01/20/22

1. in-person instruction from Feb/01
2. office hourse will be remote
3. HW# 1 solution will be posted

normal example w/ a proper prior for unknown θ & σ^2

† Maximum Likelihood Estimator (MLE)

- The maximum likelihood estimation approach is a way to implement the likelihood principle.

$$\ell(\hat{\theta}^{\text{MLE}} \mid x) \text{ is at least as great as } \ell(\theta \mid x) \text{ for every } \theta \in \Theta.$$

- A lot more in STAT 205(B)!

† Conditionality Principle

- **Def** If two experiments on the parameter θ, $\mathcal{E}_1$ and $\mathcal{E}_2$, are available and if one of these two experiments is selected with probability p, the resulting inference on θ should only depend on the selected experiment.

- Any inference should depend only on the outcome observed and not on any other outcome we might have observed.

- This sharply contrasts with a frequentist approach (which cares long-run behavior over repeated experiments). e.g. unbiasedness, significance levels, and power of tests, etc., violate the conditionality principle.

Example 14 (JB-p25)

- Suppose a substance to be analyzed can be sent either to a laboratory in New York or a laboratory in CA. The two labs seem equally good, so a fair coin is flipped to choose between them, with "heads" denoting that the lab in NY will be chosen. The coin is flipped and comes up tails, so the CA lab is used.

- After a while, the experimental results come back and a conclusion and report must be developed. Should this conclusion take into account the fact that the coin could have been heads and hence that the experiment in NY might have been performed instead?

- Common sense and conditional view point say **NO**, but the frequentist approach calls for averaging over all possible data, even the possible NY data.

• **Example 1.3.7:** In research laboratory, a physical quantity θ can be measured by a precise but often busy machine, which provides a measurement, $x_1 \sim N(\theta, 0.1)$, with probability $p = 0.5$, or through a less precise but always available machine, which gives $x_2 \sim N(\theta, 10)$. The machine being selected at random, depending on the availability of the more precise machine, the inference on θ when it has been selected should not depend on that fact that the alternative machine *could have been selected.*

$X_1 \sim N(\theta, 0.1)$ w/p 0.5

$X_2 \sim N(\theta, 10)$ w/p 0.5

conditionality principle

Suppose we know the precise machine was used,

$X_1 \pm 1.96 \cdot \sqrt{0.1}$ → $(-5.19, 5.19)$

Otherwise, $X_2 \pm 1.96 \times \sqrt{10}$.

$X = 0$

X

Find a s.t. $P_r(X < a) = 0.025$

$= \underbrace{P_r(P)}_{=0.5} \cdot \underbrace{P_r(X < a \mid P)}_{N(\theta, 0.1)} + \underbrace{(1 - P_r(P))}_{=0.5 \ = 1-0.5} \cdot \underbrace{P_r(X < a \mid NP)}_{N(\theta, 10)}$

$\Rightarrow a = -5.19$

- **Theorem 1.3.8** (Birnbaum, 1962) The Likelihood Principle is equivalent to the conjunction of the Sufficiency and the Conditionality Principles.

  See page 18 for the proof.

- Read Sec 1.3.4 for more about the likelihood principle.

STAT 206B
Chapter 2: Decision-Theoretic Foundations

Winter 2022

† Statistical Decision Theory

- CR Chapter 2 & JB Chapters 1 & 2.
- Decision theory deals the problem of making decisions
- Statistical decision theory: Making decisions in the presence of statistical knowledge (statistical knowledge explains some of the uncertainties involved in the decision problem)

JB Example 1 (page 5)

- Consider the situation of a drug company deciding whether or not to market a new pain killer. Two of the many factors affecting its decision are
  - ⋆⋆ the proportion of people for which the drug will prove effective (θ_1)
  - ⋆⋆ the proportion of market the drug will capture (θ_2)

- *Examples of decision problems:* estimate θ_1 & θ_2, decide whether or not to market the drug, how much to market, what price to charge, etc.

- θ_1 and θ_2 are unknown $\Rightarrow$ conduct experiments to obtain statistical information about them.

- This is a problem of statistical decision theory!

JB Example 1 (page 5) (contd)

- Consider the problem of estimating θ_2 (the proportion of market the drug will capture).

  $\star\star$ Let me use θ, not θ_2 from now on.

  $\Rightarrow$ Parameter space: $\theta \in \Theta = [0, 1]$.

- Goal: Estimating $\theta \Leftrightarrow$ Choosing a number from interval $[0, 1]$

  $\Rightarrow$ Your decision d will be a number $\in \mathcal{D} = [0, 1]$.

- Decisions are more commonly called "actions".

† Action and Action Space (Decision and Decision Space)

- $d \in \mathcal{D}$: d denotes an action (decision) and $\mathcal{D}$ the set of all possible actions under consideration (action space, decision space).

  e.g. Problem of estimating θ:

  $\mathcal{D} = \Theta$ and $d \in \Theta = \mathcal{D}$.

  e.g. Testing problem:

  $\mathcal{D} = \{accept, reject\}$.

- In the JB example, $\Theta = \mathcal{D}$ (true for an estimation problem, but not necessarily for other problems).

† Loss Function

θ: true state L(θ, d)
d: decision d = δ(x)

- Consider the standard estimation problem, i.e., $\mathcal{D} = \Theta$.
- **Def 2.1.1** A loss function is any function L from $\Theta \times \mathcal{D}$ in $[0, +\infty)$.

  ⋆⋆ The loss function evaluates the penalty (or error) $L(\theta, d)$ associated with the decision (action) d when the parameter takes the value θ for all $(\theta, d) \in \Theta \times \mathcal{D}$.

- Utility(U) and Loss (L)

  gain

$$L(\theta, d) = -U(\theta, d)$$

  ⋆⋆ Read CR Section 2.2 and JB Chapter 2 for details.

- Will discuss usual loss functions (Section 2.5).

• **Example 2.1.2:** Consider the problem of estimating the mean θ of a normal vector, $x \mid \theta \sim \mathrm{N}_n(\theta, \Sigma)$, where Σ is a known diagonal matrix with diagonal elements σ_i^2, $i = 1, \ldots, n$.

⋆⋆ $\mathcal{D} = \Theta = \mathbb{R}^n$

⋆⋆ Consider

$$L(\theta, \delta) = \sum_{i=1}^{n} \left(\frac{\delta_i - \theta_i}{\sigma_i} \right)^2,$$

where δ_i: an estimator of θ_i (the i-th component of $\boldsymbol{\theta}$).

* L takes its minimum at 0, e.g., $L(t) = t^2$ i.e., the global estimation error is the sum of the squared componentwise errors.

* $L(\theta, \delta)$ prevents the overall loss from being heavily affected by components with a large variance.

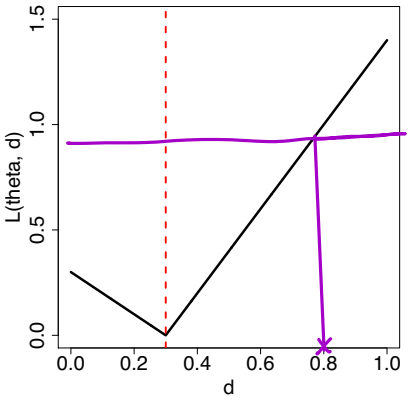
JB Example 1 (page 5) (contd)

- The company thinks that an overestimation of demand (and hence overproduction of the drug) is twice as costly as an underestimate of demand and that otherwise the loss is linear in the error.

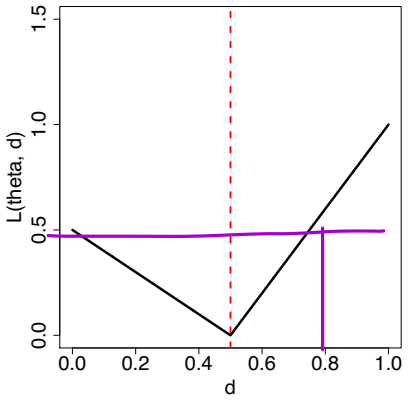
- The company might determine the loss function to be

$$L(\theta, d) = \begin{cases} |\theta - d| & \text{if } \theta \geq d \ (\textit{underestimation}), \\ 2|\theta - d| & \text{if } \theta < d \ (\textit{overestimation}), \end{cases}$$

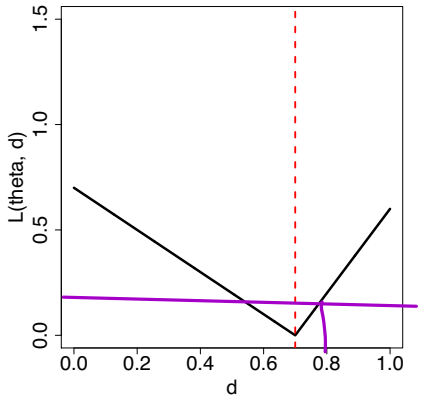
JB Example 1 (page 5) (contd)



(a) $\theta = 0.3$



(b) $\theta = 0.5$



(c) $\theta = 0.7$

JB Example 1 (page 5) (contd)

- Conduct a sample survey to obtain sample information about θ would be to.
- For example, assume n people are interviewed, and the number x who would buy the drug is observed. A reasonable choice for such x might be $x \sim \text{Bin}(n, \theta)$,

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

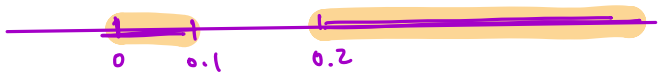
- $\mathcal{X}$: sample space (the set of all possible outcomes), x: a particular realization, $x \in \mathcal{X}$.

JB Example 1 (page 5) (contd)

- There could be considerable prior information about θ, arising from previous introductions of new similar drugs into the market.

- Suppose that new drugs tended to capture between 1/10 and 1/5 of the market, with all values between 1/10 and 1/5 being equally likely. That is,

$$\pi(\theta) = 10, \quad \text{for } \theta \in (0.1, 0.2).$$



† A fundamental basis of Bayesian Decision Theory

- Statistical inference should start with the rigorous determination of three factors;

★★ the distribution family for the observations (sampling distribution), $f(x \mid \theta)$ for $x \in \mathcal{X}$

★★ the prior distribution for the parameter $\pi(\theta)$, $\theta \in \Theta$

★★ the loss association with the decisions, $L(\theta, \delta) \in [0, +\infty)$

$\delta \in \mathcal{D}$ $\quad$ $\delta(x) = d$

$\delta(x)$

$x \longrightarrow d$

† Decision Rule (δ) $\quad \delta(x) = \bar{x} = \sum_{i=1}^{n} x_i / n \in \mathcal{D} = H$

- **JB Def 2** (p9): A (nonrandomized) decision rule $\delta(x)$ is a function from $\mathcal{X}$ into $\mathcal{D}$, i.e., the allocation of a decision to each outcome $x \sim f(x \mid \theta)$ from a random experiment.
- If x is the observed value (assumed to follow $f(x \mid \theta)$), then $\delta(x)$ is the action that will be taken.
- In estimation problems, decision rule δ, from $\mathcal{X}$ to $\mathcal{D}$ is usually called *estimator* (while the *value* $\delta(x)$ is called *estimate* of θ).
- **JB Example 1 (page 5)**: $\delta(x) = x/n$ (sample proportion): this does not incorporate the loss function or prior information

† Bayesian Approach to Decision Theory

- Minimize the expected loss of a decision d for the believed distribution of θ at the time of decision making, i.e., $\pi(\theta \mid x)$.

- The *posterior expected loss* of decision d, when the posterior distribution is $\pi(\theta \mid x)$,

$$\begin{aligned} \rho(\pi, d \mid x) &= \mathrm{E}^{\pi}\left[L(\theta, d) \mid x\right] \\ &= \int_{\Theta} L(\theta, d)\pi(\theta \mid x)d\theta. \end{aligned}$$

  $\Rightarrow$ $\rho(\pi, d \mid x)$ averages the error (loss) according to the posterior distribution of θ, conditionally on the observed data.

- A *Bayes decision*, $\delta^{\pi}(x)$ is any decision $d \in \mathcal{D}$ which minimizes $\rho(\pi, d \mid x)$.

JB Example 4(p10) Assume $X \mid \theta \sim \mathrm{N}(\theta, 1)$. The goal is estimating θ.

⋆⋆ $d \in \mathcal{D} = \mathbb{R}$.

⋆⋆ sampling distribution: $\mathrm{N}(\theta, 1)$

⋆⋆ prior distribution: $\mathrm{N}(\mu, \tau^2)$

$\Rightarrow \mathrm{N}\left(\underbrace{\left(\frac{1}{1} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{x}{1} + \frac{\mu}{\tau^2}\right)}_{= \mu_1}, \underbrace{\left(\frac{1}{1} + \frac{1}{\tau^2}\right)^{-1}}_{= \tau_1^2} \right)$

⋆⋆ loss function: squared-error loss, $L(\theta, d) = (\theta - d)^2$

Find the posterior expected loss for any $d \in \mathcal{D}$.

$(\mu_1 - d) \cdot E(\theta - \mu_1 \mid x)$

$E((X - E(X))^2) = \mathrm{Var}(X)$

Find $d^{opt} = \underset{d \in \mathcal{D}}{\operatorname{argmin}} \; \rho(\pi, d \mid x)$

$$\rho(\pi, d \mid x) = E^{\pi}(L(\theta, d) \mid x) = E^{\pi}((\theta - d)^2 \mid x)$$

$$= E^{\pi}((\theta \pm \mu_1 - d)^2 \mid x)$$

$$= E^{\pi}((\theta - \mu_1)^2 + 2(\theta - \mu_1)(\mu_1 - d) + (\mu_1 - d)^2 \mid x)$$

$$= E^{\pi}((\theta - \mu_1)^2 \mid x) + E((\mu_1 - d)^2 \mid x)$$

$$= \gamma^2 + \underbrace{E\left((\mu_1 - d)^2 | x\right)}_{\geq 0} = (\mu_1 - d)^2$$

$d = \mu_1 \rightarrow$ Bayes decision is $d = \mu_1$

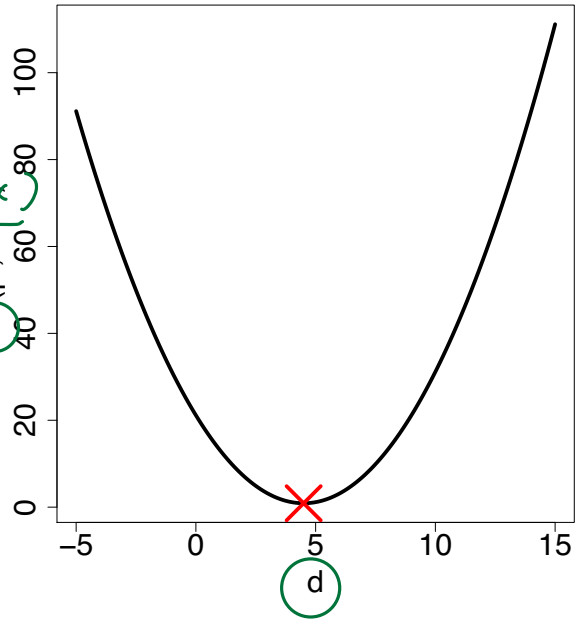
JB Example 4(p10) (contd) Suppose $x = 5$ is observed. Assume $\mu = 0$ and $\tau^2 = 9$.

$\delta(x) = \left(\frac{1}{1} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{x}{1} + \frac{\mu}{\tau^2}\right)$

$= \mu_1$

$\delta(5) = 4.5$

$\tau_1^2 + E((\mu_1 - d)^2 | x)$



- Find $\delta^\pi(x)$.

† Frequentist Risk (Average Loss, CR Section 2.3 & JB Section 1.3)

- In the frequentist paradigm, the long run performance of $\delta(x)$ by varying $x \in \mathcal{X}$ is the key.

  - **JB Def 3 (p9) & CR p61**: The *frequentist risk (or average risk)* of a decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = \mathrm{E}_\theta\left[L(\theta, \delta(x))\right] = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x \mid \theta) dx.$$

  $\Rightarrow$ The error (loss) is averaged over the different values of x proportionally to the density $f(x \mid \theta)$.

- Suppose we have multiple estimators and want to compare them (or even want to select the best estimator). *How?*

JB Example 4(p10) Assume $X \mid \theta \sim \mathrm{N}(\theta, 1)$. The goal is estimating θ under squared-error loss, $L(\theta, d) = (\theta - d)^2$. Consider the decision rule $\delta_c(x) = cx$.

- Find $R(\theta, \delta_c)$.

$$R(\theta, \delta_c) = E_\theta\left(L(\theta, \delta_c)\right)$$

$$= E_\theta\left((\theta - c\cdot x)^2\right)$$

$$\pm c\theta$$

$$= E_\theta\left(\left(c(\theta - x) + (1-c)\theta\right)^2\right)$$

$$= E_\theta\left(c^2(\theta-x)^2 + 2c(\theta-x)(1-c)\theta + (1-c)^2\theta^2\right)$$

$$= c^2 \cdot 1 + 0 + (1-c)^2\theta^2$$