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

- 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 $[0, +\infty)$.

  $\star\star$ 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)

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

  $\star\star$ 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, $x \mid \theta \sim \mathsf{N}_n(\theta, \Sigma)$, 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.

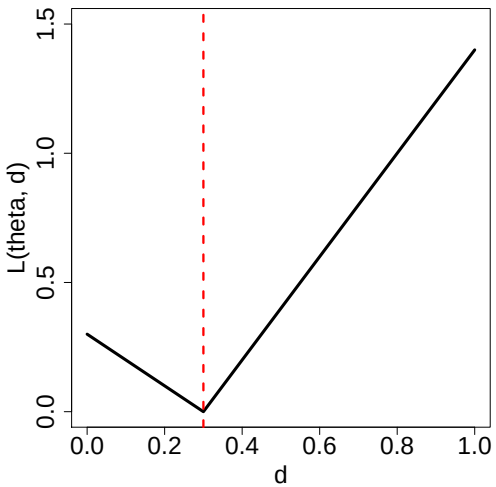
JB Example 1 (page 5) (contd)

- 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.

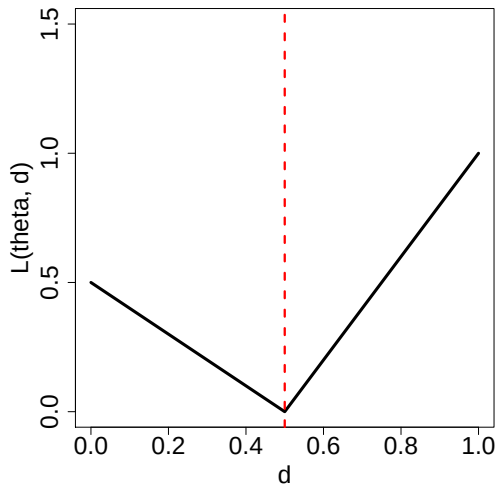
- The company might determine the loss function to be

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

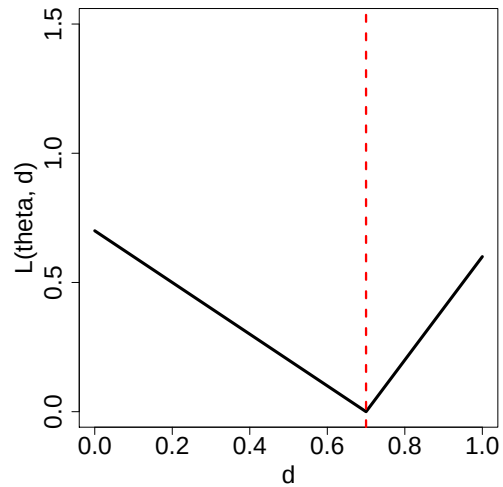
JB Example 1 (page 5) (contd)



(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, $L(\theta, \delta) \in [0, +\infty)$

† Decision Rule (δ)

- **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.

$\star\star$ $d \in \mathcal{D} = \mathbb{R}$.

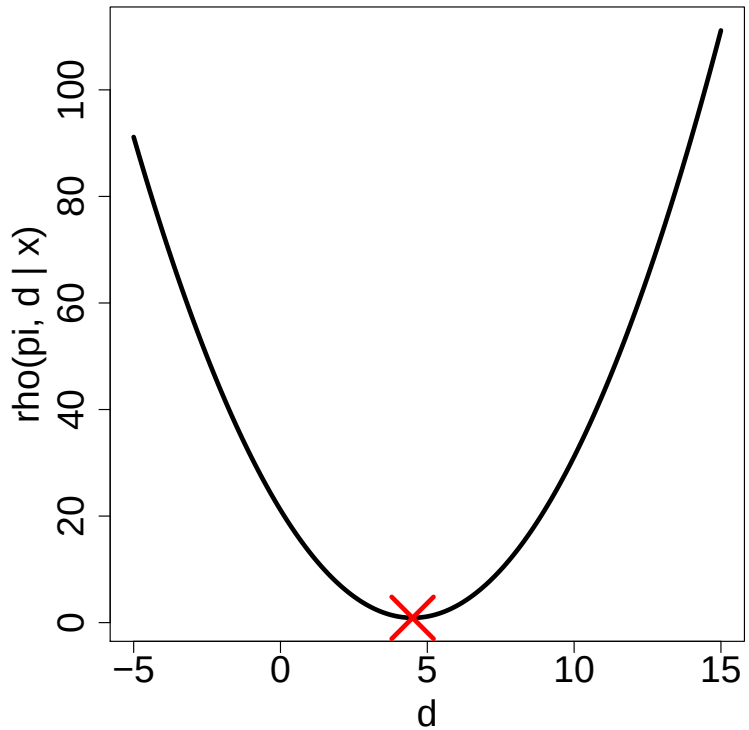
$\star\star$ sampling distribution: $\mathrm{N}(\theta, 1)$

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

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

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

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



- 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

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

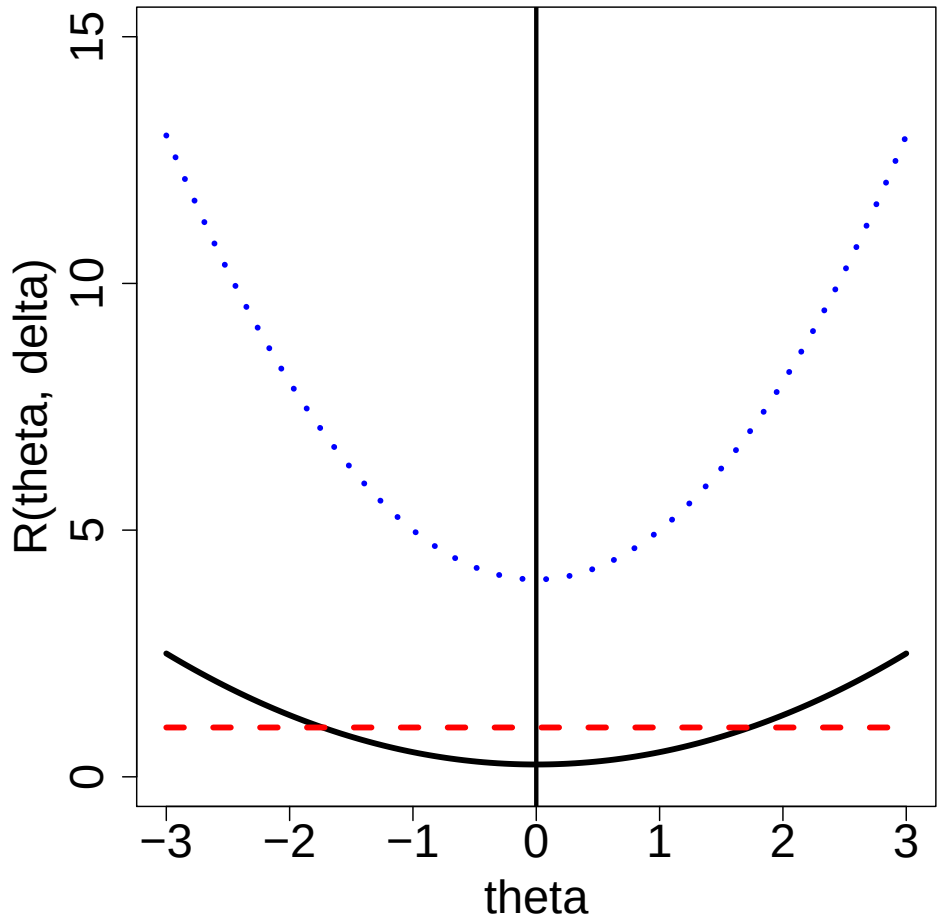
  $\Rightarrow$ The error (loss) is averaged over the different values of x proportionally to the density $f(x \mid \theta)$.

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

JB Example 4(p10) (contd) Plot of $R(\theta, \delta_c)$



- Difficulties associated with using $R(\theta, \delta)$.

  $\star\star$ For each $\theta \in \Theta$, $R(\theta, \delta)$ is the expected loss based on an average over the random $X \in \mathcal{X}$.

  $\Rightarrow$ *long-run* performance of $\delta(x)$ and **not** directly for the given observation x.

  $\star\star$ A function of $\theta \in \Theta$ & θ is unknown.

  $\Rightarrow$ The frequentist approach $R(\theta, \delta)$ does not induce a total ordering on the set of procedures.

† How can Frequentists choose δ?

- An additional principle must be introduced to select a specific rule for use.

  e.g. δ_1 is preferred to δ_3 under some concept of optimalty.

- Some important frequentist decision principles (CR 2.4 + a lot in JB)

  ⋆⋆ Bayes risk principle

  ⋆⋆ minimax

  ⋆⋆ admissibility

  ⋆⋆ restricted classes: e.g. we only consider unbiased estimators.

- Bayes estimators are *often optimal* for the frequentist concepts of optimality.

† The Bayes Risk Principle

- The frequentist risk of a decision rule $\delta(x)$ is a function of θ.

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x \mid \theta) dx.$$

- The *integrated risk* (also called Bayes Risk) is the frequentist risk averaged over Θ according to their prior $\pi(\theta)$.

$$\begin{aligned} r(\pi, \delta) &= \mathrm{E}^{\pi}\left[R(\theta, \delta)\right] \\ &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, d) f(x \mid \theta) dx \pi(\theta) d\theta. \end{aligned}$$

- $r(\pi, \delta)$ is a real number associated with estimator δ.

  $\Rightarrow$ Induces a total ordering on the set of estimators, so allows for the direct comparison of estimators.

† Any connection between $r(\pi, \delta)$ and $\rho(\pi, \delta \mid x)$?

$\Rightarrow$ They lead to the same decision.

- **Th 2.3.2** An estimator minimizing the integrated risk $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ which minimizes the posterior expected loss, $\rho(\pi, \delta \mid x)$, since

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\theta, \delta(x) \mid x) m(x) dx.$$

- **Def 2.3.3**

  ⋆⋆ **A Bayes estimator** associated with a prior distribution π and a loss function L is any estimator δ^π, which minimizes $r(\pi, \delta)$.

  ⋆⋆ For every $x \in \mathcal{X}$, it is given by $\delta^\pi(x)$ (**a Bayes action**), argument of $\min_d \rho(\pi, d \mid x)$.

  ⋆⋆ The value $r(\pi) = r(\pi, \delta^\pi)$ is then called **the Bayes risk**.

- **JB Def 9, p160** If π is an improper prior, but $\delta^\pi(x)$ is an action which minimizes $\rho(\pi, d \mid x)$ for each x with $m(x) > 0$, then δ^π is called a **generalized Bayes rule.**

† Minimaxity: Minimize the expected loss in the least favorable case ($\Leftrightarrow$ protect against the worst possible state of nature, conservative!)

- **The Minimax Principle.** **JB p18** δ_1 is preferred to δ_2 if

$$\sup_{\theta} R(\theta, \delta_1) < \sup_{\theta} R(\theta, \delta_2).$$

Example 2.4.4 The first oil-drilling platforms in the North Sea were designed according to a minimax principle. In fact, they were supposed to resist the conjugate action of the worst gale and the worst storm ever observed, at the minimal record temperature. This strategy obviously gives a comfortable margin of safety, but is quite costly. For more recent platforms, engineers have taken into account the distribution of these weather phenomenon in order to reduce the production cost.

- **Def 2.4.3** The **minimax risk** associated with a loss function L is the value

$$\bar{R} = \inf_{\delta \in \mathcal{D}} \sup_{\theta} R(\theta, \delta) = \inf_{\delta \in \mathcal{D}} \sup_{\theta} \mathrm{E}_{\theta} \left\{ L(\theta, \delta(x)) \right\},$$

and a **minimax estimator** is any (possibly randomized) estimator δ_0 such that

$$\sup_{\theta} R(\theta, \delta_0) = \bar{R}$$

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

† Let's connect the minimax optimality to the Bayesian approach (*Very high level!*)

- least favorable prior $\pi^\star$: a prior concentrated on the worst cases of θ to protect against the least favorable possible values of θ.
- **Connection!** Find a least favorable prior distribution + determine the resulting Bayes rule $\Rightarrow$ minimax rule.
- Recall that the Bayes risk for π is $r(\pi) = \inf_{\delta \in \mathcal{D}} r(\pi, \delta)$ and such δ is a Bayes estimator.
- **Lamma 2.4.13** If δ_0 is a Bayes estimator with respect to π_0 and if $R(\theta, \delta_0) \leq r(\pi_0)$ for every θ in the support of π_0, δ_0 is minimax and π_0 is the least favorable distribution.

† Connection (contd)

- Under a prior distribution, each value of the parameter cannot be equally weighted.
- The least favorable prior often induces a strong prior bias towards a few points of the sample space. So it is often unrealistic so the minimax principle is not necessarily appealing for a Bayesian point of view.
- May be relevant from a robustness point of view. That is, when the prior information is not precise enough to determine the prior distribution.

† Admissibility

- **Def 2.4.19** An estimator δ_0 is inadmissible if there eixsts an estimator δ_1 which dominates δ_0, that is, such that, for every θ

$$R(\theta, \delta_0) \geq R(\theta, \delta)$$

and, for at least one value θ_0 of the parameter,

$$R(\theta_0, \delta_0) > R(\theta_0, \delta).$$

Otherwise, δ_0 is said to be admissible.

- What is the underlying idea as a criterion?

Inadmissible estimators should not be considered at all since they can be uniformly improved!

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

$*$ Admissibility is related (stronger than minmax) to the Bayesian paradigm.

- Admissibility is automatically satisfied by most Bayes estimators.
- **Prop 2.4.22** If a prior distribution π is strictly positive on Θ, with finite Bayes risk and the risk function, $R(\theta, \delta)$, is a continuous function of θ for every δ, the Bayes estimator δ^π is admissible.
- Want to learn more? *Read CR 2 and JB 4.8*

Example 2.4.6 (Stein Phenomenon) Suppose a $p-$dimensional vector, $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\theta}, I_p)$ and consider the problem of estimating $\boldsymbol{\theta}$ (a $p-$dim vector). Assume the quadratic loss function $L(\boldsymbol{\theta}, \delta) = (\boldsymbol{\theta} - \boldsymbol{\delta})'(\boldsymbol{\theta} - \boldsymbol{\delta}) = ||\boldsymbol{\theta} - \boldsymbol{\delta}||^2$.

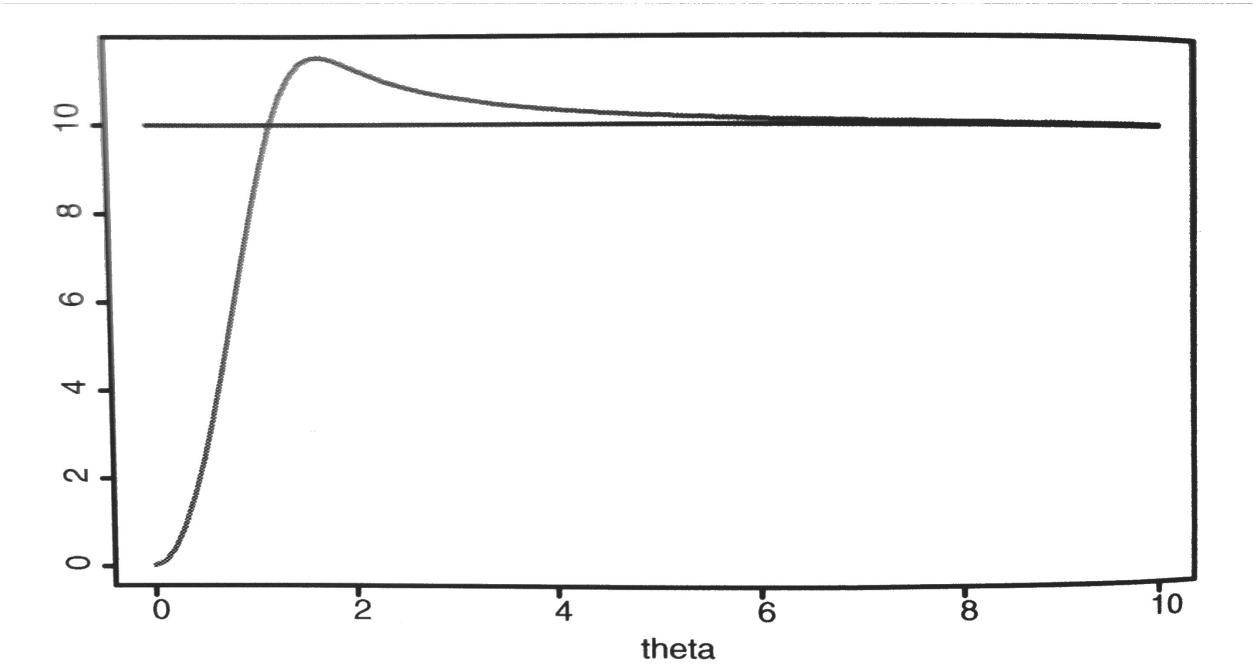
- The maximum likelihood estimator $\boldsymbol{\delta}_1(\boldsymbol{X}) = \boldsymbol{X}$
  - $\star\star$ The least squares estimator in standard regression setting
  - $\star\star$ For $p = 1$ or 2, it is admissible and the unique minimax estimator.

Example 2.4.6 (contd)

- Consider the positive part James-Stein estimator,

$$\delta_2(\boldsymbol{X}) = \begin{cases} \left(1 - \frac{2p-1}{||\boldsymbol{X}||^2}\right)\boldsymbol{X} & \text{if } ||\boldsymbol{X}||^2 \geq 2p-1, \\ \boldsymbol{0} & \text{ow.} \end{cases}$$

♣ Figure 2.4.1 Comparison of the risks of δ_1 and δ_2 for $p = 10$



Example 2.4.6 (contd)

- ⋆⋆ $\boldsymbol{\delta}_2$ cannot be minimax.
- ⋆⋆ $\boldsymbol{\delta}_2$ is definitely superior on some (the most interesting) part of the parameter space.
- ⋆⋆ "The Stein effect": allows to borrow information from the other components, even when they are independent and deal with totally different estimation problems.
- ⋆⋆ Sometimes the minimax rule is not useful! (or sometimes may not exist)
- ⋆⋆ Following James and Stein, extensive research on this has been done – *Shrinkage estimators*

† Usual loss functions (CR Section 2.5)

- Quadratic loss
- Absolute loss
- 0-1 loss
- Intrinsic loss – entropy distance (Kullback-Leibler divergence), Hellinger loss...

⋆⋆ What are the Bayesian estimators $\delta^\pi(x)$ under the classical loss functions?

† The quadratic loss

$$L(\theta, d) = (\theta - d)^2$$

- most common evaluation criterion – simplicity
- penalize large deviations too heavily

• **Prop 2.5.1** The Bayes estimator δ^π associated with the prior distribution π and with the quadratic loss $L(\theta, d) = (\theta - d)^2$, is the posterior expectation,

$$\delta^\pi(x) = \mathsf{E}^\pi(\theta \mid x) = \frac{\int_\Theta \theta f(x \mid \theta)\pi(\theta)d\theta}{\int_\Theta f(x \mid \theta)\pi(\theta)d\theta}.$$

- **Cor 2.5.2** The Bayes estimator δ^π associated with π and with the weighted quadratic loss $L(\theta, d) = w(\theta)(\theta - d)^2$, where $w(\theta)$ is a nonnegative function, is

$$\delta^\pi(x) = \frac{\mathrm{E}^\pi(w(\theta) \cdot \theta \mid x)}{\mathrm{E}^\pi(w(\theta) \mid x)}$$

- **Cor 2.5.3** When $\Theta \in \mathbb{R}^p$, the Bayes estimator δ^π associated with the prior distribution π and with the quadratic loss $L(\boldsymbol{\theta}, \boldsymbol{d}) = (\boldsymbol{\theta} - \boldsymbol{d})^t \boldsymbol{Q}(\boldsymbol{\theta} - \boldsymbol{d})$, is the posterior mean $\delta^\pi(x) = \mathrm{E}^{\boldsymbol{\theta}|x}(\theta \mid x)$, for every positive -definite symmetric $p \times p$ matrix $\boldsymbol{Q}$.

† The absolute error loss

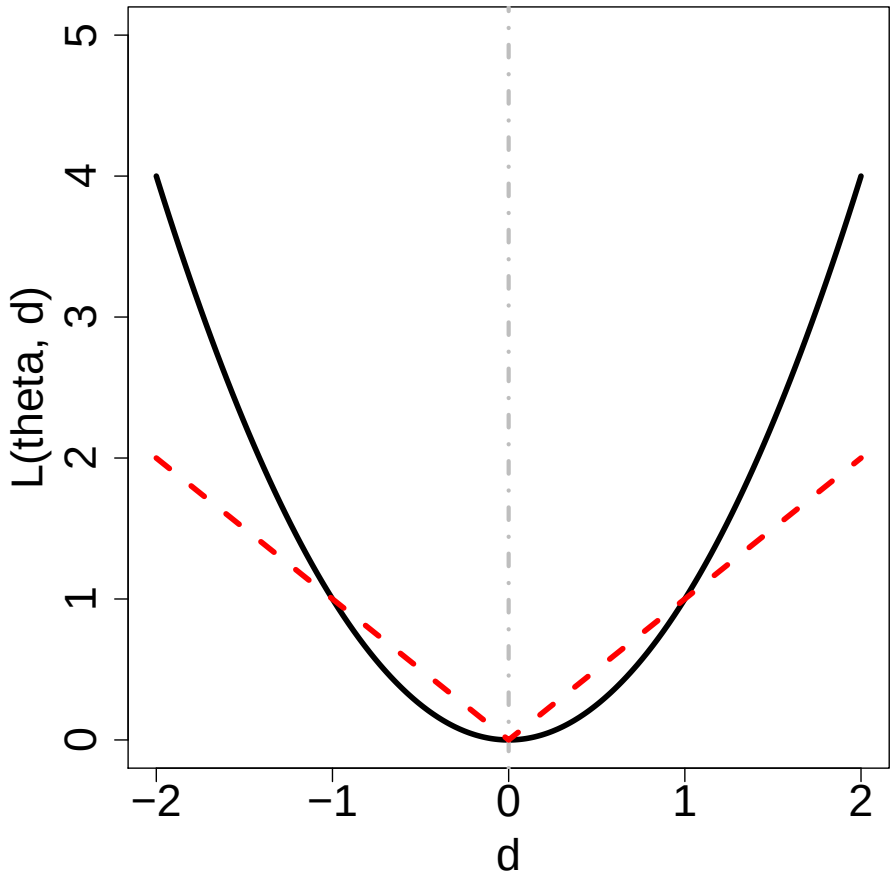
$$L(\theta, d) = |\theta - d|$$

Or multilinear function (more general)

$$L(\theta, d) = \begin{cases} k_1(\theta - d) & \text{if } \theta - d \geq 0, \\ k_2(d - \theta) & \text{if } \theta - d < 0, \end{cases}$$

- slow down the progression of the quadratic loss for large errors and has a robustifying effect.
- k_1 and k_2 reflect the relative importance of underestimation or overestimation.
- $k_1 = k_2 \Rightarrow$ the absolute error loss

♣ squared error loss vs absolute error loss



- **Prop 2.5.5** A Bayes estimator associated with the prior distribution π and the multilinear loss is a $k_1/(k_1 + k_2)$ fractile of $\pi(\theta \mid x)$.

- If $k_1 = k_2$, the Bayes estimator is the posterior median.

† The 0-1 loss: the penalty associated with an estimator δ is 0 if the answer is correct and 1 otherwise.

Example 2.5.6 Consider the test of $H0 : \theta \in \Theta_0$ vs $H1 : \theta \notin \Theta_0$. Then $\mathcal{D} = \{0, 1\}$, where 1 stands for acceptance of $H0$ and 0 for rejection. For the 0-1 loss,

$$L(\theta, d) = \begin{cases} 1 - d & \text{if } \theta \in \Theta_0, \\ d & \text{otherwise,} \end{cases}$$

the associated risk is

$$\begin{aligned} R(\theta, \delta) &= \mathsf{E}(L(\theta, \delta(x)) \\ &= \begin{cases} \Pr_\theta(\delta(x) = 0) & \text{if } \theta \in \Theta_0, \\ \Pr_\theta(\delta(x) = 1) & \text{otherwise.} \end{cases} \end{aligned}$$

• **Prop 2.5.7** The Bayes estimator associated with the prior distribution π and with the 0-1 loss is

$$\delta^{\pi}(x) = \begin{cases} 1 & \text{if } \Pr(\theta \in \Theta_0 \mid x) > \Pr(\theta \notin \Theta_0 \mid x), \\ 0 & \text{otherwise,} \end{cases}$$

i.e., $\delta^{\pi}(x)$ is equal to 1 if and only if $\Pr(\theta \in \Theta \mid x) > 1/2$.

† Intrinsic losses: $f(x \mid \theta)$ is of interest.

$$\Rightarrow \quad L(\theta, \delta) = d(f(\cdot \mid \theta), f(\cdot \mid \delta)),$$

where $d(\cdot, \cdot)$ is a distance measure.

- **entropy distance (Kullback-Leibler divergence)**:

$$L_e(\theta, \delta) = \mathsf{E}_\theta \left[\log \left(\frac{f(x \mid \theta)}{f(x \mid \delta)} \right) \right].$$

- **Hellinger Distance**

$$L_H(\theta, \delta) = \mathsf{E}_\theta \left[\left(\sqrt{\frac{f(x \mid \theta)}{f(x \mid \delta)}} - 1 \right)^2 \right].$$