

Stat 709: Mathematical Statistics

Lecture 24

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Lecture 24: Bayes rules, minimax rules, point estimators, and hypothesis tests

Approaches

The second approach to finding a good decision rule is to consider some characteristic R_T of $R_T(P)$, for a given decision rule T , and then minimize R_T over $T \in \mathfrak{S}$.

The following are two popular ways to carry out this idea.

The first method: the Bayes rule

Consider an average of $R_T(P)$ over $P \in \mathcal{P}$:

$$r_T(\Pi) = \int_{\mathcal{P}} R_T(P) d\Pi(P),$$

where Π is a known probability measure on $(\mathcal{P}, \mathcal{F}_{\mathcal{P}})$ with an appropriate σ -field $\mathcal{F}_{\mathcal{P}}$.

$r_T(\Pi)$ is called the *Bayes risk* of T w.r.t. Π .

If $T_* \in \mathfrak{S}$ and $r_{T_*}(\Pi) \leq r_T(\Pi)$ for any $T \in \mathfrak{S}$, then T_* is called a \mathfrak{S} -*Bayes rule* (or Bayes rule when \mathfrak{S} contains all possible rules) w.r.t. Π .

Lecture 24: Bayes rules, minimax rules, point estimators, and hypothesis tests

Approaches

The second approach to finding a good decision rule is to consider some characteristic R_T of $R_T(P)$, for a given decision rule T , and then minimize R_T over $T \in \mathfrak{S}$.

The following are two popular ways to carry out this idea.

The first method: the Bayes rule

Consider an average of $R_T(P)$ over $P \in \mathcal{P}$:

$$r_T(\Pi) = \int_{\mathcal{P}} R_T(P) d\Pi(P),$$

where Π is a known probability measure on $(\mathcal{P}, \mathcal{F}_{\mathcal{P}})$ with an appropriate σ -field $\mathcal{F}_{\mathcal{P}}$.

$r_T(\Pi)$ is called the *Bayes risk* of T w.r.t. Π .

If $T_* \in \mathfrak{S}$ and $r_{T_*}(\Pi) \leq r_T(\Pi)$ for any $T \in \mathfrak{S}$, then T_* is called a \mathfrak{S} -*Bayes rule* (or Bayes rule when \mathfrak{S} contains all possible rules) w.r.t. Π .

The second method: the minimax rule

Consider the worst situation, i.e., $\sup_{P \in \mathcal{P}} R_T(P)$. If $T_* \in \mathfrak{S}$ and

$$\sup_{P \in \mathcal{P}} R_{T_*}(P) \leq \sup_{P \in \mathcal{P}} R_T(P)$$

for any $T \in \mathfrak{S}$, then T_* is called a \mathfrak{S} -*minimax* rule (or minimax rule when \mathfrak{S} contains all possible rules).

Example 2.25

Consider the estimation of $\theta \in \mathcal{R}$ under loss $L(\theta, a) = (\theta - a)^2$ and

$$r_T(\Pi) = \int_{\mathcal{R}} E[\theta - T(X)]^2 d\Pi(\theta),$$

which is equivalent to $E[\vec{\theta} - T(X)]^2$, where $\vec{\theta}$ is random and distributed as Π and, given $\vec{\theta} = \theta$, the conditional distribution of X is P_θ .

Then, the problem can be viewed as a prediction problem for $\vec{\theta}$ using functions of X .

Using the result in Example 1.22, the best predictor is $E(\vec{\theta}|X)$, which is the \mathfrak{S} -Bayes rule w.r.t. Π with \mathfrak{S} being the class of rules $T(X)$ satisfying $E[T(X)]^2 < \infty$ for any θ .

The second method: the minimax rule

Consider the worst situation, i.e., $\sup_{P \in \mathcal{P}} R_T(P)$. If $T_* \in \mathfrak{S}$ and

$$\sup_{P \in \mathcal{P}} R_{T_*}(P) \leq \sup_{P \in \mathcal{P}} R_T(P)$$

for any $T \in \mathfrak{S}$, then T_* is called a \mathfrak{S} -*minimax* rule (or minimax rule when \mathfrak{S} contains all possible rules).

Example 2.25

Consider the estimation of $\theta \in \mathcal{R}$ under loss $L(\theta, a) = (\theta - a)^2$ and

$$r_T(\Pi) = \int_{\mathcal{R}} E[\theta - T(X)]^2 d\Pi(\theta),$$

which is equivalent to $E[\vec{\theta} - T(X)]^2$, where $\vec{\theta}$ is random and distributed as Π and, given $\vec{\theta} = \theta$, the conditional distribution of X is P_θ .

Then, the problem can be viewed as a prediction problem for $\vec{\theta}$ using functions of X .

Using the result in Example 1.22, the best predictor is $E(\vec{\theta}|X)$, which is the \mathfrak{S} -Bayes rule w.r.t. Π with \mathfrak{S} being the class of rules $T(X)$ satisfying $E[T(X)]^2 < \infty$ for any θ .

Example 2.25 (continued)

As a more specific example, let $X = (X_1, \dots, X_n)$ with i.i.d. components having the $N(\mu, \sigma^2)$ distribution with an unknown $\mu = \theta \in \mathcal{R}$ and a known σ^2 , and let Π be the $N(\mu_0, \sigma_0^2)$ distribution with known μ_0 and σ_0^2 .

Then the conditional distribution of $\vec{\theta}$ given $X = x$ is $N(\mu_*(x), c^2)$ with

$$\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} \quad \text{and} \quad c^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

The Bayes rule w.r.t. Π is $E(\vec{\theta}|X) = \mu_*(X)$.

In this special case we can show that the sample mean \bar{X} is minimax. For any decision rule T ,

$$\begin{aligned} \sup_{\theta \in \mathcal{R}} R_T(\theta) &\geq \int_{\mathcal{R}} R_T(\theta) d\Pi(\theta) \geq \int_{\mathcal{R}} R_{\mu_*}(\theta) d\Pi(\theta) \\ &= E\left\{[\vec{\theta} - \mu_*(X)]^2\right\} = E\left\{E\{[\vec{\theta} - \mu_*(X)]^2|X\}\right\} = E(c^2) = c^2. \end{aligned}$$

Since this result is true for any $\sigma_0^2 > 0$ and $c^2 \rightarrow \sigma^2/n$ as $\sigma_0^2 \rightarrow \infty$,

Example 2.25 (continued)

As a more specific example, let $X = (X_1, \dots, X_n)$ with i.i.d. components having the $N(\mu, \sigma^2)$ distribution with an unknown $\mu = \theta \in \mathcal{R}$ and a known σ^2 , and let Π be the $N(\mu_0, \sigma_0^2)$ distribution with known μ_0 and σ_0^2 .

Then the conditional distribution of $\vec{\theta}$ given $X = x$ is $N(\mu_*(x), c^2)$ with

$$\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} \quad \text{and} \quad c^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

The Bayes rule w.r.t. Π is $E(\vec{\theta}|X) = \mu_*(X)$.

In this special case we can show that the sample mean \bar{X} is minimax. For any decision rule T ,

$$\begin{aligned} \sup_{\theta \in \mathcal{R}} R_T(\theta) &\geq \int_{\mathcal{R}} R_T(\theta) d\Pi(\theta) \geq \int_{\mathcal{R}} R_{\mu_*}(\theta) d\Pi(\theta) \\ &= E\left\{[\vec{\theta} - \mu_*(X)]^2\right\} = E\left\{E\{[\vec{\theta} - \mu_*(X)]^2|X\}\right\} = E(c^2) = c^2. \end{aligned}$$

Since this result is true for any $\sigma_0^2 > 0$ and $c^2 \rightarrow \sigma^2/n$ as $\sigma_0^2 \rightarrow \infty$,

Example 2.25 (continued)

$$\sup_{\theta \in \mathcal{R}} R_T(\theta) \geq \frac{\sigma^2}{n} = \sup_{\theta \in \mathcal{R}} R_{\bar{X}}(\theta),$$

where the equality holds because the risk of \bar{X} under the squared error loss is σ^2/n and independent of $\theta = \mu$.

Thus, \bar{X} is minimax.

It can be seen that if both μ and σ^2 are unknown in the previous discussion, then

$$\sup_{\mu \in \mathcal{R}, \sigma^2 \in (0, \infty)} R_{\bar{X}}(\mu, \sigma^2) = \infty.$$

Hence \bar{X} cannot be minimax.

We usually try to find a Bayes rule or a minimax rule in a parametric problem where $P = P_\theta$ for a $\theta \in \mathcal{R}^k$.

A minimax rule in general may be difficult to obtain.

Bayes and minimax rules are discussed in Chapter 4.

Example 2.25 (continued)

$$\sup_{\theta \in \mathcal{R}} R_T(\theta) \geq \frac{\sigma^2}{n} = \sup_{\theta \in \mathcal{R}} R_{\bar{X}}(\theta),$$

where the equality holds because the risk of \bar{X} under the squared error loss is σ^2/n and independent of $\theta = \mu$.

Thus, \bar{X} is minimax.

It can be seen that if both μ and σ^2 are unknown in the previous discussion, then

$$\sup_{\mu \in \mathcal{R}, \sigma^2 \in (0, \infty)} R_{\bar{X}}(\mu, \sigma^2) = \infty.$$

Hence \bar{X} cannot be minimax.

We usually try to find a Bayes rule or a minimax rule in a parametric problem where $P = P_\theta$ for a $\theta \in \mathcal{R}^k$.

A minimax rule in general may be difficult to obtain.

Bayes and minimax rules are discussed in Chapter 4.

Example 2.25 (continued)

$$\sup_{\theta \in \mathcal{R}} R_T(\theta) \geq \frac{\sigma^2}{n} = \sup_{\theta \in \mathcal{R}} R_{\bar{X}}(\theta),$$

where the equality holds because the risk of \bar{X} under the squared error loss is σ^2/n and independent of $\theta = \mu$.

Thus, \bar{X} is minimax.

It can be seen that if both μ and σ^2 are unknown in the previous discussion, then

$$\sup_{\mu \in \mathcal{R}, \sigma^2 \in (0, \infty)} R_{\bar{X}}(\mu, \sigma^2) = \infty.$$

Hence \bar{X} cannot be minimax.

We usually try to find a Bayes rule or a minimax rule in a parametric problem where $P = P_\theta$ for a $\theta \in \mathcal{R}^k$.

A minimax rule in general may be difficult to obtain.

Bayes and minimax rules are discussed in Chapter 4.

Statistical inference

Three components:

- Point estimators
- Hypothesis tests
- Confidence sets

Point estimators

Let $T(X)$ be an estimator of $\vartheta \in \mathcal{R}$

Bias: $b_T(P) = E[T(X)] - \vartheta$

Mean squared error (mse):

$$\text{mse}_T(P) = E[T(X) - \vartheta]^2 = [b_T(P)]^2 + \text{Var}(T(X)).$$

Bias and mse are two common criteria for the performance of point estimators.

Statistical inference

Three components:

- Point estimators
- Hypothesis tests
- Confidence sets

Point estimators

Let $T(X)$ be an estimator of $\vartheta \in \mathcal{R}$

Bias: $b_T(P) = E[T(X)] - \vartheta$

Mean squared error (mse):

$$\text{mse}_T(P) = E[T(X) - \vartheta]^2 = [b_T(P)]^2 + \text{Var}(T(X)).$$

Bias and mse are two common criteria for the performance of point estimators.

Example 2.26

Let X_1, \dots, X_n be i.i.d. from an unknown c.d.f. F .

Suppose that the parameter of interest is $\vartheta = 1 - F(t)$ for a fixed $t > 0$.

If F is not in a parametric family, then a *nonparametric* estimator of $F(t)$ is the *empirical* c.d.f.

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, t]}(X_i), \quad t \in \mathcal{R}.$$

Since $I_{(-\infty, t]}(X_1), \dots, I_{(-\infty, t]}(X_n)$ are i.i.d. binary random variables with $P(I_{(-\infty, t]}(X_i) = 1) = F(t)$, the random variable $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$.

Consequently, $F_n(t)$ is an unbiased estimator of $F(t)$ and

$$\text{Var}(F_n(t)) = \text{mse}_{F_n(t)}(P) = \frac{F(t)[1 - F(t)]}{n}.$$

An unbiased estimator of ϑ is $U(X) = 1 - F_n(t)$, which has the same variance and mse as $F_n(t)$.

Example 2.26

Let X_1, \dots, X_n be i.i.d. from an unknown c.d.f. F .

Suppose that the parameter of interest is $\vartheta = 1 - F(t)$ for a fixed $t > 0$.

If F is not in a parametric family, then a *nonparametric* estimator of $F(t)$ is the *empirical* c.d.f.

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, t]}(X_i), \quad t \in \mathcal{R}.$$

Since $I_{(-\infty, t]}(X_1), \dots, I_{(-\infty, t]}(X_n)$ are i.i.d. binary random variables with $P(I_{(-\infty, t]}(X_i) = 1) = F(t)$, the random variable $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$.

Consequently, $F_n(t)$ is an unbiased estimator of $F(t)$ and

$$\text{Var}(F_n(t)) = \text{mse}_{F_n(t)}(P) = \frac{F(t)[1 - F(t)]}{n}.$$

An unbiased estimator of ϑ is $U(X) = 1 - F_n(t)$, which has the same variance and mse as $F_n(t)$.

Example 2.26 (continued)

The estimator $U(X) = 1 - F_n(t)$ can be improved in terms of the mse if there is further information about F .

Suppose that F is the c.d.f. of the exponential distribution $E(0, \theta)$ with an unknown $\theta > 0$.

Then $\vartheta = e^{-t/\theta}$.

The sample mean \bar{X} is sufficient for $\theta > 0$.

Since the squared error loss is strictly convex, an application of Theorem 2.5(ii) (Rao-Blackwell theorem) shows that the estimator $T(X) = E[1 - F_n(t) | \bar{X}]$, which is also unbiased, is better than $U(X)$ in terms of the mse.

Figure 2.1 shows graphs of the mse's of $U(X)$ and $T(X)$, as functions of θ , in the special case of $n = 10$, $t = 2$, and

$$F(x) = (1 - e^{-x/\theta})I_{(0, \infty)}(x).$$

Example 2.26 (continued)

The estimator $U(X) = 1 - F_n(t)$ can be improved in terms of the mse if there is further information about F .

Suppose that F is the c.d.f. of the exponential distribution $E(0, \theta)$ with an unknown $\theta > 0$.

Then $\vartheta = e^{-t/\theta}$.

The sample mean \bar{X} is sufficient for $\theta > 0$.

Since the squared error loss is strictly convex, an application of Theorem 2.5(ii) (Rao-Blackwell theorem) shows that the estimator $T(X) = E[1 - F_n(t) | \bar{X}]$, which is also unbiased, is better than $U(X)$ in terms of the mse.

Figure 2.1 shows graphs of the mse's of $U(X)$ and $T(X)$, as functions of θ , in the special case of $n = 10$, $t = 2$, and

$$F(x) = (1 - e^{-x/\theta}) I_{(0, \infty)}(x).$$

Hypothesis tests

To test the hypotheses

$$H_0 : P \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P \in \mathcal{P}_1,$$

there are two types of errors we may commit:

- rejecting H_0 when H_0 is true (called the *type I error*)
- and accepting H_0 when H_0 is wrong (called the *type II error*).

A test T : a statistic from \mathcal{X} to $\{0, 1\}$.

Probabilities of making two types of errors

Type I error rate:

$$\alpha_T(P) = P(T(X) = 1) \quad P \in \mathcal{P}_0$$

Type II error rate:

$$1 - \alpha_T(P) = P(T(X) = 0) \quad P \in \mathcal{P}_1,$$

$\alpha_T(P)$ is also called the power function of T

Power function is $\alpha_T(\theta)$ if P is in a parametric family indexed by θ .

Hypothesis tests

To test the hypotheses

$$H_0 : P \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P \in \mathcal{P}_1,$$

there are two types of errors we may commit:

- rejecting H_0 when H_0 is true (called the *type I error*)
- and accepting H_0 when H_0 is wrong (called the *type II error*).

A test T : a statistic from \mathcal{X} to $\{0, 1\}$.

Probabilities of making two types of errors

Type I error rate:

$$\alpha_T(P) = P(T(X) = 1) \quad P \in \mathcal{P}_0$$

Type II error rate:

$$1 - \alpha_T(P) = P(T(X) = 0) \quad P \in \mathcal{P}_1,$$

$\alpha_T(P)$ is also called the power function of T

Power function is $\alpha_T(\theta)$ if P is in a parametric family indexed by θ .

Remarks

- Note that these are risks of T under the 0-1 loss in statistical decision theory.
- Type I and type II error probabilities cannot be minimized simultaneously.
- These two error probabilities cannot be bounded simultaneously by a fixed $\alpha \in (0, 1)$ when we have a sample of a fixed size.

Significance tests

A common approach to finding an "optimal" test is to assign a small bound α to the type I error rate $\alpha_T(P)$, $P \in \mathcal{P}_0$, and then to attempt to minimize the type II error rate $1 - \alpha_T(P)$, $P \in \mathcal{P}_1$, subject to

$$\sup_{P \in \mathcal{P}_0} \alpha_T(P) \leq \alpha. \quad (1)$$

The bound α is called the *level of significance*.

The left-hand side of (1) is called the *size of the test* T .

The level of significance should be positive, otherwise no test satisfies (1) except the silly test $T(X) \equiv 0$ a.s. \mathcal{P} .

Remarks

- Note that these are risks of T under the 0-1 loss in statistical decision theory.
- Type I and type II error probabilities cannot be minimized simultaneously.
- These two error probabilities cannot be bounded simultaneously by a fixed $\alpha \in (0, 1)$ when we have a sample of a fixed size.

Significance tests

A common approach to finding an "optimal" test is to assign a small bound α to the type I error rate $\alpha_T(P)$, $P \in \mathcal{P}_0$, and then to attempt to minimize the type II error rate $1 - \alpha_T(P)$, $P \in \mathcal{P}_1$, subject to

$$\sup_{P \in \mathcal{P}_0} \alpha_T(P) \leq \alpha. \quad (1)$$

The bound α is called the *level of significance*.

The left-hand side of (1) is called the *size of the test* T .

The level of significance should be positive, otherwise no test satisfies (1) except the silly test $T(X) \equiv 0$ a.s. \mathcal{P} .

Example 2.28

Let X_1, \dots, X_n be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2 .

Consider the hypotheses $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, where μ_0 is a fixed constant.

Since the sample mean \bar{X} is sufficient for $\mu \in \mathcal{R}$, it is reasonable to consider the following class of tests: $T_c(X) = I_{(c, \infty)}(\bar{X})$, i.e., H_0 is rejected (accepted) if $\bar{X} > c$ ($\bar{X} \leq c$), where $c \in \mathcal{R}$ is a fixed constant.

Let Φ be the c.d.f. of $N(0, 1)$.

By the property of the normal distributions,

$$\alpha_{T_c}(\mu) = P(T_c(X) = 1) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).$$

Figure 2.2 provides an example of a graph of two types of error probabilities, with $\mu_0 = 0$.

Since $\Phi(t)$ is an increasing function of t ,

$$\sup_{\mu \in \mathcal{P}_0} \alpha_{T_c}(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right).$$

Example 2.28

Let X_1, \dots, X_n be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2 .

Consider the hypotheses $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, where μ_0 is a fixed constant.

Since the sample mean \bar{X} is sufficient for $\mu \in \mathcal{R}$, it is reasonable to consider the following class of tests: $T_c(X) = I_{(c, \infty)}(\bar{X})$, i.e., H_0 is rejected (accepted) if $\bar{X} > c$ ($\bar{X} \leq c$), where $c \in \mathcal{R}$ is a fixed constant.

Let Φ be the c.d.f. of $N(0, 1)$.

By the property of the normal distributions,

$$\alpha_{T_c}(\mu) = P(T_c(X) = 1) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).$$

Figure 2.2 provides an example of a graph of two types of error probabilities, with $\mu_0 = 0$.

Since $\Phi(t)$ is an increasing function of t ,

$$\sup_{P \in \mathcal{P}_0} \alpha_{T_c}(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right).$$

Example 2.28 (continued)

In fact, it is also true that

$$\sup_{P \in \mathcal{P}_1} [1 - \alpha_{T_c}(\mu)] = \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right).$$

If we would like to use an α as the level of significance, then the most effective way is to choose a c_α (a test $T_{c_\alpha}(X)$) such that

$$\alpha = \sup_{P \in \mathcal{P}_0} \alpha_{T_{c_\alpha}}(\mu),$$

in which case c_α must satisfy

$$1 - \Phi\left(\frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma}\right) = \alpha,$$

i.e., $c_\alpha = \sigma z_{1-\alpha} / \sqrt{n} + \mu_0$, where $z_a = \Phi^{-1}(a)$.

In Chapter 6, it is shown that for any test $T(X)$ satisfying

$$\sup_{P \in \mathcal{P}_0} \alpha_T(P) \leq \alpha,$$

$$1 - \alpha_T(\mu) \geq 1 - \alpha_{T_{c_\alpha}}(\mu), \quad \mu > \mu_0.$$