Lecture 13: Statistical decision and inference

Basic elements

- X: a sample from a population $P \in \mathscr{P}$
- Decision: an action we take after observing X
- $\mathscr{A}$: the set of allowable actions
- $(\mathscr{A}, \mathscr{F}_{\mathscr{A}})$: the action space
- $\mathscr{X}$: the range of X
- Decision rule: a statistic $T(X)$ from $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$ to $(\mathscr{A}, \mathscr{F}_{\mathscr{A}})$
- If X is observed, then we take the action $T(X) \in \mathscr{A}$
- Performance: loss function $L(P,a)$ from $\mathscr{P} \times \mathscr{A}$ to $[0,\infty)$, Borel in a
  If our action is $T(X)$, then our "loss" is $L(P,T(X))$
  It is difficult to assess $L(P,T(X))$ since it is random.
- Risk: the average (expected) loss defined as

$$R_T(P) = E[L(P,T(X))] = \int_{\mathscr{X}} L(P,T(x))dP_X(x).$$

If $\mathscr{P}$ is parametric, the loss and risk are denoted by $L(\theta,a)$, $R_T(\theta)$

Comparisons

- For decision rules T_1 and T_2, T_1 is *as good as* T_2 iff
$$R_{T_1}(P) \le R_{T_2}(P) \quad \text{for any } P \in \mathscr{P},$$
and is *better* than T_2 if, in addition, $R_{T_1}(P) < R_{T_2}(P)$ for some P.
- T_1 and T_2 are *equivalent* iff $R_{T_1}(P) = R_{T_2}(P)$ for all $P \in \mathscr{P}$.
- Optimal rule: If T_* is as good as any other rule in $\mathfrak{I}$, a class of allowable decision rules, then T_* is $\mathfrak{I}$-*optimal* (or optimal if $\mathfrak{I}$ contains all possible rules).

Randomized decision rules

A function δ on $\mathscr{X} \times \mathscr{F}_{\mathscr{A}}$; for every $A \in \mathscr{F}_{\mathscr{A}}$, $\delta(\cdot, A)$ is a Borel function and, for every $x \in \mathscr{X}$, $\delta(x, \cdot)$ is a probability measure on $(\mathscr{A}, \mathscr{F}_{\mathscr{A}})$.

- If $X = x$ is observed, we have a distribution of actions: $\delta(x, \cdot)$.
- A nonrandomized decision rule T is a special randomized decision rule with $\delta(x, \{a\}) = I_{\{a\}}(T(x))$, $a \in \mathscr{A}$, $x \in \mathscr{X}$.
- An example is a discrete distribution $\delta(x, \cdot)$ assigning probability $p_j(x)$ to a nonrandomized decision rule $T_j(x)$, $j = 1, 2, \ldots$

Loss and risk of a randomized decision rule

The loss function for a randomized rule δ is defined as

$$L(P,\delta,x) = \int_{\mathscr{A}} L(P,a)d\delta(x,a),$$

which reduces to the same loss function when δ is nonrandomized.
The risk of a randomized rule δ is then

$$R_\delta(P) = E[L(P,\delta,X)] = \int_{\mathscr{X}} \int_{\mathscr{A}} L(P,a)d\delta(x,a)dP_X(x).$$

Example 2.19

$X = (X_1,...,X_n)$ is a vector of iid measurements for a parameter $\theta \in \mathscr{R}$.
We want to estimate θ.
Action space: $(\mathscr{A},\mathscr{F}_{\mathscr{A}}) = (\mathscr{R},\mathscr{B})$.
A common loss function in this problem is the *squared error loss* $L(P,a) = (\theta - a)^2$, $a \in \mathscr{A}$.
Let $T(X) = \bar{X}$, the sample mean.
The loss for $\bar{X}$ is $(\bar{X} - \theta)^2$.

If the population has mean μ and variance $\sigma^2 < \infty$, then

$$R_{\bar{X}}(P) = E(\theta - \bar{X})^2 = (\theta - E\bar{X})^2 + E(E\bar{X} - \bar{X})^2$$
$$= (\theta - E\bar{X})^2 + \operatorname{Var}(\bar{X}) = (\mu - \theta)^2 + \frac{\sigma^2}{n}.$$

If θ is in fact the mean of the population, then

$$R_{\bar{X}}(P) = \frac{\sigma^2}{n}$$

is an increasing function of the population variance σ^2 and a decreasing function of the sample size n.
Consider another decision rule $T_1(X) = (X_{(1)} + X_{(n)})/2$.
$R_{T_1}(P)$ does not have a simple explicit form if there is no further assumption on the family $\mathscr{P}$ containing P.
For some $\mathscr{P}$, $\bar{X}$ (or T_1) is better than T_1 (or $\bar{X}$) (exercise), whereas for some $\mathscr{P}$, neither $\bar{X}$ nor T_1 is better than the other.

The problem in Example 2.19 is a special case of a general problem called *estimation*.
In an estimation problem, a decision rule T is called an *estimator*.

The following example describes another type of important problem called *hypothesis testing*.

Example 2.20

Let $\mathscr{P}$ be a family of distributions, $\mathscr{P}_0 \subset \mathscr{P}$, $\mathscr{P}_1 = \{P \in \mathscr{P} : P \notin \mathscr{P}_0\}$.
A hypothesis testing problem can be formulated as that of deciding which of the following two statements is true:

$$H_0: \ P \in \mathscr{P}_0 \qquad \text{versus} \qquad H_1: \ P \in \mathscr{P}_1.$$

H_0 is called the *null hypothesis* and H_1 is the *alternative hypothesis*.
The action space for this problem contains only two elements, i.e., $\mathscr{A} = \{0, 1\}$, where 0 is accepting H_0 and 1 is rejecting H_0.
A decision rule is called a *test*, which must have the form $I_C(X)$, where $C \in \mathscr{F}_{\mathscr{X}}$ is called the *rejection* or *critical region*.
0-1 loss: $L(P, a) = 0$ if a correct decision is made and 1 if an incorrect decision is made, which leads to the risk

$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & P \in \mathscr{P}_0 \\ P(T(X) = 0) = P(X \notin C) & P \in \mathscr{P}_1. \end{cases}$$

May use unequal losses: $L(P, j) = 0$, $P \in \mathscr{P}_j$, $L(P, j) = c_j$, $P \in \mathscr{P}_{1-j}$

Definition 2.7 (Admissibility)

Let $\Im$ be a class of decision rules (randomized or nonrandomized). A decision rule $T \in \Im$ is called $\Im$-*admissible* (or admissible when $\Im$ contains all possible rules) iff there does not exist any $S \in \Im$ that is better than T (in terms of the risk).

Remarks

- If a decision rule T is inadmissible, then there exists a rule better than T and T should not be used in principle.
- However, an admissible decision rule is not necessarily good. For example, in an estimation problem a silly estimator $T(X) \equiv$ a constant may be admissible.
- If T_* is $\Im$-optimal, then it is $\Im$-admissible.
- If T_* is $\Im$-optimal and T_0 is $\Im$-admissible, then T_0 is also $\Im$-optimal and is equivalent to T_*.
- If there are two $\Im$-admissible rules that are not equivalent, then there does not exist any $\Im$-optimal rule.
- How to check admissibility will be discussed in Chapter 4

Suppose that we have a sufficient statistic $T(X)$ for $P \in \mathscr{P}$.
Intuitively, our decision rule should be a function of T.
This is not true in general, but the following result indicates that this is true if randomized decision rules are allowed.

Proposition 2.2

Let $T(X)$ be a sufficient statistic for $P \in \mathscr{P}$ and let δ_0 be a decision rule. Then

$$\delta_1(t, A) = E[\delta_0(X, A)|T = t],$$

which is a randomized decision rule depending only on T, is equivalent to δ_0 if $R_{\delta_0}(P) < \infty$ for any $P \in \mathscr{P}$.

Note that Proposition 2.2 does not imply that δ_0 is inadmissible.
If δ_0 is a nonrandomized rule,

$$\delta_1(t, A) = E[I_A(\delta_0(X))|T = t] = P(\delta_0(X) \in A|T = t)$$

is still a randomized rule, unless $\delta_0(X) = h(T(X))$ a.s. P for some h

The following result tells us when nonrandomized rules are all we need and when decision rules that are not functions of sufficient statistics are inadmissible.

Theorem 2.5

Suppose that $\mathscr{A}$ is a convex subset of $\mathscr{R}^k$ and that for any $P \in \mathscr{P}$, $L(P,a)$ is a convex function of a.

(i) Let δ be a randomized rule satisfying $\int_{\mathscr{A}} \|a\| d\delta(x,a) < \infty$ for any $x \in \mathscr{X}$ and let $T_1(x) = \int_{\mathscr{A}} a d\delta(x,a)$.
Then $L(P,T_1(x)) \leq L(P,\delta,x)$ (or $L(P,T_1(x)) < L(P,\delta,x)$ if L is strictly convex in a) for any $x \in \mathscr{X}$ and $P \in \mathscr{P}$.

(ii) (Rao-Blackwell theorem). Let T be a sufficient statistic for $P \in \mathscr{P}$, $T_0 \in \mathscr{R}^k$ be a nonrandomized rule satisfying $E\|T_0\| < \infty$, and $T_1 = E[T_0(X)|T]$.
Then $R_{T_1}(P) \leq R_{T_0}(P)$ for any $P \in \mathscr{P}$.
If L is strictly convex in a and T_0 is not a function of T, then T_0 is inadmissible.

The proof of Theorem 2.5 is an application of Jensen's inequality.

The concept of admissibility and sufficiency helps us to eliminate some decision rules, but usually there are still too many rules left.
A $\mathfrak{I}$-optimal rule often does not exist, if $\mathfrak{I}$ is too large or too small.

Example 2.22 (finding a decision rule?)

Let $X_1,...,X_n$ be i.i.d. random variables from a population $P \in \mathscr{P}$ that is the family of populations having finite mean μ and variance σ^2.

Consider the estimation of μ ($\mathscr{A} = \mathscr{R}$) under the squared error loss. It can be shown that if we let $\mathfrak{I}$ be the class of all possible estimators, then there is no $\mathfrak{I}$-optimal rule (exercise).

Next, let $\mathfrak{I}_1$ be the class of all linear functions in $X = (X_1,...,X_n)$, i.e., $T(X) = \sum_{i=1}^n c_i X_i$ with known $c_i \in \mathscr{R}$, $i = 1,...,n$.

$$R_T(P) = \mu^2 \left(\sum_{i=1}^n c_i - 1 \right)^2 + \sigma^2 \sum_{i=1}^n c_i^2. \qquad (1)$$

If there is a $\mathfrak{I}_1$-optimal rule T_*, then $(c_1^*,...,c_n^*)$ is a minimum of the function of $(c_1,...,c_n)$ on the right-hand side of (1).
Then $c_1^*,...,c_n^*$ must be the same and equal to $\mu^2/(\sigma^2 + n\mu^2)$, which depends on P, i.e., T_* is not a statistic.

Consider now a subclass $\mathfrak{I}_2 \subset \mathfrak{I}_1$ with c_i's satisfying $\sum_{i=1}^n c_i = 1$.
From (1), $R_T(P) = \sigma^2 \sum_{i=1}^n c_i^2$ if $T \in \mathfrak{I}_2$.
Minimizing $\sigma^2 \sum_{i=1}^n c_i^2$ subject to $\sum_{i=1}^n c_i = 1$ leads to $c_i = n^{-1}$.

Thus, the sample mean $\bar{X}$ is $\Im_2$-optimal.

There may not be any optimal rule if we consider a small class of rules.

In view of the fact that an optimal rule often does not exist, statisticians adopt two approaches to choose a decision rule.

Approach I

Define a class $\Im$ of decision rules that have some desirable properties (statistical and/or nonstatistical) and then try to find the best rule in $\Im$.

In Example 2.22, for instance, any estimator T in $\Im_2$ has the property that T is linear in X and $E[T(X)] = \mu$.

In a general estimation problem, we can use the following concept.

Definition 2.8 (Unbiasedness)

In an estimation problem, the *bias* of an estimator $T(X)$ of a parameter ϑ of the unknown population is defined to be

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

(denoted by $b_T(\theta)$ when P is in a parametric family indexed by θ).
An estimator $T(X)$ is *unbiased* for ϑ iff $b_T(P) = 0$ for any $P \in \mathscr{P}$.

Remarks

- $\Im_2$ in Example 2.22 is the class of unbiased estimators linear in X.
- In Chapter 3, we study how to find a $\Im$-optimal estimator when $\Im$ contains unbiased estimators or unbiased estimators linear in X.
- Another property we may consider is *invariance* (see textbook).

Approach II

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 first method: the Bayes rule

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

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

where Π is a known probability measure on $(\mathscr{P}, \mathscr{F}_{\mathscr{P}})$.

$r_T(\Pi)$ 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: the minimax rule

Consider the worst situation, i.e., $\sup_{P\in\mathscr{P}} R_T(P)$.

If $T_* \in \Im$ and

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

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

Example 2.25

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

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

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.

The problem becomes to predict $\vec{\theta}$ using functions of X.

Using the result in Example 1.22, the best predictor is $E(\vec{\theta}|X)$, 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.

More on Bayes and minimax rules will be studied in Chapter 4.

Statistical inference

A major approach in statistical analysis does not use any loss-risk.

Three components in statistical inference

- Point estimators (Chapters 3-5)
- Hypothesis tests (Chapter 6)
- Confidence sets (Chapter 7)

Point estimators

Let $T(X)$ be an estimator of $\vartheta \in \mathscr{R}$
Bias: $b_T(P) = E[T(X)] - \vartheta$
Mean squared error (mse):

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

Bias and mse are two common criteria for the performance of point estimators, i.e., instead of considering risk functions, we use bias and mse to evaluate point estimators.

Read Example 2.26

Hypothesis tests

To test the hypotheses

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

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 $\mathscr{X}$ to $\{0,1\}$.

Probabilities of making two types of errors

Type I error rate:

$$\alpha_T(P) = P(T(X) = 1) \qquad P \in \mathscr{P}_0$$

Type II error rate:

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

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

$$\sup_{P\in\mathscr{P}_0} \alpha_T(P) \leq \alpha. \tag{2}$$

The bound α is called the *level of significance*.
The left-hand side of (2) is called the *size* of the test T.
The level of significance should be positive, otherwise no test satisfies (2) except the silly test $T(X) \equiv 0$ a.s. $\mathscr{P}$.

Confidence sets

ϑ: a k-vector of unknown parameters related to the unknown $P \in \mathscr{P}$

If a Borel set $C(X)$ (in the range of ϑ) depending only on the sample X such that

$$\inf_{P \in \mathscr{P}} P(\vartheta \in C(X)) \geq 1 - \alpha, \qquad (3)$$

where α is a fixed constant in $(0,1)$, then $C(X)$ is called a *confidence set* for ϑ with *level of significance* $1 - \alpha$.

The left-hand side of (3) is called the *confidence coefficient* of $C(X)$, which is the highest possible level of significance for $C(X)$.

A confidence set is a random element that covers the unknown ϑ with certain probability.

If (3) holds, then the *coverage probability* of $C(X)$ is at least $1 - \alpha$, although $C(x)$ either covers or does not cover ϑ whence we observe $X = x$.

The concepts of level of significance and confidence coefficient are very similar to the level of significance and size in hypothesis testing.

Confidence sets are closely related to hypothesis tests (Chapter 7).

Example 2.32

Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with both $\mu \in \mathscr{R}$ and $\sigma^2 > 0$ unknown.
Let $\theta = (\mu, \sigma^2)$ and $\alpha \in (0,1)$ be given.
Let $\bar{X}$ be the sample mean and S^2 be the sample variance.
Since $(\bar{X}, S^2)$ is sufficient (Example 2.15), we focus on $C(X)$ that is a function of $(\bar{X}, S^2)$.

From Example 2.18, $\bar{X}$ and S^2 are independent and $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}.
Since $\sqrt{n}(\bar{X}-\mu)/\sigma$ has the $N(0,1)$ distribution,

$$P\left(-\tilde{c}_\alpha \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \tilde{c}_\alpha\right) = \sqrt{1-\alpha},$$

where $\tilde{c}_\alpha = \Phi^{-1}\left(\frac{1+\sqrt{1-\alpha}}{2}\right)$ (verify).
Since the chi-square distribution χ^2_{n-1} is a known distribution, we can always find two constants $c_{1\alpha}$ and $c_{2\alpha}$ such that

Example 2.32 (continued)

$$P\left(c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = \sqrt{1-\alpha}.$$

Then

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha, c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = 1-\alpha,$$

or

$$P\left(\frac{n(\bar{X}-\mu)^2}{\tilde{c}_\alpha^2} \le \sigma^2, \frac{(n-1)S^2}{c_{2\alpha}} \le \sigma^2 \le \frac{(n-1)S^2}{c_{1\alpha}}\right) = 1-\alpha. \quad (4)$$

The left-hand side of (4) defines a set in the range of $\theta = (\mu, \sigma^2)$ bounded by two straight lines, $\sigma^2 = (n-1)S^2/c_{i\alpha}$, $i = 1, 2$, and a curve $\sigma^2 = n(\bar{X}-\mu)^2/\tilde{c}_\alpha^2$ (see the shadowed part of Figure 2.3).
This set is a confidence set for θ with confidence coefficient $1-\alpha$, since (4) holds for any θ.