Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Variance estimation for complex statistics and estimators: Linearization and residual techniques

by Jean-Claude Deville

December 1999

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 193-203
Statistics Canada, Catalogue No. 12-001

Variance estimation for complex statistics and estimators: Linearization and residual techniques

Jean-Claude Deville [1]

Abstract

In a sample survey, in the absence of external information, the total of a variable is estimated using the Horvitz-Thompson estimator. Its variance is in turn estimated by calculating a fairly complex quadratic form, generally recursively. In this paper, this problem is assumed to be solved on the basis of a software capable of carrying out the calculation automatically. In the case of complex estimators (*i.e.*, of the calibration type), and in that of non-linear statistics (substitution estimators), it is shown that the same tool may always be used provided an appropriate artificial variable is chosen. In all cases, this artificial variable provides an estimation of the variance that is approximately unbiased and constructed using the influence function technique as well as some asymptotic postulates. Many examples are provided for the use of this technique: complex but explicit functions of totals (correlation coefficient), implicit functions of totals, calibrated estimators, fractiles and rank statistics, statistics derived from factorial methods.

Key Words: Variance estimation; Complex statistics; Linearization; Substitution estimators; Residual technique; Influence function; Implicit parameters; Fractiles; Rank statistics; Factorial analysis.

1. Introduction

The formulation of results in the form of confidence intervals is the goal (rarely reached) of all sample surveys. The most common procedure consists in estimating the variance of the statistics involved for the probability distribution induced by the sampling scheme (and sometimes, for the sake of simplicity, following fairly drastic assumptions called models). Then, following the assumption, rarely contradicted by the facts when the samples are large enough, that the statistic follows a normal distribution, a confidence interval symmetric about the point estimation is derived according to simple, standard procedures.

There is abundant literature dealing with this problem, before and after the benchmark work found in the book by Wolter (1985).

The goal of this paper is to show how simple tools can be used to effectively carry out a variance estimation in complex cases by means of a unique technique, *i.e.*, linearization. We will first describe the state of the art concerning the estimation of the variance of the Horvitz-Thompson estimator for a total. After providing a definition of "linearizable statistic" and a description of the concept of influence function analogous to that used in non-parametric statistics, we will introduce the class of functional substitution estimators shown to be linearizable under fairly general assumptions. We will show how the usual rules of differential calculus can be extended to linearized variables, and how, using step-by-step procedures, they can be used to calculate fairly easily the linearized variables of fairly complex statistics. Special attention will be given to statistics using quantiles as well as those linked to the most current multivariate analysis.

This procedure is the chief component of the POULPE software used at INSEE:

- Having a tool to estimate the variance of a total using a simple expansion estimator.
- Reverting to this case, using specially constructed variables, when using a complex estimator and/or when estimating a complex statistic.

2. General framework: Simple expansion estimator

Let us consider a population U of units $k, l, ...,$ for which a sample design is defined, *i.e.*, a probability distribution p that associates with any part s of U – the sample – a probability $p(s)$ of being selected. Using the latter, probabilities of inclusion $\pi_k (\pi_k = \sum_{s \ni k} p(s))$ are defined, as are probabilities of inclusion of order 2, π_{kl} for elements of s. It is then possible to use the Horvitz-Thompson estimator, $\hat{Y} = \sum_{k \in s} y_k / \pi_k$ of the total Y of a variable of interest y. It offers the advantage of being (almost) always available and unbiased, and its variance is easily calculated:

$$\mathrm{Var}(\hat{Y}) = \sum_U \pi_k (1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + 2 \sum \sum_U (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \qquad (2.1)$$

where the second sum extends to all pairs (k, l) of population U.

A useful estimator of the variance of $\hat{Y}$ is given by:

$$\widehat{\mathrm{Var}}(\hat{Y}) = \sum_s (1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + 2 \sum \sum_s \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}} \right) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \qquad (2.2)$$

[1] Jean-Claude Deville, Laboratoire de Statistique d'Enquête, Insee/Ensai/Crest, Campus de Ker-Lann, 35170-Bruz, France.

In practice, for large samples, this variance estimator is calculated recursively since the double sum which appears has a prohibitive number of terms. Moreover, the probabilities of inclusion π_{kl} can be calculated easily only in some rare simple cases.

In fact, all known sample designs boil down to a few simple schemes: Bernoulli or Poisson sampling, simple random sampling, systematic sampling and sampling with unequal probabilities of fixed size. For the first of these, there are some closed formulas providing a variance estimate based on the sum of squares. The same applies to systematic sampling given a few assumptions that are easily verified for selection order. Finally, the variance of sampling with unequal probabilities of fixed size, for many selection methods, can be approximated in an extremely fine and general manner using the following formula, applied in POULPE (Deville 1993):

$$\hat{V} = \frac{1}{1 - \sum a_k^2} \sum_s (1 - \pi_k) \left(\frac{y_k}{\pi_k} - A \right)^2 \qquad (2.3)$$

where $a_k = \dfrac{1 - \pi_k}{\sum\limits_s (1 - \pi_k)}$ and $A = \sum\limits_s a_k \dfrac{y_k}{\pi_k}$.

These simple schemes can be combined to provide arbitrarily complex designs by means of two operations, *i.e.*, stratification and multi-stage sampling (or sub-sampling).

In terms of stratification, a variance estimate of the grand total can be obtained by adding the variances of the estimators of stratum totals.

Multi-stage sampling can be obtained when the population is divided into sub-populations U_i called "primary units". A sample s_1 of the latter is selected on the basis of a sample design p_1 applied to the population of primary units. Then, for each $U_i (i \in s_1)$, a sample is selected using a design p_i / s_1. Conditionally on s_1, these designs are independent. From them are derived the probabilities of inclusion and the following variance formula:

$$\text{Var}(\hat{Y}) = \text{Var}\left(\sum_{s_1} \frac{Y_i}{\pi_i} \right) + E\left(\sum_{s_1} \frac{\text{Var}(\hat{Y}_i)}{\pi_i^2} \Big/ s_1 \right) \qquad (2.4)$$

where Y_i is the total of y for U_i, π_i its probability of inclusion, $\hat{Y}_i$ the estimator of Y_i for design p_{i/s_1}. Finally, if $V(Y_i;\ i \in s_1)$ is the variance estimator of $\sum_{s_1} Y_i / \pi_i$ and V_i a variance estimator of $\hat{Y}_i$ conditionally on s_1, then

$$\widehat{\text{Var}}(\hat{Y}) = V(\hat{Y}_i;\ i \in s_1) + \sum_{i \in s_1} \frac{V_i}{\pi_i} \qquad (2.5)$$

is a variance estimator of $\hat{Y}$ (Durbin 1953).

Naturally, for each stratum h or each primary unit i, the sample survey p_h or p_i / s_1 can itself be stratified or become a multi-stage sampling. However, in all cases, the repetitive and recursive use of the abovementioned rules makes it possible to calculate a variance estimator using simple elements based on the sum of squares. For surveys carried out among people, it is customary for a sample design to comprise three to five selection stages.

This means that the quadratic form (2.2) can be calculated mechanically, without however any explicit computation of the terms involved in the double sum found in the formula.

To complete this overview, it should be noted that a sample is frequently selected in several stages, normally two or three. This means that a sample s is selected and used as a reference population for the selection of a second sample r, using a sample design $q(r/s)$. If it is controlled by the statistician, this design is generally a stratified design with simple random sampling for each stratum. Otherwise, the design is described using a response model that makes it possible to formalize a reweighting procedure for non-response. In all cases, there are second-stage probabilities of inclusion P_k and P_{kl} describing the inclusion in r of the unit k or of the pair (k, l). The expansion estimator is $\hat{Y}_{\text{exp}} = \sum_r y_k / \pi_k P_k$.

Its variance can be calculated fairly easily, and is estimated using the expression:

$$\widehat{\text{Var}}(\hat{Y}_{\text{exp}}) = \sum_r \sum \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}} \right) \frac{1}{P_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \sum_r \sum \left(1 - \frac{P_k P_l}{P_{kl}} \right) \frac{y_k}{P_k \pi_k} \frac{y_l}{P_l \pi_l} \qquad (2.6)$$

with $\pi_{kk} = \pi_k$ and $P_{kk} = P_k$.

In spite of a few difficulties, this variance estimator can be calculated mechanically while avoiding the prohibitive double sums. The same applies to three-stage designs which occur when a non-response stage is added to a second stage controlled by the statisticians. Such procedures are used in POULPE.

Thus, the Horvitz-Thompson estimator (or its extension, the expansion estimator) has a variance that takes on the quadratic form $Q(y_k;\ k \in U)$. The latter can be estimated without bias (or eventually with negligible bias) using the recursive calculation of a quadratic form $\hat{Q}(y_k;\ k \in s)$ (where s now represents the final sample, no matter how many stages were needed to obtain it). In the following, we will assume the availability of an "automatic" method of calculating this quadratic form.

3. Complex statistics and asymptotic postulates

We will show that it can also be applied when we use a more refined estimator than HT (involving external auxiliary information), *e.g.*, for complex statistics (means, quantile ratios, complex indexes such as GINI, the coefficient of an econometric model, a principal component analysis factor).

The results we will provide are "asymptotic" approximations. As in Isaki and Fuller (1982) or in Deville and Särndal (1992), we postulate the following scheme: in a series of sampling problems indexed by an integer v (which we will suppress so as not to overload the notation), the size N of the population tends towards infinity as does the size n (or the size expectation) of the sample. For each v, we thus also have a sample design, the associated HT estimator, a vector of invariable fixed size for variables x_k of X estimated using $\hat{X}$.

The three following propositions are postulated:

- $N^{-1}X$ has a limit. (3.1)
- $N^{-1}(\hat{X} - X)$ converges in terms of probability towards zero. (3.2)
- $n^{-1/2}N^{-1}(\hat{X} - X)$ has as a limit a multidimensional normal distribution. (3.3)

The first postulate formalizes the concept of a series of populations of increasing size extracted from a parent continuous distribution. This can also be interpreted as if the population were an i.i.d. sample of a certain infinite superpopulation. The other two postulates relate to the convergence of the HT estimator and to the fact that it leads to a central limit theorem. In practical terms, these postulates are satisfied in many cases, given certain technical assumptions: simple random sampling (Hájek 1964), Poisson sampling and randomized systematic sampling - (Rosen 1972), stratified design with the number of strata tending towards infinity (immediate application of Lindeberg conditions).

In reality, however, we can never tell whether, for example, the design used involves a number of strata tending towards infinity! What the asymptotic postulates mean is simply that certain magnitudes (technically those which are $O_p(n^{-1/2})$) are considered "small" and that the product of the two "small" quantities is a "negligible" (and therefore neglected!) quantity.

On the basis of these postulates, we will show how certain estimators and certain non-linear statistics can be approximated using HT statistics having the form $\sum_s z_k/\pi_k$ for well-chosen variables z_k.

4. Substitution estimators and functionals

Let us now consider a fairly general class of non-linear statistics of the finite population based on the concept of a measurement functional, as well as their substitution estimators.

With each unit k of the population U there is associated a point x_k of $\boldsymbol{R}^p$ for the p problem variables of interest to us. The population U is thus represented by the measure M having a unit mass in each of the points x_k.

This measure is positive, discrete and finite, and its total mass has a value of N. We assume that all the x_k are separate, without loss of generality (we can always add a dimension which is the "rank" of k in arbitrary numbering). For any variable $y_k = y(x_k)$, we thus have $\int y\,dM = \sum_U y_k$.

From an asymptotic point of view, the series of populations is a series of measures on $\boldsymbol{R}^p$. According to the first asymptotic postulate, this series behaves as if we were dealing with i.i.d. selections for a fixed probability distribution on $\boldsymbol{R}^p$.

A functional $T(M)$ associates with any measure of a class containing at least the point measurements, a real number or a vector. We also assume that all the functionals of interest are homogeneous, *i.e.*, there is a positive real number α dependent on T such that $T(tM) = t^{\alpha}T(M)$ for any positive real number t. A total is a homogeneous functional of level 1, a mean of level 0, a sum having a double index of level 2. Being limited to homogeneous functionals is not too cumbersome in practical terms.

Now let $\hat{M}$ (estimator of M) denote the measure allocating a weight w_k to any point x_k for k in s and zero to any other point, regardless of the origin of the weights (Horvitz-Thompson or calibration).

Definition: The substitution estimator of a functional $T(M)$ is $T(\hat{M})$.

In the case of a total, this definition should not be surprising since $T(\hat{M}) = \int x d\hat{M}(x) = \sum_s x_k w_k$. For "ordinary" complex statistics (ratios, means or indexes, for example), this represents the common practices of survey operations. The same applies to statistics of rank, with finer points having more to do with the estimation of the distribution function than the estimation of the fractiles (see for example Chambers, Dorfmann and Hall 1992).

A fairly general class of parameters linked to the finite population can be obtained using implicit equations which define them. Such is the case, for example, for the adjustment of a parametric model at the population level leading to an "estimating equation" derived from a broad adjustment principle (maximizing likelihood, minimizing chi-2 or "moment" methods or "generalized moment" methods).

This form of writing introduces the (eventually multidimensional) model parameter as a functional of M. Its estimator (in the sense of sampling) is the same functional for $\hat{M}$. Thus, the estimation of the least squares in the linear model is written as follows:

$$B = \arg\ \mathrm{Min} \sum_U q_k\,(y_k - x_k'B)^2.$$

The estimation (sampling) of B is $\hat{B} = \arg\ \mathrm{Min}\sum_s w_k q_k\,(y_k - x_k'B)^2$. The use of $\hat{B}$ (rather than an estimator for a model conditional upon the sample) is much more robust, and correctly accounts for the fluctuations of sampling on the result (on this point, see Binder 1983, Binder and Patak 1994, or Binder and Kovačević 1997).

Generally speaking, an "estimating equation" at the population level will be written as $T(M, \lambda) = 0$ where T is a functional of dimension p parametered by vector λ lso of dimension p. This equation will be assumed to have a unique solution for fixed M. The substitution estimator of λ is the solution of the (estimating) equation $T(\hat{M}, \hat{\lambda}) = 0$.

5. Linearizable statistic

Let us consider some statistics S dependent on the observations $(x_k;\ k \in s)$ (in fact a series of statistics defined in each of the sampling problems within the asymptotic framework). S is said to have a probability of order $f(n)$ (where f is some positive function of n), and we write $S = O_p(f(n))$ if for any $\varepsilon > 0$ there is a constant C such that

$$\Pr\left(\frac{\|S\|}{f(n)} \geq C\right) \leq \varepsilon.$$

In other words, the survivorship functions of variables $\|S\|/f(n)$ are uniformly overestimated by the survivorship function represented as $(C(\varepsilon),\ \varepsilon)$.

The third asymptotic axiom (central limit axiom!) can therefore be written as $N^{-1}(\hat{X} - X) = O_p(n^{-1/2})$, and the second as $N^{-1}\hat{X} = O_p(1)$. In the rest of this paper, we will use more or less implicitly the following well-known result (see for example Billingsley 1969):

Result: If a statistic S converges towards a certain distribution, and if $(S - T) = O_p(f(n))$ with $f(n) \to 0$, then T converges towards the same distribution. Specifically, S and T have the same limit variance.

The statistic S is said to be of degree α if $N^{-\alpha}S$ tends towards a limit. Clearly, for example, a HT estimator is of degree 1, a ratio of HT estimators is of degree 0 (or homogeneous). The third asymptotic postulate states that $nE(\hat{X} - X)^2$ is of degree 2. The substitution estimator of a homogeneous functional of degree α is a statistic of degree α.

The following definition can now be formulated:

Definition: A statistic S of degree α is linearizable if there is a synthetic variable z_k (known as the linearized variable of S) such that the variance of $\hat{Z}$ is equivalent to that of S in the sense that $n^{1/2}(N^{-\alpha}S - N^{-1}\hat{Z}) = O_p(f(n))$ with $f(n) \to 0$. In general, we will almost always have $f(n) = n^{-1/2}$.

In practice, this means that the variance, and therefore a confidence interval, will be estimated for S on the basis of the variance of $\hat{Z}$ (whether or not it is a HT estimator).

Note, on the other hand, that the definition does not imply the uniqueness of the variable linearizing a statistic. Specifically, the approximation contained in the definition can be more or less fine at two levels: that of the convergence speed $f(n)$, and, for an equal speed, that of the increment $C(\varepsilon)$.

Generally speaking, however, the linearized variable z_k cannot be computed explicitly by means of data from the sample. We are then led to replace z_k by an approximation $\tilde{z}_k$ using certain statistics estimated on the basis of the sample. This occurs in the most elementary cases. The matter of the legitimacy of this approximation must be dealt with, and this can only be done within an asymptotic framework.

Result: If quantities $\tilde{z}_k$ depend regularly on a fixed, finite number of estimated parameters, then the variance estimators $\hat{Q}(z_k;\ k \in s)$ and $\hat{Q}(\tilde{z}_k;\ k \in s)$ are equivalent, *i.e.*, their difference as normalized by factor n / N^2 is an asymptotically negligible quantity.

Proof: By "regularly" is meant that $\tilde{z}_k = z_k + \xi_k'(\hat{\Gamma} - \Gamma) + O_p(\|\hat{\Gamma} - \Gamma\|^2)$, where Γ is the p vector of the parameters, $\hat{\Gamma}$ is its vector of estimators and ξ_k is a $p-$variable. The asymptotic postulates tell us that $n / N^2 Q(z_k)$ converge towards a finite quantity just as $n / N^2 Q(z_k,\ \xi_k) = \sum_{k,\,l} \Delta_{kl} z_k \xi_l$ if the quadratic form is made explicit, and that $n / N^2 Q(\xi_k)$. We then have:

$$Q(\tilde{z}_k) = Q(z_k) + 2Q(z_k,\ \xi_k)'(\hat{\Gamma} - \Gamma) + O_p(\|\hat{\Gamma} - \Gamma\|^2).$$

As $\|\hat{\Gamma} - \Gamma\| = O_p(n^{-1/2})$, we obtain the result.

When the number of estimated parameters tends towards infinity, the situation is not perfectly clear. In practical terms, obviously, what is meant by the number of estimated parameters tending towards infinity? Theoretically, moreover, there are some difficulties as can be seen from the following two contradictory examples:

Example: Poststratification. We assume the poststrata defined on the basis of a numerical variable, and we construct m adjacent poststrata, each of which comprises about n / m surveyed units. Here vector Γ is that of the m means of poststrata $\bar{Y}_h$, $h = 1$ at m. If m increases with n, each estimated parameter $\hat{\bar{Y}}_h$ is such that $\hat{\bar{Y}}_h - \bar{Y} = O_p((n/m)^{-1/2})$. Then $\|\hat{\Gamma} - \Gamma\|$ is of the order of $m^{3/2}n^{-1/2}$. Taking $m = n^{\alpha}$ with $\alpha < 1/3$, the previous result and its proof remain valid.

Counter-example: For the estimation of inequality indexes, we are led to use statistics such as $S = \sum_s y_k \hat{F}(y_k)$ where $\hat{F}$ is an estimation of the distribution function of variable y. If R_k denotes the rank of y_k in the population, we could imagine that $z_k = 1 / N\, y_k R_k$ is a linearized statistic for S. This is completely false (Deville 1997).

The difference with respect to the previous example rests in the fact that S uses an estimated parameter per sampled unit, in which case anything can happen! The general procedure for dealing with such statistics will be described below in section 12.

6. Influence function of a functional and asymptotic variance of the substitution estimator

Definition: The influence function (if it exists) of a functional T is:

$$IT(M;\ x) = \lim_{t \to 0} \frac{1}{t} (T(M + t\delta_x) - T(M))$$

where δ_x denotes the unit mass assumed at point x.

Comment: This definition is slightly different from that used in the field of robust statistics (Hampel, Ronchetti, Rousseeuw, and Stahel 1985). It is made necessary by the fact that the total mass of M is variable, and often unknown in a statistical problem. It is nothing more than the differential as viewed by Gateaux for a Dirac mass assumed at a point x.

The essential point of this paper can now be formulated:

Result: Under broad assumptions, the substitution estimation of a functional $T(M)$ is linearizable. A linearized variable is $z_k = IT(M;\ x_k)$ where IT is the influence function of T in M.

Comment: The influence function can thus be used to estimate the variance of $T(\hat{M})$. This being said, very often the influence function includes in its definition certain functionals of M (*e.g.*, a ratio or a mean). We are thus led to choose an estimation of the influence function itself in order to compute the variance estimation. This choice is not necessarily unique.

Proof of the result: Let us provide the space of measurements on $\boldsymbol{R}^q$ with a metric d accounting for the convergence: $d(M_1, M_2) \to 0$ if and only if $N^{-1}(\int y\ dM_1 - \int y\ dM_2) \to 0$ for any variable of interest y. The asymptotic postulates mean that $d(\hat{M}/N,\ M/N)$ tends towards zero. We can visibly ensure that $d(\hat{M}/N,\ M/N)$ is $O_p(1/\sqrt{n})$ according to the third postulate. Now, let us assume that T can be derived in accordance with Fréchet, *i.e.*, for any direction of the increase, in the space of "useful" measures provided with the abovementioned metric. Thus we have:

$$N^{-\alpha}(T(\hat{M}) - T(M)) = \frac{1}{N} \sum_U z_k(w_k - 1) + o\left(d\left(\frac{\hat{M}}{M}, \frac{M}{N} \right) \right).$$

The result is that:

$$\sqrt{n} N^{-\alpha} (T(\hat{M} - T(M))) = \frac{\sqrt{n}}{N} \sum_U z_k (w_k - 1) + o_p(1).$$

Thus, according to the third postulate, the variance of the second member tends towards a limit, that of $n/N^2\, \mathrm{Var}(\hat{Z})$, and the result is obtained.

7. Examples and computing rules for influence functions

Example 1: If T is the total $T = \int x\ dM(x)$ of a variable, the influence function of T is this variable itself: $IT(M,\ x) = x.$ Specifically, if $x = 1,\ T = N$ the population size. The influence function is then constant, and its value is 1.

The rules of composition for influence functions are copied from those of differential calculus:

Rule 1: If f is a derivable function defined on the space of values for T a vector function, we have:

$$I(f(T)) = Df(T)\ IT$$

(where Df represents the matrixes of the partial derivatives of f).
The proof is immediate.

Example 2: $f(T) = 1/T$ and $T = \int x\, dM,$ scalar total. The influence function is $-x/T^2$.

Rule 2: If S and T are two functionals, we have:

$$I(S + T) = IS + IT \quad \text{and} \quad I(ST) = S\ IT + T\ IS.$$

If T and S have vector values, and if H is a matrix, we have, when the products are defined:

$$I(HT) = H\ IT \quad \text{and} \quad I(S'HT) = (IS)'HT + S'H\ IT.$$

Example 3: $R = \int y\ dM / \int x\ dM = Y/X$ a ratio of two totals. The influence function is:

$$\frac{y}{X} - \frac{Yx}{X^2} = \frac{1}{X}(y - Rx).$$

For a mean $\bar{Y} = \int y\ dM / \int dM,$ the influence function is therefore: $1/N\,(y - \bar{Y}),$ which is the usual definition, or just about, given in the robustness theory (Lecoûtre and Tassi 1987).

Rule 3: Let $S_i (i = 1, ..., q)$ denote scalar functionals and $S = \prod_{i=1}^q S_i$. We have:

$$IS = S\left(\sum_{i=1}^{q} \frac{IS_i}{S_i} \right).$$

Proof: $I(\mathrm{Log} S) = \dfrac{IS}{S}.$

Now let $T(\lambda) = T(M,\ \lambda)$ denote a family of functionals depending regularly on a parameter λ that varies in a domain of $\boldsymbol{R}^q$, with Λ a measure on this domain. This leads to:

Rule 4: $I(\int T(\lambda)\ d\Lambda(\lambda)) = \int IT(\lambda)\ d\Lambda(\lambda).$ This is elementary.

Note: The persistency conditions include the possibility of reaching the limit under the integration sign.

If, moreover, φ is a function of $\boldsymbol{R}^q$ in the domain of T measurable for all measures M of interest, we proceed as follows:

Rule 5: $I(\int T(\varphi(x))\, dM(x); \xi) = T(\varphi(\xi)) + \int IT(M, \varphi(x); \xi)\ dM(x)$.

Proof: (provided as an example): Let $S(M) = \int T(M, \varphi(x))\, dM(x)$. We have:

$$\frac{1}{t}[s(M + t\delta_\xi) - S(M)] =$$

$$\frac{1}{t}\int [T(M + t\delta_\xi, \varphi(x)) - T(M, \varphi(x))]\ dM(x)$$

$$+ \delta_\xi(T)\,(T(M + t\delta_\xi\ ), \varphi(x)).$$

The second term tends towards $T(M, \varphi(\xi))$ whenever t tends towards zero. The former can be written as follows:

$$\left[\int IT(M, \varphi(x);\ \xi) + R_{M,\, \varphi(x);\, \xi}(t)\right]\ dM(x)$$

where R is a quantity that tends towards zero (it may be assumed that the convergence is consistent at x). The result is derived immediately.

Let us now assume that $T(\lambda)$ is a functional with values in $\boldsymbol{R}^q$, regular at λ. Specifically, then, matrix $\partial T/\partial\lambda$ is reversible for any M, and, for fixed M, application $\lambda \to T(\lambda)$ is one-one and allows for a partial reciprocal function. Equation $T(\lambda) = T_0$ therefore has a unique solution for any M, defining a functional $\lambda(M)$.

Rule 6: The influence function of $\lambda(M)$ is:

$$I\lambda(M;\ \xi) = -\frac{\partial T}{\partial\lambda}(M, \lambda)^{-1}\ IT(M,\ \lambda;\ \xi).$$

Proof: $T(M + t\delta_\xi, \lambda(M + t\delta_\xi) - T(M, \lambda)) = 0$, hence:

$$IT(M,\ \lambda;\ \xi) + \frac{\partial T}{\partial\lambda} I\lambda(M;\ \xi) = 0.$$

This rule may also be needed:

Rule 7: Let S denote a functional in $\boldsymbol{R}^q$, and let T_λ denote a family of functionals regularly indexed by $\lambda \in \boldsymbol{R}^q$. We have:

$$I(T_s) = IT_{\lambda/\lambda=s} + \left(\frac{\partial T}{\partial\lambda}\right)_{\lambda=s} IS.$$

Proof: Writing everything, we have $I(T_S) = IT(S(M), M)$. The rest is obvious.

Note, finally, the interesting link between the influence function and the functional from which it is derived:

Result: If T is homogeneous of degree α we have:

$$\int IT(M; x)\ dM(x) = \sum_U IT(M; x_k) = \alpha T(M).$$

The specific case $\alpha = 0$ shows that any homogeneous functional has a zero-sum influence.

Proof: We have:

$$\frac{T((1 + h)M) - T(M)}{h} = \frac{((1 + h)^\alpha - 1)}{h} T(M)$$

from the definition of homogeneity. The result follows from the linearity of the derivation as interpreted by Gateaux and the definition of influence function.

8. Applications: Functions of totals

We have already seen that, for linear functionals, $T(M) = \sum_U y_k = \int ydM$, the influence function is y_k itself. The application of the notion of influence function becomes redundant. It will be noted, however, that it is in no way asymptotic.

Function of totals: If X is a vector of totals, the influence function of X is, naturally, the vector x_k of the variables making up X. As a result, if $T(M) = f(X) = f(\int xdM)$, the influence function of T is:

$$IT(M;\ x_k) = f'(X)\ .\ IT(\int xdM) = f'(X)\ .\ x_k$$

where $f'(X)$ is the row vector of the partial derivatives of f with respect to the coordinates of X taken at point X. We are led naturally to the classical result of Woodruff (1971).

In line with the above, the substitution estimator of $f(X)$ is $f(\hat{X})$. Its approximate variance is that of $f'(X)\ .\ x_k$, and it is numerically approximated by $f'(\hat{X})\ .\ x_k$ in compliance with common practices.

Example: The ratio $R = f(X, Y) = Y/X$ of two scalar totals is estimated using $\hat{R} = \hat{Y}/\hat{X}$. This statistic (of degree 0) allows as a linearized variable $z_k = 1/X(y_k - Rx_k)$. To numerically compute the variance estimation of $\hat{R}$, we use the approximation $\tilde{z}_k = 1/\hat{X}(y_k - \hat{R}x_k)$, an expression which depends on $\hat{Y}$ and $\hat{X}$ and therefore on s; $\tilde{z}_k$ is therefore not a linearized variable as understood in the definition.

Example: Ratio estimator.

It is $\hat{Y}_{\text{rat}} = (X/\hat{X})\hat{Y}$. If we refer to the previous example, the linearized variable of $\hat{Y}_{\text{rat}}$ is $y_k - Rx_k$, approximated by $y_k - \hat{R}x_k$. And yet it could also be said that the estimated variance of $\hat{Y}_{\text{rat}}$ must be equal to X^2 times the estimated variance of $\hat{R}$, which leads to the approximation $X/\hat{X}(y_k - \hat{R}x_k)$ which has many times been deemed more interesting than the previous one. This example shows that the choice of linearized variable is not necessarily unique once external information is used.

Nevertheless, one of the advantages of the influence function approach is to provide computations fairly easily in apparently complex cases.

Example: The correlation coefficient between x and y is written as follows:

$$\rho = \frac{\sum_U x_k y_k - \frac{1}{N}\sum_U x_k \sum_U y_k}{\sqrt{\sum_U x_k^2 - \frac{1}{N}\left(\sum x_k\right)^2}\sqrt{\sum_U y_k^2 - \frac{1}{N}\left(\sum y_k\right)^2}}$$

$$= \frac{V_{XY}}{\sqrt{V_{XX}V_{YY}}}.$$

Using the logarithmic derivatives (rule 3), we obtain:

$$\frac{(I\rho)_k}{\rho} = \frac{I(V_{XY})_k}{V_{XY}} - \frac{1}{2}\frac{I(V_{XX})_k}{V_{XX}} - \frac{1}{2}\frac{I(V_{YY})_k}{V_{YY}}.$$

The influence of $A_{XY} = 1/N \sum_U x_k \sum_U y_k$ is obtained in the same way using:

$$I(A_{XY})_k = A_{XY}\left(\frac{x_k}{X} + \frac{y_k}{Y} - \frac{1}{N}\right) = \bar{Y}x_k + \bar{X}y_k - \bar{X}\bar{Y}$$

hence: $I(V_{XY})_k = x_k y_k - I(A_{XY})_k = (x_k - \bar{X})(y_k - \bar{Y})$. Then $I(V_{XX})_k = (x_k - \bar{X})^2$, and $I(V_{YY})_k = (y_k - \bar{Y})^2$ and so, with

$$S_x^2 = \frac{V_{XX}}{N} \text{ and } \cong_Y^2 = \frac{V_{YY}}{N}:$$

$$N(I\rho)_k = \frac{(x_k - \bar{X})(y_k - \bar{Y})}{S_X S_Y} - \frac{1}{2}\rho\left(\frac{(x_k - \bar{X})^2}{S_X^2} + \frac{(y_k - \bar{Y})^2}{S_Y^2}\right).$$

And the work is done.

9. Application: Implicit parameter

Let us assume that B, a parameter with q components, is a solution to an equation of the type:

$$H(B) = \sum_U l_k(B) = 0 \qquad (9.1)$$

where the l_k are regular functions of $\boldsymbol{R}^q$ in $\boldsymbol{R}^q$. This situation frequently occurs when B is the parameter of a model assumed to be valid in the population U. Under the usual assumptions of independence, equation (9.1) can result from the application of the maximum likelihood estimation principle. It is then the equation of the score. In the case of a linear model with Gauss residuals, this leads to the normal equations that can also be derived from the least squares principle:

$$l_k(B) = \frac{1}{\sigma_k^2}x_k(y_k - x_k'B)$$

with obvious notations.

If the functional family $H(B)$ regularly depends on B, we have:

$$I(B_0)_k = -\left(\frac{\partial H}{\partial B}\right)^{-1}_{B=B_0} IT(B_0)_k$$

i.e.,:

$$I(B_0)_k = -\left(\sum_U \frac{\partial l_k}{\partial B}\right)^{-1} l_k(B_0).$$

In the case of regression, we have:

$$I(B)_k = -\left(\sum_U \frac{x_k x_k'}{\sigma_k^2}\right)^{-1}\frac{1}{\sigma_k^2}x_k e_k = -T^{-1}\frac{1}{\sigma_k^2}x_k e_k$$

with the regression residual e_k. Thus we simply have the linearized variable of the vector of regression coefficients. To numerically compute the variance estimator, we use the approximation

$$\tilde{z}_k = \hat{T}_s^{-1}\frac{1}{\sigma_k^2}x_k\tilde{e}_k \text{ where } \hat{T}_s = \sum_S \frac{x_k x_k'}{\sigma_k^2 \pi_k}$$

and $\tilde{e}_k = y_k - \hat{B}x_k$. This expression therefore depends on s through $\hat{T}_s$ and

$$\hat{A} = \sum_s \frac{x_k y_k}{\sigma_k^2 \pi_k}.$$

Example: Regression estimator.

When the constant (or the variable σ_k^2) is part of the regressors, *i.e.*, when there is a vector λ such that $x_k'\lambda = 1$ (or σ_k^2) for any x, the regression estimator takes on the simple form $\hat{Y}_{\text{reg}} = X'\hat{B}$ where X is the known vector of the total of the x_k.

Regression estimation theory (Cochran 1977, Särndal, Swensson, Wretman 1992) tells us that the residuals e_k are the linearized variable of this estimator, and that they can be approximated using the estimated empirical residuals $\tilde{e}_k$ (note that these only depend on a finite number of parameters).

However, the above leads us to think that we should have:

$$\widehat{\text{Var}}(\hat{Y}_{\text{Reg}}) = X'\,\widehat{\text{Var}}(\hat{B})X$$

and that a natural "linearized" variable for $\hat{Y}_{\text{Reg}}$ should be $X'\hat{T}_s^{-1}\,1/\sigma_k^2\,x_k\,\tilde{e}_k$. If $\hat{T}_s$ is replaced by its expectation T, we notice that $X'T^{-1} = \lambda'$ and that $1'x_k = \sigma_k^2$, and we fall back on the previous approximation. Note, finally, that the quantities $X'\hat{T}_s^{-1}$ are exactly the weight corrections (or $g-$weights) used in the regression estimator, the use of which is often recommended within the framework of variance estimation.

It is quite clear that the two linearized variables lead asymptotically to the same result. The choice should therefore be based on other criteria. In a few specific cases, the concept of conditional estimation justifies the use of these $g-$weights, notably for the poststratified estimator in the case of simple random sampling. The general case remains fairly mysterious.

In the case of a logistic regression adjustment, the dependent variable y_k has a value of 0 or 1, and the equations of the score are written as follows:

$$H(B) = \sum_U x_k(y_k - f(x_k'B)) \text{ with } f(u) = \exp u/(1 + \exp u).$$

We therefore have:

$$I(B)_k \equiv \Big(\sum_U x_k x_k' f(x_k'B)\,(1 - f(x_k'B))\Big)^{-1} x_k\,(y_k - f(x_k'B)).$$

Using this variable makes it possible to compute correctly the precision of a logistic regression, *i.e.*, taking into consideration the sampling scheme.

10. The residual technique for complex estimators

Many complex estimators commonly used nowadays can be incorporated into the general framework of external data calibration (Deville, Särndal 1992; Deville, Särndal, Sautory 1993). A vector X of totals of auxiliary variables x_k is known, and we look for new weights w_k confirming the calibration equations $\sum_s w_k\, x_k = X$ any sample s. If we look for such weights as close as possible to the HT weights, they are found to be necessarily of the type $w_k = 1/\pi_k\, F_k(x_k'\,\lambda)$ where λ is a vector of the same dimension as X solving the calibration equations. The functions F_k depend on the chosen distance and allow limited development in the form $F_k(u) = 1 + q_k u + O(u^2)$. The most frequent form is $F_k(u) = F(q_k\, u)$, F a unique function. Often, also, the q_k are all equal to 1. Thus we find in this family the ratio estimator (arbitrary F and $q_k = 1/x_k$), the poststratified estimator (with x_k a stratum indicating vector), the raking ratio estimator (with x_k a margin indicating vector and F an exponential function), and the regression estimator $(F(u) = 1 + u)$.
The asymptotic variance of these estimators can be obtained naturally by applying the rules of linearization. The calibration equations define λ using:

$$T(\hat{M}, \lambda) = \int x_k\, F_k(x_k'\,\lambda)\, d\hat{M}(k) = \sum_s \frac{1}{\pi_k}\, x_k\, F_k(x_k'\,\lambda) = X.$$

Since we have $T(M,\,0) = X$, the application of rule 6 yields:

$$I\lambda(M, x) = -T^{-1} x_k$$

with

$$T = \int x_k\, x_k'\, F_k\,(0)\, dM(k) = \sum_U q_k\, x_k\, x_k'.$$

Moreover, the calibrated estimator appears to be the substitution estimator of the functional $S(M, \lambda) = \int y_k F_k(x_k'\lambda)\, dM(k)$, which, according to rule 7, allows for the linearized variable $y_k - x_k'\, T^{-1} \int q_i\, y_i\, x_i\, dM(i) = y_k - x_k' B$ by introducing the vector B of the least squares regression parameters into the population for the weights q.

Thus the variance of the calibrated estimator is obtained by replacing, in formula (2.1), the y_k by the residuals $e_k = y_k - x_k' B$ of the regression of y on x with the weights q. For the variance estimation, we use in formula (2.2) either $\tilde{e}_k = y_k - x_k'\hat{B}$, or, as in the case of the regression estimator, $F(x_k'\hat{\lambda})\,\tilde{e}_k$.

If we now turn to a parameter $T(M)$ estimated by substitution using the weights w_k obtained by calibration, we have the following important result:

Result: If $T(M)$ allows for a linearized variable z_k, and if $T(\hat{M}_w)$ is the estimator of $T(M)$ using the weights w_k derived from calibration on a vector $X = \sum_U x_k$, then e_k, a residual of the regression z_k on x_k is a linearized variable for $T(\hat{M}_w)$.

Proof: The variance of $T(\hat{M}_w)$ is equivalent to that of $\hat{Z}_w = \sum w_k z_k$ according to the previous demonstration. However, the variance of $\hat{Z}_w$ is equivalent to that of $\sum_s 1/\pi_k\, e_k$.

Comment: Very often, *e.g.*, in the case of an explicit function of totals, z_k is a linear form $\sum_{i=1}^p A_i\, y_k^i$. We then have:

$$\text{Var} \sum_s w_k z_k = \text{Var} \sum_{i=1}^p A_i\, \hat{Y}_w^i.$$

This suggests the following procedure:

- compute the residuals ε_k^i of the regressions of y_k^i on the x_k.
- form the synthetic variable $\sum_i A_i\, \varepsilon_k^i$
- compute the variance of this variable.

It is quite clear that this corresponds to the direct computation of the residuals of z_k, which is definitely more simple.

Comment: While it may be trivial, this result is perhaps the most useful one in this paper, and this comment simply ensures that it will not go unnoticed.

11. Application: Fractiles

The distribution function $F(x) = 1/N\,\text{Card}(k;\, x_k \le x)$ is a functional family $1/N \int 1(\xi \le x)\, dM(\xi)$. The value of influence $IF(x)_k$ is therefore

$$\frac{1}{N}(\mathbf{1}\,(x_k \le x) - F(x)). \qquad (11.1)$$

For $\alpha \in\,]0, 1[$, the fractile t_α is defined by $F(t_\alpha) = \alpha$ if we are ruthless, and by t_α: $F(t_\alpha - 0) < \alpha \le F(t_\alpha)$ if we take into consideration the staircase-shape of F. If we are ruthless, the ad hoc linearized variable is therefore:

$$I(t_\alpha)_k = -\left(\frac{\partial F}{\partial x}\bigg|\, x = t\alpha\right)^{-1} IF(t_\alpha)_k$$

$$= -\frac{1}{F'(t_\alpha)} \cdot \frac{1}{N}(\mathbf{1}\,(x_k \le t_\alpha) - \alpha). \qquad (11.2)$$

The problem arises from the fact that $F'(x)$ idealizes a density of the variable at point x which does not exist because of the stairs.

The difficulty can be overcome by using the following construction:

By definition, a regulating core is a positive function $K(x, t)$, confirming, for any x, $\int K(x, t)dt = 1$, which is regular (*e.g.*, sufficiently derivable). For any x, $K(x, .)$ is a "bell" function about x, *e.g.*, a normalized indicatrix of an interval surrounding x. More generally, the support of $K(x,.)$ will be an interval containing x. We note $G(x, t) = \int^t K(x, u)\, du$ and $\bar{G}(x, t) = 1 - G(x, t)$. $G(x; .)$ is a distribution function. From an asymptotic point of view, the core K depends on the size N of the population; the "band width", *i.e.*, the "mean" width of the support of $K(x, .)$, decreases with N.

We now replace the distribution function by its smoothing $F_K(x) = \int F(t)\, K(x, t)\, dt$. For a reasonable choice of K, F_K is strictly increasing wherever its value is not 0 or 1, and very close to F so that all the fractiles $t_{K\alpha}$ are defined univocally and close to t_α no matter how they are defined. Following integration by parts, note that we also have:

$$F_K(x) = \int \bar{G}(x, t)\, dF(t) = \frac{1}{N} \sum_U \bar{G}(x, x_k).$$

We therefore have:

$$I(F_K(x), \xi) = \frac{1}{N}(\bar{G}(x, \xi) - F_K(x)) \qquad (11.3)$$

which is entirely analogous to (11.1).

Since F_K is derivable ($\bar{G}$ being so), we have:

$$It_{K\alpha}(x) = -\frac{1}{F_K'(t_{K\alpha})} \frac{1}{N} (\bar{G}(t_{K\alpha}, x) - \alpha). \quad (11.4)$$

This formula is entirely analogous to (11. 2) save that $F_K'(t_{K\alpha})$ is perfectly defined. The linearization of $t_{K\alpha}$ does not therefore cause any particular problem, and may be used approximately for the linearization of t_α itself. A combined strategy consists in using the linearized variable

$$z_k = -\frac{1}{F_K'(t_\alpha)} (\mathbf{1}(x_k \le t_\alpha) - \alpha)$$

with

$$K(x, t) = \frac{1}{b - a} \mathbf{1}(a \le t < b)$$

(where $[a, b]$ is an interval containing x, more or less arbitrary). A practically correct linearized variable would be:

$$z_k = -\frac{b - a}{F(b) - F(a)} (\mathbf{1}(x_k \le t_\alpha) - \alpha).$$

The interval $[a, b]$ will have to be large enough so that in

$$\tilde{z}_k = -\frac{b - a}{\hat{F}(b) - \hat{F}(a)} (\mathbf{1}(x_k \le \hat{t}_\alpha) - \alpha)$$

the first factor will be sufficiently insensitive to sampling fluctuations.

12. Indexes of concentration and other functionals linked to ranks

Let us consider a few examples.

(a) GINI index.

With $T_x = \int \mathbf{1}(\xi < x)\, dM(\xi)$, the Gini index can be defined as:

$$\text{GINI} = \frac{\int x T_x dM(x)}{NX}.$$

Applying rule 5, we find for the influence of the numerator $xT_x + \int \xi \mathbf{1}(x \le \xi)\, dM(\xi)$. And yet $\int \xi \mathbf{1}(x \le \xi)\, dM(\xi) = X - \int \xi \mathbf{1}(\xi \le x)\, dM(\xi) = X - T_x \bar{x}_<$ where $\bar{x}_<$ is the mean of the x_k lower than x. Since X is a constant, the numerator linearized variable is therefore $T_x(x - \bar{x}_<)$. A linearized variable for GINI is therefore:

$$\text{IGINI}_\text{k} = F(x_k) \frac{x_k - \bar{x}_<}{X} - \text{GINI} \frac{x_k}{X}.$$

(b) Population below the poverty threshold.

It is defined as the proportion (of revenues) lower than half the distribution median. For the proper weight, let α and β denote two numbers between 0 and 1, and let us consider the indicator $J_{\alpha\beta} = F(\beta\, t_\alpha)$. The usual indicator corresponds to $\alpha = \beta = 1/2$.

The linearization is obvious using the rules under section 6 and the convention for writing the distribution function derivative:

$$\begin{aligned} IJ_{\alpha\beta}(x) &= IF_{\beta q_\alpha}(x) + F'(\beta t_\alpha)\, \beta I_{t_\alpha}(x) \\ &= \frac{1}{N}(\mathbf{1}(x \le \beta t_\alpha) - F(\beta t_\alpha)) - \frac{1}{N} \frac{F'(\beta t_\alpha)}{F'(t_\alpha)} (1(x \le t_\alpha) - \alpha) \\ &= \frac{1}{N} \left[\mathbf{1}(x \le \beta t_\alpha) - \frac{F'(\beta t_\alpha)}{F'(t_\alpha)} 1(x \le t_\alpha) + (\alpha - F(\beta t_\alpha)) \right]. \end{aligned}$$

For $\beta = 1$ we are able to find $IJ_{\alpha 1} = 0$.

The variance of the indicator is therefore computed simply by using the artificial variable having a value of 1 if $x_k \le \beta \hat{t}_\alpha$,

$$1 - \frac{F'(\beta \hat{t}_\alpha)}{F'(\hat{t}_\alpha)}$$

if $\beta \hat{t}_\alpha < x_k \le \hat{t}_\alpha$ and 0 if $x_k > \hat{t}_\alpha$.

(c) Kendall's coefficient of rank correlation.

Two numerical variables x_k and y_k are linked to individual k. The ranks of x_k and y_k respectively can be written as $R_k^X = \int_{x \le x_k} dM(x, y)$ and $R_k^Y = \int_{y \le y_k} dM(x, y)$. The coefficient of rank correlation is the correlation coefficient between R_k^X and R_k^Y, *i.e.*, following some elementary simplifications:

$$r = 12 \left(\frac{1}{N^3} \int R_\xi^X R_\eta^Y dM(\xi, \eta) - \frac{1}{2} \right).$$

This expression can be linearized by applying the rules related to influence functions. For: $T = \int R_\xi^X R_\eta^Y dM(\xi,\, n)$, we have:

$$\begin{aligned} IT(x,\, y) &= R_x^X R_y^Y + \int \mathbf{1}(x \le \xi)\, R_\eta^Y \, dM(\xi,\, \eta) \\ &\quad + R_\xi^X \mathbf{1}(y \le \eta)\, dM(\xi,\, \eta) \\ &= R_x^X\ R_y^Y + A_x + B_y \end{aligned}$$

where we have assumed

$$A_x = \sum_{k \in U:\ x \le x_k} R_K^Y$$

and

$$B_y = \sum_{k \in sU:\ y \le x_k} R_k^X,$$

so that finally:

$$Ir(x,\, y) = \frac{12}{N}\left(F_x^X F_y^Y + \frac{A_x}{N^2} + \frac{B_y}{N^2} - \frac{1}{4}\left(r + \frac{1}{2}\right)\right).$$

The variance is computed as follows:

- The linearized variable is $z_k = Ir(x_k,\, y_k)$.
- $\hat{F}_{x_k}^X$ and $\hat{F}_{y_k}^Y$ are the estimators of the distribution functions of x and y respectively.
- A_x is estimated using

  $$\hat{A}_x = \sum_{k \in s:\ x_k \ge x} w_k$$

  and B_y likewise.
- In the calculation, we use the approximation of

$$z_k,\, \tilde{z}_k = \frac{12}{\hat{N}}\left(\hat{F}_{x_k}^X \hat{F}_{y_k}^Y + \frac{A_{x_k}}{\hat{N}_2} + \frac{B_{y_k}}{\hat{N}_2} - \frac{1}{4}\left(\hat{r} + \frac{1}{2}\right)\right)$$

and we calculate the variance of the total of this variable estimated using the HT estimator (formula 2.2).

13. Factorial methods

The principal components of the vectorial variable x_k are the eigenvectors u of the matrix of covariances $C = \sum_U x_k x_k' - X\bar{X}'$. They therefore confirm:

$$\begin{matrix} Cu = \lambda u \\ u'u = 1 \end{matrix} \quad \text{with } \lambda \text{ the eigenvalue.}$$

The variance of λ and that of the components of the u can be obtained fairly simply. The influence of C is $IC(x) = (x - \bar{X})(x - \bar{X})'$. The influence of $Cu - \lambda u$ is

$$ICu + CIu - I\lambda u - \lambda Iu = 0. \qquad (13.1)$$

However $(Iu)'u = 0$, and also $u'CIu = 0$ because C is a symmetric matrix. By multiplying (13.1) on the left by u' we have:

$$u'ICu = I\lambda = (u'(x - \bar{X}))^2.$$

And yet $u'(x - \bar{X})$ is equal to $\lambda\xi^2$ where ξ is the principal component associated with $(\lambda,\, u)$. From this is derived the calculation of the variance of $\hat{\lambda}$, the solution to $\hat{C}\hat{u} - \hat{\lambda}\hat{u} = 0$.

The variance of the components of u is obtained analogously. Let $(\lambda_v,\, v)$ denote another eigenvalue, eigenvector pair of C. We multiply equation (13.1) on the left by v. We have:

$$v'ICu + \lambda_v(v'Iu) - \lambda(v'Iu) = 0$$

hence:

$$(v'Iu) = \frac{(\lambda\lambda_v)^{1/2}\xi\xi_v}{\lambda - \lambda_v}$$

and therefore:

$$Iu = \sum_{v \ne u} \frac{(\lambda\lambda_v)^{1/2}\xi\xi_v}{\lambda - \lambda_v} v.$$

Correspondence analysis or multiple correspondence analysis is subject to analogous treatment.

In the case of multiple correspondence analysis (the more general case), each individual is characterized by the vector x_k which "stacks" the indicatrixes of membership in the modalities of p qualitative variables (2 in the case of correspondence analysis). If $\underline{1}$ denotes the vector all of whose components have a value of 1, we have $x_k'\,\underline{1} = p$ for any k. We then look for vectors u normed by $1/p\, N\, u'Du = 1$, with $D = \text{diag}\sum_U x_k = \sum_U \text{diag}\, x_k$ such that the variance of $\xi_k = 1/p x_k' u$ is stationary. This yields a solution to the problem of eigenvalues: $Cu - p\lambda Du = 0$ where $C = \sum_U x_k x_k'$.

The search for a linearized variable for λ and u follows the same procedure as before. We have the relationship between influences:

$$(IC - pI\lambda D - p\lambda ID)\, u - (C - p\lambda D)\, Iu = 0.$$

As $IC = xx'$, $ID = \text{diag}\ x$, and $u'DIu = 0$, we obtain through premultiplication by u':

$$I\lambda(x) = \frac{1}{N}\left(\left(\frac{x'}{p}u\right)^2 - \lambda\frac{x'}{p}uOu\right)$$

where uOv denotes the Hadamard product (*i.e.*, component by component) of u and v. We know that $u = \underline{1}$ is a eigenvector associated with the eigenvalue 1, also the largest. We check that for $u = \underline{1}$ we have $I\lambda = 0$!

In the same manner, we obtain the components of Iu on the other proper vectors v:

$$v'DIu = \frac{1}{N}\left[\left(\frac{x'}{p} \cdot u\right)\left(\frac{x'}{p} \cdot v\right) - \lambda\frac{x'}{p} \cdot uOv\right].$$

The analysis may be continued by calculating the variability of a projection onto a factorial design. If A is a subpopulation of size N_A, the coordinates of its representative point on the factorial designs are

$$\alpha_u = \left(\frac{\sum_A x_k}{\sum_A 1} \right)' \cdot u = \bar{X}'_A \cdot u.$$

We linearize α_u by using the relationship $I\alpha_u = (I\bar{X}_A)'u + \bar{X}'_A I u$ and the rest is simple.

14. Conclusion

The linearization of complex statistics has long been considered the most flexible and comprehensive method of obtaining an estimation of the variance. Specifically, this method is applicable to any sample design and to any type of estimator. The popularity of methods based on sample replications is due largely to the fact that certain statistics are considered too complex to be linearized. However, for the large class of substitution estimators, the use of influence functions and of algebraic rules governing their construction makes it possible to obtain fairly simply linearized variables by means of which variance estimation boils down to the estimation of a total estimated using the Horvitz-Thompson estimator.

Acknowledgements

I wish to thank Olivier Sautory and Yves Tillé for kindly providing critical comments concerning various drafts of this paper. The latter provided the demonstration at the beginning of section 7. I also wish to thank the two referees of *Survey Methodology*, whose comments did much to make the paper more readable, as well as the editorial board for its understanding.

References

Billingsley, P. (1969). *Convergence of Probability Measures*. New York: John Wiley & Sons, Inc.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D.A., and Kovačević, M.S. (1997). Variance estimation for measures of income inequality and polarization: The estimating equations approach. *Journal of Official Statistics*, 13, 41-58.

Binder, D.A., and Patak, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

Chambers, R.L., Dorfman, A.H. and Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, 79, 577-582.

Cochran, W. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley & Sons, Inc.

Deville, J.-C. (1993). Une formule universelle d'estimation de variance. Internal document, INSEE-UMS.

Deville, J.-C. (1997). Estimation de la variance du coefficient de Gini mesuré par sondage. In *Actes des Journées de Méthodologie Statistiques*, INSEE METHODES, 69-70-71.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Durbin, J. (1953). Some results in sampling theory when units are selected with unequal probabilities. *Journal of the Royal Statistical Society B*, 15, 262-269.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 361-374.

Hampel, F.R., Ronchetti, E., Rousseeuw, P.J. and Stahel, W. (1985). *Robust Statistics*: *The Approach Based on the Influence Function*. New York: John Wiley & Sons, Inc.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Lecoûtre, J.P., and Tassi, PH. (1987). Statistique non-paramétrique et robustesse. *Economica*.

Rosen, B. (1972). Asymptotic theory for successive sampling I and II. *Annals of Mathematical Statistics*, 43, 373-397 and 748-776.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.