EXAMPLE 46. Assume $\mathbf{X}=(X_1, X_2, \ldots, X_p)' \sim \mathcal{N}_p(\boldsymbol{\theta}, \mathbf{I}_p)$, where $\boldsymbol{\theta}=(\theta_1, \ldots, \theta_p)'$ and $\mathbf{I}_p$ is the $(p \times p)$ identity matrix. It is desired to estimate $\boldsymbol{\theta}$ under sum-of-squares error loss $(L(\boldsymbol{\theta}, \mathbf{a})=\sum_{i=1}^{p}(\theta_i-a_i)^2)$. (This could equivalently be stated as the problem of trying to simultaneously estimate p normal means from independent problems.) Since $\boldsymbol{\theta}$ is a location parameter, the noninformative prior density $\pi(\boldsymbol{\theta})=1$ is deemed appropriate. It is easy to see that the (formal) posterior density of $\boldsymbol{\theta}$ given $\mathbf{x}$ is then a $\mathcal{N}_p(\mathbf{x}, \mathbf{I}_p)$ density. The generalized Bayes estimator of $\boldsymbol{\theta}$ is the mean of the posterior (under sum-of-squares error loss, or indeed any quadratic loss), so $\boldsymbol{\delta}^0(\mathbf{x})=\mathbf{x}=(x_1, \ldots, x_p)'$ is the generalized Bayes estimator. (If a sample of vectors $\mathbf{X}^1, \ldots, \mathbf{X}^n$ was taken, the generalized Bayes estimator would just be the vector of sample means.)

This most standard of estimators is admissible for $p=1$ or 2 (see Chapter 8), but surprisingly is inadmissible for $p \geq 3$. Indeed James and Stein (1960) showed that

$$\boldsymbol{\delta}^{JS}(\mathbf{x})=\left(1-\frac{(p-2)}{\sum_{i=1}^{p} x_i^2}\right)\mathbf{x}$$

has $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS})<R(\boldsymbol{\theta}, \boldsymbol{\delta}^0)$ for all $\boldsymbol{\theta}$, if $p \geq 3$. (The proof is outlined in a more general setting in Subsection 5.4.3, where additional references are also given.)

It should be mentioned that the inadmissibility in this example is, in some sense, less serious than that in Example 45. In fact, the ratio of $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS})$ to $R(\boldsymbol{\theta}, \boldsymbol{\delta}^0)$ is very close to one over most of the parameter space. Only in a small region near zero (several standard deviations wide) will the ratio of risks be significantly smaller than one. This is in contrast to the situation of Example 45, in which the ratio of risks can be uniformly bad.

The estimator $\boldsymbol{\delta}^{JS}$ can be modified so as to adjust the region of significant improvement to coincide with prior knowledge concerning $\boldsymbol{\theta}$. Indeed, one attractive such modification is the estimator $\boldsymbol{\delta}^S$ in (4.102), which not only satisfies $R(\boldsymbol{\theta}, \boldsymbol{\delta}^S)<R(\boldsymbol{\theta}, \boldsymbol{\delta}^0)$, but also was shown to have excellent Bayesian performance when prior knowledge about $\boldsymbol{\theta}$ is available. (Such modifications also eliminate the conditionally silly feature of $\boldsymbol{\delta}^{JS}$ that, if say $|\mathbf{x}|^2=(p-2)/1000$, then $\boldsymbol{\delta}^{JS}(\mathbf{x})=-999\mathbf{x}$.) Nevertheless, if essentially *no* prior information about $\boldsymbol{\theta}$ is available, then use of $\boldsymbol{\delta}^{JS}$ (or some modification) will not be significantly beneficial, and $\boldsymbol{\delta}^0$ might as well be used. (See Subsection 5.4.3 for further discussion.)

A version of this example, where use of $\pi(\boldsymbol{\theta}) \equiv 1$ does result in a uniformly inadmissible estimator, was given in Example 35 in Subsection 4.7.9. There it was desired to estimate $\eta=|\boldsymbol{\theta}|^2$, and the generalized Bayes estimator under squared-error loss was shown to be $\delta^\pi(\mathbf{x})=|\mathbf{x}|^2+p$. But it can be shown that the estimator $\delta^c(\mathbf{x})=|\mathbf{x}|^2-c$ has

$$R(\boldsymbol{\theta}, \delta^c)=2p+(p-c)^2+4|\boldsymbol{\theta}|^2, \tag{4.134}$$

so that $R(\boldsymbol{\theta}, \delta^\pi)-R(\boldsymbol{\theta}, \delta^p)=4p^2$, a substantial uniform difference (although again $R(\boldsymbol{\theta}, \delta^\pi)/R(\boldsymbol{\theta}, \delta^p) \to 1$ as $|\boldsymbol{\theta}|^2 \to \infty$).

4.8.3. Inadmissibility and Long Run Evaluations

Inadmissibility is a frequentist concept, and hence its relevance to a Bayesian can be debated. Indeed, it would be hard to quarrel with the conditional noninformative prior Bayesian who, in say Example 46, decided *after careful thought* that his prior was essentially uniform in the region of concentration of the likelihood function, and so used $\pi(\boldsymbol{\theta}) \equiv 1$. It is quite a different matter, however, to recommend *automatic* use of a given noninformative prior when prior information is vague (or, even worse, when it is not). The point is that, in recommending repeated use of a particular prior, a Bayesian has entered into the frequentist domain; it is then perfectly reasonable to investigate how repeated use of this prior actually performs.

The natural method of evaluating long run performance of a procedure δ (here the procedure determined by repeated use of a particular noninformative prior) was discussed in Subsection 1.6.2; consider a sequence of independent problems $\{(\theta^{(1)}, X^{(1)}), (\theta^{(2)}, X^{(2)}), \ldots\}$ and a loss (or criterion function) L that measures the performance of the procedure in each problem, and look at the long run performance of δ. One reasonable method of *comparing* two procedures, δ_1 and δ_2, would be to look at

$$S_N=\sum_{i=1}^{N}\{L(\theta^{(i)}, \delta_1(X^{(i)}))-L(\theta^{(i)}, \delta_2(X^{(i)}))\},$$

and consider the limiting behavior (in some sense) of S_N. We stress that this compares actual performance in repeated use, and is not some arbitrary frequentist comparison. Not surprisingly, however, the limiting behavior of S_N is very related to risk comparisons between δ_1 and δ_2. Indeed, the following theorem establishes two such results.

Theorem 10. *Consider* $\boldsymbol{\theta}=(\theta^{(1)}, \theta^{(2)}, \ldots)$ *to be any fixed sequence of parameters* $(\theta^{(i)} \in \Theta)$, *and suppose random variables* $X^{(i)} \in \mathcal{X}$ *are independently generated from the densities* $f(x^{(i)}|\theta^{(i)})$, $i=1, 2, \ldots$ (*here* f *is the same for the entire sequence*). *Define the random variables*

$$Z_i=L(\theta^{(i)}, \delta_1(X^{(i)}))-L(\theta^{(i)}, \delta_2(X^{(i)})),$$

and assume that $E_{\theta^{(i)}}[Z_i-E_{\theta^{(i)}}(Z_i)]^2<\infty$ *for all* i.

(a) *If* $R(\theta, \delta_1)-R(\theta, \delta_2)>\varepsilon>0$ *for* <u>all</u> $\theta \in \Theta$, *then*

$$P_{\boldsymbol{\theta}}\left(\liminf_{N \to \infty} \frac{1}{N} S_N>\varepsilon\right)=1 \tag{4.135}$$

for any sequence $\boldsymbol{\theta}$.

(b) *If* $R(\theta, \delta_1)-R(\theta, \delta_2)>0$ *for all* θ, Θ *is closed, and* $R(\theta, \delta_1)$ *and* $R(\theta, \delta_2)$ *are continuous in* θ, *then* (4.135) *is valid for any* <u>bounded</u> *sequence* $\boldsymbol{\theta}$ (*although* $\varepsilon>0$ *could depend on the bound*).

PROOF. Observe that

$$\psi(\theta^{(i)}) = E_{\theta^{(i)}}[Z_i] = R(\theta^{(i)}, \delta_1) - R(\theta^{(i)}, \delta_2).$$

Thus, by the strong law of large numbers,

$$\frac{1}{N}\sum_{i=1}^{N}[Z_i - \psi(\theta^{(i)})] \to 0$$

with probability one. Under the condition on the risks in part (a), $\psi(\theta^{(i)}) > \varepsilon$ for all i, and the result is immediate. The proof of part (b) is left as an exercise. □

The moment condition in the theorem is trivially satisfied for bounded losses, and usually holds even for unbounded losses. Knowing that (4.135) holds would seem to be a serious indictment of δ_1, since this indicates that δ_1 will be inferior to δ_2 in actual practical use. The indictment in part (a) is particularly serious, since the conclusion holds for *any* sequence $\boldsymbol{\theta}$. Even the part (b) conclusion, that (4.135) holds for any bounded $\boldsymbol{\theta}$, is disturbing since, in practice, the sequence $\boldsymbol{\theta}$ probably will be bounded; we may not *know* the bound (the common reason for using an unbounded Θ in the first place), but δ_2 will beat δ_1 in repeated use *regardless of knowledge of the bound.*

EXAMPLE 46 (continued). It can be shown that the moment condition on the Z_i and the continuity condition on the risks are satisfied for either the problem of estimating means, $\boldsymbol{\theta}^{(i)}$, or of estimating $\eta^{(i)} = |\boldsymbol{\theta}^{(i)}|^2$. For estimating means, we saw that $R(\boldsymbol{\theta}, \boldsymbol{\delta}^0) - R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS}) > 0$ for all $\boldsymbol{\theta}$, so that the conclusion in part (b) of Theorem 10 applies. For estimating η, we saw that $R(\boldsymbol{\theta}, \delta^\pi) - R(\boldsymbol{\theta}, \delta^p) = 4p^2$, so that the conclusion in part (a) of Theorem 10 applies.

It is important to keep inadmissibility results, such as these, in proper perspective. Unless ε in (4.135) is quite large, δ_1 might be perfectly satisfactory in practice (as was stated to be the case for problems of estimating $\boldsymbol{\theta}$ in Example 46 when only very vague prior information is available). And, again, inadmissibility applies only to automated use of δ_1, in say a computer package. The need to at least consider admissibility when developing automated generalized Bayes rules seems strong, however. Other discussions, similar to that in this subsection, can be found in Hill (1974), Heath and Sudderth (1978), and Berger (1984d).

A repetitive but better-safe-than-sorry comment: the last two subsections should not be interpreted as demonstrating that noninformative prior Bayesian analysis is bad. Throughout the book, we have argued that such analysis is very powerful and almost always gives excellent results. An attempt has been made to expose the pitfalls of the noninformative prior Bayesian approach, but we feel that these pitfalls are far less frequent and less deep than those for other competing methods of "automatic" analysis.

Dutch Book Arguments

The use of (4.135) to compare δ_1 and δ_2 is reminiscent of so-called "Dutch book" or "betting coherency" arguments. The typical Dutch book scenario deals with evaluation of methods (usually inference methods) which produce, for each x, either a probability distribution for θ, say $q_x(\theta)$ (which could be a posterior distribution, a fiducial distribution, etc.), or a system of confidence statements $\{C(x), \alpha(x)\}$ with the interpretation that θ is felt to be in $C(x)$ with probability $\alpha(x)$. (Note that frequentist "confidence" theory is excludable from this scenario, in that it does not claim to yield anything resembling the probability that θ is in $C(x)$.) For simplicity, we will restrict ourselves to the confidence statement framework; any $\{q_x(\theta)\}$ can be at least partially evaluated through confidence statements, by choosing $\{C(x)\}$ and letting $\alpha(x)$ be the probability (with respect to q_x) that θ is in $C(x)$.

The assumption is then made (more on this later) that, since $\alpha(x)$ is thought to be the probability that θ is in $C(x)$, the statistician who proposes $\{C(x), \alpha(x)\}$ should be willing to make both the bet that θ is in $C(x)$ at odds of $(1-\alpha(x))$ to $\alpha(x)$, and the bet that θ is not in $C(x)$ at odds of $\alpha(x)$ to $(1-\alpha(x))$. One can then set up a long run evaluation scheme, where a sequence of problems $\{(\theta^{(1)}, X^{(1)}), (\theta^{(2)}, X^{(2)}), \ldots\}$ is again considered, and in which the statistician must accept any bets in each problem according to his stated odds.

Suppose the statistician's opponent bets according to the betting function $s(x)$, where (following Robinson (1979a, b)) $s(x) = 0$ means that no bet is offered; $s(x) > 0$ means that an amount $s(x)$ is bet that $\theta \in C(x)$; and $s(x) < 0$ means that the amount $|s(x)|$ is bet that $\theta \notin C(x)$. The loss to the statistician in the ith problem can then be shown to be

$$W_i = [I_{C(x^{(i)})}(\theta^{(i)}) - \alpha(x^{(i)})]s(x^{(i)}),$$

and of interest is again the limiting behavior of $S_N = \sum_{i=1}^{N} W_i$. If (4.135) holds for all $\boldsymbol{\theta}$, the statistician is called *incoherent* (or, alternatively, $s(x)$ is said to be a *super relevant* betting strategy), while, if (4.135) holds only for bounded $\boldsymbol{\theta}$, the statistician is called *weakly incoherent* (or $s(x)$ is *weakly relevant*). These concepts can be found in this or related form in such works as Buehler (1959, 1976), Wallace (1959), Freedman and Purves (1969), Bondar (1977), Heath and Sudderth (1978), Robinson (1979a, 1979b), and Lane and Sudderth (1983).

If $\{C(x), \alpha(x)\}$ is incoherent or weakly incoherent, then the statistician will for sure lose money in the repeated betting scenario, which certainly casts doubt on the validity of the probabilities $\{\alpha(x)\}$. A number of objections to this scenario have been raised, however, and careful examination of these objections is worthwhile.

Objection 1. The statistician will have no incentive to bet, unless he perceives

the odds as slightly favorable. This turns out to be no problem if incoherence is present, since the odds can be adjusted by $\varepsilon/2$ in the statistician's favor, and he will still lose. If only weak incoherence is present, it is still often possible to adjust the odds by a function $g(x)$, so that the statistician perceives the bets to be in his favor, and yet he will lose in the long run.

Objection 2. The situation is unfair to the statistician, since his opponent gets to choose when, how much, and which way to bet. Various proposals have been made to "even things up." The possibility mentioned in Objection 1 is one such proposal, but does not change the conclusions much. A more radical possibility, suggested by Fraser (1977), is to allow the statistician to decline bets. This can have a drastic effect, but strikes us as too radical, in that it gives the statistician license to state completely silly $\alpha(x)$ for some x. It is after all the $\{\alpha(x)\}$ that are being tested, and testing should be allowed for all x.

Objection 3. The most serious objection to the betting scenario is that $\{\alpha(x)\}$ is generally not selected for use in betting, but rather to communicate information about θ. It may be that there is no *better* choice of $\{\alpha(x)\}$ for communicating the desired information. Consider the following example, which can be found in Buehler (1971), and is essentially successive modifications by Buehler and H. Rubin of an earlier example of D. Blackwell.

EXAMPLE 47. Suppose X has density $f(\theta+1|\theta)=f(\theta-1|\theta)=\frac{1}{2}$, and $\theta\in\Theta=$ {integers}. We are to evaluate the confidence we attach to the sets $C(x)=\{x+1\}$ (the point $(x+1)$), and a natural choice is $\alpha(x)=\frac{1}{2}$ (since θ is either $x-1$ or $x+1$, and in the absence of fairly strong prior information about θ, either choice seems equally plausible). This choice can be beaten in the betting scenario, by betting that θ is not in $C(x)$ with probability $g(x)$, where $0<g(x)<1$ is a continuous increasing function. (Allowing a randomized betting strategy does not seem unreasonable.) Indeed, the expected gain per bet of one unit, for any bounded $\boldsymbol{\theta}$, is $\sup_{\theta_i\in\boldsymbol{\theta}}[g(\theta_i+1)-g(\theta_i-1)]>0$, so that $\alpha(x)=\frac{1}{2}$ is weakly incoherent. (A continuous version of this example, mentioned in Robinson (1979a), has $X\sim\mathcal{N}(\theta,1)$, $\Theta=\mathbb{R}^1$, $C(x)=(-\infty,x)$, and $\alpha(x)=\frac{1}{2}$.)

In this and other examples where $\{\alpha(x)\}$ loses in betting, one can ask the crucial question—Is there a better α that could be used? The question has no clear answer, because the purpose of α is not clearly defined. One possible justification for $\alpha(x)=\frac{1}{2}$, in the above example, is that it is the unique limiting probability of $C(x)$ for sequences of what could be called increasingly vague prior distributions (cf. Stone (1970)). (A more formal Bayesian justification along these lines would be a robust Bayesian justification, to the effect that the class of possible priors is so large that the range of possible posterior probabilities for $(-\infty,x)$ will include $\frac{1}{2}$ for all x.) An

alternative justification can be found by retreating to decision theory, attempting to quantify how well $\alpha(x)$ performs as an indicator of whether or not θ is in $C(x)$, and then seeing if there is any better α. For instance, using the quadratic scoring function (see Subsection 2.4.3) as an indicator of how well $\alpha(x)$ performs, would mean considering the loss function

$$L_C(\theta,\alpha(x))=(I_{C(x)}(\theta)-\alpha(x))^2.$$

(For the moment, we are considering $\{C(x)\}$ as given, and worrying only about the choice of α. Note that, for any posterior distribution on θ, the optimal choice of $\alpha(x)$ for L_C is the posterior probability of $C(x)$, so that L_C is a natural measure of the accuracy of α.) One can then ask if there is a better α, employing usual decision-theoretic ideas. The answer in the case of Example 47 is—no. It can be shown that $\alpha(x)=\frac{1}{2}$ is admissible for this loss, and hence no improvement is possible. (The same cannot necessarily be said, however, if choice of $C(x)$ is brought into the picture. For instance, a reasonable overall loss for $\{C(x),\alpha(x)\}$ is

$$L(\theta,C(x),\alpha(x))=c_1(I_{C(x)}(\theta)-\alpha(x))^2+c_2(1-I_{C(x)}(\theta))+c_3\mu(C(x)), \tag{4.136}$$

where c_i are constants and μ is a measure of the size of $C(x)$. It can be shown in Example 47 that $\{C^*(x),\alpha^*(x)\}$, with $\alpha^*(x)\equiv\frac{1}{2}$ and

$$C^*(x)=\begin{cases}\{x-1\} & \text{with probability } g(x),\\ \{x+1\} & \text{with probability } 1-g(x),\end{cases} \tag{4.137}$$

is a better procedure than the given $\{C(x),\alpha(x)\}$.)

Decision-theoretic inadmissibility, with respect to losses such as L_C, can be related to incoherency, and seems to be a criterion somewhere between weak incoherency and incoherency (cf. Robinson (1979a)). This supports the feeling that it may be a more valid criterion than the betting criterion. This is not to say that the betting scenarios are not important. Buehler, in discussion of Fraser (1977), makes the important point that, at the very least, betting scenarios show when quantities such as $\alpha(x)$ "behave differently from ordinary probabilities." And as Hill (1974) says,

> "... the desire for coherence ... is not primarily because he fears being made a sure loser by an intelligent opponent who chooses a judicious sequence of gambles ... but rather because he feels that incoherence is symptomatic of something basically unsound in his attitudes."

Nevertheless, Objection 3 often prevents betting incoherency from having a conclusive impact, and so decision-theoretic inadmissibility (with respect to an agreed upon criterion) is more often convincing.

Decision-theoretic methods of evaluating "inferences" such as $q_x(\theta)$ (i.e., distributions for θ given x) have also been proposed (cf. Eaton (1982) and Gatsonis (1984)). For the most part, however, little attention has been directed to these matters.