

SMURF 2018 NEUCHATEL

EXPONENTIAL BOUNDS FOR SURVEY SAMPLING,
TOWARDS APPLICATIONS TO BIG-DATA

P. Bertail* • S. Cléménçon

**Modal'X, Université Paris Ouest-Nanterre-La Défense*

Télécom ParisTech, UMR CNRS 5141 LTCI, Groupe TSI

SMURF Neuchatel , august 2018

Motivations

Big Data = here Large data (a lot of individuals) in opposition to high dimensional data (may also be used in this framework...)

Examples :

- Marketing data : 10^{12} clicks per months (on an average of 10^6 products) : even to calculate means the cost of retrieving all the data is so high that it is not done in practice.
- Graph data : for n individuals (nodes), n^2 possible connections (edges) : too large to compute simple statistics (Facebook, Google, etc...), optimizing likelihood is infeasible....
- Original work motivated by some problem linked to **risk minimization in statistical learning** for very large population as well as tail estimation. Bertail, Chautru, Cléménçon (2017), Scand. J. Stat. , (2015), ESAIM + PhD thesis of Emilie Chautru(2015).

Sampling ideas for big data

Survey sampling ideas have been used for a long time in computer science.

- Basic forms : Subsampling (MapReduce) : see also the so-called divide and conquer strategy in statistical literature.
- In Spark, the main method for estimating too large statistics is Poisson sampling (see later).
- For graphs, several survey sampling plans : sample nodes or edges, balls around nodes (snow ball sampling) etc...

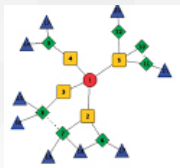


FIGURE: Snowball sampling for graphs

Some sampling ideas in the features dimension (subsampling variables for random forest).

Existing functional results in survey sampling

- Very few functional results even in the real case until recently
- Some general results for general classes of functions for sampling uniformly WR or WoR or for sampling schemes satisfying some exchangeability conditions : particular cases of weighted bootstrap.
- 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.
- J.C. Wang (2012) for empirical cdf, some problems... Boistard, Lopuhää and Ruiz-Gazen (2017), Annals of stat., for general survey sampling plans for the cdf, based on control of fourth order moments. General results for empirical process indexed by (uniformly) Donsker classes of functions, Bertail, Chautru, Cléménçon (2017), Scand. J. of stat, but only for survey sampling close to Poisson or rejective sampling. Recent generalisation by Bertail and Rebecq (2018) for negatively associated survey sampling.

- These are asymptotic results. For statistical learning need to control the risk which can be reduced to controlling the tail of an empirical process indexed by a class of function (= the algorithm) for any (at least) moderate sample size.
- For statistical learning applications in big data, need to obtain concentration inequalities or exponential bounds. Hoeffding inequality for SwOR : an improved result by Bardenet and Maillard (2015), Bernoulli + empirical bounds (that is replacing the true variance by an estimated one).
- Goal of this work : how to obtain Hoeffding or Bernstein type bounds for survey sampling. Then generalize them for general classes of function (easy with standard tools) : essentially a covering number of the class multiplied by an exponential bound. See Clemençon, Bertail and Chautru (2017), Statistics.

- 1 Introduction and notations**
 - Survey sampling plans
 - Poisson and rejective sampling
 - Negative association of survey sampling
- 2 Bernstein bounds for Poisson sampling plans**
 - Tail bounds in the independent case
 - Tail bounds for Poisson sampling
 - Tail bounds for negatively associated sampling plans
- 3 From Poisson to rejective (and more general plans)**
 - Sharp tail for rejective sampling
 - Controlling the distance to Rejective plan
- 4 Confidence bands for the distribution function**
 - Confidence bands for the distribution function
 - Some simulation results

Outlines

- 1 **Introduction and notations**
 - Survey sampling plans
 - Poisson and rejective sampling
 - Negative association of survey sampling
- 2 Bernstein bounds for Poisson sampling plans
 - Tail bounds in the independent case
 - Tail bounds for Poisson sampling
 - Tail bounds for negatively associated sampling plans
- 3 From Poisson to rejective (and more general plans)
 - Sharp tail for rejective sampling
 - Controlling the distance to Rejective plan
- 4 Confidence bands for the distribution function
 - Confidence bands for the distribution function
 - Some simulation results

Inclusion probabilities and survey sampling plans

Sampling units

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

Horvitz-Thompson estimators

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

$$\widehat{S}_\pi^{\epsilon_N} = \sum_{k \in \mathcal{U}} \frac{X_k}{\pi_k} \epsilon_k$$

- CLT's : Pioneering work of Hajek (1964), Ann Math. Stat, very difficult proofs based on coupling arguments respectively for rejective sampling, Pareto sampling Rosen(1997), Ann Stat. , sampling plans close to Rejective sampling (Rao-Sampford, Successive sampling etc...), see Berger(1998), JSPI, by controlling the L1 distance between these plans and rejective sampling.

Some examples

- USWoR (fixed size n) $\mathbb{P}(\mathcal{S} = s) = \mathbb{I}\{\#s = n\} / C_N^n$
- Poisson sampling with same inclusion probabilities $\pi_1 = \dots = \pi_N = \pi$
 $n = \sum_{i=1}^N \epsilon_i$ of random size with expectation $n_0 = N\pi$: and T_N is characterized by

$$T_N(s) = \mathbb{P}(\mathcal{S} = s) = \pi^{\text{card}(s)} (1 - \pi)^{N - \text{card}(s)}$$

- CLT immediate by independence : use Lindenberg-Feller theorem

Poisson survey sampling plan

Definition of Poisson sampling probability

ϵ_i i.i.d. $\mathcal{B}(p_i), i = 1, \dots, N$ $\mathbb{E}(n) = \sum_{i=1}^N p_i$

$$T_N(s) := \mathbb{P}(\mathcal{S} = s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)$$

Poisson survey sampling plan

Definition of Poisson sampling probability

ϵ_i i.i.d. $\mathcal{B}(p_i), i = 1, \dots, N$ $\mathbb{E}(n) = \sum_{i=1}^N p_i$

$$T_N(s) := \mathbb{P}(\mathcal{S} = s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)$$

Properties

- entirely characterized by first order inclusion probabilities 1^{er} ordre
- independence between the ϵ_i (easier!) : CLT immediate.

Poisson survey sampling plan

Definition of Poisson sampling probability

ϵ_i i.i.d. $\mathcal{B}(p_i), i = 1, \dots, N$ $\mathbb{E}(n) = \sum_{i=1}^N p_i$

$$T_N(s) := \mathbb{P}(\mathcal{S} = s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)$$

Properties

- entirely characterized by first order inclusion probabilities 1^{er} ordre
- independence between the ϵ_i (easier!) : CLT immediate.
- p_i eventually function of an auxiliary r.v. $W \sim \mathbf{P}_W$ given for the whole population \mathcal{U}_N :

$$p_i = \mathbb{E}(\epsilon_i \mid W_i) \equiv p(W_i)$$

Rejective sampling plan or Conditional Poisson

Definition

For a fixed n and given inclusion probabilities π_1^R, \dots, π_N^R the rejective sampling has a distribution given by

$$R_N := \operatorname{argmax}_{\mathbf{p}: (\pi_1, \dots, \pi_N) = (\pi_1^R, \dots, \pi_N^R)} - \sum_{\{s: \#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

Rejective sampling plan or Conditional Poisson

Definition

For a fixed n and given inclusion probabilities π_1^R, \dots, π_N^R the rejective sampling has a distribution given by

$$R_N := \operatorname{argmax}_{\mathbf{p}: (\pi_1, \dots, \pi_N) = (\pi_1^R, \dots, \pi_N^R)} - \sum_{\{s: \#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

→ sampling plan of "maximal entropy"

Rejective sampling plan or Conditional Poisson

Definition

For a fixed n and given inclusion probabilities π_1^R, \dots, π_N^R the rejective sampling has a distribution given by

$$R_N := \operatorname{argmax}_{\mathbf{p}: (\pi_1, \dots, \pi_N) = (\pi_1^R, \dots, \pi_N^R)} - \sum_{\{s: \#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

→ sampling plan of "maximal entropy", or **conditionnal Poisson**

Rejective sampling plan or Conditional Poisson

Definition

For a fixed n and given inclusion probabilities π_1^R, \dots, π_N^R the rejective sampling has a distribution given by

$$R_N := \operatorname{argmax}_{\mathbf{p}: (\pi_1, \dots, \pi_N) = (\pi_1^R, \dots, \pi_N^R)} - \sum_{\{s: \#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

➔ sampling plan of "maximal entropy", or **conditionnal Poisson**

Link to the Poisson plan

- Draw a sample \mathcal{S} according to a Poisson plan with well chosen inclusion probabilities p_1, \dots, p_N with $\sum_{i=1}^N p_i = n$ (canonical Poisson sampling)
- If $\#\mathcal{S} = n$ keep \mathcal{S} , else draw a new sample.

Rejective sampling plan or Conditional Poisson

Definition

For a fixed n and given inclusion probabilities π_1^R, \dots, π_N^R the rejective sampling has a distribution given by

$$R_N := \operatorname{argmax}_{\mathbf{p}: (\pi_1, \dots, \pi_N) = (\pi_1^R, \dots, \pi_N^R)} - \sum_{\{s: \#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

→ sampling plan of "maximal entropy", or **conditionnal Poisson**

Link to the Poisson plan

- Draw a sample \mathcal{S} according to a Poisson plan with well chosen inclusion probabilities p_1, \dots, p_N with $\sum_{i=1}^N p_i = n$ (canonical Poisson sampling)
 - If $\#\mathcal{S} = n$ keep \mathcal{S} , else draw a new sample.
- link between (p_1, \dots, p_N) and $(\pi_1^R, \dots, \pi_N^R)$ given in Hájek (1964)

Rejective sampling plan or Conditional Poisson

Definition

For a fixed n and given inclusion probabilities π_1^R, \dots, π_N^R the rejective sampling has a distribution given by

$$R_N := \operatorname{argmax}_{\mathbf{p}: (\pi_1, \dots, \pi_N) = (\pi_1^R, \dots, \pi_N^R)} - \sum_{\{s: \#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

→ sampling plan of "maximal entropy", or **conditionnal Poisson**

Link to the Poisson plan

- Draw a sample \mathcal{S} according to a Poisson plan with well chosen inclusion probabilities p_1, \dots, p_N with $\sum_{i=1}^N p_i = n$ (canonical Poisson sampling)
- If $\#\mathcal{S} = n$ keep \mathcal{S} , else draw a new sample.

→ link between (p_1, \dots, p_N) and $(\pi_1^R, \dots, \pi_N^R)$ given in Hájek (1964) considerably reduce the variance of estimators because of the fixed size.

Conditional balls and bin sampling are negatively associated

- Rejective sampling = Poisson sampling conditional to the size equal a fixed n
- Subsampling = Rejective sampling with equal inclusion probabilities
- Pareto sampling, order sampling : but too costly for big data , need to generate N uniform r.v.'s and to reorder them.
- Pivotal sampling or Srinivasan sampling
- Balanced sampling (which respects some margin conditions) using the Cube Method (Deville and Tillé, 2004) : very efficient and "almost" balanced.
- All these methods are Conditional balls and bin sampling (Dubhashi & Ranjan (1998)) and enjoy a great property : Negative association !

Negative association

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

Definition

The r.v.'s Z_1, \dots, 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 $\llbracket 1, N \rrbracket$

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

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

Negative association for survey sampling plans

- Importance of negative association 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 very clever 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 : see Patterson, Smith, Taylor, Bozorgnia(2001), *Nonlinear Analysis*. Oliveira (2012), *Asymptotics for assoc. r.v.'s*, Springer. See also recent applications in bayesian statistics in Gerger, Chopin, Whitley(2019), Arxiv

Negative association for survey sampling plans

- The usual CLT's (or functional CLT, Louichi,1999, Ann. IHP) proved for negative associated r.v's not sufficient to get CLT for estimators in survey sampling (the covariance are too big...) but works for generalized bootstrap resampling plans.
- Using results by Utev and Peligrad (2006), Ann. Proba, it is possible to get general CLT for survey sampling : Bertail and Rebecq(2018).
- Negative association leads to exponential bounds immediately : but not efficient bounds...

Outlines

- 1 Introduction and notations
 - Survey sampling plans
 - Poisson and rejective sampling
 - Negative association of survey sampling
- 2 Bernstein bounds for Poisson sampling plans**
 - Tail bounds in the independent case
 - Tail bounds for Poisson sampling
 - Tail bounds for negatively associated sampling plans
- 3 From Poisson to rejective (and more general plans)
 - Sharp tail for rejective sampling
 - Controlling the distance to Rejective plan
- 4 Confidence bands for the distribution function
 - Confidence bands for the distribution function
 - Some simulation results

Tail bounds for sums in the independent case

Hoeffding (1963).

Theorem (Hoeffding's inequality)

Let Z_1, Z_2, \dots, Z_n be independent random variables such that $a_i \leq Z_i \leq b_i$ ($i = 1, \dots, n$), then for $t > 0$

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

Tail bounds for sums in the independent case

Theorem (Bernstein's inequality)

Let Z_1, \dots, Z_n be independent random variables with variance $\sigma_i^2 < \infty$, such that, for all integers $p \geq 2$,

$$\mathbb{E}|Z_i|^p \leq p! M^{p-2} \sigma_i^2 / 2 \text{ for all } i \in \{1, \dots, n\}.$$

Then, for all $t > 0$,

$$\mathbb{P} \left(\sum_{i=1}^n (Z_i - \mathbb{E}Z_i) \geq t \right) \leq \exp \left(-\frac{t^2}{2(\sigma^2 + Mt/3)} \right),$$

where $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Remark : if the Z_i 's are bounded by M , the same results holds.

Tail bounds for Poisson sampling

Theorem

(POISSON SAMPLING) Assume that the survey scheme ϵ_N defines a Poisson sampling plan with first order inclusion probabilities $p_i > 0$, with $1 \leq i \leq N$. Then, we have almost-surely : $\forall t > 0, \forall N \geq 1$,

$$\mathbb{P} \left\{ \widehat{S}_{p_N}^{\epsilon_N} - S_N > t \right\} \leq \exp \left(- \frac{\sum_{i=1}^N \frac{1-p_i}{p_i} x_i^2}{\left(\max_{1 \leq i \leq N} \frac{x_i}{p_i} \right)^2} H \left(\frac{\max_{1 \leq i \leq N} \frac{|x_i|}{p_i} t}{\sum_{i=1}^N \frac{1-p_i}{p_i} x_i^2} \right) \right) \quad (2)$$

$$\leq \exp \left(\frac{-t^2}{2 \sum_{i=1}^N \frac{1-p_i}{p_i} x_i^2 + \frac{2}{3} \max_{1 \leq i \leq N} \frac{|x_i|}{p_i} t} \right), \quad (3)$$

where $H(t) = (1+t) \log(1+t) - t$ for $t \geq 0$.

Proof : These bounds result from Bennett inequality and Bernstein exponential inequality respectively, when applied to $(\epsilon_i/p_i)x_i$, $1 \leq i \leq N$.

Then we have (see Janson (1994), unpublished manuscript, Shao(2000), J. Theor. Probability

Theorem

Let $N \geq 1$ and $\epsilon_N^* = (\epsilon_1^*, \dots, \epsilon_N^*)$ be the vector of indicator variables related to a NA scheme on \mathcal{I}_N with first order inclusion probabilities $(\pi_1, \dots, \pi_N) \in]0, 1]^N$. Then, for any $t \geq 0$ and $N \geq 1$, we have :

$$\begin{aligned} \mathbb{P} \left\{ \widehat{S}_{\pi}^{\epsilon_N^*} - S_N \geq t \right\} &\leq 2 \exp \left(- \frac{\sum_{i=1}^N \frac{1-\pi_i}{\pi_i} x_i^2}{\left(\max_{1 \leq i \leq N} \frac{x_i}{\pi_i} \right)^2} H \left(\frac{\max_{1 \leq i \leq N} \frac{|x_i|}{\pi_i} t/2}{\sum_{i=1}^N \frac{1-\pi_i}{\pi_i} x_i^2} \right) \right) \\ &\leq 2 \exp \left(\frac{-t^2}{\frac{2}{3} \max_{1 \leq i \leq N} \frac{|x_i|}{\pi_i} t + 2 \sum_{i=1}^N \frac{1-\pi_i}{\pi_i} x_i^2} \right). \end{aligned}$$

Outlines

- 1 Introduction and notations
 - Survey sampling plans
 - Poisson and rejective sampling
 - Negative association of survey sampling
- 2 Bernstein bounds for Poisson sampling plans
 - Tail bounds in the independent case
 - Tail bounds for Poisson sampling
 - Tail bounds for negatively associated sampling plans
- 3 From Poisson to rejective (and more general plans)**
 - Sharp tail for rejective sampling
 - Controlling the distance to Rejective plan
- 4 Confidence bands for the distribution function
 - Confidence bands for the distribution function
 - Some simulation results

From Poisson to Rejective sampling

A simple interpretation of Hajek(1964)'s work and a much-much-much more simpler proof of the CLT for conditional Poisson sampling scheme including rejective sampling, successive sampling (not negatively associated), Rao-Sampford sampling, stratified sampling and so on...

Consider the regression

$$N^{-1} \sum_{i=1}^N \varepsilon_i \frac{X_i}{p_i} - \bar{X}_N = RN^{-1} \left(\sum_{i=1}^N \varepsilon_i - n \right) + \eta_N$$

with η_N orthogonal to $\sum_{i=1}^N \varepsilon_i$.

then easy calculations shows that

$$\begin{aligned} R &= \text{cov} \left(N^{-1} \sum_{i=1}^N \varepsilon_i \frac{X_i}{p_i}, \sum_{i=1}^N \varepsilon_i - n \right) / V \left(N^{-1} \sum_{i=1}^N \varepsilon_i \right) \\ &= \frac{\sum_{i=1}^N X_i (1 - p_i)}{\sum_{i=1}^N p_i (1 - p_i)} = \theta_N \end{aligned}$$

For rejective sampling (Poisson sampling conditionally to a fixed size) consider the distribution of the Horvitz Thompson mean under rejective sampling. If we denote by ε_i 's the plan corresponding to the corresponding Poisson sampling, this is given by

$$\begin{aligned}
 & P\left(\sqrt{N}\left(N^{-1}\sum_{i=1}^N\varepsilon_i\frac{X_i}{p_i}-\bar{X}_N\right)\leq x\mid N^{-1/2}\sum_{i=1}^N\varepsilon_i=N^{-1/2}n\right) \\
 &= P\left(\sqrt{N}\eta_N\leq x\mid N^{-1/2}\sum_{i=1}^N\varepsilon_i=N^{-1/2}n\right) \\
 &= \frac{P\left(\sqrt{N}\eta_N\leq x, N^{-1/2}\sum_{i=1}^N\varepsilon_i=N^{-1/2}n\right)}{P\left(N^{-1/2}\sum_{i=1}^N\varepsilon_i=N^{-1/2}n\right)}
 \end{aligned}$$

Variance reduction

Define $d_N = \sum_{i=1}^N p_i(1 - p_i)$ the variance $\text{Var}(\sum_{i=1}^N \epsilon_i)$ of the size of the Poisson plan We have the following decomposition

$$\text{Var} \left(\sum_{i=1}^N \frac{\epsilon_i}{p_i} x_i \right) = \sigma_N^2 + \theta_N^2 d_N, \quad (4)$$

where

$$\sigma_N^2 = \text{Var} \left(\sum_{i=1}^N (\epsilon_i - p_i) \left(\frac{x_i}{p_i} - \theta_N \right) \right) \quad (5)$$

is the asymptotic variance of the statistic $\widehat{S}_{p_N}^{\epsilon_N^*}$, see [Haj64].

Sharp tail for rejective sampling

Theorem

Let $N \geq 1$. Suppose that ϵ_N^* is a rejective scheme of size $n \leq N$ with canonical parameter $\mathbf{p}_N = (p_1, \dots, p_N) \in]0, 1[^N$. Set $X_N = 2 \max_{1 \leq j \leq N} |x_j|/p_j$. Then, there exist universal constants C and D such that we have for all $t > 0$ and for all $N \geq 1$,

$$\begin{aligned} \mathbb{P} \left\{ \widehat{S}_{\mathbf{p}_N}^{\epsilon_N^*} - S_N > t \right\} &\leq C \exp \left(-\frac{\sigma_N^2}{X_N^2} H \left(\frac{tX_N}{\sigma_N^2} \right) \right) \\ &\leq C \exp \left(-\frac{t^2}{2 \left(\sigma_N^2 + \frac{1}{3} tX_N \right)} \right), \end{aligned}$$

as soon as $\min\{d_N, d_N^*\} \geq 1$ and $d_N \geq D$.

Ideas of the proofs

- Idea similar to MC tail bounds obtained by Bertail and Cléménçon(2010) Probability Theory and its applications, for Markov chains.
- For the denominator : two solutions. Either use an Edgeworth expansion or sharp Berry-Esséen Bound (see for instance Deheuvels, Puri, Ralescu (1989), J. Multivariate Analysis) : unfortunately a precise analysis shows that the constant C is very big. Under some additional assumptions assuming that all the weights are of the same order n/N , then if $p_i \geq cn/N$ for all $i \in \{1, \dots, N\}$. Then, we have : $\forall N \geq 1$, $\forall n < N$, $\mathbb{P}\{\mathcal{M}_N = 0\} \geq \frac{e^{-1/6}\sqrt{c}}{\sqrt{2\pi d_N}}$.
- For the numerator, exponential change (Esscher transform P_u, N) and then get a Chernoff-bound combined again with an upper Berry Esseen Bound for $P_{u,N}\{\mathcal{M}_N = 0\}$, an idea originally introduced by Talagrand(1995), the missing factor in Hoeffding inequality, Ann. IHP.

Sharp tail for rejective sampling with inclusion prob.

Theorem

Suppose that the assumptions of preceding Theorem are fulfilled and set $M_N = (6/d_N) \sum_{i=1}^N |x_i|/\pi_i$ and $X_N = 2 \max_{1 \leq j \leq N} |x_j|/p_j$. The following assertions hold true.

(i) For all $N \geq 1$, we have almost-surely :

$$\left| \widehat{S}_{\pi_N}^{\epsilon_N^*} - \widehat{S}_{p_N}^{\epsilon_N^*} \right| \leq M_N.$$

(ii) There exist universal constants C and D such that, for all $t > M_N$ and for all $N \geq 1$, we have :

$$\begin{aligned} \mathbb{P} \left\{ \widehat{S}_{\pi_N}^{\epsilon_N^*} - S_N > t \right\} &\leq C \exp \left(-\frac{\sigma_N^2}{X_N^2} H \left(\frac{(t - M_N) X_N}{\sigma_N^2} \right) \right) \\ &\leq C \exp \left(-\frac{(t - M_N)^2}{2 \left(\sigma_N^2 + \frac{1}{3} (t - M_N) X_N \right)} \right), \end{aligned}$$

as soon as $\min\{d_N, d_N^*\} > 1$ and $d_N > D$

Idea of the proof

Analysis of the approximations in Hajek (1964), Ann. Math. Stat. shows that

Lemma

Let π_1, \dots, π_N be the first order inclusion probabilities of a rejective sampling of size n with canonical representation characterized by the Poisson weights p_1, \dots, p_N . Provided that $d_N = \sum_{i=1}^N p_i(1 - p_i) \geq 1$, we have : $\forall i \in \{1, \dots, N\}$,

$$\left| \frac{1}{\pi_i} - \frac{1}{p_i} \right| \leq \frac{6}{d_N} \times \frac{1 - \pi_i}{\pi_i}.$$

Controlling the distance between sampling plans

Same ideas as Berger(1998), JSPI

- Total Variation : $\|\tilde{T}_N - R_N\|_1 := \sum_{s \subset \mathcal{P}_N} |R_N(s) - T_N(s)|$
- Entropy : $D_{KL}(T_N \| \tilde{R}_N) := \sum_{s \subset \mathcal{P}_N} T_N(s) \log \frac{T_N(s)}{\tilde{R}_N(s)}$

Lemma

Let ϵ_N and $\tilde{\epsilon}_N$ be two schemes defined on the same probability space and drawn from plans R_N and \tilde{R}_N respectively and let $\mathbf{p}_N \in]0, 1]^N$.

Then, we have : $\forall N \geq 1, \forall t \in \mathbb{R}$,

$$\begin{aligned} \left| \mathbb{P} \left\{ \widehat{S}_{\mathbf{p}_N}^{\epsilon_N} - S_N > t \right\} - \mathbb{P} \left\{ \widehat{S}_{\mathbf{p}_N}^{\tilde{\epsilon}_N} - S_N > t \right\} \right| &\leq \|\tilde{R}_N - R_N\|_1 \\ &\leq \sqrt{2D_{KL}(R_N \| \tilde{R}_N)}. \end{aligned}$$

Outlines

- 1 Introduction and notations
 - Survey sampling plans
 - Poisson and rejective sampling
 - Negative association of survey sampling
- 2 Bernstein bounds for Poisson sampling plans
 - Tail bounds in the independent case
 - Tail bounds for Poisson sampling
 - Tail bounds for negatively associated sampling plans
- 3 From Poisson to rejective (and more general plans)
 - Sharp tail for rejective sampling
 - Controlling the distance to Rejective plan
- 4 **Confidence bands for the distribution function**
 - Confidence bands for the distribution function
 - Some simulation results

Confidence bands for the distribution function

A particular case of interest

① $\mathcal{F} = \{f_y(x) := \mathbb{I}\{x \leq y\}, (x, y) \in \mathcal{X}^2\} \rightarrow$

$$\mathbb{G}_{R_N}^{\pi(R_N)} f_y = \sqrt{N} (F_{R_N}^{\pi(R_N)}(y) - F_N(y))$$

② Functional CLT $\rightarrow \sqrt{N} \sup_{y \in \mathbb{R}} |F_{R_N}^{\pi(R_N)}(y) - F_N(y)| \xrightarrow[N \rightarrow \infty]{\mathcal{D}} \sup_{y \in \mathbb{R}} |\mathbb{G}f_y|$

③ Confidence bands of level $1 - \alpha$ for F_N

Illustration

The underlying model

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

Inclusion probabilities proportional to W

Need to have an upper bound for the maximum of the X_i 's

Illustration

The underlying model

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

Inclusion probabilities proportional to W

Need to have an upper bound for the maximum of the X_i 's

Results : an example

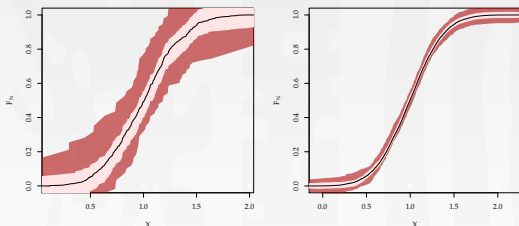











FIGURE: Example of the 95% confidence bands of the empirical distribution function in the population F_N (black line) with $c = 0.1$ (dark pink area) or with $c = 0.5$ (light pink area) for $N = 500$ (left hand plot) and $N = 10000$ (right hand

- A lot of extensions : concentration for HT-empirical process : application to empirical risk minimization based on survey sampling techniques. Tailored to rejective or conditional Poisson techniques. What about Pareto, cluster or snow ball sampling ?
- Optimal choice of weights depending on the density of (X, W) where W is an auxiliary variable observed on the whole database. How to choose the weights : depends on the variable of interest but essentially amounts to minimize the variance in the bound.
- Application to HT- gradient descent for very large datasets, with adequate choice of the weights : improve over mini-batch techniques.

A few bibliographical references

- 
- Patrice Bertail, Emilie Chautru, and Stéphan Cléménçon, *Empirical processes in survey sampling with (conditional) poisson designs*, *Scandinavian Journal of Statistics* **44** (2017), 97–111.
- 
- Y.G. Berger, *Rate of convergence to normal distribution for the Horvitz-Thompson estimator*, *J. Stat. Plan. Inf* **67** (1998), no. 2, 209–226.
- 
- H. Boistard, P. Lopuskaa, and A. Ruiz-Gazen, *Functional central limit theorems for single-stage sampling designs*, *The Annals of Statistics* **45** (2017), 1728–1758.
- 
- R. Bardenet and O.A. Maillard, *Concentration inequalities for sampling without replacement*, *Bernoulli* **21** (2015), no. 3, 1361–1385.
- 
- 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.
- 
- J. Hajek, *On the Convergence of the Horvitz-Thompson Estimator*, *The Annals of Mathematical Statistics* **35** (1964), no. 4, 1491–1523.
- 
- S. Janson, *Large deviation inequalities for sums of indicator variables*, Unpublished manuscript, available at www2.math.uu.se/~svante/papers/sj107.ps (1994).
- 
- Q.M. Shao, *A Comparison Theorem on Moment Inequalities Between Negatively Associated and Independent Random Variables*, *Journal of Theoretical Probability* **13** (2000), no. 2, 343–356.
- 
- T. Saegusa and J.A. Wellner, *Weighted likelihood estimation under two-phase sampling*, Preprint available at <http://arxiv.org/abs/1112.4951v1> (2011).