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

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 $T \in \Im$.
The following are two popular ways to carry out this idea.
The first one is to 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 $T_* \in \Im$ and $r_{T_*}(\Pi) \le r_T(\Pi)$ for any $T \in \Im$, then T_* is called a $\Im$-*Bayes rule* (or Bayes rule when $\Im$ contains all possible rules) w.r.t. Π.
The second method is to consider the worst situation, i.e., $\sup_{P\in\mathcal{P}} R_T(P)$.
If $T_* \in \Im$ and

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

for any $T \in \Im$, then T_* is called a $\Im$-*minimax* rule (or minimax rule when $\Im$ contains all possible rules).
Bayes and minimax rules are discussed in Chapter 4.

Example 2.25. We usually try to find a Bayes rule or a minimax rule in a parametric problem where $P = P_\theta$ for a $\theta \in \mathcal{R}^k$.
Consider the special case of $k = 1$ and $L(\theta, a) = (\theta - a)^2$, the squared error loss.
Note that

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

which is equivalent to $E[\boldsymbol{\theta} - T(X)]^2$, where $\boldsymbol{\theta}$ is a random variable having the distribution Π and, given $\boldsymbol{\theta} = \theta$, the conditional distribution of X is P_θ.
Then, the problem can be viewed as a prediction problem for $\boldsymbol{\theta}$ using functions of X.
Using the result in Example 1.22, the best predictor is $E(\boldsymbol{\theta}|X)$, which is the $\Im$-Bayes rule w.r.t. Π with $\Im$ being the class of rules $T(X)$ satisfying $E[T(X)]^2 < \infty$ for any θ.

As a more specific example, let $X = (X_1, ..., X_n)$ 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 $\boldsymbol{\theta}$ given $X = x$ is $N(\mu_*(x), c^2)$ with

$$\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} \qquad \text{and} \qquad c^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{1}$$

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

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

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

where $\mu_*(X)$ is the Bayes rule given in (1) and c^2 is also given in (1).
Since this result is true for any $\sigma_0^2 > 0$ and $c^2 \to \sigma^2/n$ as $\sigma_0^2 \to \infty$,

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

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.

A minimax rule in a general case may be difficult to obtain. It can be seen that if both μ and σ^2 are unknown in the previous discussion, then

$$\sup_{\theta\in\mathcal{R}\times(0,\infty)} R_{\bar{X}}(\theta) = \infty, \tag{2}$$

where $\theta = (\mu, \sigma^2)$.
Hence $\bar{X}$ cannot be minimax unless (2) holds with $\bar{X}$ replaced by any decision rule T, in which case minimaxity becomes meaningless.

Statistical inference: Point estimators, hypothesis tests, and 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 $X_1, ..., X_n$ 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.

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

Since $I_{(-\infty,t]}(X_1), ..., I_{(-\infty,t]}(X_n)$ 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 $\text{Var}(F_n(t)) = \text{mse}_{F_n(t)}(P) = F(t)[1 - F(t)]/n$.
Since any linear combination of unbiased estimators is unbiased for the same linear combination of the parameters (by the linearity of expectations), an unbiased estimator of ϑ is $U(X) = 1 - F_n(t)$, which has the same variance and mse as $F_n(t)$.
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 statistical 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 $\{0, 1\}$. Pprobabilities of making two types of errors:

$$\alpha_T(P) = P(T(X) = 1) \qquad P \in \mathcal{P}_0 \tag{3}$$

and

$$1 - \alpha_T(P) = P(T(X) = 0) \qquad P \in \mathcal{P}_1, \tag{4}$$

which are denoted by $\alpha_T(\theta)$ and $1 - \alpha_T(\theta)$ if P is in a parametric family indexed by θ.
Note that these are risks of T under the 0-1 loss in statistical decision theory.
Error probabilities in (3) and (4) cannot be minimized simultaneously.
Furthermore, these two error probabilities cannot be bounded simultaneously by a fixed $\alpha \in (0, 1)$ when we have a sample of a fixed size.

A common approach to finding an "optimal" test is to assign a small bound α to one of the error probabilities, say $\alpha_T(P)$, $P \in \mathcal{P}_0$, and then to attempt to minimize the other error probability $1 - \alpha_T(P)$, $P \in \mathcal{P}_1$, subject to

$$\sup_{P \in \mathcal{P}_0} \alpha_T(P) \le \alpha. \tag{5}$$

The bound α is called the *level of significance*.
The left-hand side of (5) is called the *size* of the test T.

The level of significance should be positive, otherwise no test satisfies (5) except the silly test $T(X) \equiv 0$ a.s. $\mathcal{P}$.

Example 2.28. Let $X_1, ..., X_n$ 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 $N(0,1)$. Then, 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,

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

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 (5),

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