STATS 300A: Theory of Statistics I **Autumn 2016/17**

Lecture 1: September 27

Lecturer: Joseph Romano *Scribes: Kenneth Tay*

1.1 Statistical Experiments

Define the following:

- Data X from some sample space $\mathcal{S}$,
- $\mathcal{E} := \sigma$-algebra on $\mathcal{S}$,
- $\{P_\theta, \theta \in \Omega\}$ a set of probability distributions that could have generated X. We call this set a *model* of possible probability distributions. Here, (Ω is a general set indexing the model.)

With this set-up, we can define the rough goal of "estimation": to construct an "estimator" (or "statistic") $\delta(X)$ which is "close" to θ (or some function $g(\theta)$).

1.2 A First Stab at the Optimality Problem

We can make the concept of "closeness" precise by defining a loss function which captures the penalty we incur when we estimate $g(\theta)$ wrongly:

Definition 1.1 *A **loss function** $L(g(\theta), \delta(X))$ is a real-valued function represents the penalty incurred when we estimate the value of $g(\theta)$ with $\delta(X)$.*

Here are 2 examples of loss functions:

- $L(\theta, \delta(X)) = |\theta - \delta(X)|$ (absolute error loss)
- $L(\theta, \delta(X)) = (\theta - \delta(X))^2$ (squared error loss)

The problem with the loss function is that we don't know which X we will get! Hence, we average over all possible realizations of X to get the risk function:

Definition 1.2 *The **risk function** of $\delta = \delta(X)$ is given by*

$$\begin{aligned} R(g(\theta), \delta) &= \mathbb{E}_\theta \left[L(g(\theta), \delta(X)) \right] \\ &= \int L(g(\theta), \delta(X)) dP_\theta(x). \end{aligned}$$

Here, the expectation/integral is taken over all X (not over all θ).

We can now try to compare estimators based on the risk function. However, when we do this, we run into a problem:

Proposition 1.3 *In general, there does not exist a procedure δ^* such that $R(\theta, \delta^*) \leq R(\theta, \delta)$ for any other δ.*

Proof: Consider the case where $g(\theta) = \theta$ with θ real, and where the loss function L is the absolute error.

For every $c \in \mathbb{R}$, let $\delta_c := c$, i.e. ignore the data and guess that $\theta = c$. Then

$$\begin{aligned} R(\theta, \delta_c) &= \mathbb{E}_\theta L(\theta, \delta_c) \\ &= L(\theta, \delta_c) \\ &= |\theta - c|. \end{aligned}$$

In particular, $R(\theta, \delta_c) = 0$ when $\theta = c$. This means that if such a δ^* existed, we must have $R(\theta, \delta^*) = 0$ for all c, which is not possible. ■

1.3 Redefining the Optimality Problem

The proposition above showed that in general, there is no procedure δ^* that "dominates" all other procedures. Hence, we will need to define what we mean by an optimal procedure in a different way. There are a number of ways we can do this.

1.3.1 Restrict the class of estimators in some way

1.3.1.1 Restricting to unbiased estimators

Definition 1.4 *The **bias** of an estimator $\delta(X)$ is given by*

$$Bias_\theta \delta(X) := \mathbb{E}_\theta \delta(X) - g(\theta).$$

Definition 1.5 *An estimator $\delta(X)$ is unbiased for $g(\theta)$ (here we take g to be real-valued) if*

$$\mathbb{E}_\theta \left[\delta(X)\right] = g(\theta)$$

for all θ.

Note that restricting the estimation problem to just unbiased estimators is a *very* severe restriction! For instance, it could be the case that there are no unbiased estimators for the problem. It could also be that biased estimators do much better overall.

For example, consider the set-up $X_1, \ldots, X_n$ iid, $X_i \sim \mathcal{N}(0, \sigma^2)$, with σ^2 unknown. The statistic to be estimated is $\theta = \sigma^2$. In this case, $\frac{1}{n}\sum X_i^2$ is an unbiased estimator, and it is the unbiased estimator with the smallest mean-squared error. However, the (biased) estimator $\frac{1}{n+1}\sum X_i^2$ has uniformly smaller mean-squared error (i.e. for all values of σ).

Why might this be the case? The mean squared error can be broken down into 2 components, variance and bias:

$$\mathbb{E}_\theta\left[(\delta(X)-g(\theta))^2\right] = \mathrm{Var}_\theta\delta(X) + \left[\mathrm{Bias}_\theta\delta(X)\right]^2.$$

Hence, it could happen that a biased estimator introduces some bias but has drastically smaller variance.

1.3.1.2 Use "symmetries" or "invariance considerations"

The basic idea here is that the inference made should not depend on units of measure. (For example, whether height data is given in feet or inches should not make a difference.)

Take, for example, data $X_1, \dots X_n$, and some fixed $c > 0$. Then, to invoke invariance, our estimator should satisfy

$$\delta(cX_1, \dots, cX_n) = c\delta(X_1, \dots, X_n)$$

for all $c > 0$. An estimator that satisfies this equation is said to be **equivariant**.

Note: Such an invariance may not be present. For example, let $X \sim \mathrm{Binom}(n, \theta)$, i.e.

$$P_\theta(X = j) = \binom{n}{j}\theta^j(1-\theta)^{n-j} \qquad \forall j = 0, 1, \dots, n.$$

There is no real symmetry or invariance to exploit in this set-up.

1.3.2 Weaker notion of optimality

1.3.2.1 Minimax estimator

Definition 1.6 *An estimator δ^* is **minimax** if $\sup_\theta R(\theta, \delta^*) \le \sup_\theta R(\theta, \delta)$ for all other δ.*

Roughly speaking, the minimax estimator minimizes the penalty in the worst-case scenario. Note that the minimax estimator may not be unique!

Example: $X \sim \mathrm{Binom}(n, \theta)$. The uniformly minimum risk unbiased estimator of θ is $\frac{X}{n}$. However, it is not minimax! The minimax estimator is $\dfrac{X + \frac{\sqrt{n}}{2}}{n + \sqrt{n}}$.

1.3.2.2 Minimize "average risk"

Let Λ be a probability distribution on Ω. We can now talk about averaging over all possible θ. The average risk of an estimator δ is given by

$$\int_{\theta\in\Omega} R(\theta, \delta)d\Lambda(\theta).$$

We say that δ^* minimizes average risk, or is a **Bayes estimator**, with respect to "prior" Λ if

$$\int R(\theta, \delta^*)d\Lambda(\theta) \le \int R(\theta, \delta)d\Lambda(\theta)$$

for any other δ.

1.4 Exponential Family Models

Definition 1.7 *A family of probability distributions $\{P_\theta\}$ is said to form an **s-parameter exponential family of distributions** if all the P_θ's have densities of the form*

$$p_\theta(x) = \exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x)$$

with respect to a dominating measure μ (i.e. $\frac{dP_\theta}{d\mu}(x) = p_\theta(x)$). Here, the η_i's, T_i's and B are real-valued functions.

The **canonical form** uses $\eta_1, \ldots, \eta_s$ as the "parameters". We can write it as

$$p(x, \eta) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x),$$

where $(\eta_1, \ldots, \eta_s)$ is the unknown vector in $\mathbb{R}^s$. Notes:

- The representation is not unique (e.g. replace η_1 and T_1 with $2\eta_1$ and $\frac{T_1}{2}$).
- $h(x)$ can be "absorbed" into the measure μ.
- There is a notion of a natural parameter space, i.e. the values of η make sense:

$$\left\{\eta = (\eta_1, \ldots, \eta_s) : \int \exp\left[\sum \eta_i T_i(x)\right] h(x)\mu(dx) < \infty\right\}.$$

- If the η_i's satisfy some linear constraint, then the number of terms in the sum can be reduced. Unless this is done, $(\eta_1, \ldots, \eta_s)$ is not **identifiable**.

Definition 1.8 *θ is **identifiable** if*

$$\theta_1 \neq \theta_2 \qquad \Rightarrow \qquad P_{\theta_1} \neq P_{\theta_2}.$$