Sampling design and statistical reliability of poverty and equity analysis using *DAD*

by

Jean-Yves Duclos

Département d'économique and CRÉFA-CIRPÉE, Université Laval, Canada

Preliminary version

 Corresponding address:
   Jean-Yves Duclos, Département d'économique, Pavillon de Sève, Université Laval, Québec, Canada, G1K 7P4; Tel.: (418) 656-7096; Fax: (418) 656-7798; Email: jyves@ecn.ulaval.ca

June 2002

Contents

1 Statistical inference with complex sample design

1.1 Sampling design

There exist in the population of interest a number of statistical units. These units are those on which we would like to know socio-demographic information such as their household composition, labor activity, income or consumption. For simplicity, we can think of these units as households or individuals. From an ethical perspective, it is usually preferable to consider individuals as statistical units of interest, but for some purposes (such as the distribution of aggregate household wellbeing) households may also be appropriate statistical units.

Since it is usually too costly to gather information on all statistical units in a large population, one would typically be constrained to obtain information on only a sample of such units. Distributive analysis is therefore usually done using survey data.

Since surveys are not censuses, we must take care to distinguish "true" population values from sample values. Sample differences across surveys are indeed due both to true population values and to sampling variability. Population values are generally not observed (otherwise, we would not need surveys). Sample values as such are rarely of interest: they would be of interest in themselves only if the statistical units which appeared by chance in a sample were also precisely those which were of ethical interest, which is usually not the case. Hence, sample values matter in as much as they can help *infer* true population values. The statistical process by which such inference is performed is called statistical inference. The sampling process should thus ideally be such that it can be used to make some statistically-sensible distributive analysis at the level of the population, not solely for the samples drawn.

Sampling errors thus arise because distributive estimates are typically made on the basis of only some of the statistical units of interest in a population. The fact that we have no information on some of the population statistical units makes us infer with sampling error the population value of the distributive indicators in which we are interested. There is an important element of randomness in the value of this sampling error. The error made when relying solely on the information content of one sample depends on the statistical units present in that sample. The drawing of other samples would generate different sampling errors. Because samples are drawn randomly, the sampling errors that arise from the use of these samples is also random.

Since the true population values are unknown, the sampling error associated

with the use of a given sample is also unknown. Statistical theory does, however, allow one to estimate the distribution of sampling errors from which actual (but unobserved) sampling errors arise. This nevertheless requires samples to be probabilistic, *viz*, that there be a known probability distribution associated to the distribution of statistical units in a sample. This also strictly means that there is absence of unquantifiable and subjective criteria in the choice of units. If this were not so, it would not be possible to assess reliably the sampling distribution of the estimators.

To draw a sample, a sampling base is used. A sampling base is made of all the sampling units (SU) from which a sample can be drawn. The base of sampling units – *e.g.*, the census of all households within in a country – is usually different from the entire population of statistical units – *e.g.*, the population of individuals, say). There are several reasons for this, an important one being that it is generally cost effective to seek information only within a limited number of clusters of statistical units, grouped geographically or socio-economically. This also facilitates the collection of cluster-level (*e.g.,* village-level) information.

A process of simple random sampling (SRS) draws sample observations randomly and directly from a base of sampling units, each with equal probability of selection. SRS is rarely used in practice to generate household surveys. Instead, a population of interest (a country, say) is often first divided into geographical or administrative zones and areas, called strata. The first stage of random selection then takes place from within a list of Primary Sampling Units (denoted as PSU's) built for each stratum. Within each stratum, a number of PSU's is then randomly selected. PSU's are often provinces, departments, villages, *etc.*. This random selection of PSU's provides "clusters" of information. The cost of surveying all statistical units un each of these clusters may be prohibitive, and it may therefore be necessary to proceed to further stages of random selection within each selected PSU.

For instance, within each province, a number of villages may be randomly selected, and within every selected village, a number of households may also be randomly selected. The final stage of random selection is done at the level of the last sampling units (LSU's). Each selected LSU may then provide information on all individuals found within that LSU. These individuals are *not* selected – information on all of them appears in the sample. They therefore do not represent LSU's in statistical terminology.

1.2 Sampling weights

Sampling weights (also called inverse probability, expansion or inflation factors) are the inverse of the sampling probabilities, *viz*, of the probabilities of a sampling unit appearing in the sample. These sampling weights are SU-specific. The sum of these weights is an estimator of the size of the population of SU's.

Samples are sometimes "self-weighted". Each sampling unit then has the same chance of being included in the survey. This arises, for instance, when the number of clusters selected in each stratum is proportional to the size of each stratum, when the clusters are randomly selected with probability proportional to their size, and when an identical number of households (or LSU) across clusters is then selected with equal probability within each cluster.

It is, however, common for the inclusion probability to differ across households. One reason comes simply from the complexity of sample designs, which makes differential sampling weights occur frequently. Another reason is that the costs of surveying SU's vary, which makes it more cost effective to survey some households (*e.g.*, urban ones) than others. Sampling precision can also be enhanced with differential probabilities of household inclusion. The aim here is to survey with greater probability those households who contribute more to the phenomenon of interest. It leads to a sampling process usually called sampling with "probability proportional to size".

Assume for instance that we are interested in estimating the value of a distribution-sensitive poverty index. The most important contributors to that index are obviously the poor households, and more precisely the poorest among them. It may be suspected that such poorest households are proportionately more likely to be found in some areas than in others. Making inclusion probabilities larger for households in these more deprived areas will then enhance the sampling precision of the estimator of the distribution-sensitive poverty index since it will gather more statistically informative data.

A reverse sample-design argument would apply for a survey intended to estimate total income in a population. The most important contributors to total income are the richest households, and it would thus be sensible to sample them with a greater probability. Yet one more consequence of the principle of "probability proportional to size" is the desirability of sampling with greater probability those households of larger sizes. Distributive analysis is normally concerned with the distribution of individual well-being. *Ceteris paribus*, larger-size households contribute more information towards such assessment, and should therefore be sampled with a greater probability (roughly speaking, with a probability propor-

tional to their size).

Omitting sampling weights in distributive analysis will systematically bias both the estimators of the values of indices and points on curves as well as the estimation of the sampling variance of these estimators. Including such weights will, however, help make the analysis free of biases. To see this, we follow Deaton (1998, p.45) and let Y be the population total of the x's, with a population of size N. An estimator of that population total is then given by

$$\hat{Y} = \sum_{i=1}^{N} t_i w_i x_i, \tag{1}$$

where t_i is the number of times unit i appears in a random sample of size n. Let π_i be the probability that unit i is selected each time an observation is drawn. Households with a low value of π_i will have a low probability of begin selected in the survey, relative to others with a higher π_i. Then, $\mathrm{E}[t_i] = n\pi_i = w_i^{-1}$ is the expected number of times unit i will appear in the sample, or roughly speaking for large n the probability of being in the sample. Hence,

$$\mathrm{E}\left[\hat{Y}\right] = \sum_{i=1}^{N} \mathrm{E}[t_i] w_i x_i = \sum_{i=1}^{N} x_i = Y \tag{2}$$

and $\hat{Y}$ is therefore an unbiased estimator of Y. An analogous argument applies to show that $\hat{N} = \sum_{i=1}^{N} t_i w_i$ is an unbiased estimator of population size N.

1.3 Stratification

The sampling base is usually stratified in a number of strata. The basic advantage of stratification is to use prior information on the distribution of the population, and to "partition" it in parts that are thought to differ significantly from each other. Sampling then draws information systematically from each of those parts of the population. With stratification, no part of the sampling base therefore goes unrepresented in the final sample.

To be more specific, a variable of interest, such as household income, often tends to be less variable within some stratum than across an entire population. This is because households within the same stratum typically share to a greater extent than in the entire population some socio-economic characteristics – such as geographical locations, climatic conditions, and demographic characteristics – that are determinants of the living standards of these households. Stratification

helps generate systematic sample information from a diversity of "socio-economic areas".

Because information from a "broader" spectrum of the population leads on average to more precise estimates, stratification generally decreases the sampling variance of estimators. For instance, suppose at the extreme that household income is the same for all households in a given stratum, and this, for all strata. In this case, supposing also that the population size of each stratum is known in advance, it is sufficient to draw only one household from each stratum to know exactly the distribution of income in the population.

1.4 Clustering (or multi-stage sampling)

Multi-stage sampling implies that SU's end up in a sample only subsequently to a process of multi-stage selection. "Groups" (or clusters) of SU's are first randomly selected within a population (which may be stratified). This is followed by further sampling within the selected groups, which may be followed by yet another process of random selection within the subgroups selected in the previous stage.

The first stage of random selection is done at the level of primary sampling units (PSU). An important assumption would seem to be that first-stage sampling be random and with replacement for the selection of a PSU to be done independently from that of another. There are many cases, however, in which this is not true.

1. First-stage sampling is typically made without replacement.

   This will not matter in practice for the estimation of the sampling variance if there is multi-stage sampling, that is, if there is an additional stage of sampling within each selected PSU. The intuitive reason is that selecting a PSU only reveals random and incomplete information on the population of statistical units within that PSU, since not all of these statistical units appear in the sample when their PSU is selected. Selecting that same PSU once more (in a process of first-stage sampling with replacement) does therefore reveal additional information, information different from that provided by the first-time selection of that PSU. This extra information is roughly of equal value to that which would have been revealed if a process of sampling *without* replacement had forced the selection of a different PSU.

   Hence, in the case of multi-stage sampling, first-stage sampling without replacement does not extract significantly more information than first-stage

sampling with replacement. It does not therefore practically lead to less variable estimators than a process of first-stage sampling with replacement.

If, however, there is no further sampling after the initial selection of PSU's, then a finite population correction (FPC) factor should be used in the computation of the sampling variance. This would generate a better estimate of the true sampling variance. If FPC factors are not used, then the sampling variance of estimators will tend to be overestimated. This means that it will be more difficult to establish statistically significant differences across distributive estimates, making the distributive analysis more conservative and less informative than it could have been.

2. Sampling is often *systematic*.

Systematic sampling can be done in various ways. For instance, a complete list of N sampling units is gathered. Letting n be the number of sampling units that are to be drawn, a "step" s is defined as $s = N/n$. A first sampling unit is randomly chosen within the first s units of the sampling list. Let the rank of that first unit be $k \in \{1, 2, \cdots, s\}$. The $n - 1$ subsequent units with ranks $k + s$, $k + 2s$, $k + 3s$,..., $k + ns$, then complete the sample.

If the order in which the sampling units appear in the sampling list is random, then such systematic sampling is equivalent to pure random sampling. If, however, this is not the case, then the effect of such systematic sampling on the sampling variance of the subsequent distributive estimators depends on how the sampling units were ordered in the sampling list in the first place.

(a) For instance, a "cyclical" ordering makes sampling units appear in cycles. "Similar" sampling units then show up in the sampling list at roughly fixed intervals. Suppose for illustrative purposes that the size of these intervals is the same as s. Then, systematic sampling will lead to a gathering of information on similar units (*e.g.*, with similar incomes), thus reducing the statistical information that is extracted from the sample. This will reduce the sampling precision of estimators, and increase their sampling variance.

(b) A cyclical ordering of sampling units suggests that there is more sampling-unit heterogeneity around a given sampling unit than across the whole sampling base (since information around sampling units is simply cyclically repeated across the sampling base). A more frequent phenomenon arises when adjacent sampling units show less heterogeneity than that

shown by the entire sampling base. A typical occurrence of this is when sampling units are ordered geographically in a sampling list. Households living close to each other appear close to each other in the list. Villages far away from each other are also far away in the sampling list. Since geographic proximity is often associated with socio-economic resemblance, the farther from each other in the list are sampling units, the more likely will they also differ in socio-economic characteristics.

Systematic sampling will then force units from across the entire sampling list to appear in the sample. Representation from *implicit* strata will thus be compelled into the sample. This will lead to a sampling feature usually called *implicit stratification*. Pure random sampling from the sampling list will not force such a systematic extraction of information, and will therefore lead to more variable estimators.

By how far implicit stratification reduces sampling variability depends on the degree of between-stratum heterogeneity which stratification allows to extract, just as for explicit stratification. The larger the heterogeneity of units far from each other, the larger the fall in the sampling variability induced by the systematic sampling's implicit stratification.

One way to account for and to detect the impact of implicit stratification in the estimation of sampling variances is to group pairs of adjacent sampling units into implicit strata. Assume again that n sampling units are selected systematically from a sampling list. Then, create $n/2$ implicit strata and compute sampling variances as if these were explicit strata. If these pairs did not really constitute implicit strata (because, say, the ordering in the sampling list had in fact been established randomly), then this procedure will not affect affect much the resulting estimate of the sampling variance. But if systematic sampling did lead to implicit stratification, then the pairing of adjacent sampling units will reduce the estimate of the sampling variance – since the variability within each implicit stratum will be found to be systematically lower than the variability across all selected sampling units.

Generally, variables of interest (such as living standards) vary less within a cluster than between clusters. Hence, *ceteris paribus*, multi-stage selection reduces the "diversity" of information generated compared to SRS and leads to a less informative coverage of the population. The impact of clustering sample ob-

servations is therefore to tend to decrease the precision of estimators, and thus to increase their sampling variance. *Ceteris paribus*, the lower the within-cluster variability of a variable of interest, the larger the loss of information that there is in sampling further within the same clusters.

To see this, suppose the extreme case in which household income happens to be the same for all households in a cluster, and this, for all clusters. In such cases, it is clearly wasteful to adopt multi-stage sampling: it would be sufficient to draw one household from each cluster in order to know the distribution of income within that cluster. More information would be gained from sampling from other clusters.

1.5 Impact of stratification, clustering, weighting and sampling without replacement on sampling variability

There are two modelling approaches to thinking about how data were initially generated. The first one, which is also the more traditional in the sampling design literature, is the finite population approach. The second approach is the super-population one: the actual population is a sample drawn from all possible populations, the infinite super-population. This second approach sometimes presents analytical advantages, and it is therefore also regularly used in econometrics.

To illustrate the impact of stratification and clustering on sampling variability, consider therefore the following "super-population model", based on Deaton (1998, p.56). Then,

$$x_{hij} = \mu + \underbrace{\alpha_h}_{\text{stratum effect}} + \underbrace{\beta_{hi}}_{\text{cluster effect}} + \underbrace{\epsilon_{hij}}_{\text{household effect}} . \qquad (3)$$

For simplicity, assume that the x_{hij} are drawn from the same number n of clusters in each of the L strata, and that the same number of LSU (or "households") m is selected in each of the clusters. The indices hij then stand for:

- $h = 1, ..., L$: stratum h

- $i = 1, ..., n$: cluster i (in stratum h)

- $j = 1, ..., m$: household j (in cluster i of stratum h).

For simplicity, also assume that α_h is distributed with mean 0 and variance σ_α^2, that β_{hi} is distributed with mean 0 and variance σ_β^2, and that ϵ_{hij} is distributed

with mean 0 and variance σ_ϵ^2. Assume moreover that these three random terms are distributed independently from each other.

1.5.1 Stratification

Say that we wish to estimate mean income μ. The estimator, $\hat{\mu}$, is given by

$$\hat{\mu} = (Lmn)^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{hij}. \tag{4}$$

Let

$$\hat{\mu}_h = (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{hij} \tag{5}$$

be the estimator of the mean of stratum h. Clearly, $\mathrm{E}[\hat{\mu}_h] = \mu + \alpha_h$ and $\mathrm{E}[\hat{\mu}] = \mu$ since by (4) and (5)

$$\mathrm{E}[\hat{\mu}] = (Lmn)^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{E}[x_{hij}] = (Lmn)^{-1}(Lmn)\mu = \mu. \tag{6}$$

and

$$\mathrm{E}[\hat{\mu}_h] = (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{E}[x_{hij}|\alpha_h] = (mn)^{-1}mn\,(\mu + \alpha_h) = \mu + \alpha_h. \tag{7}$$

Because of the independence of sampling across strata, we also have that

$$\mathrm{var}(\hat{\mu}) = \mathrm{var}\left(L^{-1} \sum_{h=1}^{L} \hat{\mu}_h\right) = L^{-2} \sum_{h=1}^{L} \mathrm{var}\left(\hat{\mu}_h\right). \tag{8}$$

The sampling variability of $\hat{\mu}$ is thus a simple average of the sampling variances of the L strata's $\hat{\mu}_h$.

Stratification can in fact be thought of as an extreme case of clustering, with the number of selected clusters corresponding to the number of population clusters, and with sampling being done without replacement to ensure that all population clusters will appear in the sample. Suppose instead that one were to select L strata randomly and with replacement, to make it possible that not all of the strata will be selected. This is in a sense what happens when stratification is dropped and clustering is introduced. Using (4) and (5), we then have that

$$\hat{\mu} = L^{-1} \sum_{h=1}^{L} t_h \hat{\mu}_h \tag{9}$$

where t_h is a random variable showing the number of times stratum h was selected. Then, denoting $\mu_h = \mu + \alpha_h$, we have approximately that

$$\hat{\mu} \cong \mu + L^{-1} \sum_{h=1}^{L} \left((t_h - \mathrm{E}\,[t_h])\, \mu_h + (\hat{\mu}_h - \mu_h)\, \mathrm{E}\,[t_h] \right). \tag{10}$$

and thus that

$$\mathrm{var}(\hat{\mu}) \cong L^{-2}\, \mathrm{var} \left(\sum_{h=1}^{L} \alpha_h t_h + \sum_{h=1}^{L} (\hat{\mu}_h - \alpha_h) \right) \tag{11}$$

since $\mu \sum_{h=1}^{L} t_h = \mu$ and $\mathrm{E}\,[t_h] = 1$. Assuming independence between $\hat{\mu}_h$ and t_h and between the $\hat{\mu}_h$, we have that

$$\mathrm{var}(\hat{\mu}) \cong L^{-2} \left\{ \mathrm{var} \left(\sum_{h=1}^{L} \alpha_h t_h \right) + \sum_{h=1}^{L} \mathrm{var}\,(\hat{\mu}_h) \right\}. \tag{12}$$

Since t_h follows a multinomial distribution, with $\mathrm{var}(t_h) = (L-1)/L$ and $\mathrm{cov}(t_h, t_i) = -1/L$, we find that

$$\mathrm{var} \left(\sum_{h=1}^{L} \alpha_h t_h \right) = \sum_{h=1}^{L} \alpha_h^2\, \mathrm{var}\,(t_h) + \sum_{h=1}^{L} \sum_{i \neq h} \alpha_h \alpha_i\, \mathrm{cov}(t_h, t_i) = L \sum_{h=1}^{L} \alpha_h^2 = L\sigma_\alpha^2. \tag{13}$$

Hence, using (11) and (13), we obtain

$$\mathrm{var}(\hat{\mu}) \cong L^{-2} \sum_{h=1}^{L} \mathrm{var}\,(\hat{\mu}_h) + L^{-1}\sigma_\alpha^2. \tag{14}$$

The last term in (14) is the effect upon sampling variability of removing stratification. The larger this term, the greater the fall in sampling variability that originates from stratification.

1.5.2   Clustering

Let us now investigate the effect of clustering on the sample variance, that is, on $\mathrm{var}\,(\hat{\mu}_h)$. We find:

$$\mathrm{var}\,(\hat{\mu}_h) \quad = \quad \mathrm{var} \left((mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{hij} \right)$$

$$
\begin{aligned}
&= (mn)^{-2} \operatorname{var}\left(\underbrace{\underbrace{\beta_{h1} + \ldots + \beta_{h1}}_{m \text{ times}} + \ldots + \underbrace{\beta_{hn} + \ldots + \beta_{hn}}_{m \text{ times}}}_{n \text{ times}} + \sum_{i=1}^{n}\sum_{j=1}^{m}\epsilon_{hij} \right) \\
&= (mn)^{-2} \operatorname{var}\left(m\sum_{i=1}^{n}\beta_{hi} + \sum_{i=1}^{n}\sum_{j=1}^{m}\epsilon_{hij} \right) \\
&= \frac{\sigma_{\beta}^{2}}{n} + \frac{\sigma_{\epsilon}^{2}}{mn}.
\end{aligned}
\tag{15}
$$

The first line of (15) follows by the definition of $\hat{\mu}_h$, and the second line follows from (3) – note that α_h is fixed for all of the x_{hij} in the same stratum h. The last line of (15) is obtained from the sampling independence between β_{hi} and ϵ_{hij}.

Hence, for a *per-stratum* given number of observations mn, it is better to have a large n to reduce sampling variability, namely, it is better to draw observations from a large number of clusters. The larger the cross-cluster variability σ_{β}^{2}, the more important it is to have a large number of clusters in order to keep $\operatorname{var}\left(\hat{\mu}_h\right)$ low. *Ceteris paribus*, for a given sample size and for a given $\sigma_{\beta}^{2} + \sigma_{\epsilon}$, the sampling variance of distributive estimators is smaller the smaller the between-cluster heterogeneity, σ_{β}^{2}, but the larger the within-cluster heterogeneity, σ_{ϵ}^{2}.

1.5.3 Finite population corrections

Sampling without replacement imposes that all of the selected sampling units are different. It therefore extracts on average more information from the sampling base than sampling with replacement, and ensures that the samples drawn are on average closer to the population of sampling units. Sampling without replacement therefore increases the precision of sample estimators. To account for this increase in sampling precision, a FPC factor can be used, although it complicates slightly the estimation of the variance of the relevant estimators.

Assume simple random sampling of n sampling units from a population of N sampling units. Thus, we have that $w_i = N/n$ for all of the n sample observations. To illustrate the derivation of an FPC factor in this simplified case, we follow Deaton (1998, p.42-44) and Cochrane (1977). An estimator $\hat{Y}$ of the population total Y of the x's is given by

$$
\hat{Y} = \frac{N}{n}\sum_{i=1}^{N} t_i x_i
\tag{16}
$$

where the random variable t_i indicates whether – and how many times – the population unit i was included in the sample. Taking the variance of (16), we find:

$$\text{var}\left(\hat{Y}\right) = \left(\frac{N}{n}\right)^2 \left(\sum_{i=1}^{N} x_i^2 \, \text{var}(t_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} x_i x_j \, \text{cov}(t_i, t_j)\right). \qquad (17)$$

Using (16) and (17), the distinction between simple random sampling *with* and *without* replacement is analogous to the distinction between a binomial and a multinomial distribution for the t_i. With sampling without replacement, the probability that any one population unit appears in the final sample is equal to n/N, *i.e.*, $\text{E}[t_i] = n/N$. Since t_i then takes either a 0 or a 1 value, it thus follows a binomial distribution with parameter n/N. The variance of t_i is then given by $\text{E}[t_i^2] - (n/N)^2 = n/N - (n/N)^2 = n/N\,(1 - n/N)$. The covariance $\text{cov}(t_i, t_j)$ can be found by noting that $\text{E}[t_i t_j] = \text{P}(t_i = t_j = 1) = n/N \cdot (n-1)/(N-1)$, and thus that

$$\text{cov}(t_i, t_j) = -\frac{n(N-n)}{N^2(N-1)}. \qquad (18)$$

Substituting $\text{var}(t_i)$ and $\text{cov}(t_i, t_j)$ into (17), and defining

$$S^2 = (N-1)^{-1} \sum_{i=1}^{N} \left(x_i - N^{-1}Y\right)^2, \qquad (19)$$

we find

$$\text{var}\left(\hat{Y}\right) = N^2(1-f)\frac{S^2}{n} \qquad (20)$$

where $1 - f = (N-n)/N$ is an FPC factor.

Take now the case of simple random sampling with replacement. We can then express t_i for any given population unit i as a sum of n independent draws t_{ij}, with $j = 1, \ldots, n$, each one t_{ij} indicating whether observation i was selected in draw j. Thus:

$$t_i = \sum_{j=1}^{n} t_{ij}. \qquad (21)$$

Since for any draw j, $\text{E}[t_{ij}] = 1/N$, the expected value of t_i is again n/N, but t_i may not take values greater than 1. The draws t_{ij} being independent, and each draw having a binomial distribution with parameter $1/N$, we have that

13

$$\text{var}(t_i) = \sum_{j=1}^{n} \text{var}(t_{ij}) = \sum_{j=1}^{n} \frac{1}{N}\left(1 - \frac{1}{N}\right) = \frac{n}{N}\left(1 - \frac{1}{N}\right), \qquad (22)$$

which is the variance of a multinomial distribution with parameters n and $1/N$. It can be checked that the covariance $\text{cov}(t_i, t_j)$ is given by $-n/N^2$. Substituting $\text{var}(t_i)$ and $\text{cov}(t_i, t_j)$ into (17) again, we now find

$$\text{var}\left(\hat{Y}\right) = N^2 \frac{(N-1)}{N} \frac{S^2}{n}. \qquad (23)$$

This is larger than (20): the difference between the two results equals

$$N^2 \frac{(n-1)}{N} \frac{S^2}{n} \qquad (24)$$

and depends on the magnitude of n relative to N. The larger the value of n relative to N, the greater the sampling precision gains that there are in sampling *without* replacement.

1.5.4   Impact of weighting on sampling variance

We follow once more the approach of Deaton (1998, pp.45-49) and Cochrane (1977). Suppose that we are again interested in estimating the variance of the estimator $\hat{Y}$ of a total Y, but for simplicity assume that sampling is done with replacement so that we can for now ignore FPC factors. $\hat{Y}$ is now defined as:

$$\hat{Y} = \sum_{i=1}^{N} t_i w_i x_i. \qquad (25)$$

Taking its variance, we find

$$\text{var}\left(\hat{Y}\right) = \sum_{i=1}^{N} w_i^2 x_i^2 \, \text{var}(t_i) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} w_i w_j x_i x_j \, \text{cov}(t_i, t_j). \qquad (26)$$

t_i follows once more a multinomial distribution, but now with $\text{var}(t_i) = n\pi_i(1 - \pi_i)$ and $\text{cov}(t_i, t_j) = -n\pi_i\pi_j$. Substituting this into (26), we find

$$\text{var}\left(\hat{Y}\right) = n^{-1}\left(\sum_{i=1}^{N} \frac{x_i^2}{\pi_i} - Y^2\right). \qquad (27)$$

14

To estimate (27), we can substitute population values by sample values and thus use the estimator

$$\widehat{\mathrm{var}}\left(\hat{Y}\right) = n^{-1} \left(\sum_{i=1}^{N} t_i w_i \frac{x_i^2}{\pi_i} - \left(\sum_{i=1}^{N} t_i w_i x_i \right)^2 \right). \tag{28}$$

Denote as $y_i = w_i x_i, i = 1, \ldots, n$ the n sample values of $w_i x_i$, and let $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$. Then, (28) leads to

$$\widehat{\mathrm{var}}\left(\hat{Y}\right) = \frac{n}{n-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 , \tag{29}$$

with the difference that a familiar $n/(n-1)$ small-sample correction factor has been introduced in (29) to correct for the small-sample bias in estimating the variance of the y_i.

1.5.5 Summary

The above calls to mind the importance for statistical offices of making available sample design information. This includes providing

- the sampling weights;

- stratum and PSU (cluster) identifying variables;

- information on the presence or not of systematic sampling (and thus of implicit stratification), including the relationship between the numbering of sampling units and the original ordering of these units in the sampling base;

- the finite population correction factors, namely, the size of the sampling bases, when appropriate.

Equipped with this information, distributive analysts can provide reliable estimates of the sampling precision of their estimators.

1.6 Formulae for computing standard errors of distributive estimators with complex sample design

We provide in this section a detailed account of the computation of sampling variances in *DAD*, taking full account of the sampling design. Let:

- $h = 1, \ldots, L$: the list of the strata (*e.g.* the geographical regions)

- $i = 1, \ldots, N_h$: the list of primary sampling units (PSU; *e.g.*, villages) in stratum h

- N_h: the population number of PSU in a stratum h

- n_h: the number of selected PSU in a stratum h

- M_{hi}: the population number of last sampling units (LSU) (*e.g.*, households) in PSU hi

- m_{hi}: the number of selected LSU in the PSU hi (for instance, the number of households from village hi that appear in the sample)

- q_{hij}: the number of observations in selected LSU hij (*e.g.*, the number of household members in a household hij whose socio-economic information is recorded in the survey, with each household member providing 1 line of information in the data file).

- w_{hij}: the sampling weight of LSU hij

- $M = \sum_{h=1}^{L} \sum_{i=1}^{N_h} M_{hi}$: the population number of LSU (*e.g.*, the number of households in the population)

- $m = \sum_{h=1}^{L} \sum_{i=1}^{n_h} m_{hi}$: the number of selected LSU (*e.g.*, the number of selected households that appear in the sample)

- X_{hijk}: the value of the variable of interest (*e.g.*, adult-equivalent income) for statistical unit $hijk$ in the population

- S_{hijk}: the size of statistical unit $hijk$ in the population (*e.g.*, if the statistical unit is a household, then S_{hijk} may be the number of persons in household hij, or alternatively the number of adult equivalents)

- $Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_{k=1}^{q_{hij}} S_{hijk} X_{hijk}$: the population total of interest

- x_{hijk}: the value of X (the variable of interest) that appears in the sample for sample observation h, i, j

- s_{hijk}: the size of selected sample observation $hijk$

- $\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \sum_{k=1}^{q_{hij}} w_{hij} s_{hijk} x_{hijk}$: the estimated population total of interest

- $\hat{M} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$: the estimated population number of LSU

- $y_{hij} = \sum_{k=1}^{q_{hij}} s_{hijk} x_{hijk}$: the relevant sum in LSU hij

- $y_{hi} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$: the relevant sum in PSU hi

- $\overline{y_h} = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$: the relevant mean in stratum h

The sampling covariance of two totals, $\hat{Y}$ and $\hat{Z}$ ($\hat{Z}$ being defined similarly to $\hat{Y}$) is then estimated by

$$\widehat{\text{cov}}_{SD}(\hat{Y}, \hat{Z}) = \sum_{h=1}^{L} n_h{}^2 (1 - f_h) \frac{S_h^{yz}}{n_h} \tag{30}$$

where

$$S_h^{yz} = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \overline{y_h})(z_{hi} - \overline{z_h}) \tag{31}$$

– note the similarity with (19) and (20) – and where f_h is a function of a user-specified FPC factor, fpc_h, for stratum h, such that,

- if a fpc_h is not specified by the user, then $f_h = 0$;

- if $fpc_h \geq n_h$, then $f_h = n_h / fpc_h$;

- if $fpc_h \leq 1$, then $f_h = fpc_h$.

Recall that setting $f_h \neq 0$ is useful only when the sampling design is of the form either of simple random sampling or of stratified random sampling with no subsequent sub-sampling within the PSU's selected. In both cases, sampling must have been done without replacement.

The variance $\widehat{V}_{SD}$ of $\hat{Y}$ is obtained from (30) simply by replacing $(z_{hi} - \overline{z_h})$ by $(y_{hi} - \overline{y_h})$.

An often-used indicator of the impact of sampling design on sampling variability is called the design effect, $deff$. The design effect is the ratio of the design-based estimate of the sampling variance ($\widehat{V}_{SD}$) over the estimate of the sampling variance assuming that we have obtained a simple random sample of m LSU without replacement. Denote this latter estimate as $\hat{V}_{SRS}$. Then,

$$deff = \frac{\widehat{V}_{SD}}{\hat{V}_{SRS}}. \tag{32}$$

17

For such a simple sampling design, we would have that

$$\hat{Y} = \frac{M}{m} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} y_{hij} \tag{33}$$

and, recalling (20), the sampling variance of $\hat{Y}$ would then equal

$$V_{SRS} = \left(\frac{M}{m}\right)^2 \text{var} \left(\sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} y_{hij} \right) = M^2 (1 - f) \frac{\text{var}(y)}{m}, \tag{34}$$

where $\text{var}(y)$ is the variance of the population y_{hij}, and where $f = m/\hat{M}$ if a FPC factor is specified for the computation of $\hat{V}_{SD}$, and $f = 0$ otherwise. V_{SRS} can then be estimated as follows:

$$\hat{V}_{SRS} = \hat{M}^2 (1 - f) \left(\frac{1}{m - 1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \frac{w_{hij}}{\hat{M}} \left(y_{hij} - \hat{Y}/\hat{M} \right)^2 \right). \tag{35}$$

Some of the above variables often take familiar forms and names:

- x_{hijk} can be thought of as an "individual-level" variable, such as height, health status, schooling, or own consumption. This variable is called the "variable of interest" in *DAD*. If x_{hijk} is indeed individual-specific, then s_{hijk} will not exceed 1 in most reasonable instances. Individual outcomes are, however, not always observed. Even if they are, we may sometimes believe that there is equal sharing in the household to which individuals belong. In those cases, x_{hijk} will typically take the form of adult-equivalent income or other household-specific measure of living standard.

- s_{hijk} gives the "size" of the sample observation $hijk$. This size may be purely demographic, such as the number of individuals in the unit whose living standard is captured by x_{hijk}. It may also be 1 even if $hijk$ represents a household, if we are interested in a household count for distributive analysis. But s_{hijk} may also be an ethical size, which depends on normative perceptions on how important the unit is in terms of some distributive analysis. Examples of such sizes include the number of adult-equivalents in the unit (if, say, we wish to assign individuals an ethical weight that is proportional to their "needs"), the number of families, the number of adults, the number of workers, the number of children, the number of citizens, the number of voters, *etc.*.

18

- q_{hij} is the number of sample observations or statistical units provided by the last sampling unit. This LSU may contain a grouping of households, of villages, *etc...* More commonly for the empirical analysis of poverty and equity, a LSU represents a household.

1.7 Computation of standard errors for complex estimators of poverty and equity

Most distributive estimators do not take the form of a simple sum of variable values for each of the sample observations, unlike the case of the estimator of a population total (recall for instance (16)). Instead, estimators of distributive indices take the following general form:

$$\hat{\theta} = g\left(\hat{\alpha_1}, \hat{\alpha_2}, \dots, \alpha_K\right), \tag{36}$$

where

- $\hat{\alpha}_k$ is asymptotically expressible as a sum of observations of $y_{k,i}$: $\alpha_k = \sum_{i=1}^{n} y_{k,i}$,

- θ can be expressed as a continuous function g of the $\alpha's$,

- n is the number of sample observations

- and $y_{k,i}$ is usually some k-specific transform of the living standard of individual or household i.

DAD uses Rao's (1973) linearization approach to derive the standard error of these distributive indices. Define $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)'$ and let G be the gradient of g with respect to the α's:

$$G = \left(\frac{\partial \theta}{\partial \alpha_1}, \frac{\partial \theta}{\partial \alpha_2}, \dots, \frac{\partial \theta}{\partial \alpha_K}\right)'. \tag{37}$$

A linearization of $\hat{\theta}$ then yields

$$\hat{\theta} \cong \theta + G'\alpha \tag{38}$$

The sampling variance of $\hat{\theta}$ can then be shown to be asymptotically equal to the variance of $\theta + G'\alpha$, which is equal to

$$G'VG \tag{39}$$

where V is the asymptotic covariance matrix of the $\hat{\alpha}_k$, given by

$$V = \begin{vmatrix} \text{var}(\hat{\alpha}_1) & \text{cov}(\hat{\alpha}_1, \hat{\alpha}_2) & \ldots & \text{cov}(\hat{\alpha}_1, \hat{\alpha}_K) \\ \text{cov}(\hat{\alpha}_2, \hat{\alpha}_1) & Var(\hat{\alpha}_2) & \ldots & \text{cov}(\hat{\alpha}_2, \hat{\alpha}_K) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{\alpha}_K, \hat{\alpha}_1) & \text{cov}(\hat{\alpha}_K, \hat{\alpha}_2) & \ldots & \text{var}(\hat{\alpha}_K) \end{vmatrix} \tag{40}$$

The gradient elements $\frac{\partial \theta}{\partial \alpha_1}, \frac{\partial \theta}{\partial \alpha_2}, ldots$, can be estimated consistently using the estimates $\frac{\partial \hat{\theta}}{\partial \hat{\alpha}_1}, \frac{\partial \hat{\theta}}{\partial \hat{\alpha}_1}, \ldots$ of the true derivatives. The elements of the covariance matrix can also be estimated consistently using the sample data, replacing $\text{var}(\hat{\alpha})$ by $\hat{\text{var}}(\hat{\alpha})$. Note that it is at the level of the estimation of these covariance elements that the full sampling design structure is taken into account.

1.8   Finite-sample properties of asymptotic results

DAD's methodology is based on asymptotic sampling theory. By this theory, all of *DAD*'s estimators are asymptotically normally distributed around their true population value. Although it is asymptotic in nature, *viz*, it is strictly valid only when the number of observations tends to infinity, we nevertheless expect this methodology to provide a good approximation to the true sampling distribution of *DAD*'s estimators for the usual sample sizes that are found in empirical analyses of poverty and equity.

It may nonetheless be instructive to compare the results of the above asymptotic approach to those of a numerical simulation approach like the bootstrap (a standard reference is Efron and Tibshirani (1993). The bootstrap (BTS) is a method for estimating the sampling distribution of an estimator which proceeds by re-sampling repetitively one's data. For each simulated sample, one recalculates the value of the estimator. One then uses the BTS distribution of simulated values of the estimators to carry out statistical inference. In finite samples, neither the asymptotic nor the BTS sampling distribution is necessarily superior to the other. In infinitely large samples, they are usually equivalent.

2 Confidence intervals and hypothesis testing

2.1 Basic principles

- θ is unknown;

- An estimator $\hat{\theta}$ is available for θ;

- By the law of large numbers and the central limit theorem, $\hat{\theta}$ is consistent and asymptotically normally distributed:

$$\hat{\theta} \sim N\left(\theta, \sigma_{\hat{\theta}}^2\right) \tag{41}$$

- Define $Z \sim N(0,1)$ and $P(Z > z_\alpha) = \alpha$;

- We do not know $\sigma_{\hat{\theta}}^2$, but we can estimate it by $\hat{\sigma}_{\hat{\theta}}^2$ – this is provided by *DAD*.

- Then, asymptotically, $\hat{\theta} \sim N\left(\theta, \hat{\sigma}_{\hat{\theta}}^2\right)$

- Using the sample value t of $\hat{\theta}$, we can do statistical testing and we can build confidence intervals.

2.2 Hypothesis testing

2.2.1 Procedures to follow:

1. Specify hypotheses to be tested and significance level of test (α).

2. Mention test statistic:

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} \tag{42}$$

3. State distribution of test statistic under the null hypothesis:

$$\frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}} \sim N(0,1) \tag{43}$$

4. Follow decision rule : reject $H_0 : \theta = \theta_0$ in favour of H_1 if

- $\frac{t - \theta_0}{\hat{\sigma}_{\hat{\theta}}} > z_\alpha$ if $H_1 : \theta > \theta_0$

- $\left| \frac{t-\theta_0}{\hat{\sigma}_{\hat{\theta}}} \right| > z_{\alpha/2}$ if $H_1 : \theta \neq \theta_0$

- $\frac{t-\theta_0}{\hat{\sigma}_{\hat{\theta}}} < -z_{\alpha}$ if $H_1 : \theta < \theta_0$

2.3 Confidence intervals

We wish to compute a $(1 - \alpha)$ confidence interval for the parameter θ.

- For a symmetric two-sided confidence interval, this is given by

$$\left[\hat{\theta} - \hat{\sigma}_{\hat{\theta}} z_{\alpha/2} \; , \; \hat{\theta} + \hat{\sigma}_{\hat{\theta}} z_{\alpha/2} \right] \tag{44}$$

- For a right-sided confidence interval, this is given by

$$\left[-\infty \; , \; \hat{\theta} + \hat{\sigma}_{\hat{\theta}} z_{\alpha} \right] \tag{45}$$

- For a left-sided confidence interval, this is given by

$$\left[\hat{\theta} - \hat{\sigma}_{\hat{\theta}} z_{\alpha} \; , \; \infty \right] \tag{46}$$

The actual confidence intervals are obtained by replacing $\hat{\theta}$ by t in the above. Some examples of z_{α}:

- $z_{0.10} = 1.28$ for a 80% confidence interval

- $z_{0.05} = 1.645$ for a 90% confidence interval

- $z_{0.025} = 1.96$ for a 95% confidence interval

- $z_{0.01} = 2.33$ for a 98% confidence interval

- $z_{0.005} = 2.575$ for a 99% confidence interval

References

[1] Asselin, Louis-Marie (1984), *Techniques de sondage avec applications à l'Afrique*, Centre canadien d'études et de coopération internationale, CECI/Gaëtan Morin éditeur.

[2] Cochrane, William G. (1977), *Sampling Techniques*, Wiley, New York.

[3] Deaton, Angus S. (1998), *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, John Hopkins University Press, 1998.

[4] Efron, Bradley and Robert J. Tibshirani (1993), *An introduction to the bootstrap*, Chapman and Hall, London.

[5] Howes, S., and J.O. Lanjouw (1998), "Does Sample Design Matter for Poverty Rate Comparisons?", *Review of Income and Wealth*, **44**, 99–109.

[6] Rao, C.R. (1973) *Linear statistical inference and its application*, Wiley, New York.