Statistical Methods MT2007

Analysis of Sampling Plans

Our notation follows Thompson (1997) most closely.

Suppose we have a finite population $\mathcal{U}$ of N units, with (possibly a vector of) measurements y_j, and we are interested in population mean μ or total T. A *sample* is a subset s of $\mathcal{U}$. A *sampling plan* or *(probability) sampling design* is then a probability distribution over S, the set of all subsets of $\mathcal{U}$. Note that this allows samples to be of different sizes, but in most cases we will only consider fixed size designs, those in which all samples s with positive probability are of size $n(s) = n$.

A sampling plan gives rise to *inclusion probabilities*: π_j is the probability that unit j will be included in the sample. Sometimes we may also consider the *joint inclusion probabilities*, where π_{jk} is the probability that both unit j and unit k are included in the sample.

The most important thing to remember is that with a sampling plan we do not have independent draws (sampling with replacement is strictly excluded by our definitions), so this needs to be taken into account in the analysis of the sample. If the *sampling fraction* n/N is small, IID sampling may be a reasonable approximation to sampling without replacement (but nowadays we can often work with the more complex exact formulae): however for more complex sampling plans we need to take the plan into account. Further details can be found in Lohr (1999), the `survey` package for R and in specialised software for survey analysis.

The analysis of a sampling plan normally amounts to providing an estimator of μ or T (equivalently if N is known) and an estimate of its variability (usually of its variance under the sampling plan).

Simple random sampling (SRS)

Simple random sampling is the fixed-size sampling plan for which all samples have equal probability. Its inclusion probabilities are obviously all equal, and (a little less obviously) $\pi_j = n/N$.

Let $\bar{y}$ denote the sample mean. Then $\bar{y}$ is an unbiased estimate of μ with variance

$$\operatorname{var} \bar{y} = (1 - f)S^2/n$$

where S^2 is the *population* variance, and $1 - f = 1 - n/N$ is known as the *finite population correction*. Note that this variance is smaller than that for IID sampling by the factor $(1 - f)$. Thus SRS is always more efficient that IID sampling.

Weighted sampling

The Horvitz–Thompson estimator

$$\hat{T} = \sum_{j \in s} y_j / \pi_j$$

is unbiased for any sampling plan. It has variance

$$\operatorname{var} \hat{T} = \frac{1}{2} \sum_j \sum_{k \neq j} (\pi_j \pi_k - \pi_{jk}) \left(\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k} \right)^2$$

which can be estimated by

$$\widehat{\operatorname{var}} \hat{T} = \frac{1}{2} \sum_j \sum_{k \neq j} (\pi_j \pi_k - \pi_{jk}) / \pi_{jk} \left(\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k} \right)^2$$

where the sums are confined to units in the sample, the Yates–Grundy–Sen variance estimator. Notice this requires $\pi_{jk} > 0$ if $\pi_i > 0, \pi_k > 0$, and that small selection probabilities lead to high variance (unless for units with small y_j).

The Horvitz–Thompson estimator is not always a very good estimator: from Basu (1971)

> The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records, and a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take $50y$ (where y is the present weight of Sambo) as an estimate of the total weight $Y = Y_1 + \cdots + Y_{50}$ of the 50 elephants.
>
> The circus statistician is horrified when he learns of the owner's purposive sampling plan. 'How can you get an unbiased estimate of Y this way?' protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally Sambo is selected and the owner is happy.
>
> 'How are you going to estimate Y?' asks the statistician. 'Why? The estimate ought to be $50y$ of course' says the owner. 'Oh! No! That cannot possibly be right' says the statistician. 'I recently read in an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz–Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators. 'What is the Horvitz–Thompson estimator in this case?' asks the owner, duly impressed. 'Since the selection probability for Sambo in our plan was 99/100', says the statistician, 'the proper estimate of Y is $100y/99$ and not $50y$.' 'And, how would you have estimated Y,' inquires the incredulous owner, 'if our sampling plan made us select, say. the big elephant Jumbo?' 'According to what I understand of the Horvitz–Thompson estimation method,' says the unhappy statistician, 'the proper estimate of Y would then have been $4900y$, where y is Jumbo's weight.'
>
> That is how the statistician lost his circus job (and perhaps became a teacher of statistics!).

Normally the inclusion probabilities are calculated from the sampling plan, but (as above) one might want to create a sampling plan to achieve specific inclusion probabilities. This is (for $n > 1$) somewhat tricky and several methods are available: see Thompson (1997, §2.8). One reason for doing so is that we have some measure x_j of 'size' for the units, and we want to sample bigger units more frequently. In particular, if $\pi_i \propto x_j$ we have *pps* sampling ('probability proportional to size').

Stratified sampling

Suppose that the units are divided into *strata* by one or more categorical variables. Stratified random sampling is to pick independently simple random samples of size n_h from each stratum.

This will in general be a weighted sample, with $\pi_j = n_h/N_h$ for units j in stratum h of size N_h, but will be equally weighted if the sizes of the stratum samples are chosen to be proportional to the stratum sizes (called *proportional allocation*). In this case the sampling design is more efficient that SRS if and only if the variances within the strata are on average smaller than the overall variance. Note that the HT estimator is only the sample mean in the case of proportional allocation (and in fact the sample mean is unbiased only in that case).

Proportional allocation is not necessarily optimal, in the sense of minimizing the variance of the estimator for a given sample size or cost. First, there may be different costs associated with sampling different strata (especially if stratification is on size), and also strata with different within-stratum variances S_h^2 should be sampled with different intensities. *Optimal allocation* has $n_h \propto N_h S_h/\sqrt{c_h}$, where c_h is the cost per unit in stratum h: *Neyman* allocation is the special case of equal costs.

Cluster sampling

In stratified sampling, we select a random sample from every group In cluster sampling we select groups at random, and include the whole group in the sample (or not). One can work out that cluster sampling is more efficient than SRS when the clusters are more heterogeneous than the population, but cluster sampling is chosen for low cost, not efficiency.

Real-life sampling plans are often complex versions of cluster sampling, with several stages of sampling. In two-stage cluster sampling, we first take an SRS of the groups (the *primary sampling units*), and then take SRSs from each of the selected groups. This has elements of both cluster and stratified sampling.

Ratio and regression estimators

Notice that in Basu's example there was a resource that was not used, a complete measurements of the weights three years ago (say x_j). Can we make use of this? Well, probably the total weight three years ago is a far better estimate than anything based on a single elephant

now! But suppose these had all been young elephants and so might be expected to have grown by a similar amount in the interim. Then we might use the sample to estimate the growth factor and apply it to the previous total weight.

Two 'obvious' estimators of the ratio T_y/T_x from a sample of size n are $r_1 = \frac{1}{n}\sum_{j\in s}(y_j/x_j)$ and $r_2 = \frac{1}{n}\sum_{j\in s} y_j / \sum_{j\in s} x_j$. Both are biased under SRS, but generally r_2 has a smaller bias and variance, and so is preferred. However, note that SRS is not necessarily the best sampling plan as we could use weighted sampling with $\pi_j \propto x_j$. In that case r_1 is proportional to the HT estimator.

Ratio estimators are used when approximately $y_j = rx_j$. Regression estimators are used when approximately $y_j = a + bx_j$, and a and b are estimated by linear regression of the sampled values of y_j on x_j. Naïve calculations show that they are more efficient than ratio estimators if and only if the regression model with intercept is a better fit than that without an intercept (which is almost always true). However, these do not take into account the estimation of the regression coefficients. For a more sophisticated approach see Thompson (1997, Chapter 5).

Model-based inference

All the results so far base inference on the randomness of the sampling plan. Alternatives are to assume a stochastic model for the generation of the y_j (called *superpopulation* modelling, since the actual population is itself regarded as a sample from a superpopulation). There are frequentist and Bayesian versions depending how the superpopulation model is fitted: inference is based on the joint distribution of the population and the sampling plan.

As ever, care is needed to build a suitably rich model. For example, superpopulation models often fail to take stratification into account and can be badly biased as a result. Both ratio and regression estimation can be viewed as model-based, but model-based inference gives slightly different standard errors.

Design-based inference cannot be used to adjust for non-response: model-based inference can although the assumptions needed can be unpalatable. Model-based inference can also be used with quota samples.

Additional References

Barnett, V. (1991) *Sample Survey Principles & Methods.* Edward Arnold.

Basu, D. (1971) An essay on the logical foundations of survey sampling, part one. In: V. P. Godambe and D. A. Sprott (eds) *Foundations of Statistical Inference*. Holt, Reinhart and Wilson, pp. 203–242.

Lohr, S. L. (1999) *Sampling Design and Analysis.* Duxbury Press.