TD3–Chapitre 3 - Sondages aléatoires simples (SAS) à probabilités inégales

Exercice 11. (**Taille des ménages**) Une population est composée de 6 ménages de tailles respectives 2, 4, 3, 9, 1 et 2. On tire 3 ménages sans remise, avec une probabilité proportionnelle à leur taille.

1. Donner les probabilités d'inclusion des 6 ménages de la base de sondage.
2. Réaliser effectivement le tirage par une méthode systématique.
3. À partir de l'échantillon obtenu en 2, donner une estimation de la taille moyenne des ménages. Le résultat était-il prévisible ?

Corrigé :

Cet exercice correspond à regarder ce qui se passe lorsque l'on prend comme variable auxilière, la variable d'intérêt.

On note Y la variable taille du ménage en tant que variable d'intérêt.

1. Soit π_k la probabilité d'inclusion d'ordre 1 de l'unité (ménage) k. On a :

$$\pi_k = n \cdot \frac{y_k}{\sum_{i\in U} x_i}$$

où n est la taille de l'échantillon, ie dans notre cas $n = 3$. Ainsi : $\pi_4 = 3 \times 9/21 = 9/7 > 1$ que l'on modifie en $\pi_4 = 1$. On recalcule ensuite les autres valeurs, aisni

$$\pi_k = 2 \cdot \frac{y_k}{\sum_{i\in U, i\neq 4} x_i} = 2y_k/12 = y_k/6$$

Cela nous donne :

$$\pi_1 = 1/3, \pi_2 = 2/3, \pi_3 = 1/2, \pi_5 = 1/6, \pi_6 = 1/3$$

On a par ailleurs, ce qui nous sera utile par la suite :

$$C_0 = 0; C_1 = 1/3; C_2 = 1; C_3 = 3.2; C_4 = 2.5; C_5 = 8/3 \simeq 2.67; C_6 = 3.$$

2. Méthode de tirage systématique. On tire d'abord un nombre u suivant la loi uniforme sur [0; 1]. Prenons par exemple $\mathrm{u} = 0; 2$. La première unité sélectionnée est alors la première car $0 \leqslant 0.2 < 0.33$. On a ensuite $1 \leqslant u + 1 = 1.2 < 1.5$ et la deuxième unité sera donc sélectionnée la numéro trois. Enfin $u + 2 = 2, 2$ et la troisième unité sélectionnée est l'unité 4 car $1.5 \leqslant 2.2 < 2.5$. On tire donc les unités 1,3 et 4 . 3. L'estimateur de la taille moyenne est :

$$\hat{Y} = \frac{1}{N} \sum_{k\in S} \frac{y_k}{\pi_k}$$

où $N = 6$ est la taille de la population. On a donc :

$$\hat{Y} = \frac{1}{6}(2/(1/3) + 3/(1/2) + 9/1) = 21/6 = 7/2$$

Calculons la taille moyenne des ménages : $Y = (2 + 4 + 3 + 9 + 1 + 2)/6 = 7$ On voit donc que l'estimation de la taille moyenne donne cette taille moyenne de manière exacte. Ce résultat était prévisible puisqu'en effet la variable auxiliaire et la variable d'intérêt sont ici confondues (les probabilités d'inclusion ont été pondérées par la variable d'intérêt Y).

Exercice 12. (**Taille d'entreprise**) On veut sélectionner un échantillon de taille 4 dans une population de 8 entreprises dont on connaît la taille, mesurée en termes d'effectif salarié. L'échantillon est tiré à probabilités proportionnelles à la taille. De même que dans l'exercice précédent, nous allons volontairement faire jouer le rôle de variable auxiliaire à la variable d'intérêt afin de regarder les conséquences sur un exemple numérique.

Entreprise	1	2	3	4	5	6	7	8
Taille	300	300	150	100	50	50	25	25

1. Donner les probabilités d'inclusion d'ordre 1 des entreprises
2. Sélectionner l'échantillon selon un tirage systématique en utilisant 0.27 comme nombre aléatoire.
3. Lister les échantillons possibles que l'on peut obtenir avec un tirage systématique, et indiquer les probabilités de tirage de chacun d'eux.
4. A partir des échantillons obtenus, donner une estimation du total de l'effectif salarié des entreprises. Le résultat était-il prévisible ?
5. Calculer la matrice des probabilités d'inclusion d'ordre 2 ? Commenter.

Corrigé :

Le tableau ci-dessous rassemble les données ainsi que les valeurs des probabilités d'inclusion d'ordre 1 et les valeurs des C_k.

Entreprise	1	2	3	4	5	6	7	8	
Taille	300	300	150	100	50	50	25	25	1000
π_k	1	1	0,75	0,5	0,25	0,25	0,125	0,125	4
C_k	1	2	2,75	3,25	3,5	3,75	3,875	4	-

1. Soit y_k la taille de l'entreprise k et $n = 4$ la taille de l'échantillon. Alors la probabilité d'inclusion est égale à :

$$\pi_k = 4\frac{y_k}{\sum_{i\in U} y_i} = 4y_k/1000$$

comme pour $k = 1$ cela donne $1.2 > 1$, on pose $\pi_1 = 1$ et on recalcule pour $k \geqslant 2$ avec

$$\pi_k = 3\frac{y_k}{\sum_{i\geqslant 2} y_i} = 3y_k/700$$

On obtient alors $\pi_2 = 900/700 > 1$, que l'on modifie donc par $\pi_2 = 1$ et de nouveau pour $k \geqslant 3$

$$\pi_k = 2\frac{y_k}{\sum_{i\geqslant 3} y_i} = 2y_k/400$$

Les valeurs restantes sont dans le tableau ci-dessus.

2. Prenons 0.27 comme nombre aléatoire, de là
   — $0.27 < C_1$, la première unité sélectionnée est donc 1 .
   — $C_1 \leqslant 1.27 < C_2$, la deuxième unité sélectionnée est donc 2.
   — $C_2 \leqslant 2.27 < C_3$, la troisième unité sélectionnée est donc 3.
   — $C_4 \leqslant 3.27 < C_5$, la quatrième unité sélectionnée est donc 5 .
   L'échantillon choisi est donc constitué des individus $1, 2, 3$ et 5.
3. Au vu des valeurs des C_k, les unités 1 et 2 seront systématiquement sélectionnées pour constituer le sondage.
   — Supposons que le nombre aléatoire u tiré soit tel que $0.875 \leqslant u < 1$, alors $1.825 \leqslant u + 1 < 2; 2.875 \leqslant u + 2 < 3$ et $3.875 \leqslant u + 3 < 4$.
   ↪L'échantillon tiré est donc dans ce cas $\{1; 2; 4; 8\}$ et la probabilité de tirer cet échantillon est 0.125 qui correspond à tirer un nombre entre 0.875 et 1 selon la loi uniforme.
   — Supposons que le nombre aléatoire u tiré soit tel que $0.75 \leqslant u < 0.875$, alors $1.75 \leqslant u+1 < 1.875\,; 2.75 \leqslant u + 2 < 2.875$ et $3.75 \leqslant u + 3 < 3.875$.
   ↪ L'échantillon tiré est dans ce cas $\{1; 2; 4; 7\}$ et la probabilité de tirer cet échantillon est 0.125 correspondant à la probabilité de tirer un nombre entre 0.75 et 0.875 selon la loi uniforme.

— Supposons que le nombre aléatoire u tiré soit tel que $0.5 \leqslant u < 0.75$, alors $1.5 \leqslant u+1 < 1.75; 2.5 \leqslant u+2 < 2.75$ et $3.5 \leqslant u+3 < 3.75$.
↪ L'échantillon tiré est dans ce cas $\{1;2;3;6\}$ et la probabilité de tirer cet échantillon est 0,25 (probabilité de tirer un nombre entre 0.5 et 0.75 selon la loi uniforme).

— Supposons que le nombre aléatoire u tiré soit tel que $0.25 \leqslant u < 0.5$, alors $1.25 \leqslant u+1 < 1.5, 2.25 \leqslant u+2 < 2.5$ et $3.25 \leqslant u+3 < 3.5$. ↪ L'échantillon tiré est dans ce cas $\{1;2;3;5\}$ et la probabilité de tirer cet échantillon est 0.25 (probabilité de tirer un nombre entre 0.25 et 0.5 selon la loi uniforme).

— Supposons que le nombre aléatoire u tiré soit tel que $0 \leqslant u < 0.25$, alors $1 \leqslant u+1 < 1.25, 2 \leqslant u+2 < 2.25$ et $3 \leqslant u+3 < 3.25$.
↪ L'échantillon tiré est dans ce cas $\{1;2;3;4\}$ et la probabilité de tirer cet échantillon est 0.25 (probabilité de tirer un nombre entre 0 et 0.25 selon la loi uniforme).

Les résultats sont résumés dans le tableau suivant :

s	$p(s)$
$\{1;2;4;8\}$	0.125
$\{1;2;4;7\}$	0.125
$\{1;2;3;6\}$	0.25
$\{1;2;3;5\}$	0.25
$\{1;2;3;4\}$	0.25

4. Rappelons l'estimateur d'Horvitz-Thompson du total d'une variable Y

$$\bar{t}_Y = \sum_{k \in S} \frac{y_k}{\pi_k}$$

On obtient alors les résultats suivants :

s	$\bar{t}_Y$
$\{1;2;4;8\}$	1000
$\{1;2;4;7\}$	1000
$\{1;2;3;6\}$	1000
$\{1;2;3;5\}$	1000
$\{1;2;3;4\}$	1000

On trouve dans chaque cas une estimation exacte du total de l'effectif salarié. Ce résultat était prévisible puisque la variable ayant servi à calculer l'estimateur est la variable d'intérêt elle-même. On ne peut pas trouver meilleure variable auxiliaire. Cependant dans la réalité cette variable n'est justement pas connue sur l'ensemble de la population. Il s'agit ici d'un cas d'école. 5. À partir des résultats de la question 3., on calcule les probabilités d'inclusion d'ordre 2. Par exemple

$$\pi_{12} = 1, \pi_{14} = 0.125 + 0.125 + 0.25 = 0.5$$

On obtient de cette manière la matrice :

$$\begin{pmatrix} 1 & 1 & 0,75 & 0,5 & 0,25 & 0,25 & 0,125 & 0,125 \\ 1 & 1 & 0,75 & 0,5 & 0,25 & 0,25 & 0,125 & 0,125 \\ 0,75 & 0,75 & 0,5 & 0,25 & 0,25 & 0,25 & 0 & 0 \\ 0,5 & 0,5 & 0,25 & 0,25 & 0 & 0 & 0,125 & 0,125 \\ 0,25 & 0,25 & 0,25 & 0 & 0,25 & 0 & 0 & 0 \\ 0,25 & 0,25 & 0,25 & 0 & 0 & 0,25 & 0 & 0 \\ 0,125 & 0,125 & 0 & 0,125 & 0 & 0 & 0,125 & 0 \\ 0,125 & 0,125 & 0 & 0,125 & 0 & 0 & 0 & 0,125 \end{pmatrix}$$

Exercice 13. (**Kilomètres d'archives**) On désire estimer à l'échelle d'un canton le nombre de kilomètres linéaires d'archives stockées dans les mairies. Pour cela, on procède à un tirage de 4 communes parmi les 9 du canton, proportionnellement à leur population.

1. Calculer les probabilités d'inclusion de chaque commune, à partir des données suivantes :

Indice	1	2	3	4	5	6	7	8	9
Nom de la commune	Val le Grand	Les Gries	Les Combres	Flins	Villiers le Lac	Fortin	Montlebon	Sanzeau	Aumont
Population	1100	650	500	2300	4000	5500	1900	200	150

2. Estimer le métrage total des archives du canton à partir des résultats suivants :

Indice	2	4	5	6
Nom de la commune	Les Gries	Flins	Villiers le Lac	Fortin
Mètres d'archives	17	38	55	70

Corrigé :

1. Soit x_k la population de la commune k et $n = 4$ la taille de l'échantillon. Alors la probabilité d'inclusion est égale à :
$$\pi_k = 4\frac{x_k}{\sum_{i\in U} x_i}$$
On calcule la valeur pour l'unité ayant la population la plus importante, c'est-à-dire l'unité 6 :
$$\pi_6 = 4\frac{x_6}{\sum_{i\in U} x_i} = 22000/16300 > 1$$
On pose donc $\pi_6 = 1$ et l'unité 6 sera donc toujours dans l'échantillon utilisé puis on recalcule les probabilités d'inclusion suivant la formule :
$$\pi_k = 3\frac{x_k}{\sum_{i\in U, i\neq 6} x_i} = 3x_k/10800$$
Pour l'unité 5, on obtient :
$$\pi_5 = 3 \times 4000/10800 = 12000/10800 > 1$$
On pose donc $\pi_5 = 1$ et on recalcule les probabilités d'inclusion suivant la formule :
$$\pi_k = 2\frac{x_k}{\sum_{i\in U, i\neq 6,5} x_i} = 2x_k/6800$$
Pour l'unité 4, on obtient :
$$\pi_5 = 2 \times 2300/6800 = 4600/6800 = 23/34 < 1$$
On calcule de la même manière les probabilités d'inclusion d'ordre 1 pour les autres unités $1, 2, 3, 7, 8, 9$. Les résultats sont recensés dans le tableau ci-dessous.

Indice	Population x_k	π_k
1	1100	$11/34$
2	650	$6,5/34$
3	500	$5/34$
4	2300	$23/34$
5	4000	1
6	5500	1
7	1900	$19/34$
8	200	$2/34$
9	150	$1,5/34$
Somme	163004	

On retrouve bien le fait que la somme des probabilités d'inclusion doit être égale à n.

2. L'estimateur du métrage total des archives du canton est l'estimateur de Horvitz-Thompson :

$$\hat{t}_Y = \sum_{k \in S} \frac{y_k}{\pi_k}$$

Indice	Km	y_k/π_k
2	17	88,92
4	38	56,17
5	55	55
6	70	70
Somme	-	270,09

Une estimation du métrage total des archives du canton est donc 270.09 km .

Exercice 14. (**Amazonie profonde**) On désire estimer le nombre moyen d'individus par régions sur 5 régions en bordure d'Amazonie profonde. Le nombre d'individus n'est pas connu pour les 5 régions. En revanche, on peut survoler les régions et évaluer les surfaces cultivées de chacune des régions. Partant de l'hypothèse que le nombre d'individus d'une région est proportionnel à la surface cultivée, on décide d'estimer le nombre moyen d'individus par un sondage de deux unités à probabilités (inégales) proportionnelles à la variable auxiliaire "surface cultivée". On dispose des données suivantes : Surface cultivée des 5 régions : $(11, 4, 2, 8, 5)$, dont on peut déduire les probabilités d'inclusion suivantes : $(0.730.270.130.530.33)$ (retrouver ces valeurs en exercice d'entraînement). Le nombre d'individus a été relevé sur les deux premières régions : 250 et 80.

1. Déterminer le nombre moyen d'individus par région.
2. Que pouvez-vous proposer comme estimation de la variance de d'estimateur de la moyenne utilisé ci-dessus ?

Corrigé :

1. Rappelons l'estimateur usuel de la moyenne de la variable Y sur la population de taille $N = 5$: $\hat{Y}_S = \left(\sum_{k \in S} y_s/\pi_k\right)/N$. Cela nous donne la valeur de

$$(250/0.73 + 80/0.27)/5 = 127.75$$

2. On a vu en cours que la variance de l'estimateur du total, dans le cas de sondage de taille fixe peut être estimée par

$$\hat{\mathrm{Var}}\left(\hat{t}_Y\right) = -\frac{1}{2} \sum_{k,l \in S} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \frac{\Delta_{kl}}{\pi_{k\ell}}$$

avec $\Delta_{kl} = \pi_{k\ell} - \pi_k \pi_\ell$. Dans notre cas cela donne, pour une moyenne $\mathrm{Var}(\hat{Y}) = \mathrm{Var}\left(\hat{t}_Y\right)/N^2$ et pour $S = \{1, 2\}$

$$\hat{\mathrm{Var}}(\hat{\bar{Y}}) = -\frac{1}{2} \left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2 \frac{\Delta_{12}}{\pi_{12}}/N^2$$

nous avons besoin de calculer π_{12}. Nous avons comme valeurs de C_k

$$0.731.001.131.661.99$$

Les individus 1 et 2 seront sélectionnés pour tout u vérifiant $u < C_1 = 0.73$ et $C_1 \leqslant 1 + u < C_2 = 1$. Cela n'est pas possible, ainsi $\pi_{12} = 0$. Nous ne pouvons donc pas utiliser l'estimateur proposé qui nécessite que $\pi_{k\ell} > 0$.
