

Weak convergence and empirical processes in survey sampling

Processi empirici nel campionamento

Pier Luigi Conti

Abstract In this paper a “special version” of the empirical process is introduced and studied in sampling finite populations under fairly general sampling designs. This allow to establish the asymptotic normality of estimators commonly used in practice. As a by-product, estimates of the variance are also obtained.

Abstract *In questo lavoro si introduce, per disegni campionari generali, una “versione speciale” del processo empirico, e si mostra la sua convergenza debole ad un processo Gaussiano. Come conseguenza, si ottiene la normalità asintotica di stimatori di uso comune.*

Key words: Sampling design, entropy, empirical processes, differentiability

1 Introductory aspects

Asymptotics in survey sampling do have a long history. The main effort of the literature consists in providing a justification of confidence intervals and variance estimates commonly used in practice. Almost all results concern the asymptotic normality of linear statistics, and are obtained under various conditions on sampling plans. Among several papers devoted to this subject, in [5] asymptotic properties of L -statistics are obtained under stratified two-stage sampling; similar results are in [3]. A completely different approach is in a fundamental paper by Hájek [4], where asymptotic properties for the Horvitz-Thompson estimator under rejective sampling are obtained, as well as [8], [1].

All the above mentioned papers (and many others, as well) are mainly confined to linear statistics, the most important one being the Horvitz-Thompson estimator

Pier Luigi Conti
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy
e-mail: pierluigi.conti@uniroma1.it

of the population mean. Asymptotic results are obtained by *ad hoc*, specific, techniques, and it is generally difficult to extend them to non-linear estimators.

The framework of survey sampling is not far from that of nonparametric statistics. The main common feature is that no special assumptions are made on the population distribution. The main difference is that in nonparametric statistics observations are usually assumed to be independent and identically distributed (*i.i.d.*), whilst the same does not hold in survey sampling, because of the sampling design.

The main tool to study asymptotic theory in nonparametric statistics is the empirical process theory (cfr., for instance, [7]), the offers the right background for a large amount of results. Our goal here is to construct an asymptotic theory for sampling finite populations that parallels the classical theory in nonparametric statistics. At an elementary level, in classical nonparametric statistics the starting point consists in observing that almost all statistics are essentially functionals of the empirical distribution function. Asymptotic results are first obtained for the empirical distribution function (this step involves empirical processes and weak convergence of stochastic processes), and then then extended to sufficiently regular functionals (this step usually requires an appropriate form of differentiability). In the sequel we will take a similar road for the important class of “high entropy” sampling designs.

2 Basic concepts and symbols

Let $\mathcal{U}_N$ be a finite population of N units, labeled by integers $1, \dots, N$. Let Y be the variable of interest, and let y_i be the value of Y for unit i ($i = 1, \dots, N$). Let further $\mathbf{y}_N = (y_1, \dots, y_N)$. The *population distribution function* (p.d.f., for short) is defined as

$$F_N(y) = \frac{1}{N} \sum_{i=1}^N I_{(-\infty, y]}(y_i), \quad y \in \mathbb{R} \quad (1)$$

where $I_A(\cdot)$ is the indicator function of the set A .

Now, for each unit i in $\mathcal{U}_N$, define a Bernoulli random variable (r.v.) D_i , such that the unit i is included in the sample if and only if (iff) $D_i = 1$, and let $\mathbf{D}_N = (D_1, \dots, D_N)$. A (unordered, without replacement) sampling design P is the probability distribution of $\mathbf{D}_N$. In particular, $\pi_i = E_P[D_i]$ ($\pi_{ij} = E_P[D_i D_j]$) is the first (second) order inclusion probability of unit i (pair of units i, j). The suffix P denotes the sampling design used to select population units.

Let $p_1, \dots, p_N$ be N real numbers, with $p_1 + \dots + p_N = n$. The sampling design is a *Poisson design* with parameters $p_1, \dots, p_N$ if the r.v.s D_i s are independent with $\pi_i = p_i$ for each unit i . The *rejective sampling*, or *normalized conditional Poisson sampling* ([4], [6]) corresponds to the probability distribution of the random vector $\mathbf{D}_N$, under Poisson design, conditionally on $n_s = n$.

The *Hellinger distance* between a sampling design P and the rejective design is defined as

$$d_H(P, P_R) = \sum_{D_1, \dots, D_N} \left(\sqrt{Pr_P(D_N)} - \sqrt{Pr_R(D_N)} \right)^2. \quad (2)$$

Our basic assumptions are listed below.

- A1. $(\mathcal{U}_N; N \geq 1)$ is a sequence of finite populations of increasing size N .
A2. For each N , $y_1, \dots, y_N$ are realizations of a superpopulation $(Y_1, \dots, Y_N)$ composed by *i.i.d.* r.v.s Y_i with common d.f. F . In the sequel, we will denote by $\mathbb{P}$ the probability distribution of r.v.s Y_i s, and by $\mathbb{E}, \mathbb{V}$ the corresponding operators of mean and variance, respectively.
A3. For each population $\mathcal{U}_N$, sample units are selected according to a fixed size sample design with inclusion probabilities $\pi_1, \dots, \pi_N$, and sample size n . Furthermore, for $d > 0, f > 0$,

$$d_N = \sum_{i=1}^N \pi_i(1 - \pi_i) \rightarrow \infty, \quad \frac{1}{N}d_N \rightarrow d, \quad \lim_{N \rightarrow \infty} \frac{n}{N} = f \text{ as } N \rightarrow \infty.$$

- A4. For each population $(\mathcal{U}_N; N \geq 1)$, let P_R be the rejective sampling design with inclusion probabilities $\pi_1, \dots, \pi_N$, and let P be the actual sampling design (having the same inclusion probabilities). Then

$$d_H(P, P_R) \rightarrow 0 \text{ as } N \rightarrow \infty.$$

- A5. There exist two positive real numbers A, B such that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \frac{1}{\pi_i} = A < \infty, \quad \lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{1}{(i\pi_i)^2} = B < \infty.$$

3 Main basic results

As already said, the first, basic step in nonparametric statistics consists in constructing the *empirical process*, namely the rescaled difference between empirical d.f and population d.f., and in studying its limiting properties.

In the present case, in order to estimate the p.d.f. F_N , we consider the Hájek estimator

$$\hat{F}_H(y) = \frac{\sum_{i=1}^N \frac{1}{\pi_i} D_i I_{(-\infty, y]}(y_i)}{\sum_{i=1}^N \frac{1}{\pi_i} D_i}. \quad (3)$$

Our main goal is to study the asymptotic “global” behaviour of the random function $\hat{F}_H(\cdot)$. To this purpose, define the (sequence of) random function(s)

$$W_N^H(\cdot) = \{\sqrt{n}(\hat{F}_H(y) - F_N(y)), y \in \mathbb{R}\}; \quad N \geq 1. \quad (4)$$

Clearly, (4) plays exactly the same role as the classical empirical process in non-parametric statistics. The main results of the present section is Proposition 1.

Proposition 1. *If the sampling design P satisfies assumptions A1-A5, with $\mathbb{P}$ -probability 1, conditionally on $\mathbf{y}_N$ the sequence $(W_N^H(\cdot); N \geq 1)$, converges weakly, in the space $l^\infty[-\infty, +\infty]$ of the bounded functions on $[-\infty, +\infty]$ equipped with the sup-norm, to a Gaussian process $W^H(\cdot) = (W^H(y); y \in \mathbb{R})$ that can be represented as*

$$W^H(y) = \sqrt{f(A-1)}B(F(y)), \quad y \in \mathbb{R} \quad (5)$$

where $(B(t); 0 \leq t \leq 1)$ is a Brownian bridge.

In particular, if the sampling design is simple random sampling (without replacement) of size n , the Hájek estimator (3) reduces to the empirical d.f. $F_n(y) = n^{-1} \sum_i D_i I_{(-\infty, y]}(y_i)$, and $f(A-1) = 1 - f$.

In classical nonparametric statistics, the empirical process $\sqrt{n}(\hat{F}_n(y) - F(y))$ converges weakly to a Gaussian process of the form $B(F(y))$. This result is apparently similar to Proposition 1, with two differences: (i) the centering factor F instead of F_N ; (ii) the absence of the finite population correction term $f(A-1)$, since in classical nonparametric statistics the observations are (realizations of) *i.i.d.* r.v.s, and essentially there is no sampling design.

The asymptotic result of Proposition 1 is obtained conditionally on the population values $\mathbf{y}_N$; hence, the involved probability *only* depends on the sample design. Furthermore, [A2] could be replaced by other hypothesis, non even involving a superpopulation model.

4 Asymptotics for Hadamard-differentiable functionals

The estimation of the p.d.f. F_N is just the initial step in studying asymptotics, because several parameters of interest are actually functionals of F_N . Noticeable examples are the *pth* population quantile

$$Q_N(p) = F_N^{-1}(p) = \inf\{y : F_N(y) \geq p\}, \quad 0 < p < 1, \quad (6)$$

the generalized Lorenz curve (provided all y_i s are positive)

$$G_N(p) = \int_0^p Q_N(t) dt, \quad 0 \leq p \leq 1 \quad (7)$$

as well as several others.

The appropriate tool to study asymptotic properties of functional of F_N , such as (6)-(7) is the notion of Hadamard-differentiability. Let $\phi(\cdot) : l^\infty[-\infty, +\infty] \rightarrow E$ be a map having as domain the normed space $l^\infty[-\infty, +\infty]$ (endowed with the sup-norm), and taking values on an appropriate normed space E with norm $\|\cdot\|_E$. The map $\phi(\cdot)$ is Hadamard-differentiable at F if there exists a continuous linear mapping

$\phi'_F : l^\infty[-\infty, +\infty] \rightarrow E$ such that

$$\left\| \frac{\phi(F + th_t) - \phi(F)}{t} - \phi'_F(h) \right\|_E \rightarrow 0 \text{ as } t \downarrow 0, \text{ for every } h_t \rightarrow h. \quad (8)$$

$\phi(\cdot)$ is the *Hadamard derivative* of ϕ at F . The following result holds (cfr. [7]).

Proposition 2. *If $\phi(\cdot)$ is (continuously) Hadamard-differentiable at F , with Hadamard derivative $\phi'_F(\cdot)$, then the asymptotic law of $\sqrt{n}(\phi(\hat{F}_H(y)) - \phi(F_N))$ coincides with the asymptotic law of $\phi'_F(W^H)$, as N increases.*

Proposition 2 also extends and makes it rigorous a technique of variance estimation originally proposed by Deville [2]. Roughly speaking, Proposition 2 shows that the asymptotic distribution of $\sqrt{n}(\phi(\hat{F}_H(y)) - \phi(F_N))$ coincides with the asymptotic distribution of $\phi'_F(\sqrt{n}(\hat{F}_H - F_N))$. Using the symbol $\overset{d}{\approx}$ to indicate the “approximate equality of distributions”, and taking $\hat{N} = \sum_i \pi_i^{-1} D_i$, we may then write

$$\begin{aligned} \sqrt{n}(\phi(\hat{F}_H(y)) - \phi(F_N)) &\overset{d}{\approx} \phi'_F(\sqrt{n}(\hat{F}_H - F_N)) \\ &= \frac{\sqrt{n}}{\hat{N}} \sum_{i=1}^N \frac{D_i}{\pi_i} \phi'_F(I_{[y_i, +\infty)}(\cdot) - F(\cdot)) \\ &\overset{d}{\approx} \frac{\sqrt{n}}{N} \sum_{i=1}^N \frac{D_i}{\pi_i} \tau_F(y_i) \end{aligned}$$

where $\tau_F(y_i) = \phi'_F(I_{[y_i, +\infty)}(\cdot) - F(\cdot))$ is the *influence function* of ϕ computed at y_i . Hence, taking $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$, the design-based variance of $\sqrt{n}(\phi(\hat{F}_H) - \phi(F_N))$ can be asymptotically approximated by:

$$\begin{aligned} V_P \left(\sqrt{n}(\phi(\hat{F}_H(y)) - \phi(F_N)) \right) &\approx \frac{n}{N^2} \sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} \tau_F(y_i) \tau_F(y_j) \\ &= -\frac{n}{N^2} \sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} (\tau_F(y_i) - \tau_F(y_j))^2. \quad (9) \end{aligned}$$

If $\phi'_F(\cdot)$ is continuous in F , then $\tau_F(y_i)$ can be approximated, in its turn, by $\hat{\tau}_{\hat{F}_H}(y_i) = \phi'_{\hat{F}_H}(I_{[y_i, +\infty)}(\cdot) - \hat{F}_H(\cdot))$, so that the following estimator of the design-based variance of $\sqrt{n}(\phi(\hat{F}_H) - \phi(F_N))$ is obtained

$$\hat{V}_{YG} = -\frac{n}{N^2} \sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} \left(\hat{\tau}_{\hat{F}_H}(y_i) - \hat{\tau}_{\hat{F}_H}(y_j) \right)^2 \frac{D_i D_j}{\pi_{ij}}. \quad (10)$$

Further simplifications can be obtained by using the Hájek approximation $\pi_{ij} \approx \pi_i \pi_j (1 - (1 - \pi_i)(1 - \pi_j)/d_N)$, d_N being defined in [A3].

5 An example

Consider the generalized Lorenz curve (7). Under mild regularity conditions, the map $\phi : F \mapsto G(\cdot)$ is Hadamard-differentiable, with Hadamard derivative

$$\phi'_F(h) = - \int_0^{Q(p)} h(y) dy, \quad 0 \leq p \leq 1. \quad (11)$$

Taking $\hat{Q}_H(p) = \inf\{y : \hat{F}_H(y) \geq y\}$ and defining $\hat{G}_H(p)$ exactly as in (7) with Q_N replaced by $\hat{Q}_H$, from Proposition 2 it follows that the sequence of stochastic processes $\sqrt{n}(\hat{G}_H(p) - G_N(p); 0 \leq p \leq 1)$ converges weakly to a Gaussian process with zero mean function and covariance kernel

$$\begin{aligned} C_G(u, v) = f(A-1) \times \{ & S(u) - G(u)(Q(u) + Q(v)) + uQ(u)Q(v) \\ & - (G(u) - uQ(u))(G(v) - vQ(v)) \} \end{aligned} \quad (12)$$

with $S(u) = \mathbb{E}[Y_1^2 I_{(-\infty, Q(u)]}(Y_1)]$. Since elementary computations show that

$$\tau_F(y_i) = \phi'_F(I_{[y_i, +\infty)}(\cdot) - F(\cdot)) = (y_i - Q(p))I_{(-\infty, Q(p)]}(y_i) - (G(p) - pQ(p))$$

it is easy to conclude that, for fixed p , $\sqrt{n}(\hat{G}_H(p) - G_N(p))$ is asymptotically normally distributed with mean 0 and variance $V_G(p) = C_G(p, p)$. Furthermore, a (design) consistent estimator of $V_G(p)$ is

$$\hat{V}_{YG} = - \frac{n}{2N^2} \sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} \frac{D_i D_j}{\pi_{ij}} \left((y_i - \hat{Q}_H(p))I_{(-\infty, \hat{Q}_H(p)]}(y_i) - (y_j - \hat{Q}_H(p))I_{(-\infty, \hat{Q}_H(p)]}(y_j) \right)^2.$$

References

1. Berger, Y. G.: Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, **67**, 209–226 (1998)
2. Deville, J.-C.: Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, **25**, 193–203 (1999)
3. Francisco, C. A., Fuller, W. A.: Quantile Estimation with a Complex Survey Design. *The Annals of Statistics*, **19**, 454–469 (1991)
4. Hájek, J.: Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population. *The Annals of Mathematical Statistics*, **35**, 1491–1523 (1964)
5. Shao, J.: L-Statistics in Complex Survey Problems. *The Annals of Statistics*, **22**, 946–967 (1994)
6. Tillé, Y.: *Sampling Algorithms*. Springer, New York (2006)
7. Van Der Vaart, A.: *Asymptotic Statistics*. Cambridge University Press, Cambridge (1998)
8. Věšek, J. A.: Asymptotic Distribution of Simple Estimate for Rejective, Sampford and Successive Sampling. In: *Contributions to Statistics* (Ed. J. Jurečková), 263–275, Reidel Publishing Company, Dordrecht, Holland (1979)