Procédures rapides pour la sélection d'échantillons à probabilités inégales à partir d'un flux

Yves Tillé [1]

[1] *Université de Neuchâtel, Bellevaux 51, 2000 Neuchâtel, Suisse*
email : yves.tille@unine.ch

Résumé. Les méthodes d'échantillonnage probabilistes ont été mises au point dans le cadre de la statistique d'enquête. Récemment, ces méthodes ont fait l'objet d'un intérêt renouvelé pour la réduction de la taille des grands ensembles de données. Une application particulière est l'échantillonnage à partir d'un flux de données. Le flux est censé être tellement important qu'il ne peut pas être stocké. Lorsqu'une nouvelle unité apparaît, la décision de la conserver ou non doit être prise directement sans examiner toutes les unités déjà apparues dans le flux. Dans cet article, nous examinons les méthodes existantes d'échantillonnage avec des probabilités inégales à partir d'un flux. Nous proposons ensuite un résultat général sur le sous-échantillonnage à partir d'un échantillon équilibré qui nous permet de proposer plusieurs nouvelles solutions pour l'échantillonnage et le sous-échantillonnage à partir d'un flux. Plusieurs nouvelles applications de ce résultat général sont développées.

Mots-clés. échantillonnage ; échantillonnage équilibré ; flux ; méthode de Chao ; méthode de Fuller.

Abstract. Probability sampling methods were developed in the framework of survey statistics. Recently sampling methods are the subject of a renewed interest for the reduction of the size of large data sets. A particular application is sampling from a data stream. The stream is supposed to be so important that it cannot be stored. When a new unit appears, the decision to conserve it or not must be taken directly without examining all the units that already appeared in the stream. In this paper, we examine the existing possible methods for sampling with unequal probabilities from a stream. Next we propose a general result about subsampling from a balanced sample that enables us to propose several new solutions for sampling and subsampling from a stream. Several new applications of this general result are developed.

Keywords. balanced sampling ; Chao method ; Fuller method ; sampling ; stream ;

1 Introduction

L'échantillonnage avec des probabilités d'inclusion inégales, une taille fixe et sans remplacement est plus compliqué qu'il n'y paraît à première vue. En effet, les solutions immédiates et intuitives sont généralement incorrectes en ce sens qu'elles ne satisfont pas aux probabilités d'inclusion prescrites. La première méthode correcte était probablement l'échantillonnage systématique à probabilités inégales proposé par Madow (1949). Dans leur livre, Brewer et Hanif (1983) décrivent plusieurs dizaines de méthodes d'échantillonnage à probabilités inégales. Beaucoup d'entre elles ne sont pas générales. Par exemple, plusieurs méthodes ne sont applicables que pour des probabilités d'inclusion particulières ou pour une taille égale à deux. Tillé (2006) présente un large ensemble de plans d'échantillonnage à probabilité inégale. Le plan de sondage est défini comme la probabilité de sélectionner un sous-ensemble de la population. Un algorithme d'échantillonnage est une procédure qui permet de mettre en œuvre un plan de sondage. Plusieurs algorithmes différents peuvent servir à mettre en œuvre le même plan de sondage.

Dans cet article, nous nous concentrons sur les algorithmes séquentiels. Pour ces algorithmes, la décision de sélectionner ou non une unité est prise irrémédiablement après examen de l'unité. Si une unité n'est pas sélectionnée, toutes les informations la concernant sont oubliées. Les algorithmes séquentiels sont évidemment pratiques pour échantillonner dans un flux parce que les unités qui ne sont pas sélectionnées ne doivent pas être stockées. Par conséquent, les méthodes d'échantillonnage intéressent de plus en plus les informaticiens pour échantillonner dans les flux (Duffield, 2004).

Nous donnons un résultat général qui permet de définir un grand nombre d'algorithmes séquentiels pour échantillonner avec des probabilités d'inclusion inégales, une taille fixe et sans remplacement. Après avoir défini la notation dans la Section 2, nous expliquons dans la Section 3 comment calculer et mettre à jour les probabilités d'inclusion. Ensuite, dans la Section 4, nous montrons pourquoi le problème est tellement compliqué en présentant des méthodes qui ne sont pas correctes au sens où les probabilités d'inclusion ne sont pas satisfaites ou la taille de l'échantillon n'est pas fixée.

Lorsque les probabilités d'inclusion ne doivent pas être mises à jour, l'échantillonnage systématique, la méthode du pivot (Deville et Tillé, 1998; Grafström et al., 2012; Chauvet, 2012) ou la méthode de Fuller (Fuller, 1970) peuvent être des solutions appropriées (Section 5). Toutefois, lorsque la taille de l'échantillon est fixe et que la taille de la population n'est pas connue à l'avance, les probabilités d'inclusion doivent être mises à jour à chaque étape. C'est la base de la méthode de Chao (1982) qui utilise à chaque étape un réservoir de taille fixe qui est mis à jour lorsqu'une nouvelle unité est examinée. Cette mise à jour ne nécessite pas la connaissance des unités déjà exclues de l'échantillon.

Dans la Section 6, nous montrons que lorsqu'un échantillon est équilibré sur certaines variables auxiliaires, un sous-échantillon peut être sélectionné avec une taille fixe et d'autres probabilités d'inclusion. Ce résultat permet de proposer plusieurs généralisations pour la méthode de Chao. Les unités peuvent être considérées par blocs et la méthode

de Chao peut être étendue à l'échantillonnage équilibré. Il est également possible de sélectionner un échantillon en deux passages ou de sélectionner un échantillon de telle sorte que les probabilités d'inclusion peuvent être modifiées (Section 7).

2 Notation

Considérons une population U_N de taille N. Un échantillon s est un sous-ensemble de U_N et un plan de sondage $p(.)$ est une distribution de probabilités sur tous les échantillons telle que

$$p(s) \geq 0 \text{ et } \sum_{s \subset U_N} p(s) = 1.$$

Soit S un échantillon aléatoire qui est une variable aléatoire dont la distribution de probabilités est donnée par le plan de sondage $\Pr(S = s) = p(s)$.

La probabilité d'inclusion π_k est la probabilité de sélectionner une unité particulière $k \in U$. Théoriquement, elle peut être dérivée du plan de sondage de la façon suivante :

$$\pi_k = \Pr(k \in S) = \sum_{\substack{s \ni k \\ s \subset U_N}} p(s).$$

La probabilité d'inclusion conjointe est la probabilité que deux unités $k, \ell \in U$ sont sélectionnées conjointement dans l'échantillon.

$$\pi_{k\ell} = \Pr(\{k, \ell\} \subset S) = \sum_{\substack{s \subset \{k,\ell\} \\ s \subset U_N}} p(s).$$

Un flux est une population dont la taille n'est pas connue à l'avance. On peut considérer un flux comme une population croissante $U_N, U_{N+1}, U_{N+2}, U_{N+2}, \ldots$ dans lequel un échantillon doit être sélectionné à chaque étape. De plus, la population est censée être si importante qu'il n'est pas possible ou difficile de sauvegarder toutes les observations. L'échantillon doit donc être sélectionné séquentiellement dans le flux et la méthode d'échantillonnage se termine lorsque le flux s'arrête.

Considérons une variable d'intérêt y et y_k la valeur prise par cette variable sur l'unité k. Soit Y le total des valeurs $Y = \sum_{k \in U_N} y_k$. Si tous les $\pi_k > 0$ $k \in U$, alors l'estimateur par expansion donné par

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

est sans biais de

$$Y = \sum_{k \in U_N} y_k$$

(Narain, 1951; Horvitz et Thompson, 1952).

3 Calcul des probabilités d'inclusion

Supposons qu'une variable auxiliaire x est connue pour toutes les unités de la population et que x_k soit la valeur prise par x sur l'unité k. L'objectif est de définir des probabilités d'inclusion proportionnelles à x_k. Il existe deux possibilités qui correspondent à deux objectifs différents.

— La première possibilité est appelée "taux d'échantillonnage fixe". Dans ce cas, la taille de l'échantillon augmente avec la taille de la population. Les probabilités d'inclusion sont calculées une fois en utilisant un coefficient de proportionnalité $\tau > 0$. Alors

$$\pi_k = \min(1, \tau x_k). \tag{1}$$

Dans ce cas, la taille de l'échantillon n'est pas contrôlée. En effet, la taille de l'échantillon n peut être aléatoire et l'espérance de la taille de l'échantillon est donnée par $\mathrm{E}(n) = \sum_{k \in U_N} \pi_k$ qui n'est pas nécessairement entier.

— La deuxième possibilité est appelée "taille fixe de l'échantillon". Dans ce cas, un échantillon-réservoir de la taille n est sélectionné pour chaque sous-population. Dans la méthode du réservoir, le premier échantillon est composé des n premières unités de la population. Pour chaque étape suivante, une unité supplémentaire est prise en considération et peut être incluse avec une probabilité donnée. Dans ce cas, une unité du réservoir est enlevée. Les probabilités d'inclusion doivent ensuite être mises à jour pour chaque taille de population U_n, U_{n+1}, …, U_i, …, U_N, …, et sont calculées de telle sorte que

$$\pi_k(U_i, n) = \min(1, x_k \, ; \tau_i), \tag{2}$$

où τ_i est obtenu en résolvant

$$\sum_{k \in U_i} \min(1, x_k \, ; \tau_i) = n.$$

Si les probabilités d'inclusion sont égales, la méthode du réservoir a été décrite dans Knuth (1981, p. 144), McLeod et Bellhouse (1983) et Vitter (1985).

4 Pourquoi le problème est compliqué ?

4.1 Échantillonnage de Poisson

La méthode d'échantillonnage la plus simple est le tirage de Poisson. Les probabilités d'inclusion sont calculées en utilisant Expression (1). Dans le flux, chaque unité est sélectionnée indépendamment de la sélection des autres unités. Ce plan peut être mis en œuvre en générant pour chaque unité une variable aléatoire continue uniforme u_k dans $[0, 1]$. Ensuite, les unités telles que $u_k < \pi_k$ sont sélectionnées. Le problème principal est

que la taille de l'échantillon n'est pas fixe. Sa distribution de probabilités est Poisson-Binomiale (Hodges Jr. et Le Cam, 1960; Stein, 1990; Chen et Liu, 1997).

4.2 Échantillonnage aléatoire pondéré

L'échantillonnage pondéré peut être résumé comme suit. À la première étape, une unité est sélectionnée parmi la population U_N avec une probabilité de tirage inégale $p_k = x_k / \sum_{k \in U_N} x_k$. Supposons que l'unité sélectionnée est désignée par j : cette unité est retirée de la population et les probabilités de tirage sont recalculées comme suit $p_k^j = p_k/(1-p_j)$. Ensuite, une deuxième unité est sélectionnée avec la probabilité p_k^j à partir de $U_N \backslash \{j\}$. Cette unité, désignée par i, est également retirée de la population et les probabilités de tirage sont rééchelonnées selon $p_k^{j,i} = p_k/(1-p_j-p_i)$ et ainsi de suite.

Cette méthode est fausse. En effet, même avec un échantillon de taille égale à $n = 2$, la probabilité de sélectionner l'unité k est égale à :

$$p_k + \sum_{j \neq k} p_j \frac{p_k}{1-p_j}.$$

Cette probabilité n'est pas proportionnelle à x_k et ne correspond pas aux probabilités définies comme suit $\pi_k(N,n) = \min(1, x_k\ ; \tau)$, avec

$$\sum_{k \in U_N} \min(1, x_k\ ; \tau) = n.$$

Le fait que cette méthode soit fausse a déjà été souligné par Narain (1951) qui proposait de calculer des probabilités de tirage *ad hoc* afin d'obtenir des probabilités d'inclusion fixes. Le calcul est cependant lourd pour des échantillons de taille supérieure à deux et implique un calcul sur l'ensemble de la population, ce qui rend impossible une implémentation séquentielle (voir aussi Brewer et Hanif, 1983, pp. 25-26).

Efraimidis et Spirakis (2006) ont proposé une mise en œuvre rapide du plan de sondage pondéré. Toutefois, ce plan ne permet pas de sélectionner un échantillon qui satisfait à des probabilités d'inclusion données. Il n'est alors pas possible d'estimer sans biais un total. Efraimidis (2015) souligne que l'échantillonnage aléatoire pondéré peut se référer à deux définitions. Dans le premier cas, le poids relatif de chaque élément détermine la probabilité que l'élément fasse partie de l'échantillon final. Dans le second, le poids de chaque élément détermine la probabilité que l'élément soit sélectionné dans chacune des sélections de la procédure d'échantillonnage. Pour cette raison, nous sommes convaincus que dans le domaine de l'échantillonnage, le mot "pondération" devrait être évité parce qu'il n'est pas clair s'il se réfère aux probabilités de tirage ou aux probabilités d'inclusion.

5 Méthodes avec un taux de sondage fixe

5.1 Échantillonnage systématique à probabilités inégales

L'échantillonnage systématique semble avoir été proposé en premier lieu par Madow (1949). En comptabilité, ce plan de sondage est appelé échantillonnage d'unités dollar (*Dollar Unit Sampling*) (Leslie et al., 1979) ou, plus généralement, échantillonnage d'unités monétaires (*Monetary Unit Sampling*). La mise en œuvre de l'échantillonnage systématique est très simple. Il exige le calcul des probabilités d'inclusion cumulées :

$$V_k = \sum_{l=1}^{k} \pi_k \text{ avec } V_0 = 0.$$

Une variable aléatoire uniforme u dans $[0, 1]$ n'est générée qu'une seule fois. Ensuite, les unités k telles que que les intervalles $[V_{k-1} - u, V_k - u]$ contient un entier sont sélectionnés. La mise en œuvre est particulièrement pratique pour l'échantillonnage dans un flux. Si la somme des probabilités d'inclusion n'est pas un nombre entier, l'échantillonnage systématique sélectionne un échantillon de taille égale à

$$\left\lfloor \sum_{k \in U_N} \pi_k \right\rfloor \text{ ou } \left\lfloor \sum_{k \in U_N} \pi_k \right\rfloor + 1,$$

où $\lfloor a \rfloor$ est le plus grand entier inférieur ou égal à a.

5.2 Méthode du pivot ordonné

Considérons une population U de taille N dans laquelle nous voulons sélectionner un échantillon de taille fixe n avec des probabilités d'inclusion inégales $0 < \pi_k < 1, k \in U$ tel que $\sum_{k \in U_N} \pi_k = n$. La méthode du pivot proposée par Deville et Tillé (1998) est particulièrement simple et consiste à choisir à chaque étape deux unités (désignées par i et j) dans la population. Leurs probabilités d'inclusion (π_i, π_j) sont transformées au hasard en $(\widetilde{\pi}_i, \widetilde{\pi}_j)$ en utilisant la randomisation suivante :

$$(\widetilde{\pi}_i, \widetilde{\pi}_j) = \begin{cases} (\min(1, \pi_i + \pi_j), \max(\pi_i + \pi_j - 1, 0)) & \text{avec une probabilité } q \\ (\min(\pi_i + \pi_j, 1), \max(0, \pi_i + \pi_j - 1)) & \text{avec une probabilité } 1 - q, \end{cases}$$

avec

$$q = \frac{\min(1, \pi_i + \pi_j) - \pi_j}{2 \min(1, \pi_i + \pi_j) - \pi_i - \pi_j}.$$

La validité de la méthode est basée sur le fait que $\mathrm{E}(\widetilde{\pi}_i, \widetilde{\pi}_j) = (\pi_i, \pi_j)$ et $\widetilde{\pi}_i + \widetilde{\pi}_j = \pi_i + \pi_j$. Cela garantit que les probabilités d'inclusion sont satisfaites et que la somme des composantes du vecteur est toujours égale à n.

Cette étape élémentaire peut être répétée sur des couples d'unités contenant des valeurs qui ne sont pas égales à 0 ou 1. Comme à chaque étape, un composant est transformé en un 0 ou un 1, en $N-1$ étapes tout au plus, l'échantillon est sélectionné. La méthode du pivot ordonné consiste simplement à prendre les unités selon leur ordre dans le vecteur des probabilités d'inclusion. La première étape est donc appliquée sur les unités $(1,2)$. L'unité qui conserve une valeur non entière est alors confrontée à l'unité 3 et ainsi de suite.

Si la somme des probabilités d'inclusion n'est pas entière, la méthode du pivot se termine avec un vecteur dont les composantes sont toutes entières sauf une qui est égale à la partie fractionnaire de la somme des probabilités d'inclusion. Ce résultat est donné par exemple par la fonction `UPpivotal` de la fonction R `sampling` package (Matei et Tillé, 2016). La méthode du pivot est un cas particulier de la méthode du cube (Deville et Tillé, 2004) avec une seule variable d'équilibrage donnée par le vecteur des probabilités d'inclusion. La méthode du pivot ordonnée est le cas particulier de la mise en œuvre rapide de la méthode du cube proposée par Chauvet et Tillé (2006).

5.3 Implémentation rapide de la méthode de Fuller

Une variante de la méthode du pivot a été proposée par Tillé (2018). La seule différence avec celle-ci est qu'une unité fantôme '0' est ajoutée au début avec une probabilité d'inclusion π_0 qui est générée aléatoirement selon une distribution uniforme en $[0,1]$. Ensuite, la méthode du pivot est appliquée. Si la somme de la probabilité d'inclusion initiale n'est pas un nombre entier, une autre unité fantôme est ajoutée à la fin avec une probabilité d'inclusion.

$$\pi_{N+1} = \left\lceil \sum_{k=1}^{N} \pi_k \right\rceil - \sum_{k=1}^{N} \pi_k,$$

où $\lceil a \rceil$ est le plus petit entier plus grand que a.

Étant donné que la partie fractionnaire de la somme

$$\sum_{k=0}^{N+1} \pi_k$$

est égal à π_0, la méthode du pivot se termine avec un vecteur de valeurs égales à zéro ou à un sauf une (désignée par j) qui est égal à π_0. Si l'unité 0 est prise, alors l'unité j est aussi sélectionnée, et l'unité j n'est pas sélectionnée dans le cas contraire. Enfin, les deux unités fantômes sont retirées de l'échantillon qu'elles aient été sélectionnées ou non.

5.4 Méthode du cube rapide

Considérons qu'un ensemble de variables auxiliaires $z_1, \ldots, z_p$ est connu pour chaque unité de la population. Soit $\mathbf{z}_k \in \mathbb{R}^p$ le vecteur des valeurs prises par ces p variables sur

l'unité k. L'objectif est de sélectionner un plan de sondage équilibré en ce sens que les équations d'équilibrage

$$\sum_{k\in S}\frac{\mathbf{z}_k}{\pi_k}\approx\sum_{k\in U_N}\mathbf{z}_k \tag{3}$$

sont approximativement satisfaites tout en respectant les probabilités d'inclusion $\pi_k, \in U_N$. Un échantillon exactement équilibré n'existe généralement pas parce que la sélection d'un échantillon est un problème en nombre entier.

La méthode du cube (Deville et Tillé, 2004; Chauvet et Tillé, 2006) permet de sélectionner des échantillons équilibrés. La méthode est composée de deux phases appelées la phase de vol et la phase d'atterrissage. Nous nous intéresserons principalement à la phase de vol qui sélectionne des quasi-échantillons équilibrés sur $\mathbf{z}_k$. La phase de vol rapide (ffph pour fast flight phase) de la méthode du cube est une fonction qui génère un vecteur aléatoire à partir d'un vecteur de probabilités d'inclusion :

$$\text{ffph}(\pi_1,\ldots,\pi_k,\ldots,\pi_N)=\boldsymbol{\psi}=(\psi_1,\ldots,\psi_k,\ldots,\psi_N)^\top$$

de telle sorte que $0\leq\psi_k\leq 1, k\in U_N$, $\mathrm{E}(\psi_k)=\pi_k$, $\text{card}\{0<\psi_k<1\}\leq p$ et

$$\sum_{k\in U_N}\frac{\psi_k\mathbf{z}_k}{\pi_k}=\sum_{k\in U_N}\mathbf{z}_k. \tag{4}$$

Le vecteur $\boldsymbol{\psi}$ est un quasi-échantillon dans le sens où presque toutes ses composantes sont égales à 0 ou 1 sauf au maximum p composantes, où p est la dimension de $\mathbf{z}_k$.

Proposition 1. *S'il existe un vecteur $\boldsymbol{\lambda}^\top\boldsymbol{z}_k=\pi_k$ pour tous les $k\in\{U_N|\pi_k<1\}$, alors*

$$\sum_{k\in U_N}\pi_k=\sum_{k\in U_N}\psi_k.$$

Démonstration. Lorsque $\pi_k=1$, $\psi_k=1$. L'Équation (4) implique donc que

$$\sum_{k\in\{U_N|\pi_k<1\}}\frac{\boldsymbol{\lambda}^\top\mathbf{z}_k\psi_k}{\pi_k}=\sum_{k\in\{U_N|\pi_k<1\}}\frac{\pi_k\psi_k}{\pi_k}=\sum_{k\in\{U_N|\pi_k<1\}}\psi_k=\sum_{k\in\{U_N|\pi_k<1\}}\boldsymbol{\lambda}^\top\mathbf{z}_k=\sum_{k\in\{U_N|\pi_k<1\}}\pi_k.$$

□

La Proposition 1 montre que l'échantillon a une taille fixe quand c'est possible. En effet, si la somme des probabilités d'inclusion n'est pas entière, la taille de l'échantillon ne peut être fixée avec exactitude.

Avec la fonction ffph(.), il est possible de proposer une implémentation en un seul passage pour un taux d'échantillonnage fixe. On prend d'abord les $p+1$ premières unités avec π_k non entier et on applique la fonction ffph(.). La fonction ffph(.) renvoie un vecteur avec au moins une valeur égale à 0 ou 1, les unités avec des valeurs 0 sont oubliées et

les unités avec la valeur 1 sont sauvegardées. Les unités avec une valeur 0 ou 1 sont retirées du vecteur. Le vecteur est ensuite complété par une ou plusieurs unités suivantes pour construire à nouveau un vecteur de taille $p+1$. La fonction ffph(.) est appliquée à nouveau et ainsi de suite. Cette implémentation, proposée par Chauvet et Tillé (2006), permet d'échantillonner dans un flux. À la fin du flux, une phase d'atterrissage (Deville et Tillé, 2004) peut être appliquée sur les composantes de $\boldsymbol{\psi}$ qui ne sont pas des entiers afin d'avoir un échantillon approximativement équilibré.

6 Méthodes de taille fixe

6.1 Méthode de Chao

Pour les probabilités d'inclusion égales, la méthode du "réservoir" est décrite par Knuth (1981, p. 144), McLeod et Bellhouse (1983) et Vitter (1985). Une méthode de taille fixe ou approximativement fixe doit tenir compte du fait que les probabilités d'inclusion décroissent quand la taille de la population croît. La seule méthode de réservoir qui satisfait réellement aux probabilités d'inclusion a été proposée par Chao (1982). (voir aussi Sugden et al., 1996). Considérons une séquence de population $U_n \subset U_{n+1} \subset \cdots \subset U_i \subset \cdots \subset U_N$. Soit $\pi_k(U_i, n)$ les probabilités d'inclusion calculées sur une population U_i de taille i pour un échantillon de taille n tel que défini dans l'Équation (2). La méthode de Chao commence par la sélection des n premières unités de population U_n qui s'appelle le réservoir. À chaque étape, le réservoir est mis à jour comme suit. Supposons qu'à l'étape $i-1$, le réservoir est désigné par S_{i-1}. À l'étape $i = n+1, \ldots, N$, l'unité i est incluse dans le réservoir avec une probabilité $\pi_i(U_i, n)$. Si l'unité i est sélectionnée, l'une des unités du réservoir est enlevée avec la probabilité :

$$a_{ki} = \frac{1}{\pi_k(U_i, n)} \left[1 - \frac{\pi_k(U_i, n)}{\pi_k(U_{i-1}, n)}\right], k = 1, \ldots, i-1.$$

Il est en effet possible de prouver que

$$\sum_{k \in S_{i-1}} a_{ki} = 1.$$

Une autre façon de présenter la méthode consiste à ajouter l'unité i dans le réservoir qui devient $A_i = S_{i-1} \cup \{i\}$. Ensuite, n unités sont sélectionnées parmi A_i avec probabilités.

$$\frac{\pi_k(U_i, n)}{\pi_k(U_{i-1}, n)}, k \in S_{i-1}, \text{ et } \pi_i(U_i, n).$$

En effet,

$$\sum_{S_{i-1}} \frac{\pi_k(U_i, n)}{\pi_k(U_{i-1}, n)} + \pi_i(U_i(U_i, n) = n. \tag{5}$$

Dans la méthode Chao originale, les probabilités d'inclusion $\pi_k(U_i, n)$ doivent être recalculées à chaque étape pour toutes les unités qui ont déjà été observées. Cependant, Cohen et al. (2009) ont montré que les unités non sélectionnées peuvent être définitivement oubliées et qu'une implémentation séquentielle est vraiment possible dans le sens où seules les unités sélectionnées à l'étape i doivent être stockées, car il est possible de calculer $\pi_k(U_i, n)$ en ne connaissant que les valeurs x_k de S_{i-1}.

6.2 Généralisation de la méthode de Chao

Les résultats de Chao (1982) et Cohen et al. (2009) peuvent être généralisés. Supposons qu'un quasi-échantillon $\boldsymbol{\psi}^1 = (\psi_1^1, \ldots, \psi_N^1)^\top$ ait été sélectionné avec un échantillonnage équilibré dans une population U_N avec probabilités d'inclusion $\boldsymbol{\pi}^1 = (\pi_1^1, \ldots, \pi_N^1)^\top$. Les probabilités d'inclusion sont proportionnelles à une variable positive x, de sorte que nous pouvons écrire $\pi_k^1 = \min(1, \tau_1 x_k)$, où τ_1 est obtenu en résolvant l'équation suivante :

$$\sum_{k \in U_N} \min(1, \tau_1 x_k) = n.$$

La théorie ci-dessous admet des valeurs non entières pour n.

Le quasi-échantillon $\boldsymbol{\psi}^1$ est supposé être équilibré sur les p variables auxiliaires dont les valeurs sur l'unité k sont les composantes du vecteur $\mathbf{z}_k$. Les équations d'équilibrage sont alors exactement satisfaites :

$$\sum_{k \in U_N} \frac{\mathbf{z}_k \psi_k^1}{\pi_k^1} = \sum_{k \in U_N} \mathbf{z}_k. \tag{6}$$

Proposition 2. *S'il existe un vecteur $\boldsymbol{\lambda}$ tel que $\boldsymbol{\lambda}^\top \boldsymbol{z}_k = x_k$, pour tous les $k \in \{k \in U_n | \pi_k^1 < 1\}$, l'Équation (6) implique : $\sum_{k \in U_N} \psi_k^1 = \sum_{k \in U_N} \pi_k^1$.*

Démonstration. Du côté gauche de l'Équation (6),

$$\begin{aligned} \boldsymbol{\lambda}^\top \sum_{k \in U_N} \frac{\mathbf{z}_k \psi_k^1}{\pi_k^1} &= \sum_{k \in U_N} \frac{x_k \psi_k^1}{\pi_k^1} = \sum_{k \in U_N | \pi_k^1 < 1} \frac{x_k \psi_k^1}{\pi_k^1} + \sum_{k \in U_N | \pi_k^1 = 1} \frac{x_k \psi_k^1}{\pi_k^1} \\ &= \sum_{k \in U_N | \pi_k^1 < 1} \frac{x_k \psi_k^1}{x_k \tau_1} + \sum_{k \in U_N | \pi_k^1 = 1} x_k = \sum_{k \in U_N | \pi_k^1 < 1} \frac{\psi_k^1}{\tau_1} + \sum_{k \in U_N | \pi_k^1 = 1} x_k. \end{aligned} \tag{7}$$

Du côté droit de l'Équation (6),

$$\boldsymbol{\lambda}^\top \sum_{k \in U_N} \mathbf{z}_k = \sum_{k \in U_N} x_k. \tag{8}$$

En égalisant (7) et (8), on obtient

$$\sum_{k\in U_N|\pi_k^1<1} \psi_k^1 = \sum_{k\in U_N|\pi_k^1<1} \tau_1 x_k = \sum_{k\in U_N|\pi_k^1<1} \pi_k^1.$$

Quand $\pi_k^1 = 1$, $\psi_k^1 = 1$, et nous obtenons la Proposition 2. □

Lors de la deuxième phase, un autre quasi-échantillon équilibré ψ_k^2 est sélectionné de telle sorte que les probabilités d'inclusion finales sont les suivantes $\boldsymbol{\pi}^2 = (\pi_1^2, \ldots, \pi_k^2 \ldots, \pi_N^2)^\top$. Les probabilités d'inclusion π_k^2 sont supposées être proportionnelles à une autre variable. v_k, i.e. $\pi_k^2 = \min(1, \tau_2 v_k)$ où τ_2 est obtenu en résolvant

$$\sum_{k\in U_N} \min(1, \tau_2 v_k) = m.$$

La variable v_k peut être égale à x_k. La théorie ci-dessous admet aussi des valeurs non entières pour m. De plus, les composantes de $\boldsymbol{\pi}^2$ doivent toutes être inférieures ou égales aux composantes de $\boldsymbol{\pi}^1$ ($\pi_k^2 \leq \pi_k^1, k \in U_N$), et donc $m \leq n$.

L'échantillon de la deuxième phase est tiré dans le quasi-échantillon $\boldsymbol{\psi}^1$. Les probabilités de tirage sont :

$$\xi_k = \frac{\pi_k^2 \psi_k^1}{\pi_k^1}. \tag{9}$$

Puisque $\mathrm{E}(\psi_k^1) = \pi_k^1$, nous avons $\mathrm{E}(\xi_k) = \pi_k^2$.

Proposition 3. *Si le quasi-échantillon ψ_1 est équilibré sur $\boldsymbol{z}_k$ et qu'il existe un vecteur $\boldsymbol{\theta} \in \mathbb{R}^p$ tel que $\boldsymbol{\theta}^\top \boldsymbol{z}_k = v_k$, alors les probabilités d'inclusion π_k^2 et donc les probabilités de tirage ξ_k peuvent être calculées à partir de v_k sans connaître les unités telles que $\psi_k^1 = 0$, en résolvant dans τ_2 :*

$$\sum_{k\in U_N} \min(1, \tau_2 v_k) \frac{\psi_k^1}{\pi_k^1} \pi_k^1 = m.$$

Démonstration. D'après les équations d'équilibrage, nous avons

$$\sum_{k\in U_N} v_k \frac{\psi_k^1}{\pi_k^1} = \sum_{k\in U_N} v_k.$$

De plus, si $\pi_k^1 < 1$, nous avons aussi $\pi_k^2 < 1$. On considére

$$\begin{aligned}
\sum_{k\in U_N} \pi_k^2 \frac{\psi_k^1}{\pi_k^1} &= \sum_{k\in U_N|\pi_k^1<1} \pi_k^2 \frac{\psi_k^1}{\pi_k^1} + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N|\pi_k^1<1} \tau_2 v_k \frac{\psi_k^1}{\pi_k^1} + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N} \tau_2 v_k \frac{\psi_k^1}{\pi_k^1} - \sum_{k\in U_N|\pi_k^1=1} \tau_2 v_k \frac{\psi_k^1}{\pi_k^1} + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N} \tau_2 v_k - \sum_{k\in U_N|\pi_k^1=1} \tau_2 v_k + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N|\pi_k^1<1} \pi_k^2 + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N} \pi_k^2.
\end{aligned}$$

On obtient ainsi

$$\sum_{k\in U_N} \min(1, \tau_2 v_k) \frac{\psi_k^1}{\pi_k^1} = \sum_{k\in U_N} \min(1, \tau_2 v_k).$$

□

La Proposition 3 nous permet d'oublier les unités qui ne sont pas sélectionnées tout en préservant la possibilité de calculer les probabilités d'inclusion de la seconde phase. Le deuxième quasi-échantillon est équilibré sur $\psi_k^1 \mathbf{z}_k / \pi_k^1$. Ainsi, si $\sum_{k\in U_N} \pi_k^1$ est un entier, $\boldsymbol{\psi}^1$ est un échantillon de taille fixe.

La Proposition 3 implique également que s'il existe un vecteur $\boldsymbol{\theta}$ tel que $\boldsymbol{\theta}^\top \mathbf{z}_k = v_k$, pour tout $k \in U_n$, alors

$$\sum_{k\in U_N} \xi_k = \sum_{k\in U_N} \pi_k^2. \tag{10}$$

L'équation (5) pour le plan de Chao est un cas particulier de l'égalité (10) lorsque $x_k = v_k$ et $\mathbf{z}_k = x_k$.

De plus, si le deuxième quasi-échantillon est équilibré sur $\psi_k^1 \mathbf{z}_k / \pi_k^1 = \psi_k^1$, nous avons aussi :

$$\sum_{k\in U_N|\psi_k^1>0} \frac{\psi_k^1 \psi_k^2}{\xi_k} = \sum_{k\in U_N} \frac{\pi_k^1 \psi_k^2}{\pi_k^2} \pi_k^2. = \sum_{k\in U_N} \psi_k^1 = \sum_{k\in U_N} \pi_k^1.$$

7 Quelques applications

Avec le résultat de la section précédente, il est possible d'imaginer plusieurs nouvelles méthodes.

7.1 Préservation de deux variables pour un échantillonnage à probabilités inégales

Supposons que deux variables $u_k > 0$ et $v_k > 0$ sont disponibles et que l'on hésite sur la variable à utiliser pour calculer les probabilités d'inclusion. L'astuce consiste à sélectionner un premier échantillon qui préserve les possibilités de sélectionner un sous-échantillon qui peut être soit avec des probabilités d'inclusion proportionnelles à u_k ou v_k.

On calcule d'abord $x_k = \max(u_k, v_k)$. Un premier quasi-échantillon ψ_k^1 peut être sélectionné avec des probabilités d'inclusion $\pi_k^1 = \min(1, \tau_1 x_k)$, où τ_1 peut être choisi librement. Cet échantillon est équilibré sur trois variables auxiliaires $\mathbf{z}_k = (u_k, v_k, x_k)^\top$. Ensuite, un deuxième échantillon peut être sélectionné avec des probabilités d'inclusion $\pi_k^2 = \min(1, \tau_2 u_k)$ ou $\pi_k^2 = \min(1, \tau_2 v_k)$, avec $\tau_2 \leq \tau_1$.

7.2 Méthode du réservoir en blocs

La méthode de Chao peut être généralisée. Au lieu de considérer à chaque étape une unité à inclure dans le réservoir, un ensemble de H unités peut être considéré. À la première étape, un bloc des H premières unités (où $H > n$) est sélectionné dans la population avec des probabilités d'inclusion $\pi_k^1 = \min(1, \tau_1 x_k)$, où

$$\sum_{k \in U_H} \min(1, \tau_1 x_k) = n.$$

Parmi ces H unités, n unités sont sélectionnées dans un échantillon ou un quasi-échantillon ψ_k^1. Ensuite, les H unités suivantes sont considérées et, parmi ces $n + H$ unités, n unités sont sélectionnées avec probabilités d'inclusion $\pi_k^2 = \min(1, \tau_2 x_k)$, où

$$\sum_{k \in U_{2H}} \min(1, \tau_2 x_k) = \sum_{k \in U_{2H}} \min(1, \tau_2 x_k) \frac{\psi_k^1}{\pi_k^1} = n.$$

Les probabilités de tirage sont calculées à l'aide de l'Équation (9) et les unités n suivantes sont sélectionnées parmi ces $n + H$ unités, et ainsi de suite.

7.3 Méthode du réservoir équilibré

La méthode de Chao peut être généralisée à un échantillonnage équilibré. Considérons que les p variables d'équilibrage sont dans $\mathbf{z}_k$ et qu'il existe un vecteur $\boldsymbol{\lambda}$ tel que $\boldsymbol{\lambda}^\top \mathbf{z}_k =$

x_k. On considère d'abord les $n+1$ premières unités de la population. Puis, on définit $\pi_k^1 = \min(1, \tau_1 x_k)$ où

$$\sum_{k \in U_{n+1}} \min(1, \tau_1 x_k) = n.$$

On sélectionne un quasi-échantillon ψ_k^1 qui est équilibré avec la taille n dans ce sous-ensemble de la population. Comme $\boldsymbol{\lambda}^\top \mathbf{z}_k = x_k$, alors $\sum_{k \in U_{n+1}} \pi_k^1 = n$.

On prend l'unité suivante (qui a le numéro $n+2$). On calcule $\pi_k^2 = \min(1, \tau_2 x_k)$ où

$$\sum_{k \in U_{n+2}} \min(1, \tau_2 x_k) = n.$$

On sélectionne un quasi-échantillon équilibré avec des probabilités de tirage données par l'Équation (9), et ainsi de suite. Cette méthode est assez lente, car un échantillon équilibré doit être sélectionné $N - n$ fois.

7.4 Méthode du réservoir en blocs équilibrée

La méthode du réservoir par blocs peut être combinée avec un échantillonnage équilibré. Un bloc des H premières unités (où $H > n$) est pris dans la population. Parmi ces H unités, un quasi-échantillon est sélectionné à l'aide d'un échantillonnage équilibré. Ensuite, on prend les H unités suivantes. Les probabilités d'inclusion sont calculées à l'aide de l'Équation (9). L'échantillonnage équilibré est de nouveau appliqué pour sélectionner un quasi-échantillon, et ainsi de suite. Cette méthode est plus rapide que la précédente car la phase de vol de la méthode du cube doit être exécutée moins fréquemment.

7.5 Méthode en deux passages

Dans le cas d'un très grand flux avec une grande taille d'échantillon, on peut imaginer une procédure en deux passages. Durant un premier passage, on sélectionne chaque unité du flux avec une probabilité calculée en fonction du nombre d'unités déjà examinées. La taille du flux est ainsi considérablement réduite Ensuite, on réalise une seconde phase d'échantillonnage afin d'obtenir un échantillon de taille fixe.

Au cours de la première phase, les n premières unités sont sélectionnées. Ensuite, un quasi-échantillon équilibré est sélectionné avec des probabilités d'inclusion $\pi_k^1 = \pi_k(U_k, n), k = n+1, \ldots, N$. $\pi_k(U_i, n) = \min(\tau_i x_k, 1)$ où

$$\sum_{k \in U_i} \min(\tau_i x_k, 1) = n.$$

Ce quasi-échantillon ψ_k^1 est équilibré sur deux variables auxiliaires $\mathbf{z}_k = (\pi_k(U_k, n), x_k)^\top$.

Si x_k est constant, le $\pi_k(U_k, n) = 1, k = 1, \ldots, n$, et $\pi_k(U_k, n) = n/k, k = n+1, \ldots, N$. Dans ce cas, l'espérance de la taille de l'échantillon est la suivante

$$\sum_{k \in U_N} \pi_k(U_k, n) = n + \sum_{k=n+1}^{N} \frac{n}{k} = n \left(1 + \sum_{k=1}^{N} \frac{1}{k} - \sum_{k=1}^{n} \frac{1}{k} \right) \approx n(1 + \ln N - \ln n) = n + n \ln \frac{N}{n}.$$

La taille de l'échantillon est considérablement réduite à la fin de la première phase.

Ensuite, le deuxième échantillon est sélectionné avec des probabilités d'inclusion $\pi_k^2 = \pi_k(U_N, n)$ pour tous les $k \in U_N$. Les probabilités d'inclusion sont données par l'Équation (9). Notons que $\pi_k(U_k, n)$ devrait être calculé à chaque étape, ce qui peut être parfois lent. Cependant, toute limite supérieure pour $\pi_k(U_k, n)$ peut également être utilisée. Par exemple, si

$$\pi_k(U_k, n) = \min(x_k \tau_k, 1),$$

alors une limite supérieure facilement calculable pourrait être

$$\pi_{k+1}(U_{k+1}, n) \leq \min \left(n x_k \frac{\sum_{k \in U_k} x_k}{\sum_{k \in U_{k+1}} x_k}, 1 \right).$$

8 Discussion

La méthode de Chao permet d'échantillonner avec des probabilités d'inclusion inégales à partir d'un flux car les probabilités d'inclusion peuvent être mises à jour sans connaître les unités qui ne sont pas sélectionnées. Ce résultat peut être généralisé pour les quasi-échantillons obtenus par la méthode du cube. De plus, ces résultats peuvent également être étendus pour les problèmes où des blocs d'unités apparaissent dans le flux. En conséquence, on peut définir plusieurs nouveaux algorithmes pour échantillonner ou sous-échantillonner à partir d'un flux tout en préservant les propriétés d'équilibrage de l'échantillon.

Références

Brewer, K. R. W. et Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer, New York.

Chao, M.-T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69 :653–656.

Chauvet, G. (2012). On a characterization of ordered pivotal sampling. *Bernoulli*, 18(4) :1099–1471.

Chauvet, G. et Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21 :9–31.

Chen, X.-H. et Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7 :875–892.

Cohen, E., Duffield, N., Kaplan, H., Lund, C., et Thorup, M. (2009). Stream sampling for variance-optimal estimation of subset sums. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1255–1264. Society for Industrial and Applied Mathematics.

Deville, J.-C. et Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85 :89–101.

Deville, J.-C. et Tillé, Y. (2004). Efficient balanced sampling : The cube method. *Biometrika*, 91 :893–912.

Duffield, N. (2004). Sampling for passive internet measurement : A review. *Statistical Science*, pages 472–498.

Efraimidis, P. S. (2015). Weighted random sampling over data streams. In Zaroliagis, C., Pantziou, G., et Kontogiannis, S., editors, *Algorithms, Probability, Networks, and Games : Scientific Papers and Essays Dedicated to Paul G. Spirakis on the Occasion of His 60th Birthday*, pages 183–195. Springer International Publishing, Cham.

Efraimidis, P. S. et Spirakis, P. G. (2006). Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5) :181–185.

Fuller, W. A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, B32 :209–226.

Grafström, A., Lundström, N. L. P., et Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2) :514–520.

Hodges Jr., J. L. et Le Cam, L. (1960). The Poisson approximation to the Poisson binomial distribution. *Annals of Mathematical Statistics*, 31 :737–740.

Horvitz, D. G. et Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 :663–685.

Knuth, D. E. (1981). *The Art of Computer Programming (Volume II) : Seminumerical Algorithms*. Addison-Wesley, Reading, MA.

Leslie, D. A., Teitlebaum, A. D., et Anderson, R. J. (1979). *Dollar-unit sampling : a practical guide for auditors*. Copp Clark Pitman.

Madow, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, 20 :333–354.

Matei, A. et Tillé, Y. (2016). The R sampling package, Version 2.8. Université de Neuchâtel `https://cran.r-project.org/web/packages/sampling/index.html`.

McLeod, A. I. et Bellhouse, D. R. (1983). A convenient algorithm for drawing a simple random sampling. *Applied Statistics*, 32 :182–184.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3 :169–174.

Stein, C. (1990). Application of Newton's identities to a generalized birthday problem and to the Poisson-Binomial distribution. Technical Report TC 354, Department of Statistics, Stanford University.

Sugden, R. A., Smith, T. M. F., et Brown, R. P. (1996). Chao's list sequential scheme for unequal probability sampling. *Journal of Applied Statistics*, 23 :413–421.

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

Tillé, Y. (2018). Fast implementation of Fuller's unequal probability sampling method. Technical report, University of Neuchâtel.

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1) :37–57.