Basic Principles of Statistics

February 5-7 , 2008

1 Decisions, Loss, and Risk

The basic idea in inferential statistics is to take an action based on a decision strategy that uses information obtained from data, X. In parametric statistics, the distribution of X depends on the choice of probabilities from a family P_θ where θ, the so-called **state of nature**, is chosen from a parameter set Θ. We will write E_θ to indicate expectation with respect to the probability θ.

In the example of determining preference between two choices based on a poll. The design is the choice of individuals selected for the poll resulting in data,

$$X = (X_1, \dots, X_n),$$

Here, the X_i's are independent and identicallly distributed, or as more commonly stated in statistics, form a **simple random sample**.

One decision is to estimate the fraction of the population that take the first position, then the action is to choose a number θ from the parameter set $[0, 1]$.

A simple decision problem is to hypothesize a preferred choice and then either to reject or to fail to reject the hypothesis.

Given data $x = (x_1, \dots, x_n) \in S^n$, we must make a decision, a choice from the **action space**, A. Thus, we introduce the **decision function** or **rule**.

$$d : S^n \to A.$$

Decisions have consequences, a measure of how seriously we view incorrect decisions. This leads to the introduction of the **loss function**,

$$\mathcal{L} : \Theta \times A \to \mathbb{R}.$$

Thus, if the state of nature is θ, then $\mathcal{L}(\theta, a)$ is the loss incurred upon taking the action a.

Example 1 (Loss Functions).

1. $\mathcal{L}_1(\theta, a) = |a - \theta|$,
2. $\mathcal{L}_2(\theta, a) = (a - \theta)^2$,
3. $\mathcal{L}_\infty(\theta, a) = 0$ *if* $\theta = a$ *and* $\mathcal{L}(\theta, a) = 1$ *if* $\theta \neq a$

The goal is to make the choice of decision function from the set of decision functions $\mathcal{D}$ that minimizes the loss on average.

Definition 2. *The* **risk function**

$$\mathcal{R} : \Theta \times \mathcal{D} \to \mathbb{R}$$

is defined by

$$\mathcal{R}(\theta, d) = E_\theta \mathcal{L}(\theta, d(X_1, \ldots, X_n)).$$

Example 3. *Our datum is the result of a single discrete random variable with mass function $p(\cdot|\theta)$ and decision function $d(x) = x$. The question is how does the parameter θ reflect a property of the mass function.*

1.

$$\begin{aligned} \mathcal{R}_1(\theta, d) &= E_\theta \mathcal{L}_1(\theta, d(X)) = E|X - \theta| = \sum_x |x - \theta| p_X(x|\theta) \\ &= \sum_{x<\theta} (\theta - x) p_X(x|\theta) + \sum_{x \geq \theta} (x - a) p_X(x|\theta) \\ &= \theta P_\theta\{X < \theta\} - \theta P_\theta\{X \geq \theta\} - \sum_{x<\theta} x p_X(x|\theta) + \sum_{x\geq\theta} x p_X(x|\theta). \end{aligned}$$

$\mathcal{R}_1$ is a continuous piecewise linear function of θ with slope

$$P\{X < \theta\} - P\{X \geq \theta\} = 1 - 2P\{X \geq \theta\}).$$

Thus, $\mathcal{R}_1$ is decreasing if $P\{X \geq \theta\} > 1/2$ and increasing if $P\{X \geq \theta\} < 1/2$. Consequently, $\mathcal{R}_1$ is minimized by taking a equal to the median.

2.

$$\mathcal{R}_2(\theta, d) = E_\theta \mathcal{L}_2(\theta, d(X)) = E(X - a)^2 = \sum_x (x - \theta)^2 p_X(x|\theta)$$

Thus,

$$\frac{\partial}{\partial a} \mathcal{R}_2(\theta, d) = -\sum_x (x - \theta) p_X(x|\theta) = -EX + \theta.$$

Thus, the minimum is achieved by taking θ equal to the mean

3.

$$\mathcal{R}_\infty(\theta, d) = E_\theta \mathcal{L}_\infty(\theta, d(X)) = 0 \cdot P\{X = \theta\} + 1 \cdot P\{X \neq \theta\} = 1 - P\{X = \theta\}.$$

This is minimized by taking a equal to the mode.

2 Minimax Rules and Bayes Rules

Given a loss function, the goal is to find a "good" decision function, one that minimizes risk. This choice has to be made without the knowledge of the state of nature. In other words, the parameter value θ is unknown.

The dilemma can be seen whenever we have to decision rules d_1 and d_2 and two parameter values θ_1 and θ_2 so that

$$\mathcal{R}(\theta_1, d_1) < \mathcal{R}(\theta_1, d_2) \text{ but } \mathcal{R}(\theta_2, d_1) > \mathcal{R}(\theta_2, d_2).$$

The two classical approach to this problem are **minimax rules** and **Bayes rules**

For a minimax case, we consider, for a given decision rule, the state of nature that has the most risk:

$$\sup_{\theta \in \Theta} \mathcal{R}(\theta, d).$$

Then, choose the decision rule d^* that minimizes this maximum risk:

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

If this rule d^* exists then it is called a **minimax** rule.

This rule leads to decisions functions that guard against those situations with the worst risk. If such cases are very rare, then we can introduce a probability distribution Π on the parameter space Θ. With with **prior distribution**, the risk is a random variable.

Definition 4. *If the prior distribution Π has density π, the the* **mean risk***,*

$$r(\Pi, d) = \int_{\Theta} \mathcal{R}(\theta, d)\pi(\theta)\ d\theta.$$

If the prior distribution Π has mass function π, the the **mean risk***,*

$$r(\Pi, d) = \sum_{\theta \in \Theta} \mathcal{R}(\theta, d)\pi(\theta).$$