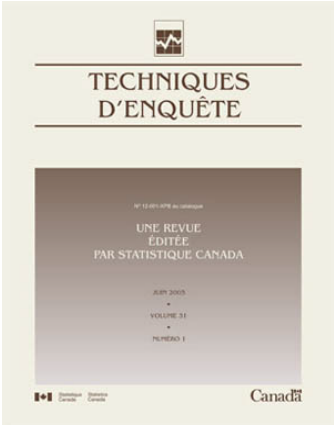
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Échantillonnage inverse à probabilités inégales



par Yves Tillé

Date de diffusion : le 20 décembre 2016

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, décembre 2016
Vol. 42, nº 2, pp. 301-313
Statistique Canada, nº 12-001-X au catalogue

Échantillonnage inverse à probabilités inégales

Yves Tillé[1]

Résumé

Dans le cadre d'une enquête économique auprès d'un échantillon d'entreprises, on sélectionne au hasard des professions dans une liste jusqu'à ce que l'on identifie un nombre r de professions présentes dans une unité locale. Il s'agit d'un problème d'échantillonnage inverse pour lequel nous proposons quelques solutions. Les plans simples avec et sans remise se traitent au moyen des distributions binomiale négative et hypergéométrique négative. On propose également des estimateurs pour le cas où les unités sont sélectionnées à probabilités inégales avec ou sans remise.

Mots-clés : Emplacement; estimateur de Horvitz-Thompson; binomiale négative; hypergéométrique négative; plan inverse; probabilité d'inclusion; salaire.

1 Le problème

Le problème est apparu dans le cadre d'une question posée concernant la nouvelle « Enquête sur les postes vacants et les salaires » de Statistique Canada. Cette enquête comporte un volet « salaires » et un volet « postes vacants ». Dans le cadre de la partie « salaires », on s'intéresse aux salaires moyens, aux salaires minimums, aux salaires maximums, et aux salaires d'entrée pour différentes professions.

L'objectif est de fournir des statistiques sur les salaires au niveau des régions économiques (qui sont une subdivision des provinces). Au premier degré, on sélectionne un échantillon de 100 000 emplacements d'entreprises (appelées aussi unités locales d'entreprises) selon un plan de Poisson stratifié par secteurs d'activité et régions économiques.

Dans ce qui suit, par simplicité, on utilisera le terme « entreprise » à la place d'« emplacement » tout en gardant à l'esprit qu'un emplacement est selon Statistique Canada « une unité de production située en un point géographique précis, où se fait l'activité économique ou à partir duquel elle s'exerce, et pour lequel il est possible d'obtenir, au minimum, des données en matière d'emploi. »

Dans un souci de gestion du fardeau de réponse, on ne peut pas inventorier toutes les professions pour chaque entreprise. Il a donc été envisagé de proposer une liste de professions et de demander si ces professions sont présentes dans une entreprise. Les professions peuvent ainsi être tirées aléatoirement dans la liste et proposées successivement au responsable de l'entreprise jusqu'à l'obtention de r professions. Comme on s'intéresse plus spécifiquement aux professions les plus courantes, il est utile de considérer les cas où les professions sont sélectionnées à probabilités inégales dans la liste proportionnellement à leur prévalence dans la population totale. Notons que cette méthode n'a pas été retenue pour l'Enquête sur les postes vacants et les salaires de Statistiques Canada qui propose plutôt aux entreprises enquêtées une liste de professions de taille fixe. Cela dit, les caractéristiques théoriques de la méthode demeurent d'intérêt.

Nous appelons échantillonnage inverse un schéma dans lequel on sélectionne des unités successivement jusqu'à l'obtention d'un nombre fixé par avance d'unités présentant une certaine caractéristique. Il ne faut

1. Yves Tillé, Institut de Statistique, Université de Neuchâtel, Rue de Bellevaux, 51, 2000 Neuchâtel, Suisse. Courriel : yves.tille@unine.ch.

pas confondre l'échantillonnage inverse avec l'échantillonnage réjectif. Dans un échantillonnage réjectif, on sélectionne un échantillon selon un plan et on rejette cet échantillon s'il n'a pas la caractéristique désirée (par exemple une taille fixe, où une moyenne égale à celle de la population). On réitère la sélection d'échantillons jusqu'à l'obtention d'un échantillon ayant la propriété désirée.

L'échantillonnage inverse pose un certain nombre de questions théoriques. Comment un tel plan peut-il être mis en oeuvre avec des probabilités d'inclusion égales ou inégales. Quelle est la probabilité d'inclusion d'une profession à l'intérieur de chaque entreprise ? Comment estimer une variable d'intérêt au moyen de l'échantillon constitué de quelques entreprises et de quelques professions au sein d'elles ? Comment estimer le nombre de professions présentes dans l'entreprise ? Plus généralement, comment peut-on mettre en oeuvre cette enquête et ensuite procéder à une estimation ?

La question essentielle est la manière de sélectionner ces professions. On peut les sélectionner selon un plan simple avec ou sans remise ou à probabilités inégales. Une option consisterait à sélectionner les unités à probabilités inégales au moyen de la méthode d'échantillonnage de Poisson séquentielle proposée par Ohlsson (1998) ou le sondage de Pareto proposé par Rosén (1997). Le problème de l'échantillonnage inverse a déjà été abordé par Murthy (1957); Sampford (1962); Pathak (1964); Chikkagoudar (1966, 1969); Salehi et Seber (2001). Cependant, le paramètre à estimer est ici particulier, puisque l'on veut estimer des moyennes de revenus entre toutes les entreprises possédant une profession particulière. Nous proposons également un nouveau plan inverse sans remise à probabilités inégales.

Cet article est organisé comme suit : Dans la section 2, le problème est posé et la notation est définie. On traite ensuite le cas à probabilités égales avec remise à la section 3 et sans remise à la section 4. Le cas avec remise et à probabilités inégales est développé à la section 5. On présente enfin une nouvelle méthode de tirage pour le cas sans remise à probabilités inégales à la section 6. Enfin, la section 7 contient une petite discussion.

2 Formalisation du problème

On utilise la notation suivante :

- U : la population de N entreprises, c'est-à-dire $U = \{1,\ldots,i,\ldots,N\}$ (U peut désigner la population d'entreprises dans une région économique),
- L : la liste de professions,
- M : le nombre de professions dans la liste, c'est-à-dire la taille de L,
- F_i : la liste des professions présentes dans l'entreprise i, avec $F_i \subset L$,
- D_i : la liste des professions absentes dans l'entreprise i, avec $D_i \subset L$, $F_i \cup D_i = L$ et $D_i \cap F_i = \varnothing$,
- Mp_i : le nombre de professions dans l'entreprise i, c'est-à-dire la taille de F_i,
- r : le nombre de professions distinctes que l'on veut obtenir dans chaque entreprise,
- X_i : le nombre d'échecs obtenus avant d'obtenir les r professions dans l'entreprise i en sélectionnant les professions selon un plan particulier.

L'objectif principal est d'estimer le salaire moyen pour une profession dans la population entière. Soient y_{ik} le salaire moyen dans la profession k au sein de l'entreprise i et z_{ik} le nombre d'employés ayant la profession k au sein de l'entreprise i. L'objectif est d'estimer le salaire moyen de la profession k donné par

$$\bar{Y}_k = \frac{\sum_{i \in U \mid F_i \ni k} z_{ik} y_{ik}}{\sum_{i \in U \mid F_i \ni k} z_{ik}}.$$

Supposons qu'un échantillon d'entreprises S_1 soit sélectionné dans U au moyen d'un plan donné quelconque avec des probabilités d'inclusion π_{1i}. Dans l'entreprise i, on sélectionne un échantillon S_i de professions au moyen d'un des plans décrits ci-dessous avec la probabilité d'inclusion $\pi_{k|i}$. Si le plan est avec remise, $\pi_{k|i}$ représente l'espérance de nombre de fois que la profession k est sélectionnée dans l'entreprise i.

On peut estimer $\bar{Y}_k$ au moyen d'un estimateur de type « quotient » (Hájek 1971) :

$$\hat{\bar{Y}}_k = \frac{\sum_{i \in S_1 \mid (S_i \cap F_i) \ni k} \frac{z_{ik} y_{ik}}{\pi_{1i} \pi_{k|i}}}{\sum_{i \in S_1 \mid (S_i \cap F_i) \ni k} \frac{z_{ik}}{\pi_{1i} \pi_{k|i}}}.$$

Il est donc nécessaire de connaître la probabilité qu'une profession soit sélectionnée au sein d'une entreprise. Cependant, avec un plan de type inverse, cette probabilité n'est pas connue et devra donc être estimée pour pouvoir procéder à une estimation de $\bar{Y}_k$. Comme les probabilités d'inclusion apparaissent au dénominateur, il est préférable d'estimer les inverses des $\pi_{k|i}$. Au sein d'une entreprise, une profession a d'autant moins de chance d'être sélectionnée que le nombre de professions qu'elle comporte est élevé. De plus, cette probabilité dépend du plan de sondage inverse utilisé au sein de chacune des entreprises.

3 Cas du tirage simple avec remise

Supposons que l'entreprise i ait une proportion p_i de professions de la liste présentes dans l'entreprise. Si l'échantillon de professions est tiré avec remise au sein de l'entreprise i jusqu'à l'identification de r professions présentes dans l'entreprise, alors X_i a une distribution binomiale négative notée $X_i \sim NB(r, p_i)$. Dans ce cas,

$$\Pr(X_i = x_i) = \binom{r + x_i - 1}{x_i} p_i^r (1 - p_i)^{x_i},$$

avec $x_i \in \mathbb{N} = \{0,1,2,3,\ldots\}$, $p_i \in [0,1]$, $r \in \mathbb{N}^* = \{1,2,3,\ldots\}$. De plus,

$$\mathrm{E}(X_i) = \frac{r(1-p_i)}{p_i} \quad \text{et} \quad \mathrm{var}(X_i) = \frac{r(1-p_i)}{p_i^2}.$$

Soit $A_{ik}, k \in L,$ le nombre de fois que l'unité k est sélectionnée dans l'échantillon prélevé dans l'entreprise i. Dans un plan simple avec remise de taille $n,$ les A_{ik} ont une distribution multinomiale. Donc

$$\Pr(A_{ik} = a_{ik}, k \in L) = \frac{n!}{M^n} \prod_{k \in L} \frac{1}{a_{ik}!},$$

où $A_{ik} = 0, \ldots, n,$ et

$$\sum_{k \in L} a_{ik} = n.$$

Si on conditionne ce vecteur multinomial sur une taille fixe dans une partie donnée de la population, on obtient

$$\begin{aligned}
\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k \in F_i} A_{ik} = r\right) &= \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ et } \sum_{k \in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k \in F_i} A_{ik} = r\right)} \\
&= \frac{\dfrac{n!(1-p_i)^{(n-r)}}{(n-r)!M^r} \displaystyle\prod_{k \in F_i} \frac{1}{a_{ik}!}}{\dfrac{n! p_i^r (1-p_i)^{n-r}}{r!(n-r)!}} \\
&= r! \left(\frac{1}{Mp_i}\right)^r \prod_{k \in F_i} \frac{1}{a_{ik}!},
\end{aligned}$$

avec

$$\sum_{k \in F_i} a_{ik} = r.$$

Ce résultat montre que si l'on conditionne la somme des A_{ik} sur une partie de la population, la distribution reste multinomiale et conditionnellement on a encore un plan simple avec remise.

Dans la procédure où l'on tire avec remise jusqu'à obtenir r professions présentes au sein de l'entreprise $i,$ on a donc

$$\mathrm{E}(A_{ik} \,|\, X_i) = \begin{cases} \dfrac{r}{Mp_i} & \text{si } k \in F_i \\ \dfrac{X_i}{M - Mp_i} & \text{si } k \in D_i. \end{cases}$$

En effet, conditionnellement à $X_i,$ dans F_i de taille $Mp_i,$ on sélectionne r professions et, dans D_i de taille $M(1-p_i),$ on sélectionne X_i professions.

Dans le cas avec remise, on ne calcule pas vraiment une probabilité d'inclusion, mais l'espérance de A_{ik} que l'on note $\pi_{k|i}$,

$$\pi_{k|i} = \mathrm{EE}(A_{ik} \mid X_i) = \frac{r}{Mp_i},$$

$k \in L$. Le problème est que l'on connaît M, r et X_i, mais pas p_i. On peut estimer p_i par la méthode des moments en résolvant $\mathrm{E}(X_i) = X_i$, ce qui donne

$$X_i = \frac{r(1-\hat{p}_i)}{\hat{p}_i}$$

et donc

$$\hat{p}_{i1} = \frac{r}{X_i + r}.$$

La méthode du maximum de vraisemblance fournit le même estimateur que la méthode des moments, mais cet estimateur est biaisé (Mikulski et Smith 1976; Johnson, Kemp et Kotz 2005, page 222). Si $r \geq 2$, l'estimateur sans biais à variance minimale de p_i est

$$\hat{p}_{i2} = \frac{r-1}{X_i + r - 1}.$$

Cependant $1/\hat{p}_{i1}$ est sans biais pour $1/p_i$.

Comme on utilise des poids qui sont des inverses des $\pi_{k|i}$, on estime alors les inverses des $\pi_{k|i}$ par :

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{M\hat{p}_{i2}}{r} = \dfrac{M(r-1)}{r(X_i + r - 1)} & \text{si } k \in F_i \\ \dfrac{M(1-\hat{p}_{i2})}{X_i} = \dfrac{M}{X_i + r - 1} & \text{si } k \in D_i. \end{cases}$$

Le cas avec remise n'est cependant pas très satisfaisant, car en sélectionnant r professions présentes avec remise, on n'obtient pas nécessairement r professions distinctes puisqu'on peut sélectionner plusieurs fois la même profession. De plus, l'échantillonnage peut s'avérer particulièrement long si Mp_i est petit. On privilégiera donc l'échantillonnage sans remise.

4 Cas du tirage simple sans remise

Pour le cas sans remise, on utilise la même notation que pour le tirage avec remise. Le nombre d'échecs X_i a alors une distribution hypergéométrique négative. Cette distribution de probabilité est un peu méconnue à tel point qu'elle a été présentée comme une distribution « oubliée » par Miller et Fridell (2007). Cette distribution est le pendant de la binomiale négative pour le tirage sans remise. Le cadre général est le suivant : On considère une population de taille M dans laquelle se trouve Mp_i unités favorables, à savoir

les professions de la liste qui sont présentes au sein de l'entreprise. Si les tirages sont réalisés à probabilités égales sans remise jusqu'à ce que r unités favorables apparaissent, alors la variable hypergéométrique négative, $X_i \sim NH(M, r, Mp_i)$, compte le nombre d'échecs avant d'obtenir r évènements favorables.

La distribution de probabilité vaut :

$$\Pr(X_i = x) = p(x; M, r, Mp_i) = \frac{\binom{x+r-1}{x}\binom{M-x-r}{Mp_i-r}}{\binom{M}{Mp_i}},$$

où $x \in \{0, \ldots, M(1-p_i)\}$, $M \in \{1, 2, \ldots\}$, $Mp_i \in \{1, 2, \ldots, M\}$, et $r \in \{1, 2, \ldots, Mp_i\}$.

$$\mathrm{E}(X_i) = \frac{Mr(1-p_i)}{Mp_i+1}, \mathrm{var}(X_i) = \frac{rM(1-p_i)(M+1)(Mp_i-r+1)}{(Mp_i+1)^2(Mp_i+2)}.$$

De nouveau, on peut noter A_{ik} le nombre de fois que l'unité k est sélectionnée dans l'échantillon. Maintenant A_{ik} ne peut prendre que les valeurs 0 et 1. Si on sélectionne n unités au moyen d'un plan simple sans remise dans L, le plan de sondage est défini par

$$\Pr(A_{ik} = a_{ik}, k \in L) = \binom{M}{n}^{-1},$$

où $a_{ik} \in \{0,1\}$, et

$$\sum_{k \in L} a_{ik} = n.$$

Si on conditionne le vecteur des A_{ik} sur une taille fixe dans une partie de la population, on obtient

$$\begin{aligned}
\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k \in F_i} A_{ik} = r\right) &= \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ et } \sum_{k \in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k \in F_i} A_{ik} = r\right)} \\
&= \left[\frac{\binom{Mp_i}{r}\binom{M-Mp_i}{n-r}}{\binom{M}{n}}\right]^{-1} \sum_{\substack{k \in D_i \\ \sum_{k \in F_i} A_{ik} = n-r \\ A_{ik} \in \{0,1\}}} \frac{1}{\binom{M}{n}} \\
&= \left[\frac{\binom{Mp_i}{r}\binom{M-Mp_i}{n-r}}{\binom{M}{n}}\right]^{-1} \frac{\binom{M-Mp_i}{n-r}}{\binom{M}{n}} \\
&= \binom{Mp_i}{r}^{-1},
\end{aligned}$$

avec

$$\sum_{k \in F_i} a_{ik} = r.$$

Ce résultat montre que si l'on conditionne la somme des A_{ik} sur une partie de la population, on a encore un plan simple sans remise. Dans la procédure où l'on tire sans remise jusqu'à obtenir r professions présentes au sein de l'entreprise i, on a donc

$$\mathrm{E}\left(A_{ik} \mid X_i\right) = \begin{cases} \dfrac{r}{Mp_i} & \text{si } k \in F_i \\ \dfrac{X_i}{M - Mp_i} & \text{si } k \in D_i. \end{cases}$$

La probabilité d'inclusion est donc

$$\pi_{k|i} = \mathrm{EE}\left(A_{ik} \mid X_i\right) = \begin{cases} \dfrac{r}{Mp_i} & \text{si } k \in F_i \\ \dfrac{\mathrm{E}\left(X_i\right)}{M - Mp_i} = \dfrac{r}{Mp_i + 1} & \text{si } k \in D_i, \end{cases}$$

pour tout $k \in L$. À nouveau, le problème est que l'on connaît M, r et X_i, mais pas p_i. On peut estimer p_i par la méthode du maximum de vraisemblance au moyen d'une méthode numérique.

Par la méthode des moments, on peut avoir une estimation en résolvant en p_i l'équation : $X_i = \mathrm{E}\left(X_i\right)$ c'est-à-dire,

$$X_i = \frac{Mr(1 - \hat{p}_i)}{M\hat{p}_i + 1}.$$

D'où

$$\hat{p}_{i1} = \frac{Mr - X_i}{M\left(r + X_i\right)}.$$

On vérifie cependant en quelques lignes que, si $r \geq 2$,

$$\hat{p}_{i2} = \frac{r - 1}{r + X_i - 1}$$

est sans biais pour p_i.

À nouveau, comme on utilise des poids qui sont des inverses des $\pi_{k|i}$. On estime alors les inverses des probabilités d'inclusion par :

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{M\hat{p}_{i2}}{r} = \dfrac{M\left(r-1\right)}{r\left(X_i + r - 1\right)} & \text{si } k \in F_i \\ \dfrac{M\left(1 - \hat{p}_{i2}\right)}{X_i} = \dfrac{M}{X_i + r - 1} & \text{si } k \in D_i. \end{cases}$$

Ces poids sont également utilisés dans l'estimateur de Murthy (1957) qui est sans biais (voir aussi Salehi et Seber 2001). Si $Mp_i < r,$ toutes les professions seront sélectionnées dans l'entreprise i et les probabilités d'inclusion estimées sont alors égales à 1.

5 Tirage à probabilités inégales avec remise

Le tirage à probabilités inégales n'est pas vraiment plus difficile à traiter quand le tirage est avec remise. Notons maintenant p_{ik} la probabilité de tirage d'une profession à chaque tirage avec

$$\sum_{k\in L} p_{ik} = 1.$$

On note P_i la somme des p_{ik} restreintes aux professions présentes au sein de l'entreprise i :

$$P_i = \sum_{k\in F_i} p_{ik}.$$

Dans ce cas, X_i a une distribution binomiale négative avec les paramètres r et P_i. Donc,

$$\mathrm{E}(X_i) = \frac{r(1-P_i)}{P_i} \quad \text{et} \quad \mathrm{var}(X_i) = \frac{r(1-P_i)}{P_i}.$$

Soit $A_{ik}, k \in L$ le nombre de fois que l'unité k est sélectionnée dans l'échantillon. Dans un plan à probabilités inégales avec remise de taille $n,$ les A_{ik} ont une distribution multinomiale. Donc

$$\Pr(A_{ik} = a_{ik}, k \in L) = n!\prod_{k\in L}\frac{p_{ik}^{a_{ik}}}{a_{ik}!},$$

où $A_{ik} = 0,\ldots,n,$ et

$$\sum_{k\in L} a_{ik} = n.$$

Si on conditionne ce vecteur multinomial sur une taille fixe dans une partie de la population, on obtient

$$\begin{aligned}\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k\in F_i} A_{ik} = r\right) &= \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ et } \sum_{k\in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k\in F_i} A_{ik} = r\right)} \\ &= \frac{\dfrac{n!(1-P_i)^{(n-r)}}{(n-r)!}\displaystyle\prod_{k\in F_i}\frac{p_{ik}^{a_{ik}}}{a_{ik}!}}{\dfrac{n!P_i^r(1-P_i)^{n-r}}{r!(n-r)!}} \\ &= r!\prod_{k\in F_i}\left(\frac{p_{ik}}{P_i}\right)^{a_{ik}}\frac{1}{a_{ik}!},\end{aligned}$$

avec

$$\sum_{k \in F_i} a_{ik} = r.$$

Ce résultat montre que si l'on conditionne la somme des A_{ik} sur une partie de la population, la distribution reste multinomiale et conditionnellement on a encore un plan à probabilités inégales avec remise.

Avec la procédure selon laquelle on tire avec remise jusqu'à obtenir r professions présentes au sein de l'entreprise i, on a donc

$$\mathrm{E}(A_{ik} \mid X_i) = \begin{cases} \dfrac{rp_{ik}}{P_i} & \text{si } k \in F_i \\ \dfrac{X_i p_{ik}}{1-P_i} & \text{si } k \in D_i. \end{cases}$$

L'espérance de A_{ik} vaut

$$\pi_{k|i} = \mathrm{EE}(A_{ik} \mid X_i) = \frac{rp_{ik}}{P_i},$$

$k \in L$. Le problème est que l'on connaît p_{ik}, r et X_i, mais pas P_i. On peut estimer P_i par la méthode des moments en résolvant $\mathrm{E}(X_i) = X_i$, ce qui donne

$$X_i = \frac{r\left(1-\hat{P}_i\right)}{\hat{P}_i}$$

et donc

$$\hat{P}_{i1} = \frac{r}{X_i + r}.$$

La méthode du maximum de vraisemblance fournit le même estimateur que la méthode des moments, mais cet estimateur est biaisé (Mikulski et Smith 1976; Johnson et coll. 2005, page 222). En effet, l'estimateur sans biais à variance minimale est

$$\hat{P}_{i2} = \frac{r-1}{X_i + r - 1}.$$

Cependant $1/\hat{P}_{i1}$ est sans biais pour P_i.

À nouveau, comme on utilise des poids qui sont des inverses des $\pi_{k|i}$. On estime alors les inverses des $\pi_{k|i}$ par :

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{\hat{P}_{i2}}{rp_{ik}} = \dfrac{r-1}{(X_i + r - 1)rp_{ik}} & \text{si } k \in F_i \\ \dfrac{1-\hat{P}_{i2}}{X_i p_{ik}} = \dfrac{1}{(X_i + r - 1)p_{ik}} & \text{si } k \in D_i. \end{cases} \tag{5.1}$$

6 Tirage à probabilités inégales sans remise

6.1 Tirage séquentiel sans remise

Pour le tirage sans remise, le premier problème est la définition du plan. Une option est d'utiliser la méthode d'Ohlsson (1995) appelée échantillonnage de Poisson séquentiel. Cette méthode consiste à générer M variable aléatoires uniformes dans l'intervalle $[0,1]$ notée u_{ik}. Ensuite on choisit les n unités correspondant aux plus petites valeurs de $u_{ik}/\pi_{k|i}$. Cette méthode a l'avantage d'être utilisable pour toute taille d'échantillon et de fournir une suite d'échantillons qui sont inclus l'un dans l'autre. Malheureusement, elle ne vérifie qu'approximativement les probabilités d'inclusion fixées. Les approximations sont cependant très précises selon les simulations données dans Ohlsson (1995).

Des méthodes ont été également proposées par Sampford (1962) et Pathak (1964). Nous proposons une solution exacte à ce problème au sens où les probabilités d'inclusion sont exactement vérifiées. On commence par calculer les probabilités d'inclusion pour un plan de taille fixe n avec des probabilités d'inclusion proportionnelles à une variable auxiliaire strictement positive $b_k, k \in L$. Les probabilités sont déterminées par

$$\pi_{k|i}(n) = \min\left(1, C_n \frac{b_k}{\sum_{\ell \in L} b_\ell}\right),$$

où C_n est déterminé de sorte que

$$\sum_{k \in L} \pi_{k|i}(n) = \sum_{k \in L} \min\left(1, C_n \frac{b_k}{\sum_{\ell \in L} b_\ell}\right) = n.$$

Un algorithme simple pour calculer ces probabilités est décrit entre autres dans Tillé (2006, page 19). Ces probabilités peuvent être calculées simplement au moyen de la fonction `inclusionprobabilities` du package R sampling.

Une méthode de tirage séquentielle doit donc sélectionner un échantillon de taille n avec des probabilités d'inclusion $\pi_{k|i}(n)$. Ensuite, elle doit permettre de passer de la taille n à la taille $n+1$ en sélectionnant simplement une unité supplémentaire de manière à ce que l'échantillon complété ait bien une probabilité d'inclusion $\pi_{k|i}(n+1)$. Il semble que la seule méthode permettant de réaliser cela est la méthode éliminatoire (Tillé 1996). La méthode éliminatoire part de la population complète (la liste des professions) et élimine une unité à chaque étape. À l'étape $j = 1, \ldots, N$, l'unité est éliminée parmi les unités restantes avec la probabilité

$$1 - \frac{\pi_{k|i}(N-j)}{\pi_{k|i}(N-j+1)}.$$

Cette méthode permet ainsi de créer une suite d'échantillons inclus l'un dans l'autre qui vérifient les probabilités d'inclusion relatifs à leur taille.

Il suffit donc d'appliquer la méthode éliminatoire pour la taille d'échantillon $n = 1$ afin que l'algorithme élimine toutes les unités successivement. En les prenant dans l'ordre inverse des éliminations, on obtient

une suite d'unités. Les n premières unités de cette suite sont bien sélectionnées avec les probabilités d'inclusion $\pi_{k|i}(n)$. L'Annexe contient une fonction en langage R qui permet de générer cette suite. Ce code est soumis à une simulation qui montre que les probabilités obtenues par simulations en appliquant cette fonction sont bien égales aux probabilités d'inclusions fixées pour toutes les tailles d'échantillon.

6.2 Plan inverse ou négatif à probabilités inégales

Maintenant que le plan est bien défini, on peut définir le plan inverse. On prend les unités dans la liste de professions au moyen de la méthode éliminatoire jusqu'à ce que r professions présentes dans l'entreprise soient sélectionnées. Dans ce cas, la distribution de probabilité du nombre d'échecs X_i semble impossible à calculer. Le calcul de la probabilité d'inclusion conditionnelle $\mathrm{E}(A_{ik} \mid X_i)$ est également problématique.

On peut cependant procéder par analogie et estimer les probabilités d'inclusion en se basant sur l'expression (5.1) développée pour le cas avec remise où l'on remplace simplement p_{ik} par

$$\frac{\pi_{k|i}(r+X_i)}{r+X_i}.$$

On obtient alors

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{(r-1)(r+X_i)}{r(X_i+r-1)\pi_{k|i}(r+X_i)} & \text{si } k \in F_i \\ \dfrac{r+X_i}{(X_i+r-1)\pi_{k|i}(r+X_i)} & \text{si } k \in D_i. \end{cases}$$

7 Discussion

Le problème du tirage peut donc être résolu pour tous les cas : avec ou sans remise et à probabilités égales ou inégales. La solution proposée sur la base de la méthode éliminatoire respecte exactement les probabilités d'inclusion, ce qui n'est pas le cas du sondage séquentiel de Ohlsson. La mise en oeuvre est particulièrement simple, car le programme donne une séquence de professions à proposer dans l'ordre jusqu'à ce que l'objectif fixé soit atteint.

La question de l'estimation est un peu plus délicate. Pour le cas sans remise avec des probabilités inégales, on doit se contenter d'une solution heuristique. On voit également qu'au deuxième degré, on a tendance à avoir des probabilités d'inclusion plus faibles dans les entreprises qui contiennent beaucoup de professions. Ceci devrait nous conduire à sélectionner avec des plus grandes probabilités les entreprises qui pourraient avoir un plus grand nombre de professions afin de ne pas sélectionner des professions avec des probabilités trop inégales.

Remerciements

L'auteur tient à remercier Pierre Lavallée de lui avoir soumis ce problème intéressant et d'avoir émis des commentaires judicieux sur une version antérieure de cet article. L'auteur remercie également Audrey-Anne Vallée pour sa relecture méticuleuse, un arbitre et un rédacteur de *Techniques d'enquête* pour leurs remarques pertinentes qui ont permis d'améliorer cet article.

Annexe

```
#
# Chargement du package sampling qui contient la fonction inclusionprobabilities().
#
library(sampling)
#
# La fonction retourne un vecteur avec les numéros d'ordre des éliminations.
# La dernière (resp. première) unité éliminée est la première (resp. dernière)
# composante du vecteur.
# La fonction donne donc les numéros des unités à présenter
# successivement pour le tirage inverse.
# L'argument x est le vecteur des valeurs prises par la variable auxiliaire utilisée pour calculer
# les probabilités d'inclusion.
#
elimination<-function(x)
       {
       pikb=x/sum(x)
       M = length(pikb)
       n = sum(pikb)
       sb = rep(1, M)
       b = rep(1, M)
       res=rep(0, M)
       for (i in 1:(M)) {
              a = inclusionprobabilities(pikb, M - i)
              v = 1 - a/b
              b = a
              p = v * sb
              p = cumsum(p)
              u = runif(1)
              for (j in 1:length(p)) if (u < p[j])
                     break
              sb[j] = 0
              res[i]=j
              }
       res[M:1]
       }
#
# 500 000 simulations avec une taille dans une liste de taille M=20.
# En prenant les m premières composantes du vecteur v, on a un échantillon
# de taille m.
#
M=20
x=runif(M)
Pik=array(0,c(M,M))
#
# Calcul des probabilités d'inclusion pour toutes les tailles d'échantillon de 1 à 20
#
for(i in 1:M) Pik[i,]=inclusionprobabilities(x, i)
rowSums(Pik)

SIM=50000
SS=array(0,c(M,M))
for(i in 1:SIM)
{
S=array(0,c(M,M))
v=elimination(x)
for(i in 1:M) S[i,v[1:i]]=1
SS=SS+S
}
SS=SS/SIM
#
# Comparaison des probabilités d'inclusion réelles et empiriques.
#
Pik
SS
SS-Pik
```

Bibliographie

Chikkagoudar, M.S. (1966). A note on inverse sampling with equal probabilities. *Sankhyā,* A28, 93-96.

Chikkagoudar, M.S. (1969). Inverse sampling without replacement. *Australian Journal of Satistic*, 11, 155-165.

Hájek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, part on by D. Basu. Dans *Foundations of Statistical Inference*, (Éds., Godambe, V.P. et Sprott, D.A.), page 326, Toronto, Canada. Holt, Rinehart, Winston.

Johnson, N.L., Kemp, A.W. et Kotz, S. (2005). *Univariate Discrete Distributions*. New York: John Wiley & Sons, Inc.

Mikulski, P.W., et Smith, P.J. (1976). A variance bound for unbiased estimation in inverse sampling. *Biometrika*, 63(1), 216-217.

Miller, G.K., et Fridell, S.L. (2007). A forgotten discrete distribution? Reviving the negative hypergeometric model. *The American Statistician*, 61(4), 347-350.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

Ohlsson, E. (1995). Sequential Poisson sampling. Rapport de recherche 182, Stockholm University, Suède.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics*, 14, 149-162.

Pathak, P.K. (1964). On inverse sampling with unequal probabilities. *Biometrika*, 51, 185-193.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

Salehi, M.M., et Seber, G.A.F. (2001). A new proof of Murthy's estimator which applies to sequential sampling. *The Australian and New Zealand Journal of Statistics*, 43, 281-286.

Sampford, M.R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika*, 49(1/2), 27-40.

Tillé, Y. (1996). An elimination procedure of unequal probability sampling without replacement. *Biometrika*, 83, 238-241.

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