Introduction to Statistics

(Lectures 5-6)

1 A Decision-Theoretic Approach to Statistics

The following two lectures (Lecture 5-6) provide an introduction to Statistics from a *decision-theoretic* perspective, as in the foundational work of Wald (1950). The decision-theoretic approach to statistics presents problems such as prediction, estimation, testing, and inference, as decision problems under uncertainty. There is emphasis on how an agent evaluates a set of decisions that can be taken. Statistical models play a central role in the analysis, because they provide the link that explains how the observed data can be useful for the decision maker. You can find interesting introductions to statistical inference from a decision-theoretic perspective in Chapter 1 of Lehmann and Romano (2005) and Chapter 1 of Ferguson (1967), p. 1-11. Section 2.5 of Le Cam and Yang (2000) presents very interesting, but brief, historical remarks.

Overview of Lectures 5-6: The objective of Lecture 5 and 6 is to present a framework to think about statistics from a decision-theoretic perspective. We will define the following objects:

a) x: Data

b) $\{P_\theta\}_{\theta\in\Theta}$: Statistical Model

c) Θ: Parameter space

d) $\mathcal{L}(a,\theta)$: Payoff or loss of an action.

e) $d(x)$: Decision rule

f) $\mathbb{E}_\theta[L(d(x),\theta)]$: Expected payoff /loss of an strategy conditional on θ.

1.1 Data and Statistical model

Let X be a random variable taking values in some set $\mathcal{X}$. A realization of X will be refered to as data. Thus, data, are nothing else than realizations of some random variable.

To conduct statistical analysis we will require some structure on the process that generated the data. Such structure is described by a statistical model:

Definition 1:A statistical model for the data is a collection of probability distributions over $\mathcal{X}$:

$$\{\mathbb{P}_\theta\}_{\theta\in\Theta}$$

indexed by an element θ in some space Θ. The index θ is referred to as parameter and the set Θ as the parameter space.

A statistical model is also often called a *statistical experiment* or simply an experiment (see Le Cam and Yang (2000) Chapter 2).[1] The notion of statistical model plays a key role in econometrics. Concepts such as "identification" and "sufficient statistics" are formally defined in terms of a statistical model. You will have a chance to appreciate this point in the problem set (Problems 1 and 2).

1.2 Statistical (Decision) Problems

It is often helpful to define a statistical decision problem as a game involving two players ("nature" and the statistician) and two stages (data generation and decision making). The extensive form description of the game is as follows.

In the first stage, nature selects a parameter $\theta \in \Theta$ and uses it to generate data according to the distribution $\mathbb{P}_\theta$. In the second stage, the econometrician/statistician/decision maker observes the data, but does not observe the parameter selected by nature. The decision maker is not under a full veil of ignorance as he/she observes the data.

Based on the realized data, the econometrician would like to take an *action* "a" whose payoff depends on the parameter selected by nature. This means that some actions are reasonable in some states but not in others. This is modeled by endowing the decision maker with a state-contingent utility or loss:

$$u(a, \theta) \quad \text{or} \quad \mathcal{L}(a, \theta).$$

The action is assumed to live in an action space $\mathcal{A}$. The *decision problem* faced by the econometrician is the selection of an action depending on the realization of the data.

This leads to the following definition:

Definiton 2: A statistical decision problem is a tuple:

$$(\Theta, A, u, \{P_\theta\})$$

containing a parameter space, an action space, a utility (or loss) function, and a statistical model. See Section 1.3 of Chapter 1 in Ferguson (1967).

[1] Le Cam and Yang (2000) seems to credit Blackwell (1953) for the definition of experiment.

1.3 Decision Rules

In a statistical game (as in any other game) the objective of the players is to think about reasonable strategies that can be played. In the (extensive form) statistical game the information sets for the statistician are the data realizations. Thus,

Definition 3: A strategy of decision rule d for the statistician in a statistical game is a function $d : \mathcal{X} \to A$.

Here are some examples of statistical problems and decision rules:

1. Estimation Problem: The action space for the econometrician is Θ. This means that after observing a data realization, the econometrician needs to decide what is the parameter θ that generated the data.

   The decision rules for this problem are usually called **estimators**.

   A typical loss for this problem is **quadratic loss**: $\mathcal{L}(a, \theta) = (a - \theta)^2$.

2. Testing Problem: The parameter space is partitioned into two sets: Θ_0 is called the null hypothesis and Θ_1 is the alternative hypothesis. The action space of the econometrician contains only two elements $\{a_0, a_1\}$ (a_0 is interpreted as expressing support for the null and a_1 as expressing support for the alternative).

   Decision rules for this problem are usually called **tests**.

   A typical loss for this problem is the so-called **0-1 loss**: $\mathcal{L}(a_1, \theta_0) = 1$ and $\mathcal{L}(a_0, \theta_1) = 1$ (with the loss being 0 otherwise).

3. Inference problem: The action for the econometrician consists of subsets of the parameter space. The interpretation is that each subset contains the best candidate values for θ.

   Decision rules for this problem are usually called **confidence sets**.

   A typical loss for this problem is:

$$\mathcal{L}(C; \theta) \equiv \delta \mathbf{1}\{\theta \notin C\} + (1 - \delta)\text{Vol}(C).$$

1.4 "Optimal" Decision Rules

Ferguson (1967) Chapter 1, p. 1:

"The fundamental problem of decision theory can be stated quite simply. Given a game $(\Theta, A, u, \{P_\theta\})$ and a random observable X whose distribution depends on $\theta \in \Theta$, what decision rule δ should the statistician use?"

For each action taken, the utility or loss function is deterministic. However, when the statistician reports a decision rule the loss associated to a particular decision, the payoff is a random variable: each action will lead to a different value of the loss function, depending on the data realization.

One way to summarize the performance of a decision rule in different points of the parameter space is by reporting the *risk function*:

Definition 4: (Ferguson p. 7) The risk function of a decision rule d is defined as:

$$R(\theta; d) = \mathbb{E}_{P_\theta}[\mathcal{L}(d(X), \theta)].$$

Another cite from Ferguson (1967) p. 7:

"It is a natural reaction to search for a best decision rule, a rule that has the smallest risk no matter what the true state of nature is. Unfortunately, situations in which the best decision rule exists are rare and uninteresting."

While a rule that has the smallest risk need not exist, it is simple to use a dominance criterion to discard bad decision rules. This is analogous to the idea of discarding dominated strategies in the analysis of games.

Definition 5: A decision rule d is dominated if there is a decision rule d' such that:

$$R(\theta; d') \leq R(\theta; d)$$

for every $\theta \in \Theta$, with strict inequality for some θ. Decision rules that are not dominated are called *admissible*. See Chapter 2 p. 54 in Ferguson (1967) for an equivalent definition.

Not all decision rules can be compared using the dominance criterion (dominance is a partial order). There are two general methods for creating a complete ordering over decision rules: "average" and "worst-case" risk . These criteria lead to decision rules that can be defined generally for any statistical problem.

1.4.1 Bayes Rules

Given a probability distribution π on Θ (a prior) we define the the Bayes risk of a decision rule d as:

$$r(\pi, d) \equiv \int_{\Theta} R(\theta, d) d\pi(\theta) \equiv \mathbb{E}_{\pi}[R(\theta; d)].$$

Definition 6: A decision rule d^* is said to be a Bayes Rule with respect to the prior distribution π (and relative to a class of decision rules D) if:

$$r(\pi, d^*) = \inf_{d} r(\pi, d).$$

That is, if it minimizes Bayes risk. See Chapter 2 p. 31 in Ferguson (1967).

By construction, Bayes rules require the specification of prior. Consequently, the properties of a Bayes decision rule will depend on the choice of such prior. However, the are some limits on how bad a Bayes rule can behave. The following result shows that under some minimal assumptions, Bayes rules are admissible; hence, it is not possible to improve upon them uniformly over Θ.

Result: Suppose that the risk function $R(d; \cdot)$ is continuous in θ for any decision rule d. Let π be any prior with full support on Θ.[2] The Bayes rule d^* corresponding to π is admissible.

PROOF: Suppose d^* is not admissible. Then there exists d' such that:

$$R(d'; \theta) \leq R(d^*; \theta),$$

with strict inequality for some $\theta^* \in \Theta$. Since $R(d^*, \theta)$ is continuous in θ, there exists a neighborhood N_{θ}^* such that

$$R(d', \theta) < R(d^*, \theta) \quad \forall \quad \theta \in N_{\theta^*}.$$

Since π has full support

$$\int_{N_{\theta^*}} R(d'; \theta) d\pi(\theta) < \int_{N_{\theta^*}} R(d'; \theta^*) d\pi(\theta).$$

Consequently:

$$r(\pi, d') < r(\pi, d^*).$$

A contradiction.

[2] A point θ is said to be on the support of a distribution π is for every neighborhood $\mathcal{N}_{\theta}$ of θ have strictly positive probability. The distribution π has full support if every $\theta \in \Theta$ is in the support of the distribution.

□

Thus, Bayes Rules are admissible under mild regularity conditions. Interestingly, a converse of this result is also true: any admissible decision rule is, with some qualifications, Bayes for some prior. This result is known as the *Complete Class Theorem*; see Chapter 2 (in particular 2.10) in Ferguson (1967), if interested.

Bayes rules are not a statistical panacea. A poorly selected prior (one that ignores certain parts of the parameter space) leads to a Bayes rule that can be strictly improved.

Bayes Rules are defined very generally, but minimizing Bayes risk is a conceptually complicated problem: we are optimizing over a space of functions (the decision rules). We now show that it is possible to simplify the optimization problem, by distinguishing between prior and posterior information.

Let $f(x;\theta)$ denote the p.d.f corresponding to $\mathbb{P}_\theta$ and assume that $\pi(\theta)$ is also a p.d.f. Note that:

$$\begin{aligned} r(\pi, d) &\equiv \int_\Theta R(d(x);\theta)\pi(\theta)d\theta, \\ &= \int_\Theta \Big(\int_\mathcal{X} \mathcal{L}(d(x);\theta) f(x|\theta) dx \Big) \pi(\theta) d\theta, \\ &\quad \text{(by definition of Risk)} \\ &= \int_\mathcal{X} \Big(\int_\Theta \mathcal{L}(d(x);\theta) f(x|\theta)\pi(\theta) d\theta \Big) dx, \\ &= \int_\mathcal{X} \Big(\int_\Theta \mathcal{L}(d(x);\theta) \pi(\theta|x) d\theta \Big) f^*(x) dx. \end{aligned}$$

where $f^*(x) = \int_\theta f(x|\theta)\pi(\theta)d\theta$ is the marginal density of the data in a model with joint p.d.f. given by $f(x|\theta)\pi(\theta)$. So, minimizing (ex-ante) Bayes Risk is the same as choosing the action $d(x) \in \mathcal{A}$ that minimizes:

$$\int_\Theta \mathcal{L}(d(x);\theta)\pi(\theta|x)d\theta.$$

The latter quantity is referred to as posterior loss. Minimizing posterior loss is an easier problem, as action spaces are typically subsets of $\mathbb{R}^n$ not spaces of functions. In the next lectures, we will work out a few examples to develop a better understanding of this result.

1.4.2 Minimax rules

An essentially different type of ordering of the decision rules may be obtained by comparing the rules according to the worst that could happen to the statistician. In other words, d_1 is weakly preferred to d_2 if

$$\sup_{\theta} R(\theta, d_1) \leq \sup_{\theta} R(\theta, d_2).$$

Definition 7: A decision rule d_0 is said to be minimax (relative to a class D of decisions) if:

$$\sup_{\theta \in \Theta} R(\theta, d_0) = \inf_{d \in D} \sup_{\theta \in \Theta} R(\theta, d).$$

See Definition 3, Chapter 1 in Ferguson (1967). The minimax rule is designed to protect the statistician against worst-case situations: the rule is chosen to minimize the worst-case risk (or to obtain the best performance in the worst possible situation).

One of the reasons minimax decision rules are appealing is that their construction does not require the user to specify any prior belief. This is an attractive property, but it comes at the cost of making the reasonability of minimax decision rules harder to justify. In general, there is no guarantee that minimax rules are admissible. We will not show it, but under some conditions *minimax rules are Bayes*; that is, there exists a prior distribution π_0 that makes the minimax rule d_0 a Bayes rule.

A further connection between Bayes and minimax rules is as follows.

Result: Suppose that d^* is a Bayes rule with constant risk; that is $R(\theta, d^*) = C$ for some $C \in \mathbb{R}$. Then d^* is minimax.

Proof: If π has constant risk for any prior π

$$\sup_{\theta \in \Theta} R(\theta, d^*) = \int_{\Theta} R(\theta, d^*) d\pi(\theta)$$

In particular, if d^* is Bayes for π^*, for any d

$$\int_{\Theta} R(\theta, d^*) d\pi^*(\theta) \leq \int_{\Theta} R(\theta, d) d\pi^*(\theta),$$

by definition. However, for any decision rule d

$$\int_{\Theta} R(\theta, d) d\pi^*(\theta) \leq \int_{\Theta} \sup_{\theta \in \Theta} R(\theta, d) d\pi^*(\theta) = \sup_{\theta \in \Theta} R(\theta, d).$$

We conclude that for any d,

$$\sup_{\theta \in \Theta} R(\theta, d^*) \leq \sup_{\theta \in \Theta} R(\theta, d).$$

Implying d^* is minimax.

1.5 Least Favorable Distribution

In the previous section we defined a minimax rule as the decision rule that protects the statistician against worst-case situations. Thinking back about the game-based definition of a statistical decision problem, it is possible to also define a worst-case strategy for nature.

Let Θ^* denote the space of all probability distributions over Θ.

Definition 7: A distribution $\pi_0 \in \Theta^*$ is said to be least favorable if

$$\inf_{d \in \mathcal{D}} r(\pi_0, d) = \sup_{\pi \in \Theta^*} \inf_{d \in \mathcal{D}} r(\pi, d).$$

As explained by Ferguson (1967), *"the name 'least favorable' derives from the fact that if the statistician were told which prior distribution nature was using he would like least to be told a distribution π_0"* as defined above.

In some decision problems it is possible to show that the *minimax theorem* holds:

$$\sup_{\pi \in \Theta^*} \inf_{d \in \mathcal{D}} r(\pi, d) = \inf_{d \in D} \sup_{\theta \in \Theta} R(\theta, d). \tag{1.1}$$

If the minimax theorem holds, and a least favorable distribution π_0 exists, then any minimax rule d_0 is Bayes with respect to π_0.

References

Blackwell, D. (1953): "Equivalent comparisons of experiments," *The annals of mathematical statistics*, 265–272.

Ferguson, T. (1967): *Mathematical Statistics: A Decision Theoretic Approach*, vol. 7, Academic Press New York.

Ferguson, T. S. (1996): *A course in large sample theory*, vol. 49, Chapman & Hall London.

Le Cam, L. and G. L. Yang (2000): *Asymptotics in Statistics: Some Basic Concepts*, Series in Statistics, Springer, second edition ed.

Lehmann, E. and J. Romano (2005): *Testing Statistical Hypotheses*, Springer Texts in Statistics, Springer Verlag.

Wald, A. (1950): *Statistical Decision Functions*, Oxford, England: Wiley.