SECTION 14

Biased Sampling

Vardi (1985) introduced a far-reaching extension of the classical model for length-biased sampling. He solved the problem of estimating a distribution function based on several independent samples, each subject to a different form of selection bias. Using empirical process theory, Gill, Vardi and Wellner (1988) developed the asymptotic theory for generalizations of Vardi's method to abstract settings. They showed that the general model includes many interesting examples as special cases. This section presents a reworking of the ideas in those two papers. It is part of a study carried out by me in collaboration with Robert Sherman of Yale University.

The general problem is to estimate a distribution P on some set S using independent samples of sizes n_{i+} from distributions Q_i, for $i = 1, \ldots, s$, where the Q_i are related to P by means of known nonnegative weight functions $W_i(\cdot)$ on S:

$$\frac{dQ_i}{dP} = \pi_i W_i(\cdot) \qquad \text{where } \pi_i = 1/PW_i.$$

Of course the normalizing constants π_i, which we must assume to be finite and strictly positive, are unknown. For example, the W_i might be indicator functions of various subdomains of S. The problem is then one of combining the different samples in order to form an estimate of P over the whole of S. The difficulty lies in deciding how to combine the information from samples whose subdomains overlap.

For the general problem, to ensure that we get information about P over the whole domain, we must assume that the union of the sets $\{W_i > 0\}$ covers S.

Vardi suggested that a so-called nonparametric maximum likelihood estimator $\widehat{P}_n$ be used. This is a discrete probability measure that concentrates on the combined observations $x_1, x_2, \ldots$ from all s samples. If x_j appears a total of n_{ij} times in the i^{th} sample, the combined empirical measure $\widehat{Q}_n$ puts mass n_{+j}/n at x_j, where

$$n_{+j} = \sum_i n_{ij} \qquad \text{and} \qquad n = \sum_{i,j} n_{ij}.$$

The estimator $\widehat{P}_n$ modifies $\widehat{Q}_n$, putting at x_j the mass $\widehat{p}_j$ defined by maximization