STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 6A: Estimation

Estimation Overview

- Define $\hat{\theta}$ as an estimator of θ where $\hat{\theta}$ is a function of the data, y.
- **Point estimation**: single value/ best guess of θ.
- Suppose Y is the number of tails after observing a coin tossed 10 times. What might you use to estimate the coin's bias, $\theta \in [0, 1]$?
- Consider $\mathrm{E}[Y]$.
- Why?

Estimation in Classical Statistics

- In classical statistics, we identify certain desirable properties of $\hat{\theta}$.
- **Sufficiency**. If $\hat{\theta} = T(Y)$, T is a sufficient statistic for θ: i.e., $f(y|t(y)) = f(y|t(y), \theta)$.
- **Unbiasedness**. $\hat{\theta}$ is an unbiased estimate of θ if $\mathrm{E}[\hat{\theta}|\theta] = \theta$.
  Ex: if $X \sim \mathrm{N}(\mu, \sigma^2)$ then $\mathrm{E}[\overline{X}] = \mu$.
- **Consistency**. $\hat{\theta}$ is a consistent estimator of θ if $\mathbb{P}(|\hat{\theta} - \theta| > \epsilon|\theta) \to 0$, as $n \to \infty, \ \forall \epsilon > 0$.
- **Efficiency**. $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $\mathrm{Var}(\hat{\theta}_1|\theta) < \mathrm{Var}(\hat{\theta}_2|\theta)$.
  Cramer Rao Inequality: $\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathrm{I}(\theta)}$ where $\mathrm{I}(\theta) = n\mathrm{E}\left[\left(\frac{\delta}{\delta\theta} \log f(x|\theta)\right)^2\right]$. Then, $e(\hat{\theta}) = \frac{\mathrm{I}(\theta)^{-1}}{\mathrm{Var}(\hat{\theta})}$.
- **Minimum MSE**. $\hat{\theta}$ has minimum MSE if $\arg\min_t \mathrm{E}[(t-\theta)^2|\theta] = \hat{\theta}$.

Example: Normal Process

- Let $\hat{\theta} = \overline{Y}$.
- $t = (\hat{\theta}, n)$ is sufficient for μ.
- $\mathrm{E}[\hat{\theta}|\mu] = \mu$, is unbiased.
- Since $\hat{\theta}$ is unbiased, and $\mathrm{Var}(\hat{\theta}|\mu) \to 0$ as $n \to \infty$, $\hat{\theta}$ is also consistent.
- It follows from unbiasedness + sufficiency that $\hat{\theta}$ also attains lowest variance among unbiased estimators (Lehman-Scheffe theorem).
- $\hat{\theta}$ attains minimum MSE.

Example: Bernoulli Process

- Let $X \sim \mathrm{Ber}(p)$.
- $\overline{X}$ is unbiased.
- $\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}X_i = \frac{1}{n}np = p$

Methods for Identifying Estimators

- Method of Moments
- Method of Maximum Likelihood

Method of Moments Example

Example

$X_1, ..., X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Find $\hat{\mu}$ and $\hat{\sigma}$.

Solution

$$\mathrm{E}X = \mu = \overline{X} \rightarrow \hat{\mu} = \overline{X}$$

$$\mathrm{E}X^2 = [\mathrm{E}X]^2 + \mathrm{Var}(X) = \mu^2 + \sigma^2 \rightarrow \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

MLE Example

Example

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\theta)$. Find $\hat{\theta}$.

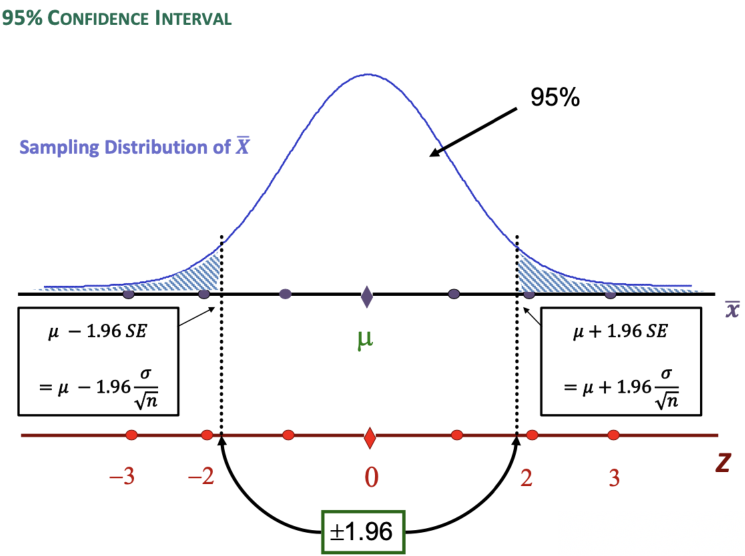
Solution

$$\begin{aligned}
\ell(\theta|\boldsymbol{x}) &= \prod_{i=1}^{n} \left[\frac{1}{\theta} e^{-x_i/\theta}\right] \\
&= \frac{1}{\theta^n} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\theta}\right) \\
&\to \log \ell(\theta|\boldsymbol{x}) = \log\left[\frac{1}{\theta^n} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\theta}\right)\right] \\
&= -n \log\theta - \frac{\sum_{i=1}^{n} x_i}{\theta} \\
&\to \frac{\partial}{\partial\theta} \log \ell(\theta|\boldsymbol{x}) = \frac{\sum_{i=1}^{n} x_i}{\theta^2} - \frac{n}{\theta} \\
&\to \sum_{i-1}^{n} x_i = n\theta \to \hat{\theta} = \overline{x}
\end{aligned}$$

Classical Interval Estimation

- Range of values for θ with a given confidence coefficient γ.
- $\gamma = \mathrm{P}(x \in C_\gamma(\theta))$ where $C_\gamma(\theta)$ is based on the sampling distribution.
- Usually we pick $C_\gamma(\theta)$ so as to cut-off $\frac{1-\gamma}{2}$ probability on both ends of the sampling distribution.
- Normal process ex: $\mu - \alpha \frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + \alpha \frac{\sigma}{\sqrt{n}}$

Normal Process Example



Normal Process Example

- Next we manipulate $C_\gamma(\theta)$ to get $C_\gamma(x)$, because we want to make a statement about θ based on the sampling distribution of the data.

$$\mu - \alpha\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + \alpha\frac{\sigma}{\sqrt{n}}$$
$$= -\alpha\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < \alpha\frac{\sigma}{\sqrt{n}}$$
$$= \overline{X} - \alpha\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + \alpha\frac{\sigma}{\sqrt{n}}$$

$\gamma = 0.95, \alpha = 1.96 \rightarrow \overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}$

- It is important to keep in mind that in the classical interval statement, X is the random variable!

Bayesian Approach

- On the Bayesian approach, the primary inferential statement about θ is the posterior distribution of θ.
- In a sense then, this class and the next class are not building the Bayesian approach further. The development is already done.
- However, we will look at tools and approaches for using the posterior distribution in order to construct point and interval estimates and hypothesis tests. And we will examine their plausibility.

Point Estimates

- If we had to make a point estimate on the Bayesian approach, what might you use?
- Consider the posterior mean,

$$\mathrm{E}[\theta|x] = \int \theta\pi(\theta|x)d\theta.$$

- Why or why not use the mean? What other suggestions?

Bayes Estimator

Bayes Estimator

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

$$\mathrm{E}[\theta|\boldsymbol{x}] = \int_{\Omega} \theta\pi(\theta|\boldsymbol{x})d\theta$$

- For example, in the case of a bernoulli process with a beta prior,

$$\begin{aligned}\hat{\theta} &= \frac{\sum_{i=1}^{n} x_i + \alpha}{\alpha+\beta+n} \\ &= \frac{\sum_{i=1}^{n} x_i}{n}\frac{n}{\alpha+\beta+n} + \frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{\alpha+\beta+n} \\ &= \overline{x}k + \mathrm{E}[\,\theta\,](1-k)\end{aligned}$$

- The point estimate is a weighted combination of the prior mean and the sample mean where the weighting depends on the strength of the prior.
- Bias: $\mathrm{E}[\mathrm{E}[\theta|x]] = \frac{n\theta+\alpha}{\alpha+\beta+n} \neq \theta$ unless $\mathrm{E}[\theta] = \theta$.
- Bias $\to 0$ as $n \to \infty$.

Estimation as a Decision Problem

- Consider $L(\hat{\theta}, \theta)$.
- This is the loss incurred by guessing that θ is $\hat{\theta}$.
- If we think of this quantity analogous to a utility function, for someone whose goal is to be as accurate as possible, than a natural decision rule is to minimize posterior expected loss after observing some data and identifying a posterior distribution.

Posterior Expected Loss

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

Minimizing Expected Loss: Part One

- Let $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. This is analogous to what we previously called the quadratic scoring rule, but now it is used to evaluate continuous point estimates.
- We want to minimize posterior expected loss, given by

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

$$\begin{aligned}
\mathrm{E}[(\hat{\theta} - \theta)^2 | \boldsymbol{X} = \boldsymbol{x}] &= \mathrm{E}\big[(\theta - \mathrm{E}[\theta|\boldsymbol{X}] + \mathrm{E}[\theta|\boldsymbol{X}] - \hat{\theta})^2 | \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \mathrm{E}\big[(\theta - \mathrm{E}[\theta|\boldsymbol{X}])^2 | \boldsymbol{X} = \boldsymbol{x}\big] + \mathrm{E}[(\mathrm{E}[\theta|\boldsymbol{X}] - \hat{\theta})^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathrm{E}\big[(\theta - \mathrm{E}[\theta|\boldsymbol{X}])^2 | \boldsymbol{X} = \boldsymbol{x}\big] + \big[\mathrm{E}[\theta|\boldsymbol{x}] - \hat{\theta}\big]^2 \\
&= \mathrm{Var}(\theta) + (\text{penalty})
\end{aligned}$$

- This is minimized when $\hat{\theta} = \mathrm{E}[\theta|\boldsymbol{x}]$, which minimizes the penalty. The remainder is irreducible expected loss due to our uncertainty about θ.

Minimizing Expected Loss: Part Two

- Let $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$. This is analogous to the absolute value score, which we said was not strictly proper.
- First, we write the expected loss, as follows.

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

Minimizing Expected Loss: Part Two

- Next we find the derivative of the loss function as follows.

$$
\begin{aligned}
\frac{\partial}{\partial \hat{\theta}} \mathrm{E}[L(\theta, \hat{\theta})] &= \frac{\partial}{\partial \hat{\theta}} \left(\int_{-\infty}^{\hat{\theta}} -(\theta - \hat{\theta}) \pi(\theta | \boldsymbol{x}) d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) \pi(\theta | \boldsymbol{x}) d\theta \right) \\
&= \frac{\partial}{\partial \hat{\theta}} \int_{-\infty}^{\hat{\theta}} -(\theta - \hat{\theta}) \pi(\theta | \boldsymbol{x}) d\theta + \frac{\partial}{\partial \hat{\theta}} \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) \pi(\theta | \boldsymbol{x}) d\theta \\
&= -(\hat{\theta} - \hat{\theta}) \pi(\hat{\theta} | \boldsymbol{x}) + \int_{-\infty}^{\hat{\theta}} \pi(\theta | \boldsymbol{x}) d\theta \\
&\quad - (\hat{\theta} - \hat{\theta}) \pi(\hat{\theta} | \boldsymbol{x}) - \int_{\hat{\theta}}^{\infty} \pi(\theta | \boldsymbol{x}) d\theta \\
&= F_{\theta | \boldsymbol{x}}(\hat{\theta}) - [1 - F_{\theta | \boldsymbol{x}}(\hat{\theta})]
\end{aligned}
$$

- Where we use Leibniz's rule for differentiation under the integral for the last line, given by

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

Minimizing Expected Loss: Part Two

- Finally, we identify the FOC.

$$\frac{\partial}{\partial\hat{\theta}}\mathrm{E}[L(\theta,\hat{\theta})] = -(\hat{\theta}-\hat{\theta})\pi(\hat{\theta}|\boldsymbol{x}) + \int_{-\infty}^{\hat{\theta}} \pi(\theta|\boldsymbol{x})d\theta$$
$$-(\hat{\theta}-\hat{\theta})\pi(\hat{\theta}|\boldsymbol{x}) - \int_{\hat{\theta}}^{\infty} \pi(\theta|\boldsymbol{x})d\theta = 0$$
$$\rightarrow \int_{-\infty}^{\hat{\theta}} \pi(\theta|\boldsymbol{x})d\theta = \int_{\hat{\theta}}^{\infty} \pi(\theta|\boldsymbol{x})d\theta$$
$$\rightarrow F_{\theta|\boldsymbol{x}}(\hat{\theta}) = 1 - F_{\theta|\boldsymbol{x}}(\hat{\theta})$$

- Therefore, $\hat{\theta}$ is the posterior median.

Summary

Squared error loss

The posterior expected loss

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

is minimized when $\hat{\theta}=\mathrm{E}[\theta|\boldsymbol{x}]$

Absolute value loss

The posterior expected loss

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

is minimized when $\hat{\theta}=$ the posterior median of $\pi(\theta|\boldsymbol{x})$.

Normal Process Example

Example

Let $X_1, \ldots X_n \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$ and suppose that the prior distribution of θ is $\mathrm{N}(\mu, \tau^2)$ where θ is the only unknown. What is the Bayes estimator under squared error loss and absolute error loss?

Solution

We know that for a normal process / normal prior with an unknown mean the posterior $\theta|\boldsymbol{X}$ is normal with

$$\mathrm{E}[\theta|\boldsymbol{x}] = \frac{\tau^2}{\tau^2 + \frac{1}{n}\sigma^2}\ \overline{\boldsymbol{x}} + \frac{\frac{1}{n}\sigma^2}{\tau^2 + \frac{1}{n}\sigma^2}\ \mu$$

$$\mathrm{Var}(\theta|\boldsymbol{X}) = \frac{\frac{1}{n}\sigma^2\tau^2}{\tau^2 + \frac{1}{n}\sigma^2}$$

Since mean = median for a normal distribution, $\hat{\theta} = \mathrm{E}[\theta|\boldsymbol{x}]$.

Credible Intervals

Credible Interval

A $(1-\alpha)100\%$ credible interval for θ is (a, b) such that,

$$P(a < \theta < b) = \int_a^b \pi(\theta|\boldsymbol{x})d\theta = 1 - \alpha$$

- Compare this to frequentist confidence interval.
- Note that this is equivalent to $F(b) - F(a)$ where

$$F(t) = \int_\infty^t \pi(\theta|\boldsymbol{x}).$$

- This makes computation with R particularly easy.

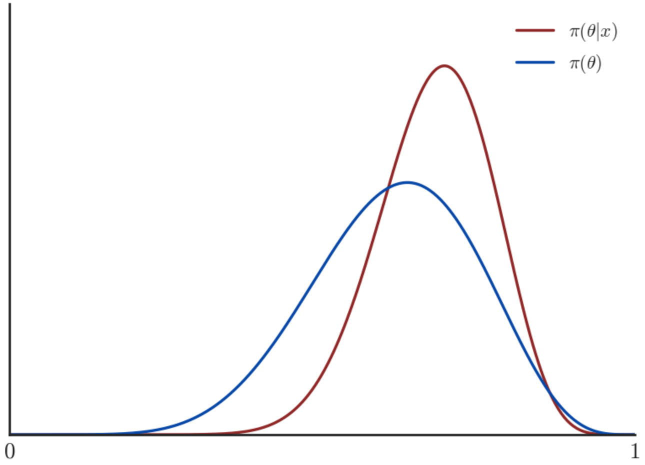
Beta Example

- Suppose we start with a $\mathrm{Beta}(8, 5)$ prior for the bias of a coin, θ.
- Observe 9 heads and 3 tails where heads corresponds to $X = 0$. What is the posterior?
- Posterior is $\mathrm{Beta}(17, 8)$.
- What is a reasonable point estimate?
- $\mathrm{E}[\theta|x] = 17/25 = 0.68$.
- Median is approximately

$$\frac{\alpha - \frac{1}{3}}{\alpha + \beta - \frac{2}{3}} = \frac{17 - \frac{1}{3}}{17 + 8 - \frac{2}{3}} = \frac{16.67}{24.34} = 0.685.$$

- Are these estimators biased? Are they asymptotically consistent?
- Yes (because of α and β) and Yes (the priors are little constants that drop out as $n \to \infty$).

Beta Example



Beta Example

- Now we can calculate a $(1-\alpha)100\%$ credible interval for θ. For example, a 95% credible interval for θ is,

$$P(a < \theta < b) = \int_a^b \pi(\theta|\boldsymbol{x}) d\theta = 0.95$$

- In our case, $a = 0.49$ and $b = 0.84$.
- R code: `qbeta(c(0.025,0.975),17,8)`
- There is really nothing special about the 95% any longer. We can set $P(a < \theta < b)$ to any $1-\alpha$.
- This gives us a probabilistic statement about any region around a point estimate.
- But now, the probabilistic statement is a statement about the parameter, not about the data, as it was in the classical approach.
- We are actually $1-\alpha$ confident that the true value of θ is in the interval.
- Not: the probability that the true θ would be captured by this interval construction procedure if we repeat the experiment many times is $(1-\alpha)100\%$.

Beta Example

- We can also easily compute the probability that θ is in any desired region of the posterior distribution.
- For example:

$$\begin{aligned}\Pr(0.4 < \theta < 0.6) &= \int_{0.4}^{0.6} \pi(\theta|\boldsymbol{x})d\theta \\ &= CDF(\theta|\boldsymbol{x})|_{\theta=0.6} - CDF(\theta|\boldsymbol{x})|_{\theta=0.4} \\ &= 0.19\end{aligned}$$

- R code: `pbeta(0.6, 17, 8) - pbeta(0.4, 17, 8)`.
- We are about 20% confident that θ is between 0.4 and 0.6.

Normal Process Example

- A 95% credible interval for θ:

$$\mathrm{E}[\theta|x] - 1.96SE < \theta < \mathrm{E}[\theta|x] + 1.96SE$$

- Suppose the posterior for θ is determined to be $\mathrm{N}(0.7, 0.1)$
- Find a 90% credible interval for θ.
- The 90% credible interval for θ is $(0.54, 0.86)$.
- R code: `qnorm(c(0.05,0.95),0.7,0.1)`.

Classical vs. Bayes