Estimation, and Decision Theory

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

02/01/2021

Point estimation

- Full posterior distribution $p(\theta \mid \mathbf{y})$ is nice...
  - But often wish to simplify inference and conclusions
- Possible *point estimates* for θ:
  - Posterior mode:
    $$\hat{\theta} = \text{argmax}_{\theta}\, p(\theta \mid \mathbf{y})$$
  - Posterior expectation:
    $$\hat{\theta} = E_{\theta \mid \mathbf{y}}\theta = \int \theta p(\theta \mid \mathbf{y})\, d\theta$$
  - Posterior median:
    $$P(\theta \leq \hat{\theta} \mid \mathbf{y}) = \int_{-\infty}^{\hat{\theta}} p(\theta \mid \mathbf{y})\, d\theta = \frac{1}{2}$$

Point estimation

- Posterior mode
  - Easy to compute because only need to work with numerator

$$p(\theta \mid \mathbf{y}) \propto p(\theta)p(\mathbf{y} \mid \theta)$$

  - Sometimes called generalized maximum likelihood estimation
- Posterior expectation uses full posterior
- Posterior median is more robust to outliers in $\mathbf{y}$ & posterior tails
- How to choose which estimate is optimal for a given application?

Loss function

- A *loss function* $l(\text{truth}, a)$ gives the loss incurred for an action a given the (usually unknown) truth.
  - Here "loss" is abstract - could be loss to society, loss in terms of model accuracy, etc.
  - want to minimize loss
- If the truth is given by parameter θ, the *posterior risk* is

$\rho(p_\theta, a)$

$$E_{\theta \mid \mathbf{y}} l(\theta \mid a) = \int l(\theta \mid a) p(\theta \mid \mathbf{y})\, d\theta$$

  - Averaging the loss over the posterior for θ
- For point estimation, our action a is given by an estimate $\hat{\theta}$: $l(\theta, \hat{\theta})$

Point estimate loss

- Squared error loss is commonly used

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

  - Posterior risk for squared error loss is minimized by posterior expectation:

$\rho(P_\theta, \hat{\theta})$

$$E_{\theta \mid \mathbf{y}}\, \theta = \text{argmin}_{\hat{\theta}}\, E_{\theta \mid \mathbf{y}} (\theta - \hat{\theta})^2$$

$$E(\theta - \hat{\theta})^2 = E(\underbrace{\theta - E\theta} + \underbrace{E\theta - \hat{\theta}})^2$$

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

$$= E(\theta - E\theta)^2 + 0 + (E\theta - \hat{\theta})^2 \geq E(\theta - E\theta)^2 = \rho(P_\theta, E\theta)$$

- Posterior risk for absolute loss $l(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ is minimized by the posterior median
  - Homework

Decision rules

- Denote the space of allowable actions $\mathcal{A}$ ($a \in \mathcal{A}$)
- Denote sample space (possible data observations) $\mathcal{Y}$ ($\mathbf{y} \in \mathcal{Y}$)
- A *decision rule* $d \in \mathcal{D} : \mathcal{Y} \to \mathcal{A}$ is a rule for determining an action based on data.
- Formal decision-theoretic framework:
  - *prior distribution:* $p(\theta)$, $\theta \in \Theta$
  - *sampling distribution:* $p(\mathbf{y} \mid \theta)$
  - *allowable actions:* $a \in \mathcal{A}$
  - *decision rules:* $d \in \mathcal{D}$: $\mathcal{Y} \to \mathcal{A}$
  - *loss function:* $l(\theta, a)$

Normal-normal point estimate

- Recall for $y_1, \ldots, y_n$ iid Normal(μ, σ^2) and $p(\mu) = \text{Normal}(\mu_0, \tau^2)$:

$$p(\mu \mid \mathbf{y}) = \text{Normal}\left(\frac{\sigma^2\mu_0 + n\tau^2\bar{y}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

- Consider decision for point estimate of μ:
  - *prior distribution:* $p(\mu) = \text{Normal}(\mu_0, \tau^2)$, $\mu \in \mathbb{R}$
  - *sampling distribution:* $y_1, \ldots, y_n$ iid Normal(μ, σ^2), $\mathbf{y} \in \mathbb{R}^n$
  - *allowable actions:* $a \in \mathcal{A} = \mathbb{R}$
  - *decision rule:* $d(\mathbf{y}) = \frac{\sigma^2\mu_0 + n\tau^2\bar{y}}{\sigma^2 + n\tau^2}$
  - *loss function:* $l(\mu, a) = (\mu - a)^2$ (or $l(\mu, a) = |\mu - a|$)

Example: Coke bottles (cont.)

- Recall:
  - Coke bottles are filled with calibration Normal$(12, 0.01)$
  - Given machine with calibration μ, bottles filled with Normal$(\mu, 0.05)$
  - For $n = 5$ and $\bar{y} = 11.88$ oz, $p(\mu \mid \mathbf{y}) = $ Normal$(11.94, 0.005)$
- Action is to estimate $\hat{\mu} = 11.94$

- Posterior risk under squared error loss is the posterior variance, 0.005

$$E_{\mu|y} L(\mu, \hat{\mu})$$

$$= \int (\mu - \hat{\mu})^2 p(\mu|y)\, d\mu$$

$$= Var_{\mu|y}(\mu)$$

Frequentist risk

- The *frequentist risk* of a decision rule d is

$$R(\theta, d) = E_{\mathbf{y} \mid \theta} l(\theta, d(\mathbf{y})) \\ = \int l(\theta, d(\mathbf{y})) p(\mathbf{y} \mid \theta) \, d\mathbf{y}$$

  - Loss averaged over $\mathbf{y}$, given θ.
  - Note: does not depend on prior $p(\theta)$

- A rule d is *inadmissible* if $\exists d^*$ with

$$R(\theta, d^*) \leq R(\theta, d) \; \forall \theta \in \Theta$$

  and $<$ for some $\theta \in \Theta$

  - Implies another rule is universally "better"
  - A rule that is not inadmissible is *admissible*
  - Admissible rules are not necessarily good

Example: Coke bottles (cont.)

- The (poor) rule $d(\mathbf{y}) = 5$ is admissible because it is unbeatable when $\mu = 5$! $R(\mu, 5) = (\mu - 5)^2$
- The rule $d(\mathbf{y}) = \bar{y}$ has frequentist risk

$$R(\mu, \bar{y}) = \text{Var}_{\mathbf{y}\,|\,\mu}\, \bar{y}$$

$$E_{y|\mu}(\mu - \bar{y})^2 = \frac{\sigma^2}{n}$$

$$= 0.05/5 = 0.01$$

  - Does not depend on μ

- Note: posterior mean has form $B\mu_0 + (1-B)\bar{y}$, where

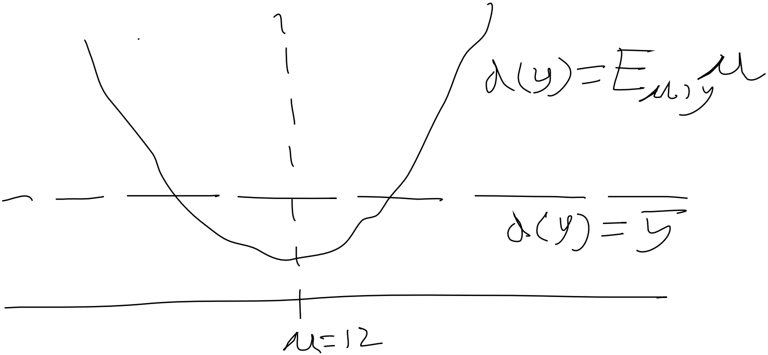
$$B = \frac{\sigma^2}{\sigma^2 + n\tau^2}.$$

  - For coke example with $n = 5$, $B = 0.5$

- The posterior mean rule $d(\mathbf{y}) = E_{\mu\,|\,\mathbf{y}}\,\mu$ has frequentist risk

$$\begin{aligned} R(\mu, E_{\mu\,|\,\mathbf{y}}\,\mu) &= B^2(\mu-\mu_0)^2 + (1-B)^2\mathrm{Var}_{\mathbf{y}\,|\,\mu}\,\bar{y} \\ &= 0.25(\mu-12)^2 + 0.0025 \end{aligned}$$

$$E_{y|\mu}(\mu - B\mu_0 - (1-B)\bar{y})^2$$
$$= E\left(B(\mu-\mu_0) + (1-B)(\mu-\bar{y})\right]^2$$
$$= B^2(\mu-\mu_0)^2 + (1-B)^2 E(\mu-\bar{y})^2$$
$$+ 0$$


$d(y) = E_{u,y} u$
$d(y) = 5$
$u=12$

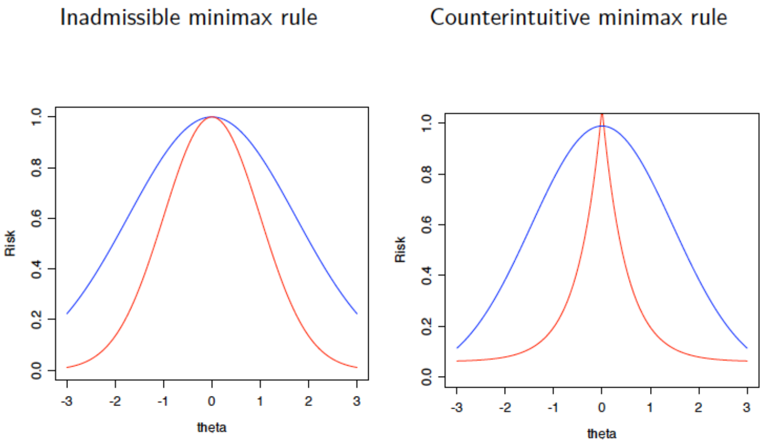
Minimax rules

- A *minimax rule d* satisfies

$$\sup_{\theta \in \Theta} R(\theta, d) \le \sup_{\theta \in \Theta} R(\theta, d^*), \ \forall d^* \in \mathcal{D}.$$

- Chose d to minimize maximum risk.
  - Prepare for "the worst case scenario"
- May not be unique, admissible, or favorable.

Minimax rules

- 

Credit: Victor Panaretos

Bayes risk

- The *Bayes risk* for given decision rule, for prior p_θ, is

$$r(p_\theta, d) = E_\theta E_{\mathbf{y}\,|\,\theta} l(\theta, d(\mathbf{y})) = \int R(\theta, d) p(\theta) d\theta$$

- Equivalently,

$$r(p_\theta, d) = E_{\mathbf{y}} E_{\theta\,|\,\mathbf{y}} l(\theta, d(\mathbf{y})) = \int \rho(p_\theta, d(\mathbf{y})) p(\mathbf{y}) d\mathbf{y}$$

  where $p(\mathbf{y})$ is the marginal distribution of $\mathbf{y}$.

- Also called "preposterior risk"
  - The expected risk before obtaining any data

Risks galore!

- *Loss function* $l(\theta, d(\mathbf{y}))$
  - Function of θ and $\mathbf{y}$
- *Posterior risk* $\rho(p_\theta, d(\mathbf{y}))$
  - Function of $\mathbf{y}$, averaged over θ
- *Frequentist risk* $R(\theta, d)$
  - Function of θ, averaged over $\mathbf{y}$
- *Bayes risk* $r(p_\theta, d)$
  - Averaged over both $\mathbf{y}$ and θ

- A *Bayes decision rule* minimizes Bayes risk:

$$\underset{d \in \mathcal{D}}{\operatorname{argmin}}\; r(p_\theta, d)$$

- Bayes rules are generally admissable
  - If a Bayes rule is unique, it is admissible.

If d is inadmissable,

$\exists\, d^*$ s.t. $R(\theta, d^*) \le R(\theta, d) \;\; \forall \theta$

$\rightarrow R(\theta, d^*)p(\theta) \le R(\theta, d)p(\theta)$

$\rightarrow \int \checkmark \, d\theta \le \int \checkmark \, d\theta$

$\rightarrow r(p_\theta, d^*) \le r(p_\theta, d)$, $\therefore d$ not a unique BR

Normal-normal model

- For the normal-normal model with $l(\mu, \hat{\mu}) = (\mu - \hat{\mu})^2$, we've shown

$$\underset{d \in \mathcal{D}}{\text{argmin}}\ \rho(p_\theta, d(\mathbf{y}))$$

  is given by the posterior mean for any $\mathbf{y}$

- So, it is a Bayes decision rule.
- The Bayes risk is given by

$$r(p_\theta, d) = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}$$

$$\int \rho(p_\mu, d(y))\, p(y)\, dy$$

$$\int \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\, p(y)\, dy =$$

Example: Coke bottles (cont.)

- For the Coke bottling example,

$$r(p_\theta, d) = 0.005$$

- This is expected loss before checking any bottles.
- Recall expected risk after collecting bottles (posterior risk) was also 0.005
  - Equivalent in this case, because posterior risk does not depend on $\mathbf{y}$.
- Bayes risk of $d(\mathbf{y}) = \bar{y}$ is 0.01 – twice as large.

$$\int R(\mu, d)\, p(\mu) = \frac{\sigma^2}{n} \underbrace{\int p(\mu)\, d\mu}_{1} = 0.01$$