Definition A rule δ is a minimax rule if $\max_\theta R(\theta, \delta) \leq \max_\theta R(\theta, \delta')$ for any other rule δ'. It minimizes the maximum risk.

Since minimax minimizes the maximum risk (ie, the loss averaged over all possible data $X \sim f$) the choice of rule is not influenced by the actual data $X = x$ (though given the rule δ, the action $\delta(x)$ is data-dependent). It makes sense when the maximum loss scenario must be avoided, but can can lead to poor performance on average.

Definition Suppose we have a prior probability $\pi = \pi(\theta)$ for θ. Denote by

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

the Bayes risk of rule δ. A Bayes rule is a rule that minimizes the Bayes risk.

Let $\pi(\theta|x) = \dfrac{L(x;\theta)\pi(\theta)}{h(x)}$ denote the posterior following from likelihood L and prior π. Denote by $\int L_S(\theta, \delta(\mathbf{x}))\pi(\theta|x)d\theta$ the expected posterior loss.

A Bayes rule minimizes the EPL.

$$
\begin{aligned}
\int R(\theta,\delta)\pi(\theta)d\theta &= \int\int L_S(\theta,\delta(x))L(\theta;x)\pi(\theta)dx\,d\theta \\
&= \int\int L_S(\theta,\delta(x))\pi(\theta|x)h(x)dx\,d\theta \\
&= \int h(x)\left(\int L_S(\theta,\delta(x))\pi(\theta|x)d\theta\right)dx
\end{aligned}
$$

That is for each x we choose $\delta(x)$ to minimize the integral

$$\int L_S(\theta, \delta(\mathbf{x}))\pi(\theta|x)d\theta$$

Bayes rules for Point estimation

Zero-one loss Minimize

$$\begin{aligned}\int_{-\infty}^{\infty} \pi(\theta|x) L_S(\theta, \widehat{\theta}) d\theta &= \int_{\widehat{\theta}+b}^{\infty} \pi(\theta|x) d\theta + \int_{-\infty}^{\widehat{\theta}-b} \pi(\theta|x) d\theta \\ &= 1 - \int_{\widehat{\theta}-b}^{\widehat{\theta}+b} \pi(\theta|x) d\theta\end{aligned}$$

That is we want to maximize

$$\int_{\widehat{\theta}-b}^{\widehat{\theta}+b} \pi(\theta|x) d\theta$$

If $\pi(\theta|x)$ is unimodal the maximum is attained by choosing $\widehat{\theta}$ to be the mid-point of the interval of length $2b$ for which $\pi(\theta|b)$ has the same value at both ends.

If $\pi(\theta|x)$ is unimodal and symmetric, the optimal $\widehat{\theta}$ is the median (equal to the mean and mode) of the posterior distribution. As $b \longrightarrow 0$, $\widehat{\theta} \longrightarrow$ the global mode of the posterior distribution.

Absolute error loss

$$\int |\widehat{\theta}-\theta|\pi(\theta|x)d\theta = \int_{-\infty}^{\widehat{\theta}}(\widehat{\theta}-\theta)\pi(\theta|x)d\theta + \int_{\widehat{\theta}}^{\infty}(\theta-\widehat{\theta})\pi(\theta|x)d\theta.$$

Differentiate wrt $\widehat{\theta}$ and equate to zero.

$$\int_{-\infty}^{\widehat{\theta}}\pi(\theta|x)d\theta - \int_{\widehat{\theta}}^{\infty}\pi(\theta|x)d\theta = 0$$

That is, the optimal $\widehat{\theta}$ is the median of the posterior distribution.

Quadratic loss

Minimize

$$\mathbb{E}_{\theta|\mathbf{x}}[(\widehat{\theta}-\theta)^2] = [(\widehat{\theta}-\bar{\theta})^2] + \mathbb{E}[(\theta-\bar{\theta})^2]$$

where $\bar{\theta}$ is the posterior mean of θ. Note that $\hat{\theta}$ and $\bar{\theta}$ are constants in the posterior distribution of θ so that $(\hat{\theta}-\bar{\theta})\mathbb{E}(\theta - \bar{\theta}) = 0$. The Quadratic loss function is minimized when $\widehat{\theta} = \bar{\theta}$, the posterior mean.

Example

X is Binomial (n, θ), and the prior $\pi(\theta)$ is a Beta (α, β) distribution.

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha + \beta}$$

The distribution is unimodal if $\alpha, \beta > 1$ with mode

$$\frac{\alpha - 1}{\alpha + \beta - 2}.$$

The posterior distribution of $\theta \mid x$ is Beta $(\alpha + x, \beta + n - x)$. With zero-one loss and $b \to 0$ the Bayes estimator is $(\alpha + x - 1)/(\alpha + \beta + n - 2)$. For a quadratic loss function, the Bayes estimator is $(\alpha + x)/(\alpha + \beta + n)$, and for an absolute error loss function is the median of the posterior loss function.

Finding minimax rules

If δ is a Bayes rule for prior π, with $r(\pi, \delta) = C$, and δ_0 is a rule for which $\max_\theta R(\theta, \delta_0) = C$, then δ_0 is minimax.

Proof: (Y&S Ch 2) if for some other rule δ', $\max_\theta R(\theta, \delta') = C - \epsilon$ for some $\epsilon > 0$ (so δ_0 is not minimax), then $r(\pi, \delta') \leq C - \epsilon$ (the mean is less than or equal the maximum) and $r(\pi, \delta') < r(\pi, \delta)$ so δ is not the Bayes rule for π, a contradiction.

[*This is an informal treatment which assumes the min and max exist - see Y&S Ch 2 Sec 2.6*]

If δ is a Bayes rule for prior π with the property that $R(\theta, \delta)$ does not depend on θ, then δ is minimax.

Proof: (Y&S Ch 2) Let $R(\theta, \delta) = C$ (no θ dependence). For δ' as above, $r(\pi, \delta') \leq C - \epsilon$. But $r(\pi, \delta) = C$ so δ is not the Bayes rule for π, a contradiction.

This result is useful, as it gives an approach to finding minmax rules. Bayes rules are sometimes easy to compute, so if we find a prior that yields a Bayes rule with constant risk for all θ we have the minimax rule.

Bayes rules are 'nearly always' admissible (see GJJ Sec 6.2, Y&S Sec 2.7).

Exercise Suppose θ takes one of K possible values, with K finite, and π is a prior that puts non-zero probability on each possible θ. Show that if δ is a Bayes rule with respect to π, then δ is admissible.

Application: finding a minimax estimator for quadratic loss

The risk function is

$$\begin{aligned}
\mathbb{E}_{X|\theta}[(\widehat{\theta}-\theta)^2] &= [\text{Bias}(\widehat{\theta})]^2 + \text{Var}[\widehat{\theta}] \\
&= \left[\theta - \mathbb{E}\left(\frac{\alpha+X}{\alpha+\beta+n}\right)\right]^2 + \text{Var}\left[\frac{\alpha+X}{\alpha+\beta+n}\right] \\
&= \left[\theta - \left(\frac{\alpha+n\theta}{\alpha+\beta+n}\right)\right]^2 + \frac{n\theta(1-\theta)}{(\alpha+\beta+n)^2} \\
&= \frac{[\theta(\alpha+\beta)-\alpha]^2 + n\theta(1-\theta)}{[\alpha+\beta+n]^2}
\end{aligned}$$

The Bayes estimator with constant risk is minimax, for this to hold coefficients of θ and θ^2 in the numerator must be zero. That is $\alpha = \beta = \sqrt{n}/2$, so the minimax estimator using quadratic loss is $(\alpha + x)/(\alpha + \beta + n) = (x + \sqrt{n}/2)/(n + \sqrt{n})$.

Hypothesis testing with loss functions

$X_i \sim f(x;\theta)$ iid for $i = 1, 2, ..., n$: test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

Decision rule is δ_C when we use a critical region C so $\delta_C(x) = H_1$ if $x \in C$ and otherwise $\delta_C(x) = H_0$.

Loss function:

$$L_S(\theta, \delta_C(x)) = \begin{cases} a & \theta = \theta_0, x \in C \\ b & \theta = \theta_1, x \notin C \end{cases}$$

so the loss for Type I error is a (H_0 holds and we accept H_1) and the loss for Type II error is b (H_1 holds and we accept H_0).

The risk function for the rule δ_C is

$$\begin{aligned} R(\theta_0; \delta_C) &= \int L_S(\theta_0, \delta_C(x)) f(x; \theta_0) dx \\ &= \int a \mathbb{I}(x \in C) f(x; \theta_0) dx \\ &= a\alpha \end{aligned}$$

as $\alpha = P(X \in C | H_0)$ is the probability for Type I error, and

$$R(\theta_1; \delta_C) = b\beta,$$

as $\beta = P(X \notin C | H_1)$ is the probability for a Type II error.

Calculate the Bayes risk $r(\pi, \delta_C)$. Let $\pi(\theta_0) = p_0$ and $\pi(\theta_2) = p_1$ be the prior probabilities that H_0 and H_1 hold. The Bayes risk is

$$
\begin{aligned}
r(\pi, \delta_C) &= \sum_{\theta \in \{\theta_0, \theta_1\}} R(\theta; \delta_C)\pi(\theta) \\
&= p_0 a\alpha(C) + p_1 b\beta(C).
\end{aligned}
$$

The Bayes test chooses the critical region C to minimize the Bayes risk. Notice that, as we vary C, the levels α and β vary, so the level of the Bayes test is determined from the requirement that C minimizes the Bayes risk.

The Neyman-Pearson lemma states that the best test of size α of H_0 vs H_1 is a likelihood ratio test with critical region

$$C' = \left\{ x; \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} \geq A \right\}$$

for some constant $A > 0$ chosen so that $P(X \in C' | H_0) = \alpha$.

The following theorem tells us how to choose the critical region to minimize the Bayes risk, in the same way that the Neyman-Pearson lemma tells us how to maximize the power at fixed size.

Theorem (GJJ p129) the critical region for the Bayes test is the critical region for a LR test with

$$A = \frac{p_0 a}{p_1 b}$$

Every LR test is a Bayes test for some p_0, p_1.

Example Let $X_1, \ldots, X_n$ be N(μ, σ^2) with σ^2 known, and we want to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1$, with $\mu_1 > \mu_0$. The critical region for the LR test

$$\frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} \geq A$$

becomes

$$\bar{x} \geq \frac{\sigma^2 \log(A)}{n(\mu_1 - \mu_0)} + \frac{1}{2}(\mu_0 + \mu_1) = B(\text{ say})$$

In the classical case we ignore the exact value of B, but in a Bayes test $A = p_0 a/(p_1 b)$ and we substitute into B. As an example take

$$\mu_0 = 0, \mu_1 = 1, \sigma^2 = 1, n = 4, a = 2, b = 1, p_0 = 1/4, p_1 = 3/4$$

Then the Bayes test has critical region

$$\bar{x} \geq \frac{1}{4}\log(\frac{1}{3} \times \frac{2}{1}) + \frac{1}{2} = \frac{1}{4}\log(\frac{2}{3}) + \frac{1}{2} = 0.399$$

For this test (using the fact that $\bar{X}$ is $N(\mu, 1/4)$)

$$\alpha = P(\bar{X} \geq 0.3999 \mid \mu = 0, \sigma^2/n = 1/4) = 0.212$$

and

$$\beta = P(\bar{X} < 0.3999 \mid \mu = 1, \sigma^2/n = 1/4) = 0.115$$

In a classical approach fixing $\alpha = 0.05$, $B = 1.645\sqrt{1/4} = 0.822$, so

$$\beta = P(\bar{X} < 0.822 \mid \mu = 1, \sigma^2 = 1/4) = 0.363$$

In the Bayes test α has been increased and β decreased.

Stein's paradox and the James-Stein Estimator

Let $X_i \sim N(\mu_i, 1)$, $i = 1, 2, ..., p$ be jointly independent so we have one data point for each of the p μ_i-parameters. Let $X = (X_1, ..., X_p)$ and $\mu = (\mu_1, ..., \mu_p)$. The MLE $\hat{\mu}$ for μ is $\hat{\mu}_{MLE} = X$. This estimator is inadmissable for quadratic loss.

This is a paradox which forces us to think about the meaning of admissibility, and the implications of Quadratic Loss.

Proof (of the Stein paradox): Consider the alternative estimator

$$\hat{\mu} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X \quad \text{(the James-Stein estimator)}$$

We will show that if $a = p - 2$ then $R(\mu, \hat{\mu}) < R(\mu, \hat{\mu}_{MLE})$ for every $\mu \in R^n$, so that the MLE is inadmissable in this case.

First, the risk for $\hat{\mu}_{MLE}$ is

$$\begin{aligned} R(\mu, \hat{\mu}_{MLE}) &= \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_{MLE,i}|^2) \\ &= \sum_{i=1}^{p} \mathbb{E}(|\mu_i - X_i|^2) \\ &= p \end{aligned}$$

recognizing $\mathrm{Var}(X_i) = 1$.

In order to calculate $R(\mu, \hat{\mu})$, it is convenient to use Stein's Lemma, for Normal RV,

$$\mathbb{E}((X_i - \mu)h(X)) = \mathbb{E}\left(\frac{\partial h(X)}{\partial X_i}\right).$$

This can be shown by integrating by parts. Noting

$$\int (x_i - \mu)e^{-(x_i-\mu_i)^2/2}dx = -e^{-(x_i-\mu_i)^2/2}$$

we have

$$\int_{-\infty}^{\infty} (x_i - \mu_i)h(x)e^{-(x_i-\mu_i)^2/2}dx_i = -h(x)e^{-(x_i-\mu_i)^2/2}\Big|_{x_i=-\infty}^{x_i=\infty} + \int_{-\infty}^{\infty} \frac{\partial h(x)}{\partial x_i}e^{-(x_i-\mu_i)^2/2}dx_i$$

The first term is zero if $h(x)$ (for eg) is bounded, giving the lemma.

Now

$$R(\mu,\hat{\mu}) \;=\; \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) \quad \text{with} \quad \hat{\mu}_i = \left(1 - \frac{a}{\sum_i X_i^2}\right) X_i$$

$$\begin{aligned}
\mathbb{E}(|\mu_i - \hat{\mu}_i|^2) &= \mathbb{E}(|\mu_i - X_i|^2) - 2a\mathbb{E}\left(\frac{(X_i-\mu_i)X_i}{\sum_j X_j^2}\right) \\
&\qquad + a^2\mathbb{E}\left(\frac{X_i^2}{(\sum_j X_j^2)^2}\right) \\
\mathbb{E}\left(\frac{(X_i-\mu_i)X_i}{\sum_j X_j^2}\right) &= \mathbb{E}\left(\frac{\partial}{\partial X_i}\frac{X_i}{\sum_j X_j^2}\right) \quad \text{Stein's lemma} \\
&= \mathbb{E}\left(\frac{\sum_j X_j^2 - 2X_i^2}{(\sum_j X_j^2)^2}\right) = \mathbb{E}\left(\frac{1}{\sum_j X_j^2} - 2\frac{X_i^2}{(\sum_j X_j^2)^2}\right)
\end{aligned}$$

Putting the pieces together,

$$\begin{aligned}\sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) &= R(\mu, \hat{\mu}_{MLE}) - (2ap - 4a)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) + \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \\ &= p - (2ap - 4a - a^2)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\end{aligned}$$

and this is less than p if $2ap - 4a - a^2 > 0$ and in particular at $a = p - 2$, which minimizes the risk over $a \in R$.

We have shown that the obvious (MLE) estimator is inadmissable, and we do better to use an estimator in which data for different μ_i influence the value at other μ_i. This is surprising, given that the data are independent.

We have not shown that the James Stein estimator is admissible.