Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Une propriété intéressante de l'entropie de certains plans d'échantillonnage

par Yves Tillé et David Haziza

Décembre 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2010
Vol. 36, N° 2, pp. 247-249
Statistique Canada, N° 12-001-X au catalogue

Une propriété intéressante de l'entropie de certains plans d'échantillonnage

Yves Tillé et David Haziza [1]

Résumé

Dans cette note brève, nous montrons que l'échantillonnage aléatoire sans remise et l'échantillonnage de Bernoulli ont à peu près la même entropie quand la taille de la population est grande. Nous donnons un exemple empirique en guise d'illustration.

Mots clés : Échantillonnage de Poisson conditionnel ; entropie ; échantillonnage aléatoire simple ; échantillonnage de Poisson.

1. Introduction

Considérons une population finie de taille N et soit $U = \{1, ..., k, ..., N\}$ l'ensemble des étiquettes de cette population. Un échantillon s est un sous-ensemble de U et un plan d'échantillonnage est une loi de probabilité $p(.)$ sur les sous-ensembles de U telle que $p(s) \geq 0$ pour tout $s \subset U$, et

$$\sum_{s \subset U} p(s) = 1.$$

Soit $\pi_k = P(k \in s)$ la probabilité d'inclusion de premier ordre de l'unité k dans l'échantillon :

$$\pi_k = \sum_{\substack{s \subset U \\ s \ni k}} p(s).$$

De même, soit $\pi_{k\ell} = P(k \in s \text{ et } \ell \in s)$ la probabilité d'inclusion de deuxième ordre des unités k et ℓ dans l'échantillon :

$$\pi_{k\ell} = \sum_{\substack{s \subset U \\ s \ni k,\, \ell}} p(s).$$

L'entropie d'un plan d'échantillonnage $p(.)$, désignée par $I(p)$, est définie comme étant

$$I(p) = -\sum_{s \in Q} p(s) \log p(s), \qquad (1)$$

où $Q = \{s \mid p(s) > 0\}$ est le support du plan d'échantillonnage $p(.)$. Un plan d'échantillonnage possède une entropie élevée quand le degré d'incertitude ou de surprise est grand en ce qui concerne l'échantillon qui sera sélectionné. Autrement dit, quand un plan d'échantillonnage possède une entropie élevée, il est très difficile de prédire le type d'échantillon que l'on obtiendra. De nombreux plans d'échantillonnage utilisés en pratique sont des plans à entropie élevée. L'échantillonnage systématique, dont l'entropie est très faible, est une exception notable. Le concept d'entropie est utile dans le contexte de l'estimation de la variance. Lorsque l'entropie d'un plan d'échantillonnage est élevée, il est possible d'obtenir une approximation des probabilités d'inclusion de deuxième ordre, $\pi_{k\ell}$, en fonction des probabilités d'inclusion de premier ordre, ce qui simplifie considérablement le problème d'estimation de la variance dans le contexte de l'échantillonnage avec probabilités inégales ; voir, par exemple, Brewer et Donadio (2003), Matei et Tillé (2005), Henderson (2006) et Haziza, Mecatti et Rao (2008).

Il est bien connu que le plan d'échantillonnage à entropie maximale est l'échantillonnage de Poisson :

$$p_{\text{poiss}}(s) = \left(\prod_{k \in s} \pi_k \right) \left(\prod_{k \in U \setminus s} (1 - \pi_k) \right) \qquad (2)$$

pour tout $s \in Q$; voir par exemple, Tillé (2006). Un cas particulier de l'échantillonnage de Poisson est l'échantillonnage de Bernoulli, que l'on obtient à partir de (2) en posant que $\pi_k = \pi \in (0, 1)$, ce qui mène à

$$p_{\text{bern}}(s) = \pi^{n_s} (1 - \pi)^{N - n_s}, \text{ pour tout } s \subset U,$$

où n_s est la taille aléatoire de s. En utilisant (1) et en notant que $\sum_{s \in Q} n_s p(s) = N\pi$, l'entropie de l'échantillonnage de Bernoulli est donnée par

$$I(p_{\text{bern}}) = -N(1 - \pi) \log(1 - \pi) - N\pi \log \pi, \qquad (3)$$

qui est maximale quand $\pi = 1/2$. Dans ce cas, nous avons $I(p_{\text{bern}}) = N \log 2$.

Si nous nous limitons à la classe des plans d'échantillonnage avec taille fixe et probabilités d'inclusion de premier ordre π_k, $k \in U$, le plan dont l'entropie est maximale

1. Yves Tillé, Institut de Statistique, Université de Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Suisse ; David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, QC, Canada, H3C 3J7. Courriel : haziza@dms.umontreal.ca.

est celui connu sous le nom d'échantillonnage de Poisson conditionnel (EPC) ; (voir Chen, Dempster et Liu 1994 ; Deville 2000 ; Tillé 2006). Le plan EPC peut être mis en œuvre en sélectionnant des échantillons à plusieurs reprises conformément à l'échantillonnage de Poisson jusqu'à l'obtention de la taille d'échantillon souhaitée, disons n. Quand $\pi_k = n/N$ pour tout $k \in U$, le plan EPC se réduit à l'échantillonnage aléatoire simple sans remise :

$$p_{\text{eas}}(s) = \binom{N}{n}^{-1}$$

pour tout $s \in Q$. Il découle de (1) que l'entropie de l'échantillonnage aléatoire simple est donnée par

$$I(p_{\text{eas}}) = \log N! - \log n! - \log(N-n)!. \qquad (4)$$

Autrement dit, l'échantillonnage aléatoire simple sans remise est le plan à entropie maximale dans la classe des plans d'échantillonnage équiprobable avec taille fixe.

Les plans d'échantillonnage n'ont pas tous une entropie élevée. Par exemple, le plan d'échantillonnage systématique 1 sur G possède une très faible entropie. Ici, on suppose que le nombre d'échantillons, $G = N/n$, est un entier. Puisque $p_{\text{syst}}(s) = 1/G$ pour tout $s \in Q$, l'entropie de l'échantillonnage systématique est donnée par

$$I(p_{\text{syst}}) = \log N - \log n,$$

qui est beaucoup plus faible que (4), surtout pour les grandes valeurs de N.

2. Résultat principal

À la présente section, nous comparons l'entropie de l'échantillonnage de Bernoulli à celle de l'échantillonnage aléatoire simple sans remise. Puisque le support des plans d'échantillonnage de Bernoulli est beaucoup plus grand que celui des plans d'échantillonnage aléatoire simple sans remise, nous nous attendons à ce que l'entropie de l'échantillonnage de Bernoulli soit beaucoup plus élevée que celle de l'échantillonnage aléatoire simple sans remise. Le tableau 1 donne l'entropie de l'échantillonnage aléatoire simple et de l'échantillonnage de Bernoulli pour diverses valeurs de N et de π. Étonnamment, nous constatons que l'entropie des deux plans d'échantillonnage pour les mêmes probabilités d'inclusion et la même taille d'échantillon est à peu près la même. L'examen du tableau 1 montre clairement que les deux plans d'échantillonnage ont des entropies similaires, même pour des tailles de population modestes (par exemple, $N = 100$), indépendamment de la valeur de π. Ce résultat est un peu curieux, étant donné la réduction importante du nombre d'échantillons possibles lorsque l'on fixe la taille de l'échantillon. En effet, rappelons que la taille du support est $\binom{N}{n}$ pour l'échantillonnage aléatoire simple sans remise, tandis qu'elle est égale à 2^N pour l'échantillonnage de Bernoulli. Par exemple, pour $N = 100$ et $n = 20$, la taille du support pour l'échantillonnage aléatoire simple sans remise est égale à $\binom{100}{20} \approx 5{,}36 \times 10^{20}$, tandis qu'elle est égale à $2^{100} \approx 1{,}26 \times 10^{30}$ pour l'échantillonnage de Bernoulli. Autrement dit, la taille du support de l'échantillonnage de Bernoulli est environ $2{,}36 \times 10^{9}$ fois plus grande que celle du support de l'échantillonnage aléatoire simple sans remise.

Résultat 1. *Soit* $I(p_{\text{bern}})$ *et* $I(p_{eas})$ *les entropies de l'échantillonnage de Bernoulli et de l'échantillonnage aléatoire simple sans remise, respectivement, données par* (3) *et* (4). *Alors*,

$$\lim_{N\to\infty} \frac{I(p_{eas})}{I(p_{\text{bern}})} = 1.$$

Preuve. En partant de la formule de Stirling (voir Abramowitz et Stegun 1964, page 257)

$$\lim_{n\to\infty} \frac{n\log n - n}{\log n!} = 1,$$

nous obtenons

$$\lim_{\substack{N\to\infty \\ n\to\infty \\ N-n\to\infty}} \frac{N\log N - n\log n - (N-n)\log(N-n)}{\log\binom{N}{n}} = 1,$$

d'où nous obtenons

$$\lim_{N\to\infty} = \frac{\log\binom{N}{N\pi}}{-N(1-\pi)\log(1-\pi) - N\pi\log\pi} = 1.$$

3. Conclusion

Dans la présente note, nous avons montré que l'échantillonnage de Bernoulli et l'échantillonnage aléatoire simple sans remise ont des entropies fort semblables, même pour des tailles de population modestes. Nous conjecturons que nous ferions la même constatation en comparant le plan d'échantillonnage de Poisson et le plan d'échantillonnage de Poisson conditionnel pour un ensemble donné de probabilités d'inclusion de premier ordre. Toutefois, la preuve de ce résultat semble être considérablement plus complexe.

Tableau 1
Entropie des plans (échantillonnage de Bernoulli ; échantillonnage aléatoire simple)

N	$\pi = 0,1$	$\pi = 0,2$	$\pi = 0,3$	$\pi = 0,4$	$\pi = 0,5$
10	(3,3 ; 2,3)	(5 ; 3,8)	(6,1 ; 4,8)	(6,7 ; 5,3)	(6,9 ; 5,5)
100	(32,5 ; 30,5)	(50 ; 47,7)	(61,1 ; 58,6)	(67,3, 64,8)	(69,3 ; 66,8)
1 000	(325,1 ; 321,9)	(500,4 ; 496,9)	(610,9 ; 607,3)	(673, 669,4)	(693,1 ; 689,5)
10 000	(3 250,8 ; 3 246,5)	(5 004 ; 4 999,4)	(6 108,6 ; 6 103,9)	(6 730,1 ; 6 725,3)	(6 931,5 ; 6 926,6)
100 000	(32 508,3 ; 32 502,8)	(50 040,2 ; 50 034,5)	(61 086,4 ; 61 080,5)	(67 301,2 ; 67 295,2)	(69 314,7 ; 69 308,7)
1 000 000	(325 083 ; 325 076)	(500 402 ; 500 396)	(610 864 ; 610 857)	(673 012 ; 673 005)	(693 147 ; 693 140)

Remerciements

Nous remercions un rédacteur associé et un examinateur de leurs commentaires constructifs. Les travaux de David Haziza ont été financés en partie par des bourses de recherche du Conseil de recherches en sciences naturelles et en génie du Canada.

Bibliographie

Abramowitz, M., et Stegun, I.A. (1964). *Handbook of Mathematical Functions*. New York : Dover.

Brewer, K.R.W., et Donadio, M.E. (2003). La variance sous grande entropie de l'estimateur de Horvitz-Thompson. *Techniques d'enquête*, 29, 213-220.

Chen, S.X., Dempster, A.P. et Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Rapport technique, CREST-ENSAI, Rennes.

Haziza, D., Mecatti, F. et Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron*, 66, 91-108.

Henderson, T. (2006). Estimating the variance of the Horvitz-Thompson estimator. Thèse de maîtrise, School of Finance and Applied Statistics, The Australian National University.

Matei, A., et Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 4, 543-570.

Tillé, Y. (2006). *Sampling Algorithms*. New York : Springer.