12
Statistical Decision Theory

12.1 Preliminaries

We have considered several point estimators such as the maximum likelihood estimator, the method of moments estimator, and the posterior mean. In fact, there are many other ways to generate estimators. How do we choose among them? The answer is found in **decision theory** which is a formal theory for comparing statistical procedures.

Consider a parameter θ which lives in a parameter space Θ. Let $\widehat{\theta}$ be an estimator of θ. In the language of decision theory, an estimator is sometimes called a **decision rule** and the possible values of the decision rule are called **actions**.

We shall measure the discrepancy between θ and $\widehat{\theta}$ using a **loss function** $L(\theta, \widehat{\theta})$. Formally, L maps $\Theta \times \Theta$ into $\mathbb{R}$. Here are some examples of loss functions:

$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$	squared error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert$	absolute error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert^p$	L_p loss,
$L(\theta, \widehat{\theta}) = 0$ if $\theta = \widehat{\theta}$ or 1 if $\theta \neq \widehat{\theta}$	zero–one loss,
$L(\theta, \widehat{\theta}) = \int \log\left(\frac{f(x;\theta)}{f(x;\widehat{\theta})}\right) f(x;\theta)dx$	Kullback–Leibler loss.

Bear in mind in what follows that an estimator $\widehat{\theta}$ is a function of the data. To emphasize this point, sometimes we will write $\widehat{\theta}$ as $\widehat{\theta}(X)$. To assess an estimator, we evaluate the average loss or risk.

12.1 Definition. *The* **risk** *of an estimator* $\widehat{\theta}$ *is*

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\left(L(\theta, \widehat{\theta})\right) = \int L(\theta, \widehat{\theta}(x)) f(x; \theta) dx.$$

When the loss function is squared error, the risk is just the MSE (mean squared error):

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \text{MSE} = \mathbb{V}_\theta(\widehat{\theta}) + \mathsf{bias}^2_\theta(\widehat{\theta}).$$

In the rest of the chapter, if we do not state what loss function we are using, assume the loss function is squared error.

12.2 Comparing Risk Functions

To compare two estimators we can compare their risk functions. However, this does not provide a clear answer as to which estimator is better. Consider the following examples.

12.2 Example. Let $X \sim N(\theta, 1)$ and assume we are using squared error loss. Consider two estimators: $\widehat{\theta}_1 = X$ and $\widehat{\theta}_2 = 3$. The risk functions are $R(\theta, \widehat{\theta}_1) = \mathbb{E}_\theta(X - \theta)^2 = 1$ and $R(\theta, \widehat{\theta}_2) = \mathbb{E}_\theta(3 - \theta)^2 = (3 - \theta)^2$. If $2 < \theta < 4$ then $R(\theta, \widehat{\theta}_2) < R(\theta, \widehat{\theta}_1)$, otherwise, $R(\theta, \widehat{\theta}_1) < R(\theta, \widehat{\theta}_2)$. Neither estimator uniformly dominates the other; see Figure 12.1. ■

12.3 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Consider squared error loss and let $\widehat{p}_1 = \overline{X}$. Since this has 0 bias, we have that

$$R(p, \widehat{p}_1) = \mathbb{V}(\overline{X}) = \frac{p(1-p)}{n}.$$

Another estimator is

$$\widehat{p}_2 = \frac{Y + \alpha}{\alpha + \beta + n}$$

where $Y = \sum_{i=1}^n X_i$ and α and β are positive constants. This is the posterior mean using a Beta (α, β) prior. Now,

$$R(p, \widehat{p}_2) \quad = \quad \mathbb{V}_p(\widehat{p}_2) + (\mathsf{bias}_p(\widehat{p}_2))^2$$

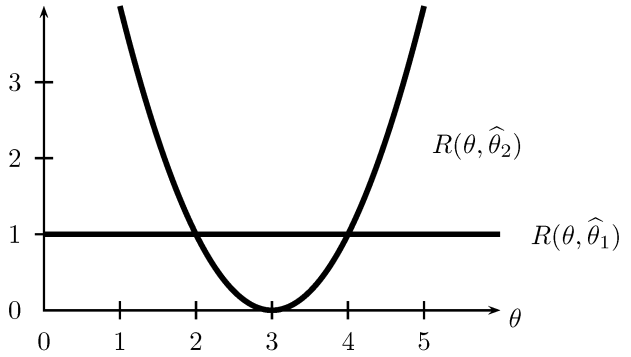

FIGURE 12.1. Comparing two risk functions. Neither risk function dominates the other at all values of θ.

$$= \mathbb{V}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2$$
$$= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2.$$

Let $\alpha = \beta = \sqrt{n/4}$. (In Example 12.12 we will explain this choice.) The resulting estimator is

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and the risk function is

$$R(p, \widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

The risk functions are plotted in figure 12.2. As we can see, neither estimator uniformly dominates the other.

These examples highlight the need to be able to compare risk functions. To do so, we need a one-number summary of the risk function. Two such summaries are the maximum risk and the Bayes risk.

12.4 Definition. *The* **maximum risk** *is*

$$\overline{R}(\widehat{\theta}) = \sup_\theta R(\theta, \widehat{\theta}) \tag{12.1}$$

and **the Bayes risk** *is*

$$r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta) d\theta \tag{12.2}$$

where $f(\theta)$ is a prior for θ.

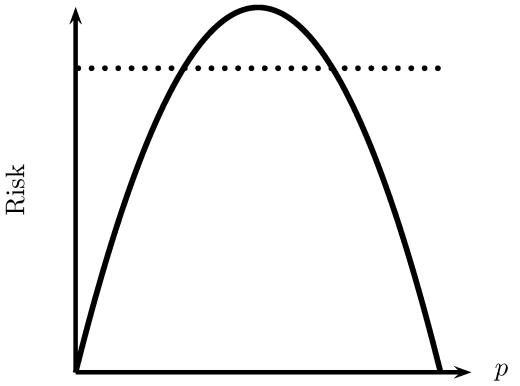

FIGURE 12.2. Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 12.3. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

12.5 Example. Consider again the two estimators in Example 12.3. We have

$$\overline{R}(\widehat{p}_1) = \max_{0\leq p\leq 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

and

$$\overline{R}(\widehat{p}_2) = \max_{p} \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}.$$

Based on maximum risk, $\widehat{p}_2$ is a better estimator since $\overline{R}(\widehat{p}_2) < \overline{R}(\widehat{p}_1)$. However, when n is large, $\overline{R}(\widehat{p}_1)$ has smaller risk except for a small region in the parameter space near $p = 1/2$. Thus, many people prefer $\widehat{p}_1$ to $\widehat{p}_2$. This illustrates that one-number summaries like maximum risk are imperfect. Now consider the Bayes risk. For illustration, let us take $f(p) = 1$. Then

$$r(f, \widehat{p}_1) = \int R(p, \widehat{p}_1) dp = \int \frac{p(1-p)}{n} dp = \frac{1}{6n}$$

and

$$r(f, \widehat{p}_2) = \int R(p, \widehat{p}_2) dp = \frac{n}{4(n+\sqrt{n})^2}.$$

For $n \geq 20$, $r(f, \widehat{p}_2) > r(f, \widehat{p}_1)$ which suggests that $\widehat{p}_1$ is a better estimator. This might seem intuitively reasonable but this answer depends on the choice of prior. The advantage of using maximum risk, despite its problems, is that it does not require one to choose a prior. ■

These two summaries of the risk function suggest two different methods for devising estimators: choosing $\widehat{\theta}$ to minimize the maximum risk leads to

minimax estimators; choosing $\widehat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

12.6 Definition. *A decision rule that minimizes the Bayes risk is called a* **Bayes rule**. *Formally, $\widehat{\theta}$ is a Bayes rule with respect to the prior f if*

$$r(f, \widehat{\theta}) = \inf_{\widetilde{\theta}} r(f, \widetilde{\theta}) \tag{12.3}$$

where the infimum is over all estimators $\widetilde{\theta}$. An estimator that minimizes the maximum risk is called a **minimax rule**. *Formally, $\widehat{\theta}$ is minimax if*

$$\sup_{\theta} R(\theta, \widehat{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta} R(\theta, \widetilde{\theta}) \tag{12.4}$$

where the infimum is over all estimators $\widetilde{\theta}$.

12.3 Bayes Estimators

Let f be a prior. From Bayes' theorem, the posterior density is

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{m(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \tag{12.5}$$

where $m(x) = \int f(x, \theta)d\theta = \int f(x|\theta)f(\theta)d\theta$ is the **marginal distribution** of X. Define the **posterior risk** of an estimator $\widehat{\theta}(x)$ by

$$r(\widehat{\theta}|x) = \int L(\theta, \widehat{\theta}(x))f(\theta|x)d\theta. \tag{12.6}$$

12.7 Theorem. *The Bayes risk $r(f, \widehat{\theta})$ satisfies*

$$r(f, \widehat{\theta}) = \int r(\widehat{\theta}|x)m(x)\, dx.$$

Let $\widehat{\theta}(x)$ be the value of θ that minimizes $r(\widehat{\theta}|x)$. Then $\widehat{\theta}$ is the Bayes estimator.

Proof. We can rewrite the Bayes risk as follows:

$$\begin{aligned} r(f, \widehat{\theta}) &= \int R(\theta, \widehat{\theta})f(\theta)d\theta = \int \left(\int L(\theta, \widehat{\theta}(x))f(x|\theta)dx \right) f(\theta)d\theta \\ &= \int \int L(\theta, \widehat{\theta}(x))f(x, \theta)dxd\theta = \int \int L(\theta, \widehat{\theta}(x))f(\theta|x)m(x)dxd\theta \\ &= \int \left(\int L(\theta, \widehat{\theta}(x))f(\theta|x)d\theta \right) m(x)\, dx = \int r(\widehat{\theta}|x)m(x)\, dx. \end{aligned}$$

If we choose $\widehat{\theta}(x)$ to be the value of θ that minimizes $r(\widehat{\theta}|x)$ then we will minimize the integrand at every x and thus minimize the integral $\int r(\widehat{\theta}|x)m(x)dx$. ■

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

12.8 Theorem. *If $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x) = \int \theta f(\theta|x)d\theta = \mathbb{E}(\theta|X = x). \tag{12.7}$$

If $L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$ then the Bayes estimator is the median of the posterior $f(\theta|x)$. If $L(\theta, \widehat{\theta})$ is zero–one loss, then the Bayes estimator is the mode of the posterior $f(\theta|x)$.

Proof. We will prove the theorem for squared error loss. The Bayes rule $\widehat{\theta}(x)$ minimizes $r(\widehat{\theta}|x) = \int(\theta - \widehat{\theta}(x))^2 f(\theta|x)d\theta$. Taking the derivative of $r(\widehat{\theta}|x)$ with respect to $\widehat{\theta}(x)$ and setting it equal to 0 yields the equation $2\int(\theta - \widehat{\theta}(x))f(\theta|x)d\theta = 0$. Solving for $\widehat{\theta}(x)$ we get 12.7. ■

12.9 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is known. Suppose we use a $N(a, b^2)$ prior for μ. The Bayes estimator with respect to squared error loss is the posterior mean, which is

$$\widehat{\theta}(X_1, \ldots, X_n) = \frac{b^2}{b^2 + \frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{b^2 + \frac{\sigma^2}{n}}a. \quad ■$$

12.4 Minimax Rules

Finding minimax rules is complicated and we cannot attempt a complete coverage of that theory here but we will mention a few key results. The main message to take away from this section is: Bayes estimators with a constant risk function are minimax.

12.10 Theorem. *Let $\widehat{\theta}^f$ be the Bayes rule for some prior f:*

$$r(f, \widehat{\theta}^f) = \inf_{\widehat{\theta}} r(f, \widehat{\theta}). \tag{12.8}$$

Suppose that

$$R(\theta, \widehat{\theta}^f) \leq r(f, \widehat{\theta}^f) \text{ for all } \theta. \tag{12.9}$$

Then $\widehat{\theta}^f$ is minimax and f is called a **least favorable prior**.

PROOF. Suppose that $\widehat{\theta}^f$ is not minimax. Then there is another rule $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}^f)$. Since the average of a function is always less than or equal to its maximum, we have that $r(f, \widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$r(f, \widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}^f) \leq r(f, \widehat{\theta}^f)$$

which contradicts (12.8). ■

12.11 Theorem. *Suppose that $\widehat{\theta}$ is the Bayes rule with respect to some prior f. Suppose further that $\widehat{\theta}$ has constant risk: $R(\theta, \widehat{\theta}) = c$ for some c. Then $\widehat{\theta}$ is minimax.*

PROOF. The Bayes risk is $r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta) d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq r(f, \widehat{\theta})$ for all θ. Now apply the previous theorem. ■

12.12 Example. Consider the Bernoulli model with squared error loss. In example 12.3 we showed that the estimator

$$\widehat{p}(X^n) = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}$$

has a constant risk function. This estimator is the posterior mean, and hence the Bayes rule, for the prior Beta(α, β) with $\alpha = \beta = \sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax. ■

12.13 Example. Consider again the Bernoulli but with loss function

$$L(p, \widehat{p}) = \frac{(p - \widehat{p})^2}{p(1-p)}.$$

Let

$$\widehat{p}(X^n) = \widehat{p} = \frac{\sum_{i=1}^n X_i}{n}.$$

The risk is

$$R(p, \widehat{p}) = E\left(\frac{(\widehat{p} - p)^2}{p(1-p)}\right) = \frac{1}{p(1-p)}\left(\frac{p(1-p)}{n}\right) = \frac{1}{n}$$

which, as a function of p, is constant. It can be shown that, for this loss function, $\widehat{p}(X^n)$ is the Bayes estimator under the prior $f(p) = 1$. Hence, $\widehat{p}$ is minimax. ■

A natural question to ask is: what is the minimax estimator for a Normal model?

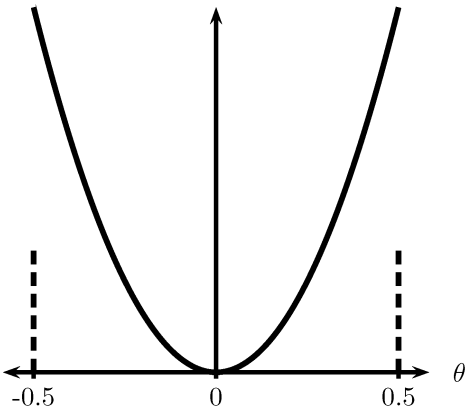

FIGURE 12.3. Risk function for constrained Normal with m=.5. The two short dashed lines show the least favorable prior which puts its mass at two points.

12.14 Theorem. *Let* $X_1, \ldots, X_n \sim N(\theta, 1)$ *and let* $\widehat{\theta} = \overline{X}$. *Then* $\widehat{\theta}$ *is minimax with respect to any well-behaved loss function.* [1] *It is the only estimator with this property.*

If the parameter space is restricted, then the theorem above does not apply as the next example shows.

12.15 Example. Suppose that $X \sim N(\theta, 1)$ and that θ is known to lie in the interval $[-m, m]$ where $0 < m < 1$. The unique, minimax estimator under squared error loss is

$$\widehat{\theta}(X) = m \tanh(mX)$$

where $\tanh(z) = (e^z - e^{-z})/(e^z + e^{-z})$. It can be shown that this is the Bayes rule with respect to the prior that puts mass $1/2$ at m and mass $1/2$ at $-m$. Moreover, it can be shown that the risk is not constant but it does satisfy $R(\theta, \widehat{\theta}) \leq r(f, \widehat{\theta})$ for all θ; see Figure 12.3. Hence, Theorem 12.10 implies that $\widehat{\theta}$ is minimax. ■

[1] "Well-behaved" means that the level sets must be convex and symmetric about the origin. The result holds up to sets of measure 0.

12.5 Maximum Likelihood, Minimax, and Bayes

For parametric models that satisfy weak regularity conditions, the maximum likelihood estimator is approximately minimax. Consider squared error loss which is squared bias plus variance. In parametric models with large samples, it can be shown that the variance term dominates the bias so the risk of the MLE $\widehat{\theta}$ roughly equals the variance:[2]

$$R(\theta, \widehat{\theta}) = \mathbb{V}_\theta(\widehat{\theta}) + \mathsf{bias}^2 \approx \mathbb{V}_\theta(\widehat{\theta}).$$

As we saw in Chapter 9, the variance of the MLE is approximately

$$\mathbb{V}(\widehat{\theta}) \approx \frac{1}{nI(\theta)}$$

where $I(\theta)$ is the Fisher information. Hence,

$$nR(\theta, \widehat{\theta}) \approx \frac{1}{I(\theta)}. \tag{12.10}$$

For any other estimator θ', it can be shown that for large n, $R(\theta, \theta') \geq R(\theta, \widehat{\theta})$. More precisely,

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} \sup_{|\theta - \theta'| < \epsilon} n\, R(\theta', \widehat{\theta}) \geq \frac{1}{I(\theta)}. \tag{12.11}$$

This says that, in a local, large sample sense, the MLE is minimax. It can also be shown that the MLE is approximately the Bayes rule.

In summary:

> In most parametric models, with large samples, the MLE is approximately minimax and Bayes.

There is a caveat: these results break down when the number of parameters is large as the next example shows.

12.16 Example (Many Normal means)**.** Let $Y_i \sim N(\theta_i, \sigma^2/n)$, $i = 1, \ldots, n$. Let $Y = (Y_1, \ldots, Y_n)$ denote the data and let $\theta = (\theta_1, \ldots, \theta_n)$ denote the unknown parameters. Assume that

$$\theta \in \Theta_n \equiv \left\{ (\theta_1, \ldots, \theta_n) : \sum_{i=1}^n \theta_i^2 \leq c^2 \right\}$$

[2] Typically, the squared bias is order $O(n^{-2})$ while the variance is of order $O(n^{-1})$.

for some $c > 0$. In this model, there are as many parameters as observations. [3] The MLE is $\widehat{\theta} = Y = (Y_1, \ldots, Y_n)$. Under the loss function $L(\theta, \widehat{\theta}) = \sum_{i=1}^n (\widehat{\theta}_i - \theta_i)^2$, the risk of the MLE is $R(\theta, \widehat{\theta}) = \sigma^2$. It can be shown that the minimax risk is approximately $\sigma^2/(\sigma^2 + c^2)$ and one can find an estimator $\widetilde{\theta}$ that achieves this risk. Since $\sigma^2/(\sigma^2 + c^2) < \sigma^2$, we see that $\widetilde{\theta}$ has smaller risk than the MLE. In practice, the difference between the risks can be substantial. This shows that maximum likelihood is not an optimal estimator in high dimensional problems. ■

12.6 Admissibility

Minimax estimators and Bayes estimators are "good estimators" in the sense that they have small risk. It is also useful to characterize bad estimators.

12.17 Definition. *An estimator $\widehat{\theta}$ is* **inadmissible** *if there exists another rule $\widehat{\theta}'$ such that*

$$\begin{aligned} R(\theta, \widehat{\theta}') &\leq R(\theta, \widehat{\theta}) \text{ for all } \theta \text{ and} \\ R(\theta, \widehat{\theta}') &< R(\theta, \widehat{\theta}) \text{ for at least one } \theta. \end{aligned}$$

Otherwise, $\widehat{\theta}$ is **admissible**.

12.18 Example. Let $X \sim N(\theta, 1)$ and consider estimating θ with squared error loss. Let $\widehat{\theta}(X) = 3$. We will show that $\widehat{\theta}$ is admissible. Suppose not. Then there exists a different rule $\widehat{\theta}'$ with smaller risk. In particular, $R(3, \widehat{\theta}') \leq R(3, \widehat{\theta}) = 0$. Hence, $0 = R(3, \widehat{\theta}') = \int (\widehat{\theta}'(x) - 3)^2 f(x; 3) dx$. Thus, $\widehat{\theta}'(x) = 3$. So there is no rule that beats $\widehat{\theta}$. Even though $\widehat{\theta}$ is admissible it is clearly a bad decision rule. ■

12.19 Theorem (Bayes Rules Are Admissible)**.** *Suppose that $\Theta \subset \mathbb{R}$ and that $R(\theta, \widehat{\theta})$ is a continuous function of θ for every $\widehat{\theta}$. Let f be a prior density with full support, meaning that, for every θ and every $\epsilon > 0$, $\int_{\theta-\epsilon}^{\theta+\epsilon} f(\theta) d\theta > 0$. Let $\widehat{\theta}^f$ be the Bayes' rule. If the Bayes risk is finite then $\widehat{\theta}^f$ is admissible.*

PROOF. Suppose $\widehat{\theta}^f$ is inadmissible. Then there exists a better rule $\widehat{\theta}$ such that $R(\theta, \widehat{\theta}) \leq R(\theta, \widehat{\theta}^f)$ for all θ and $R(\theta_0, \widehat{\theta}) < R(\theta_0, \widehat{\theta}^f)$ for some θ_0. Let

[3] The many Normal means problem is more general than it looks. Many nonparametric estimation problems are mathematically equivalent to this model.

$\nu = R(\theta_0, \widehat{\theta}^f) - R(\theta_0, \widehat{\theta}) > 0$. Since R is continuous, there is an $\epsilon > 0$ such that $R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) > \nu/2$ for all $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$. Now,

$$\begin{aligned} r(f, \widehat{\theta}^f) - r(f, \widehat{\theta}) &= \int R(\theta, \widehat{\theta}^f) f(\theta) d\theta - \int R(\theta, \widehat{\theta}) f(\theta) d\theta \\ &= \int \left[R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) \right] f(\theta) d\theta \\ &\geq \int_{\theta_0 - \epsilon}^{\theta_0 + \epsilon} \left[R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) \right] f(\theta) d\theta \\ &\geq \frac{\nu}{2} \int_{\theta_0 - \epsilon}^{\theta_0 + \epsilon} f(\theta) d\theta \\ &> 0. \end{aligned}$$

Hence, $r(f, \widehat{\theta}^f) > r(f, \widehat{\theta})$. This implies that $\widehat{\theta}^f$ does not minimize $r(f, \widehat{\theta})$ which contradicts the fact that $\widehat{\theta}^f$ is the Bayes rule. ∎

12.20 Theorem. *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Under squared error loss, $\overline{X}$ is admissible.*

The proof of the last theorem is quite technical and is omitted but the idea is as follows: The posterior mean is admissible for any strictly positive prior. Take the prior to be $N(a, b^2)$. When b^2 is very large, the posterior mean is approximately equal to $\overline{X}$.

How are minimaxity and admissibility linked? In general, a rule may be one, both, or neither. But here are some facts linking admissibility and minimaxity.

12.21 Theorem. *Suppose that $\widehat{\theta}$ has constant risk and is admissible. Then it is minimax.*

Proof. The risk is $R(\theta, \widehat{\theta}) = c$ for some c. If $\widehat{\theta}$ were not minimax then there exists a rule $\widehat{\theta}'$ such that

$$R(\theta, \widehat{\theta}') \leq \sup_\theta R(\theta, \widehat{\theta}') < \sup_\theta R(\theta, \widehat{\theta}) = c.$$

This would imply that $\widehat{\theta}$ is inadmissible. ∎

Now we can prove a restricted version of Theorem 12.14 for squared error loss.

12.22 Theorem. *Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Then, under squared error loss, $\widehat{\theta} = \overline{X}$ is minimax.*

Proof. According to Theorem 12.20, $\widehat{\theta}$ is admissible. The risk of $\widehat{\theta}$ is $1/n$ which is constant. The result follows from Theorem 12.21. ∎

Although minimax rules are not guaranteed to be admissible they are "close to admissible." Say that $\widehat{\theta}$ is **strongly inadmissible** if there exists a rule $\widehat{\theta}'$ and an $\epsilon > 0$ such that $R(\theta, \widehat{\theta}') < R(\theta, \widehat{\theta}) - \epsilon$ for all θ.

12.23 Theorem. *If $\widehat{\theta}$ is minimax, then it is not strongly inadmissible.*

12.7 Stein's Paradox

Suppose that $X \sim N(\theta, 1)$ and consider estimating θ with squared error loss. From the previous section we know that $\widehat{\theta}(X) = X$ is admissible. Now consider estimating two, unrelated quantities $\theta = (\theta_1, \theta_2)$ and suppose that $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$ independently, with loss $L(\theta, \widehat{\theta}) = \sum_{j=1}^{2}(\theta_j - \widehat{\theta}_j)^2$. Not surprisingly, $\widehat{\theta}(X) = X$ is again admissible where $X = (X_1, X_2)$. Now consider the generalization to k normal means. Let $\theta = (\theta_1, \ldots, \theta_k)$, $X = (X_1, \ldots, X_k)$ with $X_i \sim N(\theta_i, 1)$ (independent) and loss $L(\theta, \widehat{\theta}) = \sum_{j=1}^{k}(\theta_j - \widehat{\theta}_j)^2$. Stein astounded everyone when he proved that, if $k \geq 3$, then $\widehat{\theta}(X) = X$ is inadmissible. It can be shown that the **James-Stein estimator** $\widehat{\theta}^S$ has smaller risk, where $\widehat{\theta}^S = (\widehat{\theta}_1^S, \ldots, \widehat{\theta}_k^S)$,

$$\widehat{\theta}_i^S(X) = \left(1 - \frac{k-2}{\sum_i X_i^2}\right)^+ X_i \tag{12.12}$$

and $(z)^+ = \max\{z, 0\}$. This estimator shrinks the X_i's towards 0. The message is that, when estimating many parameters, there is great value in shrinking the estimates. This observation plays an important role in modern nonparametric function estimation.

12.8 Bibliographic Remarks

Aspects of decision theory can be found in Casella and Berger (2002), Berger (1985), Ferguson (1967), and Lehmann and Casella (1998).

12.9 Exercises

1. In each of the following models, find the Bayes risk and the Bayes estimator, using squared error loss.

   (a) $X \sim \text{Binomial}(n, p)$, $p \sim \text{Beta}(\alpha, \beta)$.

(b) $X \sim \text{Poisson}(\lambda)$, $\lambda \sim \text{Gamma}(\alpha, \beta)$.

(c) $X \sim N(\theta, \sigma^2)$ where σ^2 is known and $\theta \sim N(a, b^2)$.

2. Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ and suppose we estimate θ with loss function $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2/\sigma^2$. Show that $\overline{X}$ is admissible and minimax.

3. Let $\Theta = \{\theta_1, \ldots, \theta_k\}$ be a finite parameter space. Prove that the posterior mode is the Bayes estimator under zero–one loss.

4. (Casella and Berger (2002).) Let $X_1, \ldots, X_n$ be a sample from a distribution with variance σ^2. Consider estimators of the form bS^2 where S^2 is the sample variance. Let the loss function for estimating σ^2 be

$$L(\sigma^2, \widehat{\sigma}^2) = \frac{\widehat{\sigma}^2}{\sigma^2} - 1 - \log\left(\frac{\widehat{\sigma}^2}{\sigma^2}\right).$$

Find the optimal value of b that minimizes the risk for all σ^2.

5. (Berliner (1983).) Let $X \sim \text{Binomial}(n, p)$ and suppose the loss function is

$$L(p, \widehat{p}) = \left(1 - \frac{\widehat{p}}{p}\right)^2$$

where $0 < p < 1$. Consider the estimator $\widehat{p}(X) = 0$. This estimator falls outside the parameter space $(0, 1)$ but we will allow this. Show that $\widehat{p}(X) = 0$ is the unique, minimax rule.

6. (Computer Experiment.) Compare the risk of the MLE and the James-Stein estimator (12.12) by simulation. Try various values of n and various vectors θ. Summarize your results.