STA7346: Statistical Inference

Instructor: Prof. Zhihua Su
Taken by Yu Zheng

Chapter 1: Decision Theory, Bayesian Analysis

1. General Non-sequential Decision Problem

The general non-sequential decision theory consists of three basic elements.

(a) A nonempty set of Θ of possible states of nature, sometimes referred to as the parameter space;

(b) A nonempty set of $\mathcal{A}$ of actions available to the statistician;

(c) A loss function $L(\theta, a)$, a map from $\Theta \times \mathcal{A}$ to the reals.

The triplet $(\Theta, \mathcal{A}, L)$ defines what is called a **Game**, and is now interpreted as follows. Nature chooses a point θ in Θ, and the statistician without being informed of the choice of nature, chooses an action a in $\mathcal{A}$. As a consequence of these two choices, the statistician loses an amount $L(\theta, a)$.

- Example 1: (Prisoner's dilemma)

  Two prisoners were partners in a crime, and they were questioned in separate rooms. Each prisoner had a choice of confessing to the crime, and there by implicating the other, or denying that he had participated in the crime. If only one prisoner confessed, then he would go free, and the authorities would throw the book at the other prisoner, requiring him to spend 6 months in prison. If both prisoners denied being involved, then both would be held 1 month on a technicality, and if both prisoners confessed, they would both be held for 3 months.

  To create $(\Theta, \mathcal{A}, L)$ out of this, label prisoner 1 as nature, and prisoner 2 as statistician. Denoting by 1 and 2 the respective decision to confess or deny. One gets $\Theta = \{1, 2\}$, $\mathcal{A} = \{1, 2\}$. The loss function is now demonstrated in the following table:

$\mathcal{A} \setminus \Theta$	1	2
1	3	0
2	6	1

  That is, $L(1,1) = 3, L(2,1) = 0, L(1,2) = 6, L(2,2) = 1$.

- Example 2: Suppose $\Theta = \{\theta_1 = \text{no rain}, \theta_2 = \text{light rain}, \theta_3 = \text{heavy rain}\}$,
  $\mathcal{A} = \{a_1 = \text{carrying an umbrella}, a_2 = \text{not carrying an umbrella}\}$.

$\mathcal{A} \setminus \Theta$	θ_1	θ_2	θ_3
a_1	1	-1	-4
a_2	0	1	3

The above are examples of no-data decision problems. In statistics, however, we usually consider a random variable X assuming values $x \in X$ with a family of possible distributions $\{P_\theta, \theta \in \Theta\}$. If we observe $X = x$, we take the action $\delta(x) \in \mathcal{A}$ according to some rule δ. Such a δ is called a **Non-randomized Decision Rule**.

Def: A *non-randomized decision rule (function)* δ is a map from $\mathcal{X}$ to $\mathcal{A}$.

If $X = x$ is observed, and θ is the true state of nature, or true parameter, the loss incurred is $L(\theta, \delta(x))$.

Note that $L(\theta, \delta(x))$ is a random variable. The **average loss** or **risk** is then given by

$$R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(x)) = \int L(\theta, \delta(x)) f_\theta(x)\, dx \quad \text{or} \quad \sum_x L(\theta, \delta(x)) f_\theta(x)$$

according as $f_\theta(x)$ is the pdf or pmf of X.

Let $\mathcal{D}$ = class of all non-randomized decision rules δ for which $R(\theta, \delta)$ is finite for all $\theta \in \Theta$.

- Example 3: Consider the prisoners dilemma in Example 1. Suppose that before the prisoners are questioned, the statistician asked the nature if he was going to confess or deny. Assume that the nature answers truthfully with probability $3/4$. Then the statistician observes a random variable X (the answer nature gives) taking the values 1 or 2. Note that $f_1(1) = f_2(2) = 3/4$, $f_1(2) = f_2(1) = 1/4$.

  There are four possible nonrandomized decision rules $\delta_1, \delta_2, \delta_3, \delta_4$ from $\mathcal{X} = \{1, 2\}$ to $\mathcal{A} = \{1, 2\}$.

  $$\begin{gathered}\delta_1(1) = \delta_1(2) = 1 \\ \delta_2(1) = 1, \delta_2(2) = 2 \\ \delta_3(1) = 2, \delta_3(2) = 1 \\ \delta_4(1) = \delta_4(2) = 2\end{gathered}$$

  Rules δ_1 and δ_4 ignores the value of X. Rule δ_2 reflects the belief of the statistician that the nature is telling the truth, and rule δ_3 that the nature is not telling the truth.

  $$R(\theta, \delta_1) = L(\theta, \delta_1(1)) f_\theta(1) + L(\theta, \delta_1(2)) f_\theta(2)$$

  $$\begin{aligned} R(1, \delta_1) &= L(1,1) f_1(1) + L(1,1) f_1(2) = L(1,1) = 3 \\ R(2, \delta_1) &= L(2,1) f_2(1) + L(2,1) f_2(2) = L(2,1) = 0 \end{aligned}$$

  Proceeding in this way, we complete the following table:

$\Theta \setminus \mathcal{D}$	δ_1	δ_2	δ_3	δ_4
1	3	15/4	$\cdots$	$\cdots$
2	0	3/4	$\cdots$	$\cdots$

  Notice that the original triplet $(\Theta, \mathcal{A}, L)$ has now been replaced by the new triplet $(\Theta, \mathcal{D}, R)$ in which the space $\mathcal{D}$ and the function R have an underlying structure depending on $\mathcal{A}$, L and the distribution of X.

Two important categories of mathematical statistics can be brought under the above framework.

- Hypothesis testing (Page 8 in "Inference typed notes")
- Point estimation (Page 8 in "Inference typed notes")

The fundamental problem of decision theory can be stated as follows:

Given $(\Theta, \mathcal{A}, L)$ and a random variable X with distribution depending on $\theta \in \Theta$, what decision rule should a statistician use? The natural reaction is to search for the best decision rule, a rule that has the smallest risk uniformly for all $\theta \in \Theta$. Such a decision rule usually does not exist as the following example shows:

- Example: Let $X = (X_1, \cdots, X_{16})$ constitute a random sample of size 16 from $\mathcal{N}(\theta, 1)$. The problem is point estimation of θ. Assume squared error loss, i.e. $L(\theta, a) = (\theta - a)^2$. The following two estimators are suggested:

$$\delta_1(X) = \sum_{i=1}^{16} X_i = \bar{X}, \quad \delta_2(X) = 0$$

$$R(\theta, \delta_1) = \frac{1}{16}, \quad R(\theta, \delta_2) = \theta^2$$

  The conclusion is that neither δ_1 nor δ_2 is uniformly better than the other.

In the absence of a uniformly best decision rule, two general methods are proposed for arriving at decision rule which are quite satisfactory.

Method 1 (Restricting the Available Rules)

The reason a uniformly best rule does not exist is that there are too many others available, some of which are not really good because they guard against some specific states of nature extremely well while neglecting the others. The example in the previous page illustrate this. This suggests restriction of the rules, and find the best in the restricted class. Two such restrictions are (i) unbiasedness and (ii) invariance.

Method 2 (Ordering the Decision Rules)

The statistician may, if he likes, invent a principle which leads to an ordering of the available decision rules. Two important and useful principle in decision theory are (i) The Bayes Principle and (ii) The Minimax Principle.

(i) The Bayes Principle

The Bayes principle involves the notion of a distribution on the parameter space Θ, called a Prior Distribution.

Def: The *Bayes risk* of a decision rule δ w.r.t. a prior distribution ξ denoted by $r(\xi, \delta)$ is defined by $r(\xi, \delta) = \mathbb{E}R(w, \delta)$ where w is a random variable assuming values $\theta \in \Theta$ with a distribution ξ.

Def: A decision rule δ_ξ is said to *Bayes* w.r.t. a prior distribution ξ if $r(\xi, \delta_\xi) = \inf_{\delta \in D} r(\xi, \delta)$.

The value on the right hand side of the above is known as the **Minimum Bayes Risk**. Bayes rules may not exist even if the minumum Bayes risk is defined and is finite. In such a case, a statistician has to be satisfied with a rule whose Bayes risk is close to the minimum value.

Def: Let $\varepsilon > 0$. A decision rule δ_ξ is said to be ε-Bayes w.r.t. the prior distribution ξ if $r(\xi, \delta_\xi) \leq \inf_{\delta \in D} r(\xi, \delta) + \varepsilon$.

(ii) The Minimax Principle

An essential different type of ordering of the decision rules is obtained by ordering the rules according to the worst that could happen to the statistician.

In other words, a decision rule δ_1 is **preferred** to a decision rule δ_2 if $\sup_\theta R(\theta, \delta_1) < \sup_\theta R(\theta, \delta_2)$.

Def: A decision rule δ_D is said to be *Minimax* if $\sup_{\theta \in \Theta} R(\theta, \delta_D) = \inf_{\delta \in D} \sup_{\theta \in \Theta} R(\theta, \delta)$.

The value on the right hand side is called the Minimax value or the Upper value of the game. If “inf” is replaced by “min” and “sup” is replaced by “max”, it is easy to see why the word minimax is coined.

Even if the minimax value is finite, there may not be a minimax decision rule so that the statistician has to be satisfied with a rule whose maximum risk is within ε of the minimax value.

Def: Let $\varepsilon > 0$. A decision rule δ_D is said to be ε-minimax if $\sup_{\theta \in \Theta} R(\theta, \delta_D) \leq \inf_{\delta \in D} \sup_{\theta \in \Theta} R(\theta, \delta) + \varepsilon$.

2. The Geometry of Bayes and Minimax Rules when the Parameter Space contains a Finite Number of Elements

We first need to bring in the notion of **Randomized Decision Rules**.

Suppose $\mathcal{D}$ is the space of all non-randomized decision rules. We extend $\mathcal{D}$ to $\mathcal{D}^*$ where $\mathcal{D}^*$ is the space of all probability distributions over $\mathcal{D}$.

Forexample, suppose $\mathcal{D} = \{\delta_1, \delta_2, \delta_3, \delta_4\}$. A typical element of $\mathcal{D}^*$ is a probability distribution δ^* such that δ^* assigns probability p_i to δ_i $(i = 1, 2, 3, 4)$, $p_i \geq 0$ and $\sum_{i=1}^4 p_i = 1$. In general, we shall write the risk function corresponding to δ^* as $R(\theta, \delta^*) = \mathbb{E} R(\theta, Y)$, where Y is a random variable assuming values in $\mathcal{D}$ given by δ^*.

E.g. Let $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{A} = \{a_1, a_2\}$. Let $\mathcal{D} = \{\delta_1, \delta_2, \delta_3, \delta_4\}$, where

$$\delta_1(x_1) = \delta_1(x_2) = a_1, \delta_2(x_1) = a_1, \delta_2(x_2) = a_2$$

$$\delta_3(x_1) = a_2, \delta_3(x_2) = a_1, \delta_4(x_1) = \delta_4(x_2) = a_2$$

Then $\mathcal{D}^*$ is in one-to-one correspondence with $\{(\zeta_1, \zeta_2, \zeta_3, \zeta_4) : \zeta_i \geq 0,\ \sum_{i=1}^4 \zeta_i = 1\}$, where δ_i is chosen with probability ζ_i. Also, let $\xi = \xi(\theta) = P_\theta(X = x_1) = 1 - P_\theta(X = x_2)$. Then

$$\begin{aligned} R(\theta, \delta^*) &= E_\theta L(\theta, \delta^*(x)) \\ &= \zeta_1 E_\theta L(\theta, \delta_1(x)) + \zeta_2 E_\theta L(\theta, \delta_2(x)) + \zeta_3 E_\theta L(\theta, \delta_3(x)) + \zeta_4 E_\theta L(\theta, \delta_4(x)) \\ &= \zeta_1 L(\theta, a_1) + \zeta_2 \left[\xi L(\theta, a_1) + (1 - \xi) L(\theta, a_2)\right] + \zeta_3 \left[\xi L(\theta, a_2) + (1 - \xi) L(\theta, a_1)\right] + \zeta_4 L(\theta, a_2) \\ &= \xi \left[L(\theta, a_1)(\zeta_1 + \zeta_2) + L(\theta, a_2)(\zeta_3 + \zeta_4)\right] + (1 - \xi) \left[L(\theta, a_1)(\zeta_1 + \zeta_3) + L(\theta, a_2)(\zeta_2 + \zeta_4)\right] \end{aligned}$$

We are now in a position to dicuss the geometry of Bayes and minimax rules when the parameter apace contains a finite number of elements.

Suppose $\Theta = \{\theta_1, \cdots, \theta_k\}$. Consider the **Risk Set** $S \in \mathbb{R}^k$ defined by

$$S = \{(y_1, \cdots, y_k) :\ y_j = R(\theta_j, \delta^*) \text{ for some } \delta^* \in \mathcal{D}^*;\ j = 1, \cdots, k\}.$$

We first show that the risk set S is a convex subset of $\mathbb{R}^k$.

Proof Let $\boldsymbol{Y} = (Y_1, \cdots, Y_k) \in S$, $\boldsymbol{Z} = (Z_1, \cdots, Z_k) \in S$. Hence $\exists$ decision rules δ_1^*, δ_2^* each $\in \mathcal{D}^*$ such that $y_j = R(\theta_j, \delta_1^*)$, $z_j = R(\theta_j, \delta_2^*)$, $j = 1, \cdots, k$. Let $0 \leq \alpha \leq 1$. Consider $\delta_\alpha \in \mathcal{D}^*$ which chooses non-randomized decision rules in $\mathcal{D}$ according to the distribution $\delta_\alpha = \alpha\delta_1^* + (1 - \alpha)\delta_2^*$. Then $R(\theta_j, \delta_\alpha) = \alpha R(\theta_j, \delta_1^*) + (1 - \alpha) R(\theta_j, \delta_2^*),\ j = 1, \cdots, k$, i.e. $\alpha\boldsymbol{Y} + (1 - \alpha)\boldsymbol{Z} \in S$. Hence S is a convex set. □

Geometry of Bayes rules and Minimax rules is referable to Page 12 & 13 in "Inference typed notes"

One advantage of the Bayes procedure is that the search for good decision rules may be restricted to the class of non-randomized decision rules.

More precisely, if a Bayes rule with respect to a prior distribution ξ exists, there exists a non-randomized Bayes rule with respect to ξ. A non-rigorous proof of this is as follows.

proof: Suppose $\delta_\xi \in \mathcal{D}^*$ is a Bayes rule with respect to a prior distribution ξ. Then

$$r(\xi, \delta_\xi) = \int_\Theta R(\theta, \delta_\xi)\, d\xi(\theta)$$

Let Y denote the random variable assuming values in $\mathcal{D}$ with distribution function given by δ_ξ. Then

$$R(\theta, \delta_\xi) = \mathbb{E}R(\theta, Y) = \int_{\mathcal{D}} R(\theta, \delta)\, d\delta_\xi(\delta)$$

$$\begin{aligned} r(\xi, \delta_\xi) &= \int_\Theta \int_{\mathcal{D}} R(\theta, \delta)\, d\delta_\xi(\delta) d\xi(\theta) \\ &= \int_{\mathcal{D}} \int_\Theta R(\theta, \delta)\, d\xi(\theta) d\delta_\xi(\delta) \quad (\text{assume that we can}) \\ &= \int_{\mathcal{D}} r(\xi, \delta)\, d\delta_\xi(\delta) \\ &= \int_{\mathcal{D}} r(\xi, \delta_\xi)\, d\delta_\xi(\delta) + \int_{\mathcal{D}} (r(\xi, \delta) - r(\xi, \delta_\xi))\, d\delta_\xi(\delta) \\ &= r(\xi, \delta_\xi) + \int_{\mathcal{D}} (r(\xi, \delta) - r(\xi, \delta_\xi))\, d\delta_\xi(\delta) \end{aligned}$$

Now δ_ξ is the Bayes rule w.r.t. ξ, $r(\xi, \delta_\xi) \leq r(\xi, \delta)$ for every $\delta \in \mathcal{D}$ (non-randomized). Thus $r(\xi, \delta) = r(\xi, \delta_\xi)$ with probability 1. This implies that δ is Bayes with respect to ξ.

3. Finding Bayes Rules

First consider the no data problem. Suppose the loss involved in taking action a when θ is the true parameter is $L(\theta, a)$. Let the prior pdf be given by $g(\theta)$. Then the Bayes procedure consists in finding and a_0 (if possible) which minimizes $\int L(\theta, a) g(\theta)\, d\theta$ w.r.t. a.

Suppose now X is a random variable with pdf $f_\theta(x)$. If ξ denote the prior distribution with pdf $g(\theta)$, the Bayes risk of a decision rule δ w.r.t. the prior ξ is given by

$$r(\xi, \delta) = \int_\Theta R(\theta, \delta) g(\theta)\, d\theta = \int_\Theta \int_{\mathcal{X}} L(\theta, \delta(x)) f_\theta(x) g(\theta)\, dx d\theta.$$

In this case, we interpret $f_\theta(x)$ as the conditional pdf of X given θ. Hence $f_\theta(x) g(\theta) =$ joint pdf of X and W, where W is a random variable assuming values θ in Θ. An alternate way of writing $f_\theta(x) g(\theta)$ is $p(\theta|x) h(x)$ where $p(\theta|x)$ is the conditional pdf of W given $X = x$, where $h(x)$ is the marginal pdf of X. To find a decision rule δ_ξ for which the Bayes risk is minimized, it suffices in finding a $\delta_\xi(x)$ (if possible) which minimizes $\int_\Theta L(\theta, \delta(x)) p(\theta|x)\, d\theta$ for each $x \in \mathcal{X}$.

If for each x, $\delta_\xi(x)$ is unique and $r(\xi, \delta_\xi)$ is finite then we say the Bayes decision rule is **unique**.

The problem of determining Bayes procedure arises in a number of contexts.

(i) As a way of Utilizing Past Experience

It is frequently reasonable to treat the parameter θ of a statistical problem as the realization of a random variable W, rather than an unknown constant. Suppose, for example, that we wish to estimate the probability of a penny showing heads when spun on a flat surface. So far we would have considered n spins of the penny as a set of n binomial trials with an unknown probability p of showing heads. Suppose, however, we have had considerable experience with spinning pennies, which has provided us with approximate values of p for a large number of similar experiments. if we believe this experience to be relevant to the present penny, it might be reasonable to represent past knowledge as a probability distribution for p, the approximate shape of which is suggested by the earlier data.

This is not unlike the modeling which is usually done in statistics. An assumption such as the random variables representing the outcomes of our experiments have normal, Poisson, exponential distributions, and so on, we also draw on past experience. Futhermore, we also realize that these methods are in no sense exact, but at least represent reasonable approximations. There is the difference that in non-Bayesian parametric models, normally only the shape of the distribution is assumed to be known, but the values of the parameters are not assuming to be known. In a Bayesian analysis, however, the prior distribution is usually taken as completely specified. However, this is a difference in degree rather than in kind, and may be quite reasonable if the past experience is sufficiently extensive.

A difficulty, of course, is the assumption that past experience is relevant to the present case. Perhaps, the mint has recently changed its manufacturing process, and the present coin, though similar in appearance as the earlier ones, has totally different spinning properties. But, a similar difficulty is associated with a non-Bayesian model.

The choice of a prior distribution $\xi(\theta)$ is typically made by combining experience with conve- nience. When we assume that the amount of rainfall has a gamma distribution, we don't do so because we really believe this to be the case, but because the gamma family is a two-parameter family which seems to fit such a prior distribution by starting with a flexible family that is math- ematically easy to handle and selecting a member from this family which approximates our past experience. Such an approach, in which the model incorporates a prior distribution for θ to reflect past experience is useful in fields in which a large amount of past experience is available. It can be brought to beer, for example, in many applications in education, business and medicine.

There is one important difference between the modeling of the distribution P_θ and the prior distribution $\xi(\theta)$ of θ. Typically, we have a number of observations from P_θ, and can use these to check the assumption of the form of the distribution. Such a check of ξ is not possible on the basis of one experiment because the value of θ under study represents only a single observation from this distribution. This requires special safeguards in the case of a Bayesian analysis.

Another difference concerns the meaning of a replication of the experiment. In a non-Bayesian analysis, a replication consists of drawing another set of observations from P_θ with the same value of θ. In a Bayesian analysis, we replicate the experiment by first drawing another value say θ' from Θ under ξ, and then a set of observations from $P_{\theta'}$.

(ii) As a description of a State of Mind

A formally similar approach is adopted by the so-called neo-Bayesian school which interprets ξ as expressing the subjective's feeling about the likelihood of different θ-values. In the presence of a large amount of previous experience, the chosen ξ is often close to the one discussed before, but the subjective approach can be applied even when little or no prior knowledge is available. Thus, with the present interpretations, the prior distribution ξ models the state of ignorance about θ. The subjective Bayesian uses the observations to modify prior beliefs. After $X = x$ is observed, the belief about θ is expressed by the posterior (i.e. conditional) distribution of θ given $X = x$.

(iii) As Mathematical Tools

In addition to the above, Bayesian tools have become essential in many decision theoretics analysis. We shall see this later in this Chapter in the context of minimax and admissible proce- dures.

In the example considered earlier, for the no-data problem in finding the Bayes risk, one minimizes $\int_\Theta L(\theta, a)g(\theta)\,d\theta$ w.r.t. a, whereas if $X = x$ is observed, one minimizes $\int L(\theta, a)P(\theta|x)\,d\theta$ w.r.t. a. In the second case, the desired a will be a function of x as one might anticipate.

Estimation Problem:

$$\theta \in \Theta \subset \mathbb{R}^1, \quad a \in \mathcal{A} \subset \mathbb{R}^1$$
$$\text{Squared Error Loss: } L(\theta, a) = (a-\theta)^2$$
$$\text{Absolute Error Loss: } L(\theta, a) = |a-\theta|$$

1. Squared Error Loss: Assuming that the prior distribution has finite second moment. Let $m = \text{prior mean of } \xi = \int_\Theta \theta g(\theta)\,d\theta$, g being the prior pdf. Then

$$\begin{aligned}\int_\Theta (\theta-a)^2 g(\theta)\,d\theta &= \int_\Theta (\theta - m + m - a)^2 g(\theta)\,d\theta \\ &= \int_\Theta (\theta-m)^2 g(\theta)\,d\theta + (m-a)^2 \\ &\geq \int_\Theta (\theta-m)^2 g(\theta)\,d\theta\end{aligned}$$

   with equality if and only if $a = m$.

   *Remark: Thus, for the no-data problem, the Bayes estimate of $\theta = \int \theta g(\theta)\,d\theta$, the prior mean. If $X = x$ is available, the Bayes estimate of θ is $\int \theta p(\theta|x)\,d\theta$, the posterior mean. In general, if we want to estimate $\gamma(\theta)$, some real-valued function of θ, the Bayes estimate is given by

$$\int_\Theta \gamma(\theta) g(\theta)\,d\theta \quad \text{for the on-data problem}$$
$$\int_\Theta \gamma(\theta) p(\theta|x)\,d\theta \quad \text{if } X = x \text{ is observed}$$

   - Eg1: $X_1, \cdots, X_n$ i.i.d. Bernoulli(θ), $\theta \in [0,1]$, $\gamma(\theta) = \theta$. Suppose the prior distribution is Beta(α, β), $g(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$. Find the Bayes estimate of θ.

     $f_\theta(x_1, \cdots, x_n) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$. Then $p(\theta|x_1, \cdots, x_n) \propto \theta^{\sum x_i + \alpha - 1}(1-\theta)^{n-\sum x_i + \beta - 1}$. The posterior distribution is Beta($\sum x_i + \alpha, n - \sum x_i + \beta$).

     The Bayes estimate of θ under squared error loss is $\frac{\sum x_i + \alpha}{n+\alpha+\beta}$, which can be written as

$$\frac{\sum x_i + \alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta}\frac{\sum x_i}{n} + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta}$$

     which is a weighted average of the sample mean and the prior mean. The estimate leans more towards the sample mean if $\alpha + \beta$ is small compared to n, and to the prior mean $\frac{\alpha}{\alpha+\beta}$ if n is relatively small compared to $\alpha + \beta$.

   - Eg2: $X_1, \cdots, X_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$, where $\theta \in \mathbb{R}$ is unknown, but $\sigma^2 (> 0)$ is known. Suppose the prior distribution is $\mathcal{N}(\mu, \tau^2)$. Find the Bayes estimate of θ.

$$\begin{aligned} f_\theta(x_1,\cdots,x_n) &\propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2} \\ &= e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i-\bar{x})^2+n(\bar{x}-\theta)^2\right]} \\ &\propto e^{-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}} \end{aligned}$$

$$p(\theta|x_1,\cdots,x_n) \propto e^{-\frac{1}{2}\left[\frac{n(\bar{x}-\theta)^2}{\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}\right]} \propto e^{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta-\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)^2}$$

The posterior distribution is $\mathcal{N}(\frac{\frac{n\bar{X}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}})$. Under the squared error loss, the Bayes estimate of θ is $\frac{\frac{n\bar{X}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} = (1-B)\bar{x}+B\mu, B=\frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}$.

- Remark: In the above example, the posterior distribution of the parameters depends on the X_i's only through the value of the minimual sufficient statistic.

We go back to Eg2. In this example, $\bar{X}$ is sufficient for θ and $\bar{X}|\theta \sim \mathcal{N}(\theta, \sigma^2/n)$, while $\theta \sim \mathcal{N}(\mu,\tau^2)$.

$$\begin{pmatrix}\bar{X}\\ \theta\end{pmatrix} \overset{d}{=} \begin{pmatrix}\theta+\varepsilon\\ \theta\end{pmatrix}, \quad \text{where } \varepsilon\sim\mathcal{N}(0,\sigma^2/n), \theta \text{ and } \varepsilon \text{ are independent.}$$

Then

$$\begin{pmatrix}\bar{X}\\ \theta\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix}, \begin{pmatrix}\tau^2+\sigma^2/n & \tau^2\\ \tau^2 & \tau^2\end{pmatrix}\right)$$

$$\theta|\bar{X} \sim \mathcal{N}\left(\mu+\frac{\tau^2}{\tau^2+\sigma^2/n}(\bar{X}-\mu), \tau^2-\frac{\tau^4}{\tau^2+\sigma^2/n}\right)$$

i.e.

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

- Eg3: Suppose $X_1,\cdots,X_n (n\geq 2)$ are i.i.d. $\mathcal{N}(\theta,\sigma^2)$, where $\theta\in\mathbb{R}$ and $\sigma^2>0$ are both unknown. $r=\sigma^{-2}$ is called the "precision" parameter.

  Suppose the prior of r follows a gamma distribution $\left(\frac{\beta}{2}, \frac{2}{\alpha}\right)$, and the conditional (prior) distribution of θ given $R=r$ is $\mathcal{N}(\mu, \frac{1}{\lambda r})$, where $\lambda>0$ and $\mu\in\mathbb{R}^n$ are known.

  Find the Bayes estimator for θ and r.

$$g_{\alpha,\beta}(r) \propto \exp\left\{-\frac{1}{2}\alpha r\right\} r^{\beta/2-1}, \quad r>0, \alpha>0, \beta>0$$

$$\text{prior distribution: } g_{\lambda,\alpha,\beta}(\theta|r) \propto (\lambda r)^{1/2}\exp\left\{-\frac{\lambda r}{2}(\theta-\mu)^2\right\}, \quad \lambda>0,\alpha>0,\beta>0$$

  Note that $\left(\bar{X}, \sum_{i=1}^n (X_i-\bar{X})^2\right)$is the minimal sufficient statistic for (θ, r). Let $U=\bar{X}, S=\sum_{i=1}^n (X_i-\bar{X})^2$.

$$f(u,s) \propto r^{1/2} e^{-\frac{nr}{2}(u-\theta)^2} \cdot e^{-\frac{rs}{2}} s^{\frac{n-1}{2}-1} r^{\frac{n-1}{2}}, \quad s>0, u\in\mathbb{R}$$

$$p(\theta, r|u,s) \propto r^{n/2+\beta/2-1+1/2}\exp\left\{-\frac{r}{2}[n(u-\theta)^2+\lambda(\theta-\mu)^2+s+\alpha]\right\}$$

  Now,

$$\begin{aligned} n(u-\theta)^2+\lambda(\theta-\mu)^2 &= (n+\lambda)\theta^2-2\theta(nu+\lambda\mu)+nu^2+\lambda\mu^2 \\ &= (n+\lambda)\left(\theta-\frac{nu+\lambda\mu}{n+\lambda}\right)^2+nu^2+\lambda\mu^2-\frac{(nu+\lambda\mu)^2}{n+\lambda} \\ &= (n+\lambda)\left(\theta-\frac{nu+\lambda\mu}{n+\lambda}\right)^2+\frac{n\lambda}{n+\lambda}(u-\mu)^2 \end{aligned}$$

So,

$$p(\theta,r|u,s)\propto r^{\frac{n+\beta+1}{2}-1}\exp\left\{-\frac{r}{2}(n+\lambda)\left(\theta-\frac{nu+\lambda\mu}{n+\lambda}\right)^2-\frac{r}{2}\left[s+\alpha+\frac{n\lambda}{n+\lambda}(u-\mu)^2\right]\right\}$$

Conditioned on $R=r, U=u$ and $S=s, \theta\sim\mathcal{N}\left(\frac{nu+\lambda\mu}{n+\lambda},\frac{1}{r(n+\lambda)}\right)$, and conditioned on $U=u$ and $S=s$, $R\sim Gamma\left(\frac{1}{2}(n+\beta),\left[\frac{1}{2}\left(\alpha+s+\frac{n\lambda}{n+\lambda}(u-\mu)^2\right)\right]^{-1}\right)$.

The Bayes estimator of θ is $\frac{nu+\lambda\mu}{n+\lambda}$, and the Bayes estimator of r is $\frac{n+\beta}{\alpha+s+\frac{\lambda n}{n+\lambda}(u-\mu)^2}$.

- Eg4: Suppose $X_1,\cdots,X_n$ are i.i.d. uniform on $(0,\theta)$, where $\theta>0$. The prior distribution of θ is Pareto with parameter α,β. Find the Bayes estimator of θ under squared error loss,

$$g_{\alpha,\beta}(\theta)=\frac{\beta\alpha^\beta}{\theta^{\beta+1}}\mathbb{1}_{\theta\geq\alpha}$$

The minimal sufficient statistic for θ is $T=\max(X_1,\cdots,X_n)$.

$$f_\theta(t)=\frac{nt^{n-1}}{\theta^n}\mathbb{1}_{(0<t<\theta)}$$

$$p(\theta|t)\propto\theta^{-(n+\beta+1)}\mathbb{1}_{(\theta\geq\max(\alpha,t))}$$

This is Pareto distribution with parameters $\max(\alpha,t)$ and $n+\beta$.

Under squared error loss, the Bayes estimator of θ is $\frac{(n+\beta)\max(\alpha,t)}{n+\beta-1}$.

*Sidenote: For a Pareto(α, β) random variable X, $\mathbb{E}(X)=\begin{cases}\frac{\beta\alpha}{\beta-1}, & \text{if } \beta>1\\ \infty, & \text{if } \beta=1\end{cases}$.

- Eg5: Suppose $X\sim Uniform(0,|\theta|^{-1})$, $1\leq|\theta|<\infty$. Suppose the prior distribution has pdf

$$g(\theta)=\begin{cases}\frac{1}{2}|\theta|^{-2}, & 1\leq|\theta|<\infty\\ 0, & \text{otherwise}\end{cases}$$

Find the Bayes estimator of θ under the squared error loss.

$$p(\theta|x)=\frac{1}{2}|\theta|^{-1}(-\log x)^{-1}\mathbb{1}_{(x\leq|\theta|^{-1}\leq1)}$$

$$\begin{aligned}\mathbb{E}(\theta|X) &= -\frac{1}{2\log X}\int_{x\leq|\theta|^{-1}\leq1}\theta|\theta|^{-1}\,d\theta \\ &= -\frac{1}{2\log X}\left(\int_1^{\frac{1}{X}}\theta\theta^{-1}\,d\theta+\int_{-\frac{1}{X}}^{-1}\theta(-\theta)\,d\theta\right) \\ &= \frac{1}{2\log X}\left(\frac{1}{X}-1-\left(\frac{1}{X}-1\right)\right) \\ &= 0\end{aligned}$$

So assuming squared error loss, the Bayes estimator of θ is 0 for all X.

Note that $R(\theta, \hat{\theta}) = (0-\theta)^2 = \theta^2$, then the Bayes risk of $\hat{\theta}$ w.r.t. the prior $g(\theta)$ is

$$\int_{1\le|\theta|<\infty} \theta^2 g(\theta)\, d\theta = \int_1^\infty \theta^2 \frac{1}{2}|\theta|^{-2}\, d\theta + \int_{-\infty}^1 \theta^2 \frac{1}{2}|\theta|^{-2}\, d\theta = \infty$$

Hence, although in this case, the posterior risk is minimized uniquely for every x, there does not exist a unique Bayes estimator of θ.

2. Quadratic Loss (Weighted Squared Error Loss):

   $L(\theta, a) = w(\theta)(\theta - a)^2$ where $w(\theta) > 0$ for all $\theta \in \Theta$. Assuming the prior pdf $g(\theta)$ for the no data problem, the Bayes estimate of θ is obtained by minimizing $\int w(\theta)(\theta-a)^2 g(\theta)\, d\theta$ w.r.t. a.

   Assuming this integral is well-defined, the Bayes estimator of θ is

   $$a_0 = \frac{\int_\Theta \theta w(\theta) g(\theta)\, d\theta}{\int_\Theta w(\theta) g(\theta)\, d\theta} \quad \text{provided the denominator is finite}$$

   If $X = x$ is observed, $\int w(\theta)(\theta-a)^2 p(\theta|x)\, d\theta$ is minimized at

   $$a_0 = \frac{\int_\Theta \theta w(\theta) p(\theta|x)\, d\theta}{\int_\Theta w(\theta) p(\theta|x)\, d\theta} \quad \text{provided the denominator is finite}$$

   - Eg1: Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(θ) where $\theta \in (0,1)$. Let $L(\theta, a) = \frac{(\theta-a)^2}{\theta(1-\theta)}$. So $w(\theta) = \frac{1}{\theta(1-\theta)}$. Assuming Beta($\alpha, \beta$) prior, we know already that $\theta|X_i = x_i (i = 1, \cdots, n) \sim Beta(\sum_{i=1}^n x_i + \alpha, n - \sum_{i=1}^n x_i + \beta)$, $w(\theta)p(\theta|x) \propto \theta^{\sum_{i=1}^n x_i + \alpha - 2}(1-\theta)^{n - \sum_{i=1}^n x_i + \beta - 2}$.

     (a) If $\sum_{i=1}^n x_i + \alpha - 1 > 0$ and $n - \sum_{i=1}^n x_i + \beta - 1 > 0$. This is $Beta(\sum_{i=1}^n x_i + \alpha - 1, n - \sum_{i=1}^n x_i + \beta - 1)$. In this case, the Bayes estimator of θ is $\frac{\sum_{i=1}^n x_i + \alpha - 1}{n + \alpha + \beta - 2}$.

     (b) If $0 < \alpha \le 1$, $\sum_{i=1}^n x_i = 0$, $w(\theta)p(\theta|x) \propto \theta^{\alpha-2}(1-\theta)^{n+\beta-2}$. We want to minimize $\int_0^1 (\theta-a)^2 w(\theta)p(\theta|x)\, d\theta$. Note that

     $$\int_0^1 (\theta-a)^2 \theta^{\alpha-2}(1-\theta)^{n+\beta-2}\, d\theta \begin{cases} = \infty, & \text{if } a \neq 0 \\ < \infty, & \text{if } a = 0 \end{cases}$$

     So the Bayes estimate of $\theta = 0$.

     (c) If $0 < \beta \le 1$, $\sum_{i=1}^n x_i = n$, $w(\theta)p(\theta|x) \propto \theta^{n+\alpha-2}(1-\theta)^{\beta-2}$. We want to minimize $\int_0^1 (\theta-a)^2 w(\theta)p(\theta|x)\, d\theta$. Note that

     $$\int_0^1 (\theta-a)^2 \theta^{n+\alpha-2}(1-\theta)^{\beta-2}\, d\theta \begin{cases} = \infty, & \text{if } a \neq 1 \\ < \infty, & \text{if } a = 1 \end{cases}$$

     So the Bayes estimate of $\theta = 1$.

     (d) In the special case of Uniform$(0,1)$ prior, i.e., $\alpha = \beta = 1$, if follows that the Bayes estimate of θ is $\sum_{i=1}^n x_i / n$.

3. Absolute Error Loss:

   $L(\theta, a) = |\theta - a|$

   If the loss is absolute error loss, for the no-data problem, $\int |\theta - a| g(\theta)\, d\theta$ is minimized w.r.t. a is a is a median of the prior distribution. When $X = x$ is observed, the Bayes estimate is a median of the posterior distribution.

We give below the definition of a median and then prove the result that ensures the previous statements.

Def: Let Z be a real-valued random varibale. M is called a median of z is $P(Z < M) \leq 1/2 \leq P(Z \leq M)$.

Lemma: Suppose Z has a median M and $\mathbb{E}|Z| < \infty$. Then $\mathbb{E}|Z - M| = \inf_a \mathbb{E}|Z - a|$.

proof: Since $\mathbb{E}|Z| < \infty, \mathbb{E}|Z - a| \leq \mathbb{E}(|Z| + |a|) = \mathbb{E}|Z| + |a| < \infty$ for every real a.

Consider first the case when $a > M$. Then

$$|Z - a| - |Z - M| = \begin{cases} M - a, & \text{if } M < a \leq Z \\ M + a - 2Z, & \text{if } M < Z < a \\ a - M & \text{if } Z \leq M < a \end{cases}$$

Hence,

$$\begin{aligned} \mathbb{E}(|Z - a| - |Z - M|) &= (M - a)P(Z \geq a) + \int_M^a (M + a - 2Z)\, dF(Z) + (a - M)P(Z \leq M) \\ &\geq (M - a)P(Z \geq a) + \int_M^a (M - a)\, dF(Z) + (a - M)P(Z \leq M) \\ &= (M - a)P(Z > M) + (a - M)P(Z \leq M) \\ &= (a - M)[2P(Z \leq M) - 1] \\ &\geq (a - M)(2 \cdot \frac{1}{2} - 1) = 0 \end{aligned}$$

The case $a < M$ is similar; we will also have $\mathbb{E}(|Z - a| - |Z - M|) \geq 0$.

Thus $\mathbb{E}|Z - a|$ is minimized at $a = M$.

- Eg1: Suppose $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\theta, \sigma^2)$, where $\theta \in \mathbb{R}$ is unknown but $\sigma^2 (> 0)$ is known. Assuming the $\mathcal{N}(\mu, \tau^2)$ prior, the posterior distribution of θ given $X_i = x_i (1 \leq i \leq n)$ is $\mathcal{N}\left(\frac{\frac{n\bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right)$.

  Hence, using the absolute error loss, the Bayes estimate of θ is $\frac{\frac{n\bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$.

Bayesness and Unbiasedness

Theorem (Blackwell and Girshirk): Let Θ be an open or closed interval in the real line (can be $(-\infty, \infty)$), and let $L(\theta, a) = w(\theta)(\theta - a)^2$ ($w(\theta) > 0$ for all $\theta \in \Theta$). If a Bayes estimator of $\delta_\xi(x)$ of θ w.r.t. ξ is also an unbiased estimator of θ and $\mathbb{E}_\xi[w(\theta)] < \infty$, then the Bayes risk of δ_ξ w.r.t. ξ, namely $r(\xi, \delta_\xi) = 0$.

*Remark 1: In Particular, if $w(\theta) = 1$ for all $\theta \in \Theta$, i.e., the loss is squared error loss, the above theorem says that if a Bayes estimator is unbiased, its Bayes risk is zero.

*Remark 2: The message of the above theorem is that unbiased estimator are not typically Bayes w.r.t. any proper prior. As an example, let $X_1, \cdots, X_n$ be i.i.d. $\mathcal{N}(\theta, 1)$. Then $\bar{X}$ is an unbiased estimator of θ and $\mathbb{E}_\theta(\bar{X} - \theta)^n = \frac{1}{n}$, so that the Bayes risk of $\bar{X}$ w.r.t. any proper prior is $\frac{1}{n} \neq 0$, under squared error loss. Hence, $\bar{X}$ cannot be a Bayes estimator of θ w.r.t. any proper prior.

*Remark 3: The above theorem is not true if the condition $\mathbb{E}_\xi(w(\theta)) < \infty$ is dropped. Consider again the Bernoulli situation, but suppose that the loss is $L(\theta, a) = \frac{(\theta - a)^2}{\theta(1 - \theta)}$. In this case, $w(\theta) = 1/\theta(1 - \theta)$. Hence, assuming Uniform$(0, 1)$ prior, $\mathbb{E}_\xi(w(\theta)) = \int_0^1 \frac{1}{\theta(1-\theta)}\, d\theta = \infty$. We have already seen that in this case, the Bayes estimator of θ is $\bar{X}$.

*Remark 4: This result was generalized later to any arbitrary convex loss by Bicked & Blackwell (Ann. Math. Stat., 1967).

proof:

$$
\begin{aligned}
r(\xi, \delta_\xi) &= \int_\Theta \int_\mathcal{X} w(\theta)[\delta_\xi(x) - \theta]^2 f_\theta(x)\xi(\theta)\, dxd\theta \\
&= \int_\Theta \int_\mathcal{X} w(\theta)\delta_\xi^2(x) f_\theta(x)\xi(\theta)\, dxd\theta + \int_\Theta \int_\mathcal{X} w(\theta)\theta^2 f_\theta(x)\xi(\theta)\, dxd\theta - 2\int_\Theta \int_\mathcal{X} w(\theta)\theta\delta_\xi(x) f_\theta(x)\xi(\theta)\, dxd\theta \quad (1)
\end{aligned}
$$

$$
\begin{aligned}
\int_\Theta \int_\mathcal{X} w(\theta)\theta^2 f_\theta(x)\xi(\theta)\, dxd\theta &= \int_\Theta w(\theta)\theta^2\xi(\theta)\, d\theta \\
&= \int_\Theta w(\theta)\theta \int_\mathcal{X} \delta_\xi(x) f_\theta(x)\, dx\, \xi(\theta)\, d\theta \\
&= \int_\Theta \int_\mathcal{X} w(\theta)\theta\delta_\xi(x) f_\theta(x)\xi(\theta)\, dxd\theta \quad (2) \\
&= \int_\Theta \int_\mathcal{X} w(\theta)\theta\delta_\xi(x) f_\theta(x) p(\theta|x) h(x)\, dxd\theta \\
&= \int_\mathcal{X} \delta_\xi(x)\delta_\xi(x) \int_\Theta w(\theta) p(\theta|x)\, d\theta\, h(x)\, dx \\
&= \int_\Theta \int_\mathcal{X} w(\theta)\delta_\xi^2(x) p(\theta|x) h(x)\, dxd\theta \\
&= \int_\Theta \int_\mathcal{X} w(\theta)\delta_\xi^2(x) f_\theta(x)\xi(\theta)\, dxd\theta \quad (3)
\end{aligned}
$$

Substitute (2) and (3) to (1), we have $r(\xi, \delta_\xi) = 0$.

4. Convex Loss and Non-Randomized Decision Rules

Result: The function $f(x) = \|x\|^r$, $r \geq 1$ is a convex function, where $\|\cdot\|$ denote the Euclidean norm.

proof: First show that $\|\alpha x + (1-\alpha)y\| \leq \alpha\|x\| + (1-\alpha)\|y\|$.

Note that

$$
\begin{aligned}
&\|\alpha x + (1-\alpha)y\|^2 - (\alpha\|x\| + (1-\alpha)\|y\|)^2 \\
=&\alpha^2\|x\|^2 + (1-\alpha)^2\|y\|^2 + 2\alpha(1-\alpha)\langle x, y\rangle - \left[\alpha^2\|x\|^2 + (1-\alpha)^2\|y\|^2 + 2\alpha(1-\alpha)\|x\|\|y\|\right] \\
=&2\alpha(1-\alpha)\left(\langle x, y\rangle - \|x\|\|y\|\right) \leq 0
\end{aligned}
$$

Thus $\|\alpha x + (1-\alpha)y\| \leq \alpha\|x\| + (1-\alpha)\|y\|$.

When $r \geq 1$, it suffices to show that $[\alpha\|x\| + (1-\alpha)\|y\|]^r \leq \alpha\|x\|^r + (1-\alpha)\|y\|^r$.

Define a random variable Z such that $P(Z = \|x\|) = \alpha = 1 - P(Z = \|y\|)$. Then , we have

$$
\begin{aligned}
[\alpha\|x\| + (1-\alpha)\|y\|]^r &= (\mathbb{E}Z)^r \\
\alpha\|x\|^r + (1-\alpha)\|y\|^r &= \mathbb{E}(Z^r)
\end{aligned}
$$

Since Z^r is a convex function of $Z(\geq 0)$ for $r \geq 1$, applying Jensen's inequality, one gets $\mathbb{E}(Z^r) \geq (\mathbb{E}Z)^r$.

Typically the loss function that we consider are of the form $L(\theta, a) = \|\theta - a\|^r (r \geq 1)$, which for fixed θ are convex functions of a. Typically $r = 2$ or 1.

Of course, one can view $L(\theta, a)$ as a convex function of θ for a fixed due to the symmetry of the loss.

It can be shown that for convex loss, or more specifically, for $L(\theta, a)$ which are convex in a for fixed θ, any randomized rule can be improved (in terms of risk) uniformly by a non-randomized rule. Hence, for such a loss, in particular, for squared error loss, we will only use non-randomized rules.

(Justification: Let δ^* be a randomized decision rule in $\mathcal{D}^*$ for which $\int |a| \, dF(\delta) < \infty$, where $F(\delta)$ is the distribution on $\mathcal{D}^*$ that corresponds to δ^*. Let $\delta(x) = \int a \, dF(\delta)$. Then $L(\theta, \delta(x)) \leq \int L(\theta, a) \, dF(\delta) = L(\theta, \delta^*(x))$.)

For a convex loss, it is also possible to prove a generalized version of the Rao-Blackwell Theorem.

Theorem (Rao-Blackwell): Let W be any unbiased estimator of $r(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = E(W|T)$. Then $\mathbb{E}_\theta \phi(T) = r(\theta)$ and $Var(\phi(T)) \leq Var(W)$.

Theorem: Let $\gamma(\theta)$ be an estimable parameter function. Ler T be a complete sufficient statistic for θ. Then there exists a unique unbiased estimator of $\gamma(\theta)$ based on T, which has the smallest risk under any convex loss among all unbiased estimator of $\gamma(\theta)$.

proof: Suppose $g(X)$ is the unbiased estimator of $\gamma(\theta)$ and let $h(T) = \mathbb{E}(g(X)|T)$. Then for any convex loss $L(\gamma(\theta), a)$, convex in a,

$$\begin{aligned} R(\gamma(\theta), g(X)) &= \mathbb{E}(L(\gamma(\theta), g(X))) \\ &= \mathbb{E}\mathbb{E}[L(\gamma(\theta), g(X))|T] \\ &\geq \mathbb{E}[L(\gamma(\theta, \mathbb{E}(g(X)|T)))] \\ &= \mathbb{E}[L(\gamma(\theta), h(T))] \\ &= R(\gamma(\theta), h(T)). \end{aligned}$$

Suppose we have $h_1(T)$ and $h_2(T)$. $\mathbb{E}(h_1(T)) = \mathbb{E}(h_2(T)) = \gamma(\theta)$, i.e., $\mathbb{E}(h_1(T) - h_2(T)) = 0$. By completeness, $h_1(T) = h_2(T)$.

5. Minimax Decision Rules

Recall the definition of a minimax decision rule. The statistician wants to find a decision rule δ_0^* which minimizes the worst that can happen to him, i.e., a δ_0^* which minimizes $\sup_{\theta \in \Theta} R(\theta, \delta^*)$. For simplicity, assume the supremum is attained at $\theta' \in \Theta$. Let Ξ be the class of all prior distribution on Θ. Then the following lemma holds.

Lemma: $\sup_{\theta \in \Theta} R(\theta, \delta^*) = \sup_{\xi \in \Xi} \int_\Theta R(\theta, \delta^*) \, d\xi(\theta) \overset{def}{=} \sup_{\xi \in \Xi} r(\xi, \delta^*)$.

proof:

$$\int_\Theta R(\theta, \delta^*) \, d\xi(\theta) \leq \int_\Theta \sup_{\theta \in \Theta} R(\theta, \delta^*) \, d\xi(\theta) = \sup_{\theta \in \Theta} R(\theta, \delta^*).$$

This implies

$$\sup_{\xi \in \Xi} r(\xi, \delta^*) = \sup_{\xi \in \Xi} \int R(\theta, \delta^*) \, d\xi(\theta) \leq \sup_{\theta \in \Theta} R(\theta, \delta^*) \qquad (1)$$

By assumption, $\sup_{\theta \in \Theta} R(\theta, \delta^*) = R(\theta', \delta^*)$. Let ξ' denote a prior such that $\xi'(\{\theta'\}) = 1$. Hence

$$R(\theta', \delta^*) = \int_\Theta R(\theta, \delta^*)\, d\xi'(\theta) \leq \sup_{\xi \in \Xi} \int_\Theta R(\theta, \delta^*)\, d\xi(\theta)$$

i.e.

$$\sup_{\theta \in \Theta} R(\theta, \delta^*) \leq \sup_{\xi \in \Xi} r(\xi, \delta^*) \qquad (2)$$

Combining (1) and (2), the lemma follows.

Thus a statistician's aim is to find a δ_0^* minimizing $\sup_{\xi \in \Xi} \int_\Theta R(\theta, \delta^*)\, d\xi(\theta) = \sup_{\xi \in \Xi} r(\xi, \delta^*)$.

Conversely, nature's aim is to choose a ξ which maximizes $\inf_{\delta^* \in \mathcal{D}^*} \int_{\theta \in \Theta} R(\theta, \delta^*)\, d\xi(\theta) = \inf_{\delta^* \in \mathcal{D}^*} r(\xi, \delta^*)$.

Def: A distribution $\xi_0 \in \Xi$ is said to be *least favorable* if

$$\inf_{\delta^* \in \mathcal{D}^*} r(\xi_0, \delta^*) = \sup_{\xi \in \Xi} \inf_{\delta^* \in \mathcal{D}^*} r(\xi, \delta^*).$$

The value on the RHS is called the *maximin or lower value of the game.*

The name "least favorable" derives from the fact that if the statistician were told which prior distribution nature was using, he would like least to be told a distribution ξ_0 satistisfying the above definition. We may note that a least favorable distribution ξ_0 does not necessarily exist.

The fundamental theorem of game (minimax theorem) states that under certain assumption (Von Newmann (1928), Sion (1958) On general minimax theorem)

$$\inf_{\delta^* \in \mathcal{D}^*} \sup_{\xi \in \Xi} r(\xi, \delta^*) = \sup_{\xi \in \Xi} \inf_{\delta^* \in \mathcal{D}^*} r(\xi, \delta^*),$$

i.e.,

$$\inf_{\delta^* \in \mathcal{D}^*} \sup_{\theta \in \Theta} R(\theta, \delta^*) = \sup_{\xi \in \Xi} \inf_{\delta^* \in \mathcal{D}^*} r(\xi, \delta^*).$$

Thus **a Bayes rule corresponding to a least favorable prior is minimax**. However, a least favorable prior need not necessarily exist, or even though it may exist, there is no direct way of putting hands on it. Hence, a general approach of finding minimax rules may not exist. We propose below a few techniques.

Theorem (The Bayes method): Suppose that there is a distribution ξ over Θ such that
$r(\xi, \delta_\xi) = \int R(\theta, \delta_\xi)\, d\xi(\theta) = \sup_{\theta \in \Theta} R(\theta, \delta_\xi)$. Then

1. δ_ξ is minimax
2. If δ_ξ is unique Bayes w.r.t. ξ, then it is the unique minimax procedure
3. ξ is least favorable

proof:

1. Suppose δ^* is a decision rule (randomized or non-randomized). Then

$$\sup_{\theta\in\Theta} R(\theta,\delta_\xi) = \int_\Theta R(\theta,\delta_\xi)\,d\xi(\theta) \le \int_\Theta R(\theta,\delta^*)\,d\xi(\theta) \le \sup_{\theta\in\Theta} R(\theta,\delta^*).$$

Hence, δ_ξ is minimax.

2. The proof follows by replacing $\le$ by $<$ in the first inequality in the first inequality above.
3. Let ξ^* be a prior different from ξ, and let δ_{ξ^*} denote the corresponding Bayes rule. Then

$$\begin{aligned} r(\xi^*,\delta_{\xi^*}) &= \int_\Theta R(\theta,\delta_{\xi^*})\,d\xi^*(\theta) \\ &\le \int_\Theta R(\theta,\delta_\xi)\,d\xi^*(\theta) \\ &\le \sup_{\theta\in\Theta} R(\theta,\delta_\xi) = r(\xi,\delta_\xi) \end{aligned}$$

Hence, ξ is least favorable.

Corollary: If a Bayes rule δ_ξ w.r.t. a prior ξ has constant risk $R(\theta,\delta_\xi) = r$ for all $\theta\in\Theta$, then δ_ξ is minimax, and ξ is least favorable.

Example: Let $X_1,\cdots,X_n$ be i.i.d. Bernoulli(θ), $\theta\in[0,1]$. Assuming the squared error loss and the Beta(α,β) prior, the Bayes estimator of θ is $\delta = \frac{n\bar X+\alpha}{n+\alpha+\beta}$, where $\bar X = \sum_{i=1}^n X_i/n$. Then

$$\begin{aligned} R(\theta,\delta) &= \mathbb{E}\left[\frac{n\bar X+\alpha}{n+\alpha+\beta} - \theta\right]^2 \\ &= V\left(\frac{n\bar X+\alpha}{n+\alpha+\beta}\right) + \left(\mathbb{E}\left(\frac{n\bar X+\alpha}{n+\alpha+\beta}\right) - \theta\right)^2 \\ &= \frac{n\theta(1-\theta)}{(n+\alpha+\beta)^2} + \left(\frac{n\theta+\alpha}{n+\alpha+\beta} - \theta\right)^2 \\ &= \frac{n\theta(1-\theta) + (\alpha-\theta(\alpha+\beta))^2}{(n+\alpha+\beta)^2} \\ &= \frac{((\alpha+\beta)^2-n)\theta^2 + (n-2\alpha(\alpha+\beta))\theta + \alpha^2}{(n+\alpha+\beta)^2} \end{aligned}$$

In order that $R(\theta,\delta)$ is free from θ, we must have

$$\begin{cases} (\alpha+\beta)^2 = n \\ 2\alpha(\alpha+\beta) = n \end{cases} \Rightarrow \begin{cases} \alpha = \sqrt{n}/2 \\ \beta = \sqrt{n}/2 \end{cases}$$

Then $\frac{n\bar X+\sqrt{n}/2}{n+\sqrt{n}}$ is the Bayes estimator of θ w.r.t. the Beta$(\sqrt{n}/2,\sqrt{n}/2)$ prior and it has constant risk $\frac{n}{4(n+\sqrt{n})^2} = \frac{1}{4(\sqrt{n}+1)^2}$. So $\frac{n\bar X+\sqrt{n}/2}{n+\sqrt{n}}$ is a minimax estimator of θ.

Note that the UMVUE of θ is $\bar X$ amd it has risk $\frac{\theta(1-\theta)}{n}$. The corresponding maximum risk over $\theta\in[0,1]$ occurs at $\theta = 1/2$ and is given by $\frac{1}{4n} > \frac{1}{4(\sqrt{n}+1)^2}$. Hence $\bar X$ is not a minimax estimator of θ under squared error loss.

*Remark: Suppose, however we consider the weighted squared error loss $L(\theta,a) = (\theta-a)^2/\theta(1-\theta)$. We have already noted that the UMVUE $\bar X$ is the Bayes estimator of θ w.r.t. the Uniform$[0,1]$ prior. Also, under the above loss, $\bar X$ has constant risk $1/n$. Hence $\bar X$ is a minimax estimator of θ under the above loss.

Lemma 1: Suppose δ_0 is minimax when $\theta\in\Theta_0\subset\Theta$ and $\sup_{\theta\in\Theta} R(\theta,\delta_0) = \sup_{\theta\in\Theta_0} R(\theta,\delta_0)$. Then δ_0 is minimax even for $\theta\in\Theta$.

proof: For any decision rule δ',

$$\sup_{\theta\in\Theta} R(\theta,\delta') \geq \sup_{\theta\in\Theta_0} R(\theta,\delta') \geq \sup_{\theta\in\Theta_0} R(\theta,\delta_0) = \sup_{\theta\in\Theta} R(\theta,\delta_0).$$

Hence, δ_0 is minimax even when $\theta \in \Theta$.

Lemma 2: If X is a minimax estimator of θ under the squared error loss, $aX+b$ is a minimax estimator of $a\theta+b$ under squared error loss, where $a(\neq 0)$ and b are known constants.

proof: Call $\delta_0(X) = aX+b$. Suppose δ_0 is not a minimax estimator of $a\theta+b$ under squared error loss. Then there exists a $\delta(X)$ such that

$$\sup_{\theta\in\Theta} R(a\theta+b,\delta(X)) < \sup_{\theta\in\Theta} R(a\theta+b,\delta_0(X)),$$

i.e.,

$$\sup_{\theta\in\Theta} \mathbb{E}(\delta(X)-(aX+b))^2 < \sup_{\theta\in\Theta} \mathbb{E}(\delta_0(X)-(a\theta+b))^2 = a^2 \sup_{\theta\in\Theta} \mathbb{E}(X-\theta)^2$$

$$\Rightarrow \sup_{\theta\in\Theta} \mathbb{E}\left(\frac{\delta(X)-b}{a}-\theta\right)^2 < \sup_{\theta\in\Theta} \mathbb{E}(X-\theta)^2$$

which contradicts that X is a minimax estimator of θ under square error loss.

*Note: However, it is **NOT** necessarily true that if X is a minimax estimator of θ_1, Y is a minimax estimator of θ_2, then $aX+bY$ is a minimax estimator of $a\theta_1+b\theta_2$ where a and b are known constants. The following example illustrate this.

Example: Suppose $X \sim$ binomial(n,θ_1), $Y \sim$ binomial(n,θ_2), where $(\theta_1,\theta_2) \in [0,1]\times[0,1]$. X and Y are independent. Assuming squared error loss, a minimax estimator is wanted for $\theta_1-\theta_2$.

In view of a previous example, one might expect $\frac{X+\sqrt{n}/2}{n+\sqrt{n}} - \frac{Y+\sqrt{n}/2}{n+\sqrt{n}} = \frac{X-Y}{n+\sqrt{n}}$ is a minimax estimator of $\theta_1-\theta_2$. We shall see that this is not the case.

Consider the estimator $a(X-Y)$ for $\theta_1-\theta_2$, where a is a constant. The corresponding risk function is given by

$$\begin{aligned}\mathbb{E}[a(X-Y)-(\theta_1-\theta_2)]^2 &= a^2V(X-Y)+(an-1)^2(\theta_1-\theta_2)^2 \\ &= na^2\left(\theta_1(1-\theta_1)+\theta_2(1-\theta_2)\right)+(an-1)^2(\theta_1-\theta_2)^2 \\ &= g(\theta_1,\theta_2)\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial g}{\partial\theta_1} &= na^2(1-2\theta_1)+2(an-1)^2(\theta_1-\theta_2) \\ \frac{\partial g}{\partial\theta_2} &= na^2(1-2\theta_2)+2(an-1)^2(\theta_2-\theta_1)\end{aligned}$$

Setting $\frac{\partial g}{\partial\theta_1} = \frac{\partial g}{\partial\theta_2} = 0$ leads to $\theta_1+\theta_2 = 1$.

This suggests that the maximum of the risk occurs for those θ_1, θ_2 such that $\theta_1+\theta_2=1$. So we try to find Bayes estimators having constant risk over the subset Θ_0 given by $\Theta_0 = \{(\theta_1,\theta_2) : 0 \leq \theta_1 = 1-\theta_2 \leq 1\}$, hoping that the supremum of the risk such an estimator over Θ is attained in this subset.

Now for $(\theta_1, \theta_2) \in \Theta_0$, minimal sufficient statistic for θ_1 is $Z = X + n - Y$. Note that X and $n - Y$ are i.i.d. Binomial(n, θ_1) for $(\theta_1, \theta_2) \in \Theta_0$. Hence, for $(\theta_1, \theta_2) \in \Theta_0$, $Z \sim$ binomial$(2n, \theta_1)$. Using a previous example, a Bayes estimator of θ_1 having constant risk is given by $\frac{Z+\sqrt{2n}/2}{2n+\sqrt{2n}}$. Now using Lemma 2, the Bayes estimator of $2\theta_1 - 1 = \theta_1 - \theta_2$ is

$$\begin{aligned} 2\frac{Z + \sqrt{2n}/2}{2n + \sqrt{2n}} - 1 &= \frac{2(X + n - Y)}{2n + \sqrt{2n}} + \frac{\sqrt{2n}}{2n + \sqrt{2n}} - 1 \\ &= \frac{2(X - Y)}{2n + \sqrt{2n}} \\ &= \frac{\sqrt{2n}}{\sqrt{2n} + 1}\left(\frac{X}{n} - \frac{Y}{n}\right). \end{aligned}$$

Next we show that the supremum of the risk of the estimator $\frac{\sqrt{2n}}{\sqrt{2n}+1}\left(\frac{X}{n} - \frac{Y}{n}\right)$ of $\theta_1 - \theta_2$ over $\Theta = [0, 1] \times [0, 1]$ is attained at the subset Θ_0. With this end, first compute

$$\begin{aligned} \mathbb{E}\left[\frac{\sqrt{2n}}{\sqrt{2n} + 1}\left(\frac{X}{n} - \frac{Y}{n}\right) - (\theta_1 - \theta_2)\right]^2 &= V\left(\frac{\sqrt{2n}}{\sqrt{2n} + 1}\left(\frac{X}{n} - \frac{Y}{n}\right)\right) + \left(\mathbb{E}\left(\frac{\sqrt{2n}}{\sqrt{2n} + 1}\left(\frac{X}{n} - \frac{Y}{n}\right)\right) - (\theta_1 - \theta_2)\right)^2 \\ &= \frac{2n}{(\sqrt{2n} + 1)^2} \cdot \frac{\theta_1(1 - \theta_1) + \theta_2(1 - \theta_2)}{n} + \left(\frac{\sqrt{2n}}{\sqrt{2n} + 1}(\theta_1 - \theta_2) - (\theta_1 - \theta_2)\right)^2 \\ &= \frac{2\theta_1(1 - \theta_1) + 2\theta_2(1 - \theta_2) + (\theta_1 - \theta_2)^2}{(\sqrt{2n} + 1)^2} \\ &= \frac{1 - (1 - \theta_1 - \theta_2)^2}{(\sqrt{2n} + 1)^2} \end{aligned}$$

which is maximzed at $\theta_1 + \theta_2 = 1$. Hence $\frac{\sqrt{2n}}{\sqrt{2n}+1}\left(\frac{X}{n} - \frac{Y}{n}\right)$ is a minimax estimator of $\theta_1 - \theta_2$.

Example (Nonparametric Minimax Estimation): Let $X_1, \cdots, X_n$ be i.i.d. each having the common distribution function $F \in \mathcal{F} =$ the totality of all possible distributions on $[0, 1]$. Let $\theta(F) = \mathbb{E}_F(X_1)$. Assuming squared error loss, we need a minimax estimator of $\theta(F)$.

Let $\mathcal{F}_0 (\subset \mathcal{F})$ be the totality of binomial distribution for which $P(X_i = 1) = \theta = 1 - P(X_i = 0), \theta \in [0, 1]$. Using a previous example for $F \in \mathcal{F}_0$, the estimator $\frac{\sum X_i + \sqrt{n}/2}{n + \sqrt{n}}$ is a Bayes estimator of θ with constant risk $\frac{1}{4(\sqrt{n}+1)^2}$.

Now for any $F \in \mathcal{F}$,

$$\begin{aligned} \mathbb{E}_F\left[\frac{\sum_{i=1}^n X_i + \sqrt{n}/2}{n + \sqrt{n}} - \theta(F)\right]^2 &= V_F\left(\frac{\sum_{i=1}^n X_i + \sqrt{n}/2}{n + \sqrt{n}}\right) + \left[\mathbb{E}_F\frac{\sum_{i=1}^n X_i + \sqrt{n}/2}{n + \sqrt{n}} - \theta(F)\right]^2 \\ &= \frac{nV_F(X_1)}{(n + \sqrt{n})^2} + \frac{n}{(n + \sqrt{n})^2}\left(\frac{1}{2} - \theta(F)\right)^2 \\ &= \frac{V_F(X_1) + (1/2 - \theta(F))^2}{(\sqrt{n} + 1)^2} \end{aligned}$$

Since $V_F(X_1) = \mathbb{E}_F(X_1^2) - (\mathbb{E}_F(X_1))^2 \leq \mathbb{E}_F(X_1) - (\mathbb{E}_F(X_1))^2 = \theta(F) - \theta^2(F)$, we have

$$\mathbb{E}_F\left[\frac{\sum_{i=1}^n X_i + \sqrt{n}/2}{n + \sqrt{n}} - \theta(F)\right]^2 \leq \frac{\theta(F) - \theta^2(F) + (1/2 - \theta(F))^2}{(\sqrt{n} + 1)^2} = \frac{1}{4(\sqrt{n} + 1)^2}$$

Hence, supremum (over $F \in \mathcal{F}$) of the risk of the estimator $\frac{\sum_{i=1}^{n} X_i + \sqrt{n}/2}{n+\sqrt{n}}$ of $\theta(F)$ is attained when $F \in \mathcal{F}_0$. Using Lemma 1, $\frac{\sum_{i=1}^{n} X_i + \sqrt{n}/2}{n+\sqrt{n}}$ is a minimax estimator of $\theta(F)$.

As already noted, unbiased estimators are not typically Bayes w.r.t. any proper prior when the loss is squared error. Hence, using the Theorem on Bayes Method, minimaxity of such estimator **cannot** usually be proved. It turns out that very often unbiased estimators can be viewed as the limit of a sequence of Bayes estimators. In such cases, the following theorem will be useful in proving Minimaxity.

Theorem: Let $\{\xi_1, \xi_2, \cdots\}$ be a sequence of priors on Θ. Let $\delta_m = \delta_{\xi_m}$ be a Bayes rule w.r.t. ξ_m having Bayes risk $r_m = r(\xi_m, \delta_m)$. If δ_0 is a rule with $\sup_{\theta \in \Theta} R(\theta, \delta_0) = r$ and $\lim_{m \to \infty} r_m \geq r$, then δ_0 is minimax.

proof: For any rule $\delta^* \in \mathcal{D}^*$, $r_m = \int R(\theta, \delta_m)\, d\xi_m(\theta) \leq \int R(\theta, \delta^*)\, d\xi_m(\theta) \leq \sup_{\theta \in \Theta} R(\theta, \delta^*)$.

$r = \sup_{\theta \in \Theta} R(\theta, \delta_0) \leq \lim_{m \to \infty} r_m \leq \sup_{\theta \in \Theta} R(\theta, \delta^*)$. So δ_0 is minimax.

Corollary: If $r_m \to r = \sup_{\theta \in \Theta} R(\theta, \delta_0)$, then δ_0 is minimax.

Example: Let $X_1, \cdots, X_n$ be i.i.d. $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. Under the squared error loss, find a minimax estimator for θ.

Consider the sequence of prior ξ_m which are $\mathcal{N}(0, m)$, $m \geq 1$. Writing $X = (X_1, \cdots, X_n)$, under the squared error loss, the Bayes estimator of θ w.r.t. ξ_m is $\delta_m = \frac{n\bar{X}}{n+1/m}$.

$$\begin{aligned} r(\xi_m, \delta_m) &= \mathbb{E}\mathbb{E}[\delta_m(X) - \theta]^2 \\ &= \mathbb{E}\mathbb{E}[(\theta - \delta_m(X))^2 | X] \\ &= \mathbb{E}(V(\theta | X)) \\ &= \mathbb{E}\left(\frac{1}{n + 1/m}\right) \\ &= \frac{1}{n + 1/m} \to \frac{1}{n} = \mathbb{E}(\bar{X} - \theta)^2 \end{aligned}$$

So, $\bar{X}$ is a minimax estimator of θ.

*Remark: The Bayes risk r_m are often troublesome to calculate. The minimax δ_0 is usually obtained as pointwise limit of the δ_m 's and we might hope by continuity considerations that $\delta_m \to \delta_0$ implies $r_m \to r$ making calculations for r_m unnecessary. The following example shows that although $\delta_m \to \delta_0$ (pointwise), $r_m \nrightarrow r$.

Example: Suppose $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\theta, 1)$. Consider the sequence of priors $\xi_m$which are $\mathcal{N}(bm, m)$. Assuming squared error loss, the Bayes estimator of θ under ξ_m is $\delta_m(X) = \frac{n\bar{X} + bm/m}{n+1/m} = \frac{n\bar{X}+b}{n+1/m} \to \bar{X} + b/n$ as $m \to \infty$.

But

$$
\begin{aligned}
r(\xi_m, \delta_m) &= \mathbb{E}\mathbb{E}\left(\theta - \frac{n\bar{X}+b}{n+1/m}\right)^2 \\
&= \mathbb{E}\left[\mathbb{E}\left(\left(\theta - \frac{n\bar{X}+b}{n+1/m}\right)^2 \middle| X\right)\right] \\
&= \mathbb{E}\left(\frac{1}{n+1/m}\right) \\
&= \frac{1}{n+1/m} \to \frac{1}{n} \text{ as } m \to \infty.
\end{aligned}
$$

However, $\mathbb{E}(\bar{X}+b/n-\theta)^2 = V(\bar{X}) + \Big[\mathbb{E}(\bar{X}+b/n) - \theta\Big]^2 = 1/n + b^2/n^2 > 1/n.$

Thus $\bar{X}+b/n$ is not a minimax estimator of θ.

*Remark: In all the above examples, the loss was squared error loss or weighted squared error and thus convex. Hence, one could restrict attention to non-randomized decision rules in finding Minimax rules. However, this need not be the case, if the loss is not convex.

Example: (A randomized estimate better than a non-randomized estimate using the minimaxity criterion)

Let $X \sim$ Bernoulli(θ), $\theta \in \Theta = [0,1]$, $\mathcal{A} = [0,1]$. Suppose

$$
L(\theta, \delta(X)) = \begin{cases} 1, & \text{if } |\delta(X) - \theta| \geq 1/4, \\ 0, & \text{if } |\delta(X) - \theta| < 1/4. \end{cases}
$$

Note that L is not convex in θ.

$$
R(\theta, \delta(x)) = P(|\delta(x) - \theta| \geq 1/4)
$$

We first show that for any non-randomized estimator $\delta(X)$ of θ, $\sup_{\theta \in \Theta} R(\theta, \delta) = 1$.

Case I:

$$
\frac{1}{4} \leq \delta(0) \leq 1 \quad \frac{1}{4} \leq \delta(1) \leq 1 \qquad \text{take } \theta = 0
$$

Case II:

$$
\left\{ \begin{matrix} 0 \leq \delta(0) \leq \frac{3}{4} & 0 \leq \delta(1) \leq \frac{1}{4} \\ 0 \leq \delta(0) \leq \frac{1}{4} & 0 \leq \delta(1) \leq \frac{3}{4} \end{matrix} \right\} \qquad \text{take } \theta = 1
$$

Case III:

$$
\left\{ \begin{matrix} \frac{3}{4} < \delta(0) \leq 1 & 0 \leq \delta(1) \leq \frac{1}{4} \\ \frac{3}{4} < \delta(1) \leq 1 & 0 \leq \delta(0) \leq \frac{1}{4} \end{matrix} \right\} \qquad \text{take } \theta = \frac{1}{2}
$$

A particular randomized estimator δ^* would be: ignore X and use the estimator Y for θ under Y is Uniform[0, 1]. Then

$$R(\theta, \delta^*) = P(|Y-\theta| \geq \frac{1}{4}) = P(Y \geq \theta + \frac{1}{4}) + P(Y \leq \theta - \frac{1}{4})$$
$$= \begin{cases} 3/4 - \theta, & \text{if } 0 \leq \theta \leq 1/4 \\ 1/2, & \text{if } 1/4 \leq \theta \leq 3/4 \\ \theta - 1/4, & \text{if } 3/4 \leq \theta \leq 1 \end{cases}$$

So $\sup_{\theta \in \Theta} R(\theta, \delta^*) = 3/4$.

6. Admissibility

Def: A decision rule δ_1 is said to be *at least as good as* δ_2 if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$; δ_1 is said to be *better* than δ_2 if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$ with strict inequality for some $\theta \in \Theta$; δ_1 is said to be *risk equivalent* to δ_2 if $R(\theta, \delta_1) = R(\theta, \delta_2)$ for all $\theta \in \Theta$.

Def: A decision rule δ_0 is said to be *admissible* if there does not exist any decision rule δ such that $R(\theta, \delta) \leq R(\theta, \delta_0)$ for all $\theta \in \Theta$ with strict inequality for some $\theta \in \Theta$, i.e., there does not exist any rule better than δ_0.

*Better than -- Dominate; Admissible -- No rule dominating it

Def: A class $\mathcal{C}(\subset \mathcal{D}^*)$ of decision rules is said to be *complete* if given any $\delta \in \mathcal{D}^*$ such that $\delta \notin \mathcal{C}$, there exists a rule $\delta_0 \in \mathcal{C}$ which is better than δ. A class $\mathcal{C}(\subset \mathcal{D}^*)$ of decision rules is said to be *essentially complete* if given any rule $\delta \in \mathcal{D}^*$ such that $\delta \notin \mathcal{C}$, there exists a rule $\delta_0 \in \mathcal{C}$ which is at least as good as δ

The difference between complete and essentially complete class of decision rules may be illuminated by the following two lemmas.

Lemma 1: If $\mathcal{C}$ is a complete class, and $\mathcal{A}$ denotes the class of admissible rules, then $\mathcal{A} \subset \mathcal{C}$.

proof: Suppose $\mathcal{A} \not\subset \mathcal{C}$. Then there exists $\delta_0 \in \mathcal{A}$, but $\delta_0 \notin \mathcal{C}$. Here, there is a $\delta \in \mathcal{C}$ which is better than δ_0 contradicting the admissibility of δ_0.

Lemma 2: If $\mathcal{C}$ is an essentially complete class, and there exists an admissible $\delta \notin \mathcal{C}$, then there exists a $\delta' \in \mathcal{C}$ which is risk equivalent to δ.

proof: Since $\mathcal{C}$ is essentially complete, and $\delta \notin \mathcal{C}$, there is a $\delta' \in \mathcal{C}$ such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all $\theta \in \Theta$. Since δ is admissible, strict inequality cannot occur for any $\theta \in \Theta$. Hence, $R(\theta, \delta') = R(\theta, \delta)$ for all $\theta \in \Theta$.

Def: A class $\mathcal{C}$ of decision rules is said to be *minimal complete* if $\mathcal{C}$ is complete, and no proper subclass of $\mathcal{C}$ is complete. A class $\mathcal{C}$ of decision rules is said to be *minimal essentially complete* if $\mathcal{C}$ is essentially complete and no proper subclass of $\mathcal{C}$ is essentially complete.

*Remark: It is not necessary that minimal complete or minimal essentially complete classes exist. If the statistician can find an essentially complete class, there is no need for him to look outside this class to find a decision rule, for he can do just as well inside the class. Thus, if the statistician can find a small (essentially) complete class from which to make his choice, his task is greatly simplified. The smallest class may not exist, but if it does, it is called a minimal (essentially) complete class.

The following clarifies the relationship between admissible rules and minimal complete classes.

Theorem: If a minimal complete class exists, it consists exactly of the (all) admissible rules

proof: Let $\mathcal{C}$ denote the minimal complete class, and $\mathcal{A}$ denote the class of all admissible rules. We need to show $\mathcal{C} = \mathcal{A}$. We have already proved that $\mathcal{A} \subset \mathcal{C}$ since a minimal complete class is complete. So we need to show $\mathcal{C} \subset \mathcal{A}$.

Suppose there exists a $\delta_0 \in \mathcal{C}$, but $\delta_0 \notin \mathcal{A}$. Thus

Step 1: There exists a $\delta_1 \in \mathcal{C}$ which is better than δ_0.

Step 2: Show $\mathcal{C}_1 = \mathcal{C} - \{\delta_0\}$ is complete (which contradicts the minimal completeness of $\mathcal{C}$)

- proof of Step 1: Since δ_0 is inadmissible, there exists a δ better than δ_0. If $\delta \in \mathcal{C}$, take $\delta_1 = \delta$. If not, there exists $\delta_1 \in \mathcal{C}$ better than δ, and hence better than δ_0.
- proof of Step 2: Note that $\mathcal{C}_1 \cup \{\delta_1\}$ is complete because of Step 1, any rule improved by δ_0 is also improved by δ_1. But $\delta_1 \in \mathcal{C} - \{\delta_0\}$. Hence $\mathcal{C}_1 \cup \{\delta_1\} = \mathcal{C}_1$ is complete.

*Remark: It is not necessarily true that the class of all admissible decision rules is essentially complete. Consider for example the situation when $\mathcal{D} = \{\delta_0, \delta_1, \delta_2, \cdots\}$. Let $\Theta = \{\theta_1, \theta_2\}$ and suppose $R(\theta_1, \delta_0) = 0$, $R(\theta_2, \delta_0) = 1$ and $R(\theta_1, \delta_i) = R(\theta_2, \delta_i) = 2^{-i}$, $i = 1, 2, \cdots$. Then the only admissible rule is δ_0, but $\mathcal{A} = \{\delta_0\}$ is not essentially complete.

Theorem: If a class of admissible rules is complete, then it is minimal complete.

proof: Suppose $\mathcal{A}$ is a class of admissible rules which is complete. If $\mathcal{A}$ is not minimal complete, then $\exists$ a proper subset $\mathcal{A}_1$ of $\mathcal{A}$ which is complete. Suppose $\delta \in \mathcal{A} - \mathcal{A}_1$. Then since $\mathcal{A}_1$ is complete, there exists a $\delta_1 \in \mathcal{A}_1$ which is better than δ which contradicts the admissibility of δ.

The following theorem gives further insight into the relationship between admissibility and completeness.

Theorem: Consider the triplet $(\Theta, \mathcal{D}^*, \mathbb{R})$. Let $\mathcal{A}$ be the set of all admissible rules in $\mathcal{D}^*$, and let $\mathcal{C} = \cap_{\gamma \in \Gamma} \mathcal{C}_\gamma$ where $\{\mathcal{C}_\gamma : \gamma \in \Gamma\}$ is the collection of all complete subsets of $\mathcal{D}^*$. Then $\mathcal{A} = \mathcal{C}$.

Moreover, $\mathcal{A} = \mathcal{C}$ is complete if and only if there exists a minimal complete subset $\mathcal{C}_m$ of $\mathcal{D}^*$ in which case $\mathcal{A} = \mathcal{C} = \mathcal{C}_m$.

proof: In view of Lemma 1, $\mathcal{A} \subset \mathcal{C}_\gamma$, for all $\gamma \in \Gamma$. Hence $\mathcal{A} \subset \mathcal{C}$. We need to prove that $\mathcal{C} \subset \mathcal{A}$. Suppose not. Then there exist $\delta_0 \in \mathcal{D}^*$ that $\delta_0 \in \mathcal{C} - \mathcal{A}$. Since δ_0 is inadimissible, there exists $\delta_1 \in \mathcal{D}^*$ which is better than δ_0. Then for any fixed $\gamma \in \Gamma$, $\mathcal{C}'_\gamma = (\mathcal{C}_\gamma - \{\delta_0\}) \cup \{\delta_1\}$ is complete, because any rule improvable by δ_0 is improvable by δ_1. Hence, there exists a $\gamma_1 \in \Gamma$, such that $\mathcal{C}_{\gamma_1} = \mathcal{C}'_\gamma$. But $\delta_0 \notin \mathcal{C}_{\gamma_1}$. Hence $\delta_0 \notin \mathcal{C}$ where $\mathcal{C} = \cap_{\gamma \in \Gamma} \mathcal{C}_\gamma$, a contradiction to $\delta_0 \in \mathcal{C} - \mathcal{A}$. So $\mathcal{C} \subset \mathcal{A}$. Thus $\mathcal{C} = \mathcal{A}$.

Now assume $\mathcal{A} = \mathcal{C}$ is complete. If $\mathcal{A}_0$ is a complete subset of $\mathcal{A}$, then $\mathcal{A}_0 \supset \mathcal{C} = \mathcal{A}$. Hence $\mathcal{A} = \mathcal{A}_0$. Hence $\mathcal{A}$ is the minimal complete subset. Conversely, if a minimal complete subset $\mathcal{C}_m$ exists, by Theorem (the one before the last Theorem), $\mathcal{C}_m = \mathcal{A}$. So $\mathcal{C} = \mathcal{A} = \mathcal{C}_m$.

Methods of Finding Admissible Rules

Theorem: If for any prior distribution ξ, a Bayes rule δ_ξ is unique up to risk equivalence (i.e. for any δ, $r(\xi,\delta) = r(\xi,\delta_\xi) \Rightarrow R(\theta,\delta) = R(\theta,\delta_\xi)$ for all $\theta \in \Theta$), then it is admissible, provided that the Bayes risk is finite.

prrof: Suppose not. Then there exists some δ' such that $R(\theta,\delta) \leq R(\theta,\delta_\xi)$ for all $\theta \in \Theta$, having strict inequality for some $\theta \in \Theta$. Thus,

$$r(\xi,\delta) = \int_\Theta R(\theta,\delta)\, d\xi(\theta) \leq \int_\Theta R(\theta,\delta_\xi)\, d\xi(\theta) = r(\xi,\delta_\xi) \leq r(\xi,\delta).$$

Then $r(\xi,\delta) = r(\xi,\delta_\xi)$. Thus $R(\theta,\delta) = R(\theta,\delta_\xi)$ for all $\theta \in \Theta$, which is a contradiction.

Theorem: Assume that $\Theta = \{\theta_1,\cdots,\theta_k\}$, and δ_ξ is a Bayes rule w.r.t. a prior $\xi = \{\xi_1,\cdots,\xi_k\}$ exists, where ξ_i is the prior probability assigned to θ_i by ξ, $1 \leq i \leq k$. If $\xi_i > 0$ for all $1 \leq i \leq k$, then δ_ξ is admissible.

proof: Suppose not. Then there exists a δ such that $R(\theta,\delta) \leq R(\theta,\delta_\xi)$ for all $\theta \in \Theta$ with strict inequality for some $\theta \in \Theta$. Then

$$r(\xi,\delta) = \sum_{i=1}^k R(\theta_i,\delta)\xi_i < \sum_{i=1}^k R(\theta_i,\delta_\xi)\xi_i = r(\xi,\delta_\xi),$$

which contradicts the Bayesness of δ_ξ w.r.t. ξ.

*Remark: The above theorem ceases to hold if the condition $\xi_i > 0$ for all $1 \leq i \leq k$ is dropped. To see this, consider the example $\Theta = \{\theta_1,\theta_2\}$ and suppose that $1 \leq R(\theta_1,\delta) \leq 2$, $0 \leq R(\theta_2,\delta) \leq 1$ for all decision rules δ under consideration. Consider the prior ξ: $(\xi_1,\xi_2) = (1,0)$. Then $\sum_{i=1}^2 \xi_i R(\theta_i,\delta) = R(\theta_1,\delta) \geq 1$. Now any decision rule δ_0 satisfying $R(\theta_1,\delta_0) = 1$ is Bayes w.r.t. ξ. Suppose however that $R(\theta_2,\delta_0) = 1/2$. Consider now a decision rule δ with $R(\theta_1,\delta) = 1$, $R(\theta_2,\delta) < 1/2$. Then δ dominates (is better than) δ_0.

(Note that δ_0 is inadmissible, but is Bayes w.r.t. ξ)

The extension of the above theorem when Θ is NOT finite requires a new concept.We consider the case $\Theta = \mathbb{R}$. A point $\theta_0 \in \mathbb{R}$ is said to be in the *support* of a distribution function F on the real line if for every $\varepsilon > 0$, $F(\theta_0+\varepsilon) - F(\theta_0-\varepsilon) > 0$.

Examples:

- $X \sim$ Binomial(n,θ), $\theta \in [0,1]$. Support of X is $\{0,1,\cdots,n\}$
- $X \sim$ Uniform$[0,1]$. Supposrt of X is $[0,1]$
- (A discrete distribution with an interval as support) Let $\{r_1,r_2,\cdots\}$ denote the set of rationals in $[0,1]$ with $P(X = r_n) = 2^{-n}$, $n = 1,2,\cdots$. Then the support of X is the unit interval $[0,1]$

We shall also need the concept of generalized Bayes estimators.

Def: A rule δ_0 is said to be *Generalized Bayes* w.r.t. a prior (proper or improper) ξ if for every $x \in \mathcal{X}$, $\int_\Theta L(\theta,\delta(x))p(\theta|x)\, d\theta$ takes on a finite minimum value when $\delta = \delta_0$.

(A prior ξ on Θ is said to be improper if $\xi(\theta) \geq 0$ for all $\theta \in \Theta$ and $\int_\Theta d\xi(\theta) = \infty$)

Example 1: Suppose $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\theta, 1)$, where $\theta \in \Theta = \mathbb{R}$. Suppose the prior is uniform $(-\infty, \infty)$. Then

$$p(\theta|x) \propto e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2} \cdot 1 \propto e^{-\frac{n}{2}(\theta-\bar{x})^2}$$

Hence posterior distribution of θ is $\mathcal{N}(\bar{X}, 1/n)$. Hence assuming squared error loss, $\int_{-\infty}^{\infty}(\theta-a)^2 p(\theta|x)\, d\theta$ is minimized at $\delta(x) = \bar{x}$. Hence $\bar{X}$ is the generalized Bayes estimator of θ under the uniform (improper) prior over the real line.

Example 2: Suppose $X_1, \cdots, X_n$ are i.i.d. Bernoulli(θ), where $\theta \in \Theta = (0, 1)$. Consider the improper prior ξ on $(0, 1)$ with pdf $g(\theta) = \frac{1}{\theta(1-\theta)}$, $0 < \theta < 1$. Then $p(\theta|x) \propto \theta^{\sum_{i=1}^n x_i - 1}(1-\theta)^{n-\sum_{i=1}^n x_i - 1}$.

We have already noted in a previous example that $\int_0^1 (\theta-a)^2 p(\theta|x)\, d\theta$ is minimized at $a = \bar{X}$. Hence $\bar{X}$ is the generalized Bayes estiator of θ under the improper prior $g(\theta)$.

Example 3: Suppose $X_1, \cdots, X_n$ are i.i.d. Poisson(θ), where $\theta \in (0, \infty)$. Consider the improper prior ξ with pdf $g(\theta) = \frac{1}{\theta}$, $\theta \in (0, \infty)$. Then $p(\theta|x) \propto e^{-n\theta}\theta^{\sum_{i=1}^n x_i - 1}$.

Note that $\int_0^\infty (\theta-a)^2 p(\theta|x)$ is minimized at $a = \bar{X}$. Hence $\bar{X}$ is the generalized Bayes estimator of θ under the improper prior $g(\theta)$.

Theorem: Suppose Θ is an open subset of $\mathbb{R}$ and assume that $R(\theta, \delta)$ is continuous in θ, for every fixed δ. If δ_ξ is generalized Bayes w.r.t. ξ on Θ for which $|r(\xi, \delta_\xi)| < \infty$, and the support of ξ is Θ, then δ_ξ is admissible.

proof: Suppose not. Then there exists a δ such that $R(\theta, \delta) \leq R(\theta, \delta_\xi)$ for all $\theta \in \Theta$ with strict inequality for some $\theta_0 \in \Theta$.

Since $R(\theta, \delta)$ is continuous in θ for each fixed δ, given any $\eta > 0$, there exists an $\varepsilon > 0$ such that

$$|R(\theta, \delta) - R(\theta_0, \delta)| < \frac{1}{4}\eta$$
$$|R(\theta, \delta_\xi) - R(\theta_0, \delta_\xi)| < \frac{1}{4}\eta$$

whenever $|\theta - \theta_0| < \varepsilon$. Hence, for $|\theta - \theta_0| < \varepsilon$,

$$\begin{aligned} R(\theta, \delta_\xi) - R(\theta, \delta) &> R(\theta_0, \delta_\xi) - \frac{1}{4}\eta - (R(\theta_0, \delta) + \frac{1}{4}\eta) \\ &= R(\theta_0, \delta_\xi) - R(\theta_0, \delta) - \frac{\eta}{2}. \end{aligned}$$

Take $\eta = R(\theta_0, \delta_\xi) - R(\theta_0, \delta) > 0$, then

$$R(\theta, \delta_\xi) - R(\theta, \delta) > \eta - \eta/2 = \eta/2 \text{ whenever } |\theta - \theta_0| < \varepsilon.$$

Hence

$$\begin{aligned} r(\xi, \delta_\xi) - r(\xi, \delta) &= \int_\Theta [R(\theta, \delta_\xi) - R(\theta, \delta)]\, d\xi(\theta) \\ &\geq \int_{\theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]} [R(\theta, \delta_\xi) - R(\theta, \delta)]\, d\xi(\theta) \\ &> \int_{\theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]} \frac{\eta}{2}\, d\xi(\theta) \\ &= \frac{\eta}{2}[\xi(\theta_0 + \varepsilon) - \xi(\theta_0 - \varepsilon)] > 0 \end{aligned}$$

since $\theta_0 \in \Theta$, the support of ξ. This is a contradiction of the Bayesness of δ_ξ. Thus δ_ξ is admissible.

Example: Suppose $X_1, \cdots, X_n$ are i.i.d. Bernoulli(θ), where $\theta \in (0, 1)$. If the prior ξ has pdf $g(\theta) = \frac{1}{\theta(1-\theta)}$, $0 < \theta < 1$, $\bar{X}$ is the generalized Bayes estimator of θ w.r.t. squared error loss. Note that $R(\theta, \bar{X}) = \frac{\theta(1-\theta)}{n}$, $\forall 0 < \theta < 1$.

$$r(\xi, \delta_\xi) = r(\xi, \bar{X}) = \int_0^1 R(\theta, \bar{X}) g(\theta)\, d\theta = \int_0^1 \frac{\theta(1-\theta)}{n} \frac{1}{\theta(1-\theta)}\, d\theta = \frac{1}{n} < \infty.$$

Hence $\bar{X}$ is an admissible estimator of θ under squared error loss.

*Remark: Note that this theorem cannot be used to prove admissibility of $\bar{X}$ for estimating θ, the normal mean. Note that $\bar{X}$ is the generalized Bayes estimator of θ under the uniform $(-\infty, \infty)$ prior (say ξ) and squared error loss. However, $R(\theta, \bar{X}) = 1/n$ for all $\theta \in \mathbb{R}$, then $r(\xi, \bar{X}) = \int_{-\infty}^{\infty} R(\theta, \bar{X}) \cdot 1\, d\theta = \int_{-\infty}^{\infty} \frac{1}{n}\, d\theta = \infty$. The same problem arises in the Poisson case because $\int_0^\infty R(\theta, \bar{X})\, d\theta = \int_0^\infty \frac{\theta}{n} \frac{1}{\theta}\, d\theta = \infty$.

A Unified Admissibility Proof (Brown & Hwang, 1983; 3rd Purdue Symposium)

Bartlett's identity:

$$\mathbb{E}\left(\frac{\partial}{\partial \theta} \log f_\theta(X)\right) = 0$$

proof: $\int f_\theta(x)\, dx = 1$

$$\begin{aligned} \Rightarrow 0 = \frac{\partial}{\partial \theta} \int f_\theta(x)\, dx &= \int \frac{\partial}{\partial \theta} f_\theta(x)\, dx = \int \frac{\frac{\partial}{\partial \theta} f_\theta(x)}{f_\theta(x)} f_\theta(x)\, dx \\ &= \int \frac{\partial}{\partial \theta} \log f_\theta(x) \cdot f_\theta(x)\, dx = \mathbb{E}\left(\frac{\partial}{\partial \theta} \log f_\theta(X)\right) \end{aligned}$$

Second Bartlett's identity:

$$Var\left[\frac{\partial}{\partial \theta} \log f_\theta(X)\right] + \mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X)\right] = 0$$

proof: $\int f_\theta(x)\, dx = 1 \Rightarrow \int \frac{\partial^2 f_\theta(x)}{\partial \theta^2}\, dx = 0 \Rightarrow \mathbb{E}\left(\frac{f''_\theta(X)}{f_\theta(X)}\right) = 0$. Because $\frac{\partial^2}{\partial \theta^2} \log f_\theta = \frac{f''_\theta}{f_\theta} - \frac{(f'_\theta)^2}{f_\theta^2}$, we have

$$\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X)\right] + \mathbb{E}\left[\left(\frac{f'_\theta(X)}{f_\theta(X)}\right)^2\right] = 0,$$

which is equivalent to the second Bartlett's identity.

Brown and Hwang provided sufficient conditions to prove admissibility of generalized Bayes estimators for mean vectors in the multiparameter regular exponential family of distributions. (Here "regular" means the support of the distribution does not depend on the parameters)

Suppose that X has pdf

$$f_\theta(x) = \exp(\theta^T x - \psi(\theta) + h(x))$$

here x and θ are both vectors. Assume that $\theta \in \Theta := \{\theta : \int \exp(\theta^T x + h(x))\, dx < \infty\}$, and Θ is an open subset in $\mathbb{R}^p$.

Then

$$\mathbb{E}_\theta(X) = \nabla\psi(\theta) = (\partial\psi(\theta)/\partial\theta_1, \cdots, \partial\psi(\theta)/\partial\theta_p)^T$$

proof: Use the Bartlett's identity, $\mathbb{E}\left(\frac{\partial \log f_\theta(x)}{\partial\theta}\right) = 0$. That is, $\mathbb{E}_\theta(X - \nabla\psi(\theta)) = 0$.

Consider now the loss $L(\nabla\psi(\theta), a) = \|a - \nabla\psi(\theta)\|^2 = \|a - \mathbb{E}_\theta(X)\|^2$. Then the risk of a non-randomized estimator $\delta(X)$ of $\nabla\psi(\theta)$ is given by $\mathbb{E}_\theta\|\delta(X) - \nabla\psi(\theta)\|^2$. We now find expression for a generalized Bayes estimator of $\nabla\psi(\theta)$ under an arbitraty prior pdf $g(\theta)$.

We will use the generic notation for any function $u(\theta)$. Let $I_x(u) = \int u(\theta) f_\theta(x)\, d\theta$.

Lemma: Suppose $I_x(\nabla g) < \infty$ for all $x \in \mathcal{X}$, under a prior (proper or improper) g, the generalized Bayes estimator $\delta_g(X)$ of $\nabla\psi(\theta)$ is given by $\delta_g(X) = X + \frac{I_X(\nabla g)}{I_X(g)} = \frac{I_X(g\nabla\psi)}{I_X(g)}$.

proof: (One parameter case) Suppose $\theta \in (\underline{\theta}, \overline{\theta})$.

$$\delta_g(x) = \frac{\int_{\underline{\theta}}^{\overline{\theta}} \nabla\psi(\theta) \exp[\theta x - \psi(\theta)] g(\theta)\, d\theta}{\int_{\underline{\theta}}^{\overline{\theta}} \exp[\theta x - \psi(\theta)] g(\theta)\, d\theta} = \frac{I_x(g\nabla\psi(\theta))}{I_x(g)} \quad (\star)$$

Now integrating the numerator by parts

$$\int_{\underline{\theta}}^{\overline{\theta}} \nabla\psi(\theta) \exp[\theta x - \psi(\theta)] g(\theta)\, d\theta = -\exp[-\psi(\theta) + \theta x]\Big|_{\underline{\theta}}^{\overline{\theta}} + \int_{\underline{\theta}}^{\overline{\theta}} \frac{d}{d\theta}\{\exp(\theta x) g(\theta)\} \exp(-\psi(\theta))\, d\theta$$

The first term is zero (for the reason see the lemma below this proof).

The second term simplifies to

$$\int_{\underline{\theta}}^{\overline{\theta}} [x \exp(\theta x) g(\theta) + \exp(\theta x) \nabla g(\theta)] \exp(-\psi(\theta))\, d\theta$$

$$= x \int_{\underline{\theta}}^{\overline{\theta}} \exp(\theta x - \psi(\theta)) g(\theta)\, d\theta + \int_{\underline{\theta}}^{\overline{\theta}} \nabla g(\theta) \exp(\theta x - \psi(\theta))\, d\theta \quad (\star\star)$$

Combining $(\star)$ and $(\star\star)$, we have $\delta_g(x) = x + \frac{I_x(\nabla g)}{I_x(g)}$. In the multiparameter case, work separately with each partial derivative.

Lemma (Woodroofe 1981): Let $h(\theta) = \exp(\theta^T x - \psi(\theta))$, $\theta \in (\underline{\theta}, \overline{\theta})$, $h(\theta) \to 0$ as $\theta \to \underline{\theta}$ or $\theta \to \overline{\theta}$ (Diaconis & Ylvisaker, 1979, Conjugate priors for exponential families, Theorem 2). If g is a continuously differentiable function with $\int_{\underline{\theta}}^{\overline{\theta}} |g'| d\theta < \infty$. Then $h(\theta)g(\theta) \to 0$ as $\theta \to \underline{\theta}$ or $\theta \to \overline{\theta}$.

proof: Note that $\frac{\partial \log h}{\partial \theta} = x - \psi'(\theta)$, $\frac{\partial^2 \log h}{\partial \theta^2} = -\psi''(\theta) < 0$. So $\log h(\theta)$ is a strictly concave function. Thus there exists a $\theta^* \in (\underline{\theta}, \overline{\theta})$ such that $\log h(\theta)$ is decreasing in $(\theta^*, \overline{\theta})$, and there exists a $\theta^0 \in (\underline{\theta}, \overline{\theta})$ such that $\log h(\theta)$ is increasing in $(\underline{\theta}, \theta^0)$.

If $\theta^* < \theta_1 < \theta_2 < \overline{\theta}$, $h(\theta)$ is decreasing.

$$\begin{aligned}
|g(\theta_2)|h(\theta_2) &= |g(\theta_1) + g(\theta_2) - g(\theta_1)|h(\theta_2) \\
&\leq |g(\theta_1)|h(\theta_2) + \left| \int_{\theta_1}^{\theta_2} g'(\theta)\, d\theta \right| h(\theta_2) \\
&\leq |g(\theta_1)|h(\theta_2) + h(\theta_2) \int_{\theta_1}^{\theta_2} |g'(\theta)|\, d\theta \\
&\leq |g(\theta_1)|h(\theta_2) + \int_{\theta_1}^{\theta_2} |g'(\theta)|h(\theta)\, d\theta
\end{aligned}$$

Let $\theta_2 \to \overline{\theta}$,

$$\limsup_{\theta \to \overline{\theta}} |g(\theta)|h(\theta) \leq \int_{\theta_1}^{\overline{\theta}} |g'(\theta)|h(\theta)\, d\theta$$

Now let $\theta_1 \to \overline{\theta}$, then $|g(\theta)|h(\theta) \to 0$. This behavior of $g(\theta)h(\theta)$ at the lower end-point may be analyzed similarly.

A few notations

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

For a sequence $\{h_n\}$, $h_n : \Theta \to [0, 1]$ of absolutely continuous functions, let $h_n(\theta) = 1$ if $\theta \in S$ for some set S with $\int_S g(\theta)\, d\theta > 0$ and $h_n(\theta) = 0$ when $\|\theta\| > n$. Consider the sequence of priors $\{g_n(\theta)\}$ such that $g_n(\theta) = g(\theta)h^2(\theta)$. Let δg_n denote generalized Bayes estimator of $\nabla\psi(\theta)$ under the prior g_n. Define $\Delta_n = r(g_n, \delta_g) - r(g_n, \delta_{g_n})$.

Blyth's Method: δ_g is admissible if there exists $\{h_n\}$ such that $\Delta_n \to 0$.

proof: Suppose δ_g is not admissible, and let δ' be a decision rule such that $R(\theta, \delta') \leq R(\theta, \delta_g)$ and $\delta' \neq \delta_g$ for almost all x. Let $\delta'' = \frac{\delta' + \delta_g}{2}$. Then by Jensen's inequality and the strict convexity of squared error loss

$$R(\theta, \delta'') < \frac{1}{2}\left[R(\theta, \delta') + R(\theta, \delta_g)\right] \leq R(\theta, \delta_g).$$

Now,

$$\begin{aligned}\Delta_n &= \sup_{\delta}\{r(g_n,\delta_g) - r(g_n,\delta)\} \\ &\geq r(g_n,\delta_g) - r(g_n,\delta'') \\ &= \int [R(\theta,\delta_g) - R(\theta,\delta'')]g_n(\theta)\,d\theta \\ &\geq \int_S [R(\theta,\delta_g) - R(\theta,\delta'')]g_n(\theta)\,d\theta \\ &= \int_S [R(\theta,\delta_g) - R(\theta,\delta'')]g(\theta)\,d\theta > 0\end{aligned}$$

Hence $\Delta_n \nrightarrow 0$ as $n \to \infty$, a contradiction.

We are now in a position to state the main theorem of Brown and Hwang. We need the following regularity conditions.

1. (The growth condition)

$$\int_{\mathbb{R}^p - S} \frac{g(\theta)}{\|\theta\|^2\{\log(\max(\|\theta\|,2))\}^2}\,d\theta < \infty, \qquad \text{where } S = \{\theta : \|\theta\| \leq 1\}.$$

2. (Asymptotic flatness condition)

$$\int I_x\left[g\|\frac{\nabla g}{g} - \frac{I_x(\nabla g)}{I_x(g)}\|^2\right]dx < \infty$$

3. $$\sup_K\{R(\theta,\delta_g) : \theta \in K\} < \infty \quad \text{for all compact sets } K \subset \Theta$$

*Remark: A sufficient condition for the second condition to hold is that

$$\int_{\mathbb{R}^p} \frac{\|\nabla g(\theta)\|^2}{g(\theta)}\,dx < \infty$$

proof:

$$\begin{aligned}I_x\left[g\left\|\frac{\nabla g}{g} - \frac{I_x(\nabla g)}{I_x(g)}\right\|^2\right] &= I_x\left[\frac{\|\nabla g\|^2}{g}\right] + \frac{\|I_x(\nabla g)\|^2}{I_x(g)} - 2I_x\left[g\langle\frac{\nabla g}{g}, \frac{I_x(\nabla g)}{I_x(g)}\rangle\right] \\ &= I_x\left[\frac{\|\nabla g\|^2}{g}\right] + \frac{\|I_x(\nabla g)\|^2}{I_x(g)} - 2\frac{I_x(\|\nabla g\|^2)}{I_x(g)} \leq I_x\left[\frac{\|\nabla g\|^2}{g}\right]\end{aligned}$$

Theorem: Assume three conditions above. Then δ_g is admissible under squared error loss.

Corollary 1: Suppose $\Theta = \mathbb{R}^p$ with $p = 1$ and 2. Then the estimator X is an admissible estimator of $\nabla\psi(\theta)$.

proof: Let $g(\theta) = 1, \forall\theta$. Then $\nabla g = 0$ and $\delta_g(X) = X$. Hence $\int \frac{\|\nabla g(\theta\|^2)}{g(\theta)}\,d\theta = 0$, which implies condition 2.

Note also that

$$\begin{aligned}\int_{\mathbb{R}^p - S} \frac{g(\theta)}{\|\theta\|^2[\log(\max(\|\theta\|,2))]^2}\,d\theta &= \int_{\|\theta\|\geq 1} \frac{1}{\|\theta\|^2[\log(\max(\|\theta\|,2))]^2} \\ &\leq \int_{1\leq\|\theta\|\leq 2} \frac{1}{(\log 2)^2}\,d\theta + \int_{\|\theta\|\geq 2} \frac{1}{\|\theta\|^2(\log\|\theta\|)^2}\,d\theta.\end{aligned}$$

Use the polar transformation

$$\theta_1 = r\cos\phi_1, \theta_2 = r\sin\phi_1\cos\phi_2, \cdots, \theta_{p-1} = r\sin\phi_1\cdots\sin\phi_{p-2}\cos\phi_{p-1}, \theta_p = r\sin\phi_1\cdots\sin\phi_{p-2}\sin\phi_{p-1}$$

Then $\|\theta\|^2 = r^2$. The Jacobian of transformation is $r^{p-1}\sin^{p-2}\phi_1\cdots\sin\phi_{p-1}$. Hence the right hand side is bounded above by

$$\frac{1}{(\log 2)^2}\int_{1\le r\le 2} r^{p-1}\,dr + \int_2^\infty \frac{r^{p-1}}{r^2(\log r)^2} < \infty$$

for $p = 1$ or 2.

*Remark: Corollary 1 includes the results of Blyth (1951), Karlin (1958) for $p = 1$ and Stein (1956) for $p = 2$. Both Blyth and Stein considered the normal example. Karline (1958) considered the result for the one-parameter exponential family. Karlin (1958) actually proved the result for this estimator $\frac{X}{\lambda+1}$ of $\nabla\psi(\theta)$ with $\lambda > -1$. The latter is generalized Bayes with respect to the prior $g(\theta) = \exp[-\lambda\psi(\theta)]$.

Adimissibility of $\bar{X}$ for the one-parameter exponential family for estimating $\mathbb{E}_\theta(X)$ under squared error loss

Let $X_1, \cdots, X_n$ be i.i.d. with common pdf $f_\theta(x) = \exp(\theta x - \psi(\theta) + h(x))$. Note that $\frac{\partial \log f}{\partial\theta} = x - \psi'(\theta)$, $\frac{\partial^2 \log f}{\partial\theta^2} = -\psi''(\theta)$. So $I(\theta) = \mathbb{E}_\theta\left(-\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2}\right) = \psi''(\theta)$. From the first Bartlett identity one gets $\mathbb{E}(X) = \psi'(\theta)$, while from the second Bartlett identity, $\mathbb{E}_\theta\left[\frac{\partial \log f_\theta(x)}{\partial\theta}\right]^2 = \mathbb{E}_\theta\left[-\frac{\partial^2 \log f(\theta)}{\partial\theta^2}\right] = \psi''(\theta)$, i.e., $\mathbb{E}_\theta(X - \psi'(\theta))^2 = \psi''(\theta)$, i.e., $V_\theta(X) = \psi''(\theta)$.

We will prove admissibility of $\bar{X}$ for estimating $\psi'(\theta)$ under squared error loss.

proof: Suppose that $\bar{X}$ is inadmissible. Due to the convexity of the loss and sufficiency of $\bar{X}$, there exists an estimator $\delta(\bar{X})$ of $\psi'(\theta)$ such that

$$\mathbb{E}_\theta[\delta(\bar{X}) - \psi'(\theta)]^2 \le \mathbb{E}_\theta[\bar{X} - \psi'(\theta)]^2 = \psi''(\theta)/n$$

with strict inequality for some θ.

But by the Cramer-Rao inequality,

$$\begin{aligned}\mathbb{E}_\theta[\delta(\bar{X}) - \psi'(\theta)]^2 &= V_\theta(\delta(\bar{X})) + [\mathbb{E}_\theta(\delta(\bar{X})) - \psi'(\theta)]^2 \\ &\ge \frac{((\psi''(\theta)) + b'(\theta))^2}{n\psi''(\theta)} + b^2(\theta)\end{aligned}$$

where $b(\theta) = \mathbb{E}_\theta(\delta(\bar{X})) - \psi'(\theta)$.

Then we have

$$\begin{aligned}\frac{\psi''}{n} + \frac{2b'(\theta)}{n} + \frac{(b'(\theta))^2}{n\psi''(\theta)} + b^2(\theta) \le \frac{\psi''(\theta)}{n} &\Rightarrow \frac{2b'(\theta)}{n} + \frac{(b'(\theta))^2}{n\psi''(\theta)} + b^2(\theta) \le 0 \\ &\Rightarrow \frac{2b'(\theta)}{n} + b^2(\theta) \le 0\end{aligned}$$

We shall show that $b(\theta) = 0$ for all θ, i.e., $\mathbb{E}_\theta[\delta(\bar{X})] = \mathbb{E}_\theta[\bar{X}]$, for all θ, which implies $\delta(\bar{X}) = \bar{X}$ a.s., a contradiction to the supposition.

Note that $b'(\theta) \leq 0$, i.e. $b(\theta)$ is non-increasing in θ.

First suppose $b(\theta_0) > 0$ for some $\theta = \theta_0$. Then $b(\theta) > 0$ for all $\theta \leq \theta_0$. Now since $2b'(\theta)/nb^2(\theta) \leq -1$ for all $\theta \leq \theta_0$, it follows that

$$\begin{aligned}
\int_\theta^{\theta_0} \frac{2}{n} \frac{b'(\theta)}{b^2(\theta)}\, d\theta &\leq -\int_\theta^{\theta_0} d\theta \\
\Rightarrow \frac{2}{n}\left(-\frac{1}{b(\theta)}\right)\Big|_\theta^{\theta_0} &\leq \theta - \theta_0 \\
\Rightarrow \frac{1}{b(\theta_0)} - \frac{1}{b(\theta)} &\geq \frac{n}{2}(\theta_0 - \theta) \\
\Rightarrow \frac{1}{b(\theta_0)} \geq \frac{n}{2}(\theta_0 - \theta) &\to \infty \text{ as } \theta \to -\infty
\end{aligned}$$

a contradiction.

Conversely, if $b(\theta_0) < 0$ for some $\theta = \theta_0$, then $b(\theta) < 0$ for all $\theta > \theta_0$.

Note that $\frac{2b'(\theta)}{nb^2(\theta)} \leq -1$ for all $\theta > \theta_0$. Using the same trick, we can get $\frac{1}{b(\theta_0)} \leq -\frac{n}{2}(\theta - \theta_0) \to -\infty$ as $n \to +\infty$, a contradiction.

Thus $b(\theta) = 0$ for all θ.

A complete class result

Characterizing Admissible Linear Estimators Of the Binomial Parameter

Suppose $X_1, \cdots, X_n$ are i.i.d. Bernoulli(p). Let $\bar{X} = \sum_{i=1}^n x_i/n$. We are interested in characterizing admissible linear estimators of the form $a\bar{X} + b$ of p (a and b are constants) under squared error loss.

Theorem: $a\bar{X} + b$ is an admissible estimator of p under squared error loss if and only if $0 \leq a \leq 1, b \geq 0$ and $a + b \leq 1$.

The proof is a consequence of the following results.

Lemma 1 (Does not require any distributional assumptions): LetY be a random variable with mean ρ and finite variance. Then assuming squared error loss, $aY + b$ is an inadmissible estimator of ρ whenever

$$(i)\ a > 1 \quad \text{or } (ii)\ a < 0 \quad \text{or } (iii)\ a = 1, b \neq 0.$$

proof:

(i) $\mathbb{E}(aY + b - \rho)^2 = a^2V(Y) + (a\rho + b - \rho)^2 \geq a^2V(Y) > V(Y) = \mathbb{E}(Y - \rho)^2$ So Y dominates $aY + b$.

(ii) $\mathbb{E}(aY + b - \rho)^2 > (a\rho + b - \rho)^2 \geq (a-1)^2(-\rho - \frac{b}{a-1}) > (-\frac{b}{a-1} - \rho)^2$. So $-\frac{b}{a-1}$ dominates $aY + b$.

(iii) $\mathbb{E}(Y + b - \rho)^2 = V(Y) + b^2 > V(Y) = \mathbb{E}(Y - \rho)^2$. So Y dominates $aY + b$.

Lemma 2 (Does not require any distributional assumptions): Let Y be a random variable with mean ρ and finite variance. Then assuming squared error loss, $aY + b\,(0 \leq a < 1)$ is an inadmissible estimator of ρ if

$$(i)\ \rho \geq 0 \text{ and } b < 0 \quad \text{or } (ii)\rho \leq k \text{ and } ak + b > k$$

proof:

(i) $\mathbb{E}(aY + b - \rho)^2 - \mathbb{E}(aY - b - \rho)^2 = 4b(a-1)\rho \geq 0$, with strict inequality for $\rho > 0$. So $aY - b$ dominates $aY + b$.

(ii) $\mathbb{E}(aY + b - \rho)^2 = \mathbb{E}(-aY - b + \rho)^2 = \mathbb{E}(aY' + k - ak - b - \rho')^2$ where $Y' = k - Y$ and $\rho' = k - \rho \geq 0$. Using part (i), $aY' + k - ak - b$ is inadmissible for estimating ρ' if $k - ak - b < 0$, i.e., $ak + b > k$. Hence $aY + b$ is inadmissible for estimating ρ under the same condition.

Using Lemma 1 and Lemma 2 with $Y = \bar{X}$ and $\rho = p$, a necessary condition for admissibility of $a\bar{X} + b$ will be

$$(i) 0 < a \leq 1, b \geq 0, a + b \leq 1$$
$$(ii) a = 0 \text{ and } 0 \leq b \leq 1.$$

To prove the sufficiency of (i), first note that when $a = 1$, $b = 0$, we have proved already the admissibility of $\bar{X}$. For $0 < a < 1$, $b \geq 1$, $a + b \leq 1$, note that under squared error loss and the Beta($\frac{nb}{a}$, $\frac{n(1-a-b)}{a}$) prior, $a\bar{X} + b$ is the generalized Bayes estimator of p with finite Bayes risk $\frac{ab(1-a-b)}{(1-a)(n(1-a)+a)}$. Hence $a\bar{X} + b$ is an admissible estimator of p in this case.

To prove the sufficiency of (ii), i.e., when $a = 0$ and $0 \leq b \leq 1$, we need to prove the admissibility of the constant estimator $b \in [0, 1]$ of p. This follows immediately by assigning a prior degenerate of b because then b is the Bayes estimator of p with Bayes risk $\mathbb{E}(b - p)^2 < \infty$. Hence b is an admissible estimator of p in this case.

Next we prove a few miscellaneous results regarding admissibility of $\bar{X}$ and some related functions of $\bar{X}$ in the normal case.

- Example 1: Suppose $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\theta, 1)$, but $\theta \in [a, b]$, where $-\infty < a < b < \infty$, and a and b are specified. Then $\bar{X}$ is an inadimissible estimator of θ under squared error loss.

  $\bar{X}$ is dominated by

$$\delta(\bar{X}) = \begin{cases} \bar{X}, & \text{if } a \leq \bar{X} \leq b \\ a, & \text{if } \bar{X} < a \\ b, & \text{if } \bar{X} > b \end{cases}.$$

- Example 2: Suppose $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\theta, 1)$, $\theta \in (-\infty, \infty)$. The UMVUE of θ^2 in this case is $\bar{X}^2 - 1/n$. This is not an admissible estimato of θ^2 under squared error loss.

  $\bar{X}^2 - 1/n$ is dominated by

$$\delta(\bar{X}) = \begin{cases} \bar{X}^2 - \frac{1}{n}, & \text{if } |\bar{X}| \geq \frac{1}{\sqrt{n}} \\ 0, & \text{if } |\bar{X}| < \frac{1}{\sqrt{n}} \end{cases}.$$

- Example 3: Suppose $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\theta, \sigma^2)$, where both $\theta \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. Assuming squared error loss, we want to prove admissibility of $\bar{X}$.

First observe that if $\sigma^2(>0)$ were known. $\bar{X}$ is an admissible estimator of θ under squared error loss. To prove the present result, suppose $\bar{X}$ is an inadmissible estimator of θ. Then there exists $\delta = \delta(X_1, \cdots, X_n)$ such that $R((\theta, \sigma^2), \delta) \leq R((\theta, \sigma^2), \bar{X})$ for all (θ, σ^2) with strict inequality for some (θ_0, σ_0^2). Then for the restricted problem when $\theta \in \mathbb{R}$ and $\sigma^2 = \sigma_0^2$, $R((\theta, \sigma_0^2), \delta) \leq R((\theta, \sigma_0^2), \bar{X})$ for all θ, with strict inequality at $\theta = \theta_0$. This contradicts the admissibility of $\bar{X}$ as an estimator of θunder squared error loss when $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\theta, \sigma_0^2)$ $(\theta \in \mathbb{R})$.

Chapter 2: Multiparameter Estimation