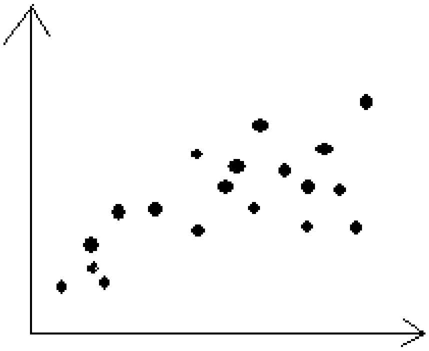
Optimal inclusion probabilities and estimators when sampling with varying probabilities

by
Daniel Thorburn,
Department of Statistics
Stockholm University
Ventspils August 2006

We discuss optimal allocation of inclusion probabilities in the presence of auxiliary information. In most situations one should use itboth when deciding the inclusion probabilities and in the estimator. The Horvitz-Thompson-(HT)-estimator is seldom optimal. We will only look at large sample theory and use a modelassisted designbased approach. The observations y_i; $i \in U$ are iid rv.

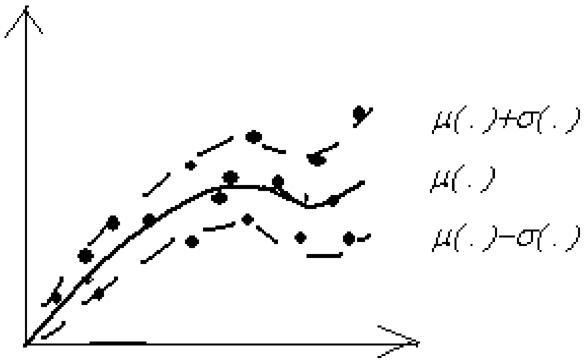
$$\mathrm{E}(Y_i|x_i) = \mu(x_i); \quad \mathrm{Var}(Y_i|x_i) = \sigma^2(x_i)$$

where μ is a nice function and x_i is the auxiliary information (perhaps multidimensional). In this talk we will assume that x is one-dimensional but generalisations to the multidimensional situation is straight-forward. If the population looks as follows (but with more data-points)$\mu(x_i)$ and $\sigma(x_i)$ may look as follows



We assume that μ varies slowly so that it can be estimated fairly well from the sample (e.g. with a moving average other kernel estimators or spline functions). Its estimator is denoted by $\mu^*(\bullet)$. A natural estimator of the total is then

$$\sum_U \mu^*(x_i) + \sum_s \frac{1}{\pi_i}(y_i - \mu^*(x_i))$$



The first part is a model-based estimator and the last part is an estimator of the design-bias. We call this estimator a generalised difference estimator. It is approximately unbiased for large samples and its variance can be estimated by the ordinary Sen-Yates-Grundy estimator

$$\sum_U \sum_U \frac{(\pi_{i,j} - \pi_i \pi_j)}{\pi_i \pi_j} (y_i - \mu(x_i))(y_j - \mu(x_j))$$

From the assumed independence the expected variance under the model this is approximately

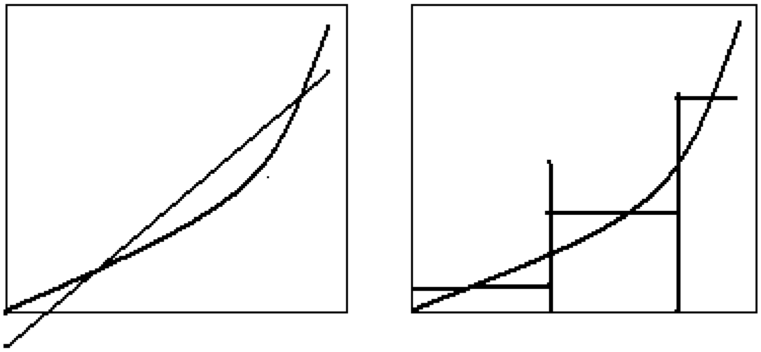
$$\sum_U \frac{1 - \pi_i}{\pi_i} \sigma^2(x_i)$$

Minimising this expression under a cost constraint gives that the inclusion probabilities (at least for large n and N) should be chosen

$$\pi_i \propto \frac{\sigma(x_i)}{c_i^{1/2}}$$

if the marginal costs are c_i. Those who have seen Neyman allocation recognises this expression.

We have not assumed anything about the procedure selecting the sample, i.e. the second order inclusion probabilities are unimportant, but $\mu(\bullet)$ must be estimated consistently and the above estimator used. With a more rigid model like polynomial regression with a bounded degree, a dependence may appear and the second order inclusion probabilities become important. In the next picture it is illustrated by a straight line regression and a curved mean value function. The residuals for two close x-values will mostly have the same sign. In that case one ought to choose a πps-design which spreads the observations so that $\pi_{i,j} < \pi_i * \pi_j$ if x_i and x_j are close. A similar effect occurs when stratifying with a limited number of strata or using splines with a bounded number of nodes.



With the above asymptotically optimal estimator the πps-method did not matter. The second order inclusion probabilities disappeared in the approximate variance. Then you can choose any sampe design like: systematic πps, Pareto-πps or Poisson-sampling. But if you intend to use a non-optimal method, like the ordinary HT-estimator, the design matters. Most commonly used πps-methods try to get a high independence i.e. they tries to mimic SRS, when $\pi(\bullet)$, is constant. In most cases this is a silly choice. It is e.g known that variants of systematic sampling and stratified sampling are better even when the inclusion probabilities are constant if the mean value function varies slowly. This holds here too. Systematic πps, ordering the observations after $\pi(x)$ or x or other sensible background variables is better. If you intend to use the HT-estimator and wants an asymptotically small variance you should avoid methods like Sampford, Pareto-πps or Poisson-sampling.

Systematic πps has, as we said, the advantage that you get a good and representative sample. But it has the disadvantage that the variance cannot be estimated exactly. But this is not a really a valid counterargument. Because everything you can estimate with e.g. SRS or Sampford, you can still estimate with a never larger variance. Thus you can give an upper bound on the variance-which is the variance with with Pareto-πps, say. There also exist list-sequential methods which have the same asymptotically optimal behaviour as systematic sampling and where the variance is possible to estimate. Another way to obtain fairly good sampling schemes is to use stratified sampling with decreasing strata widths.

If one uses a "silly" estimator like the HT-estimator or the regression estimator, the above inclusion probabilities are not optimal. Instead one must add a residual term to the variance getting

$$\pi_i \propto \frac{\left(\sigma^2(x_i) + (\mu(x_i) - \mathrm{E}(\mu^*(x_i)))^2\right)^{\frac{1}{2}}}{c_i^{1/2}}$$

where $\mathrm{E}(\mu^*(x_i))$ e.g. is 0 for the HT-estimator and the best regression line $\mathrm{E}(a^* + b^* x_i)$ for the regression estimator. If the variance function $\sigma^2(x_i)$ is small compared to the size, $\mu(x_i)$, this formula says that for the HT-estimator, inclusion probabilities proportional to size are optimal. This is a well-known fact, which often is used to motivate πps.