Biostat 602 Winter 2017

Lecture Set 13

Loss Function

Reading: CB 7.3.4

Bayesian Inference – Recap

- Allows making inference on the distribution of θ given data.
- Available information (from prior experiments) about θ can be utilized.
- Uncertainty of θ can be formally quantified.
- Misleading prior can result in misleading inference.
- Bayesian inference (especially the prior formulation) can be highly "subjective".
- Bayesian inference can be computationally intensive.

Ingredients

- **Prior** of θ : $\theta \sim \pi(\theta)$.
- **Sampling distribution** of $\mathbf{X}$ given θ.

$$\mathbf{X}|\theta \sim f(\mathbf{x}|\theta)$$

- Marginal distribution of $\mathbf{X}$

$$m(\mathbf{x}) = \int_{\theta\in\Omega} f(\mathbf{x},\theta)d\theta = \int_{\theta\in\Omega} f(\mathbf{x}|\theta)\pi(\theta)d\theta$$

- Bayesian inference is based on **Posterior distribution** of θ (conditional distribution of θ given $\mathbf{X}$)

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x},\theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \qquad \text{(Bayes' Rule)}$$

Bayes Estimator

Bayes Estimator of θ is defined as the posterior mean of θ.

$$E(\theta|\mathbf{x}) = \int_{\theta\in\Omega} \theta\pi(\theta|\mathbf{x})d\theta$$

We shall generalize this definition in this Lecture Set, but this is the most commonly accepted definition of Bayes estimator.

Conjugate Family

Definition 7.2.15: Let $\mathcal{F}$ denote the class of pdfs or pmfs for $f(x|\theta)$. A class Π of prior distributions is a conjugate family of $\mathcal{F}$, if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, and all priors in Π, and all $x \in \mathcal{X}$.

Example 1: Normal Bayes Estimators Let $X \sim \mathcal{N}(\theta, \sigma^2)$ and suppose that the prior distribution of θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ^2, τ^2 are all known, it follows, that

$$\begin{aligned}
\pi(\theta) &= \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\tau^2}\right] \\
f(x|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]
\end{aligned}$$

$$
\begin{aligned}
\pi(\theta|x) &\propto \pi(\theta) f(x|\theta) \\
&\propto \exp\left[-\frac{(\theta-\mu)^2}{2\tau^2} - \frac{(x-\theta)^2}{2\sigma^2}\right] \\
&= \exp\left[-\frac{\sigma^2(\theta-\mu)^2 + \tau^2(x-\theta)^2}{2\tau^2\sigma^2}\right] \\
&= \exp\left[-\frac{(\sigma^2+\tau^2)\theta^2 - 2(\sigma^2\mu + \tau^2 x)\theta + \sigma^2\mu^2 + \tau^2 x^2}{2\tau^2\sigma^2}\right] \\
&= \\
&\propto
\end{aligned}
$$

So $\theta|x$ also becomes normal, with mean and variance given by

$$
\begin{aligned}
\mathrm{E}[\theta|x] &= \frac{\tau^2}{\sigma^2+\tau^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu \\
\mathrm{Var}(\theta|x) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}
\end{aligned}
$$

- The normal family is its own conjugate family.
- The Bayes estimator for θ is a weighted average of the prior and sample means.
- As the prior variance τ^2 approaches to infinity (prior information becomes more vague), the Bayes estimator tends towards sample mean.

Loss/Risk Function

A **Loss Function** associated with point estimation is a real-valued non-negative function of the estimate and estimator, that is typically an increasing function of the distance between the two.

Let $\hat{\theta}$ be an estimator of θ and let $L(\hat{\theta}, \theta)$ be a function of θ and $\hat{\theta}$. Following are some examples of loss functions.

Squared error loss

$$L(\hat{\theta}, \theta) \;=\; (\hat{\theta} - \theta)^2$$

Weighted squared error loss

$$L(\hat{\theta}, \theta) \;=\; \omega(\theta)(\hat{\theta} - \theta)^2$$

where $\omega(\theta) \geq 0$ is a weight function.

Absolute error loss

$$L(\hat{\theta}, \theta) \;=\; |\hat{\theta} - \theta|$$

Asymmetric loss function

$$L(\theta, \hat{\theta}) \;=\; (\hat{\theta} - \theta)^2 I(\hat{\theta} < \theta) + 10(\hat{\theta} - \theta)^2 I(\hat{\theta} \geq \theta)$$

A loss that penalties overestimation more than underestimation

Relative squared error loss

$$L(\theta, \hat{\theta}) = \frac{(\hat{\theta} - \theta)^2}{|\theta| + 1}$$

This is a special case of weighted squared error loss. This loss penalizes errors in estimation more if θ is near 0 than if $|\theta|$ is large.

Stein's loss in variance estimation

$$L(\sigma^2, \hat{\sigma}^2) = \frac{\hat{\sigma}^2}{\sigma^2} - 1 - \log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right)$$

This loss is more complicated than squared error loss, but it has some reasonable properties. For any fixed value of σ^2, $L(\sigma^2, \hat{\sigma}^2) \to \infty$ as $\hat{\sigma}^2 \to 0$ or $\hat{\sigma}^2 \to \infty$. Thus, gross underestimation is penalized just as heavily as gross overestimation.

- All loss functions are non-negative
- The loss is zero when the estimator matches the parameter value

Risk Function

Definition: Risk function is expected loss of an estimator.

$$R(\theta, \hat{\theta}) \;=\; \mathrm{E}[L(\theta, \hat{\theta}(\mathbf{X}))|\theta]$$

Highlights on risk function

- If $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$, $R(\theta, \hat{\theta})$ is MSE.
- Loss and risk functions are not restricted to the Bayesian framework. It can be applied to any estimators.
- For example, UMVUE minimizes the risk function for squared error loss among all unbiased estimators, across all θ.
- Across all possible estimators, uniformly minimizing risk function across all θ is extremely difficult and often impossible (e.g. MSE).
- However, under the Bayesian framework where the distribution of θ is given, finding the best estimator is possible.

Bayes Risk

Bayes risk is defined as the average risk across all values of θ given prior $\pi(\theta)$

$$\int_{\Omega} R(\theta, \hat{\theta})\pi(\theta)d\theta$$

The Bayes rule with respect to a prior π is the optimal estimator with respect to a Bayes risk, which is defined as the one that minimize the Bayes risk.

Alternative definition of Bayes Risk

$$\begin{aligned}
\int_\Omega R(\theta,\hat{\theta})\pi(\theta)d\theta &= \int_\Omega \mathrm{E}[L(\theta,\hat{\theta}(\mathbf{X}))]\pi(\theta)d\theta \\
&= \int_\Omega \left[\int_\mathcal{X} f(\mathbf{x}|\theta)L(\theta,\hat{\theta}(\mathbf{x}))d\mathbf{x}\right]\pi(\theta)d\theta \\
&= \int_\Omega \left[\int_\mathcal{X} f(\mathbf{x}|\theta)L(\theta,\hat{\theta}(\mathbf{x}))\pi(\theta)d\mathbf{x}\right]d\theta \\
&= \int_\Omega \left[\int_\mathcal{X} \pi(\theta|\mathbf{x})m(\mathbf{x})L(\theta,\hat{\theta}(\mathbf{x}))d\mathbf{x}\right]d\theta \\
&= \int_\mathcal{X} \left[\int_\Omega \pi(\theta|\mathbf{x})L(\theta,\hat{\theta}(\mathbf{x}))d\theta\right]m(\mathbf{x})d\mathbf{x}
\end{aligned}$$

The quantity in square brackets is a function of $\mathbf{x}$ only. Minimizing the Bayes risk is equivalent to minimizing for each given $\mathbf{x} \in \mathcal{X}$, the quantity inside the bracket, which is called the *posterior expected loss*.

Posterior Expected Loss

$$\int_\Omega R(\theta,\hat{\theta})\pi(\theta)d\theta = \int_\mathcal{X} \left[\int_\Omega \pi(\theta|\mathbf{x})L(\theta,\hat{\theta}(\mathbf{x}))d\theta\right]m(\mathbf{x})d\mathbf{x}$$

Posterior expected loss is defined as

$$\mathrm{E}\left[L(\theta,\hat{\theta})|X=\mathbf{x}\right] = \int_\Omega \pi(\theta|\mathbf{x})L(\theta,\hat{\theta}(\mathbf{x}))d\theta$$

Bayes estimator is the estimator that minimizes the posterior expected loss.

Bayes Estimator based on squared error loss

$$
\begin{aligned}
L(\hat{\theta}, \theta) &= (\hat{\theta} - \theta)^2 \\
\text{Posterior expected loss} &= \int_{\Omega} (\theta - \hat{\theta})^2 \pi(\theta|\mathbf{x}) d\theta \\
&= \mathrm{E}[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}]
\end{aligned}
$$

So, the goal is to minimize $\mathrm{E}[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}]$

$$
\begin{aligned}
\mathrm{E}\left[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}\right] &= \mathrm{E}\left[\left(\theta - \mathrm{E}(\theta\mathbf{X}) + \mathrm{E}(\theta|\mathbf{X}) - \hat{\theta}\right)^2 \middle| \mathbf{X} = \mathbf{x}\right] \\
&= \mathrm{E}\left[(\theta - \mathrm{E}(\theta|\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] + \mathrm{E}\left[\left(\mathrm{E}(\theta|\mathbf{X}) - \hat{\theta}\right)^2 \middle| \mathbf{X} = \mathbf{x}\right] \\
&= \mathrm{E}\left[(\theta - \mathrm{E}(\theta|\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] + \left[\mathrm{E}(\theta|\mathbf{x}) - \hat{\theta}\right]^2
\end{aligned}
$$

which is minimized when $\hat{\theta} = \mathrm{E}(\theta|\mathbf{x})$.

Example 2 - Binomial Bayes estimator Let $X_1, \cdots, X_n$ be i.i.d. $Bernoulli(p)$, $p \sim \mathrm{Beta}(\alpha, \beta)$. Recall that

$$
\hat{p} = \frac{\sum_{i=1}^n X_i}{n} \qquad \hat{p}_B = \frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}
$$

are MLE and Bayes estimators of p, respectively. Assuming squared error loss,

1. What is the risk function of $\hat{p}$?
2. What is the risk function of $\hat{p}_B$?

3. Compare the Bayes risk between $\hat{p}$ and $\hat{p}_B$.

4. In the absence of good prior information about p, if we want to make risk function of $\hat{p}_B$ constant (based on squared error loss), what should be α and β?

5. Compare the risk functions between $\hat{p}$ and $\hat{p}_B$ from the previous problem, when $n = 4$ and $n = 400$.

Solution: For squared error loss, risk function is MSE. Now MSE of $\hat{p} = \overline{X}$ is

$$\mathrm{E}[\hat{p} - p]^2 = \mathrm{Var}(\overline{X}) = \frac{p(1-p)}{n}$$

On the other hand, risk function of $\hat{p}_B$ equals

$$\begin{aligned}
\mathrm{E}[\hat{p}_B - p]^2 &= \mathrm{Var}(\hat{p}_B) + [\mathrm{Bias}(\hat{p}_B)]^2 \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}\right) + \left[\mathrm{E}\left(\frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}\right) - p\right]^2 \\
&= \frac{np(1-p)}{(\alpha + \beta + n)^2} + \left[\frac{np + \alpha}{\alpha + \beta + n} - p\right]^2
\end{aligned}$$

Bayes Risk

For MLE $\hat{p}$

$$\begin{aligned} R(\hat{p},p) &= \mathrm{E}[\hat{p}-p]^2 = \mathrm{Var}(\overline{X}) = \frac{p(1-p)}{n} \\ \int_0^1 R(\hat{p},p)\pi(p)dp &= \int_0^1 \frac{p(1-p)}{n}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}dp \\ &= \tfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\tfrac{\Gamma(\alpha+1)\Gamma(\beta+1)}{n\Gamma(\alpha+\beta+2)}\int_0^1 \tfrac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)}p^{\alpha}(1-p)^{\beta}dp \\ &= \frac{\alpha\beta}{n(\alpha+\beta+1)(\alpha+\beta)} \end{aligned}$$

For Bayes estimator $\hat{p}_B$

$$\begin{aligned} R(\hat{p}_B,p) &= \mathrm{E}[\hat{p}_B-p]^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left[\frac{np+\alpha}{\alpha+\beta+n} - p\right]^2 \\ &= \frac{np(1-p)+\alpha^2(1-p)^2-2\alpha\beta p(1-p)+\beta^2p^2}{(\alpha+\beta+n)^2} \\ \mathrm{E}[R] &= \tfrac{\Gamma(\alpha+\beta)\left[(n-2\alpha\beta)\Gamma(\alpha+1)\Gamma(\beta+1)+\alpha^2\Gamma(\alpha)\Gamma(\beta+2)+\beta^2\Gamma(\alpha+2)\Gamma(\beta)\right]}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+2)(\alpha+\beta+n)^2} \\ &= \tfrac{\alpha\beta[n-2\alpha\beta+\alpha(\beta+1)+\beta(\alpha+1)]}{(\alpha+\beta+1)(\alpha+\beta)(\alpha+\beta+n)^2} \\ &= \frac{(n+\alpha+\beta)\alpha\beta}{(\alpha+\beta+n)^2(\alpha+\beta+1)(\alpha+\beta)} \\ &= \frac{\alpha\beta}{(\alpha+\beta+n)(\alpha+\beta+1)(\alpha+\beta)} \end{aligned}$$

Comparing two Bayes risks

$$\begin{aligned}
\textstyle\int_0^1 R(\hat{p},p)\pi(p)dp &= \frac{\alpha\beta}{n(\alpha+\beta+1)(\alpha+\beta)} \\
\textstyle\int_0^1 R(\hat{p}_B,p)\pi(p)dp &= \frac{\alpha\beta}{(\alpha+\beta+n)(\alpha+\beta+1)(\alpha+\beta)} \\
\frac{1}{(\alpha+\beta+n)} &\le \frac{1}{n}
\end{aligned}$$

$\hat{p}_B$ always has smaller Bayes risk than $\hat{p}$.

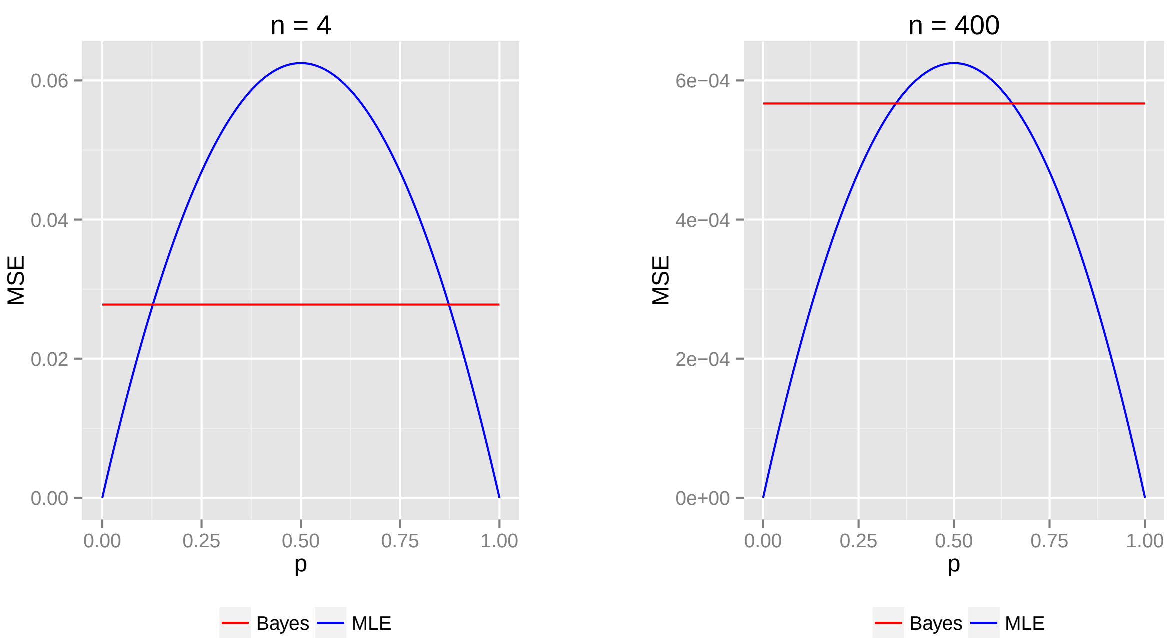
<u>Condition for constant risk function</u>

$$\begin{aligned}
\mathrm{E}[\hat{p}_B - p]^2 &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left[\frac{np+\alpha}{\alpha+\beta+n} - p\right]^2 \\
&= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left[\frac{\alpha-(\alpha+\beta)p}{\alpha+\beta+n}\right]^2 \\
&= \frac{[(\alpha+\beta)^2-n]p^2 + [n-2\alpha(\alpha+\beta)]p + \alpha^2}{(\alpha+\beta+n)^2}
\end{aligned}$$

$$\begin{aligned}
\alpha+\beta &= \sqrt{n} \\
\alpha &= \frac{n}{2(\alpha+\beta)} = \frac{1}{2}\sqrt{n} \\
\beta &= \sqrt{n} - \alpha = \frac{1}{2}\sqrt{n}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[\hat{p} - p]^2 &= \frac{p(1-p)}{n} \\
\mathrm{E}[\hat{p}_B - p]^2 &= \frac{[(\alpha+\beta)^2-n]p^2 + [n-2\alpha(\alpha+\beta)]p + \alpha^2}{(\alpha+\beta+n)^2} \\
&= \frac{n}{4(n+\sqrt{n})^2}
\end{aligned}$$

Comparing Bayesian risk functions



- There is no uniform winner. As p is closer to the boundaries of its domain, $\hat{p}_B$ is better than $\hat{p}$.
- As the sample size grows larger, there is a larger range of p for which $\hat{p}_B$ is superior to $\hat{p}$.

Different Bayes Estimators

Bayes estimators are minimizers of expected loss, and hence depend directly on the choice of loss function. Consider a point estimation problem for real-valued parameter θ.

Squared error loss

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})$$

The posterior expected loss is

$$\int_{\Omega} (\theta - \hat{\theta})^2 \pi(\theta|\mathbf{x}) d\theta \;\; = \;\; \mathrm{E}[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}]$$

This expected value is minimized by $\hat{\theta}_B = \mathrm{E}(\theta|\mathbf{x})$. So the Bayes estimator is the mean of the posterior distribution.

Absolute error loss

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

The posterior expected loss is

$$\begin{aligned}
\mathrm{E}[L(\theta, \hat{\theta})|\mathbf{x}] &= \mathrm{E}[|\theta - \hat{\theta}||\mathbf{X} = \mathbf{x}] \\
&= \int_{\Omega} |\theta - \hat{\theta}(\mathbf{x})| \pi(\theta|\mathbf{x}) d\theta \\
&= \int_{-\infty}^{\hat{\theta}} -(\theta - \hat{\theta}) \pi(\theta|\mathbf{x}) d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) \pi(\theta|\mathbf{x}) d\theta
\end{aligned}$$

In order to minimize the posterior expected loss, we make use of Leibnitz's rule

$$\frac{\partial}{\partial \theta} \int_{a(\theta)}^{b(\theta)} f(x|\theta) dx = f(b(\theta)|\theta) b'(\theta) - f(a(\theta)|\theta) a'(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x|\theta) dx$$

where the formula includes $a(\theta) = -\infty, \; b(\theta) = \infty$. Taking derivative with respect to $\hat{\theta}$ and setting it equal to zero, we have (using Leibnitz's rule)

$$\begin{aligned}
\frac{\partial}{\partial \hat{\theta}} \mathrm{E}[L(\theta, \hat{\theta}(\mathbf{x}))] &= -(\hat{\theta} - \hat{\theta}) \pi(\hat{\theta}|\mathbf{x}) + \int_{-\infty}^{\hat{\theta}} \pi(\theta|\mathbf{x}) d\theta \\
&\quad -(\hat{\theta} - \hat{\theta}) \pi(\hat{\theta}|\mathbf{x}) - \int_{\hat{\theta}}^{\infty} \pi(\theta|\mathbf{x}) d\theta = 0
\end{aligned}$$

The solution $\hat{\theta}_B$ satisfies

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

Thus, $\hat{\theta}_B$ is the <u>posterior median</u>. That it is the unique minimizer is easily verified by observing

$$\frac{\partial}{\partial\hat{\theta}}\left[\int_{-\infty}^{\hat{\theta}} \pi(\theta|\mathbf{x})d\theta - \int_{\hat{\theta}}^{\infty} \pi(\theta|\mathbf{x})d\theta\right] = 2\pi(\hat{\theta}|\mathbf{x}) > 0$$

Example 3: Normal Bayes Estimators Let $X_1, \cdots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ and suppose that the prior distribution of θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ^2, τ^2 are all known, what is the Bayes estimator based on (a) squared error loss and (b) the absolute error loss?

Solution: The posterior distributon of θ given $\mathbf{x}$ is normal with

$$\begin{aligned} \mathrm{E}[\theta|\mathbf{x}] &= \frac{\tau^2}{\tau^2 + \frac{1}{n}\sigma^2}\overline{x} + \frac{\frac{1}{n}\sigma^2}{\tau^2 + \frac{1}{n}\sigma^2}\mu \\ \mathrm{Var}(\theta|\mathbf{x}) &= \frac{\frac{1}{n}\sigma^2\tau^2}{\tau^2 + \frac{1}{n}\sigma^2} \end{aligned}$$

- For squared error loss, the Bayes estimator is $\hat{\theta} = \mathrm{E}[\theta|\mathbf{x}]$.
- For absolute error loss, the Bayes estimator is also $\hat{\theta} = \mathrm{E}[\theta|\mathbf{x}]$ (why?)

Example 4: Let $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$ and $\pi(p) \sim \text{Beta}(\alpha, \beta)$. What is the Bayes estimator with respect to (a) squared error loss and (b) absolute error loss?

Solution:

- The posterior distribution follows $\text{Beta}(\sum x_i + \alpha, n - \sum x_i + \beta)$.
- Bayes estimator that minimizes posterior expected squared error loss is the posterior mean

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

Bayes estimator that minimizes posterior expected absolute error loss is the posterior median satisfying

$$\int_0^{\hat{\theta}} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\sum x_i + \alpha)\Gamma(n - \sum x_i + \beta)} p^{\sum x_i + \alpha - 1}(1-p)^{n - \sum x_i + \beta - 1} dp = \frac{1}{2}$$

There is no closed form solution for $\hat{\theta}$, but it can be represented in terms of incomplete beta function.

Example 5: Let $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Consider an estimator of σ^2,

$$\sigma_b^2 \;=\; b s_{\mathbf{X}}^2 = \frac{b \sum_{i=1}^n (X_i - \overline{X})^2}{n-1},$$

i.e. consider an estimator in the class of scale multiples of the sample variance.

1. Using squared error loss, what is the b that minimizes Bayes risk?
2. Using Stein's loss function,

$$L(\sigma^2, \sigma_b^2) \;=\; \frac{\sigma_b^2}{\sigma^2} - 1 - \log \frac{\sigma_b^2}{\sigma^2}$$

   what is the b that minimizes Bayes risk?