OPTIMAL SAMPLE SIZES FOR ALTERNATIVE LOSS FUNCTIONS

Dhiren Ghosh, Synectics for Mngmt Decisions & Andrew Vogt, Georgetown U.
Dhiren Ghosh, SMD, 3030 Clarendon Blvd, # 305, Arlington, VA 22201

Keywords: asymmetric loss, cost, risk

0. INTRODUCTION The loss function L under simple random sampling is usually assumed to be of the form:

$$L = C|\overline{X} - \mu|,$$

where C is cost per unit absolute error. Here $\overline{X}$ is the sample mean of a sample of size n, used to estimate the true mean μ. The loss function is a linear function of the absolute error, symmetric with respect to both positive and negative errors.

The expected value of the loss function, the risk function, is:

$$E(L) = CE(|\overline{X} - \mu|).$$

When the underlying variable X has standard deviation σ, and $\overline{X}$ is approximately normal with negligible finite population correction factor, then

$$E(|\overline{X} - \mu|) = \sqrt{\frac{2}{\pi}}\frac{\sigma}{\sqrt{n}}. \qquad (1)$$

If C can be quantified - and this is often difficult, the sample size n can be chosen to minimize the expected cost function:

$$A + Bn + CE(|\overline{X} - \mu|).$$

The value that accomplishes this is a positive integer n approximately equal to:

$$(\frac{C\sigma}{\sqrt{2\pi}B})^{\frac{2}{3}}. \qquad (2)$$

In practice the loss function's dependence on error need be neither symmetric nor linear. If the functional form of the dependence as well as the magnitudes of the cost factors can be identified, an alternative expected cost function can be calculated and used to choose the optimal sample size. In the case of an asymmetric loss function, not only is the choice of the sample size affected but even the choice of an estimator. As we shall see, for a given sample size a biased estimator may have a smaller expected loss than the sample mean.

We illustrate the situation with some examples. In most cases we assume that the estimator is the sample mean $\overline{X}$, that this estimator is approximately normal with standard deviation $\frac{\sigma}{\sqrt{n}}$, and that the finite population correction factor can be ignored.

1. ASYMMETRY The simplest case to consider is a slight modification of $C|\overline{X} - \mu|$ that allows for asymmetry, namely

$$L_1 = \begin{cases} C_1|\overline{X} - \mu| & \text{if } \overline{X} > \mu \\ C_2|\overline{X} - \mu| & \text{if } \overline{X} < \mu, \end{cases} \qquad (3)$$

where C_1 and C_2 are costs per unit positive error and negative error respectively.

In this case the expected loss is:

$$\frac{(C_1 + C_2)}{2}\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sqrt{n}},$$

instead of C times the expression in (1). It has the same form as in the traditional case.

2. BIAS However, the asymmetry suggests a new way of looking at estimation. Instead of adopting the *a priori* goal of minimizing mean square error (and thus using the sample mean, a linear unbiased estimator), we propose to choose an estimator that minimizes the risk for a given sample size. We consider estimators of the form $\overline{X} + b$ where b is an undetermined constant. Replacing $\overline{X}$ in (3) by $\overline{X} + b$, we find that the expected loss is:

$$(C_1+C_2)\frac{\sigma}{\sqrt{n}}\phi(\frac{\sqrt{n}b}{\sigma}) - C_2b + (C_1+C_2)b\Phi(\frac{\sqrt{n}b}{\sigma}),$$

where ϕ is the standard normal density function and Φ is its associated cumulative distribution. It is apparent that when $C_1 = C_2$ and $b = 0$ this reduces to the usual expression. Now we vary b so as to minimize this quantity and obtain a minimum at:

$$b = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\frac{C_2}{C_1 + C_2}).$$

This reduces to 0 when $C_1 = C_2$ and is negative when $C_2 < C_1$.

3. NONLINEARITIES Now we consider a symmetric but nonlinear loss function. Let

$$L_2 = \begin{cases} 0 & \text{if } |\overline{X} - \mu| < T \\ M & \text{if } |\overline{X} - \mu| > T. \end{cases}$$

Here M and T are positive constants. There is no penalty unless the error exceeds a certain size. In this case the expected loss is:

$$2M(1 - \Phi(\frac{\sqrt{n}T}{\sigma})).$$

A variation of this is:

$$L_3 = \begin{cases} 0 & \text{if } |\overline{X} - \mu| \leq T \\ M(|\overline{X} - \mu| - T) & \text{if } |\overline{X} - \mu| > T, \end{cases}$$

in which case the expected loss is:

$$2M\left(\sigma\phi(\frac{\sqrt{n}T}{\sigma}) - T(1 - \Phi(\frac{\sqrt{n}T}{\sigma}))\right).$$

We consider two more nonlinear loss functions.

Let

$$L_4 = \begin{cases} 0 & \text{if sign } \overline{X} = \text{sign } \mu \\ M & \text{otherwise.} \end{cases}$$

Although this loss function is asymmetric, for simplicity we do not permit a biased estimator, but retain $\overline{X}$ as our estimator. The expected loss is:

$$M\Phi(\frac{-|\sqrt{n}\mu|}{\sigma}),$$

which tends to 0 as n tends to ∞ and to $\frac{1}{2}$ as n goes to 0.

4. COMPARATIVE LOSS Finally consider:

$$L_5 = \begin{cases} M_1 & \text{if } |\overline{X} - \mu| > |X' - \mu| \\ -M_2 & \text{if } |\overline{X} - \mu| < |X' - \mu|. \end{cases}$$

Here X' is another independent estimator of μ, assumed to be normal and unbiased with standard error e'. The quantities M_1 and M_2 are assumed to be positive. Hence either a penalty or a reward may result from the estimation process. The expected loss is:

$$M_1 - \left((M_1 + M_2)\frac{2}{\pi}\arctan\frac{e'\sqrt{n}}{\sigma}\right).$$

This tends to M_1 for small values of n and to $-M_2$ for large values.

Reference

DeGroot, M. H. *Optimal statistical decisions*, McGraw-Hill, New York. 1970.