Stat210A: Theoretical Statistics **Lecture Date: August 28, 2014**

Lecture 1: Statistical Decision Theory

Lecturer: Michael I. Jordan *Scribe: K. Jarrod Millman*

1 Notation

Statistical inference concerns learning from data. In general, we will consider data $X = (X_1, X_2, ..., X_n)$ for $n \in \{1, 2, ...\}$ to be a random vector (or variable). In other words, we assume that there exists some distribution P_θ such that $X \sim P_\theta$ where P_θ belongs to a set $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ of probability distributions. We call $\mathcal{P}$ our statistical model for the distribution of X and Ω our parameter space (i.e., the set of all possible values of θ). This semester, we focus on parametric models where $\theta \subset \mathbf{R}^P$. Nonparametric models will be covered in 210B.

A statistic $\delta(X)$ is a function of the data X which may be real- or function-valued. In statistical estimation the goal is to find a statistic δ such that $\delta(X)$ is "close to" $g(\theta)$, where θ is the "true" parameter that indexes X's distribution P_θ. We say that δ is an estimator of $g(\theta)$.

2 Decision Theory Framework

To make precise what we mean by "close to" we introduce statistical decision theory. Statistical decision theory was first developed by Abraham Wald and was inspired by work in economics.

We begin by introducing the notion of a loss function L, which given a parameter θ and an estimate $\delta(x)$ returns a non-negative real number that is the loss associated with estimating $g(\theta)$ with $\delta(x)$. We also require that there is no loss for estimating $g(\theta)$ with the correct answer (i.e., $L(\theta, g(\theta)) = 0$). For example, we may be interested in the squared error loss function

$$L(\theta, \delta(x)) = (g(\theta) - \delta(x))^2.$$

In practice, the "true" parameter θ is unknown and the statistic $\delta(X)$ is random. From the frequentist perspective, θ is assumed fixed (while unknown) so the focus is on finding statistics that are "good" over lots of different x. For instance, if you are writing a general purpose software package, then you want to provide some guarantee that your code will perform well on lots of different data. Alternatively, Bayesians condition on the data and treat the parameter θ as random.

2.1 Frequentist risk

The frequentist risk is defined as

$$R(\theta, \delta) = E_\theta L(\theta, \delta(X))$$

where E_θ means that we are taking the expectation with respect to P_θ. Even though X has been integrated out the frequentist risk R is still a function of θ.

In certain situations, comparing the risk of two estimators is straightforward. Consider the situation in Figure 1. Here $R(\theta, \delta_2)$ always has lower risk than $R(\theta, \delta_1)$. This motivates the following definitions.

Definition 1. Given a risk function R and two estimators δ_1 and δ_2, we say δ_2 dominates δ_1 if $R(\theta, \delta_2) \leq R(\theta, \delta_1)$ for all θ and there exists a θ such that $R(\theta, \delta_2) < R(\theta, \delta_1)$.

Definition 2. Given risk function R, an estimator δ_1 is said to be inadmissible if there exists an estimator δ_2 which dominates it.

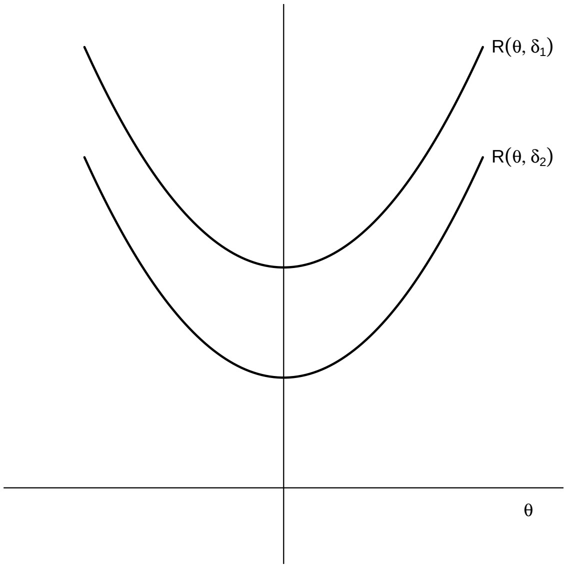

Figure 1: Two frequentist risk functions where one is dominated by the other.

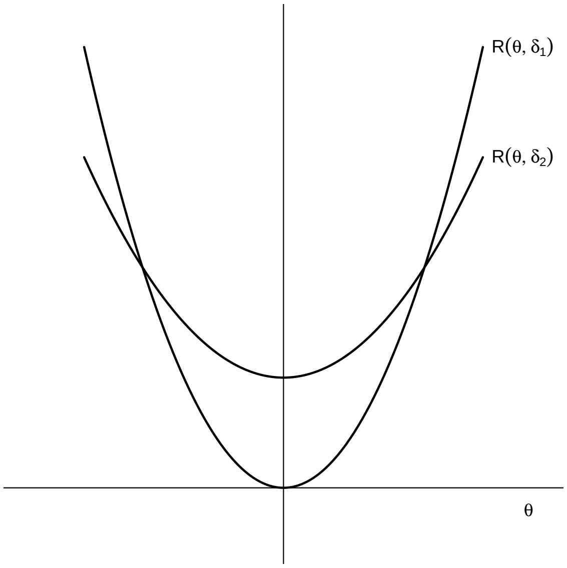

Figure 2: Two frequentist risk functions where neither is dominated by the other.

To compare two admissible statistics (see Figure 2), we need to impose additional constraints.

1. Initially much attention focused on unbiased estimators. An unbiased estimator is one where $E_\theta \delta(X) = g(\theta)$. It was often found that there exists a best (i.e., lowest variance) unbiased estimator. However, many good procedures are biased and some biased ones are inadmissible. For instance, in Stein et al. (1956), they show that the sample mean is an inadmissible estimator of the mean of a multivariate Gaussian once you have three or more dimensions.

2. Equivariance is another useful criteria, which requires estimators to change in a coherent way when the data and parameter change in a compatible way. For example a shift invariant estimator for a location parameter should change by the same amount if all data values are increased by a given amount.

3. Minimax focuses on the worst case performance. In other words, the estimator with lowest maximal risk is chosen.

4. Constraints such as run time are increasingly used in practice.

2.2 Example

For concreteness, consider (Keener, 2010, Example 3.1, p. 40-41) the following random variable

$$X \sim Bin(100, \theta)$$

for $\theta \in [0,1]$. We now examine the frequentist risk associated with the following three estimators of $g(\theta) = \theta$ under the squared error loss function:

$$\delta_1(X) = \frac{X}{100}, \qquad \delta_2(X) = \frac{X+3}{100}, \qquad \text{and} \quad \delta_3(X) = \frac{X+3}{106}.$$

We have the following risk functions for our three estimators

$$\begin{aligned} R(\theta,\delta_1) &= \frac{\theta(1-\theta)}{100} \\ R(\theta,\delta_2) &= \frac{9+100\theta(1-\theta)}{100^2} \\ R(\theta,\delta_3) &= \frac{(9-8\theta)(1+8\theta)}{106^2} \end{aligned}$$

for $\theta \in [0,1]$. The three risk functions are plotted in Figure 3.

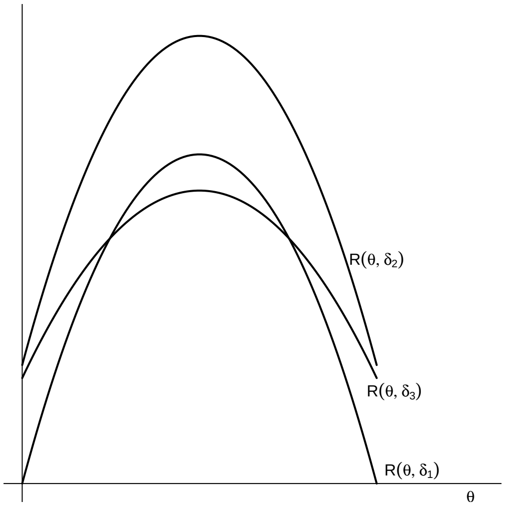

Figure 3: Frequentist risk functions for δ_1, δ_2, and δ_3

It should be clear from the figure that δ_2 is inadmissible as both of the other risk functions dominate it. According to the minimax criteria, you would choose δ_3 over δ_1 since it has better worse case risk. Using the traditional criteria of unbiasedness, you would choose δ_1 over δ_3 since δ_1 is unbiased while δ_3 is biased. If we knew that θ should be close to $1/2$ then δ_3 would give us lower frequentist risk, while if we knew that θ should be closer to 0 or 1 then δ_1 would give us lower risk.

3 Bayes risk

If we are only interested in particular region of θ, it may be desirable to weight those values more heavily. This leads to Bayes' risk:

$$R(\delta) = \int R(\theta,\delta)\pi(\theta)d\theta. \tag{1}$$

To make comparison of different statistics meaningful, $\pi(\theta)$ is required to be a non-negative function of θ which integrates to 1. For frequentist, $\pi(\theta)$ is viewed as a weighting function. For a Bayesian, $\pi(\theta)$ is viewed as a prior on θ. Regardless of how it is viewed, if everyone agrees on a particular $\pi(\theta)$ then they can agree on which estimator δ is best.

References

Keener, R. (2010). *Theoretical Statistics: Topics for a Core Course.* Springer, New York, NY.

Stein, C. et al. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206.