REVUE DE STATISTIQUE APPLIQUÉE

J. VALEMBOIS

Les nombres de Stirling et de Bernoulli dans un problème de sondage

Revue de statistique appliquée, tome 19, n° 1 (1971), p. 87-94

<http://www.numdam.org/item?id=RSA_1971__19_1_87_0>

LES NOMBRES DE STIRLING ET DE BERNOULLI DANS UN PROBLÈME DE SONDAGE [1]

Madame VALEMBOIS

Département Mathématiques, Centre Universitaire Dauphine

1 - INTRODUCTION

Il s'agit d'estimer la moyenne d'une population de taille N finie lorsque l'on tire un échantillon de taille n fixée, avec remise. Ce problème ne se pose que lorsqu'un sondage exhaustif se révèle impossible matériellement, l'estimateur, moyenne de l'échantillon, étant alors meilleur que tout autre. L'article de P. THIONET [1] (1967) sur ce sujet est complété ici par l'utilisation des nombres de Bernoulli. Rappelons que le principe est de ne garder que les observations distinctes pour obtenir un échantillon S de taille n_d aléatoire. Il y a alors au moins deux estimations possibles pour m :

$$\bar{x}' = \frac{1}{n_d} \sum_{S_d} x_i \qquad \bar{x}'' = \frac{1}{E(n_d)} \sum_{S_d} x_j$$

Cet article concerne l'étude de $\bar{x}'$ et accessoirement celle de $\bar{x}''$.

2 - RAPPELS

2.1 - Loi de n_d

L'expression de la probabilité d'avoir n_d observations distinctes utilise l'une des propriétés des nombres de Stirling de seconde espèce.

$S(n, n_d)$ = nombre de partitions de n éléments en n_d parties distinctes et l'on a :

$$\boxed{\text{Prob}\ (n_d) = \frac{1}{N^n} S(n, n_d)\ n_d\ !\ C_N^{n_d}}$$

Des calculs simples donnent :

$$E(n_d) = N\left[1 - \left(1 - \frac{1}{N}\right)^n\right]$$

$$V(n_d) = N\left(1 - \frac{1}{N}\right)^n - N^2\left(1 - \frac{1}{N}\right)^{2n} + N\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)^n$$

---

(1) Mémoire de D.E.A., rédigé en juin 1968.

En faisant des développements limités pour N grand et n fixé, on obtient :

$$E(n_d) \# n\left(1 - \frac{n-1}{2N}\right)$$

$$V(n_d) \# \frac{n(n-1)}{2N}$$

ce qui permet de dire que la méthode conduit en général à une valeur n_d proche de n si $n^2 \ll N$.

2.2 - Espérance mathématique de $\bar{x}'$

Conditionnellement à n_d connu l'espérance mathématique de $\bar{x}'$ est :

$$E(\bar{x}'/n_d) = m$$

D'où :

$$E(\bar{x}') = \sum_{n_d=1}^{n} E(\bar{x}'/n_d)\ \text{Prob}(n_d) = m$$

Donc $\bar{x}'$ est un estimateur sans biais de m.

2.3 - Variance de $\bar{x}'$

De même

$$V(\bar{x}') = \sum_{n_d=1}^{n} V(\bar{x}'/n_d)\ \text{Prob}\ (n_d)$$

$V(\bar{x}'/n_d)$ se calcule comme dans le cas d'un tirage sans remise (il n'y a que des observations distinctes)

$$V(\bar{x}') = \sum_{n_d=1}^{n} \frac{1}{n_d} \frac{\sigma^2(N - n_d)}{N-1}\ \text{Prob}\ (n_d) = \frac{\sigma^2}{N-1}\left[N\,E\left(\frac{1}{n_d}\right) - 1\right]$$

Le calcul de $E\left(\frac{1}{n_d}\right)$ fait par P. THIONET [1] donne :

$$\boxed{E\left(\frac{1}{n_d}\right) = \frac{1}{N^n}[N^{n-1} + (N-1)^{n-1} + \ldots\ldots + 1^{n-1}]}$$

3 - LES NOMBRES DE BERNOULLI

3.1 -

C'est ici qu'interviennent les nombres de Bernoulli : la somme des puissances de N − 1 premiers entiers s'exprime sous la forme d'un polynome en N par :

$$(N-1)^{n-1} + (N-2)^{n-1} + \ldots\ldots + 1^{n-1} = \frac{1}{n}\,[B_n(N) - B_n]$$

dont la principale propriété est $B_n(N+1) - B_n(N) = n\,N^{n-1}$

$$B_n(N) = B_0 N + C_n^1 B_1 N^{n-1} + \ldots + C_n^p B_p N^{n-p} + \ldots + C_n^n B_n$$

où les B_n sont les nombres de Bernoulli [2, p. 52].

$$B_0 = 1,\ B_1 = -\frac{1}{2},\ B_2 = \frac{1}{6},\ B_4 = -\frac{1}{30},\ B_6 = \frac{1}{42},\ B_8 = -\frac{1}{30},\ B_{10} = \frac{5}{66},$$

$$B_{12} = -\frac{691}{2730} \text{ et } B_{2p+1} = 0 \text{ pour } p \geq 1$$

tous les nombres non nuls ont des signes alternés.

Il s'agit des nombres $B^{(n)}$ pour n = 1 [2, p. 287-288], leur fonction génératrice exponentielle est :

$$G(1,t) = \frac{t}{e^t - 1}$$

On a donc :

$$E\left(\frac{1}{n_d}\right) = \frac{1}{n\,N^n}\left[B_n(N) - B_n + n\,N^{n-1}\right]$$

$$= \frac{1}{N^n}\left[\frac{N^n}{n} + \frac{N^{n-1}}{2} + \frac{n-1}{12} N^{n-2} + \ldots + \frac{C_n^p B_p}{n} N^{n-p} + \ldots + B_{n-1} N\right]$$

3.2 - Variance

Mais revenons à la variance de $\bar{x}'$

$$V(\bar{x}') = \frac{\sigma^2}{N-1}\left[N\,E\left(\frac{1}{n_d}\right) - 1\right]$$

Donc V(x') est égal à :

$$V(\bar{x}') = \frac{\sigma^2}{n\,N^{n-1}(N-1)}\left[B_n(N) - B_n\right]$$

$$= \frac{\sigma^2}{(N-1)N^{n-1}}\left[\frac{N^n}{n} - \frac{N^{n-1}}{2} + \frac{n-1}{12} N^{n-2} + \ldots + \frac{C_n^p B_p}{n} N^{n-p} + \ldots + B_{n-1} N\right]$$

Pour N grand et n fixé ≥ 3, on retrouve alors l'expression approchée donnée par RAO [3] :

$$V(\bar{x}') \# \frac{\sigma^2}{N-1}\left[\frac{N}{n} - \frac{1}{2} + \frac{n-1}{12N}\right]$$

le terme suivant étant négatif on peut dire dans les mêmes conditions que :

$$V(\bar{x}') \leq \frac{\sigma^2}{N-1}\left[\frac{N}{n} - \frac{1}{2} + \frac{n-1}{12N}\right]$$

3.3 - Amélioration du résultat

En fait cette inégalité est vraie quels que soient N et $n \geq 3$:

Soit

$$B_n(N) - B_n = N - \frac{n\,N^{n-1}}{2} + \frac{n(n-1)}{12}\,N^{n-2} + R\,(N) \quad \text{pour } n \geq 3$$

et

$$\delta_n(N) = R_n(N+1) - R_n(N)$$

On a

$$\delta_n(N) = B_n(N+1) - B_n(N) - [(N+1)^n - N^n] + \frac{n}{2}[(N+1)^{n-1} - N^{n-1}]$$

$$- \frac{n(n-1)}{12}[(N+1)^{n-2} - N^{n-2}]$$

$$\delta_n(N) = n\,N^{n-1} - \sum_{j=1}^{n} C_n^j\,N^{n-j} + \frac{n}{2}\sum_{j=1}^{n-1} C_{n-1}^j\,N^{n-1-j} - \frac{n(n-1)}{12}\sum_{j=1}^{n-2} C_{n-2}^j N^{n-2-j}$$

$$\delta_n(N) = n\,N^{n-1} - n\,N^{n-1} - \frac{n(n-1)}{2}\,N^{n-2} + \frac{n(n-1)}{2}\,N^{n-2}$$

$$- \sum_{j=3}^{n}\left[C_n^j - \frac{n}{2}\,C_{n-1}^{j-1} + \frac{n(n-1)}{12}\,C_{n-2}^{j-2}\right] N^{n-j}$$

$$\delta_n(N) = -\frac{1}{12}\sum_{j=3}^{n} C_n^j\,(j-3)\,(j-4)\,N^{n-j} \leq 0$$

Donc quels que soient N et n, indépendamment de l'inégalité $N \geq n$

$$R_n(N+1) \leq R_n(N) \quad \ldots\ldots \leq R_n(0) = 0$$

ce qui démontre l'inégalité

$$\boxed{V(\bar{x}') < \frac{\sigma^2}{N-1}\left[\frac{N}{n} - \frac{1}{2} + \frac{n+1}{12N}\right]}$$

avec égalité pour n = 3 et 4

Pour n = 1 $V(\bar{x}') = \sigma^2$. Pour n = 2 $V(\bar{x}') = \frac{\sigma^2}{2}$

Avec le même principe de démonstration on obtient un encadrement de $V(\bar{x}')$

$$\frac{\sigma^2}{N-1}\left[\frac{N}{n} - \frac{1}{2}\right] \leq V(\bar{x}') \leq \frac{\sigma^2}{N-1}\left[\frac{N}{n} - \frac{1}{2} + \frac{n-1}{12N}\right]$$

Si l'on voulait un encadrement plus précis il suffirait de prendre plus de termes dans le développement polynomial de $V(\bar{x}')$. Les démonstrations

se font de la même façon jusqu'à ce que l'on prenne le septième terme non nul, mais malheureusement il n'apparaît pas de recurrence pour un nombre quelconque de termes.

L'erreur relative faite en employant l'approximation de Ro vaut :

$$E = \frac{\frac{(n-1)\,\sigma^2}{12N\,(N-1)}}{\frac{\sigma^2}{n(N-1)N^{n-1}}\,[B_n(N) - B_n]}$$

Or

$$B_n(N) - B_n \geq N^n - \frac{n\,N^{n-1}}{2}$$

Donc

$$E \leq \frac{n(n-1)}{12\left(N^2 - \frac{nN}{2}\right)} = \frac{n(n-1)}{6N(2N-n)}$$

L'approximation de Rao ne sera donc applicable que si $n \ll N$. Par contre si $n = 1/2\ N$ l'erreur relative sera de l'ordre de 1/36 ce qui n'est guère satisfaisant.

4 - COMPARAISON AVEC LES SONDAGE EXHAUSTIF ET BERNOULLIEN

4.1 - Sondage exhaustif.

En ce qui concerne le sondage exhaustif la variance de $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$ est $\frac{\sigma^2}{n}\frac{N-n}{N-1}$.

Comparons $V(\bar{x}_1)$ et $V(\bar{x}')$ pour $n \geq 2$

$$V(\bar{x}') - V(\bar{x}_1) \geq \frac{\sigma^2}{N-1}\left[\frac{N}{n} - \frac{1}{2}\right] - \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{\sigma^2}{2(N-1)} \geq 0$$

Donc on retrouve un résultat prévisible c'est-à-dire que l'estimateur du sondage exhaustif est meilleur que $\bar{x}'$.

L'efficacité est minorée par

$$\frac{V(x')}{V(x_1)} > \frac{\frac{\sigma^2}{n(N-1)N^{n-1}}\left(N^n - \frac{n}{2}\,N^{n-1}\right)}{\frac{\sigma^2}{n}\frac{N-n}{n-1}} = \frac{N - \frac{n}{2}}{N-n} = 1 + \frac{n}{2(N-n)}$$

4.2 - Sondage bernoullien

Comparons maintenant $\bar{x}'$ et $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ d'un sondage bernoullien

$$V(\bar{x}) - V(\bar{x}') \geq \frac{\sigma^2}{n} - \frac{\sigma^2}{n(N-1)}\left[N - \frac{n}{2} + \frac{n(n-1)}{12N}\right] \quad \text{pour } n \geq 3$$

$$V(\bar{x}) - V(\bar{x}') > \frac{\sigma^2}{n(N-1)}\left[\frac{n-1}{2} - \frac{n(n-1)}{12N}\right]$$

Cette expression est positive donc $\bar{x}'$ est meilleur que l'estimateur du sondage bernoullien quels que soient n et N (Pour n = 1 et n = 2 les variances sont égales).

L'efficacité de $\bar{x}'$ par rapport à $\bar{x}$ est minorée par :

$$\frac{V(\bar{x})}{V(\bar{x}')} \geq \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n(N-1)}\left[N - \frac{n}{2} + \frac{n(n-1)}{12N}\right]} = \frac{N-1}{N - \frac{n}{2} + \frac{n(n-1)}{12N}}$$

donc

$$\frac{V(\bar{x})}{V(\bar{x}')} \geq 1 + \frac{\frac{n-2}{2} - \frac{n(n-1)}{12N}}{N - \frac{n}{2} + \frac{n(n-1)}{12N}}$$

Dans le cas où l'approximation de Rao s'applique c'est-à-dire si $n \ll N$, on a :

$$\frac{V(\bar{x}')}{V(\bar{x}_1)} \# 1 + \frac{n}{2(N-n)} \# 1 + \frac{n}{2N}$$

et

$$\left(\frac{V(\bar{x})}{V(\bar{x}')} \# 1 + \frac{n}{2N}\right)$$

Le gain en efficacité de $\bar{x}'$ par rapport à $\bar{x}$ est donc la moitié du gain en efficacité de $\bar{x}_1$ par rapport à $\bar{x}$ lorsque $n \ll N$.

5 - ETUDE DE $\bar{x}'' = \frac{1}{E(n_d)} \sum_{s_d} x_i$

Rappelons que $E(n_d) = N\left[1 - \left(1 - \frac{1}{N}\right)^n\right]$

Nous n'approfondirons pas l'étude de cet estimateur car son expression nécessite la connaissance précise de la taille de la population.

Cet estimateur est sans biais

$$E(\bar{x}'') = \sum_{n_d=1}^{n} E(\bar{x}''/n_d) \text{ Prob } n_d = \sum_{n_d=1}^{n} \frac{n_d m}{E(x_d)} \text{ Prob}(n_d) = m$$

Sa variance est :

$$V(\bar{x}'') = \frac{1}{[E(n_d)]} \sum_{n_d=1}^{n} n_d \, \sigma^2 \, \frac{N - n_d}{N - 1} \, \text{Prob}(n_d)$$

$$V(\bar{x}'') = \frac{1}{[E(n_d)]} \; \frac{\sigma^2}{N - 1} \, [N \, E(n_d) - [E(n_d)]^2]$$

Comme

$$V(n_d) = N\left(1 - \frac{1}{N}\right)^n - N^2\left(1 - \frac{1}{N}\right)^{2n} + N\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)^n$$

Après calculs il vient :

$$V(\bar{x}'') = \frac{\sigma^2}{N} \; \frac{\left(1 - \frac{1}{N}\right)^n - \left(1 - \frac{2}{N}\right)^n}{\left[1 - \left(1 - \frac{1}{N}\right)^n\right]^2}$$

Pour n fixé et N grand un développement limité donne :

$$V(\bar{x}'') \# \frac{\sigma}{n} \left[1 - \frac{n - 1}{2N} + \frac{(n - 1)\,(n - 11)}{12N^2}\right]$$

Donc pour $n \ll N$. $\quad V(\bar{x}'') \leqslant \frac{\sigma^2}{n}\left[1 - \frac{n - 1}{2N}\right]$

On voit immédiatement que $\bar{x}''$ est un meilleur estimateur que $\bar{x}$ pour $n \ll N$.

D'autre part on retrouve que $\bar{x}''$ est moins bon que l'estimateur du sondage exhaustif quels que soient n et N.

$$V(\bar{x}'') = \frac{1}{[E(n_d)]^2} \sum_{n_d=1}^{n} n_d \, \sigma^2 \, \frac{N - n_d}{N - 1} \, \text{Prob}(n_d) \geq \frac{\sigma^2}{n} \, \frac{N - n}{N - 1} \, \frac{E(n_d^2)}{[E(n_d)]^2}$$

$$V(\bar{x}'') \geq \frac{\sigma^2}{n} \, \frac{N - n}{N - 1}\left[1 + \frac{V(n_d)}{[E(n_d)]^2}\right] \geq \frac{\sigma^2}{n} \, \frac{N - n}{N - 1}$$

Comparons maintenant $\bar{x}'$ et $\bar{x}''$

$$V(\bar{x}') - V(\bar{x}'') \# \frac{\sigma^2}{N - 1}\left[\frac{N}{n} - \frac{1}{2} + \frac{n - 1}{12N}\right] - \frac{\sigma^2}{n}\left[1 - \frac{n - 1}{2N}\right]$$

$$\# \frac{\sigma^2}{N - 1}\left[\frac{1}{2n} + \frac{(n - 1)\,(n - 6)}{12nN}\right]$$

Donc pour des valeurs de N suffisamment grandes par rapport à n, $\bar{x}''$ est meilleur que $\bar{x}'$.

6 - CONCLUSION

Dans un sondage bernoullien le fait de ne retenir que les observations distinctes améliore l'estimation de m à condition de remplacer n par le nom-

bre des observations distinctes ou par l'espérance mathématique de ce nombre. Nous avons démontré que $\bar{x}' = \frac{1}{n_d} \sum_{s_d} x_i$ est meilleur ou équivalent à $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i$ quels que soient n et N et que d'autre part $\bar{x}'' = \frac{1}{E(n_d)} \sum_{s_d} x_i$ est meilleur que $\bar{x}'$ pour des valeurs suffisamment grandes de N par rapport à n. Et nous avons vérifié que ces deux estimateurs ne peuvent en aucun cas atteindre la précision de l'estimateur d'un sondage exhaustif.

En l'état actuel de la statistique, ces résultats ne sont plus que des cas particuliers d'une théorie générale (des estimations "suffisantes") ; mais ils ont eu une réelle importance historique en aidant à découvrir que $\bar{x}$ n'était pas le meilleur estimateur possible dans le contexte non paramétrique (absence de loi théorique de distribution pour les x_i).

REFERENCES

[1] THIONET P. - Application des nombres de Stirling de deuxième espèce à un problème de sondage (1967)
Revue de Statistique Appliquée

[2] DAVID F.N., BARTON D.E. - Combinatorial Chance (1962, p. 294-295)

[3] RAO J.N.K. - On the comparison of sampling with and without replacement
Revue de l'I.I.S. (1966) 34, 2, p. 125-138

[4] RIORDAN J. - An Introduction to Combinatorial Analysis (1958, p. 32, 42 à 45, 75, 91 à 99)

[5] BECHENBACH - Applied Combinatorial Mathematics (1964, p. 67 à 72, 82 à 88)