De nouvelles macros SAS d'échantillonnage équilibré

Guillaume Chauvet

Yves Tillé

La méthode du Cube

Algorithmes de tirage

Qu’est ce qu’un échantillon équilibré?

Un échantillon est dit équilibré sur une variable y s’il fournit un π-estimateur de total égal au vrai total. Autrement dit, si :

$$\sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} y_k$$

pour tout échantillon s sélectionnable

Une technique : la méthode du Cube

Mise au point par Deville et Tillé (2000) et basée sur une représentation géométrique du plan de sondage

Permet de tirer un échantillon équilibré sur un nombre quelconque de variables, à probabilités inégales

Principe de l’algorithme

Au plus N étapes de tirage ; à chacune, on décide de retenir ou d'écarter définitivement une unité **parmi toutes les unités restantes**

Complexité de l'ordre de N^2p

Principe de l'algorithme rapide

Autre implémentation de la méthode du CUBE

Au plus N étapes de tirage ; à chacune, l’algorithme décide de retenir ou d'écarter définitivement une unité **parmi les p+1 premières unités restantes**

Complexité de l'ordre de Np^2

Principe de l'algorithme rapide (suite)

Le même principe a été utilisé pour programmer la macro CUBE du Nouveau Recensement

L'implémentation proposée ici est plus rapide : on ne prend véritablement en compte que p+1 individus à chaque itération.

Mise en œuvre de l’algorithme rapide

Programmes et applications

La macro d’échantillonnage équilibré

Elle permet de tirer un échantillon équilibré, selon l'algorithme rapide de la méthode du Cube

Les trois options d'atterrissage proposées par Deville et Tillé sont disponibles.

Exemple 1 :

Sélection d'un échantillon de 1500 adresses, avec probabilités proportionnelles au nombre de logements

Base de sondage : logements de la ville de Lyon, source RP 99

Equilibrage sur des variables de type logement et socio-démographiques

Ecart obtenu entre π-estimateur et vrai total pour les variables d'équilibre

Variable	Estimateur HT	Vrai total	Ecart relatif (%)
Nb de logements	251 279	251 279	0,00%
Nb de logements collectifs	243 071	243 381	0,13%
Hommes de moins de 20 ans	46 596	46 395	-0,43%
Hommes de 20 à 39 ans	75 091	75 116	0,03%
Hommes de 40 à 59 ans	46 195	46 078	-0,25%
Hommes de 60 à 74 ans	20 733	20 726	-0,03%
Hommes de 75 ans et plus	10 495	10 435	-0,57%
Femmes de moins de 20 ans	46 196	46 156	-0,09%
Femmes de 20 à 39 ans	83 966	83 957	-0,01%
Femmes de 40 à 59 ans	51 983	51 881	-0,20%
Femmes de 60 à 74 ans	28 644	28 637	-0,02%
Femmes de 75 ans et plus	21 512	21 421	-0,42%
Actifs	206 834	206 732	-0,05%
Inactifs	224 576	224 070	-0,23%
Français de naissance	376 919	376 326	-0,16%
Français par acquisition	21 906	21 833	-0,33%
Etrangers hors UE	22 993	22 978	-0,07%
Etrangers de l'UE	9 591	9 665	0,76%

Ecart obtenu entre π-estimateur et vrai total pour les variables d'équilibre

Variable	Estimateur HT	Vrai total	Ecart relatif (%)
Nb de logements	251 279	251 279	0,00%
Nb de logements collectifs	243 071	243 381	0,13%
Hommes de moins de 20 ans	46 596	46 395	-0,43%
Hommes de 20 à 39 ans	75 091	75 116	0,03%
Hommes de 40 à 59 ans	46 195	46 078	-0,25%
Hommes de 60 à 74 ans	20 733	20 726	-0,03%
Hommes de 75 ans et plus	10 495	10 435	-0,57%
Femmes de moins de 20 ans	46 196	46 156	-0,09%
Femmes de 20 à 39 ans	83 966	83 957	-0,01%
Femmes de 40 à 59 ans	51 983	51 881	-0,20%
Femmes de 60 à 74 ans	28 644	28 637	-0,02%
Femmes de 75 ans et plus	21 512	21 421	-0,42%
Actifs	206 834	206 732	-0,05%
Inactifs	224 576	224 070	-0,23%
Français de naissance	376 919	376 326	-0,16%
Français par acquisition	21 906	21 833	-0,33%
Etrangers hors UE	22 993	22 978	-0,07%
Etrangers de l'UE	9 591	9 665	0,76%

La macro d'échantillonnage équilibré (suite)

Une nouvelle option permet de laisser le complémentaire de l'échantillon équilibré

Applications possibles :

- Tirage successif de plusieurs échantillons équilibrés (Ex : groupes de rotation du Nouveau Recensement)
- Tirage d'un échantillon complémentaire

Exemple 2 :

Sélection de 3 échantillons de 500 adresses, avec probabilités proportionnelles au nombre de logements

(même base de sondage que précedemment, variables d'équilibrage identiques)

		Echantillon 1	Echantillon 2	Echantillon 3
Variable	**Vrai total**	**Ecart relatif**	**Ecart relatif**	**Ecart relatif**
Nb de logements	251 279	0,00%	0,00%	0,00%
Nb de logements collectifs	243 381	0,06%	0,06%	0,13%
Hommes de moins de 20 ans	46 395	-0,31%	-0,33%	-0,43%
Hommes de 20 à 39 ans	75 116	0,23%	0,10%	0,03%
Hommes de 40 à 59 ans	46 078	-1,32%	-0,33%	-0,25%
Hommes de 60 à 74 ans	20 726	0,05%	0,19%	-0,03%
Hommes de 75 ans et plus	10 435	3,27%	-1,08%	-0,57%
Femmes de moins de 20 ans	46 156	-1,05%	-0,60%	-0,09%
Femmes de 20 à 39 ans	83 957	-0,13%	-0,20%	-0,01%
Femmes de 40 à 59 ans	51 881	0,25%	-0,56%	-0,20%
Femmes de 60 à 74 ans	28 637	-0,97%	0,55%	-0,02%
Femmes de 75 ans et plus	21 421	0,71%	-0,28%	-0,42%
Actifs	206 732	-0,22%	-0,08%	-0,05%
Inactifs	224 070	-0,16%	-0,34%	-0,23%
Français de naissance	376 326	0,03%	-0,49%	-0,16%
Français par acquisition	21 833	-2,76%	2,63%	-0,33%
Etrangers hors UE	22 978	-1,97%	0,69%	-0,07%
Etrangers de l’UE	9 665	1,16%	1,89%	0,76%

		Echantillon 1	Echantillon 2	Echantillon 3
Variable	**Vrai total**	**Ecart relatif**	**Ecart relatif**	**Ecart relatif**
Nb de logements	251 279	0,00%	0,00%	0,00%
Nb de logements collectifs	243 381	0,06%	0,06%	0,13%
Hommes de moins de 20 ans	46 395	-0,31%	-0,33%	-0,43%
Hommes de 20 à 39 ans	75 116	0,23%	0,10%	0,03%
Hommes de 40 à 59 ans	46 078	-1,32%	-0,33%	-0,25%
Hommes de 60 à 74 ans	20 726	0,05%	0,19%	-0,03%
Hommes de 75 ans et plus	10 435	3,27%	-1,08%	-0,57%
Femmes de moins de 20 ans	46 156	-1,05%	-0,60%	-0,09%
Femmes de 20 à 39 ans	83 957	-0,13%	-0,20%	-0,01%
Femmes de 40 à 59 ans	51 881	0,25%	-0,56%	-0,20%
Femmes de 60 à 74 ans	28 637	-0,97%	0,55%	-0,02%
Femmes de 75 ans et plus	21 421	0,71%	-0,28%	-0,42%
Actifs	206 732	-0,22%	-0,08%	-0,05%
Inactifs	224 070	-0,16%	-0,34%	-0,23%
Français de naissance	376 326	0,03%	-0,49%	-0,16%
Français par acquisition	21 833	-2,76%	2,63%	-0,33%
Etrangers hors UE	22 978	-1,97%	0,69%	-0,07%
Etrangers de l’UE	9 665	1,16%	1,89%	0,76%

La macro d’échantillonnage équilibré stratifié

Elle permet de tirer un échantillon stratifié :

- Globalement équilibré
- Approximativement équilibré dans chaque strate
- De taille fixe dans chacune des strates

Exemple 3 :

Les adresses de Lyon sont stratifiées en 36 Iris

Tirage d'un échantillon d'1/20ème des adresses, équilibré sur les nombres de logements et les variables croisées âge-sexe.

Contraintes : dans chaque strate

- Allocation proportionnelle
- Probabilités proportionnelles au nb de logements
- Equilibrage approximatif

Ecart obtenu entre π-estimateur et vrai total pour les variables d'équilibre

Variable	Estimateur HT	Vrai total	Ecart relatif
Nb de logements	251 279	251 279	0,0%
Nb de logements collectifs	243 388	243 381	0,0%
Hommes de moins de 20 ans	46 299	46 395	0,2%
Hommes de 20 à 39 ans	74 901	75 116	0,3%
Hommes de 40 à 59 ans	46 168	46 078	-0,2%
Hommes de 60 à 74 ans	20 658	20 726	0,3%
Hommes de 75 ans et plus	10 452	10 435	-0,2%
Femmes de moins de 20 ans	46 043	46 156	0,2%
Femmes de 20 à 39 ans	84 106	83 957	-0,2%
Femmes de 40 à 59 ans	51 900	51 881	0,0%
Femmes de 60 à 74 ans	28 591	28 637	0,2%
Femmes de 75 ans et plus	21 337	21 421	0,4%

Résultat :

Bon équilibrage global

La taille fixe est exactement respectée dans chaque strate

Equilibrage souvent bon au niveau des strates, mais pas toujours

Conclusion

Possibilité de tirer des échantillons équilibrés sur de grandes bases de sondage (2 heures pour une base de 300 000 unités, avec 100 variables d'équilibrage)

Possibilité de tirer des échantillons équilibrés avec une répartition par strate contrôlée

Bibliographie

- Deville, J.-C., Tillé, Y. (2004) : Efficient Balanced Sampling : the Cube method. *Biometrika*, 91: 893-912
- Deville, J.-C., Tillé, Y. (2005) : Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128 : 411-425
- Rousseau, S., Tardieu, F. (2004) : La macro SAS CUBE d'échantillonnage équilibré – *Documentation de l'utilisateur*
- Tillé, Y. (2001). Théorie des Sondages : échantillonnage et estimation en populations finies. Dunod, Paris
- Tillé, Y., Favre, A.-C. (2004) : Coordination, combination and extension of balanced samples. *Biometrika*, 91 : 913-928