Contents

1 Statistical inference **1**

2 The estimation problem **2**

3 The testing problem **4**

4 Loss, decision rules and risk **6**

4.1 Statistical decision problems 7

4.2 Decision rules and risk 8

4.3 Why risk? 8

5 Statistical decision theory **11**

This material is similar to that in Lehmann and Casella [1998], section 1.1 and Ferguson [1967], sections 1.1-1.4.

1 Statistical inference

$$\begin{aligned} &X \sim P, && P \in \mathcal{P}, && \text{infer } P \text{ from } X \\ &X \sim P_\theta, && \theta \in \Theta, && \text{infer } \theta \text{ from } X \end{aligned}$$

This is *induction*: reasoning from the specific (X) to the general (θ).

Examples:

1. survey sampling:
   - θ: a population characteristic
   - P_θ: a distribution depending on θ and the sampling mechanism
   - X: a sample characteristic
2. experiment
   - θ: a physical quantity
   - P_θ: a distribution depending on θ and the measurement process
   - X: a measurement

In both cases, the goal is to infer something about θ from X.

2 The estimation problem

Data : $X \in \mathcal{X}$, to-be observed (e.g. $X = (X_1, \ldots, X_n)$).

Model : $X \sim P_\theta$, $\theta \in \Theta$ (e.g. $X_1, \ldots, X_n \sim$ i.i.d. $p(x|\theta)$).

Estimand : $g(\theta)$, some known function of θ (e.g., $g(\theta) = \theta$ or $g(\theta) = \int h(x) P_\theta(dx)$).

Goal : Identify a good estimator $\delta(x)$ of $g(\theta)$.

What is an estimator δ?

δ is a $\sigma(\mathcal{X})$-measurable function.

$\delta(\cdot)$ is the estimator.

$\delta(x)$ is the estimate when $X = x$.

Ideally, $\delta(X)$ is "close" to $g(\theta)$ when $X \sim P_\theta$.

Example (mean estimation):

$$X = (X_1, \ldots, X_n),\ X_1, \ldots, X_n \sim \text{i.i.d.} P_\theta$$
$$\mu(\theta) = \int x P_\theta(dx) = \text{population mean}$$

Some estimators:

$\delta_1(X) = \bar{X}$

$\delta_2(X) = \frac{n}{n+1}\bar{X} + \frac{1}{n+1}\mu_0$

$\delta_3(X) = \mu_0$

Will any of these be "close" to θ? How do we define "close"?

$$\begin{aligned}\text{MSE}(\theta, \delta) &= \int (\delta(X) - \mu(\theta))^2 P_\theta(dX) \\ &= \text{E}_{X|\theta}[(\delta(X) - \mu(\theta))^2] \\ &= \text{E}_{X|\theta}[(\delta(X) - \text{E}_{X|\theta}[\delta(X)])^2] + (\text{E}_{X|\theta}[\delta(X)] - \mu(\theta))^2 \\ &= \text{Var}_{X|\theta}[\delta(X)] + \text{Bias}^2_{X|\theta}[\delta(X)]\end{aligned}$$

MSE is the average squared distance between the estimator and the estimand, where the "average" is with respect to the population P_θ.

If $\int X^2 P_\theta(dX) < \infty$ let $\sigma^2(\theta) = \text{Var}_{X|\theta}[X]$. The MSE of $\delta_w(X) = w\bar{X} + (1-w)\mu_0$ is

$$\text{MSE}(\theta, \delta_w) = w^2 \frac{\sigma^2(\theta)}{n} + (1-w)^2(\mu(\theta) - \mu_0)^2$$

See Figure 1. For the three estimators above, we have

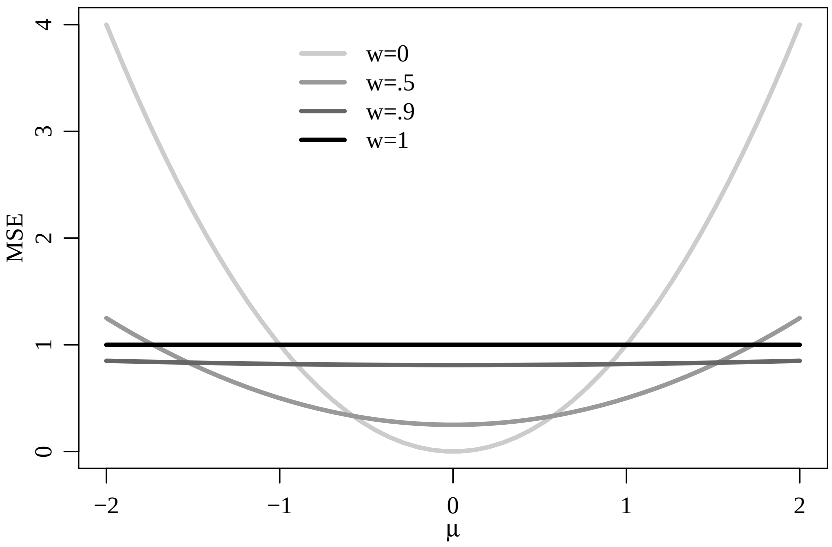

Figure 1: Some MSE functions when $\sigma^2(\theta)/n = 1$, constant for all μ.

$\text{MSE}(\theta, \delta_1) = \frac{\sigma^2(\theta)}{n}$

$\text{MSE}(\theta, \delta_2) = (\frac{n}{n+1})^2 \frac{\sigma^2(\theta)}{n} + \frac{1}{(n+1)^2}(\mu(\theta) - \mu_0)^2.$

$\text{MSE}(\theta, \delta_3) = (\mu(\theta) - \mu_0)^2$

Discuss: How do these estimators differ asymptotically? When would each be appropriate? How would you pick w?

3 The testing problem

Data : $X \in \mathcal{X}$, to-be observed (e.g. $X = (X_1, \ldots, X_n)$).

Model : $X \sim P_\theta$, $\theta \in \Theta$ (e.g. $X_1, \ldots, X_n \sim$ i.i.d. $p(x|\theta)$).

Hypotheses : $\mathrm{H}_0 : \theta \in \Theta_0$, $\mathrm{H}_1 : \theta \notin \Theta_0$.

Goal : Identify a good test function $\delta(X)$ of H_0 versus H_1.

What is a test function?

$\delta : \mathcal{X} \to [0,1]$.

$\delta(x)$ is the probability with which you reject H_0 and accept H_1 when $X = x$.

A nonrandomized test is one for which $\delta(X) \in \{0,1\}$ with probability 1.

Ideally, $\delta(X)$ is small (with high probability) when $\theta \in \Theta_0$, and large (with high probability) when $\theta \in \Theta_0$.

Example (simple versus simple hypotheses):

$$X_1, \ldots, X_n \sim \text{ i.i.d. } p_\theta(x), \ \theta \in \{0,1\}$$

- p_0 is the standard normal density ($\mathrm{H}_0 : \theta = 0$);
- p_1 is the standard Cauchy density ($\mathrm{H}_1 : \theta = 1$).

Consider tests of the form

$$\delta_c(x) = \begin{cases} 1 & \text{if } \prod_{i=1}^n \{p_1(x_i)/p_0(x_i)\} > c \\ 0 & \text{if } \prod_{i=1}^n \{p_1(x_i)/p_0(x_i)\} < c, \end{cases}$$

where $c \in \{0\} \cup \mathbb{R}^+ \cup \{\infty\}$ (this class is equal to the set of admissible tests).

How should we evaluate such tests? Suppose we lose \$1 if the test makes an incorrect decision. Let

$$LR(X) = \prod_{i=1}^n p_1(X_i)/p_0(X_i).$$

Our expected loss for a given test δ_c is then

$$\Pr(\delta_c(X) \neq \theta)|\theta) = \Pr(LR(X) < c|\theta = 1)1(\theta = 1) + \Pr(LR(X) > c|\theta = 0)1(\theta = 0).$$

See Figure 2.

Discuss: How would you choose c? How does this relate to p-values, level and power?

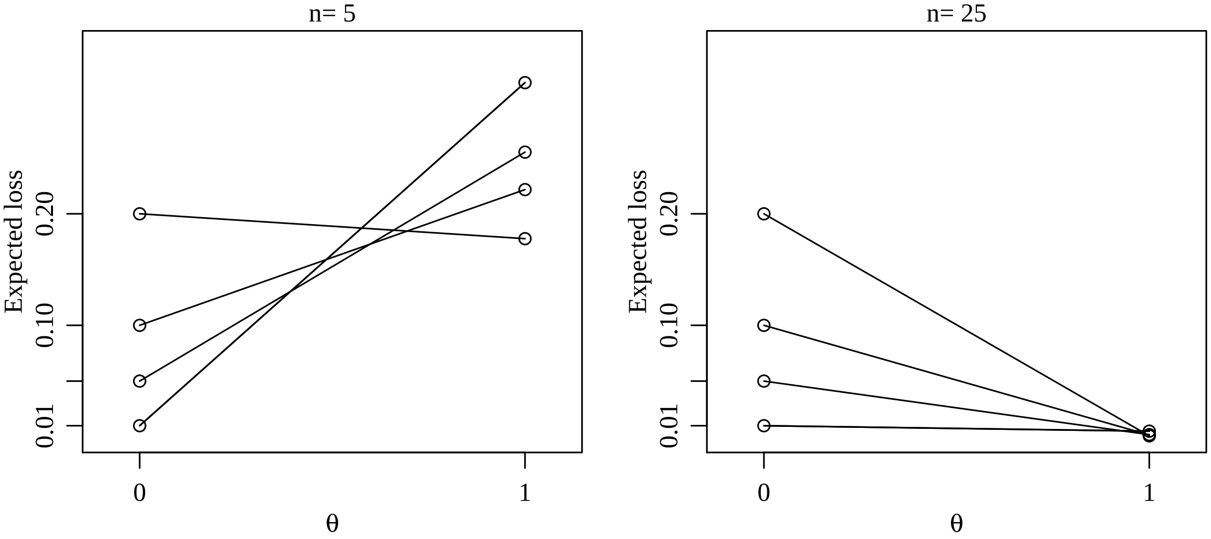

Figure 2: Expected loss under $\theta \in \{0, 1\}$ for $n = 5$ on the left, $n = 25$ on the right.

4 Loss, decision rules and risk

4.1 Statistical decision problems

A *statistical decision problem* consists of

1. an unobservable process P from which observable data X are sampled ($X \sim P$);
2. a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ which we hope includes P;
3. a decision/loss structure: $\{\Theta, D, L\}$:

   Θ, the parameter space, indexes the possible processes;

   D, the decision space, is the set of decisions available;

   $L : \Theta \times D \to \mathbb{R}$, the loss function, denotes the loss incurred for each combination of decision and parameter value.

Example (testing):

$\Theta_0 \cup \Theta_1 = \Theta,\ \ \Theta_0 \cap \Theta_1 = \phi$

$D = \{d_0, d_1\} = \{$ "say Θ_0", "say Θ_1" $\}$

A simple loss function:

	d_0	d_1
$\theta \in \Theta_0$	0	l_0
$\theta \in \Theta_1$	l_1	0

Example (estimation):

$D = g(\Theta)$

$L(\theta, g(\theta)) = 0\ \forall \theta \in \Theta$

$L(\theta, d) \geq 0\ \forall (\theta, d) \in \Theta \times g(\Theta)$

Decision process:

1. $X \sim P_\theta$, θ unknown.

2. Decision maker sees X.

3. Decision maker makes decision $d \in D$, which may depend on X.

4.2 Decision rules and risk

Definition (decision rule)**.** *A non-randomized decision rule is a function* $\delta : \mathcal{X} \to D$.

We will refer to the set of decision rules as $\mathcal{D}$, so $d \in D$, $\delta \in \mathcal{D}$ and $\delta(x) \in D$.

Intuitively, a decision rule δ prescribes a course of action for every observable dataset $X \in \mathcal{X}$. One way to evaluate the performance of a decision rule is in terms of its pre-experimental expected loss, or risk $R(\theta, \delta)$.

$$\begin{aligned}
R(\theta, \delta) &= \mathrm{E}_{X|\theta}[L(\theta, \delta)] \\
&= \int_{\mathcal{X}} L(\theta, \delta(X)) P_\theta(dX) \\
&= \text{ pre-experimental expected loss} \\
&= \text{ average loss under repeated use of } \delta\text{, under } \theta
\end{aligned}$$

<u>Ideal</u>: Use a $\delta(X)$ with a low risk at the true value of θ.
<u>Problem</u>: We may know $R(\theta, \delta)$ for all θ and δ, but we don't know which θ is true.

Thus when evaluating different decision rules, we must consider their risk as a function of θ.

4.3 Why risk?

Throughout this class we will evaluate estimators based on their risk, which is their expected loss:

$$R(\theta, \delta) = \mathrm{E}_{X|\theta}[L(\theta, \delta(X))].$$

You may wonder to yourself, "why not evaluate based on median loss, or some quantile of the distribution of $L(\theta, \delta(X))$?" Well, if you are a fan of evaluating procedures based on hypothetical repetitions, then risk can be related to the long-run average (or total) loss. Otherwise, it turns out there is some philosophical justification for evaluating procedures based on expected loss, which you may or may not find more compelling.

Let us assume that, if θ were known to you, that you could provide a preference ordering to the possible decisions. In a testing situation for example, if we knew $\theta \in \Theta_0$ then we would prefer the decision $d_0 =$"say $\theta \in \Theta_0$" to d_1="say $\theta \notin \Theta_0$," so we write

$$d_0 \prec d_1.$$

In an estimation problem, we generally prefer the decision $d_0 =$"say $\theta = \theta_0$" to $d_1 =$"say $\theta = \theta_1$" if θ_0 is closer in some way to θ_1. For example, in the case that $\Theta = D = \mathbb{R}$ and $\theta = 0$ we prefer d_1 to d_2 if $|d_1| < |d_2|$. At the very least, we wouldn't prefer d_2 to d_1, and so we write

$$d_1 \preceq d_2.$$

Based on our preferences over D, we might have preferences over randomized decisions. Again, suppose we are estimating θ, and are considering how bad different decisions are when $\theta = 0$. Two examples of randomized decisions are the following:

$$D_1 = \begin{cases} .1 \text{ w.p. } .9 \\ .6 \text{ w.p. } .1 \end{cases} \qquad D_2 = \begin{cases} .2 \text{ w.p. } .7 \\ .3 \text{ w.p. } .3 \end{cases}$$

If it is really important that that we don't say θ is over $1/2$ when $\theta = 0$, then we might prefer D_2 to D_1. On the other hand, maybe we stand to gain greatly if the decision is within a 10th of θ, in which case we may prefer D_1.

Both of these randomized decisions correspond to probability distributions over the decision space. Calling them P_1 and P_2, we either

- prefer P_1 (so $P_1 \prec P_2$), or

- prefer P_2 (so $P_2 \prec P_1$) or
- are indifferent (so $P_1 \sim P_2$).

Now consider our preferences over all such distributions on D. Some would call us irrational if our preferences did not form a partial ordering, that is if they did not satisfy the following condition:

$$\text{If } P_1 \preceq P_2 \text{ and } P_2 \preceq P_3 \text{ then } P_1 \preceq P_3$$

These same people might also call us irrational if our preferences didn't satisfy the following additional "axioms of rationality":

A1: If $P_1 \preceq P_2$ then

$$\lambda P_1 + (1-\lambda)P_3 \preceq \lambda P_2 + (1-\lambda)P_3 \;\; \forall \lambda \in (0,1], P_3.$$

A2: If $P_1 < P_2 < P_3$ then there exists λ_a and λ_b in (0,1) such that

$$\lambda_a P_1 + (1-\lambda_a)P_3 \preceq P_2 \preceq \lambda_b P_1 + (1-\lambda_b)P_3.$$

The first axiom seems reasonable, although it has been critiqued because it suggests that aversion to uncertainty is irrational (see Allais' paradox). Axiom 2 essentially says that there is no decision infinitely preferable than another.

If our preferences over probability distributions on D are rational, then the following representation holds:

Theorem 1. *If a partial ordering on distributions over D satisfies A1 and A2, then there exists a function $L(\theta, d)$ such that*

$$P_1 \preceq P_2 \quad \Leftrightarrow \quad \mathrm{E}_{D|P_2}[L(\theta, D)] \leq \mathrm{E}_{D|P_2}[L(\theta, D)].$$

In words, rationality implies your preferences over random decisions can be thought of as a preference to minimize risk.

Now let's relate this back to statistical decision making. Each estimator/test/decision function δ is a function from $\mathcal{X}$ to D, and so if $X \sim P_\theta$, then each $\delta(X)$ corresponds to a probability distribution P_δ over D. The theorem says that if we have rational preferences over distributions on D, then our preferences among estimators will correspond to their risks (see Ferguson [1967, Section 1.4] for a bit more discussion and a proof of the theorem).

5 Statistical decision theory

Statistical decision theory concerns the evaluation of decision rules based on their risk functions.

Example (mean estimation):

$$X = (X_1, \ldots, X_n),\ X_1, \ldots, X_n \sim \text{i.i.d.} P_\theta$$
$$\mu(\theta) = \int x P_\theta(dx) = \text{population mean}$$

- $D = \mu(\Theta)$
- $L(\theta, d) = (\mu(\theta) - d)^2\ \forall d \in \mu(\Theta)$ (squared error/quadratic loss)
- $R(\theta, \delta) = \mathrm{E}_{X|\theta}[(\mu(\theta) - d)^2] = \mathrm{MSE}(\theta, \delta)$.

Some estimators:

- $\delta_1(X) = \bar{X}$
- $\delta_2(X) = \frac{n}{n+1}\bar{X} + \frac{1}{n+1}\mu_0$
- $\delta_3(X) = \mu_0$

Could one of these (or another estimator) uniformly minimize the risk across $\theta \in \Theta$?

Consider the risk of $\delta_3(X)$:

$$\begin{aligned} R(\theta, \delta_3) &= 0 && \theta \in \mu^{-1}(\mu_0) \\ R(\theta, \delta_3) &= (\mu(\theta) - \mu_0)^2 \geq 0 && \text{in general} \end{aligned}$$

δ_3 is typically unbeatable if $\mu(\theta) = \mu_0$, but is a poor estimator for away from μ_0.

Example (testing): Recall our class of tests for simple-versus-simple hypotheses:

$$\delta_c(x) = \begin{cases} 1 & \text{if } \prod_{i=1}^n \{p_1(x_i)/p_0(x_i)\} > c \\ 0 & \text{if } \prod_{i=1}^n \{p_1(x_i)/p_0(x_i)\} < c. \end{cases}$$

Is there a choice of c that minimizes the risk

$$\Pr(\delta_c(X) \neq \theta)|\theta) = \Pr(LR(X) < c|\theta = 1)1(\theta = 1) + \Pr(LR(X) > c|\theta = 0)1(\theta = 0),$$

for both values of θ?

Which estimator has the best risk function?

Generally, there is no estimator or decision rule with uniformly minimum risk: If there were such an rule, it would have to have the same risk as the rule $\delta(X) = g(\theta_0)$ (which generally has zero risk) *for every* $\theta_0 \in \Theta$. The question "which rule has the best risk function" is therefore ill-posed. There are two basic strategies for formulating well-posed versions of this question:

1. Global risk comparisons (LC chapters 4 and 5, LR chapter 8)

    (a) Admissible rules: Consider only rules that are not globally dominated.

    (b) Bayes risk: Compare risk functions averaged over Θ.

    (c) Minimax risk: Compare supremums of risk functions.

2. Decision rule restrictions

(a) Invariant rules:

- UMRE estimation (LC chapter 3);
- UMPI tests (LR chapter 6).

(b) Unbiased rules:

- UMRU estimation (LC chapter 2);
- UMPU tests (LR chapter 4).

One shouldn't look at these as five approaches as unrelated. Sometimes the UMRE is UMRU. Sometimes the UMRU is minimax. Interestingly, in many situations the admissible rules are the Bayes rules, and minimax, equivariant and unbiased rules can often be interpreted as Bayes rules, or approximately so, under particular prior distributions. These connections are what we will study in 581.

References

Thomas S. Ferguson. *Mathematical statistics: A decision theoretic approach.* Probability and Mathematical Statistics, Vol. 1. Academic Press, New York, 1967.

E. L. Lehmann and George Casella. *Theory of point estimation.* Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98502-6.