Biostatistics 602 - Statistical Inference
Lecture 16
Evaluation of Bayes Estimator

Hyun Min Kang

March 14th, 2013

Last Lecture

- What is a Bayes Estimator?
- Is a Bayes Estimator the best unbiased estimator?
- Compared to other estimators, what are advantages of Bayes Estimator?
- What is conjugate family?
- What are the conjugate families of Binomial, Poisson, and Normal distribution?

Recap - Bayes Estimator

- θ : parameter
- $\pi(\theta)$: prior distribution
- $\mathbf{X}|\theta \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$: sampling distribution
- Posterior distribution of $\theta|\mathbf{x}$

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{\text{Joint}}{\text{Marginal}} = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \\ m(\mathbf{x}) &= \int f(\mathbf{x}|\theta)\pi(\theta)\,d\theta \qquad \text{(Bayes' rule)}\end{aligned}$$

- Bayes Estimator of θ is

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

Recap - Example

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$
- $\pi(p) \sim \text{Beta}(\alpha, \beta)$
- Prior guess : $\hat{p} = \frac{\alpha}{\alpha+\beta}$.
- Posterior distribution : $\pi(p|\mathbf{x}) \sim \text{Beta}(\sum x_i + \alpha, n - \sum x_i + \beta)$
- Bayes estimator

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

Loss Function Optimality

The mean squared error (MSE) is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[\hat{\theta} - \theta]^2$$

Let $\hat{\theta}$ is an estimator.

- If $\hat{\theta} = \theta$, it makes a correct decision and loss is 0
- If $\hat{\theta} \neq \theta$, it makes a mistake and loss is not 0.

Loss Function

Let $L(\theta, \hat{\theta})$ be a function of θ and $\hat{\theta}$.

- Squared error loss

$$\begin{aligned} L(\hat{\theta}, \theta) &= (\hat{\theta} - \theta)^2 \\ \mathrm{MSE} &= \text{Average Loss} = \mathrm{E}[L(\theta, \hat{\theta})] \end{aligned}$$

  which is the expectation of the loss if $\hat{\theta}$ is used to estimate θ.

- Absolute error loss

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

- A loss that penalties overestimation more than underestimation

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

Risk Function - Average Loss

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

If $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$, $R(\theta, \hat{\theta})$ is MSE.
An estimator with smaller $R(\theta, \hat{\theta})$ is preferred.

Definition : 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

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

Posterior Expected Loss

Posterior expected loss is defined as

$$\int_\Omega \pi(\theta|\mathbf{x}) L(\theta, \hat{\theta}(\mathbf{x}))\, d\theta$$

An alternative definition of Bayes rule 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 $\mathrm{E}[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}]$

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

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

Summary so far

Loss function $L(\theta, \hat{\theta})$
- e.g. $(\hat{\theta} - \theta)^2$, $|\hat{\theta} - \theta|$

Risk function $R(\theta, \hat{\theta})$ is average of $L(\theta, \hat{\theta})$ across all $x \in \mathcal{X}$
- For squared error loss, risk function is the same to MSE.

Bayes risk Average risk across all θ, based on the prior of θ.
- Alternatively, average posterior error loss across all $x \in \mathcal{X}$.

Bayes estimator $\hat{\theta} = \mathrm{E}[\theta|\mathbf{x}]$. Based on squared error loss,
- Minimize Bayes risk
- Minimize Posterior Expected Loss

Bayes Estimator based on absolute error loss

Suppose that $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$. The posterior expected loss is

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

$\frac{\partial}{\partial \hat{\theta}} \mathrm{E}[L(\theta, \hat{\theta}(\mathbf{x}))] = 0$, and $\hat{\theta}$ is posterior median.

Two Bayes Rules

Consider a point estimation problem for real-valued parameter θ.

For squared error loss, 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 $\hat{\theta} = \mathrm{E}(\theta|\mathbf{x})$. So the Bayes rule estimator is the mean of the posterior distribution.

For absolute error loss, the posterior expected loss is $\mathrm{E}(|\theta - \hat{\theta}||\mathbf{X} = \mathbf{x})$. As shown previously, this is minimized by choosing $\hat{\theta}$ as the median of $\pi(\theta|\mathbf{x})$.

Example

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(p)$.
- $\pi(p) \sim \mathrm{Beta}(\alpha, \beta)$
- The posterior distribution follows $\mathrm{Beta}(\sum x_i + \alpha, n - \sum x_i + \beta)$.
- 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

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

Asymptotic Evaluation of Point Estimators

When the sample size n approaches infinity, the behaviors of an estimator are unknown as its *asymptotic* properties.

Definition - Consistency

Let $W_n = W_n(X_1, \cdots, X_n) = W_n(\mathbf{X})$ be a sequence of estimators for $\tau(\theta)$. We say W_n is consistent for estimating $\tau(\theta)$ if $W_n \xrightarrow{\mathrm{P}} \tau(\theta)$ under P_θ for every $\theta \in \Omega$.

$W_n \xrightarrow{\mathrm{P}} \tau(\theta)$ (converges in probability to $\tau(\theta)$) means that, given any $\epsilon > 0$.

$$\lim_{n\to\infty} \Pr(|W_n - \tau(\theta)| \geq \epsilon) = 0$$
$$\lim_{n\to\infty} \Pr(|W_n - \tau(\theta)| < \epsilon) = 1$$

When $|W_n - \tau(\theta)| < \epsilon$ can also be represented that W_n is close to $\tau(\theta)$. Consistency implies that the probability of W_n close to $\tau(\theta)$ approaches to 1 as n goes to ∞.

Tools for proving consistency

- Use definition (complicated)
- Chebychev's Inequality

$$\begin{aligned} \Pr(|W_n - \tau(\theta)| \geq \epsilon) &= \Pr((W_n - \tau(\theta))^2 \geq \epsilon^2) \\ &\leq \frac{\mathrm{E}[W_n - \tau(\theta)]^2}{\epsilon^2} \\ &= \frac{\mathrm{MSE}(W_n)}{\epsilon^2} = \frac{\mathrm{Bias}^2(W_n) + \mathrm{Var}(W_n)}{\epsilon^2} \end{aligned}$$

Need to show that both $\mathrm{Bias}(W_n)$ and $\mathrm{Var}(W_n)$ converges to zero

Theorem for consistency

Theorem 10.1.3

If W_n is a sequence of estimators of $\tau(\theta)$ satisfying

- $\lim_{n->\infty} \text{Bias}(W_n) = 0$.
- $\lim_{n->\infty} \text{Var}(W_n) = 0$.

for all θ, then W_n is consistent for $\tau(\theta)$

Weak Law of Large Numbers

Theorem 5.5.2

Let $X_1, \cdots, X_n$ be iid random variables with $\text{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2 < \infty$. Then $\overline{X}_n$ converges in probability to μ. i.e. $\overline{X}_n \xrightarrow{\text{P}} \mu$.

Consistent sequence of estimators

Theorem 10.1.5

Let W_n is a consistent sequence of estimators of $\tau(\theta)$. Let a_n, b_n be sequences of constants satisfying

1. $\lim_{n\to\infty} a_n = 1$
2. $\lim_{n\to\infty} b_n = 0$.

Then $U_n = a_n W_n + b_n$ is also a consistent sequence of estimators of $\tau(\theta)$.

Continuous Map Theorem

If W_n is consistent for θ and g is a continuous function, then $g(W_n)$ is consistent for $g(\theta)$.

Example

Problem

$X_1, \cdots, X_n$ are iid samples from a distribution with mean μ and variance $\sigma^2 < \infty$.

1. Show that $\overline{X}_n$ is consistent for μ.
2. Show that $\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ is consistent for σ^2.
3. Show that $\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is consistent for σ^2.

Example - Solution

Proof: $\overline{X}_n$ is consistent for μ

By law of large numbers, $\overline{X}_n$ is consistent for μ.

- $\text{Bias}(\overline{X}_n) = \text{E}(\overline{X}_n) - \mu = \mu - \mu = 0.$
- $\text{Var}(\overline{X}_n) = \text{Var}\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \sigma^2/n.$
- $\lim_{n\to\infty} \text{Var}(\overline{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$

By Theorem 10.1.3. $\overline{X}$ is consistent for μ.

Solution - consistency for σ^2

$$\begin{aligned}\frac{\sum(X_i - \overline{X})^2}{n} &= \frac{\sum(X_i^2 + \overline{X}^2 - 2X_i\overline{X})}{n} \\ &= \frac{\sum X_i^2 + n\overline{X}^2 - 2\overline{X}\sum_{i=1}^n X_i}{n} \\ &= \frac{\sum X_i^2}{n} - \overline{X}^2\end{aligned}$$

By law of large numbers,

$$\frac{1}{n}\sum X_i^2 \xrightarrow{\text{P}} \text{E}X^2 = \mu^2 + \sigma^2$$

Note that $\overline{X}^2$ is a function of $\overline{X}$. Define $g(x) = x^2$, which is a continuous function. Then $\overline{X}^2 = g(\overline{X})$ is consistent for μ^2. Therefore,

$$\frac{\sum(X_i - \overline{X}_n)^2}{n} = \frac{\sum X_i^2}{n} - \overline{X}^2 \xrightarrow{\text{P}} (\mu^2 + \sigma^2) - \mu^2 = \sigma^2$$

Solution - consistency for σ^2 (cont'd)

From the preious slide, $\sum(X_i - \overline{X}_n)^2/n$ is consistent for σ^2.
Define $S_n^2 = \frac{1}{n-1}\sum(X_i - \overline{X}_n)^2$, and $(S_n^*)^2 = \frac{1}{n}\sum(X_i - \overline{X}_n)^2$.

$$S_n^2 = \frac{1}{n-1}\sum(X_i - \overline{X}_n)^2 = (S_n^*)^2 \cdot \frac{n}{n-1}$$

Because $(S_n^*)^2$ was shown to be consistent for σ^2 previously, and $a_n = \frac{n}{n-1} \to 1$ as $n \to \infty$, by Theorem 10.1.5, S_n^2 is also consistent for σ^2.

Example - Exponential Family

Problem

Suppose $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\beta)$.

1. Propose a consistent estimator of the median.
2. Propose a consistent estimator of $\Pr(X \leq c)$ where c is constant.

Consistent estimator for the median

First, we need to express the median in terms of the parameter β.

$$\begin{aligned}
\int_0^m \frac{1}{\beta} e^{-x/\beta} dx &= \frac{1}{2} \\
-e^{-x/\beta}\Big|_0^m &= \frac{1}{2} \\
1 - e^{-m/\beta} &= \frac{1}{2} \\
\text{median} &= m = \beta \log 2
\end{aligned}$$

By law of large numbers, $\overline{X}_n$ is consistent for $\mathrm{E}X = \beta$.
Applying continuous mapping Theorem to $g(x) = x \log 2$, $g(\overline{X}) = \overline{X}_n \log 2$ is consistent for $g(\beta) = \beta \log 2$ (median).

Consistent estimator of $\Pr(X \leq c)$

$$\begin{aligned}
\Pr(X \leq c) &= \int_0^c \frac{1}{\beta} e^{-x/\beta} dx \\
&= 1 - e^{-c/\beta}
\end{aligned}$$

As $\overline{X}$ is consistent for β, $1 - e^{-c/\beta}$ is continuous function of β.
By continuous mapping Theorem, $g(\overline{X}) = 1 - e^{-c/\overline{X}}$ is consistent for $\Pr(X \leq c) = 1 - e^{-c/\beta} = g(\beta)$

Consistent estimator of $\Pr(X \leq c)$ - Alternative Method

Define $Y_i = I(X_i \leq c)$. Then $Y_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ where $p = \Pr(X \leq c)$.

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n I(X_i \leq c)$$

is consistent for p by Law of Large Numbers.

Summary

Today

- Bayes Risk Functions
- Consistency
- Law of Large Numbers

Next Lecture

- Central Limit Theorem
- Slutsky Theorem
- Delta Method