Lecture 2: Statistical Decision Theory (Part I)

Hao Helen Zhang

Outline of This Note

- Part I: Statistics Decision Theory (*from Statistical Perspectives - "Estimation"*)
  - loss and risk
  - MSE and bias-variance tradeoff
  - Bayes risk and minimax risk
- Part II: Learning Theory for Supervised Learning (*from Machine Learning Perspectives - "Prediction"*)
  - optimal learner
  - empirical risk minimization
  - restricted estimators

Statistical Inference

Assume data $Z = (Z_1, \cdots, Z_n)$ follow the distribution $f(z|\theta)$.

- $\theta \in \Theta$ is the parameter of interest, but unknown. It represents uncertainties.
- θ is a scalar, vector, or matrix
- Θ is the set containing all possible values of θ.

The goal is to estimate θ using the data.

Statistical decision theory is concerned with the problem of making decisions.

- It combines the sampling information (data) with a knowledge of the consequences of our decisions.

Three major types of inference:

- point estimator ("educated guess"): $\hat{\theta}(\mathrm{Z})$
- confidence interval $P(\theta \in [L(\mathrm{Z}), U(\mathrm{Z})]) = 95\%$
- hypotheses testing $H_0 : \theta = 0$ vs $H_1 : \theta = 1$

Early works in decision theoy was extensively done by Wald (1950).

Bias and Variance

- Bias of $\hat{\theta}$

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- Variance of $\hat{\theta}$

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

Loss Function

How to measure the quality of $\hat{\theta}$? Use a loss function

$$L(\theta, \hat{\theta}(Z)) : \quad \Theta \times \Theta \longrightarrow R.$$

- The loss is non-negative

$$L(\theta, \hat{\theta}) \geq 0, \quad \forall \theta, \hat{\theta}.$$

- known as *gains* or *utility* in economics and business.
- A loss quantifies the consequence for each decision $\hat{\theta}$, for various possible values of θ.

In decision theory,

- θ is called the *state of nature*
- $\hat{\theta}(Z)$ is called an *action*.

Examples of Loss Functions

For regression,

- squared loss function: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
- absolute error loss: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$
- L_p loss: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|^p$

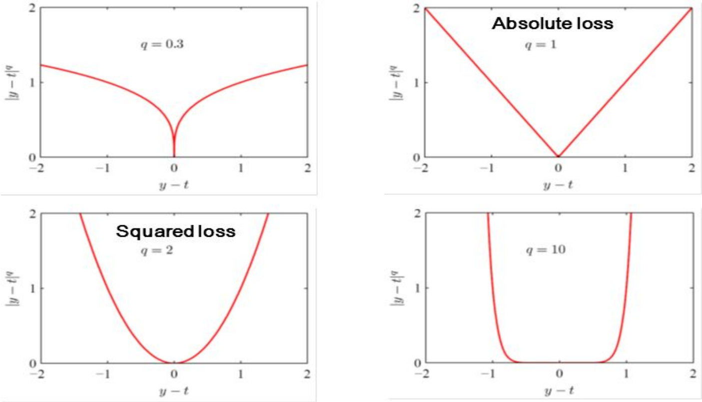
For classification

- 0-1 loss function: $L(\theta, \hat{\theta}) = I(\theta \neq \hat{\theta})$

Density estimation

- Kullback-Leibler loss: $L(\theta, \hat{\theta}) = \int \log \left(\frac{f(z|\theta)}{f(z|\hat{\theta})} \right) f(z|\theta) dz$

Other loss functions



Risk Function

Note that $L(\theta, \hat{\theta}(Z))$ is a function of Z (which is random)

- Intuitively, we prefer decision rules with small "expected loss" or "long-term average loss", resulted from the use of $\hat{\theta}(Z)$ repeatedly with varying Z.
- This leads to the *risk function* of a decision rule.

The **risk function** of an estimator $\hat{\theta}(Z)$ is

$$R(\theta, \hat{\theta}(Z)) = E_\theta[L(\theta, \hat{\theta}(Z))] = \int_{\mathcal{Z}} L(\theta, \hat{\theta}(z)) f(z|\theta) dz,$$

where $\mathcal{Z}$ is the sample space (the set of possible outcomes) of Z.

- The expectation is taken over data Z; θ is fixed.

About Risk Function (Frequenst Interpretation)

The risk function

- $R(\theta, \hat{\theta})$ is a deterministic function of θ.
- $R(\theta, \hat{\theta}) \geq 0$ for any θ.

We use the risk function

- to evaluate the overall performance of one estimator/action/decision rule
- to compare two estimators/actions/decision rules
- to find the best (optimal) estimator/action/decision rule

Example: Consider the squared loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta}(\mathrm{Z}))^2$. Its risk is

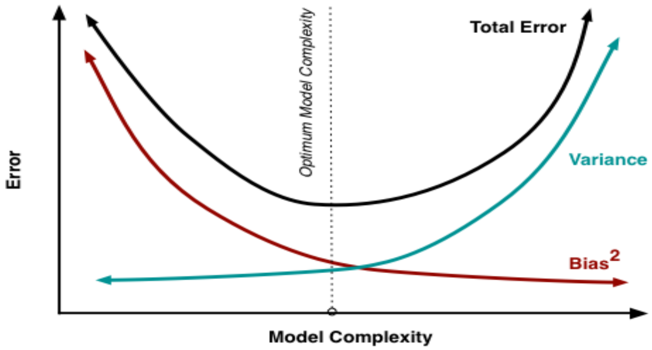
$$R(\theta, \hat{\theta}) = E[\theta - \hat{\theta}(\mathrm{Z})]^2,$$

which is called **mean squared error** (**MSE**).

The MSE is the sum of **squared bias** of $\hat{\theta}$ and **its variance**.

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{E}_\theta[\theta - \hat{\theta}(\mathrm{Z})]^2 \\
&= \mathrm{E}_\theta[\theta - E_\theta\hat{\theta}(\mathrm{Z}) + E_\theta\hat{\theta}(\mathrm{Z}) - \hat{\theta}(\mathrm{Z})]^2 \\
&= \mathrm{E}_\theta[\theta - E_\theta\hat{\theta}(\mathrm{Z})]^2 + E_\theta[\hat{\theta}(\mathrm{Z}) - E_\theta\hat{\theta}(\mathrm{Z})]^2 + 0 \\
&= [\theta - E_\theta\hat{\theta}(\mathrm{Z})]^2 + E_\theta[\hat{\theta}(\mathrm{Z}) - E_\theta\hat{\theta}(\mathrm{Z})]^2 \\
&= \mathrm{Bias}^2_\theta[\hat{\theta}(\mathrm{Z})] + \mathrm{Var}_\theta[\hat{\theta}(\mathrm{Z})].
\end{aligned}$$

Both bias and variance contribute to the risk.


Total Error
Variance
Bias^2
Error
Optimum Model Complexity
Model Complexity

Risk Comparison: Which Estimator is Better

Given $\hat{\theta}_1$ and $\hat{\theta}_2$, we say $\hat{\theta}_1$ is the preferred estimator if

$$R(\theta, \hat{\theta}_1) < R(\theta, \hat{\theta}_2), \quad \forall \theta \in \Theta.$$

- We need compare two curves as functions of θ.
- If the risk of $\hat{\theta}_1$ is uniformly dominated by (smaller than) that of $\hat{\theta}_2$, then $\hat{\theta}_1$ is the winner!

Example 1

The data $Z_1, \cdots, Z_n \sim N(\theta, \sigma^2), n > 3$. Consider

- $\hat{\theta}_1 = Z_1$,
- $\hat{\theta}_2 = \frac{Z_1+Z_2+Z_3}{3}$

Which is a better estimator under the squared loss?

Example 1

The data $Z_1, \cdots, Z_n \sim N(\theta, \sigma^2), n > 3$. Consider

- $\hat{\theta}_1 = Z_1$,
- $\hat{\theta}_2 = \frac{Z_1+Z_2+Z_3}{3}$

Which is a better estimator under the squared loss?

Answer: Note that

$$R(\theta, \hat{\theta}_1) = \text{Bias}^2(\hat{\theta}_1) + \text{Var}(\hat{\theta}_1) = 0 + \sigma^2 = \sigma^2,$$

$$R(\theta, \hat{\theta}_2) = \text{Bias}^2(\hat{\theta}_2) + \text{Var}(\hat{\theta}_2) = 0 + \sigma^2/3 = \sigma^2/3.$$

Since

$$R(\theta, \hat{\theta}_2) < R(\theta, \hat{\theta}_1), \ \forall\theta$$

$\hat{\theta}_2$ is better than $\hat{\theta}_1$.

Best Decision Rule (Optimality)

We say the estimator $\hat{\theta}^*$ is **best** if it is better than any other estimator. And $\hat{\theta}^*$ is called the **optimal** decision rule.

- In principle, the best decision rule $\hat{\theta}^*$ has uniformly the smallest risk R for all values of $\theta \in \Theta$.
- In visualization, the risk curve of $\hat{\theta}^*$ is uniformly the lowest among all possible risk curves over the entire Θ.

However, in many cases, such a best solution does not exist.

- One can always reduce the risk at a specific point θ_0 to zero by making $\hat{\theta}$ equal to θ_0 for all z.

Example 2

Assume a single observation $Z \sim N(\theta, 1)$. Consider two estimators:

- $\hat{\theta}_1 = Z$
- $\hat{\theta}_2 = 3$.

Using the squared error loss, direct computation gives

$$\begin{aligned} R(\theta, \hat{\theta}_1) &= E_\theta(Z - \theta)^2 = 1. \\ R(\theta, \hat{\theta}_2) &= E_\theta(3 - \theta)^2 = (3 - \theta)^2. \end{aligned}$$

Which has a smaller risk?

Example 2

Assume a single observation $Z \sim N(\theta, 1)$. Consider two estimators:

- $\hat{\theta}_1 = Z$
- $\hat{\theta}_2 = 3.$

Using the squared error loss, direct computation gives

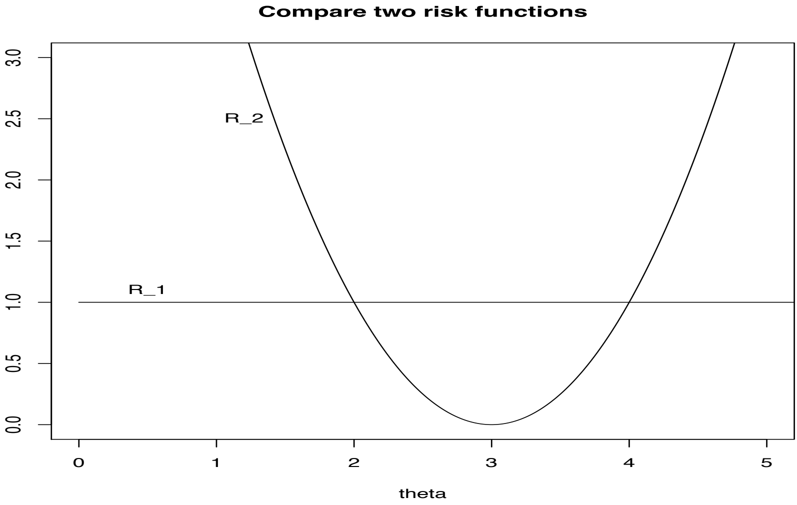
$$\begin{aligned} R(\theta, \hat{\theta}_1) &= E_\theta (Z - \theta)^2 = 1. \\ R(\theta, \hat{\theta}_2) &= E_\theta (3 - \theta)^2 = (3 - \theta)^2. \end{aligned}$$

Which has a smaller risk?

Comparison:

- If $2 < \theta < 4$, then $R(\theta, \hat{\theta}_2) < R(\theta, \hat{\theta}_1)$, so $\hat{\theta}_2$ is better.
- Otherwise, $R(\theta, \hat{\theta}_1) < R(\theta, \hat{\theta}_2)$, so $\hat{\theta}_1$ is better.

Two risk functions cross. Neither estimator uniformly dominates the other.



Compare two risk functions
R_2
R_1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
theta

Best Decision Rule from a Class

In general, there exists no *uniformly best* estimator which simultaneously minimizes the risk for all values of θ.

How to avoid this difficulty?

Best Decision Rule from a Class

In general, there exists no *uniformly best* estimator which simultaneously minimizes the risk for all values of θ.

How to avoid this difficulty?
One solution is to

- restrict the estimators within a class $\mathcal{C}$, which rules out estimators that overly favor specific values of θ at the cost of neglecting other possible values.

Commonly used restricted classes of estimators:

- $\mathcal{C}$=\{unbiased estimators\}, i.e., $\mathcal{C} = \{\hat{\theta} : E_\theta[\hat{\theta}(\mathrm{Z})] = \theta\}$.
- $\mathcal{C}$=\{linear decision rules\}

Uniformly Minimum Variance Unbiased Estimator (UMVUE)

Example 3: The data $Z_1, \cdots, Z_n \sim N(\theta, \sigma^2), n > 3$. Compare three estimators

- $\hat{\theta}_1 = Z_1$
- $\hat{\theta}_2 = \frac{Z_1+Z_2+Z_3}{3}$
- $\hat{\theta}_3 = \bar{Z}$.

Which is the best **unbiased** estimator under the squared loss?

All the three are unbiased for θ. So their risk is equal to variance,

$$R(\theta, \hat{\theta}_j) = \text{Var}(\hat{\theta}_j), \quad j = 1, 2, 3.$$

Since $\text{Var}(\hat{\theta}_1) = \sigma^2, \text{Var}(\hat{\theta}_2) = \frac{\sigma^2}{3}, \text{Var}(\hat{\theta}_3) = \frac{\sigma^2}{n}$, so $\hat{\theta}_3$ is the best.

Actually, $\hat{\theta}_3 = \bar{Z}$ is the best in $\mathcal{C} =$ {unbiased estimators}. Call it **UMVUE**.

BLUE (Best Linear Unbiased Estimator)

The data $Z_i = (\boldsymbol{X}_i, Y_i)$ follows the model

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \cdots n,$$

- $\boldsymbol{\beta}$ is a vector of non-random unknown parameters
- X_{ij} are "explanatory variables"
- ε_i's are uncorrelated, random error terms following Gaussian-Markov assumptions: $E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2 < \infty$.

$\mathcal{C}$ ={unbiased, linear estimators}. The "linear" means $\widehat{\boldsymbol{\beta}}$ is linear in Y.

Gauss-Markov Theorem: The ordinary least squares estimator (OLS) $\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'$y is best <u>linear unbiased estimator</u> (BLUE) of $\boldsymbol{\beta}$.

Alternative Optimality Measures

The risk R is a function of θ, not easy to use.

Alternative ways for comparing the estimators?

Alternative Optimality Measures

The risk R is a function of θ, not easy to use.

Alternative ways for comparing the estimators?

In practice, we sometimes use a one-number summary of the risk.

- Maximum Risk
$$\bar{R}(\hat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \hat{\theta}).$$
- Bayes Risk
$$r_B(\pi, \hat{\theta}) = \int_\Theta R(\theta, \hat{\theta}) \pi(\theta) d\theta,$$
where $\pi(\theta)$ is a prior for θ.

They lead to optimal estimators under different senses.

- the **minimax** rule: consider the worse-case risk (conservative)
- the **Bayes** rule: the average risk according to the prior beliefs about θ.

Minimax Rule

A decision rule that minimizes the maximum risk is called a **minimax** rule, also known as **MinMax** or **MM**

$$\bar{R}(\hat{\theta}^{MinMax}) = \inf_{\hat{\theta}} \bar{R}(\hat{\theta}),$$

where the infimum is over all estimators $\hat{\theta}$. Or, equivalently,

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}^{MinMax}) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta}).$$

- The MinMax rule focuses on the worse-case risk.
- The MinMax rule is a very conservative decision-making rule.

Example 4: Maximum Binomial Risk

Let $Z_1, \cdots, Z_n \sim Bernoulli(p)$. Under the square loss,

- $\hat{p}_1 = \bar{Z}$,
- $\hat{p}_2 = \frac{\sum_{i=1}^n Z_i + \sqrt{n/4}}{n+\sqrt{n}}$.

Then their risk is

$$R(p, \hat{p}_1) = \text{Var}(\hat{p}_1) = \frac{p(1-p)}{n}.$$

and

$$R(p, \hat{p}_2) = \text{Var}(\hat{p}_2) + [\text{Bias}(\hat{p}_2)]^2 = \frac{n}{4(n+\sqrt{n})^2}.$$

Note: $\hat{p}_2$ is the Bayes estimator obtained by using a Beta(α, β) prior for p (to be discussed in Example 6).

Example: Maximum Binomial Risk (cont.)

Consider their maximum risk

$$\bar{R}(\hat{p}_1) = \max_{0 \le p \le 1} \frac{p(1-p)}{n} = \frac{1}{4n}.$$
$$\bar{R}(\hat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

Based on the maximum risk, $\hat{\theta}_2$ is better than $\hat{\theta}_1$.

Note that $R(\hat{p}_2)$ is a constant. (Draw a picture)

Maximum Binomial Risk (continued)

The ratio of two risk functions is

$$\frac{R(p, \hat{p}_1)}{R(p, \hat{p}_2)} = 4p(1-p)\frac{(n+\sqrt{n})^2}{n^2},$$

- When n is large, $R(p, \hat{p}_1)$ is smaller than $R(p, \hat{p}_2)$ except for a small region near $p = 1/2$.
- Many people prefer $\hat{p}_1$ to $\hat{p}_2$.
- Considering the worst-case risk only can be conservative.

Bayes Risk

Frequentist vs Bayes Inferences:

- Classical approaches ("frequentist") treat θ as a fixed but unknown constant.
- By contrast, Bayesian approaches treat θ as a *random* quantity, taking value from Θ.
  - θ has a probability distribution $\pi(\theta)$, which is called the *prior* distribution.

The decision rule derived using the Bayes risk is called the **Bayes** decision rule or **Bayes estimator**.

Bayes Estimation

- θ follows a prior distribution $\pi(\theta)$

$$\theta \sim \pi(\theta).$$

- Given θ, the distribution of a sample z is

$$\mathrm{z}|\theta \sim f(\mathrm{z}|\theta).$$

The marginal distribution of z:

$$m(\mathrm{z}) = \int f(\mathrm{z}|\theta)\pi(\theta)d\theta$$

- After observing the sample, the prior $\pi(\theta)$ is updated with sample information. The updated prior is called the *posterior* $\pi(\theta|\mathrm{z})$, which is the conditional distribution of θ given z,

$$\pi(\theta|\mathrm{z}) = \frac{f(\mathrm{z}|\theta)\pi(\theta)}{m(\mathrm{z})} = \frac{f(\mathrm{z}|\theta)\pi(\theta)}{\int f(\mathrm{z}|\theta)\pi(\theta)d\theta}.$$

Bayes Risk and Bayes Rule

The **Bayes risk** of $\hat{\theta}$ is defined as

$$r_B(\pi, \hat{\theta}) = \int_{\Theta} R(\theta, \hat{\theta})\pi(\theta)d\theta,$$

where $\pi(\theta)$ is a prior, $R(\theta, \hat{\theta}) = E[L(\theta, \hat{\theta})|\theta]$ is the frequentist risk.

- Bayes risk is the weighted average of $R(\theta, \hat{\theta})$, where the weight is specified by $\pi(\theta)$.

Bayes Rule $\hat{\theta}_{\pi}^{Bayes}$: the decision rule that minimizes the Bayes risk

$$r_B(\pi, \hat{\theta}_{\pi}^{Bayes}) = \inf_{\hat{\theta}} r_B(\pi, \hat{\theta}),$$

where the infimum is over all estimators $\tilde{\theta}$.

- The Bayes rule depends on the prior π.

Posterior Risk

Assume $Z \sim f(z|\theta)$ and $\theta \sim \pi(\theta)$.

For any estimator $\hat{\theta}$, define its **posterior risk**

$$r(\hat{\theta}|z) = \int L(\theta, \hat{\theta}(z))\pi(\theta|z)d\theta.$$

- The posterior risk is a function only of z not a function of θ.

Connections: Risk, Bayes Risk, and Posterior Risk

The risk is defined as

$$R(\theta, \hat{\theta}) = E_{Z|\theta} L(\theta, \hat{\theta}(Z)).$$

The Bayes risk is defined as

$$\begin{aligned} r_B(\pi, \hat{\theta}) &= E_\theta R(\theta, \hat{\theta}) \\ &= E_\theta E_{\mathrm{Z}|\theta} L(\theta, \hat{\theta}(\mathrm{Z})) \\ &= E_{\theta,\mathrm{Z}} L(\theta, \hat{\theta}(\mathrm{Z})) \\ &= E_{\mathrm{Z}} E_{\theta|\mathrm{Z}} L(\theta, \hat{\theta}(\mathrm{Z})) \\ &= E_{\mathrm{Z}} r(\hat{\theta}|\mathrm{Z}), \end{aligned}$$

Remark: For any two random variables X and Y, we have

$$E_{X,Y} f(X, Y) = E_X E_{Y|X} f(X, Y) = E_Y E_{Y|X} f(X, Y).$$

Alternative Proof

Theorem: The Bayes risk $r_B(\pi, \hat{\theta})$ can be expressed as

$$r_B(\pi, \hat{\theta}) = \int r(\hat{\theta}|\mathrm{z}) m(\mathrm{z}) d\mathrm{z}.$$

Alternative Proof

Theorem: The Bayes risk $r_B(\pi, \hat{\theta})$ can be expressed as

$$r_B(\pi, \hat{\theta}) = \int r(\hat{\theta}|\mathrm{z}) m(\mathrm{z}) d\mathrm{z}.$$

Proof:

$$\begin{aligned} r_B(\pi, \hat{\theta}) &= \int_{\Theta} R(\theta, \hat{\theta}) \pi(\theta) d\theta = \int_{\Theta} \left[\int_{\mathcal{Z}} L(\theta, \hat{\theta}(\mathrm{z})) f(\mathrm{z}|\theta) d\mathrm{z} \right] \pi(\theta) d\theta \\ &= \int_{\Theta} \int_{\mathcal{Z}} L(\theta, \hat{\theta}(\mathrm{z})) f(\mathrm{z}|\theta) \pi(\theta) d\mathrm{z} d\theta \\ &= \int_{\Theta} \int_{\mathcal{Z}} L(\theta, \hat{\theta}(\mathrm{z})) m(\mathrm{z}) \pi(\theta|\mathrm{z}) d\mathrm{z} d\theta \\ &= \int_{\mathcal{Z}} \left[\int_{\Theta} L(\theta, \hat{\theta}(\mathrm{z})) \pi(\theta|\mathrm{z}) d\theta \right] m(\mathrm{z}) d\mathrm{z} = \int_{\mathcal{Z}} r(\hat{\theta}|\mathrm{z}) m(\mathrm{z}) d\mathrm{z}. \end{aligned}$$

Interpretation

- Loss $L(\theta, \hat{\theta}(Z))$ is a fucntion of θ and Z
- Risk R is the average L over data Z (given θ); it is a function of θ
- Posterior risk r is the average L over θ (given Z), it is a function of Z
- Bayes risk r_B is a single-number quantity
  - r_B is equal to the average R over θ (w.r.t. its prior distribution)
  - r_B is equal to the average r over z (w.r.t to its marginal distribution)

How to Compute the Bayes Rule

Using the connection between r_B and r, we can find $\hat{\theta}_{Bayes}$ by

$$\min_{\theta} r(\hat{\theta}|\mathrm{z}) = \int L(\theta, \hat{\theta}(\mathrm{z}))\pi(\theta|\mathrm{z})d\theta,$$

for each fixed z.

- Minimizing the posterior risk r for each z guarantees minimizing the integral r_B.

Examples of Optimal Bayes Rules

Theorem:

- If $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, then the posterior risk is

$$r(\hat{\theta}|\mathrm{z}) = \int [\theta - \hat{\theta}(\mathrm{z})]^2 \pi(\theta|\mathrm{z}) d\theta.$$

  Then the Bayes rule

$$\begin{aligned} \hat{\theta}_{\pi}^{Bayes}(\mathrm{z}) &= \arg\min r(\hat{\theta}|\mathrm{z}) \\ &= \int \theta \pi(\theta|\mathrm{z}) d\theta = E(\theta|\mathrm{Z} = \mathrm{z}), \end{aligned}$$

  which is the **posterior mean** of θ.
- If $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, then $\hat{\theta}_{\pi}^{Bayes}$ is the median of $\pi(\theta|\mathrm{z})$.
- If $L(\theta, \hat{\theta})$ is zero-one loss, then $\hat{\theta}_{\pi}^{Bayes}$ is the mode of $\pi(\theta|\mathrm{z})$.

Example 5: Normal Example

Let $Z_1, \cdots, Z_n \sim N(\mu, \sigma^2)$, where μ is unknown and σ^2 is known. Suppose the prior of μ is $N(a, b^2)$, where a and b are known.

- prior distribution: $\mu \sim N(a, b^2)$
- sampling distribution: $Z_1, \cdots, Z_n | \mu \sim N(\mu, \sigma^2)$.
- posterior distribution:

$$\mu | Z_1, \cdots, Z_n \sim N\left(\frac{b^2}{b^2 + \sigma^2/n} \bar{Z} + \frac{\sigma^2/n}{b^2 + \sigma^2/n} a, \left(\frac{1}{b^2} + \frac{n}{\sigma^2}\right)^{-1} \right)$$

Then the Bayes rule with respect to the squared error loss is

$$\hat{\theta}^{Bayes}(\mathrm{Z}) = E(\theta|\mathrm{Z}) = \frac{b^2}{b^2 + \sigma^2/n} \bar{Z} + \frac{\sigma^2/n}{b^2 + \sigma^2/n} a.$$

Example 6 (revisted Example 4): Binomial Risk

Let $Z_1, \cdots, Z_n \sim Bernoulli(p)$. Consider two estimators:

- $\hat{p}_1 = \bar{Z}$ (Maximum Likelihood Estimator, MLE).
- $\hat{p}_2 = \frac{\sum_{i=1}^n Z_i + \alpha}{\alpha + \beta + n}$ (Bayes estimator using a Beta(α, β) prior).

Using the squared error loss, direct calculation gives (Homework 2)

$$R(p, \hat{p}_1) = \frac{p(1-p)}{n}$$

$$R(p, \hat{p}_2) = \mathrm{V}_p(\hat{p}_2) + \mathrm{Bias}_p^2(\hat{p}_2) = \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2$$

Consider the special choice, $\alpha = \beta = \sqrt{n/4}$. Then

$$\hat{p}_2 = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}, \quad R(p, \hat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

Bayes Risk for Binomial Example

Assume the prior for p is $\pi(p) = 1$. Then

$$\begin{aligned} r_B(\pi, \hat{p}_1) &= \int_0^1 R(p, \hat{p}_1)dp = \int_0^1 \frac{p(1-p)}{n} dp = \frac{1}{6n}, \\ r_B(\pi, \hat{p}_2) &= \int_0^1 R(p, \hat{p}_2)dp = \frac{n}{4(n+\sqrt{n})^2}. \end{aligned}$$

If $n \geq 20$, then

- $r_B(\pi, \hat{p}_2) > r_B(\pi, \hat{p}_1)$, so $\hat{p}_1$ is better in terms of Bayes risk.
- This answer depends on the choice of prior.

In this case, the Minimax rule is $\hat{p}_2$ (shown in Example 4) and the Bayes rule under uniform prior is $\hat{p}_1$. They are different.