Functional CLT for negatively associated sampling plans

Patrice Bertail*, Antoine Rebecq

Modal'X, Université Paris Nanterre

Workshop Empirical processes, Besançon , May 2019

Outlines

1. **NA of survey sampling and resampling plans**
   - A few notations and NA
   - Examples of NA sampling plans
   - A taylor made CLT for NA r.v's
2. **Empirical process for NA sampling plans**
   - An simple Hoeffding inequality for NA r.v's
   - General assumptions
   - A functional CLT for NA survey samples
3. **Some applications**
   - Confidence bands for the distribution function
   - Tail index estimators based on survey sampling

Outlines

Inclusion probabilities and survey sampling plans

Sampling units

- $\mathcal{S} \subsetneq \mathcal{U}_N$ of size $n << N$ taken at Random
- Inclusion variable : $\epsilon_i := \mathbb{I}\{i \in \mathcal{S}\}$ $\quad i \in \mathcal{U}_N$
- Inclusion probability: $\pi_i := \mathbb{P}(\epsilon_i = 1) = \mathbb{E}(\epsilon_i)$ $\quad i \in \mathcal{U}_N$
- Second order inclusion probability : $\pi_{i,j} := \mathbb{P}(\epsilon_i = 1, \epsilon_j = 1) = \mathbb{E}(\epsilon_i \epsilon_j)$ $(i,j) \in \mathcal{U}_N^2$
- The survey sampling plan is characterized by a distribution $\mathbf{R}_N$ on $\mathcal{S} \equiv (\epsilon_1, \ldots, \epsilon_N)$

Links with resampling plans

- Survey sampling plans can be seen as particular cases (without replacement) of general resampling plans appearing in the weighted bootstrap literature (Mason and Newton, 1992, Ann. Stat, Praestgaard and Wellner, 1993, Ann. Probab., Barbe and Bertail, 1995, Springer) $(W_{1,N}, W_{2,N}, ...W_{N,N})$.
- Naive bootstrap $Mult(N, (1/N..., 1/N))$, n out of N bootstrap $Mult(n, (1/N, ..., 1/N))$.
- Bayesian bootstrap (Dirichlet weights).
- Bootstrap bayesian clones (Lo, 1991, Ann. Stat.) $(Y_1/S_N, ..., Y_N/S_N)$ with $S_N = \sum_{i=1}^{N} Y_i$.
- Exchangeability is generally assumed! in bootstrap litterature

Negative association

Negative and Positive association (see Joag-Dev and Proschan, 1983, Annals of Stat.) : frequently used in time series. See Oliveira(2012), Springer for details, main properties and applications to time series

Definition

The r.v.'s $Z_1, \ldots, Z_n$ are said to be negatively associated (NA) iff for any pair of disjoint subsets A_1 and A_2 of the index set $[\![1, N]\!]$

$$Cov\left(f((Z_i)_{i\in A_1}),\ g((Z_j)_{j\in A_2})\right) \leq 0, \tag{1}$$

for any real valued measurable functions $f : E^{\#A_1} \to \mathbb{R}$ and $g : E^{\#A_2} \to \mathbb{R}$ that are both increasing in each variable.

Remark : NA implies negative correlation. This property plays an important role in survey sampling : $\pi_{i,j} - \pi_i\pi_j \leq 0$ (known as Sen-Yates-Grundy property) -> special form of the variance of Horvitz-Thompson estimator.

Examples of NA survey sampling plans

- Poisson sampling = indep. $B(1,\pi)$ r.v. (random sample size with mean $\sum_i^n \pi_i$)
- Rejective sampling = Poisson sampling conditional to the size equal a fixed n
- Subsampling = Rejective sampling with equal inclusion probabilities (SWoR)
- Pareto sampling, order sampling : select the n biggest values of well chosen indep. Pareto distribution.
- Tillé's Pivotal sampling or Srinivasan sampling : a sequential game between candidates (see Dubhashi & Ranjan, 1998, Jonasson, 2012, Electronic Com. Probab.)

Examples of NA survey sampling plans

- Determinantal sampling (cf Kulesza and Taskar,2012, Machine Learning J., Loonis and Mary, 2018, JSPI) : probability of a sample proportional to the determinant of sub-matrix (with inclusion probability on diagonal)
- Balanced sampling (which respects to some margin conditions) using the Cube Method (Deville and Tillé, 2004) : very efficient and "almost" balanced. Not always negatively associated (plan and inclusion probabilities depends on the original order) : apply a random permutation first to get NA property.
- Systematic sampling, cluster sampling, stratified sampling are NOT negatively associated. But most of the times it is possible to write estimators as sums of negatively associated variables by aggregrating over clusters or stratas.

Examples of NA resampling plans

- Naive bootstrap, n out of N bootstrap are NA : already in the paper by Joag-Dev and Proschan, 1983, Annals of Stat.
- Subsampling (without replacement even with unequal probability) is NA (particular case of rejective sampling)
- Bayesian bootstrap is NA
- Bootstrap Bayesian clones based on a log-concave distribution are NA : also in Joag-Dev and Proschan, 1983
- Double bootstrap is NA

Remark : Exchangeability of the weights (normalized to 1 or n) always implies negative correlation.

Negative association for survey sampling plans

- Importance of negative association for sampling plans stressed recently by Borcea and Brändén(2009), Inventiones Mathematicae, Brändén and Jonasson (2012), Scandinavian Journal of Statistics, based on works by Pemantle(2004) Math. Phys. 41, Joag-Dev and Proscan (1983), Annals of Stat.
- Borcea and Brändén(2009) propose criteria to prove NA (strongly Raleigh property)
- Many properties of resampling procedure (including weighted bootstrap) may be derived by proving Negative Association including CLT, deviation inequalities.

Horvitz-Thompson estimators

- Classical Horvitz-Thompson estimator of the mean of some characteristic X
- Parameter of interest $\overline{X}_N = N^{-1} \sum_{i=1}^{N} X_i$
- Horvitz Thompson estimator

$$\hat{\overline{X}}_n = N^{-1} \sum_{k \in \mathcal{U}} \frac{X_k}{\pi_k} \epsilon_k$$

- CLT for this estimator for general sampling plans . Pioneering work of Hajek (1964), An. Math. Statist, Rosen(1997), An. Stat. : very difficult proofs based on coupling arguments respectively for rejective sampling, Pareto sampling, immediate for sampling plan close to Rejective sampling (Sampford, Successive sampling etc...), see Berger(1998), JSPI, by controlling the L1 distance between this plan and rejective sampling. Simpler proofs in B., Chautru and Clémençon (2017), Scand. J. of Stat. (based on conditional CLT).

CLT based on negative association

Patterson, Smith, Taylor, Bozorgnia(2001), Nonlinear Analysis. Oliveira (2012), Asymptotics for assoc. r.v.'s, Springer

Theorem

Consider a triangular array $(X_{i,N})_{1\leq i\leq N}$ *of centered negatively correlated random variables then, under the conditions*

(i) $S_N^2 = Var(\sum_{i=1}^N X_{i,N}) \to \infty$ *as* $N \to \infty$

(ii) $\frac{1}{S_N^2}\sum_{i=1}^N \sum_{i<j} cov(X_{i,N}, X_{j,N}) \to 0$ *as* $N \to \infty$

(iii) for any $\varepsilon > 0$, $\frac{1}{S_N^2}\sum_{i=1}^N E(X_{i,N}^2 1_{\{|X_{i,N}|>\varepsilon S_N\}}) \to 0$ *as* $N \to \infty$

we have

$$\frac{1}{S_N}\sum_{i=1}^N X_{i,N} \underset{N\to\infty}{\overset{L}{\to}} N(0,1).$$

Problem : this basic CLT does not hold for most survey sampling plans.

Application to m out N bootstrap

the multinomial distribution $Mult(m, 1/N, ..., 1/N))$ and we have

$$E(W_{i,N}) = \frac{m}{N}$$

$$V(W_{i,N}) = \frac{m}{N}(1 - 1/N)$$

$$cov(W_{i,N}, W_{j,N}) = -V(W_{i,N})/(N-1) = -m/N^2$$

It follows that if we consider the weighted bootstrap sums with $Z_{i,N} = (W_{i,N} - \frac{m}{N})Y_i$ where the Y_i 's are positive i.i.d. r.v.'s with $E(Y_i^{2+\eta}) < \infty$ for some $\eta > 0$, then the $Z_{i,N}$'s are negatively associated and we get

$$\sum_{i=1}^{N} V(Z_{i,N}) = \sum_{i=1}^{N} (Y_i - \overline{Y}_N)^2 V(W_{i,N}) = m\sigma^2(1 - 1/N)(1 + o(1))$$

$$\sum_{i=1}^{N}\sum_{i<j} cov(Z_{i,N}, Z_{j,N}) = -m\left(\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y}_N)\right)^2 + \frac{1}{N^2}\sum_{i=1}^{N}(Y_i - \overline{Y}_N)^2\right)$$

A Taylor made CLT, B. and Rebecq(2017)

Theorem

Consider a triangular array $(X_{i,N})_{1\leq i\leq N}$ *of centered negatively associated random variables then, under the Lindeberg-Feller conditions (iii) of the Theorem by Patterson et al.(2001) and assuming that we have*

$$(iv) \quad 0 < lim_N \frac{S_N^2}{N} < \infty,$$

then we have

$$\frac{1}{S_N} \sum_{i=1}^{N} X_{i,N} \underset{N\to\infty}{\overset{L}{\to}} N(0,1).$$

A Taylor made CLT, B. and Rebecq(2017)

Proof : adaptation of Yuan, Su and Hu(2003),J. Theor. Prob. and Peligrad and Utev(1997), Ann. Probab. to non-stationary variables. Block of block techniques ensuring that we capture the covariance terms. This result yields

a CLT for all the sampling plan seen before. Allows to generalize the approach of Bertail, Chautru and Clemençon (2017), Scand J. Stat.

Outlines

Horvitz-Thompson empirical processes

Horvitz-Thompson empirical measure (not a probability)

$$: \mathbf{P}_N^{\pi} := \frac{1}{N} \sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i} \delta_{X_i}$$

Our Goal : Donsker Theorem for survey data

Prove a Donsker theorem for a version of the Horvitz Thompson empirical process indexed by classes of positive functions $\mathcal{F}$ under some natural conditions on the sampling plan and the class of functions (measurability issues evacuated in this talk), essentially existence of a $L^{2+\eta}(P)$ integrable envelop and some uniform entropy condition (over discrete probability measures).

- In the i.i.d. case $Z = \sqrt{N}\{P_N f - Pf\}_{,f\in\mathcal{F}}$. see van der Vaart and Wellner(1996), van de Geer(2009)
- In the survey sampling case (for some sampling plan R_N we will be interested in $\mathbb{G}_{R_N}^{\pi(R_N)} = (\mathbb{G}_{R_N}^{\pi(R_N)} f)_{f\in\mathcal{F}}$, where

$$\mathbb{G}_{R_N}^{\pi(R_N)} f = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{\epsilon_i}{\pi_i(R_N)} - 1 \right) f(X_i)$$

Existing functional results in survey sampling

- Very few functional results even in the real case : Cardot, Goga, Lardin(2013), E.J.S for the mean of functional data (over time), Wang (2012), CSDA , corrected by Boistard, Lopuhäa and Ruiz-Gazen (2017), Annals of Stat, for general survey sampling plans in the case of the repartition function, based on control of fourth order moments (conditions on fourth order inclusion probabilities), Bertail, Chautru, Clémençon(2017), Scand. J. Stat. for general class but only conditional plans (under conditions on 2d order inclusion probabilities).
- Some general results for general class of functions for sampling uniformly WR or WoR or for sampling schemes satisfying some exchangeability conditions : particular cases of weighted bootstrap. Not very interesting for real applications.
- Functional version for stratified survey sampling plan (UWR or UWoR in each strata) -> same as independent bootstrap or subsampling in each strata), Breslow & Wellner, 2008 and Saegusa & Wellner, 2012.

A few notations and assumptions A0

Envelop of the class. There exists a measurable function $H : \mathcal{X} \to \mathbb{R}$ such that there exists $\eta > 0$ such that $H(x) > \eta$ for every x and $\int_{x \in \mathcal{X}} H^{2+\eta}(x) P(dx) < \infty$ for some η and $|f(x)| \leq H(x)$ for all $x \in \mathcal{X}$ and any $f \in \mathcal{F}$.

Donsker classes. $\mathcal{F}$ is a Donsker class of function (Hoffmann-Jorgensen weak convergence). The set of probability measures P may be considered as a subset of $l_\infty(\mathcal{F})$, *i.e.* the space of all maps $\Phi : \mathcal{F} \to \mathbb{R}$ such that $\|\Phi\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\Phi(f)| < +\infty)$ with the uniform convergence norm $\|P - Q\|_{\mathcal{F}} = d_{\mathcal{F}}(P, Q) = \sup_{h \in \mathcal{F}} |\int h dP - \int h Q|$.

Uniform covering and entropy number condition

$$\int_0^1 \sup_{Q \in \mathcal{D}} \sqrt{log(N(\varepsilon \|H\|_{2,Q}, \mathcal{F}, \|.\|_{2,Q}))}\, d\varepsilon < \infty,$$

where $\mathcal{D}$ is the set of all discrete probability measures Q such that $0 < \int H^2 dQ < \infty$

Empirical process for NA sampling plans

Recent results of Bertail and Clémençon, 2017, to appear Bernoulli) make use of Negative association to prove bounds. But some simple ones were already available in Janson(1992).

Theorem

Let $Y_1, \ldots, Y_n$ be negatively associated random variables such that $a_i \leq Y_i \leq b_i$ a.s. and $\mathbb{E}[Y_i] = 0$ for $1 \leq i \leq n$. Then, for all $t > 0$, we have: $\forall n \geq 1$,

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq t\right) \leq \exp\left(-\frac{2nt^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

The first order inclusion probabilities $\pi = (\pi_1, \dots, \pi_N)$ are now supposed to depend on some auxiliary variable W through the link function $p : \mathcal{W} \to [p_\star, 1]$, $p_\star > 0$. For this link function we will thus write

$$\pi_i = p(W_i)$$

and

$$\mathbb{G}^{\boldsymbol{p}}_{R_N} f := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{\epsilon_i}{p(W_i)} - 1 \right) f(X_i).$$

Because the variance of this process also depends on the second order inclusion probability we will also assume that:

Assumption

there exists a constant K (independent of N and $W_i's$) such that for all i, j

$$|\pi_{i,j} - \pi_i \pi_j| \leq \frac{K}{N}.$$

Assumption

Assume that we can write for some symmetric bounded positive function $0 < h(.,.) \leq K$,

$$\pi_i \pi_j - \pi_{i,j} = \frac{h(W_i, W_j)}{N}, i \neq j.$$

Under this condition and the negative association condition of the plan

$$S_N^2(f) = V(\mathbb{G}_{R_N}^{\boldsymbol{p}} f) = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} h(W_i, W_j) \left(\frac{f(X_i)}{p(W_i)} - \frac{f(X_j}{p(W_j)} \right)^2.$$

Under the preceding assumptions we have, for any f

$$S_N^2(f) \leq \frac{K}{2} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} \left(\frac{f(X_i)}{p(W_i)} - \frac{f(X_j)}{p(W_j)} \right)^2$$

Assumption

The random vectors $(X_1, W_1), \ldots, (X_N, W_N)$ *are exchangeable random vectors with common marginal distribution* $\mathbb{P}_{X,W}$, *such that* $E\frac{H(X_i)^2}{p(W_i)^2} < \infty$.

Under this condition, we define the covariance operator

$$\Sigma^p(f,g) := E_{P_{X,W}}\left(h(W_i, W_j)\left(\frac{f(X_i)}{p(W_i)} - \frac{g(X_j)}{p(W_j)}\right)\right), (f,g) \in \mathcal{F}^2.$$

A Lindeberg-Feller condition

Assumption

Lindeberg-Feller type condition hold : $\forall \eta > 0$,

$$\mathbb{E}_{\mathbb{P}_{X,W}}\left(\mathbb{E}_{T_N}\left(\mathcal{X}_{N,i}^2\,\mathbb{I}\left\{\mathcal{X}_{N,i} > \eta\sqrt{N}\right\}\right)\right) \underset{N\to\infty}{\longrightarrow} 0,$$

with $\mathcal{X}_{N,i} := \left|\frac{\epsilon_i}{p(W_i)} - 1\right| \sup_{f\in\mathcal{F}} |f(X_i)|.$

A functional CLT for NA survey samples

Theorem

Let $\mathcal{F}$ be a set of function satisfying conditions A0 (entropy condition + envelop), the Lindeberg-Feller condition and let the sampling plan satisfy the conditions before. Then, there exists a $\rho_{\mathbb{P}}$-equicontinuous Gaussian process $\mathbb{G}^p$ in $\ell^\infty(\mathcal{F})$ with covariance operator $\Sigma^p(f,g)$ such that , almost surely along the sequence,

$$\mathbb{G}^{\mathbf{p}}_{R_N} \Rightarrow \mathbb{G}^p \text{ weakly in } \ell^\infty(\mathcal{F}), \text{ as } N \to \infty.$$

Proof of the result

1) Let $\mathbf{f}_K = (f_1, \ldots, f_K)$ be any vector of functions in $\mathcal{F}^K$, $K \geq 1$. Then the finite dimensional marginals $\mathbb{G}^{\mathbf{p}}_{R_N}\mathbf{f}_K := (\mathbb{G}^{\mathbf{p}}_{R_N}f_1, \ldots, \mathbb{G}^{\mathbf{p}}_{R_N}f_K)$ are asymptotically Gaussian conditionally on $\mathcal{D}_N$, with limiting covariance matrix $\Sigma^p_{\mathbf{f}_K} := (\Sigma^p(f_k, f_\ell))_{1\leq k,\ell\leq K}$ $a.s.$ for any sequence in $\mathcal{D}_N$. This can be checked by tacking any linear combination $a\prime\mathbb{G}^{\mathbf{p}}_{R_N}\mathbf{f}_K$, $a \in \mathbb{R}^{+K}$. This amounts to check the asymptotic normality of $\mathbb{G}^{\mathbf{p}}_{R_N}a'\mathbf{f}_K$. But under the given hypotheses (Lindeberg-Feller assumption) this is direct a consequence of the negative association property of the sampling plan and of the CLT given before.

Proof of the result

2) Control of the increments

For this observe that we have for all $(f,g) \in \mathcal{F}^2$, for some constant $C_1, C_2, ...,$

$$\mathbb{E}_{R_N}\left((\mathbb{G}^{\mathbf{p}}_{R_N}(f-g))^2\right)$$

$$\leq C_1 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (\pi_i \pi_j - \pi_{i,j}) \left(\frac{f(X_i) - g(X_i)}{p_i} - \frac{f(X_j) - g(X_j)}{p_j} \right)^2$$

$$\leq C_2 \|f-g\|^2_{2,\mathbb{P}_N}$$

Proof of the result

The process $\{\mathbb{G}^{\mathbf{p}}_{R_N} f : f \in \mathcal{F}\}$ has conditionally sub-Gaussian tails with respect to the semi-metric $\rho^2_{\mathbb{P}_N}$. Then Corollary 2.2.8 in van der Vaart and Wellner(1996), Springer yields

$$\mathbb{E}_{R_N}\left(\sup_{\rho^2_{\mathbb{P}}(f,g)\leq\delta} |\mathbb{G}^{\mathbf{p}}_{R_N} f - \mathbb{G}^{\mathbf{p}}_{R_N} g|\right) \leq K \int_{\epsilon=0}^{\delta} \sqrt{\log\left(\mathcal{N}(\epsilon/2, \mathcal{F}_\delta, L_2(\mathbb{P}_N))\right)} d\epsilon \tag{2}$$

where $K < +\infty$ is a constant and $\mathcal{F}_\delta = \{f - g,\ \rho^2_{\mathbb{P}_N}(f,g) \leq \delta\}$.

Outlines

Confidence bands for the distribution function

A particular case of interest

1. $\mathcal{F} = \{f_y(x) := \mathbb{I}\{x \leq y\}, (x,y) \in \mathcal{X}^2\}$ ➡
   $\mathbb{G}_{R_N}^{\pi(R_N)} f_y = \sqrt{N}\left(F_{R_N}^{\pi(R_N)}(y) - F_N(y)\right)$
2. Fonctional CLT ➡ $\sqrt{N} \sup_{y\in\mathbb{R}} \left|F_{R_N}^{\pi(R_N)}(y) - F_N(y)\right| \xrightarrow[N\to\infty]{\mathcal{D}} \sup_{y\in\mathbb{R}} |\mathbb{G} f_y|$
3. Confidence bands of level $1-\alpha$ for F_N :

$$CB_{1-\alpha} := \left[F_{R_N}^{\pi(R_N)} - \frac{q_{1-\alpha}}{\sqrt{N}},\ F_{R_N}^{\pi(R_N)} + \frac{q_{1-\alpha}}{\sqrt{N}}\right],$$

$q_{1-\alpha}$ quantile of order $1-\alpha$ of $\sup_{y\in\mathbb{R}} |\mathbb{G} f_y|$

Confidence bands for the distribution function

A particular case of interest

1. $\mathcal{F} = \{f_y(x) := \mathbb{I}\{x \leq y\}, (x,y) \in \mathcal{X}^2\}$ ➡
   $\mathbb{G}_{R_N}^{\pi(R_N)} f_y = \sqrt{N}\left(F_{R_N}^{\pi(R_N)}(y) - F_N(y)\right)$
2. Fonctional CLT ➡ $\sqrt{N} \sup_{y\in\mathbb{R}} \left|F_{R_N}^{\pi(R_N)}(y) - F_N(y)\right| \xrightarrow[N\to\infty]{\mathcal{D}} \sup_{y\in\mathbb{R}} |\mathbb{G}f_y|$
3. Confidence bands of level $1-\alpha$ for F_N :

$$CB_{1-\alpha} := \left[F_{R_N}^{\pi(R_N)} - \frac{q_{1-\alpha}}{\sqrt{N}},\ F_{R_N}^{\pi(R_N)} + \frac{q_{1-\alpha}}{\sqrt{N}}\right],$$

$q_{1-\alpha}$ quantile of order $1-\alpha$ of $\sup_{y\in\mathbb{R}} |\mathbb{G}f_y|$

Practically

$q_{1-\alpha}$ unknown (limiting distribution not pivotal as in i.i.d case for continous F) ➡ simulation of the limiting process $\mathbb{G}f_y$

some simulation results

The underlying model

$$X = W + U \bullet W \rightsquigarrow \mathcal{TN}(\mu, \sigma_W^2, w_\star, w^\star) \bullet U \rightsquigarrow \mathcal{N}(0, \sigma_U^2) \bullet W \perp U$$

Inclusion probabilities proportional to W

some simulation results

The underlying model

$$X = W + U \bullet W \rightsquigarrow \mathcal{TN}(\mu, \sigma_W^2, w_\star, w^\star) \bullet U \rightsquigarrow \mathcal{N}(0, \sigma_U^2) \bullet W \perp U$$

Inclusion probabilities proportional to W

Results : an example

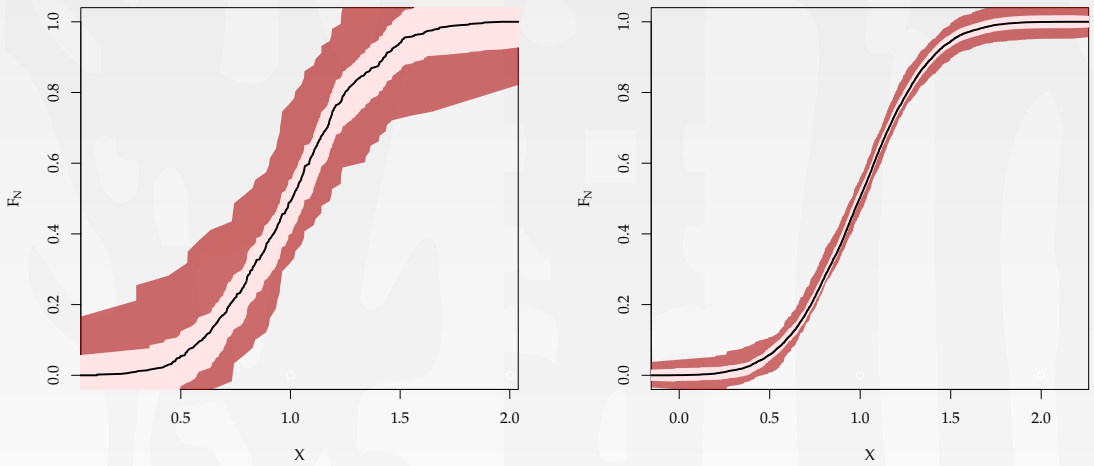

Figure: Example of the 95% confidence bands of the empirical distribution function in the population F_N (black line) with $n/N = 0.1$ (dark pink area) or with $n/N = 0.5$ (light pink area) for $N = 500$ (left hand plot) and $N = 10000$ (right hand plot)

Functionals of the HT-empirical probability

Use the plug-in estimator. Linearize with the influence function... Standard approach.

Tail index estimators based on survey sampling

1. $1 - F(x) = \overline{F}(x) = x^{-1/\gamma}\, L(x)$ $L(x)$ slowly varying function
2. Hill estimator on the whole population $H_{K,N} := \frac{1}{K} \sum_{i=1}^{K} \log\left(\frac{X_{N-i+1,N}}{X_{N-K,N}}\right)$ based on the order statistics.
3. Empirical version of $\gamma = \lim_{x\to\infty} \int_x^{+\infty} \frac{\overline{F}(u)}{\overline{F}(x)} \frac{du}{u}$
4. Horvitz Thompson version

$$\widehat{\gamma} = \int_{X_{n-k,n}}^{+\infty} \frac{\overline{F}_{R_N}^{\boldsymbol{\pi}(R_N)}(u)}{\overline{F}_{R_N}^{\boldsymbol{\pi}(R_N)}(X_{n-k,n})} \frac{du}{u} = \sum_{i=1}^{k} \int_{X_{n-i,n}}^{X_{n-i+1,n}} \frac{\overline{F}_{R_N}^{\boldsymbol{\pi}(R_N)}(u)}{\overline{F}_{R_N}^{\boldsymbol{\pi}(R_N)}(X_{n-k,n})} \frac{du}{u}$$

$$= \left(\sum_{j=1}^{k} \frac{1}{\pi_{n-j+1,n}}\right)^{-1} \sum_{i=1}^{k} \frac{1}{\pi_{n-i+1,n}} \log\left(\frac{X_{n-i+1,n}}{X_{n-k,n}}\right)$$

$$=: H_{k,n}^{\boldsymbol{\pi}}.$$

5. Equivalently

$$H_{k^*,N}^{\pi} = \left(\sum_{j=1}^{K} \frac{\varepsilon_{(N,N-j+1)}}{\pi_{(N,N-j+1)}}\right)^{-1} \sum_{i=1}^{K} \frac{\varepsilon_{(N,N-i+1)}}{\pi_{(N,N-i+1)}} \log\left(\frac{X_{(N,N-i+1)}}{X_{(N,N-K)}}\right)$$

Conditions for Asymptotic normality of the HT-Hill est

Under the von Mises condition The regularly varying survivor function $\overline{F} \in RV_{-\alpha}$ with $\alpha > 0$ is such that there is a real parameter $\rho < 0$, referred to as the *second order parameter*, and a measurable function a of constant sign, whose absolute value lies in RV_ρ such that for any $t > 0$,

$$\lim_{x\to\infty} \frac{\overline{F}(tx)/\overline{F}(x) - t^{-\alpha}}{a(x)} = t^{-\alpha}\frac{t^\rho - 1}{\rho}.$$

i) The marginal cdf F is absolutely continuous with density f.

ii) The joint cdf $F_{X,\mathbf{W}}$ is absolutely continuous with Lebesgue-integrable density $f_{X,\mathbf{W}}$ such that for all $(x, \mathbf{w}) \in (0, +\infty] \times \mathcal{W}$,

$$f_{X,\mathbf{W}}(x, \mathbf{w}) := c\left(F(x), F_{\mathbf{W}}(\mathbf{w})\right) f(x) f_{\mathbf{W}}(\mathbf{w}),$$

for some copula density $c : \mathbb{R}_+^\star \times \mathbb{R}^d \to \mathbb{R}$.

iii) The following integral is finite :

$$\int_{[0,1]^d} c(1, \mathbf{v})\, d\mathbf{v} < \infty.$$

Asymptotic normality of HT-Hill estimator

extension of Theorem 2, Bertail, Chautru, Clémençon, ESAIM, 2015

Suppose that all the conditions required before hold. Then, for

$$\sigma_p^2 := \int_{\mathcal{W}} \frac{1}{p(\mathbf{w})}\, c(1, F_{\mathbf{W}}(\mathbf{w}))\, f_{\mathbf{W}}(\mathbf{w})\, d\mathbf{w}$$

and provided that $k \to +\infty$ as $N \to +\infty$ so that $\sqrt{k} A(N/k) \to \lambda$ for some constant $\lambda \in \mathbb{R}$, we have the convergence in distribution as $N \to +\infty$:

$$\sqrt{k}\,(H_{k,N}^{\pi} - \gamma) \Rightarrow \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\, \sigma_p^2\right). \tag{3}$$

Optimal choice of the weights for Tail estimation

This variational problem may be clearly translated in terms of the simpler finite population problem (for large N)

$$min \frac{1}{N} \sum_{i=1}^{N} \frac{1}{p\left(W_{i,N}\right)} c(1, \frac{i}{N}) \text{ subject to } \sum_{i=1}^{N} p\left(W_{i,N}\right) = n,\ 0 < p\left(W_{i,N}\right) < 1.$$

The Kuhn and Tucker theorem leads to

$$p\left(W_{i,N}\right) = A\ c\left(1, \frac{i}{N}\right)^{1/2}$$

and

$$\sum_{i=1}^{N} p\left(W_{i,N}\right) = A \sum_{i=1}^{N}\ c\left(1, \frac{i}{N}\right)^{1/2} = n = AN\sigma_p^{*2}(1 + o(1))$$

with

$$\sigma_p^{*2} = \int c\left(1, \mathbf{v}\right)^{1/2} d\mathbf{v} = \eta.$$

A few bibliographical references

Survey sampling

- Y.G. Berger(1998) Rate of convergence to normal distribution for the Horvitz-Thompson estimator, J. Stat. Plan. Inf 67, no. 2, 209-226.
- J. Hajek(1964) On the Convergence of the Horvitz-Thompson Estimator, The Annals of Mathematical Statistics 35, no. 4, 1491-1523.
- P. Lopuhäa H. Boistard and A. Ruiz-Gazen(2012) Approximation of rejective sampling inclusion probabilities and application to high order correlations, Elect. J. of Statitics. http://arxiv.org/pdf/1207.5654.pdf.
- Mirakhmedov, Sh. M., Rao Jammalamadaka, S. and Mohamed, I. B. (2014). On Edgeworth expansions in generalized urn models. J. Theoretical Probability 27, 725-753.
- Y. Tillé (2006) Sampling algorithms, Springer Series in Statistics.

A few bibliographical references

Negative dependence

- Borcea, J., Brändén, P. and Liggett, T. M. (2009), Negative dependence and the geometry of polynomials, J. Amer. Math. Soc. 22, 521-567.
- Dubhashi, D. and Ranjan, D. (1998), Balls and bins: A study in negative dependence, Random Structures Algorithms 13, 99-124.
- Joag-Dev, K. and Proschan, F. (1983), Negative association of random variables with applications, Ann. Statist. 11, 286-295.
- Oliveira P.E. (2012). Asymptotics for associated random variables, Springer.
- Patterson, R.F., Smith,W. D., Taylor, R. L. and Bozorgnia, A. (2001), Limit theorems for negatively dependent random variables, Nonlinear Analysis 47, 1283-1295.
- Pemantle, R. (2000), Towards a theory of negative dependence, J. Math. Phys. 41, 1371-1390.
- Yuan M., C. Su and T. Hu (2003), A central limit theorem for random fields of negatively associated processes, J. Theoret. Probab. 16, 309-323

A few bibliographical references

Empirical processes

- P. Bertail, E. Chautru, and S. Clémençon (2017), Empirical processes in survey sampling, Scandinavian Journal of Statistics.
- N.E. Breslow and J.A. Wellner, A Z-theorem with estimated nuisance parameters and correction note for Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression, Scandinavian Journal of Statistics 35 (2008), 186-192.
- T. Saegusa and J.A. Wellner (2013) Weighted likelihood estimation under two-phase sampling, Annals of Statistics, 41, 269-295.
- S. van de Geer, Empirical processes in M-estimation, Cambridge Univ. Press, 2009.
- A. van der Vaart and J. Wellner, Weak convergence and empirical processes: with applications to statistics, Springer

Definition

Strong Rayleigh property (SR). Assume now that E is a discrete (countable) set. Denote F the probability-generating function of the discrete measure μ taking its value on the set E^N, defined by:

$$F(z) = \sum_{x=(x_1,..,x_N)\in E^N} \mu(x) z^x,$$

where $z = (z_1, \ldots, z_N) \in \mathbb{C}^N$ and

$$z^x = \prod_{i=1}^{N} z_i^{x_i} .$$

Then μ is said to be **strongly Rayleigh** (SR) if for all $x \in \mathbb{R}^N$ and $1 \leq i < j \leq N$:

$$F(x)\frac{\partial^2 F}{\partial x_i x_j}(x) \leq \frac{\partial F}{\partial x_i}(x)\frac{\partial F}{\partial x_j}(x).$$

Theorem (Pemantle,2000)