13 Decision Theory

13.1 Preliminaries

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

Consider a parameter θ which lives in a parameter space Θ. Let a be a guess at θ. We refer to a as an *action*. The set of possible actions is called the *action space* and is denoted $\mathcal{A}$. Usually, $\mathcal{A} = \Theta$ but in general they can be different. Assume that $\mathcal{A} = \Theta$ unless otherwise specified.

We shall measure how good a guess a is using a *loss function* $L(\theta, a)$. Formally, L maps $\Theta \times \mathcal{A}$ into $\mathcal{R}$. Here are some examples of loss functions:

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

A *decision rule* $\delta(x)$ assigns an action to each outcome x. Think of a decision rule as an estimator. Formally, if $\mathcal{X}$ denotes the set of possible outcomes, then δ maps $\mathcal{X}$ into $\mathcal{A}$. Examples of decision rules are $\delta(x) = x$, $\delta(x) = 2x$, $\delta(x) = mle$ and $\delta(x_1, \ldots, x_n) = n^{-1}\sum_{i=1}^n x_i$. We shall often use the terms "decision rule" and "estimator" interchangeably.

To assess a decision rule, we evaluate the average loss or *risk:*

$$R(\theta, \delta) = E_\theta\left[L(\theta, \delta(X))\right] = \int L(\theta, \delta(x)) f(x; \theta)dx.$$

When the loss function is squared error, the risk is just the MSE (mean squared error), i.e.

$$R(\theta, \delta) = E_\theta(\delta(X) - \theta)^2 = MSE = \mathrm{Var}_\theta(\delta(X)) + \mathrm{Bias}_\theta^2(\delta(X))$$

where $\mathrm{Bias}_\theta = E_\theta\delta(X) - \theta$.

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.

EXAMPLE 13.1 *Let $X \sim N(\theta, 1)$ and assume we are using squared error loss. Consider two estimators (decision rules): $\delta_1(X) = X$ and $\delta_2(X) = 3$. The risk functions are $R(\theta, \delta_1) = E_\theta(X - \theta)^2 = 1$ and $R(\theta, \delta_2) = E_\theta(3 - \theta)^2 = (3 - \theta)^2$. Notice that, if $2 < \theta < 4$ then $R(\theta, \delta_2) < R(\theta, \delta_1)$ otherwise $R(\theta, \delta_1) < R(\theta, \delta_2)$. Neither estimator dominates the other.*

EXAMPLE 13.2 *Let $X_1, \ldots, X_n \sim$ Bernoulli(p). Consider squared error loss and let $\delta_1(X^n) = \overline{X}$. Since this has 0 bias, we have that*

$$R(p, \delta_1) = Var(\overline{X}) = \frac{p(1-p)}{n}.$$

Another estimator is

$$\delta_2(X^n) = \frac{Y + \alpha}{\alpha + \beta + n}$$

where $Y = \sum_i X_i$ and α and β are positive constants. This is just the posterior mean using a Beta (α, β) prior. Now,

$$\begin{aligned} R(p, \delta_2) &= Var_p \delta_2 + (Bias_p(\delta_2))^2 \\ &= Var_p\left(\frac{Y + \alpha}{\alpha + \beta + n}\right) + \left(E_p\left(\frac{Y + \alpha}{\alpha + \beta + n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha + \beta + n)^2} + \left(\frac{np + \alpha}{\alpha + \beta + n} - p\right)^2. \end{aligned}$$

Now let us choose $\alpha = \beta = \sqrt{n/4}$. Later we shall see that there is a reason for this choice. The resulting estimator is

$$\delta_2(X^n) = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and risk function is

$$R(p, \delta_2) = \frac{n}{4(n + \sqrt{n})^2}.$$

The risk functions are plotted in figure 1. As we can see, neither estimator uniformly dominates the other.

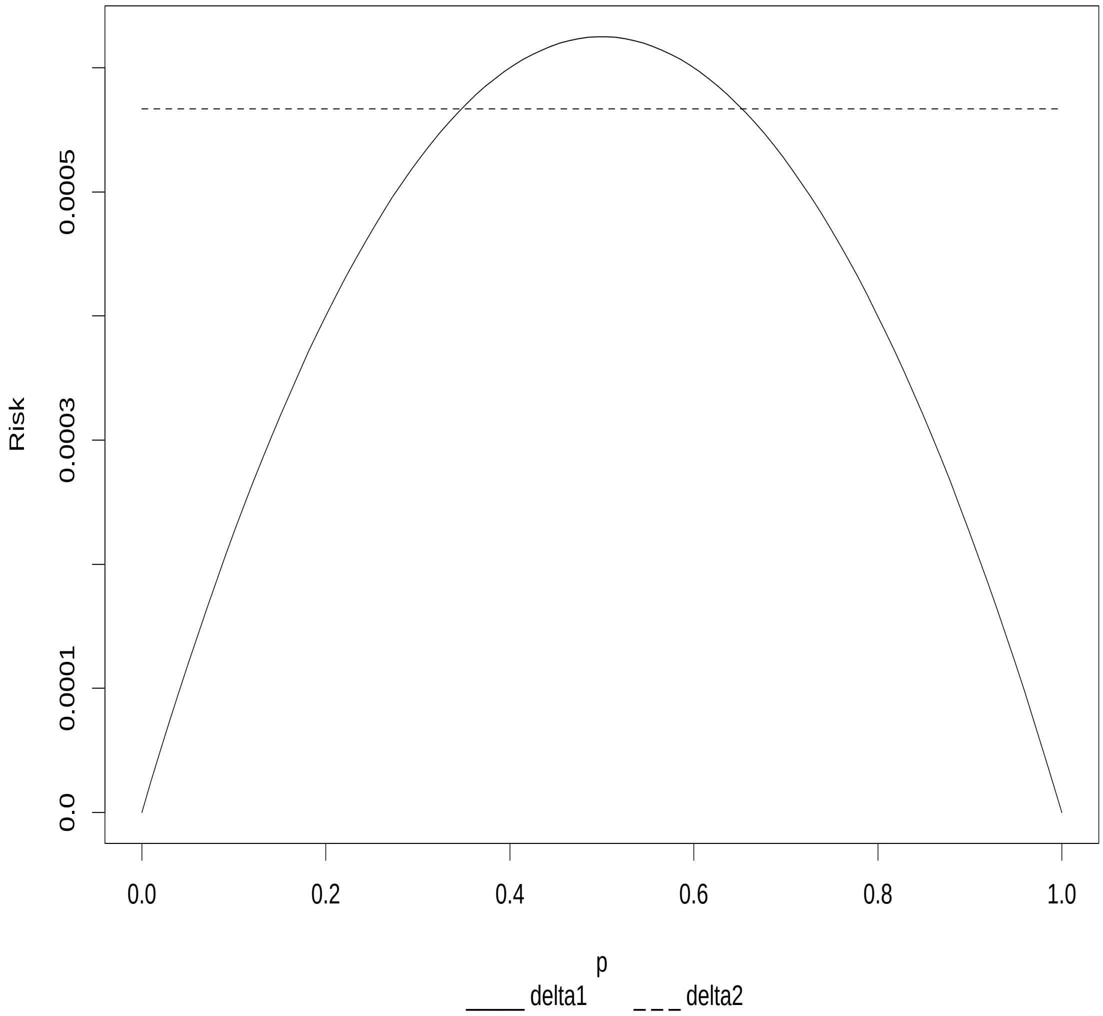

Figure xxxx. Risk functions for δ_1 and δ_2.

13.2 Comparing Risk Functions

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* [10]

$$\sup_\theta R(\theta,\delta)$$

and *the Bayes risk*

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

where $\pi(\theta)$ is a prior for θ. Decision rules that minimize the maximum risk are called *minimax rules.* Decision rules that minimize the Bayes risk are called *Bayes rules.* As it turns out, there are some connections between the two. Roughly, the minimax rule is the Bayes rule for a specially chosen prior called the *least favorable prior.*

13.3 Bayes Estimators

Given a prior $\pi(\theta)$, the Bayes risk is defined by

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

The *Bayes rule* is the rule δ^π that minimizes the Bayes risk:

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

Recall that from Bayes' theorem the posterior density is

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

where $m(x) = \int f(x,\theta)d\theta = \int f(x|\theta)\pi(\theta)d\theta$ is the Bayesian marginal distribution of X. We can re-write the Bayes risk as follows:

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

[10] The expression "sup" means "supremum" which is defined to be he least upper bound. You can think of "sup" as "max."

$$
\begin{aligned}
&= \int\int L(\theta,\delta(x))f(x|\theta)dx\pi(\theta)d\theta \\
&= \int\int L(\theta,\delta(x))f(x,\theta)dxd\theta \\
&= \int\int L(\theta,\delta(x))f(\theta|x)m(x)dxd\theta \\
&= \int\int L(\theta,\delta(x))f(\theta|x)d\theta m(x)dx \\
&= \int r(\delta|x)m(x)dx
\end{aligned}
$$

where

$$r(\delta|x) = \int L(\theta,\delta(x))f(\theta|x)d\theta$$

is the *posterior risk.* The Bayes rule is thus the rule $\delta(x)$ that minimizes the posterior risk. (If we minimize the inside integral for every x then we minimize the whole integral.) More precisely, we have:

THEOREM 13.1 *Let $\delta^\pi(x)$ minimize the posterior risk $r(\delta|x)$. Then δ^π is the Bayes rule.*

Consider squared error loss $L(\theta,a) = (\theta-a)^2$. The Bayes rule minimizes $r(a|x) = \int(\theta-a)^2 f(\theta|x)d\theta$. If we take the derivative of $r(a|x)$ with respect to a and set it equal to 0, we get that the Bayes rule is

$$\delta(x) = a = \int \theta f(\theta|x)d\theta = E(\theta|x).$$

In other words, **the Bayes rule under squared error loss is the posterior mean.** It can be shown the Bayes rule under absolute loss $L(\theta,a) = |\theta - a|$ is the posterior median. And the Bayes' rule under the 0-1 loss $L(\theta,a) = 1$ if $\theta \neq a$ and 0 otherwise, is the posterior mode.

EXAMPLE 13.3 *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*

$$\frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a.$$

13.4 Minimax Rules

If we do not want to include a prior, then we can summarize an estimator by the maximum value of its risk. This leads us to choose the estimator with the smallest maximum value. Formally, a decision rule δ^* is a *minimax rule* if

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

The problem of finding minimax rules is complicated and we cannot attempt a complete coverage of that theory here. But we will mention a few key results.

THEOREM 13.2 *Let δ^π be the Bayes rule for some prior π, i.e.*

$$r(\pi, \delta^\pi) = \inf_\delta r(\pi, \delta). \tag{8}$$

Suppose that

$$R(\theta, \delta^\pi) \leq r(\pi, \delta^\pi) \quad \text{for all } \theta. \tag{9}$$

Then δ^π is minimax and π is called a least favorable prior.

PROOF. Suppose that δ^π is not minimax. Then there is another rule δ_0 such that $\sup_\theta R(\theta, \delta_0) < \sup_\theta R(\theta, \delta^\pi)$. Since the average of a function is always less than or equal to its maximum, we have that $r(\pi, \delta_0) \leq \sup_\theta R(\theta, \delta_0)$. Hence,

$$\begin{aligned} r(\pi, \delta_0) &\leq \sup_\theta R(\theta, \delta_0) \\ &< \sup_\theta R(\theta, \delta^\pi) \\ &\leq r(\pi, \delta^\pi) \end{aligned}$$

which contradicts (8). $\diamond$

This leads immediately to the following useful theorem.

THEOREM 13.3 *Suppose that δ is the Bayes rule with respect to some prior π. Suppose further that δ has constant risk, i.e. $R(\theta, \delta) = c$ for some c. Then δ is minimax.*

PROOF. The Bayes risk is $r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)d\theta = c$ and hence $R(\theta, \delta) \leq r(\pi, \delta)$ for all θ. Now apply the previous theorem. $\diamond$

EXAMPLE 13.4 *Consider the Bernoulli with squared error loss. In example 13.1 we showed that the estimator*

$$\delta(X^n) = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

where $Y = \sum_i X_i$ has a constant risk function. Also, this estimator is the posterior mean (and hence the Bayes rule) for the prior Beta(α, β) with $\alpha = \beta = \sqrt{n/4}$. Hence, it is minimax.

EXAMPLE 13.5 *Consider again the Bernoulli but with loss function*

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

Let

$$\delta(X^n) = \widehat{p} = \frac{Y}{n}.$$

The risk is

$$\begin{aligned} R(p, \delta) &= E\left(\frac{(\widehat{p}-p)^2}{p(1-p)}\right) \\ &= \frac{1}{p(1-p)}\left(\frac{p(1-p)}{n}\right) \\ &= \frac{1}{n}. \end{aligned}$$

Hence the risk function is constant. Also, it can be shown that for this loss function $\delta(X^n)$ *is the Bayes estimator with the prior $\pi(p) = 1$. (You will show this in the homework.) Hence, $\widehat{p}$ is minimax.*

Here is a very useful result. We cannot prove it using our current machinery. We will state the result here and prove (a version of) it later.

THEOREM 13.4 *Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Let $\widehat{\theta} = \overline{X}$. Then $\widehat{\theta}$ is minimax with respect to any well-behaved*[11] *loss function. It is the only*[12] *estimator with this property.*

EXAMPLE 13.6 *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*

$$\delta(X) = m \tanh(mX)$$

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 $R(\theta, \delta) \le r(\pi, \delta)$ for all θ. Hence, Theorem 1 implies it is minimax. The risk is plotted in Figure 2 for $m = .5$.*

13.5 Maximum Likelihood, Minimax and Bayes

There is a sense in which maximum likelihood estimation is minimax. This is a bit technical; feel free to skip this if you want. Consider squared error loss which is squared bias plus variance. In parametric models with large samples, the variance term dominates the bias so the risk of the mle $\widehat{\theta}$ roughly equals the variance: $R(\theta, \widehat{\theta}) \approx Var_\theta(\widehat{\theta})$. As we saw earlier, the variance is approximately $Var(\widehat{\theta}) \approx 1/(nI(\theta))$ where $I(\theta)$ is the Fisher information. Hence, $nR(\theta, \widehat{\theta}) \approx 1/I(\theta)$. For any other estimator δ_n it can be shown that

$$\lim_{\epsilon \to 0} \text{limsup}_{n \to \infty} \sup_{|\theta - \theta'| < \epsilon} nR(\theta', \delta_n) \ge 1/I(\theta).$$

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.

13.6 Admissibility

Minimax rules and Bayes rules are "good rules" in the sense that they have small risk. Sometimes it is also useful to characterize bad rules. Specifically,

[11] The level sets must be convex and symmetric about the origin.

[12] Up to sets of measure 0.

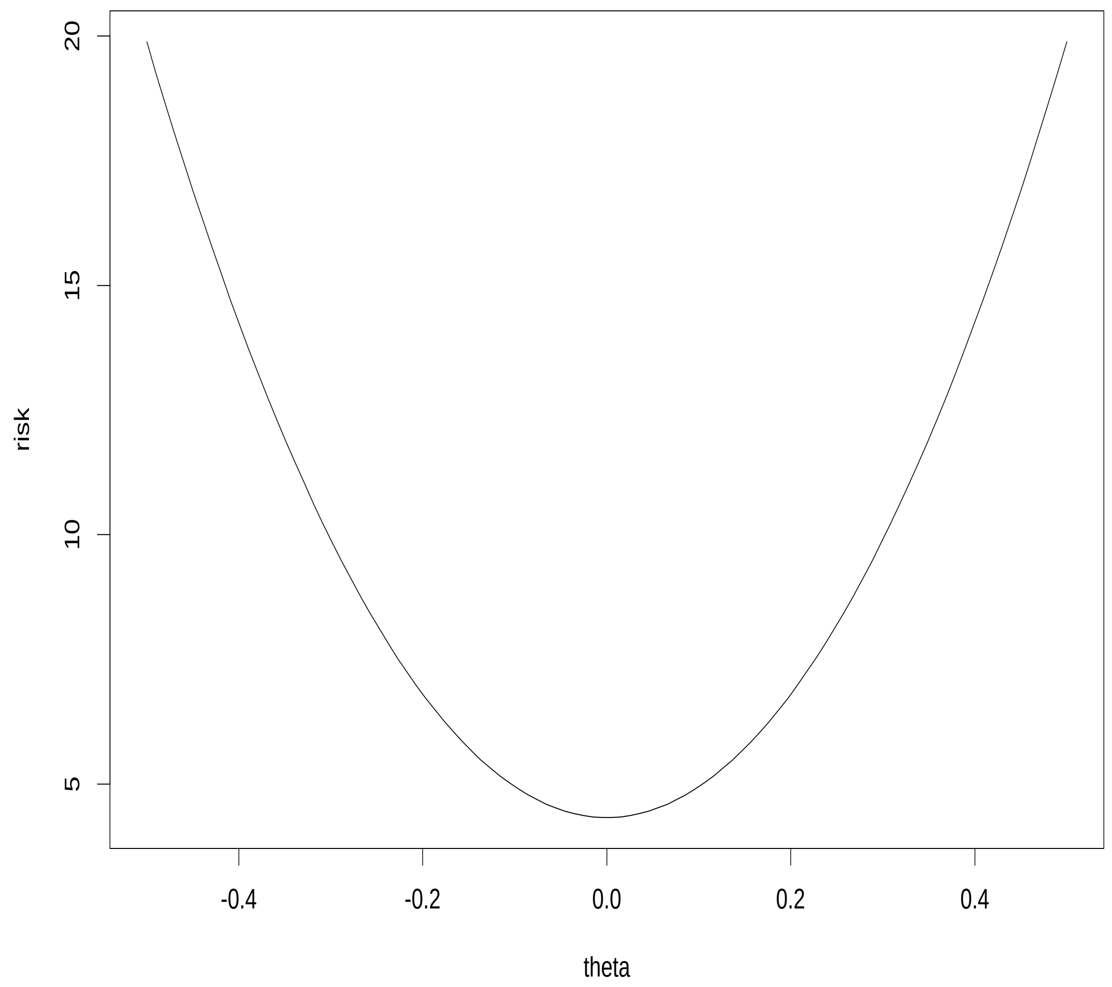

Figure 2. Risk functions for constrained Normal with m=.5.

we say that a decision rule δ is *inadmissible* if there exists another rule δ' such that

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

EXAMPLE 13.7 *Let $X \sim N(\theta,1)$ and consider estimating θ with squared error loss. Let $\delta(X) = 3$. We will show that δ is admissible. Suppose not. Then there exists a different rule δ' with smaller risk. In particular, $R(3,\delta') \leq R(3,\delta) = 0$. Hence, $0 = R(3,\delta') = \int(\delta'(x) - 3)^2 f(x;3)dx$. Thus, $\delta'(x) = 3$. So there is no rule that beats δ. Even though δ is admissible it is clearly a bad decision rule.*

A prior density has *full support* if for every θ and every $\epsilon > 0$, $\int_{\theta-\epsilon}^{\theta+\epsilon} \pi(\theta)d\theta > 0$.

THEOREM 13.5 (Bayes' rules are admissible.) *Suppose that $\Theta \subset \mathcal{R}$ and that $R(\theta,\delta)$ is a continuous function of θ for every δ. Let π be a prior density with full support and let δ^π be the Bayes' rule. If the Bayes risk is finite then δ^π is admissible.*

PROOF. Suppose δ^π is inadmissible. Then there exists a better rule δ such that $R(\theta,\delta) \leq R(\theta,\delta^\pi)$ for all θ and $R(\theta_0,\delta) < R(\theta_0,\delta^\pi)$ for some θ_0. Let $\nu = R(\theta_0,\delta^\pi) - R(\theta_0,\delta) > 0$. Since R is continuous, there is an $\epsilon > 0$ such that $R(\theta,\delta^\pi) - R(\theta,\delta) > \nu/2$ for all $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$. Now,

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

This implies that δ^π does not minimize $r(\pi,\delta)$ which contradicts the fact that δ^π is the Bayes rule.

THEOREM 13.6 *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.

THEOREM 13.7 *Suppose that δ has constant risk and is admissible. Then it is minimax.*

PROOF. The risk is $R(\delta, \delta) = c$ for some c. If δ were not minimax then there exists a rule δ' such that

$$R(\theta, \delta') \leq \sup_\theta R(\theta, \delta') < \sup_\theta R(\delta, \delta) = c.$$

This would imply that δ is inadmissible.

Now we can prove Theorem 13.4, at least for squared error loss, which says that the sample mean is admissible. This follows from the previous theorem and 13.5.

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

THEOREM 13.8 *If δ is minimax then it is not badly inadmissible.*

13.7 Stein's Paradox

Suppose that $X \sim N(\theta, 1)$ and consider estimating θ with squared error loss. We know that $\delta(X) = X$ 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 $L(\theta, a) = \sum_{j=1}^2 (\theta_j - a_j)^2$. Let $X = (X_1, X_2)$. Not surprisingly, $\delta(X) = X$ is again admissible. 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 $L(\theta, a) = \sum_{j=1}^k (\theta_j - a_j)^2$. Stein proved that if $k \geq 3$ then $\delta(X) = X$ is inadmissible. It can be shown that the following rule has smaller risk everywhere:

$$\delta_S(X) = \left(1 - \frac{k-2}{\sum_i X_i^2}\right)^+ X_i$$

where $(z)^+ = \max\{z, 0\}$. This is a startling, deep fact. Note that the estimator essentially shrinks these values towards each other. The message is that, when estimating many parameters, there is great value in "shrinking" the estimates. Surprisingly, this observation plays an important role in modern nonparametric function estimation.