Ch 7. Point estimation

Intro

Definition

A function of random variables $X_1, \cdots, X_n$ is called *Statistic*.

Definition

The probability distribution of a statistic T is called the sampling distribution of T.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Let $T = \bar{X}$, then sampling distribution of T is

$$N(\qquad\qquad , \qquad\qquad).$$

If $T = S^2$ then sampling distribution of S^2 is

$$(n-1)S^2/\sigma^2 \sim \qquad\qquad .$$

Ch 7. Point estimation

Intro

Definition

An *estimator* is a function of random variables $X_1, \cdots, X_n$, $T = W(X_1, \cdots, X_n)$.

$\lhd$ Note:

1. Estimator is actually a statistic.
2. Estimator is also random.
3. An *estimate* is a function of realized values of $X_1 = x_1, \cdots, X_n = x_n$. $t = W(x_1, \cdots, x_n)$

$\rhd$ Example: $T = \bar{X}$, $\hat{F}_n(x_0) = n^{-1} \sum_{i=1}^{n} I[X_i \leq x_0]$, $T = (\bar{X}, S^2)$.

How to estimate the parametric function $\tau(\boldsymbol{\theta})$ using the random sample $X_1, \cdots, X_n$? : MME, MLE, BE so on
Let $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k) \in \Theta$.

Definition

j^{th} Population moment: $\mu_j(\boldsymbol{\theta}) = E\left(X^j\right)$ | $C\mu_j = E\left[(X-\mu)^j\right]$

j^{th} Sample moment: $m_j = \dfrac{1}{n}\sum_{i=1}^{n} X_i^j$ | $Cm_j = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^j$

◁ Note:

1. μ_j is a function of $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$.
2. $E\left[m_j\right] = E\left[n^{-1}\sum_{i=1}^{n} X_i^j\right] = \mu_j$.

Definition

MME of $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$, denoted by $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_k)$, is defined as a solution of the system of equations

$$m_j = \mu_j(\theta_1, \cdots, \theta_k),\ j = 1, \cdots .l$$

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find MME of $\boldsymbol{\theta} = (\mu, \sigma^2)$.

$$m_1 = \bar{X} = \mu,\ m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu^2 + \sigma^2.$$

$$\tilde{\mu} = \bar{X} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2.$$

▷ Example: $X_1, \cdots, X_n \stackrel{iid}{\sim}$ Gamma(α, β). Find MME of $\boldsymbol{\theta} = (\alpha, \beta)$.

Let $S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$

$$\theta_1\theta_2 = \bar{X}, \qquad \theta_1\theta_2^2 = S^{*2},$$

$$\Rightarrow \quad \tilde{\theta}_1 = \frac{\bar{X}^2}{S^{*2}} \quad \tilde{\theta}_2 = \frac{S^{*2}}{\bar{X}}.$$

$\triangleleft$ Note:

1. MM equations may have multiple solutions or no solution. The solution may fall outside of the parameter space. (Example 7.2.2)
2. MME may not be applicable if the population moments do not exist such as Cauchy distribution.
3. One may not successful considering the first k–moments.

   $\triangleright$ Example: $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\beta)$, where

$$f(x|\beta) = \frac{1}{2\beta} e^{-|x|/\beta}.$$

Ch 7. Point estimation

Methods: MLE

Given the sample observation (measurement, data), choose as the estimator of the population parameter the value that makes the observed sample result the most likely, i.e. maximize the *likelihood*

▷ Let X be a single observation from a discrete distribution taking the values { 0,1, 2 }.

The data comes from one of two distributions:

	$P(X=0)$	$P(X=1)$	$P(X=2)$
$\theta=\theta_0$	.8	.1	.1
$\theta=\theta_1$	.2	.3	.5

$$\text{then } \hat{\theta}=\begin{cases}\theta_0 & \text{if } x=0\\ \theta_1 & \text{if } x=1,2\end{cases}.$$

Definition

Let $f(x_1, \cdots, x_n|\boldsymbol{\theta})$ be the joint pdf/pmf of $X_1, \cdots, X_n$. For a fixed $x_1, \cdots, x_n$,

$$L(\boldsymbol{\theta}) = f(x_1, \cdots, x_n|\boldsymbol{\theta})$$

as a function of $\boldsymbol{\theta}$, is called the likelihood function. $\ln[L(\boldsymbol{\theta})]$ is called the log likelihood function.

With discrete random variable,

$$L(\boldsymbol{\theta}) = P(X_1 = x_1, \cdots, X_n = x_n).$$

$X_1, \cdots, X_n \overset{iid}{\sim} f(x|\boldsymbol{\theta})$. $L(\boldsymbol{\theta}) =$

Definition

Let $f(x_1, \cdots, x_n|\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta$ be the joint pdf/pmf of $X_1, \cdots, X_n$. Then for a given set of observations $(x_1, \cdots, x_n)$, the *maximum likelihood estimate* of $\boldsymbol{\theta}$ is a point $\boldsymbol{\theta}_0 = h(x_1, \cdots, x_n)$ satisfying

$$f(x_1, \cdots, x_n|\boldsymbol{\theta}_0) = \max_{\theta \in \Theta} f(x_1, \cdots, x_n|\boldsymbol{\theta}).$$

The *maximum likelihood estimator (MLE)* is defined as

$$\hat{\boldsymbol{\theta}} = h(X_1, \cdots, X_n).$$

Ch 7. Point estimation

Methods: MLE

How to find MLE ?

- Using differentiation, - Direct maximization, - Numerical evaluation

- Assume $L(\theta)$ is twice differentiable in the interior points of Θ. Then $\hat{\theta}$ maximizes $L(\theta)$ if
  1. $\hat{\theta}$ is the unique value satisfying
  $$\frac{dL(\theta)}{d\theta} \left(\frac{d\ln[L(\theta)]}{d\theta} \right) \Big|_{\hat{\theta}} = 0$$
  $$\frac{d^2L(\theta)}{d\theta^2} \left(\frac{d^2\ln[L(\theta)]}{d\theta^2} \right) \Big|_{\hat{\theta}} < 0$$
  2. The maximizer does not occur at the boundary.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Known σ and Unknown σ

$$L\left(\theta,\sigma^2 \middle| \mathbf{x}\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2}\sum\left(x_i-\theta\right)^2/\sigma^2\right\}$$

$$l\left(\theta,\sigma^2 \middle| \mathbf{x}\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum\left(x_i-\theta\right)^2/\sigma^2$$

$$0 = \frac{\partial l}{\partial\theta} = \frac{1}{\sigma^2}\sum\left(x_i-\theta\right) \Rightarrow \sum x_i - n\theta = 0 \Rightarrow \hat{\theta} = \bar{x};$$

$$0 = \frac{\partial l}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum\left(x_i-\theta\right)^2.$$

$$\hat{\sigma}^2 = n^{-1}\sum\left(x_i-\bar{x}\right)^2.$$

$\triangleright$ Example: $X_{ij},\ i = 1, \cdots, s;\ j = 1, \cdots, n$ independently distributed as normal distribution with mean μ_i and variance σ^2. Find the mle of μ_i and σ^2.

$$L = (\sqrt{2\pi\sigma^2})^{-ns} \exp\left(-\sum_{i=1}^{s}\sum_{j=1}^{n} \frac{(x_{ij}-\mu_i)^2}{2\sigma^2} \right)$$

$$l = \ln L = -\frac{ns}{2}\ln(2\pi) - \frac{ns}{2}\ln(\sigma^2) - \sum_{i=1}^{s}\sum_{j=1}^{n} \frac{(x_{ij}-\mu_i)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial(\mu_i)} = \frac{1}{\sigma^2}\sum_{i=1}^{s}\sum_{j=1}^{n}(x_{ij}-\mu_i) = 0,\ i = 1, 2, \cdots, s,$$

$$\frac{\partial l}{\partial(\sigma^2)} = -\frac{ns}{2\sigma^2} + \sum_{i=1}^{s}\sum_{j=1}^{n} \frac{(x_{ij}-\mu_i)^2}{2(\sigma^2)^2} = 0$$

$$\hat{\mu}_i = \bar{X}_{i\bullet},\quad \hat{\sigma}^2 = \frac{1}{ns}\sum_{i=1}^{s}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i\bullet})^2$$

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} Bernoulli(p)$.

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$l(p|\mathbf{x}) = \sum x_i \log p + \left(n - \sum x_i\right)\log(1-p)$$

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{\hat{p}} - \frac{n-\sum x_i}{1-\hat{p}} = 0 \text{ (if } \hat{p} \notin \{0,1\}\text{)}. \quad \Rightarrow \hat{p} = \overline{x}\,.$$

Theorem

Let $\hat{\boldsymbol{\theta}}$ be the mle of $\boldsymbol{\theta}$. Then for any parametric function $\tau(\boldsymbol{\theta})$, the mle of $\tau(\boldsymbol{\theta})$ is defined to be $\tau(\hat{\boldsymbol{\theta}})$

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find mle of e^μ, μ^2, σ/μ and $P[X \leq a]$.

◁ Note:

1. It is possible that the likelihood equations do not have closed-form solution. May need a numerical method.
2. When the likelihood function is not differentiable, we may maximize $L(\boldsymbol{\theta})$ directly.

- Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Gamma(α, β).

$\alpha = \theta1, \ \beta = \theta2$

$$\Gamma(t) \overset{\Delta}{=} \int_0^\infty y^{t-1} e^{-y} dy$$

$$L(\theta_1, \theta_2 | x_1, \ldots, x_n) = \left[\frac{1}{\Gamma(\theta_1)\theta_2^{\theta_1}}\right]^n (x_1 x_2 \cdots x_n)^{\theta_1 - 1} \exp\left(-\sum_1^n x_i/\theta_2\right)$$

The likelihood equations do not have closed-form solution.

Example (Uniform Distribution). Let $X_1, \ldots, X_n$ be iid with the uniform $(0, \theta)$ density, i.e.,

$$f(x;\theta) = \begin{cases} 1/\theta, & 0 < x \le \theta \\ 0, & elsewhere \end{cases}$$

Find the $\hat{\theta}_{MLE}$.

- We have

$$L(\theta; x_1, x_2, \cdots, x_n) = \frac{1}{\theta^n}, \qquad 0 < x_i \le \theta$$

which is an ever-decreasing function of θ.

- The maximum of such functions cannot be found by differentiation. Note that

$$0 < x_i \le \theta \quad \Longrightarrow \quad \theta \ge \max(x_i)$$

- L can be made no larger than $\frac{1}{[\max(x_i)]^n}$ and the unique MLE $\hat{\theta}$ of θ is the nth order statistic $\max(X_i)$.

$$E[\max(X_i)] = n\theta/(n+1).$$

- The MLE of the parameter θ is biased.

$$E[\max(X_i)] = \int_0^\theta x \cdot nf(x)F^{n-1}(x)dx$$

$$= n\int_0^\theta x \cdot \frac{1}{\theta}(\frac{x}{\theta})^{n-1}dx = \frac{n}{n+1}\theta.$$

So far, $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is **unknown** and is assumed **fixed**. In **Bayesian** framework, $\boldsymbol{\theta}$ is assumed **random**. The distribution of $\boldsymbol{\theta}$ is called **prior distribution**, denoted by

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

▷ Example: Machine that stamps out parts for cars. θ=fraction of defective. On a certain day, 10 pieces are examined.

$$X_i = \begin{cases} 1, & \text{if } i\text{th piece is defective,} \\ 0, & \text{otherwise,} \end{cases}$$

$i = 1, \cdots, 10$. MME or MLE ?

▷ Example -continued: Now assume that mechanic knows something about θ and gives a statistical model for θ

$$\pi(\theta) = 6\theta(1-\theta),\ 0 \le \theta \le 1.$$

Prior distribution of θ is ?

- In Baysian frame, what is the goal of the inference about θ ?
- Then, How should we use the data $X_1 = x_1, \cdots, X_n = x_n$ to achieve the goal ?

The conditional distribution of $\boldsymbol{\theta}$ conditioning on $X_1 = x_1, \cdots, X_n = x_n$ is called **posterior distribution** of $\boldsymbol{\theta}$.

$$\pi(\boldsymbol{\theta}|x_1, \cdots, x_n) = \frac{f(x_1, \cdots, x_n, \boldsymbol{\theta})}{m(x_1, \cdots, x_n)}$$

$$= \begin{cases} \dfrac{\pi(\boldsymbol{\theta}) f(x_1, \cdots, x_n|\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) f(x_1, \cdots, x_n|\boldsymbol{\theta})} \\ \\ \dfrac{\pi(\boldsymbol{\theta}) f(x_1, \cdots, x_n|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) f(x_1, \cdots, x_n|\boldsymbol{\theta}) d\boldsymbol{\theta}} \end{cases}$$

• Any Bayesian inference is based on this posterior distribution of $\boldsymbol{\theta}$.

Example 1

Let $X_1,\ldots,X_n$ be i.i.d. Bernoulli(p).

$$\pi(p)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}\ (p\in(0,1))$$

$$f(\mathbf{x}|p)\sim p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

We determine the Bayes estimator of p.

$$\pi(p|\mathbf{x})=\frac{f(\mathbf{x}|p)\pi(p)}{f(\mathbf{x})}$$

$$=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}$$

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

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma\left(\sum x_i+\alpha\right)\Gamma\left(n-\sum x_i+\beta\right)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}.$$

$$p|\mathbf{x} \sim \text{Beta}\left(\sum x_i+\alpha,\ n-\sum x_i+\beta\right)$$

So the Bayes estimator is $E\left(p|\mathbf{x}\right)=\dfrac{\sum x_i+\alpha}{n+\alpha+\beta}$.

Note: We must choose α and β somehow. (Often used is $\alpha=\beta=1$ which makes $\pi(p)$ uniform. In that case we are not using prior information about p).

Our prior has mean $\dfrac{\alpha}{\alpha+\beta}$, so without seeing the data we use this. The MLE is $\dfrac{\sum x_i}{n}$.

Then the Bayes estimate

$$\frac{\sum x_i+\alpha}{n+\alpha+\beta}=\left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)\left(\frac{\alpha}{\alpha+\beta}\right)+\left(\frac{n}{n+\alpha+\beta}\right)\frac{\sum x_i}{n},$$

a weighted average of the prior mean and the MLE.

Note that

$$\pi(p|\mathbf{x}) = \frac{f(\mathbf{x}|p)\pi(p)}{f(\mathbf{x})} \propto f(\mathbf{x}|p)\pi(p)$$

$$\propto p^{\sum x_i}(1-p)^{n-\sum x_i} \; p^{\alpha-1}(1-p)^{\beta-1}$$

$$\propto p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}.$$

$$p|\mathbf{x} \sim \text{Beta}\left(\sum x_i + \alpha,\; n-\sum x_i+\beta\right)$$

Example 2

Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ).

$$\pi(\lambda) = \frac{\lambda^{\alpha-1} e^{-\lambda/\beta}}{\Gamma(\alpha)\beta^{\alpha}},\ 0 < \lambda < \infty;\ Gamma(\alpha, \beta)$$

$$f(\mathbf{x}|\lambda) \sim \frac{e^{-n\lambda} \lambda^{\sum x_i}}{x_1! x_2! \cdots x_n!}$$

We determine the Bayes estimator of λ.

$$\pi(\lambda|\mathbf{x}) = \frac{f(\mathbf{x}|\lambda)\pi(\lambda)}{f(\mathbf{x})} \propto f(\mathbf{x}|\lambda)\pi(\lambda)$$

$$\propto e^{-n\lambda} \lambda^{\sum x_i} \lambda^{\alpha-1} e^{-\lambda/\beta}$$

$$\propto \lambda^{\sum x_i + \alpha - 1} e^{-\lambda(n + \frac{1}{\beta})}$$

$$\sim \text{Gamma}(\textstyle\sum x_i + \alpha,\ \frac{1}{n+1/\beta})$$

$$E(\lambda|\mathbf{X}) = \frac{\sum x_i + \alpha}{n + 1/\beta}$$

In each of these examples, the prior and posterior distributions have the same form.

In such cases, the distribution of the prior and sampling distribution of $\mathbf{x}$ are said to be **conjugate**.

Choosing a conjugate prior has the advantage of allowing us to write a closed form expression for the posterior and for the estimator. It makes the computation easier.

Examples of conjugate priors

- Bernoulli model with a Beta prior on $0 < p < 1$
- Poisson model with a Gamma prior on $0 < \lambda < \infty$
- Normal model with a Normal prior on $-\infty < \mu < \infty$
- Normal model with an Inverse Gamma prior on $-\infty < \sigma^2 < \infty$ (i.e. a Gamma prior on $0 < \frac{1}{\sigma^2} < \infty$).
- Gamma model with a Gamma prior on $0 < \beta < \infty$

Example 3

Let $X_1,\ldots,X_n$ be i.i.d. $N\left(\mu,\sigma^2\right)$.

$$\pi\left(\mu\right) \sim N\left(\eta,\tau^2\right)$$

$$f\left(\mathbf{x}|\mu\right) \sim N\left(\mu,\sigma^2\right)$$

Here σ^2, η, and τ^2 are all assumed to be known. We determine the Bayes estimator of μ.

- May have more than one choice of the estimator of the parameter.
  - Need to evaluate the estimators so that we can choose the best one.
  - Need a criterion to evaluate the estimator. (Unbiasedness, MSE, Consistency, BLUE, UMVUE)

Definition

An estimator $W(X_1, \cdots, X_n)$ of a parametric function $\tau(\theta)$ is said to be an *unbiased* estimator (UE) if

$$E_\theta(W) = \tau(\theta), \quad \text{for all} \quad \theta \in \Theta.$$

Definition

The function

$$\mathrm{BIAS}_\theta(W) = E_\theta(W) - \tau(\theta)$$

is called the *bias* of W as an estimator of $\tau(\theta)$.

$\triangleright$ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find an unbiased estimator for μ and σ^2. What about the MLE of σ^2 ?

Q) If W is an UE of θ, then is $\tau(W)$ UE of $\tau(\theta)$? Yes/No

$\triangleright$ Example-Continued: S^2 is an UE of σ^2. Is S unbiased for σ ?

Definition

The *Mean Squared Error (MSE)* of an estimator W of a parameter θ is the function of θ defined by

$$\mathsf{MSE}_\theta(W) = E_\theta(W - \theta)^2.$$

$\triangleleft$ Note: $\mathsf{MSE}_\theta(W) = Var_\theta(W) + [\mathsf{BIAS}_\theta(W)]^2$

MSE has two components, one that measures precision (variance), and one that measures accuracy (bias).

Example

Let $X_1, \ldots, X_n$ be i.i.d. $N\left(\mu, \sigma^2\right)$. We $E\left(\bar{X}\right) = \mu$ and $E\left(S^2\right) = \sigma^2$ for all μ and σ^2, so they are both unbiased.

The MSEs of these estimators are

$Var\left(\bar{X}\right) = \sigma^2 / n$, $Var\left(S^2\right) = 2\sigma^4 / \left(n-1\right)$.

(Recall that $\dfrac{\left(n-1\right)S^2}{\sigma^2} \sim \chi^2_{n-1}$, and that $Var\left(\chi^2_v\right) = 2v$).

Which one is better?

MLE $\hat{\sigma}^2 = n^{-1}(n-1)S^2, \quad \tilde{\sigma}^2 = S^2 \quad ?$

For the MLE (and MOM) estimator of σ^2,

$$\hat{\sigma}^2 = \frac{1}{n}\sum\left(x_i - \overline{x}\right)^2 \ ,$$

$$E\left(\hat{\sigma}^2\right) = \frac{n-1}{n}E\left(S^2\right) = \frac{n-1}{n}\sigma^2.$$

Thus

$$Bias\left(\hat{\sigma}^2\right) = E\left(\hat{\sigma}^2\right) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{-\sigma^2}{n}$$

$$Var\left(\hat{\sigma}^2\right) = \left(\frac{n-1}{n}\right)^2 Var\left(S^2\right) = 2\left(n-1\right)\sigma^4 / n^2$$

Hence, $MSE\left(\hat{\sigma}^2\right) = Var\ \hat{\sigma}^2 + \left(Bias\ \hat{\sigma}^2\right)^2$

$$= 2\left(n-1\right)\sigma^4 / n^2 + \left(-\sigma^2 / n\right)^2$$

$$= \left[\left(2/n\right) - \left(1/n^2\right)\right]\sigma^4 < 2\sigma^4 / \left(n-1\right)$$

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Bernoulli(p). With Beta prior, the Bayes estimator is

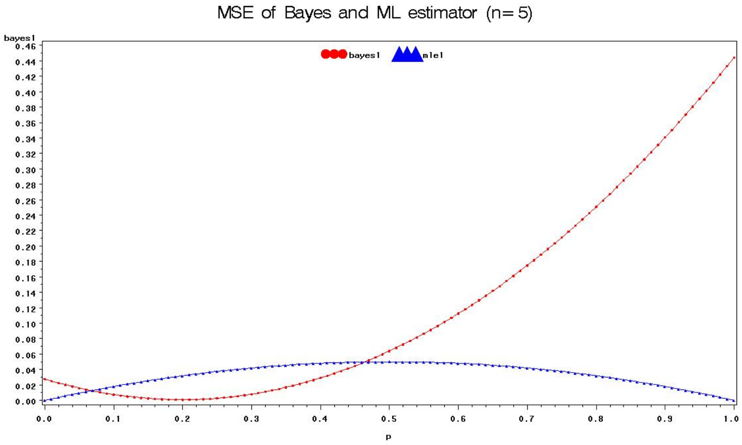
$$\hat{p}_B = \frac{n\bar{X} + \alpha}{\alpha + \beta + n}.$$

The MLE of p is $\hat{p} = \bar{X}$. Compare the MSE of two estimators.

e.g. α=5, β=20

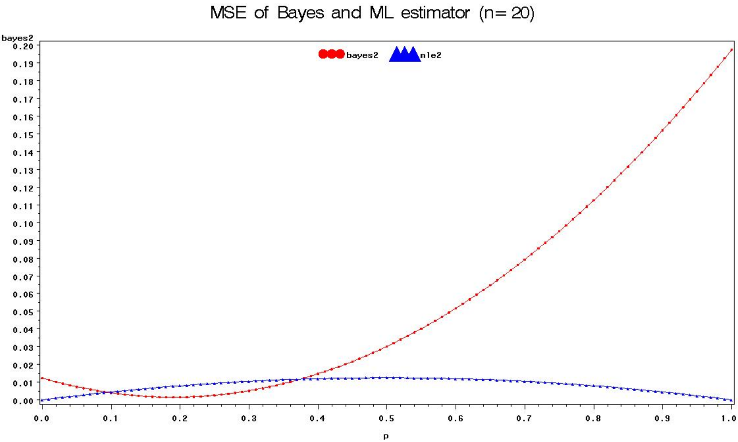
Ch 7. Point estimation

Mean Squared Error: MSE



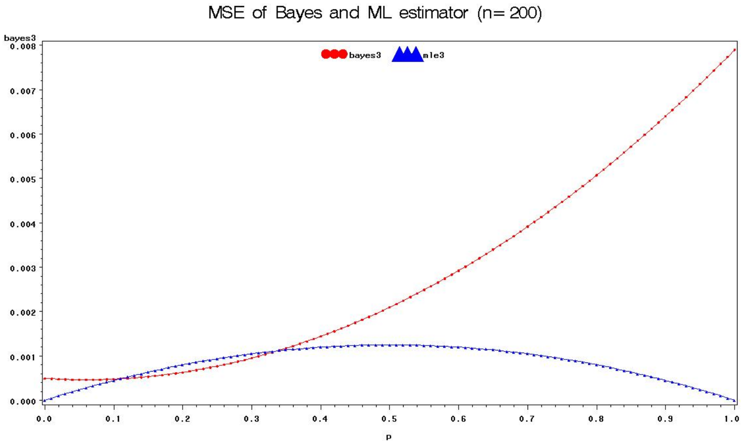
Ch 7. Point estimation

Mean Squared Error: MSE



Ch 7. Point estimation

Mean Squared Error: MSE



Definition

Let $f(\mathbf{x}|\theta) = f(x_1, \cdots, x_n|\theta)$ be the pdf/pmf of $X_1, \cdots, X_n$. An estimator W is said to be *Uniform Minimum Variance Unbiased Estimator* (UMVUE) for $\tau(\theta)$ if

1. $E_\theta(W) = \tau(\theta)$
2. $Var_\theta(W) < \infty$
3. For any other UE of $\tau(\theta)$, say $\tilde{W}$,

$$Var_\theta(W) \leq Var_\theta(\tilde{W}) \text{ for all } \theta \in \Theta$$

$\triangleleft$ Note: UMVUE may not exist. If it does, it is essentially unique.

Then, how to get the UMVUE ?

1. Using Cramér-Rao Lower Bound (CRLB)
2. Using complete and sufficient statistic and Rao-Blackwell, Lehmann-Scheffe Theorem.

- Idea of using CRLB;
  Show that for any UE, $\tilde{W}$, of $\tau(\theta)$,

$$Var_\theta(\tilde{W}) \geq c(\theta) \quad \text{for all} \quad \theta \in \Theta$$

and if we can find an UE, W, such that

$$Var_\theta(W) = c(\theta) \quad \text{for all} \quad \theta \in \Theta$$

then we can conclude W is the UMVUE of $\tau(\theta)$.

Theorem

Let $f(\mathbf{x}|\theta)$ be the pdf/pmf of $X_1, \cdots, X_n$. Assume

1. *Θ is an open space(subset) of R.*
2. *$\{\mathbf{x} : f(\mathbf{x}|\theta) > 0\}$ does not depend on θ.*
3. *$\partial f(\mathbf{x}|\theta)/\partial\theta$ exist on Θ*
4. *For any estimator $\tilde{W}$ with $E_\theta \tilde{W}^2 < \infty$, for all $\theta \in \Theta$, we have*

$$\frac{\partial}{\partial\theta} E_\theta \tilde{W} = \begin{cases} \int \ \tilde{W} \ [\frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)] d\mathbf{x} \\ \sum \ \tilde{W} \ [\frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)] \end{cases}$$

5.

$$E_\theta \left[\left(\frac{\partial \ln f(\mathbf{x}|\theta)}{\partial\theta} \right)^2 \right] < \infty$$

Theorem (Continued)

Then for any UE of a differentiable parametric function $\tau(\theta)$,

$$Var_\theta(W) \geq \frac{\left[\tau(\theta)'\right]^2}{E_\theta\left[\left(\frac{\partial \ln f(\mathbf{x}|\theta)}{\partial \theta}\right)^2\right]}.$$

◁ Note:

- The five conditions are called CR regularity conditions.
- $$I_n(\theta) = E_\theta\left[\left(\frac{\partial \ln f(\mathbf{x}|\theta)}{\partial \theta}\right)^2\right]$$

  is called the information number or fisher information of the sample.

Under regularity conditions,

$$1 = \int_{-\infty}^{\infty} f(x;\theta)dx = \int_S f(x;\theta)dx \Rightarrow 0 = \int_S \frac{\partial f(x;\theta)}{\partial \theta} dx.$$

$$\Rightarrow 0 = \int_S \frac{\partial f(x;\theta)/\partial \theta}{f(x;\theta)} f(x;\theta)dx,$$

$$\Rightarrow \quad 0 = \int_S \frac{\partial \ln f(x;\theta)}{\partial \theta} f(x;\theta)dx \qquad (1)$$

$$\Rightarrow E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right] = 0; \qquad (2)$$

We differentiate (1) again,

$$0=\int_S \frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2} f(x;\theta)dx+\int_S \frac{\partial \ln f(x;\theta)}{\partial\theta}\frac{\partial \ln f(x;\theta)}{\partial\theta} f(x;\theta)dx.$$

Let $$I(\theta)=E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right], \qquad (3)$$

we call $I(\theta)$ **Fisher information**.

$$I(\theta)=-\int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2} f(x;\theta)dx. \qquad (4)$$

Fisher information $I(\theta)$ is also

$$I(\theta) = Var\left(\frac{\partial \ln f(X;\theta)}{\partial \theta}\right).$$

$$\boxed{I(\theta) = E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial \theta}\right)^2\right] = E\left[-\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right]}$$

We call $\left(\frac{\partial \ln f(X;\theta)}{\partial \theta}\right)$ the score function.

Proof: The proof is an application of the Cauchy-Schwartz inequality, which says that $\left[Cov(Y,Z)\right]^2 \le (VarY)(VarZ)$ or

$$\frac{\left[Cov(Y,Z)\right]^2}{VarZ} \le VarY.$$

We let $Y \equiv W(\mathbf{X})$ and $Z \equiv \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)$.

First we show that $\frac{d}{d\theta}E_\theta W(\mathbf{X}) = E_\theta\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right]$, which is a covariance from (2)

$$E_\theta\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right] = 0.$$

- When
$$\frac{\partial^2 \ln f(\mathbf{x}|\theta)}{\partial \theta^2}$$
exist, then
$$I_n(\theta) = -E_\theta\left[\left(\frac{\partial^2 \ln f(\mathbf{x}|\theta)}{\partial \theta^2}\right)\right]$$
- The five conditions are usually satisfied with exponential family.
- If $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\theta)$ then
$$Var_\theta(W) \geq \frac{\left[\tau(\theta)'\right]^2}{nE_\theta\left[\left(\frac{\partial \ln f(x|\theta)}{\partial \theta}\right)^2\right]}.$$

Here

$$
\begin{aligned}
I_n(\theta) &= nE_\theta\left[\left(\frac{\partial \ln f(x|\theta)}{\partial\theta}\right)^2\right] \\
&= -nE_\theta\left[\left(\frac{\partial^2 \ln f(x|\theta)}{\partial\theta^2}\right)\right] \\
&\doteq nI_1(\theta)
\end{aligned}
$$

Proof:

$$
I_n(\theta) = Var\left(\frac{\partial \ln f(\theta;\mathbf{X})}{\partial\theta}\right) = Var\left(\sum_{i=1}^{n}\frac{\partial \ln f(X_i;\theta)}{\partial\theta}\right) = nI(\theta).
$$

▷ Example: $X_1, \cdots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, σ^2 in known. Find the CRLB and the UMVUE.

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\{-\frac{(x-\theta)^2}{2\sigma_0^2}\}.$$

$$\ln f(x, \theta) = -\ln(\sqrt{2\pi}\sigma_0) - \frac{(x-\theta)^2}{2\sigma_0^2}.$$

$$\frac{\partial \ln f(x, \theta)}{\partial \theta} = \frac{x-\theta}{\sigma_0^2}, \quad \frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2} = -\frac{1}{\sigma_0^2}.$$

$$I(\theta) = \frac{1}{\sigma_0^2}.$$

CRLB=?, UMVUE=?

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Poisson(λ). Find the CRLB and the UMVUE.

$$I_n(\lambda) = nI(\lambda) = -nE_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{e^{-\lambda}\lambda^X}{X!}\right)\right]$$

$$= -nE_\lambda\left[\frac{\partial^2}{\partial\lambda^2}(-\lambda + X\log\lambda - \log X!)\right] = \frac{n}{\lambda}.$$

CRLB=?, UMVUE=?

Ch 7. Point estimation

UMVUE: CRLB

Theorem

$X_1, \cdots, X_n \stackrel{iid}{\sim} f(x|\theta)$. Let W be an unbiased estimator of $\tau(\theta)$. Then $Var_\theta(W)$ attains the CRLB if and only if

$$a(\theta)\,[W - \tau(\theta)] = \frac{\partial}{\partial \theta} \ln[f(x_1, \cdots, x_n|\theta)]$$

for some function $a(\theta)$.

▷ Example: $X_1, \cdots, X_n \stackrel{iid}{\sim} N(\mu, \ ^2)$. UMVUE of μ, μ^2?

▷ Example: $X_1, \cdots, X_n \stackrel{iid}{\sim}$ Uniform$(0, \theta)$.

The Cramer-Rao inequality is not applicable to this pdf.

(See the Leibniz rule, Theorem 2.4.1, Page 69)

Theorem (Rao-Blackwell)

Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\phi(T) = E_\theta(W|T)$. Then

$$E_\theta[\phi(T)] = \tau(\theta)$$

and

$$Var_\theta[\phi(T)] \leq Var_\theta[W]$$

for all θ. That is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$ than W.

Proof: $E[Y] = E[E(Y|X)]$

$$\text{var}[Y] = \text{var}[E(Y|X)] + E[Y - E(Y|X)]^2$$

- Properties of the best unbiased estimator or UMVUE

  If W is a best estimator of $\tau(\theta)$, then W is unique.

Proof: Suppose W' is another best unbiased estimator, and consider the estimator $W^* = \frac{1}{2}(W + W')$. Note that $E_\theta W^* = \tau(\theta)$ and

$$
\begin{aligned}
\operatorname{Var}_\theta W^* &= \operatorname{Var}_\theta \left(\frac{1}{2}W + \frac{1}{2}W'\right) \\
&= \frac{1}{4}\operatorname{Var}_\theta W + \frac{1}{4}\operatorname{Var}_\theta W' + \frac{1}{2}\operatorname{Cov}_\theta(W, W') \\
&\le \frac{1}{4}\operatorname{Var}_\theta W + \frac{1}{4}\operatorname{Var}_\theta W' + \frac{1}{2}[(\operatorname{Var}_\theta W)(\operatorname{Var}_\theta W')]^{1/2} \\
&= \operatorname{Var}_\theta W.
\end{aligned}
$$

we can have equality only if $W' = a(\theta)W + b(\theta)$.

$$E_\theta W' = E_\theta W \qquad \operatorname{Var}_\theta W = \operatorname{Var}_\theta W'$$

Hence $a(\theta) = 1$, $b(\theta) = 0$ and $W = W'$, showing that W is unique.

Theorem (Lehmann-Scheffe)

Let $X_1, \cdots, X_n$ have joint pmf/pdf $f(\mathbf{x} : \theta)$, $\theta \in \Theta$. Suppose T is a complete and sufficient statistic. If $\phi(T)$ is an unbiased estimator of $\tau(\theta)$ and it is a function of T only then $\phi(T)$ is the UMVUE of $\tau(\theta)$.

$\triangleleft$ Note:

1. If we can find an unbiased estimator $\phi(T)$ of $\tau(\theta)$ which is a function of CSS T only then it is the UMVUE
2. For any unbiased estimator of $\tau(\theta)$, W, $E(W|T)$ is the UMVUE of $\tau(\theta)$.

Example

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N\left(\mu, \sigma^2\right)$, where both μ and σ^2 are unknown.

Then $\left(\sum X_i, \sum X_i^2\right)$ is a complete sufficient statistic (and thus so is $\left(\bar{X}, S^2\right)$).

Since S^2 is unbiased for σ^2, and is a function of the complete sufficient statistic, it is UMVUE.

Example

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $Unif(0,\theta)$. $Y = X_{(n)}$ is a complete sufficient statistic.

The density of Y is

$$f_Y(y) = \frac{ny^{n-1}}{\theta^n},\ 0 < y < \theta.$$

So $EY = \frac{n}{n+1}\theta$. Hence $\frac{n+1}{n}X_{(n)}$ is an unbiased estimator of θ. Since $X_{(n)}$ is complete and sufficient, this is the UMVUE.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Binomial(k, θ). Find the UMVUE of θ and $k\theta(1-\theta)^{k\text{-}1}$.

$$\tau(\theta) = P(X=1) = k\theta(1-\theta)^{k-1}$$

The CSS $\sum_{i=1}^{n} X_i \sim Binomial(kn, \theta)$.

An UE $I(X_1 = 1) = \begin{cases} 1 & \text{if } X_1 = 1 \\ 0 & \text{otherwise} \end{cases}$.

$$\phi\left(\sum X_i\right) = E\left[I(X_1 = 1) \middle| \sum X_i\right] = \ldots\ldots = \frac{k\dbinom{(n-1)k}{\left(\sum_{i=1}^{n} X_i\right)-1}}{\dbinom{nk}{\sum_{i=1}^{n} X_i}}$$

$$\phi\left(\sum X_i\right) = E\left[I\left(X_1 = 1\right) \middle| \sum X_i = t\right]$$

$$= P\left[\left(X_1 = 1 \middle| \sum X_i = t\right)\right]$$

$$= \frac{P\left[X_1 = 1 \text{ and } \sum X_i = t\right]}{P\left[\sum X_i = t\right]}$$

$$= \frac{P\left[X_1 = 1 \text{ and } \sum_{i=2}^{n} X_i = t - 1\right]}{P\left[\sum_{i=1}^{n} X_i = t\right]}$$

$$= \frac{P\left[X_1 = 1\right] P\left[\sum_{i=2}^{n} X_i = t - 1\right]}{P\left[\sum_{i=1}^{n} X_i = t\right]}$$

$$= \frac{\left[k\theta(1-\theta)^{k-1}\right]\left[\binom{(n-1)k}{t-1}\theta^{t-1}(1-\theta)^{(n-1)k-(t-1)}\right]}{\binom{nk}{t}\theta^{t}(1-\theta)^{nk-t}}$$

$$= \frac{k\binom{(n-1)k}{t-1}}{\binom{nk}{t}} = \frac{k\binom{(n-1)k}{\left(\sum_{i=1}^{n} X_i\right)-1}}{\binom{nk}{\sum_{i=1}^{n} X_i}}$$

is UMVUE.

Ch 7. Point estimation

Decision Theory: Loss function optimality

- Data: $\mathbf{X}$
- Model(Distribution): $f(\mathbf{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$
- Action space: $\mathcal{A}$
  Point estimation: $\mathcal{A} = \Theta$
  Testing: $\mathcal{A}$={Reject H_0, Accept H_0}
- Loss function: $L(\boldsymbol{\theta}, a)$
- Decision rule: $\delta(\mathbf{x})$: Sample space $\rightarrow \mathcal{A}$
- Risk function: Expected loss

$$R(\boldsymbol{\theta}, \delta) = E\left[L(\boldsymbol{\theta}, \delta(\mathbf{X}))\right] = \int L(\boldsymbol{\theta}, \delta(\mathbf{x})) f(\mathbf{x}|\boldsymbol{\theta}) dx$$

- Goal: Find $\delta(\mathbf{x})$ that has small risk somehow.

Definition

A real valued function $L(\boldsymbol{\theta}, a)$ satisfying

1. $L(\boldsymbol{\theta}, a) \geq 0$ for all θ, a
2. $L(\boldsymbol{\theta}, a) = 0$ for $a = \theta$

is called a *loss function* of the action a.

Definition

Let $\delta(\mathbf{X})$ be an estimator of a parametric function $\tau(\boldsymbol{\theta})$. Then

$$R(\boldsymbol{\theta}, \delta) = E\left[L(\boldsymbol{\theta}, \delta(\mathbf{X}))\right]$$

is called the *risk function* of $\delta(\mathbf{X})$ in estimating $\tau(\boldsymbol{\theta})$.

▷ Example:

1. Squared error loss

$$L(\theta, \delta(\mathbf{X})) = (\delta(\mathbf{X}) - \theta)^2,\ R(\boldsymbol{\theta}, \delta) = E\left[(\delta(\mathbf{X}) - \theta)^2\right]$$

2. Absolute error loss

$$L(\theta, \delta(\mathbf{X})) = |\delta(\mathbf{X}) - \theta|,\ R(\boldsymbol{\theta}, \delta) = E\left[|\delta(\mathbf{X}) - \theta|\right]$$

3. Stein's loss

$$L(\theta, \delta(\mathbf{X})) = \frac{\delta(\mathbf{X})}{\theta} - 1 - \ln\left(\frac{\delta(\mathbf{X})}{\theta}\right)$$

Definition

1. An estimator $\delta_1(\mathbf{X})$ is said to be at least *as good as* another estimator $\delta_2(\mathbf{X})$ if

$$R(\theta, \delta_1(\mathbf{X})) \leq R(\theta, \delta_2(\mathbf{X}))$$

   for all $\theta \in \Theta$.

2. An estimator $\delta_1(\mathbf{X})$ is *better than* $\delta_2(\mathbf{X})$ if

$$R(\theta, \delta_1(\mathbf{X})) \leq R(\theta, \delta_2(\mathbf{X}))$$

   for all $\theta \in \Theta$ and

$$R(\theta, \delta_1(\mathbf{X})) < R(\theta, \delta_2(\mathbf{X}))$$

   for at least one $\theta \in \Theta$.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Bernoulli(p).
$L(\theta, \delta(\mathbf{X})) = (\delta(\mathbf{X}) - \theta)^2$. $\delta_1(\mathbf{X}) = \bar{X}$,

$$\delta_2(\mathbf{X}) = \frac{n\bar{X} + \sqrt{n/4}}{n + \sqrt{n}}.$$

Figure 7.3.1 (page 333)

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, 1)$. $L(\theta, \delta(\mathbf{X})) = (\delta(\mathbf{X}) - \theta)^2$.
$\delta_1(\mathbf{X}) = c$, $\delta_2(\mathbf{X}) = \bar{X}$.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Estimation of σ^2 using different loss function. We restrict our estimator of the form $\delta_b(\mathbf{X}) = bS^2$. The loss function considered are squared error loss, Stein's loss.

Let $Y = \sum_{i=1} (X_i - \bar{X})^2$,

Since $Y/\sigma^2 \sim \chi^2(n-1)$, $E(Y) = (n-1)\sigma^2$, $\mathrm{var}(Y) = 2(n-1)\sigma^4$

$$E[(cY - \sigma^2)^2] = \mathrm{var}(cY) + [E(cY) - \sigma^2]^2$$

$$= 2c^2(n-1)\sigma^4 + [c(n-1)-1]^2\sigma^4 = [(n^2-1)c^2 - 2(n-1)c + 1]\sigma^4$$

$$\arg\min_{c} E[(cY - \sigma^2)^2] = 1/(n+1)$$

- In general, there does not exist an estimator $\delta(\mathbf{X})$ such that for any other estimator $\tilde{\delta}(\mathbf{X})$ we have

$$R(\theta, \delta(\mathbf{X})) \leq R(\theta, \tilde{\delta}_2(\mathbf{X}))$$

  for all $\theta \in \Theta$.
- To define the best estimator w.r.t. the given loss function, we can proceed two ways.
  1. Restrict attention to a smaller class of estimators such as unbiased estimators or linear estimators
  2. Define a criterion for comparing estimators such minimax or Bayes rule

Definition

An estimator $\delta(\mathbf{X})$ is called a *minimax estimator* if

$$\max_{\theta\in\Theta} R(\theta, \delta(\mathbf{X})) \leq \max_{\theta\in\Theta} R(\theta, \tilde{\delta}(\mathbf{X}))$$

for all other estimator $\tilde{\delta}(\mathbf{X})$.

Definition

The *Bayes risk* of an estimator $\delta(\mathbf{X})$ w.r.t. prior distribution $\pi(\theta)$ is defined as

$$B(\pi, \delta(\mathbf{X})) = E_\pi \left[R(\theta, \delta(\mathbf{X}))\right] = \int R(\theta, \delta(\mathbf{X}))\pi(\theta) d\theta$$

Definition

An estimator $\delta(\mathbf{X})^\pi$ is said to be a *Bayes estimator* w.r.t. prior distribution $\pi(\theta)$ if it minimizes Bayes risk over all estimators. That is

$$B(\pi, \delta(\mathbf{X})^\pi) = \inf_{\tilde{\delta}} B(\pi, \tilde{\delta}(\mathbf{X}))$$

Note that

$$B(\pi, \delta(\mathbf{X})^\pi = \int_\Theta R(\theta, \delta)\pi(\theta)\, d\theta = \int_\Theta \left(\int_{\mathcal{X}} L(\theta, \delta(\mathbf{x})) f(\mathbf{x}|\theta)\, d\mathbf{x} \right) \pi(\theta)\, d\theta.$$

$$= \int_{\mathcal{X}} \left[\int_\Theta L(\theta, \delta(\mathbf{x}))\pi(\theta|\mathbf{x})\, d\theta \right] m(\mathbf{x})\, d\mathbf{x}.$$

Theorem

Consider a point estimation problem for a real-valued parameter θ. The Bayes estimator is $E(\theta|\mathbf{X})$ for squared error loss and median of $\pi(\theta|\mathbf{X})$ for absolute error loss.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Bernoulli(p). $\theta \sim$ Uniform$(0, 1)$. Find the Bayes estimator with squared error loss function.

$$E(p|\mathbf{x}) = \frac{\sum x_i + \alpha}{n + \alpha + \beta}. \qquad \alpha = \beta = 1$$

Homework: p355~364

7.6, 7.7, 7.10, 7.24, 7.37, 7.44, 7.45(a)&(c), 7.48