We saw in Example 6.3.2 that this model was invariant. Furthermore, since $g_c(X_1, \ldots, X_n)$ is a random sample from a $\mathrm{n}(\mu + c, \sigma^2)$ population, the transformation $\overline{g_c}$ is $\overline{g_c}(\mu, \sigma^2) = (\mu + c, \sigma^2)$. Noting how the transformation $\overline{g_c}$ affects the parameter, the corresponding transformation of the action is $\tilde{g}_c(a) = a + c$. To verify that this is the correct transformation, note that

$$L((\mu, \sigma^2), a) = (\mu - a)^2 = (\mu + c - (a + c))^2 = L(\overline{g_c}(\mu, \sigma^2), \tilde{g}_c(a)).$$

Thus Definition 10.6.1 is verified and this problem is invariant under this group.

Any estimator that satisfies $T(x_1 + c, \ldots, x_n + c) = T(x_1, \ldots, x_n) + c$, for all c and for all $(x_1, \ldots, x_n)$, is an invariant estimator. To verify Definition 10.6.2 we note that

$$\begin{aligned} T(g_c(\boldsymbol{x})) &= T(x_1 + c, \ldots, x_n + c) \\ &= T(x_1, \ldots, x_n) + c \\ &= \tilde{g}_c(T(\boldsymbol{x})). \end{aligned}$$

Estimators that are invariant in this problem include the sample mean and the sample median. ||

Example 10.6.2: Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population. Consider estimating σ^2 using squared error loss. Use the scale group $\mathcal{G} = \{g_c(\boldsymbol{x})\colon 0 < c < \infty\}$ where

$$g_c(x_1, \ldots, x_n) = (cx_1, \ldots, cx_n).$$

Then $g_c(X_1, \ldots, X_n)$ is a random sample from a $\mathrm{n}(c\mu, c^2\sigma^2)$ population. Thus the model is invariant under this group and $\overline{g_c}(\mu, \sigma^2) = (c\mu, c^2\sigma^2)$. If this problem were invariant under this group, then there would be a $\tilde{g}_c$ such that

$$(\sigma^2 - a)^2 = L((\mu, \sigma^2), a) = L(\overline{g_c}(\mu, \sigma^2), \tilde{g}_c(a)) = (c^2\sigma^2 - \tilde{g}_c(a))^2.$$

This is true only if $\tilde{g}_c(a) = c^2\sigma^2 + \sigma^2 - a$ or $\tilde{g}_c(a) = c^2\sigma^2 - \sigma^2 + a$. But according to Definition 10.6.1, $\tilde{g}_c(a)$ can depend only on g_c and a, not σ^2. Thus this is not an invariant decision problem. ||

Example 10.6.3: Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population. Consider testing $H_0\colon \mu \le 0$ versus $H_1\colon \mu > 0$ using 0–1 loss. This problem is invariant under the scale transformation group defined in Example 10.6.2. Notice that according to the general decision theoretic definition of invariance, we need check only that the whole model is invariant under the group. We do not need to check that each subset of distributions, H_0 and H_1, is invariant as was required in Definition 8.2.4. The hypothesis testing invariance we discussed in Chapter 8 is a special case of the general notion of invariance we are now discussing.

Recall that $\overline{g_c}(\mu, \sigma^2) = (c\mu, c^2\sigma^2)$. Since $c > 0$, if $(\mu, \sigma^2) \in \Theta_0$, that is, $\mu \le 0$, then $\overline{g_c}(\mu, \sigma^2) \in \Theta_0$ and if $(\mu, \sigma^2) \in \Theta_0^c$ then $\overline{g_c}(\mu, \sigma^2) \in \Theta_0^c$. Thus action $a_i, i = 0$ or 1, is correct or incorrect for (μ, σ^2) and $\overline{g_c}(\mu, \sigma^2)$ simultaneously. This suggests that the only transformation of the sample space needed is the identity transformation, $\tilde{g}(a_i) = a_i, i = 0$ or 1. Unlike in the previous point estimation examples, here $\tilde{g}$ does not depend on which $g_c \in \mathcal{G}$ we are considering. To verify that this $\tilde{g}$ is correct and Definition 10.6.1 is satisfied, note that if $\mu \le 0$ then $c\mu \le 0$ so that

$$L((\mu, \sigma^2), a_0) = 0 = L(\overline{g_c}(\mu, \sigma^2), a_0) = L(\overline{g_c}(\mu, \sigma^2), \tilde{g}(a_0))$$

and

$$L((\mu, \sigma^2), a_1) = 1 = L(\overline{g_c}(\mu, \sigma^2), a_1) = L(\overline{g_c}(\mu, \sigma^2), \tilde{g}(a_1)).$$

Similar equalities hold if $\mu > 0$. Thus the conditions of Definition 10.6.1 are satisfied and the decision problem is invariant.

Let $\phi(\boldsymbol{x})$ be the test function for a test. Since $\tilde{g}$ is just the identity transformation, Definition 10.6.2 says that a test is invariant if

$$\phi(g_c(\boldsymbol{x})) = \tilde{g}(\phi(\boldsymbol{x})) = \phi(\boldsymbol{x}).$$

This matches Definition 8.2.3. Thus the tests considered in Example 8.2.6, tests that depend on the sample only through the statistic $\overline{X}/\sqrt{S^2/n}$, are invariant decision rules. But there are other hypothesis testing problems in which tests that are invariant according to Definition 10.6.2 are not invariant according to Definition 8.2.4. (See Exercise 10.38.) Thus the type of invariance introduced in this section provides a more general concept of invariance for hypothesis testing problems than that considered in Chapter 8. ||

10.7 Stein's Paradox

In this section we will consider a special multivariate estimation problem, one that has some rather counterintuitive features. The problem to be considered is that of estimating several normal means simultaneously and is actually a special case of the statistical problem considered in Chapter 11. An excellent introduction to Stein's paradox is given in Efron and Morris (1977).

Let $X_i \sim \mathrm{n}(\theta_i, 1), i = 1, \ldots, p$, where $p \ge 3$. (This restriction will be addressed later.) Assume $X_1, \ldots, X_p$ are mutually independent. Notice that the X_is are not iid. They come from normal populations with possibly different means, but the problems will be tied together in that there will be one loss function for the p problems. Formally, we want to estimate $\theta = (\theta_1, \ldots, \theta_p)$, using an estimator $\delta(\boldsymbol{X}) = (\delta_1(\boldsymbol{X}), \ldots, \delta_p(\boldsymbol{X}))$. The loss function is

$$L(\theta, \delta(\boldsymbol{X})) = \sum_{i=1}^{p} (\theta_i - \delta_i(\boldsymbol{X}))^2. \tag{10.7.1}$$

This loss function is the sum of squared error loss functions, but there is one important point to see. Each δ_i can be a function of $(X_1,\ldots,X_p)$, so all of the data can be used in estimating each mean. Since the X_is are independent, we might think that restricting δ_i just to be a function of X_i would be enough. However, the X_is are tied together in the loss function, and we will see that this matters.

The situation described here is not too farfetched and can be used as a model for a number of situations. For example, suppose a company needs to estimate average crop yield θ_i based on data X_i for a number of different crops in different places. Although each estimation problem is separate, they all affect the company. So it is reasonable for the loss function to tie them together. Realize that good estimation overall takes precedence over doing well in any particular problem. The results obtained in the simple model considered here have been obtained in much more generality; see Berger (1985) or Lehmann (1983) for some generalizations.

Using Exercise 10.30, we will now show that the estimator $\boldsymbol{X}=(X_1,\ldots,X_p)$ is minimax. (In more generality, the sample mean is minimax in the combined problem. See Exercise 10.39.) An estimator $\delta(\boldsymbol{X})=(\delta_1(\boldsymbol{X}),\ldots,\delta_p(\boldsymbol{X}))$ of the parameter $\theta=(\theta_1,\ldots,\theta_p)$, using the loss function (10.7.1), has risk function

$$R(\theta,\delta)=\mathrm{E}_\theta\left(\sum_{i=1}^{p}(\theta_i-\delta_i(\boldsymbol{X}))^2\right). \tag{10.7.2}$$

Furthermore, if $\pi(\theta)$ is a prior on θ, then the Bayes risk of an estimator is

$$B(\pi,\delta)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}R(\theta,\delta)\pi(\theta)\,d\theta_1\cdots d\theta_p. \tag{10.7.3}$$

Suppose now that we take a prior that is a product of independent priors, that is,

$$\pi(\theta)=\prod_{i=1}^{p}\pi_i(\theta_i),\quad \pi_i(\theta_i)\text{ a n}(0,\tau^2)\text{ pdf.}$$

For estimating θ_i with loss $(\theta_i-\delta_i(\boldsymbol{X}))^2$, the Bayes rule against π_i,δ_i^π, is

$$\delta_i^\pi(\boldsymbol{X})=\delta_i^\pi(X_i)=\frac{\tau^2}{\tau^2+1}X_i, \tag{10.7.4}$$

with Bayes risk

$$B(\pi_i,\delta_i^\pi(X_i))=\frac{\tau^2}{\tau^2+1}.$$

(These types of calculations are used in Section 10.4.3 and Exercise 10.8.) Since the priors are independent, it follows (see Exercise 10.10) that

$$\delta^\pi(\boldsymbol{X})=(\delta_1^\pi(X_1),\ldots,\delta_p^\pi(X_p))$$

is Bayes against the prior $\pi(\theta)=\prod_{i=1}^{p}\pi_i(\theta_i)$ using the loss (10.7.1). The Bayes risk is

$$\begin{aligned}B(\pi,\delta^\pi(\boldsymbol{X}))&=\sum_{i=1}^{p}\frac{\tau^2}{\tau^2+1}\\&=p\frac{\tau^2}{\tau^2+1}\\&\to p,\quad\text{as }\tau^2\to\infty,\\&=R(\theta,\boldsymbol{X}),\end{aligned}$$

and hence, by Exercise 10.30, $\boldsymbol{X}$ is minimax.

Even though $\boldsymbol{X}$ is minimax, $\boldsymbol{X}$ is not unique minimax and, since the risk of $\boldsymbol{X}$ is constant at the minimax value, any other minimax estimator will be better than $\boldsymbol{X}$. Unlike the one-dimensional problem where X is admissible, it is not admissible in higher dimensions (three or more).

This was established by Stein (1955) who showed that, if the dimension of the problem was at least three, then there exists a better procedure. (It was shown in Stein (1955) and James and Stein (1961) that the sample mean is admissible in one and two dimensions.) More importantly, in James and Stein (1961), a better estimator was exhibited. That seemingly nonintuitive estimator (but also see Exercise 10.40) is given by $\delta^S(\boldsymbol{X})=(\delta_1^S(\boldsymbol{X}),\ldots,\delta_p^S(\boldsymbol{X}))$, where

$$\delta_i^S(\boldsymbol{X})=\left(1-\frac{p-2}{\sum_{j=1}^{p}X_j^2}\right)X_i. \tag{10.7.5}$$

The original proof that δ^S dominates $\boldsymbol{X}$ is quite long and cumbersome, relying on representations of noncentral chi squared distributions. A more elegant and useful proof, however, was given by Stein using his Lemma (Stein, 1973, 1981). This use of Stein's Lemma, or more accurately, employment of integration by parts, was discovered independently by Berger (1975), who also used it to establish minimaxity of a class of estimators.

Recall Stein's Lemma, given in Chapter 4. If $X\sim\mathrm{n}(\theta,\sigma^2)$, then

$$\mathrm{E}(g(X)(X-\theta))=\sigma^2\mathrm{E}g'(X),$$

provided the expectations exist. Using this identity, computation of the risk of δ^S is relatively easy. We have

$$\begin{aligned}R(\theta,\delta^S)&=\mathrm{E}_\theta\left[\sum_{i=1}^{p}(\theta_i-\delta_i^S(\boldsymbol{X}))^2\right] &&\text{(definition of risk)}\\&=\sum_{i=1}^{p}\mathrm{E}_\theta\left[\theta_i-\delta_i^S(\boldsymbol{X})\right]^2 &&\text{(property of expectation)}\end{aligned}$$

$$= \sum_{i=1}^{p} \mathrm{E}_\theta \left[\theta_i - \left(1 - \frac{p-2}{\sum_{j=1}^{p} X_j^2} \right) X_i \right]^2 \qquad \left(\begin{array}{c} \text{definition} \\ \text{of estimator} \end{array} \right)$$

$$= \sum_{i=1}^{p} \mathrm{E}_\theta \left[(\theta_i - X_i) + \frac{p-2}{\sum_{j=1}^{p} X_j^2} X_i \right]^2$$

$$(10.7.6) \qquad = \sum_{i=1}^{p} \mathrm{E}_\theta (\theta_i - X_i)^2 + 2 \sum_{i=1}^{p} \mathrm{E}_\theta \left((\theta_i - X_i) \frac{p-2}{\sum_{j=1}^{p} X_j^2} X_i \right)$$

$$+ \sum_{i=1}^{p} \mathrm{E}_\theta \left(\frac{p-2}{\sum_{j=1}^{p} X_j^2} X_i \right)^2 . \qquad \text{(expand the square)}$$

The first expectation in (10.7.6) is equal to p, since it is the risk of $\boldsymbol{X}$, and simple manipulation will show that the third expectation is equal to $(p-2)^2 \mathrm{E}_\theta (1/\sum_{j=1}^{p} X_j^2)$. For the middle term we use Stein's Lemma:

$$\sum_{i=1}^{p} \mathrm{E}_\theta \left((\theta_i - X_i) \frac{p-2}{\sum_{j=1}^{p} X_j^2} X_i \right) = -(p-2) \sum_{i=1}^{p} \mathrm{E}_\theta \left(\frac{\partial}{\partial X_i} \frac{X_i}{\sum_{j=1}^{p} X_j^2} \right).$$

Differentiating and gathering terms gives

$$-(p-2) \sum_{i=1}^{p} \mathrm{E}_\theta \left(\frac{\partial}{\partial X_i} \frac{X_i}{\sum_{j=1}^{p} X_j^2} \right) = -(p-2) \sum_{i=1}^{p} \mathrm{E}_\theta \left(\frac{\sum_{j=1}^{p} X_j^2 - 2X_i^2}{(\sum_{j=1}^{p} X_j^2)^2} \right)$$

$$= -(p-2)^2 \mathrm{E}_\theta \left(\frac{1}{\sum_{j=1}^{p} X_j^2} \right).$$

Putting this all together, we have the risk of the Stein estimator to be

$$R(\theta, \delta^S) = p - 2(p-2)^2 \mathrm{E}_\theta \left(\frac{1}{\sum_{j=1}^{p} X_j^2} \right) + (p-2)^2 \mathrm{E}_\theta \left(\frac{1}{\sum_{j=1}^{p} X_j^2} \right)$$

$$= p - (p-2)^2 \mathrm{E}_\theta \left(\frac{1}{\sum_{j=1}^{p} X_j^2} \right)$$

$$< p = R(\theta, \boldsymbol{X}).$$

Thus the risk of δ^S is smaller than the risk of $\boldsymbol{X}$ and $\boldsymbol{X}$ is inadmissible. The above inequality is valid as long as the expectation exists, and the expectation exists as long as $p \geq 3$. If $p = 1$ or 2, $\mathrm{E}_\theta (1/\sum_{j=1}^{p} X_j^2) = \infty$.

The estimator δ^S is one of a family of estimators defined by

$$(10.7.7) \qquad \delta_i^c(\boldsymbol{X}) = \left(1 - \frac{c}{\sum_{j=1}^{p} X_j^2} \right) X_i, \quad i = 1, \ldots, p.$$

Any such estimator with $0 < c < 2(p-2)$ is better than $\boldsymbol{X}$, but the choice $c = p-2$ is optimal (see Exercise 10.41). These estimators, however, can also be uniformly improved upon in a simple way (Efron and Morris, 1973) by using a *positive-part* estimator,

$$(10.7.8) \qquad \delta_i^+(\boldsymbol{X}) = \left(1 - \frac{p-2}{\sum_{j=1}^{p} X_j^2} \right)^+ X_i, \quad i = 1, \ldots, p,$$

where we define the notation $(x)^+ = \max(0, x)$. Hence the coordinates of the positive-part estimator cannot have a different sign from the coordinates of $\boldsymbol{X}$. Also, the positive-part estimator alleviates the strange behavior of the Stein estimator near zero. (Note that as $\boldsymbol{X} \to 0$, $\delta_i^S \to -\infty$ or $+\infty$. Although this behavior does not adversely affect the risk, it would make an experimenter uncomfortable if a small $\boldsymbol{X}$ were observed.) The risk functions of $\boldsymbol{X}$, δ^S, and δ^+, which depend on θ only through $\sum_{i=1}^{p} \theta_i^2$, are shown in Figure 10.7.1. Notice that the biggest risk improvement is obtained near $\theta = 0$, because these estimators all shrink $\boldsymbol{X}$ toward the point $(0, 0, \ldots, 0)$. There is nothing magic about zero, however, and these estimators can shrink toward any point.

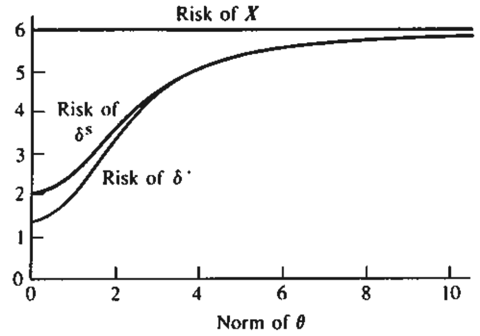

FIGURE 10.7.1 Risk functions for Stein-type estimators

Interestingly, even though δ^+ is a very good estimator (Efron and Morris, 1973), the results of Brown (1971), which generalized the work of Sacks (1963), show that δ^+ is inadmissible. Thus, there are estimators that uniformly dominate δ^+. Even though admissible estimators for this problem have been found (Strawderman, 1971; Berger, 1976), no one has found an estimator that dominates δ^+. Although, practically speaking, δ^+ cannot be improved upon by very much, finding an estimator that dominates it would be a theoretical achievement.

Finally, we note that the Stein Paradox carries over to set estimation in that, in three or more dimensions, the usual confidence set for a vector of normal means is inadmissible. There exist confidence sets centered at Stein-type estimators that have the same volume and higher coverage probability, or smaller volume and the same confidence coefficient. Brown (1966) and Joshi (1967) independently proved the existence of a dominating procedure for $p \geq 3$, and Joshi (1969) later proved the admissibility of the usual confidence set if $p = 1$ or 2. Hwang and Casella (1982) first exhibited a dominating set and Casella and Hwang (1983, 1987) explored the set estimation problem further.

EXERCISES

10.1 Let X have a $\mathrm{n}(\theta, 1)$ distribution, and consider testing $H_0: \theta \geq \theta_0$ versus $H_1: \theta < \theta_0$. Use the loss function (10.2.5) and investigate the three tests that reject H_0 if $X < -z_\alpha + \theta_0$ for $\alpha = .1, .3$, and .5.

a. For $b = c = 1$, graph and compare their risk functions.

b. For $b = 3, c = 1$, graph and compare their risk functions.

c. Graph and compare the power functions of the three tests to the risk functions in parts (a) and (b).

10.2 Consider testing H_0: $p \leq \frac{1}{3}$ versus H_1: $p > \frac{1}{3}$ where $X \sim \text{binomial}(5, p)$ using 0–1 loss. Graph and compare the risk functions for the following two tests. Test I rejects H_0 if $X = 0$ or 1. Test II rejects H_0 if $X = 4$ or 5.

10.3 Consider the binomial estimation problem in Example 10.4.1 for $n = 10$. Graph and compare the risk functions for these two estimators, $\delta(x) = \frac{1}{3}$ and $\delta'(x) = x/10$.

10.4 Show that the log of the likelihood function for estimating σ^2, based on observing $S^2 \sim \sigma^2 \chi^2_\nu / \nu$, can be written in the form

$$\log L(\sigma^2 | s^2) = K_1 \frac{s^2}{\sigma^2} - K_2 \log \frac{s^2}{\sigma^2} + K_3,$$

where K_1, K_2, and K_3 are constants, not dependent on σ^2. Relate the above log likelihood to the loss function discussed in Example 10.2.3. See Anderson (1984a) for a discussion of this relationship.

10.5 Let $X \sim \mathrm{n}(\mu, \sigma^2), \sigma^2$ known. For each $c \geq 0$, define an interval estimator for μ by $C(x) = [x - c\sigma, x + c\sigma]$ and consider the loss in (10.2.7).

a. Show that the risk function, $R(\mu, C)$, is given by

$$R(\mu, C) = b(2c\sigma) - P(-c \leq Z \leq c).$$

b. Using the Fundamental Theorem of Calculus, show that

$$\frac{d}{dc} R(\mu, C) = 2b\sigma - \frac{2}{\sqrt{2\pi}} e^{-c^2/2}$$

and, hence, the derivative is an increasing function of c for $c \geq 0$.

c. Show that if $b\sigma > 1/\sqrt{2\pi}$, the derivative is positive for all $c \geq 0$ and, hence, $R(\mu, C)$ is minimized at $c = 0$. That is, the best interval estimator is the point estimator $C(x) = [x, x]$.

d. Show that if $b\sigma \leq 1/\sqrt{2\pi}$, the c that minimizes the risk is $c = \sqrt{-2\log(b\sigma\sqrt{2\pi})}$. Hence, if b is chosen so that $c = z_{\alpha/2}$ for some α, then the interval estimator that minimizes the risk is just the usual $1-\alpha$ confidence interval.

10.6 Let $X \sim \mathrm{n}(\mu, \sigma^2)$, but now consider σ^2 unknown. For each $c \geq 0$, define an interval estimator for μ by $C(x) = [x - cs, x + cs]$, where s^2 is an estimator of σ^2 independent of X, $\nu S^2/\sigma^2 \sim \chi^2_\nu$ (for example, the usual sample variance). Consider a modification of the loss in (10.2.7),

$$L((\mu, \sigma), C) = \frac{b}{\sigma} \mathrm{Len}(C) - I_C(\mu).$$

a. Show that the risk function, $R((\mu, \sigma), C)$, is given by

$$R((\mu, \sigma), C) = b(2cM) - [2P(T \leq c) - 1],$$

where $T \sim t_\nu$ and $M = \mathrm{E}S/\sigma$.

b. If $b \leq 1/\sqrt{2\pi}$, show that the c that minimizes the risk satisfies

$$b = \frac{1}{\sqrt{2\pi}} \left(\frac{\nu}{\nu + c^2} \right)^{(\nu+1)/2}$$

c. Reconcile this problem with the known σ^2 case. Show that as $\nu \to \infty$, the solution here converges to the solution in the known σ^2 problem. (Be careful of the rescaling done to the loss function.)

10.7 The decision theoretic approach to set estimation can be quite useful (Exercise 10.34) but it can also give some unsettling results, showing the need for thoughtful implementation. Consider again the case of $X \sim \mathrm{n}(\mu, \sigma^2), \sigma^2$ unknown, and suppose that we have an interval estimator for μ by $C(x) = [x - cs, x + cs]$, where s^2 is an estimator of σ^2 independent of $X, \nu S^2/\sigma^2 \sim \chi^2_\nu$. This is, of course, the usual t interval, one of the great statistical procedures that has withstood the test of time. Consider the loss

$$L((\mu, \sigma), C) = b\mathrm{Len}(C) - I_C(\mu),$$

similar to that used in Exercise 10.6, but without scaling the length. Construct another procedure C' as

$$C' = \begin{cases} [x - cs, x + cs] & \text{if } s < K \\ \emptyset & \text{if } s \geq K \end{cases},$$

where K is a positive constant. Notice that C' does *exactly the wrong thing*. When s^2 is big and there is a lot of uncertainty, we would want the interval to be wide. But C' is empty! Show that we can find a value of K so that

$$R((\mu, \sigma), C') \leq R((\mu, \sigma), C), \quad \text{for every } (\mu, \sigma),$$

with strict inequality for some (μ, σ).

10.8 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, σ^2 known. Consider estimating θ using squared error loss. Let $\pi(\theta)$ be a $\mathrm{n}(\mu, \tau^2)$ prior distribution on θ and let δ^π be the Bayes estimator of θ. Verify the following formulas for the risk function and Bayes risk.

For $p = \frac{1}{2}$, $L(\frac{1}{2}, a_0) = L(\frac{1}{2}, a_1) = 0$. Consider the group $\{g_1, g_2\}$ from Example 6.3.1 that has only two elements,

$$g_1(x) = n - x \quad \text{and} \quad g_2(x) = x.$$

a. Show that this testing problem is invariant under this group if and only if the constants from the loss function satisfy $c_{\mathrm{I}} = c_{\mathrm{II}}$. What are $\overline{g_1}$, $\overline{g_2}$, $\tilde{g}_1$, and $\tilde{g}_2$? Why was the modification of the loss for $p = \frac{1}{2}$ necessary?
b. Explain why the conditions of Definition 8.2.4 are not satisfied.
c. Suppose n is odd. Let $c_{\mathrm{I}} = c_{\mathrm{II}} = 1$. Show that a test ϕ is invariant for this problem if and only if, for every $x = 0, \ldots, n$, ϕ takes the opposite actions at x and $n - x$.
d. Show that no invariant test exists if n is even. (*Hint*: The point $x = n/2$ creates problems.)
e. Explain why the invariant tests in part (c) do not satisfy Definition 8.2.3.

10.39 a. Adapt the argument of Section 10.7, and the result of Exercise 10.30, to show that if we observe

$$X_{ij} \sim \mathrm{n}(\theta_i, \sigma^2), \quad i = 1, \ldots, p, \quad j = 1, \ldots, n, \quad \sigma^2 \text{ known},$$

all independent, and we compute $\overline{X}_i = \frac{1}{n}\sum_j X_{ij}$, then the estimator $\overline{X} = (\overline{X}_1, \ldots, \overline{X}_p)$ is minimax.
b. Independently of part (a), show that if we observe $X_{ij} \sim \mathrm{n}(\theta_i, \sigma^2)$, $i = 1, \ldots, p$, and $j = 1, \ldots, n$, then, by sufficiency, we can reduce the problem to that considered in Section 10.7, that of $n = 1$. Hence, the estimator $\overline{X} = (\overline{X}_1, \ldots, \overline{X}_p)$ is minimax.

10.40 The form of the Stein estimator of (10.7.5) can be justified somewhat by an *empirical Bayes* argument, given in Efron and Morris (1972). Such an argument was probably known by Stein, although he makes no mention of it. The empirical Bayes explanation is quite useful, especially in data analysis (see Efron and Morris, 1973, 1975; or Casella, 1985). Let $X_i \sim \mathrm{n}(\theta_i, 1)$, $i = 1, \ldots, p$, and θ_i be iid $\mathrm{n}(0, \tau^2)$.
a. Show that the X_is, marginally, are iid $\mathrm{n}(0, \tau^2 + 1)$, hence, $\sum X_i^2/(\tau^2 + 1) \sim \chi^2_p$.
b. Using the marginal distribution, show that $\mathrm{E}(1 - ((p - 2)/\sum_{j=1}^p X_j^2)) = \tau^2/(\tau^2 + 1)$ if $p \geq 3$. Thus, the ith component of the Stein estimator,

$$\delta_i^S(\boldsymbol{X}) = (1 - ((p-2)/\sum_{j=1}^{p} X_j^2))X_i,$$

is an empirical Bayes version of the ith component of the Bayes estimator $\delta_i^{\pi}(\boldsymbol{X}) = (\tau^2/(\tau^2 + 1))X_i$.

10.41 Consider the class of Stein estimators given by (10.7.7),

$$\delta_i^c(\boldsymbol{X}) = \left(1 - \frac{c}{\sum_{j=1}^p X_j^2}\right) X_i, \quad i = 1, \ldots, p, \quad 0 < c < 2(p-2).$$

Let $X_i \sim \mathrm{n}(\theta_i, 1)$, $i = 1, \ldots, p$.
a. Using sum of squared errors loss, find an expression for the risk of $\delta^c(\boldsymbol{X}) = (\delta_1^c(\boldsymbol{X}), \ldots, \delta_p^c(\boldsymbol{X}))$.
b. Show that for any constant c satisfying $0 < c < 2(p - 2)$, $\delta^c(\boldsymbol{X})$ is better than $\boldsymbol{X}$.
c. Compute the risk of $\delta^c(\boldsymbol{X})$ at $\theta = \mathbf{0}$ and find the value of c that minimizes this risk.
d. Show that the value of $c = p - 2$ minimizes the risk within the class of estimators $\{\delta^c(\boldsymbol{X}) : 0 < c < 2(p - 2)\}$.

10.42 Suppose we observe $X_{ij} \sim \mathrm{n}(\theta_i, \sigma^2)$, σ^2 known, $i = 1, \ldots, p, j = 1, \ldots, n$, all independent, and we form the Stein estimator

$$\delta_i^S(\overline{\boldsymbol{X}}) = \left(1 - \frac{(p-2)(\sigma^2/n)}{\sum_{j=1}^p \overline{X}_j^2}\right) \overline{X}_i, \quad i = 1, \ldots, p,$$

where $\overline{X}_i = \frac{1}{n}\sum_j X_{ij}$ and $\overline{\boldsymbol{X}} = (\overline{X}_1, \ldots, \overline{X}_p)$. Show that $\delta^S(\overline{\boldsymbol{X}})$ is minimax.

10.43 For $X_i \sim \mathrm{n}(\theta_i, 1)$, $i = 1, \ldots, p$, consider the class of Stein estimators that shrink toward an arbitrary point,

$$\delta_i^S(\boldsymbol{X}, \theta^0) = \theta_i^0 + \left(1 - \frac{c}{\sum_{j=1}^p (X_j - \theta_i^0)^2}\right)(X_i - \theta_i^0), \quad i = 1, \ldots, p,$$

where $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)$ is constant and $0 < c < 2(p - 2)$.
a. Show that under sum of squared errors loss, $\delta^S(\boldsymbol{X}, \theta^0)$ dominates $\boldsymbol{X}$ for any value of θ^0.
b. Show that the risk of $\delta^S(\boldsymbol{X}, \theta^0)$ at $\theta = \theta^0$ is the same as the ordinary Stein estimator (10.7.5) at $\theta = 0$.

Miscellanea

Game Theory

A topic closely related to decision theory is *game theory*, a formal mathematical study of games in which two or more players compete. The simplest type of game is a two-person zero-sum game. Player I picks a strategy $a \in \mathcal{A}$. Player II picks a strategy $\theta \in \Theta$. Then Player I pays Player II an amount $L(\theta, a)$, where negative values of $L(\theta, a)$ correspond to payments from Player II to Player I and positive values of $L(\theta, a)$ correspond to payments from Player I to Player II.

Player I may gain some information about what strategy Player II will use by observing a random variable X whose distribution depends on Player II's strategy θ. Player I can use this information to decide what strategy a to use. From the notation we have used, the similarity between these elements of a game and the corresponding elements in a decision problem is evident. Minimax strategies make good sense in a game since there is an intelligent opponent. If Player II always knows what strategy Player I will use, then Player II can choose θ to maximize Player I's expected losses. Player I, knowing that Player II will do this, should use a minimax strategy. Such a strategy will minimize Player I's maximum expected loss, the loss Player I knows he will incur if Player II plays in the way we have described.

As mentioned in Section 5, in statistical problems Nature is not considered to be an adversary and the minimaxity criterion is not so compelling. A classic treatment of game theory and decision theory is given by Blackwell and Girshick (1954), including the famous theorem by von Neumann on the existence of minimax strategies. A later reference is Thomas (1984).

The Hunt–Stein Theorem

The Hunt–Stein Theorem is one of the great items of statistical folklore, as the original paper by Hunt and Stein was never published. However, the theorem due to them is quite real and represents one of the deepest results in mathematical statistics. The most readable (for statisticians) article about this theorem is by Bondar and Milnes (1981), with one of the most general developments given by Kiefer (1957). Lehmann (1986) discusses this theorem in the testing context.