.

We will consider situations (b) and (c).

Gross error models

Most measurement and recording processes are subject to *gross errors*, anomalous values that arise because of human error (often in recording) or instrument malfunction. To be a bit formal, suppose that if n measurements $\mathbf{X}^* \equiv (X_1^*, \ldots, X_n^*)$ could be taken without gross errors then $P^* \in \mathcal{P}^*$ would be an adequate approximation to the distribution of $\mathbf{X}^*$ (i.e., we could suppose $\mathbf{X}^* \sim P^* \in \mathcal{P}^*$). However, if gross errors occur, we observe not $\mathbf{X}^*$ but $\mathbf{X} = (X_1, \ldots, X_n)$ where most of the $X_i = X_i^*$, but there are a few

wild values. Now suppose we want to estimate $\theta(P^*)$ and use $\widehat{\theta}(X_1, \ldots, X_n)$ knowing that $\widehat{\theta}(X_1^*, \ldots, X_n^*)$ is a good estimate. Informally $\widehat{\theta}(X_1, \ldots, X_n)$ will continue to be a good or at least reasonable estimate if its value is not greatly affected by the $X_i \neq X_i^*$, the gross errors. Again informally we shall call such procedures *robust.* Formal definitions require model specification, specification of the gross error mechanism, and definitions of insensitivity to gross errors. Most analyses require asymptotic theory and will have to be postponed to Chapters 5 and 6. However, two notions, the sensitivity curve and the breakdown point, make sense for fixed n. The breakdown point will be discussed in Volume II. We next define and examine the sensitivity curve in the context of the Gaussian location model, Example 1.1.2, and then more generally.

Consider the one-sample symmetric *location* model $\mathcal{P}$ defined by

$$X_i = \mu + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3.5.1}$$

where the errors are independent, identically distributed, and symmetric about 0 with common density f and d.f. F. If the error distribution is normal, $\bar{X}$ is the best estimate in a variety of senses.

In our new formulation it is the X_i^* that obey (3.5.1). A reasonable formulation of a model in which the possibility of gross errors is acknowledged is to make the ε_i still i.i.d. but with common distribution function F and density f of the form

$$f(x) = (1 - \lambda)\frac{1}{\sigma}\varphi\left(\frac{x}{\sigma}\right) + \lambda h(x). \tag{3.5.2}$$

Here h is the density of the gross errors and λ is the probability of making a gross error. This corresponds to,

$$\begin{aligned} X_i &= X_i^* \text{ with probability } 1 - \lambda \\ &= Y_i \text{ with probability } \lambda \end{aligned}$$

where Y_i has density $h(y - \mu)$ and (X_i^*, Y_i) are i.i.d. Note that this implies the possibly unreasonable assumption that committing a gross error is independent of the value of X^*. Further assumptions that are commonly made are that h has a particular form, for example, $h = \frac{1}{K\sigma}\varphi\left(\frac{x}{K\sigma}\right)$ where $K \gg 1$ or more generally that h is an unknown density symmetric about 0. Then the gross error model is semiparametric, $\mathcal{P}_\delta \equiv \{f(\cdot - \mu) : f$ satisfies (3.5.2) for some h such that $h(x) = h(-x)$ for all $x\}$. $P_{(\mu,f)} \in \mathcal{P}_\delta$ iff $X_1, \ldots, X_n$ are i.i.d. with common density $f(x - \mu)$, where f satisfies (3.5.2). The advantage of this formulation is that μ remains identifiable. That is, it is the center of symmetry of $P_{(\mu,f)}$ for all such P. Unfortunately, the assumption that h is itself symmetric about 0 seems patently untenable for gross errors. However, if we drop the symmetry assumption, we encounter one of the basic difficulties in formulating robustness in situation (b). Without h symmetric the quantity μ is not a parameter, so it is unclear what we are estimating. That is, it is possible to have $P_{(\mu_1,f_1)} = P_{(\mu_2,f_2)}$ for $\mu_1 \neq \mu_2$ (Problem 3.5.18). Is μ_1 or μ_2 our goal? On the other hand, in situation (c), we do not need the symmetry assumption. We return to these issues in Chapter 6.

The sensitivity curve

At this point we ask: Suppose that an estimate $T(X_1, \ldots, X_n) = \theta(\widehat{F})$, where $\widehat{F}$ is the empirical d.f., is appropriate for the symmetric location model, $\mathcal{P}$, in particular, has the plug-in property, $\theta(P_{(\mu,f)}) = \mu$ for all $P_{(\mu,f)} \in \mathcal{P}$. How sensitive is it to the presence of gross errors among $X_1, \ldots, X_n$? An interesting way of studying this due to Tukey (1972) and Hampel (1974) is the *sensitivity curve* defined as follows for plug-in estimates (which are well defined for all sample sizes n).

We start by defining the sensitivity curve for general plug-in estimates. Suppose that $X \sim P$ and that $\theta = \theta(P)$ is a parameter. The empirical plug-in estimate of θ is $\widehat{\theta} = \theta(\widehat{P})$ where $\widehat{P}$ is the empirical probability distribution. See Section 2.1.2. The *sensitivity curve* of $\widehat{\theta}$ is defined as

$$SC(x; \widehat{\theta}) = n[\widehat{\theta}(x_1, \ldots, x_{n-1}, x) - \widehat{\theta}(x_1, \ldots, x_{n-1})],$$

where $x_1, \ldots, x_{n-1}$ represents an observed sample of size $n-1$ from P and x represents an observation that (potentially) comes from a distribution different from P. We are interested in the shape of the sensitivity curve, not its location. In our examples we shall, therefore, shift the sensitivity curve in the horizontal or vertical direction whenever this produces more transparent formulas. Often this is done by fixing $x_1, \ldots, x_{n-1}$ as an "ideal" sample of size $n-1$ for which the estimator $\widehat{\theta}$ gives us the right value of the parameter and then we see what the introduction of a potentially deviant nth observation x does to the value of $\widehat{\theta}$.

We return to the location problem with θ equal to the mean $\mu = E(X)$. Because the estimators we consider are location invariant, that is, $\widehat{\theta}(X_1, \ldots, X_n) - \mu = \widehat{\theta}(X_1 - \mu, \ldots, X_n - \mu)$, and because $E(X_j - \mu) = 0$, we take $\mu = 0$ without loss of generality. Now fix $x_1, \ldots, x_{n-1}$ so that their mean has the ideal value zero. This is equivalent to shifting the SC vertically to make its value at $x = 0$ equal to zero. See Problem 3.5.14. Then

$$SC(x; \bar{x}) = n\left(\frac{x_1 + \cdots + x_{n-1} + x}{n}\right) = x.$$

Thus, the sample mean is arbitrarily sensitive to gross error—a large gross error can throw the mean off entirely. Are there estimates that are less sensitive?

A classical estimate of location based on the order statistics is the *sample median* $\widehat{X}$ which we write as

$$\begin{aligned} \widehat{X} &= X_{(k+1)} && \text{if } n = 2k+1 \\ &= \tfrac{1}{2}(X_{(k)} + X_{(k+1)}) && \text{if } n = 2k \end{aligned}$$

where $X_{(1)}, \ldots, X_{(n)}$ are the order statistics, that is, $X_1, \ldots, X_n$ ordered from smallest to largest. See Section 1.3 and (2.1.17).

The sample median can be motivated as an estimate of location on various grounds.

(i) It is the empirical plug-in estimate of the population median ν (Problem 3.5.4), and it splits the sample into two equal halves.

(ii) In the symmetric location model (3.5.1), ν coincides with μ and $\widehat{x}$ is an empirical plug-in estimate of μ.

(iii) The sample median is the MLE when we assume the common density $f(x)$ of the errors $\{\varepsilon_i\}$ in (3.5.1) is the Laplace (double exponential) density

$$f(x) = \frac{1}{2\tau}\exp\{-|x|/\tau\},$$

a density having substantially heavier tails than the normal. See Problems 2.2.32 and 3.5.9.

The sensitivity curve of the median is as follows:

If, say, $n = 2k+1$ is odd and the median of $x_1, \ldots, x_{n-1} = (x^{(k)} + x^{(k+1)})/2 = 0$, we obtain

$$\begin{aligned} SC(x;\widehat{x}) &= nx^{(k)} = -nx^{(k+1)} && \text{for } x < x^{(k)} \\ &= nx && \text{for } x^{(k)} \le x \le x^{(k+1)} \\ &= nx^{(k+1)} && \text{for } x > x^{(k+1)} \end{aligned}$$

where $x^{(1)} \le \cdots \le x^{(n-1)}$ are the ordered $x_1, \ldots, x_{n-1}$.

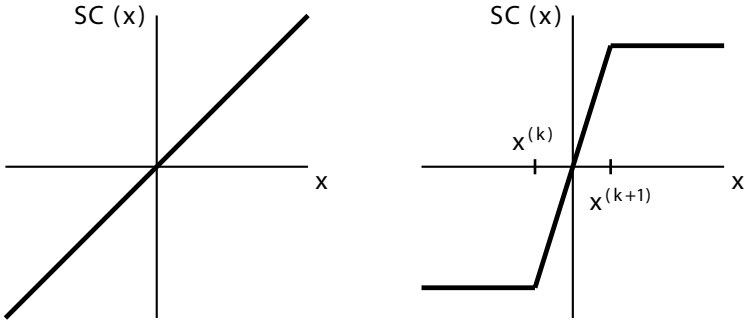

Figure 3.5.1. The sensitivity curves of the mean and median.

Although the median behaves well when gross errors are expected, its performance at the normal model is unsatisfactory in the sense that its variance is about 57% larger than the variance of $\bar{X}$. The sensitivity curve in Figure 3.5.1 suggests that we may improve matters by constructing estimates whose behavior is more like that of the mean when x is near μ. A class of estimates providing such intermediate behavior and including both the mean and

the median has been known since the eighteenth century. Let $0 \leq \alpha < \frac{1}{2}$. We define the α trimmed mean, $\bar{X}_\alpha$, by

$$\bar{X}_\alpha = \frac{X_{([n\alpha]+1)} + \cdots + X_{(n-[n\alpha])}}{n - 2[n\alpha]} \tag{3.5.3}$$

where $[n\alpha]$ is the largest integer $\leq n\alpha$ and $X_{(1)} < \cdots < X_{(n)}$ are the ordered observations. That is, we throw out the "outer" $[n\alpha]$ observations on either side and take the average of the rest. The estimates can be justified on plug-in grounds (see Problem 3.5.5). For more sophisticated arguments see Huber (1981). Note that if $\alpha = 0$, $\bar{X}_\alpha = \bar{X}$, whereas as $\alpha \uparrow \frac{1}{2}$, $\bar{X}_\alpha \to \widehat{X}$. For instance, suppose we take as our data the differences in Table 3.5.1.

If $[n\alpha] = [(n-1)\alpha]$ and the trimmed mean of $x_1, \ldots, x_{n-1}$ is zero, the sensitivity curve of an α trimmed mean is sketched in Figure 3.5.2. (The middle portion is the line $y = x(1 - 2[n\alpha]/n)^{-1}$.)

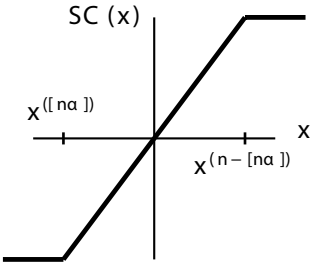

Figure 3.5.2. The sensitivity curve of the trimmed mean.

Intuitively we expect that if there are no gross errors, that is, $f = \varphi$, the mean is better than any trimmed mean with $\alpha > 0$ including the median, which corresponds approximately to $\alpha = \frac{1}{2}$. This can be verified in terms of asymptotic variances (MSEs)—see Problem 5.4.1. However, the sensitivity curve calculation points to an equally intuitive conclusion. If f is symmetric about 0 but has "heavier tails" (see Problem 3.5.8) than the Gaussian density, for example, the Laplace density, $f(x) = \frac{1}{2}e^{-|x|}$, or even more strikingly the Cauchy, $f(x) = 1/\pi(1 + x^2)$, then the trimmed means for $\alpha > 0$ and even the median can be much better than the mean, infinitely better in the case of the Cauchy—see Problem 5.4.1 again.

Which α should we choose in the trimmed mean? There seems to be no simple answer. The range $0.10 \leq \alpha \leq 0.20$ seems to yield estimates that provide adequate protection against the proportions of gross errors expected and yet perform reasonably well when sampling is from the normal distribution. See Andrews, Bickel, Hampel, Haber, Rogers, and Tukey (1972). There has also been some research into procedures for which α is chosen using the observations. For a discussion of these and other forms of "adaptation," see Jaeckel (1971), Huber (1972), and Hogg (1974).

Gross errors or outlying data points affect estimates in a variety of situations. We next consider two estimates of the spread in the population as well as estimates of quantiles; other examples will be given in the problems. If we are interested in the spread of the values in a population, then the variance σ^2 or standard deviation σ is typically used. A fairly common quick and simple alternative is the IQR (interquartile range) defined as $\tau = x_{.75} - x_{.25}$, where x_α has 100α percent of the values in the population on its left (formally, x_α is any value such that $P(X \leq x_\alpha) \geq \alpha$, $P(X \geq x_\alpha) \geq 1 - \alpha$). x_α is called a αth *quantile* and $x_{.75}$ and $x_{.25}$ are called the *upper* and *lower quartiles.* The IQR is often calibrated so that it equals σ in the $\mathcal{N}(\mu, \sigma^2)$ model. Because $\tau = 2 \times (.674)\sigma$, the scale measure used is $0.742(x_{.75} - x_{.25})$.

Example 3.5.1. *Spread.* Let $\theta(P) = \mathrm{Var}(X) = \sigma^2$ denote the variance in a population and let $X_1, \ldots, X_n$ denote a sample from that population. Then $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is the empirical plug-in estimate of σ^2. To simplify our expression we shift the horizontal axis so that $\sum_{i=1}^{n-1} x_i = 0$. Write $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i = n^{-1}x$, then

$$\begin{aligned}
SC(x; \widehat{\sigma}^2) &= n(\widehat{\sigma}_n^2 - \widehat{\sigma}_{n-1}^2) \\
&= \sum_{i=1}^{n-1}(x_i - n^{-1}x)^2 + (x - n^{-1}x)^2 - n\widehat{\sigma}_{n-1}^2 \\
&= \sum_{i=1}^{n-1} x_i^2 + (n-1)(n^{-1}x)^2 + [(n-1)/n]^2 x^2 - n\widehat{\sigma}_{n-1}^2 \\
&= (n-1)\widehat{\sigma}_{n-1}^2 + \left\{\left(\frac{n-1}{n}\right)^2 + \frac{n-1}{n^2}\right\} x^2 - n\widehat{\sigma}_{n-1}^2 \\
&= \left\{\left(\frac{n-1}{n}\right)^2 + \frac{n-1}{n^2}\right\} x^2 - \widehat{\sigma}_{n-1}^2 \cong x^2 - \widehat{\sigma}_{n-1}^2.
\end{aligned}$$

It is clear that $\widehat{\sigma}_n^2$ is very sensitive to large outlying $|x|$ values. Similarly,

$$\begin{aligned}
SC(x; \widehat{\sigma}) &= n(\widehat{\sigma}_n - \widehat{\sigma}_{n-1}) \\
&= n[\{\widehat{\sigma}_{n-1}^2 + (\widehat{\sigma}_n^2 - \sigma_{n-1}^2)\}^{\frac{1}{2}} - \widehat{\sigma}_{n-1}] \\
&= n\widehat{\sigma}_{n-1}\left[\left\{1 + \frac{\widehat{\sigma}_n^2}{\widehat{\sigma}_{n-1}^2} - 1\right\}^{\frac{1}{2}} - 1\right] \\
&\cong \widehat{\sigma}_{n-1}\frac{n}{2}\left(\frac{\widehat{\sigma}_n^2}{\widehat{\sigma}_{n-1}^2} - 1\right) \\
&= SC(x; \widehat{\sigma}^2)/2\widehat{\sigma}_{n-1}
\end{aligned} \tag{3.5.4}$$

where the approximation is valid for x fixed, $n \to \infty$ (Problem 3.5.10). □

Example 3.5.2. *Quantiles and the IQR.* Let $\theta(P) = x_\alpha$ denote a αth quantile of the distribution of X, $0 < \alpha < 1$, and let $\widehat{x}_\alpha$ denote the αth sample quantile (see 2.1.16).

If $n\alpha$ is an integer, say k, the αth sample quantile is $\widehat{x}_\alpha = \frac{1}{2}[x_{(k)} + x_{(k+1)}]$, and at sample size $n-1$, $\widehat{x}_\alpha = x^{(k)}$, where $x^{(1)} \leq \cdots \leq x^{(n-1)}$ are the ordered $x_1, \ldots, x_{n-1}$,

thus, for $2 \leq k \leq n-2$,

$$\begin{aligned} SC(x;\widehat{x}_\alpha) &= \frac{1}{2}[x^{(k-1)} - x^{(k)}],\ x \leq x^{(k-1)} \\ &= \frac{1}{2}[x - x^{(k)}],\ x^{(k-1)} \leq x \leq x^{(k+1)} \\ &= \frac{1}{2}[x^{(k+1)} - x^{(k)}],\ x \geq x^{(k+1)}. \end{aligned} \tag{3.5.5}$$

Clearly, $\widehat{x}_\alpha$ is not sensitive to outlying x's.

Next consider the *sample IQR*

$$\widehat{\tau} = \widehat{x}_{.75} - \widehat{x}_{.25}.$$

Then we can write

$$SC(x;\widehat{\tau}) = SC(x;\widehat{x}_{.75}) - SC(x;\widehat{x}_{.25})$$

and the sample IQR is robust with respect to outlying gross errors x. □

Remark 3.5.1. The sensitivity of the parameter $\theta(F)$ to x can be measured by the *influence function*, which is defined by

$$IF(x;\theta,F) = \lim_{\epsilon \downarrow 0} IF_\epsilon(x;\theta,F)$$

where

$$I_\epsilon(x;\theta,F) = \epsilon^{-1}[\theta((1-\epsilon)F + \epsilon\Delta_x) - \theta(F)]$$

and Δ_x is the distribution function of point mass at x ($\Delta_x(t) = 1[t \geq x]$). It is easy to see that (Problem 3.5.15)

$$SC(x;\widehat{\theta}) = IF_{\frac{1}{n}}(x;\theta,\widehat{F}_{n-1})$$

where $\widehat{F}_{n-1}$ denotes the empirical distribution based on $x_1,\ldots,x_{n-1}$. We will return to the influence function in Volume II. It plays an important role in functional expansions of estimates.

Discussion. Other aspects of robustness, in particular the breakdown point, have been studied extensively and a number of procedures proposed and implemented. Unfortunately these procedures tend to be extremely demanding computationally, although this difficulty appears to be being overcome lately. An exposition of this point of view and some of the earlier procedures proposed is in Hampel, Ronchetti, Rousseuw, and Stahel (1983).

Summary. We discuss briefly nondecision theoretic considerations for selecting procedures including interpretability, and computability. Most of the section focuses on robustness, discussing the difficult issues of identifiability. The rest of our very limited treatment focuses on the sensitivity curve as illustrated in the mean, trimmed mean, median, and other procedures.

3.6 PROBLEMS AND COMPLEMENTS

Problems for Section 3.2

1. Show that if $X_1, \ldots, X_n$ is a $\mathcal{N}(\theta, \sigma^2)$ sample and π is the improper prior $\pi(\theta) = 1$, $\theta \in \Theta = R$, then the improper Bayes rule for squared error loss is $\delta^*(\mathbf{x}) = \bar{x}$.

2. Let $X_1, \ldots, X_n$ be the indicators of n Bernoulli trials with success probability θ. Suppose $l(\theta, a)$ is the quadratic loss $(\theta - a)^2$ and that the prior $\pi(\theta)$ is the beta, $\beta(r, s)$, density. Find the Bayes estimate $\widehat{\theta}_B$ of θ and write it as a weighted average $w\theta_0 + (1-w)\bar{X}$ of the mean θ_0 of the prior and the sample mean $\bar{X} = S/n$. Show that $\widehat{\theta}_B = (S+1)/(n+2)$ for the uniform prior.

3. In Problem 3.2.2 preceeding, give the MLE of the Bernoulli variance $q(\theta) = \theta(1-\theta)$ and give the Bayes estimate of $q(\theta)$. Check whether $q(\widehat{\theta}_B) = E(q(\boldsymbol{\theta}) \mid \mathbf{x})$, where $\widehat{\theta}_B$ is the Bayes estimate of θ.

4. In the Bernoulli Problem 3.2.2 with uniform prior on the probabilility of success θ, we found that $(S+1)/(n+2)$ is the Bayes rule. In some studies (see Section 6.4.3), the parameter $\lambda = \theta/(1-\theta)$, which is called the *odds ratio* (for success), is preferred to θ. If we put the (improper) uniform prior $\Pi(\lambda) = 1$ $(\lambda > 0)$ on λ and use quadratic loss $(\lambda - a)^2$, under what condition on S does the Bayes rule exist and what is the Bayes rule?

5. Suppose $\boldsymbol{\theta} \sim \pi(\theta)$, $(X \mid \boldsymbol{\theta} = \theta) \sim p(x \mid \theta)$.

(a) Show that the joint density of X and θ is

$$f(x, \theta) = p(x \mid \theta)\pi(\theta) = c(x)\pi(\theta \mid x)$$

where $c(x) = \int \pi(\theta)p(x \mid \theta)d\theta$.

(b) Let $l(\theta, a) = (\theta - a)^2/w(\theta)$ for some weight function $w(\theta) > 0$, $\theta \in \Theta$. Show that the Bayes rule is

$$\delta^* = E_{f_0}(\boldsymbol{\theta} \mid x)$$

where

$$f_0(x, \theta) = p(x \mid \theta)[\pi(\theta)/w(\theta)]/c$$

and

$$c = \int\int p(x \mid \theta)[\pi(\theta)/w(\theta)]d\theta dx$$

is assumed to be finite. That is, if π and l are changed to $a(\theta)\pi(\theta)$ and $l(\theta, a)/a(\theta)$, $a(\theta) > 0$, respectively, the Bayes rule does not change.

Hint: See Problem 1.4.24.

(c) In Example 3.2.3, change the loss function to $l(\theta, a) = (\theta - a)^2/\theta^\alpha(1-\theta)^\beta$. Give the conditions needed for the posterior Bayes risk to be finite and find the Bayes rule.

6. Find the Bayes risk $r(\pi, \delta)$ of $\delta(\mathbf{x}) = \bar{X}$ in Example 3.2.1. Consider the relative risk $e(\delta, \pi) = R(\pi)/r(\pi, \delta)$, where $R(\pi)$ is the Bayes risk. Compute the limit of $e(\delta, \pi)$ as

(a) $\tau \to \infty$, **(b)** $n \to \infty$, **(c)** $\sigma^2 \to \infty$.

7. For the following problems, compute the posterior risks of the possible actions and give the optimal Bayes decisions when $x = 0$.

(a) Problem 1.3.1(d);

(b) Problem 1.3.2(d)(i) and (ii);

(c) Problem 1.3.19(c).

8. Suppose that $N_1, \ldots, N_r$ given $\boldsymbol{\theta} = \theta$ are multinomial $\mathcal{M}(n, \theta)$, $\theta = (\theta_1, \ldots, \theta_r)^T$, and that $\boldsymbol{\theta}$ has the Dirichlet distribution $\mathcal{D}(\alpha)$, $\alpha = (\alpha_1, \ldots, \alpha_r)^T$, defined in Problem 1.2.15. Let $q(\theta) = \sum_{j=1}^r c_j \theta_j$, where $c_1, \ldots, c_r$ are given constants.

(a) If $l(\theta, a) = [q(\theta) - a]^2$, find the Bayes decision rule δ^* and the minimum conditional Bayes risk $r(\delta^*(x) \mid x)$.

Hint: If $\boldsymbol{\theta} \sim \mathcal{D}(\alpha)$, then $E(\boldsymbol{\theta}_j) = \alpha_j/\alpha_0$, $\mathrm{Var}(\boldsymbol{\theta}_j) = \alpha_j(\alpha_0 - \alpha_j)/\alpha_0^2(\alpha_0 + 1)$, and $\mathrm{Cov}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = -\alpha_i\alpha_j/\alpha_0^2(\alpha_0 + 1)$, where $\alpha_0 = \sum_{j=1}^r \alpha_j$. (Use these results, do not derive them.)

(b) When the loss function is $l(\theta, a) = (q(\theta) - a)^2 / \prod_{j=1}^r \theta_j$, find necessary and sufficient conditions under which the Bayes risk is finite and under these conditions find the Bayes rule.

(c) We want to estimate the vector $(\theta_1, \ldots, \theta_r)$ with loss function $l(\theta, a) = \sum_{j=1}^r (\theta_j - a_j)^2$. Find the Bayes decision rule.

9. *Bioequivalence trials* are used to test whether a generic drug is, to a close approximation, equivalent to a name-brand drug. Let $\theta = \mu_G - \mu_B$ be the difference in mean effect of the generic and name-brand drugs. Suppose we have a sample $X_1, \ldots, X_n$ of differences in the effect of generic and name-brand effects for a certain drug, where $E(X) = \theta$. A regulatory agency specifies a number $\epsilon > 0$ such that if $\theta \in (-\epsilon, \epsilon)$, then the generic and brand-name drugs are, by definition, bioequivalent. On the basis of $\mathbf{X} = (X_1, \ldots, X_n)$ we want to decide whether or not $\theta \in (-\epsilon, \epsilon)$. Assume that given θ, $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\theta, \sigma_0^2)$, where σ_0^2 is known, and that $\boldsymbol{\theta}$ is random with a $\mathcal{N}(\eta_0, \tau_0^2)$ distribution.

There are two possible actions:

$$\begin{aligned} a &= 0 \Leftrightarrow \text{ Bioequivalent} \\ a &= 1 \Leftrightarrow \text{ Not Bioequivalent} \end{aligned}$$

with losses $l(\theta, 0)$ and $l(\theta, 1)$. Set

$$\lambda(\theta) = l(\theta, 0) - l(\theta, 1)$$

= difference in loss of acceptance and rejection of bioequivalence. Note that $\lambda(\theta)$ should be negative when $\theta \in (-\epsilon, \epsilon)$ and positive when $\theta \notin (-\epsilon, \epsilon)$. One such function (Lindley, 1998) is

$$\lambda(\theta) = r - \exp\left\{-\frac{1}{2c^2}\theta^2\right\},\ c^2 > 0$$

where $0 < r < 1$. Note that $\lambda(\pm\epsilon) = 0$ implies that r satisfies

$$\log r = -\frac{1}{2c^2}\epsilon^2.$$

This is an example with two possible actions 0 and 1 where $l(\theta, 0)$ and $l(\theta, 1)$ are not constant. Any two functions with difference $\lambda(\theta)$ are possible loss functions at $a = 0$ and 1.

(a) Show that the Bayes rule is equivalent to

$$\text{"Accept bioequivalence if } E(\lambda(\boldsymbol{\theta}) \mid \mathbf{X} = \mathbf{x}) < 0\text{"} \tag{3.6.1}$$

and show that (3.6.1) is equivalent to

$$\text{"Accept bioequivalence if } [E(\boldsymbol{\theta} \mid \mathbf{x})]^2 < (\tau_0^2(n) + c^2)\{\log(\tfrac{c^2}{\tau_0^2(n)+c^2}) + \tfrac{\epsilon^2}{c^2}\}\text{"}$$

where

$$E(\boldsymbol{\theta} \mid \mathbf{x}) = w\eta_0 + (1-w)\bar{x},\ w = \tau_0^2(n)/\tau_0^2,\ \tau_0^2(n) = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}.$$

Hint: See Example 3.2.1.

(b) It is proposed that the preceding prior is "uninformative" if it has $\eta_0 = 0$ and τ_0^2 large ("$\tau_0^2 \to \infty$"). Discuss the preceding decision rule for this "prior."

(c) Discuss the behavior of the preceding decision rule for large n ("$n \to \infty$"). Consider the general case (a) and the specific case (b).

10. For the model defined by (3.2.16) and (3.2.17), find

(a) the linear Bayes estimate of Δ_1.

(b) the linear Bayes estimate of μ.

(c) Is the assumption that the Δ's are normal needed in (a) and (b)?

Problems for Section 3.3

1. In Example 3.3.2 show that $L(\mathbf{x}, 0, v) \geq \pi/(1-\pi)$ is equivalent to $T \geq t$.

2. Suppose $g : S \times T \to R$. A point (x_0, y_0) is a *saddle point* of g if

$$g(x_0, y_0) = \sup_S g(x, y_0) = \inf_T g(x_0, y).$$

Suppose S and T are subsets of R^m, R^p, respectively, $(\mathbf{x}_0, \mathbf{y}_0)$ is in the interior of $S \times T$, and g is twice differentiable.

(a) Show that a necessary condition for $(\mathbf{x}_0, \mathbf{y}_0)$ to be a saddle point is that, representing $\mathbf{x} = (x_1, \ldots, x_m)$, $\mathbf{y} = (y_1, \ldots, y_p)$,

$$\frac{\partial g}{\partial x_i}(\mathbf{x}_0, \mathbf{y}_0) = \frac{\partial g}{\partial y_j}(\mathbf{x}_0, \mathbf{y}_0) = 0,$$

and

$$\frac{\partial^2 g}{\partial x_a \partial x_b}(\mathbf{x}_0, \mathbf{y}_0) \leq 0, \ \frac{\partial^2 g(\mathbf{x}_0, \mathbf{y}_0)}{\partial y_c \partial y_d} \geq 0$$

for all $1 \leq i, a, b \leq m$, $1 \leq j, c, d \leq p$.

(b) Suppose $S_m = \{\mathbf{x} : x_i \geq 0, \ 1 \leq i \leq m, \ \sum_{i=1}^m x_i = 1\}$, the simplex, and $g(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^m \sum_{j=1}^p c_{ij} x_i y_j$ with $\mathbf{x} \in S_m$, $\mathbf{y} \in S_p$. Show that the von Neumann minimax theorem is equivalent to the existence of a saddle point for any g as above.

3. Suppose $\Theta = \{\theta_0, \theta_1\}$, $\mathcal{A} = \{0, 1\}$, and that the model is regular. Suppose

$$l(\theta_i, i) = 0, \ l(\theta_i, j) = w_{ij} > 0, \ i, j = 0, 1, \ i \neq j.$$

Let $L_X(\theta_0, \theta_1) = p(X, \theta_1)/p(X, \theta_0)$ and suppose that $L_X(\theta_0, \theta_1)$ has a continuous distribution under both P_{θ_0} and P_{θ_1}. Show that

(a) For every $0 < \pi < 1$, the test rule δ_π given by

$$\begin{aligned} \delta_\pi(X) &= 1 \text{ if } L_X(\theta_0, \theta_1) \geq \tfrac{(1-\pi)w_{01}}{\pi w_{10}} \\ &= 0 \text{ otherwise} \end{aligned}$$

is Bayes against a prior such that $P[\boldsymbol{\theta} = \theta_1] = \pi = 1 - P[\boldsymbol{\theta} = \theta_0]$, and

(b) There exists $0 < \pi^* < 1$ such that the prior π^* is least favorable against δ_{π^*}, that is, the conclusion of von Neumann's theorem holds.

Hint: Show that there exists (a unique) π^* so that

$$R(\theta_0, \delta_{\pi^*}) = R(\theta_1, \delta_{\pi^*}).$$

4. Let $S \sim \mathcal{B}(n, \theta)$, $l(\theta, a) = (\theta - a)^2$, $\delta(S) = \bar{X} = S/n$, and

$$\delta^*(S) = (S + \frac{1}{2}\sqrt{n})/(n + \sqrt{n}).$$

(a) Show that δ^* has constant risk and is Bayes for the beta, $\beta(\sqrt{n}/2, \sqrt{n}/2)$, prior. Thus, δ^* is minimax.

Hint: See Problem 3.2.2.

(b) Show that $\lim_{n\to\infty}[R(\theta, \delta^*)/R(\theta, \delta)] > 1$ for $\theta \neq \frac{1}{2}$; and show that this limit equals 1 when $\theta = \frac{1}{2}$.

5. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ and $l(\sigma^2, d) = \left(\frac{d}{\sigma^2} - 1\right)^2$.

(a) Show that if μ is known to be 0

$$\delta^*(X_1, \ldots, X_n) = \frac{1}{n+2} \sum X_i^2$$

is minimax.

(b) If $\mu = 0$, show that δ^* is uniformly best among all rules of the form $\delta_c(\mathbf{X}) = c\sum X_i^2$. Conclude that the MLE is inadmissible.

(c) Show that if μ is unknown, $\delta(\mathbf{X}) = \frac{1}{n+1}\sum(X_i - \bar{X})^2$ is best among all rules of the form $\delta_c(\mathbf{X}) = c\sum(X_i - \bar{X})^2$ and, hence, that both the MLE and the estimate $S^2 = (n-1)^{-1}\sum(X_i - \bar{X})^2$ are inadmissible.

Hint: (a) Consider a gamma prior on $\theta = 1/\sigma^2$. See Problem 1.2.12. (c) Use (B.3.29).

6. Let $X_1, \ldots, X_k$ be independent with means $\mu_1, \ldots, \mu_k$, respectively, where

$$(\mu_1, \ldots, \mu_k) = (\mu^0_{i_1}, \ldots, \mu^0_{i_k}),\ \mu^0_1 < \cdots < \mu^0_k$$

is a known set of values, and $i_1, \ldots, i_k$ is an arbitrary unknown permutation of $1, \ldots, k$. Let $\mathcal{A} = \{(j_1, \ldots, j_k) : \text{Permutations of } 1, \ldots, k\}$

$$l((i_1, \ldots, i_k), (j_1, \ldots, j_k)) = \sum_{l,m} 1(i_l < i_m, j_l > j_m).$$

Show that the minimax rule is to take

$$\delta(X_1, \ldots, X_k) = (R_1, \ldots, R_k)$$

where R_j is the rank of X_j, that is, $R_j = \sum_{l=1}^k 1(X_l \leq X_j)$.

Hint: Consider the uniform prior on permutations and compute the Bayes rule by showing that the posterior risk of a permutation $(i_1, \ldots, i_k)$ is smaller than that of $(i'_1, \ldots, i'_k)$, where $i'_j = i_j$, $j \neq a, b$, $a < b$, $i'_a = i_b$, $i'_b = i_a$, and $R_a < R_b$.

7. Show that X has a Poisson (λ) distribution and $l(\lambda, a) = (\lambda - a)^2/\lambda$. Then X is minimax.

Hint: Consider the gamma, $\Gamma(k^{-1}, 1)$, prior. Let $k \to \infty$.

8. Let X_i be independent $\mathcal{N}(\mu_i, 1)$, $1 \leq i \leq k$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$. Write $\mathbf{X} = (X_1, \ldots, X_k)^T$, $\mathbf{d} = (d_1, \ldots, d_k)^T$. Show that if

$$l(\boldsymbol{\mu}, \mathbf{d}) = \sum_{i=1}^k (d_i - \mu_i)^2$$

then $\boldsymbol{\delta}(\mathbf{X}) = \mathbf{X}$ is minimax.

Remark: Stein (1956) has shown that if $k \geq 3$, $\mathbf{X}$ is no longer unique minimax. For instance,

$$\delta^*(\mathbf{X}) = \left(1 - \frac{k-2}{|\mathbf{X}|^2}\right)\mathbf{X}$$

is also minimax and $R(\boldsymbol{\mu}, \delta^*) < R(\boldsymbol{\mu}, \boldsymbol{\delta})$ for all $\boldsymbol{\mu}$. See Volume II.

9. Show that if $(N_1, \ldots, N_k)$ has a multinomial, $\mathcal{M}(n, p_1, \ldots, p_k)$, distribution, $0 < p_j < 1$, $1 \leq j \leq k$, then $\frac{\mathbf{N}}{n}$ is minimax for the loss function

$$l(\mathbf{p}, \mathbf{d}) = \sum_{j=1}^k \frac{(d_j - p_j)^2}{p_j q_j}$$

where $q_j = 1 - p_j$, $1 \leq j \leq k$.

Hint: Consider Dirichlet priors on $(p_1, \ldots, p_{k-1})$ with density defined in Problem 1.2.15. See also Problem 3.2.8.

10. Let $X_i (i = 1, \ldots, n)$ be i.i.d. with unknown distribution F. For a given x we want to estimate the proportion $F(x)$ of the population to the left of x. Show that

$$\delta = \frac{\text{No. of } X_i \leq x}{\sqrt{n}} \cdot \frac{1}{1+\sqrt{n}} + \frac{1}{2(1+\sqrt{n})}$$

is minimax for estimating $F(x) = P(X_i \leq x)$ with squared error loss.

Hint: Consider the risk function of δ. See Problem 3.3.4.

11. Let $X_1, \ldots, X_n$ be independent $\mathcal{N}(\mu, 1)$. Define

$$\begin{aligned} \delta(\bar{X}) &= \bar{X} + \frac{d}{\sqrt{n}} \text{ if } \bar{X} < -\frac{d}{\sqrt{n}} \\ &= 0 \text{ if } |\bar{X}| \leq \frac{d}{\sqrt{n}} \\ &= \bar{X} - \frac{d}{\sqrt{n}} \text{ if } \bar{X} > \frac{d}{\sqrt{n}}. \end{aligned}$$

(a) Show that the risk (for squared error loss) $E(\sqrt{n}(\delta(\bar{X}) - \mu))^2$ of these estimates is bounded for all n and μ.

(b) How does the risk of these estimates compare to that of $\bar{X}$?

12. Suppose that given $\boldsymbol{\theta} = \theta$, X has a binomial, $\mathcal{B}(n, \theta)$, distribution. Show that the Bayes estimate of θ for the Kullback–Leibler loss function $l_p(\theta, a)$ is the posterior mean $E(\boldsymbol{\theta} \mid X)$.

13. Suppose that given $\boldsymbol{\theta} = \theta = (\theta_1, \ldots, \theta_k)^T$, $X = (X_1, \ldots, X_k)^T$ has a multinomial, $\mathcal{M}(n, \theta)$, distribution. Let the loss function be the Kullback–Leibler divergence $l_p(\theta, a)$ and let the prior be the uniform prior

$$\pi(\theta_1, \ldots, \theta_{k-1}) = (k-1)!,\ \theta_j \geq 0,\ \sum_{j=1}^{k} \theta_j = 1.$$

Show that the Bayes estimate is $(X_i + 1)/(n + k)$.

14. Let $K(p_\theta, q)$ denote the KLD (Kullback–Leibler divergence) between the densities p_θ and q and define the *Bayes KLD* between $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ and q as

$$k(q, \pi) = \int K(p_\theta, q)\pi(\theta)d\theta.$$

Show that the marginal density of X,

$$p(x) = \int p_\theta(x)\pi(\theta)d\theta,$$

minimizes $k(q, \pi)$ and that the minimum is

$$I_{\boldsymbol{\theta},X} \equiv \int \left[E_\theta \left\{ \log \frac{p_\theta(X)}{p(X)} \right\} \right] \pi(\theta) d\theta.$$

$I_{\boldsymbol{\theta},X}$ is called the *mutual information* between $\boldsymbol{\theta}$ and X.

Hint: $k(q,\pi) - k(p,\pi) = \int \left[E_\theta \left\{ \log \frac{p(X)}{q(X)} \right\} \right] \pi(\theta) d\theta \geq 0$ by Jensen's inequality.

15. *Jeffrey's "Prior."* A density proportional to $\sqrt{I_p(\theta)}$ is called Jeffrey's prior. It is often improper. Show that in the $\mathcal{N}(\theta, \sigma_0^2)$, $\mathcal{N}(\mu_0, \theta)$ and $\mathcal{B}(n, \theta)$ cases, Jeffrey's priors are proportional to 1, θ^{-1}, and $\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$, respectively. Give the Bayes rules for squared error in these three cases.

Problems for Section 3.4

1. Let $X_1, \ldots, X_n$ be the indicators of n Bernoulli trials with success probability θ. Show that $\bar{X}$ is an UMVU estimate of θ.

2. Let $\mathcal{A} = R$. We shall say a loss function is *convex*, if $l(\theta, \alpha a_0 + (1-\alpha)a_1) \leq \alpha l(\theta, a_0) + (1-\alpha) l(\theta, a_1)$, for any $a_0, a_1, \theta, 0 < \alpha < 1$. Suppose that there is an unbiased estimate δ of $q(\theta)$ and that $T(\mathbf{X})$ is sufficient. Show that if $l(\theta, a)$ is convex and $\delta^*(\mathbf{X}) = E(\delta(\mathbf{X}) \mid T(\mathbf{X}))$, then $R(\theta, \delta^*) \leq R(\theta, \delta)$.

Hint: Use *Jensen's inequality:* If g is a convex function and X is a random variable, then $E(g(X)) \geq g(E(X))$.

3. *Equivariance.* Let $X \sim p(x, \theta)$ with $\theta \in \Theta \subset R$, suppose that assumptions I and II hold and that h is a monotone increasing differentiable function from Θ onto $h(\Theta)$. Reparametrize the model by setting $\eta = h(\theta)$ and let $q(x, \eta) = p(x, h^{-1}(\eta))$ denote the model in the new parametrization.

(a) Show that if $I_p(\theta)$ and $I_q(\eta)$ denote the Fisher information in the two parametrizations, then

$$I_q(\eta) = I_p(h^{-1}(\eta))/[h'(h^{-1}(\eta))]^2.$$

That is, Fisher information is not equivariant under increasing transformations of the parameter.

(b) *Equivariance of the Fisher Information Bound.* Let $B_p(\theta)$ and $B_q(\eta)$ denote the information inequality lower bound $(\psi')^2/I$ as in (3.4.12) for the two parametrizations $p(x, \theta)$ and $q(x, \eta)$. Show that $B_q(\eta) = B_p(h^{-1}(\eta))$; that is, the Fisher information lower bound is equivariant.

4. Prove Proposition 3.4.4.

5. Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\mu = \mu_0$ known. Show that

(a) $\widehat{\sigma}_0^2 = n^{-1} \sum_{i=1}^n (X_i - \mu_0)^2$ is a UMVU estimate of σ^2.

(b) $\widehat{\sigma}_0^2$ is inadmissible under squared error loss.

Hint: See Problem 3.3.5(b).

(c) if μ_0 is not known and the true distribution of X_i is $\mathcal{N}(\mu, \sigma^2)$, $\mu \neq \mu_0$, find the bias of $\widehat{\sigma}_0^2$.

6. Show that assumption I implies that if $A \equiv \{x : p(\mathbf{x}, \theta) > 0\}$ doesn't depend on θ, then for any set B, $P_\theta(B) = 1$ for some θ if and only if $P_\theta(B) = 1$ for all θ.

7. In Example 3.4.4, compute $\mathrm{Var}(\widehat{\theta})$ using each of the three methods indicated.

8. Establish the claims of Example 3.4.8.

9. Show that $S^2 = (\mathbf{Y} - \mathbf{Z}_D\widehat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{Z}_D\widehat{\boldsymbol{\beta}})/(n-p)$ is an unbiased estimate of σ^2 in the linear regression model of Section 2.2.

10. Suppose $\widehat{\theta}$ is UMVU for estimating θ. Let a and b be constants. Show that $\widehat{\lambda} = a + b\widehat{\theta}$ is UMVU for estimating $\lambda = a + b\theta$.

11. Suppose $Y_1, \ldots, Y_n$ are independent Poisson random variables with $E(Y_i) = \mu_i$ where $\mu_i = \exp\{\alpha + \beta z_i\}$ depends on the levels z_i of a covariate; $\alpha, \beta \in R$. For instance, z_i could be the level of a drug given to the ith patient with an infectious disease and Y_i could denote the number of infectious agents in a given unit of blood from the ith patient 24 hours after the drug was administered.

(a) Write the model for $Y_1, \ldots, Y_n$ in two-parameter canonical exponential form and give the sufficient statistic.

(b) Let $\boldsymbol{\theta} = (\alpha, \beta)^T$. Compute $I(\boldsymbol{\theta})$ for the model in (a) and then find the lower bound on the variances of unbiased estimators $\widehat{\alpha}$ and $\widehat{\beta}$ of α and β.

(c) Suppose that $z_i = \log[i/(n+1)]$, $i = 1, \ldots, n$. Find $\lim n^{-1} I(\boldsymbol{\theta})$ as $n \to \infty$, and give the limit of n times the lower bound on the variances of $\widehat{\alpha}$ and $\widehat{\beta}$.

Hint: Use the integral approximation to sums.

12. Let $X_1, \ldots, X_n$ be a sample from the beta, $\mathcal{B}(\theta, 1)$, distribution.

(a) Find the MLE of $1/\theta$. Is it unbiased? Does it achieve the information inequality lower bound?

(b) Show that $\bar{X}$ is an unbiased estimate of $\theta/(\theta+1)$. Does $\bar{X}$ achieve the information inequality lower bound?

13. Let $\mathcal{F}$ denote the class of densities with mean θ^{-1} and variance $\theta^{-2}(\theta > 0)$ that satisfy the conditions of the information inequality. Show that a density that minimizes the Fisher information over $\mathcal{F}$ is $f(x, \theta) = \theta e^{-\theta x} 1(x > 0)$.

Hint: Consider $T(X) = X$ in Theorem 3.4.1.

14. Show that if $(X_1, \ldots, X_n)$ is a sample drawn without replacement from an unknown finite population $\{x_1, \ldots, x_N\}$, then

(a) $\bar{X}$ is an unbiased estimate of $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$.

(b) The variance of $\bar{X}$ is given by (3.4.4).

15. Suppose $u_1, \ldots, u_N$ are as in Example 3.4.1 and u_j is retained independently of all other u_j with probability π_j where $\sum_{j=1}^N \pi_j = n$. Show that if M is the expected sample size, then

$$E(M) = \sum_{j=1}^{N} \pi_j = n.$$

16. Suppose the sampling scheme given in Problem 15 is employed with $\pi_j \equiv \frac{n}{N}$. Show that the resulting unbiased Horvitz–Thompson estimate for the population mean has variance strictly larger than the estimate obtained by taking the mean of a sample of size n taken without replacement from the population.

17. *Stratified Sampling.* (See also Problem 1.3.4.) Suppose the u_j can be relabeled into strata $\{x_{ki}\}$, $1 \leq i \leq I_k$, $k = 1, \ldots, K$, $\sum_{k=1}^K I_k = N$. Let $\pi_k = \frac{I_k}{N}$ and suppose $\pi_k = \frac{m_k}{n}$, $1 \leq k \leq K$.

(a) Take samples with replacement of size m_k from stratum $k = \{x_{k1}, \ldots, x_{kI_k}\}$ and form the corresponding sample averages $\bar{X}_1, \ldots, \bar{X}_K$. Define

$$\bar{\bar{X}} \equiv \frac{1}{K} \sum_{k=1}^{K} \pi_k \bar{X}_k.$$

Show that $\bar{\bar{X}}$ is unbiased and if $\bar{X}$ is the mean of a simple random sample without replacement from the population then

$$\text{Var } \bar{\bar{X}} \leq \text{Var } \bar{X}$$

with equality iff $x_{k\cdot} = I_k^{-1} \sum_{i=1}^{I_k} x_{ki}$ doesn't depend on k for all k such that $\pi_k > 0$.

(b) Show that the inequality between Var $\bar{\bar{X}}$ and Var $\bar{X}$ continues to hold if $\frac{m_k - 1}{I_k - 1} \geq \frac{n-1}{N-1}$ for all k, even for sampling without replacement in each stratum.

18. Let X have a binomial, $\mathcal{B}(n, p)$, distribution. Show that $\frac{p}{1-p}$ is not unbiasedly estimable. More generally only polynomials of degree n in p are unbiasedly estimable.

19. Show that $\widehat{X}_k$ given by (3.4.6) is **(a)** unbiased and **(b)** has smaller variance than $\bar{X}$ if $b < 2 \text{ Cov}(\bar{U}, \bar{X})/\text{Var}(\bar{U})$.

20. Suppose X is distributed according to $\{P_\theta : \theta \in \Theta \subset R\}$ and π is a prior distribution for $\boldsymbol{\theta}$ such that $E(\boldsymbol{\theta}^2) < \infty$.

(a) Show that $\delta(X)$ is both an unbiased estimate of $\boldsymbol{\theta}$ and the Bayes estimate with respect to quadratic loss, if and only if, $P[\delta(X) = \boldsymbol{\theta}] = 1$.

(b) Deduce that if $P_\theta = \mathcal{N}(\theta, \sigma_0^2)$, X is not a Bayes estimate for any prior π.

(c) Explain how it is possible if P_θ is binomial, $\mathcal{B}(n, \theta)$, that $\frac{X}{n}$ is a Bayes estimate for θ.

Hint: Given $E(\delta(X) \mid \boldsymbol{\theta}) = \boldsymbol{\theta}$, $E(\boldsymbol{\theta} \mid X) = \delta(X)$ compute $E(\delta(X) - \boldsymbol{\theta})^2$.

21. Prove Theorem 3.4.4.

Hint: It is equivalent to show that, for all $\mathbf{a}_{d\times 1}$,

$$\begin{aligned} \mathrm{Var}(\mathbf{a}^T\widehat{\boldsymbol{\theta}}) &\geq \mathbf{a}^T(\dot{\boldsymbol{\psi}}(\boldsymbol{\theta})I^{-1}(\boldsymbol{\theta})\dot{\boldsymbol{\psi}}^T(\boldsymbol{\theta}))\mathbf{a} \\ &= [\dot{\boldsymbol{\psi}}^T(\boldsymbol{\theta})\mathbf{a}]^T I^{-1}(\boldsymbol{\theta})[\dot{\boldsymbol{\psi}}^T(\boldsymbol{\theta})\mathbf{a}]. \end{aligned}$$

Note that $\dot{\boldsymbol{\psi}}^T(\boldsymbol{\theta})\mathbf{a} = \nabla E_{\boldsymbol{\theta}}(\mathbf{a}^T\widehat{\boldsymbol{\theta}})$ and apply Theorem 3.4.3.

22. *Regularity Conditions are Needed for the Information Inequality.* Let $X \sim \mathcal{U}(0,\theta)$ be the uniform distribution on $(0,\theta)$. Note that $\log p(x,\theta)$ is differentiable for all $\theta > x$, that is, with probability 1 for each θ, and we can thus define moments of $T = \partial/\partial\theta \log p(X,\theta)$. Show that, however,

(i) $E\left(\dfrac{\partial}{\partial\theta}\log p(X,\theta)\right) = -\dfrac{1}{\theta} \neq 0$

(ii) $\mathrm{Var}\left(\dfrac{\partial}{\partial\theta}\log p(X,\theta)\right) = 0$ and $I(\theta) = ET^2 = \mathrm{Var}\,T + (ET)^2 = 1/\theta^2$.

(iii) $2X$ is unbiased for θ and has variance $(1/3)\theta^2 < (1/I(\theta)) = \theta^2$.

Problems for Section 3.5

1. If $n = 2k$ is even, give and plot the sensitivity curve of the median.

2. If $\alpha = 0.25$ and $n\alpha = k$ is an integer, use (3.5.5) to plot the sensitivity curve of the IQR.

3. If $\alpha = 0.25$ and $(n-1)\alpha$ is an integer, give and plot the sensitivity curves of the lower quartile $\widehat{x}_{.25}$, the upper quartile $\widehat{x}_{.75}$, and the IQR.

4. Show that the sample median $\widehat{X}$ is an empirical plug-in estimate of the population median ν.

5. Show that the α trimmed mean $\bar{X}_\alpha$ is an empirical plug-in estimate of

$$\mu_\alpha = (1-2\alpha)^{-1}\int_{x_{1-\alpha}}^{x_\alpha} x dF(x).$$

Here $\int x dF(x)$ denotes $\int xp(x)dx$ in the continuous case and $\sum xp(x)$ in the discrete case.

6. An estimate $\delta(\mathbf{X})$ is said to be *shift* or *translation* equivariant if, for all $x_1,\ldots,x_n,c$,

$$\delta(x_1+c,\ldots,x_n+c) = \delta(x_1,\ldots,x_n) + c.$$

It is *antisymmetric* if for all $x_1, \ldots, x_n$

$$\delta(x_1, \ldots, x_n) = -\delta(-x_1, \ldots, -x_n).$$

(a) Show that $\widehat{X}, \bar{X}, \bar{X}_\alpha$ are translation equivariant and antisymmetric.

(b) Suppose $X_1, \ldots, X_n$ is a sample from a population with d.f. $F(x - \mu)$ where μ is unknown and $X_i - \mu$ is symmetrically distributed about 0. Show that if δ is translation equivariant and antisymmetric and $E_0(\delta(\mathbf{X}))$ exists and is finite, then

$$E_\mu(\delta(\mathbf{X})) = \mu$$

(i.e., δ is an unbiased estimate of μ). Deduce that $\bar{X}, \bar{X}_\alpha, \widehat{X}$ are unbiased estimates of the center of symmetry of a symmetric distribution.

7. The *Hodges–Lehmann (location)* estimate $\widehat{x}_{HL}$ is defined to be the median of the $\frac{1}{2}n(n+1)$ pairwise averages $\frac{1}{2}(x_i + x_j)$, $i \leq j$. Its properties are similar to those of the trimmed mean. It has the advantage that there is no trimming proportion α that needs to be subjectively specified.

(a) Suppose $n = 5$ and the "ideal" ordered sample of size $n - 1 = 4$ is $-1.03, -.30, .30, 1.03$ (these are expected values of four $\mathcal{N}(0,1)$-order statistics). For $x \geq .3$, plot the sensitivity curves of the mean, median, trimmed mean with $\alpha = 1/4$, and the Hodges–Lehmann estimate.

(b) Show that $\widehat{x}_{HL}$ is translation equivariant and antisymmetric. (See Problem 3.5.6.)

8. The *Huber estimate* $\widehat{X}_k$ is defined implicitly as the solution of the equation

$$\sum_{i=1}^{n} \psi_k\left(\frac{X_i - \widehat{X}_k}{\widehat{\sigma}}\right) = 0$$

where $0 \leq k \leq \infty$, $\widehat{\sigma}$ is an estimate of scale, and

$$\begin{aligned} \psi_k(x) &= x \text{ if } |x| \leq k \\ &= k \text{ if } x > k \\ &= -k \text{ if } x < -k. \end{aligned}$$

One reasonable choice for k is $k = 1.5$ and for $\widehat{\sigma}$ is,

$$\widehat{\sigma} = \operatorname*{med}_{1 \leq i \leq n} |X_i - \widehat{X}|/0.67.$$

Show that

(a) $k = \infty$ corresponds to $\bar{X}$, $k \to 0$ to the median.

(b) If $\widehat{\sigma}$ is replaced by a known σ_0, then $\widehat{X}_k$ is the MLE of θ when $X_1, \ldots, X_n$ are i.i.d. with density $f_0((x-\theta)/\sigma_0)$ where

$$\begin{aligned} f_0(x) &= \frac{1-\varepsilon}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \text{for } |x| \le k \\ &= \frac{1-\varepsilon}{\sqrt{2\pi}} e^{k^2/2-k|x|}, \quad \text{for } |x| > k, \end{aligned}$$

with k and ε connected through

$$\frac{2\varphi(k)}{k} - 2\Phi(-k) = \frac{\varepsilon}{1-\varepsilon}.$$

(c) $\widehat{x}_k$ exists and is unique when $k > 0$. Use a fixed known σ_0 in place of $\widehat{\sigma}$.

(d) $\widehat{x}_k$ is translation equivariant and antisymmetric (see Problem 3.5.6).

(e) If $k < \infty$, then $\lim_{|x|\to\infty} SC(x; \widehat{x}_k)$ is a finite constant.

9. If $f(\cdot)$ and $g(\cdot)$ are two densities with medians ν zero and identical scale parameters τ, we say that $g(\cdot)$ has *heavier tails* than $f(\cdot)$ if $g(x)$ is above $f(x)$ for $|x|$ large. In the case of the Cauchy density, the standard deviation does not exist; thus, we will use the IQR scale parameter $\tau = x_{.75} - x_{.25}$. In what follows adjust f and g to have $\nu = 0$ and $\tau = 1$.

(a) Find the set of $|x|$ where $g(|x|) \ge \varphi(|x|)$ for g equal to the Laplace and Cauchy densities $g_L(x) = (2\eta)^{-1} \exp\{-|x|/\eta\}$ and $g_C(x) = b[b^2 + x^2]^{-1}/\pi$.

(b) Find the tail probabilities $P(|X| \ge 2)$, $P(|X| \ge 3)$ and $P(|X| \ge 4)$ for the normal, Laplace, and Cauchy distributions.

(c) Show that $g_C(x)/\varphi(x)$ is of order $\exp\{x^2\}$ as $|x| \to \infty$.

10. Suppose $\sum_{i=1}^{n-1} x_i = 0$. Show that $SC(x, \widehat{\sigma}_n) \xrightarrow{P} (2\sigma)^{-1}(x^2 - \sigma^2)$ as $n \to \infty$.

11. Let μ_0 be a hypothesized mean for a certain population. The (student) t-ratio is defined as $t = \sqrt{n}(\bar{x} - \mu_0)/s$, where $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$. Let $\mu_0 = 0$ and choose the ideal sample $x_1, \ldots, x_{n-1}$ to have sample mean zero. Find the limit of the sensitivity curve of t as

(a) $|x| \to \infty$, n is fixed, and

(b) $n \to \infty$, x is fixed.

12. For the ideal sample of Problem 3.5.7(a), plot the sensitivity curve of

(a) $\widehat{\sigma}_n$, and

(b) the t-ratio of Problem 3.5.11.

This problem may be done on the computer.

13. *Location Parameters.* Let X be a random variable with continuous distribution function F. The functional $\theta = \theta_X = \theta(F)$ is said to be *scale* and *shift (translation) equivariant*

if $\theta_{a+bX} = a + b\theta_X$. It is *antisymmetric* if $\theta_X = \theta_{-X}$. Let Y denote a random variable with continuous distribution function G. X is said to be *stochastically smaller* than Y if $F(t) = P(X \le t) \ge P(Y \le t) = G(t)$ for all $t \in R$. In this case we write $X \overset{st}{\le} Y$. θ is said to be order preserving if $X \overset{st}{\le} Y \Rightarrow \theta_X \le \theta_Y$. If θ is scale and shift equivariant, antisymmetric, and order preserving, it is called a *location parameter.*

(a) Show that if F is symmetric about c and θ is a location parameter, then $\theta(F) = c$.

(b) Show that the mean μ, median ν, and trimmed population mean μ_α (see Problem 3.5.5) are location parameters.

(c) Let $\mu^{(k)}$ be the solution to the equation $E\left(\psi_k\left(\frac{X-\mu}{\tau}\right)\right) = 0$, where τ is the median of the distribution of $|X - \nu|/0.67$ and ψ_k is defined in Problem 3.5.8. Show that $\mu^{(k)}$ is a location parameter.

(d) For $0 < \alpha < 1$, let $\nu_\alpha = \nu_\alpha(F) = \frac{1}{2}(x_\alpha + x_{1-\alpha})$, $\underline{\nu}(F) = \inf\{\nu_\alpha(F) : 0 < \alpha \le 1/2\}$ and $\bar{\nu}(F) = \sup\{\nu_\alpha(F) : 0 < \alpha \le 1/2\}$. Show that ν_α is a location parameter and show that any location parameter $\theta(F)$ satisfies $\underline{\nu}(F) \le \theta(F) \le \bar{\nu}(F)$.

Hint: For the second part, let $H(x)$ be the distribution function whose inverse is $H^{-1}(\alpha) = \frac{1}{2}[x_\alpha - x_{1-\alpha}]$, $0 < \alpha < 1$, and note that $H(x - \bar{\nu}(F)) \le F(x) \le H(x - \underline{\nu}(F))$. Also note that $H(x)$ is symmetric about zero.

(e) Show that if the support $S(F) = \{x : 0 < F(x) < 1\}$ of F is a finite interval, then $\underline{\nu}(F)$ and $\bar{\nu}(F)$ are location parameters. ($[\underline{\nu}(F), \bar{\nu}(F)]$ is the *location parameter set* in the sense that for any continuous F the value $\theta(F)$ of any location parameter must be in $[\underline{\nu}(F), \bar{\nu}(F)]$ and, if F is also strictly increasing, any point in $[\underline{\nu}(F), \bar{\nu}(F)]$ is the value of some location parameter.)

14. An estimate $\widehat{\theta}_n$ is said to be *shift and scale equivariant* if for all $x_1, \ldots, x_n, a, b > 0$,

$$\widehat{\theta}_n(a + bx_1, \ldots, a + bx_n) = a + b\widehat{\theta}_n(x_1, \ldots, x_n).$$

(a) Show that the sample mean, sample median, and sample trimmed mean are shift and scale equivariant.

(b) Write the SC as $SC(x, \widehat{\theta}, \mathbf{x}_{n-1})$ to show its dependence on $\mathbf{x}_{n-1} = (x_1, \ldots, x_{n-1})$. Show that if $\widehat{\theta}$ is shift and location equivariant, then for $a \in R$, $b > 0$, $c \in R$, $d > 0$,

$$SC(a + bx, c + d\widehat{\theta}, a + b\mathbf{x}_{n-1}) = bdSC(x, \widehat{\theta}, \mathbf{x}_{n-1}).$$

That is, the SC is shift invariant and scale equivariant.

15. In Remark 3.5.1:

(a) Show that $SC(x, \widehat{\theta}) = IF_{\frac{1}{n}}(x; \theta, \widehat{F}_{n-1})$.

(b) In the following cases, compare $SC(x; \theta, F) \equiv \lim_{n\to\infty} SC(x, \widehat{\theta})$ and $IF(x; \theta, F)$.

(i) $\theta(F) = \mu_F = \int x dF(x)$.

(ii) $\theta(F) = \sigma_F^2 = \int (x - \mu_F)^2 dF(x)$.

(iii) $\theta(F) = x_\alpha$. Assume that F is strictly increasing.

(c) Does $n^{-\frac{1}{2}}[SC(x, \widehat{\theta}) - IF(x, \theta, F)] \xrightarrow{P} 0$ in the cases (i), (ii), and (iii) preceding?

16. Show that in the bisection method, in order to be certain that the Jth iterate $\widehat{\theta}_J$ is within ϵ of the desired $\widehat{\theta}$ such that $\psi(\widehat{\theta}) = 0$, we in general must take on the order of $\log \frac{1}{\epsilon}$ steps. This is, consequently, also true of the method of coordinate ascent.

17. Let $d = 1$ and suppose that ψ is twice continuously differentiable, $\psi' > 0$, and we seek the unique solution $\widehat{\theta}$ of $\psi(\theta) = 0$. The Newton–Raphson method in this case is

$$\widehat{\theta}^{(j)} = \widehat{\theta}^{(j-1)} - \frac{\psi(\widehat{\theta}^{(j-1)})}{\psi'(\widehat{\theta}^{(j-1)})}.$$

(a) Show by example that for suitable ψ and $|\widehat{\theta}^{(0)} - \widehat{\theta}|$ large enough, $\{\widehat{\theta}^{(j)}\}$ do not converge.

(b) Show that there exists, $C < \infty$, $\delta > 0$ (depending on ψ) such that if $|\widehat{\theta}^{(0)} - \widehat{\theta}| \leq \delta$, then $|\widehat{\theta}^{(j)} - \widehat{\theta}| \leq C|\widehat{\theta}^{(j-1)} - \widehat{\theta}|^2$.

Hint: (a) Try $\psi(x) = A \log x$ with $A > 1$.

(b)

$$|\widehat{\theta}^{(j)} - \widehat{\theta}| = \left| \widehat{\theta}^{(j-1)} - \widehat{\theta} - \frac{1}{\psi'(\widehat{\theta}^{(j-1)})}(\psi(\widehat{\theta}^{(j-1)}) - \psi(\widehat{\theta})) \right|.$$

18. In the gross error model (3.5.2), show that

(a) If h is a density that is symmetric about zero, then μ is identifiable.

(b) If no assumptions are made about h, then μ is not identifiable.

3.7 NOTES

Note for Section 3.3

(1) A technical problem is to give the class $\mathcal{S}$ of subsets of $\mathcal{F}$ for which we can assign probability (the measurable sets). We define $\mathcal{S}$ as the σ-field generated by $\mathcal{S}_{A,B} = \{F \in \mathcal{F} : P_F(A) \in B\}$, $A, B \in \mathcal{B}$, where $\mathcal{B}$ is the class of Borel sets.

Notes for Section 3.4

(1) The result of Theorem 3.4.1 is commonly known as the Cramér–Rao inequality. Because priority of discovery is now given to the French mathematician M. Frêchet, we shall

follow the lead of Lehmann and call the inequality after the Fisher information number that appears in the statement.

(2) Note that this inequality is true but uninteresting if $I(\theta) = \infty$ (and $\psi'(\theta)$ is finite) or if $\text{Var}_\theta(T(X)) = \infty$.

(3) The continuity of the first integral ensures that

$$\frac{\partial}{\partial\theta}\left[\int_{\theta_0}^{\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}T(x)\frac{\partial}{\partial\lambda}p(x,\lambda)dxd\lambda\right] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}T(x)\left[\frac{\partial}{\partial\theta}p(x,\theta)\right]dx$$

for all θ whereas the continuity (or even boundedness on compact sets) of the second integral guarantees that we can interchange the order of integration in

$$\int_{\theta_0}^{\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}T(x)\left[\frac{\partial}{\partial\lambda}p(x,\lambda)\right]dxd\lambda.$$

(4) The finiteness of $\text{Var}_\theta(T(X))$ and $I(\theta)$ imply that $\psi'(\theta)$ is finite by the covariance interpretation given in $(3.4.8)$.

3.8 REFERENCES

ANDREWS, D. F., P. J. BICKEL, F. R. HAMPEL, P. J. HUBER, W. H. ROGERS, AND J. W. TUKEY, *Robust Estimates of Location: Survey and Advances* Princeton, NJ: Princeton University Press, 1972.

APOSTOL, T. M., *Mathematical Analysis*, 2nd ed. Reading, MA: Addison–Wesley, 1974.

BERGER, J. O., *Statistical Decision Theory and Bayesian Analysis* New York: Springer, 1985.

BERNARDO, J. M., AND A. F. M. SMITH, *Bayesian Theory* New York: Wiley, 1994.

BICKEL, P., AND E. LEHMANN, "Unbiased Estimation in Convex Families," *Ann. Math. Statist., 40*, 1523–1535 (1969).

BICKEL, P., AND E. LEHMANN, "Descriptive Statistics for Nonparametric Models. I. Introduction," *Ann. Statist., 3*, 1038–1044 (1975a).

BICKEL, P., AND E. LEHMANN, "Descriptive Statistics for Nonparametric Models. II. Location," *Ann. Statist., 3*, 1045–1069 (1975b).

BICKEL, P., AND E. LEHMANN, "Descriptive Statistics for Nonparametric Models. III. Dispersion," *Ann. Statist., 4*, 1139–1158 (1976).

BÜHLMANN, H., *Mathematical Methods in Risk Theory* Heidelberg: Springer Verlag, 1970.

DAHLQUIST, G., A. BJÖRK, AND N. ANDERSON, *Numerical Analysis* New York: Prentice Hall, 1974.

DE GROOT, M. H., *Optimal Statistical Decisions* New York: McGraw–Hill, 1969.

DOKSUM, K. A., "Measures of Location and Asymmetry," *Scand. J. of Statist., 2*, 11–22 (1975).

DOWE, D. L., R. A. BAXTER, J. J. OLIVER, AND C. S. WALLACE, *Point Estimation Using the Kullback–Leibler Loss Function and MML*, in *Proceedings of the Second Pacific Asian Conference on Knowledge Discovery and Data Mining* Melbourne: Springer–Verlag, 1998.

HAMPEL, F., "The Influence Curve and Its Role in Robust Estimation," *J. Amer. Statist. Assoc., 69*, 383–393 (1974).

HAMPEL, F., E. RONCHETTI, P. ROUSSEUW, AND W. STAHEL, *Robust Statistics: The Approach Based on Influence Functions* New York: J. Wiley & Sons, 1986.

HANSEN, M. H., AND B. YU, "Model Selection and the Principle of Minimum Description Length," *J. Amer. Statist. Assoc.*, **96**, 746–774 (2001).

HOGG, R., "Adaptive Robust Procedures," *J. Amer. Statist. Assoc., 69*, 909–927 (1974).

HUBER, P., *Robust Statistics* New York: Wiley, 1981.

HUBER, P., "Robust Statistics: A Review," *Ann. Math. Statist., 43*, 1041–1067 (1972).

JAECKEL, L. A., "Robust Estimates of Location," *Ann. Math. Statist., 42*, 1020–1034 (1971).

JEFFREYS, H., *Theory of Probability*, 2nd ed. London: Oxford University Press, 1948.

KARLIN, S., *Mathematical Methods and Theory in Games, Programming, and Economics* Reading, MA: Addison–Wesley, 1959.

LEHMANN, E. L., *Testing Statistical Hypotheses* New York: Springer, 1986.

LEHMANN, E. L., AND G. CASELLA, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

LINDLEY, D. V., *Introduction to Probability and Statistics from a Bayesian Point of View*, Part I: *Probability*; Part II: *Inference*, Cambridge University Press, London, 1965.

LINDLEY, D.V., "Decision Analysis and Bioequivalence Trials," *Statistical Science, 13*, 136–141 (1998).

NORBERG, R., "Hierarchical Credibility: Analysis of a Random Effect Linear Model with Nested Classification," *Scand. Actuarial J.*, 204–222 (1986).

RISSANEN, J., "Stochastic Complexity (With Discussions)," *J. Royal Statist. Soc. B, 49*, 223–239 (1987).

SAVAGE, L. J., *The Foundations of Statistics* New York: J. Wiley & Sons, 1954.

SHIBATA, R., "Boostrap Estimate of Kullback–Leibler Information for Model Selection," *Statistica Sinica, 7*, 375–394 (1997).

STEIN, C., "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution," *Proc. Third Berkeley Symposium on Math. Statist. and Probability, 1*, University of California Press, 197–206 (1956).

TUKEY, J. W., *Exploratory Data Analysis* Reading, MA: Addison–Wesley, 1972.

WALLACE, C. S., AND P. R. FREEMAN, "Estimation and Inference by Compact Coding (With Discussions)," *J. Royal Statist. Soc. B, 49*, 240–251 (1987).

WIJSMAN, R. A., "On the Attainment of the Cramér–Rao Lower Bound," *Ann. Math. Statist., 1*, 538–542 (1973).