Inégalités de concentration pour le sondage aléatoire simple

Daniel Bonnery

INÉGALITÉS DE CONCENTRATION POUR LE SONDAGE ALÉATOIRE SIMPLE

Daniel Bonnéry

ENSAI - Campus de Ker-Lann,
Rue Blaise Pascal - BP 37203,
35172 BRUZ cedex

Résumé

En théorie des sondages, la loi de l'estimateur de Horvitz Thompson de la moyenne d'un caractère d'une population est, en pratique, approximée par une loi normale. Cette approximation est justifiée par un théorème de Hàjek ([1]), qui démontre la convergence vers une loi normale de l'estimateur de Horvitz Thompson sous certaines conditions. Il existe aussi des résultats à distance finie, notamment ceux de Hoeffding ([2]) et Serfling ([3]) qui permettent d'obtenir des inégalités de concentration étant donné la dispersion du caractère sur la population. Toutefois, les inégalités obtenues s'avèrent très larges et sont peu utilisées en pratique.

Lors de l'exposé, on s'attachera à montrer comment obtenir des inégalités de concentration plus fines pour ce problème particulier en utilisant des outils moins puissants, mais aussi moins généraux. Dans un premier temps en utilisant les symétries de l'espace des échantillons de taille n, et dans un second temps, en munissant l'espace des échantillons de la distance de Hamming, et en considérant la mesure des boules centrées en un échantillon.

Enfin, on comparera les inégalités obtenues avec les inégalités de Hoeffding et Serfling.

Mots clef : Sondages - Statistique mathématique

The normal approximation of the Horvitz Thompson estimator distribution is widely used by survey statisticians. This approximation, under some asymptotic assumptions, is allowed by a theorem from Hàjek ([1]). Nevertheless, finite distance results do exist, especially concentration inequalities from Hoeffding ([2]) and Serfling ([3]). But those inequalities prove to be too wide to be prefered to the normal approximation.

The aim of the talk is to show how to obtain saturated concentration inequalities in this specific problems by using less powerful but less general tools than the ones used by Hoeffding and Serfling. First by using the symetries of the set of n-sized samples and second, by using the Hamming distance among samples.

1 Notations, définitions

Il s'agit de calculer, $\forall x \in \mathbb{R}$:

$$max \left\{ P\left(\frac{n\left(\hat{\bar{t}}_y - \bar{t}_y\right)}{\sigma_y} \geq x \right) \middle| y \in \mathbb{R}^N, \sigma_y > 0 \right\} = \frac{\Psi(x)}{C_N^n}$$

comme une fonction de n et N où $\hat{\bar{t}}_y$ est l'estimateur de Horvitz-Thompson de la moyenne $\bar{t}_y = \frac{\sum_{k=1}^N y_k}{N}$ du paramètre y associé au sondage aléatoire simple de n unités parmi N, et $\sigma_y = \sqrt{\frac{\sum_{k=1}^N (y_k - \bar{t}_y)^2}{N-1}}$

On note $\mathscr{S}_n$ la famille des sous ensembles de $[\![1, N]\!]$ de taille n. s désignera un élément de $\mathscr{S}_n$.

P est la probabilité uniforme sur $\mathscr{S}_n$

Si $S \subset \mathscr{S}_n$, alors $P(S) = \frac{\#S}{\#\mathscr{S}_n} = \frac{\#S}{C_N^n}$

Pour $y \in \mathbb{R}^N$, $\hat{\bar{t}}_y(s) = \frac{\sum_{k \in s} y_k}{n}$ désigne la moyenne de y sur l'échantillon s. Quand $s \sim P$, alors $\hat{\bar{t}}_y(s)$ est l'estimateur de Horvitz Thompson de la moyenne $\bar{t}_y = \frac{\sum_{k=1}^N y_k}{N}$

On assimile $s \in \mathscr{S}$ au vecteur $s \in \mathbb{R}^N$ défini par $s_k = 1$ si $k \in s$, 0 sinon. On a alors : $n\hat{\bar{t}}_y(s) = {}^t sy$ si $s \in \mathscr{S}_n$ et quand $s \sim P$ et $\bar{t}_y = 0$: $P(n\hat{\bar{t}}_y \geq x) = \frac{\#\{s \in \mathscr{S} | {}^t sy \geq x\}}{C_N^n}$

Soit $\mathscr{B}$ le sous ensemble suivant de $\mathbb{R}^N$:

$$\mathscr{B} = \left\{ y \in \mathbb{R}^N | \ \|y\| = 1, \bar{y} = 0 \right\}$$

Une définition équivalente de Ψ est alors :

$$\begin{array}{rcll} \Psi & : & \bar{\mathbb{R}} & \to [\![0, C_N^n]\!] \\ & & x & \mapsto \Psi(x) \quad = C_N^n \max\left\{ P\left({}^t sy \geq x\right) | y \in \mathscr{B} \right\} \\ & & & \qquad\quad = \max\left\{ \#\left\{ s \in \mathscr{S}_n | {}^t sy \geq x \right\} | y \in \mathscr{B} \right\} \end{array}$$

2 Utilisation des symétries de $\mathscr{S}_n$

On constate que le calcul de $\Psi(x)$ consiste à recherche le $\frac{1}{2}$ espace de $\mathbb{1}^\perp$ éloigné d'une distance x de 0 et contenant un nombre maximal d'éléments de $\mathscr{S}_n$ Ce demi-espace est caractérisé par un vecteur $y \in \mathscr{B}$ normal à sa frontière, et un sous ensemble S de $\mathscr{S}_n$ contenu dans ce demi-espace.

En utilisant les symétries de S, c'est à dire le stabilisateur G_S de S par l'action du groupe symétrique d'ordre N sur $\mathbb{R}^N$ par permutation des coordonnées, on en déduit que G_S est aussi le stabilisateur de $\{ y \in \mathscr{B} | \Psi(x) = \min\{ {}^t sy | s \in S \} \}$

Ces propriétés permettent de calculer Ψ pour certaines valeurs de x.

Ce raisonnement peut être illustré très clairement à partir de l'étude de $\mathscr{S}_n$ pour $N = 4$ et $n = 2$ ou $n = 3$: dans ce cas, $\mathscr{S}_n$ est respectivement le tétraèdre régulier ou l'octaèdre régulier, et la détermination de y et S en fonction de x se fait naturellement à partir de l'observation de la figure.

3 Utilisation d'une distance sur $\mathscr{S}_n$

On définit la distance de Hamming sur $\mathscr{S}_n$ par

$$\begin{aligned} d(s,s') &= \frac{\#(s \setminus s' \cup s' \setminus s)}{2} \\ &= \frac{\sum_{k=1}^{N} |s_k - s'_k|}{2} \end{aligned}$$

On définit aussi, pour $S \subset \mathscr{S}, \varepsilon \in \mathbb{R}^+$:

$$\begin{aligned} d(.,S) &: s \in \mathscr{S} \mapsto d(s,S) = \min\{d(s,s') \,|s' \in \mathscr{S}_n\} \\ S_{>\varepsilon} &= \{s \in \mathscr{S}_n \,|d(s,S) > \varepsilon\} \end{aligned}$$

Soit $y \in \mathscr{B}$ tel que $y_1 \leq \ldots \leq y_N$
Alors $\min\{{}^t sy | s \in \mathscr{S}_n\} = \sum_{k=1}^{n} y_k = {}^t s_y y$
avec $s_y = [\![1, n]\!]$
Et pour $k \in [\![0, n]\!]$, on a l'inégalité suivante :

$$d(s_y, s) = k \Rightarrow |{}^t sy - {}^t s_y y| = {}^t sy - {}^t s_y y \leq -\sum_{j=1}^{k} y_j + \sum_{j=N-k+1}^{N} y_j$$

qui est équivalente à :

$${}^t sy > \sum_{j=k+1}^{n} y_j + \sum_{j=N-k+1}^{N} y_j \Leftarrow s \in \{s_y\}_{>k} = \{s_y\}_{\geq k+1}$$

Par ailleurs, pour

$$k \leq \min\{n-1, N-n-1\}, \#\{s_y\}_{>k} = \sum_{j=k+1}^{min\{n,N-n\}} C_n^{n-j} C_{N-n}^{j}$$

Nous avons donc directement l'inégalité : $\forall k \in [\![0, n-1]\!]$:

$$P\left({}^t sy > \sum_{j=k+1}^{n} y_j + \sum_{j=N-k+1}^{N} y_j\right) \leq M_k = \frac{\sum_{j=k+1}^{min\{n,N-n\}} C_n^{n-j} C_{N-n}^{j}}{C_N^n}$$

Il suffit donc de maximiser $\sum_{j=k+1}^{n} y_j + \sum_{j=N-k+1}^{N} y_j$ sur $\mathscr{B}$ pour obtenir une majoration de $\Psi(X_k)$ avec $X_k = \max\{\sum_{j=k+1}^{n} y_j + \sum_{j=N-k+1}^{N} y_j | \mathscr{B}\}$

Le maximum est obtenu pour y de la forme : $y = -\sqrt{\frac{N-q}{Nq}}\mathbb{1}_q + \sqrt{\frac{q}{(N-q)(N)}}(\mathbb{1}_N - \mathbb{1}_q)$, $q \leq n$ où $\mathbb{1}_q$ désigne le vecteur de $\mathbb{R}^N$ dont les q premières coordonnées sont égales à 1, et les $N - q$ dernières à 0

Conclusion

Lorsqu'on ne parvient pas à calculer explicitement Ψ, on arrive à le majorer à partir d'une fonction prenant n valeurs distinctes. L'inégalité obtenue s'avère meilleure que les inégalités de Serfling et Hoeffding, et fournit des résultats à distance finie non triviaux.

Bibliographie

[1] Wassily Hoeffding. (1963) *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association, 58(301):13-30
[2] R. J. Serfling. (1974), *Probability inequalities for the sum in sampling without replacement*, The Annals of Statistics, 2(1):39-48