Méthodes de sondage
Echantillonnage et Redressement

Guillaume Chauvet

École Nationale de la Statistique et de l'Analyse de l'Information

27 avril 2015

Panorama du cours

1. Echantillonnage en population finie
2. Méthodes d'échantillonnage
3. Méthodes de redressement

Objectifs du cours

- Présenter les méthodes d'inférence dans le cas d'une population finie d'individus.
- Donner les principales méthodes d'échantillonnage utilisées dans les enquêtes.
- Décrire les méthodes de redressement qui permettent d'utiliser une information auxiliaire au moment de l'estimation.
- Donner des exemples pratiques.

Echantillonnage en population finie

Notations

Notations

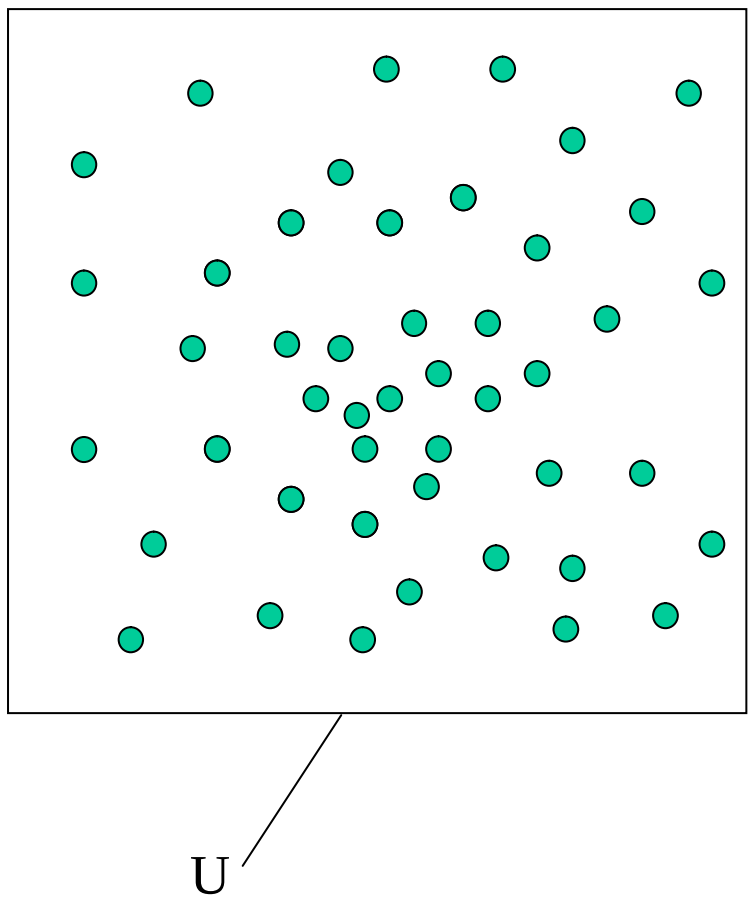
On se place dans le cadre d'une population finie U d'*individus* ou *unités statistiques*, supposées identifiables par un label. On notera simplement

$$U = \{1, \ldots, k, \ldots, N\}$$

où N désigne la taille de la population U.

On s'intéresse à une *variable d'intérêt* y (éventuellement vectorielle), qui prend la valeur y_k sur l'individu k de U. La variable y est vue ici comme non aléatoire : la population U étant fixée, **la valeur prise par y sur chaque individu est parfaitement définie et déterministe**.

On souhaite disposer d'indicateurs pour la population U (totaux, moyennes, fractiles, indices, ...).


U

Notations

Essentiellement pour des considérations pratiques, la variable d'intérêt n'est pas observée sur l'ensemble de la population :

- effectuer un *recensement* coûte cher, et suppose de disposer d'une *base de sondage* donnant la liste de l'ensemble des individus de la population,
- même dans le cas d'un recensement traditionnel, l'ensemble des données recueillies est rarement exploité,
- augmenter la taille d'un questionnaire augmente le *fardeau de réponse*, et diminue les taux de réponse,
- de façon générale, la *non-réponse* diminue la taille de l'échantillon effectivement observé.

Exemples

Exemple 1 : Les enquêtes-ménages de l'Insee visent à décrire les conditions de vie des ménages (emploi, logement, patrimoine, ...). Les ménages enquêtés sont sélectionnés dans un échantillon de zones appelé l'*Echantillon-Maître*.

Exemple 2 : Les enquêtes-entreprises sont réalisées à l'aide d'une base de sondage (répertoire SIRENE) et de sources externes.

Exemple 3 : Enquête auprès d'un échantillon de personnes pour connaître une opinion politique, les habitudes en termes de media, l'avis sur un produit ... On utilise souvent dans ce cas des *méthodes de tirage empiriques* (échantillonnage par quotas, échantillonnage de volontaires).

Paramètre d'intérêt

On s'intéresse à un *paramètre d'intérêt* de la forme

$$\theta(y_k \ , \ k \in U) \equiv \theta.$$

Un **estimateur** de ce paramètre sera de la forme

$$\begin{aligned} \hat{\theta}(y_k \ , \ k \in S) &\equiv \hat{\theta}(S) \\ &\equiv \hat{\theta}, \end{aligned}$$

où S désigne l'échantillon aléatoire finalement observé.

Paramètre d'intérêt

Total et moyenne

On peut s'intéresser au total

$$t_y = \sum_{k \in U} y_k$$

d'une variable quantitative sur la population, ou encore à sa valeur moyenne

$$\mu_y = \frac{1}{N} \sum_{k \in U} y_k.$$

Exemple :
Chiffre d'affaires total des entreprises d'un secteur d'activité, pourcentage d'étudiants fumeurs, ...

Paramètre d'intérêt

Estimation sur domaine

Un cas particulier important est celui de l'estimation sur une sous-population U_d (appelée *domaine*) d'un total

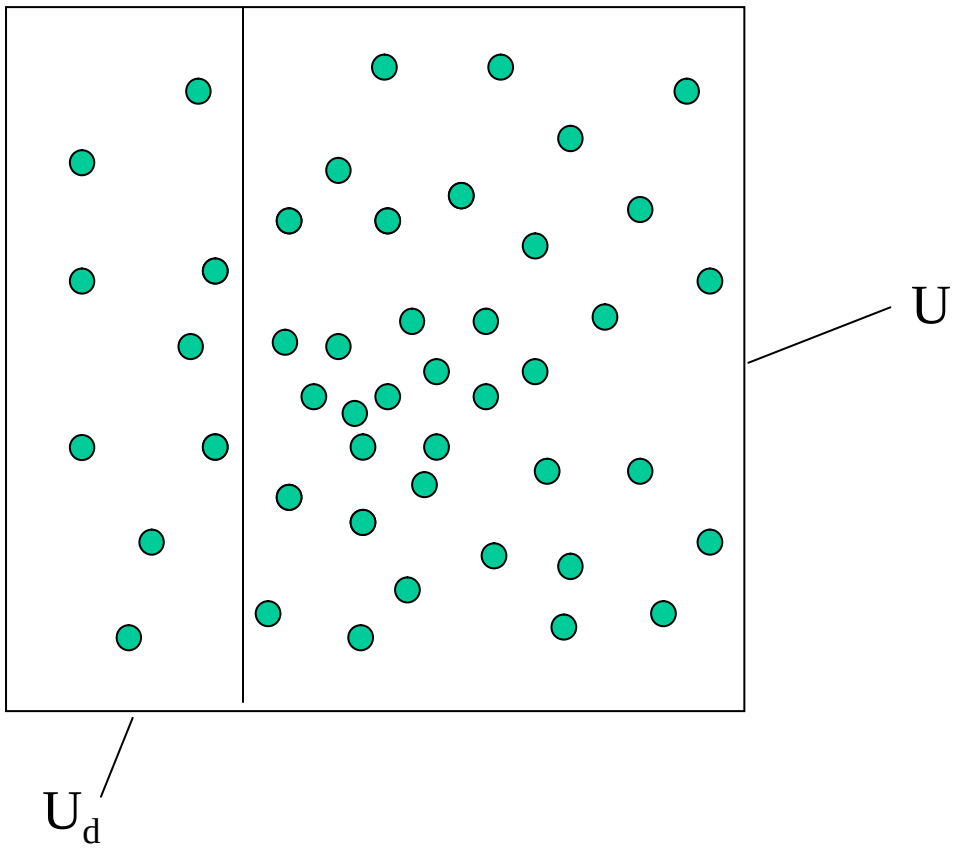
$$t_{yd} = \sum_{k \in U_d} y_k$$

ou d'une moyenne

$$\mu_{yd} = \frac{1}{N_d} \sum_{k \in U_d} y_k$$

avec N_d la taille du domaine.

Il peut s'agir d'un domaine au sens géographique (habitants d'une région), socio-démographique (individus de moins de 20 ans), temporel (individus présents à une date donnée), ...


U
U_d

Paramètre d'intérêt

Estimation par substitution

Savoir estimer un total permet de traiter le cas de très nombreux paramètres qui peuvent s'exprimer comme des fonctions de totaux. C'est le cas d'un ratio, d'un coefficient de corrélation, d'une variance, d'un coefficient de régression, ...

Ces paramètres sont estimés par substitution, en remplaçant chaque total inconnu par son estimateur.

Exemple 1 :

$$R = \frac{t_y}{t_x} \quad \text{estimé par} \quad \hat{R} = \frac{\hat{t}_y}{\hat{t}_x}.$$

Paramètre d'intérêt

Estimation par substitution

Exemple 2 :

$$OR = \frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}} \quad \text{estimé par} \quad \widehat{OR} = \frac{\frac{\hat{p}_A}{1-\hat{p}_A}}{\frac{\hat{p}_B}{1-\hat{p}_B}}.$$

Il est également possible d'estimer des paramètres plus complexes tels que des fractiles (médianes), ou des indices (Gini, utilisé comme indicateur d'inégalité).

Plan de sondage

La sélection de l'échantillon aléatoire S se fait à l'aide d'un *plan de sondage* p sur U, c'est à dire à l'aide d'une loi de probabilité sur les parties de U :

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

On note S l'échantillon aléatoire, et on distinguera

- l'*estimateur* $\hat{\theta}(y_k \ , \ k \in S) \equiv \hat{\theta}(S)$,
- l'*estimation* $\hat{\theta}(y_k \ , \ k \in s) \equiv \hat{\theta}(s)$.

On appelle *algorithme d'échantillonnage* une méthode pratique permettant de sélectionner un échantillon selon le plan de sondage choisi.

Exemple

Soit la population $U = \{1, 2, 3, 4\}$, et $p(\cdot)$ le plan de sondage défini par :

$$\begin{array}{lllllll} p(\{1,2\}) & = & 0.2 & p(\{1,4\}) & = & 0.1 & p(\{3,4\}) = 0.3 \\ p(\{1,2,3\}) & = & 0.3 & p(\{2,3,4\}) & = & 0.1 & \end{array}$$

La variable aléatoire S prend ses valeurs dans

$$\{\{1,2\},\{1,4\},\{3,4\},\{1,2,3\},\{2,3,4\}\}.$$

On a par exemple

$$\mathbb{P}(S = \{1,2\}) = p(\{1,2\}) = 0.2$$

A la différence des lois de probabilités classiques (normale, exponentielle, binomiale, ...) l'aléatoire ne porte pas sur la variable mais sur le sous-ensemble d'individus observés.

Comparaison avec une variable aléatoire réelle

Soit X une variable aléatoire distribuée selon une loi de Poisson $\mathcal{P}(\lambda)$. La variable aléatoire X prend ses valeurs dans

$$\mathbb{N} = \{0, 1, 2, \ldots\}.$$

On a pour $k \in \mathbb{N}$:

$$\mathbb{P}(X = k) \quad = \quad \exp^{-\lambda} \times \frac{\lambda^k}{k!}.$$

L'espérance de X correspond à la valeur moyenne de ses valeurs possibles, pondérées par leurs probabilités :

$$\begin{aligned} E[X] &= \sum_{k \in \mathbb{N}} k \times \mathbb{P}(X = k) \\ &= \lambda. \end{aligned}$$

Mesures de précision

L'espérance d'un estimateur $\hat{\theta}(S)$ se définit de façon analogue :

$$\begin{aligned} E_p\left[\hat{\theta}(S)\right] &= \sum_{s \subset U} \hat{\theta}(s) \times \mathbb{P}(S = s) \\ &= \sum_{s \subset U} p(s)\ \hat{\theta}(s). \end{aligned}$$

Le biais d'un estimateur $\hat{\theta}(S)$ correspond à son erreur moyenne :

$$\begin{aligned} B_p\left[\hat{\theta}(S)\right] &= E_p\left[\hat{\theta}(S) - \theta\right] \\ &= \sum_{s \subset U} p(s) \left[\hat{\theta}(s) - \theta\right]. \end{aligned}$$

Mesures de précision

On s'intéressera également à la Variance

$$\begin{aligned} V_p\left[\hat{\theta}(S)\right] &= E_p\left[\left\{\hat{\theta}(S) - E_p[\hat{\theta}(S)]\right\}^2\right] \\ &= \sum_{s \subset U} p(s)\left\{\hat{\theta}(s) - E_p[\hat{\theta}(S)]\right\}^2, \end{aligned}$$

et à l'Erreur Quadratique Moyenne (EQM)

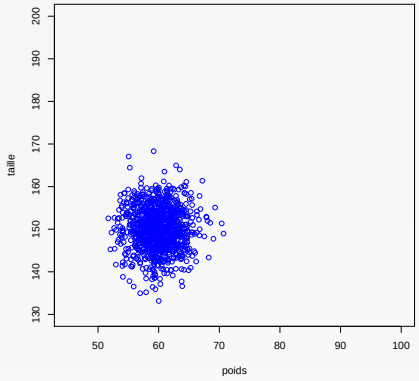
$$\begin{aligned} EQM_p\left[\hat{\theta}(S)\right] &= E_p\left[\left\{\hat{\theta}(S) - \theta\right\}^2\right] \\ &= B_p\left[\hat{\theta}(S)\right]^2 + V_p\left[\hat{\theta}(S)\right]. \end{aligned}$$

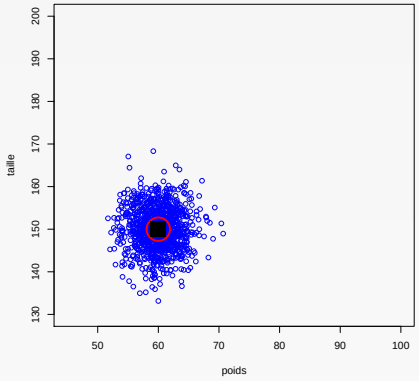
Quelques simulations

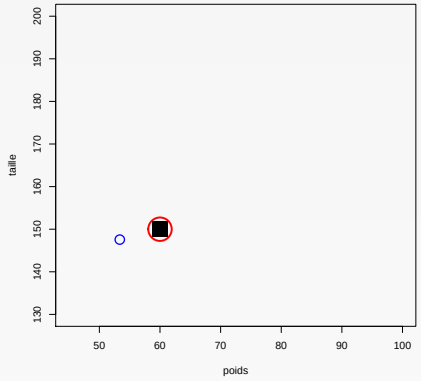
Pour illustrer la notion de biais et de variance, on considère l'exemple d'une population de $N = 1\ 000$ individus âgés de 15 à 20 ans.

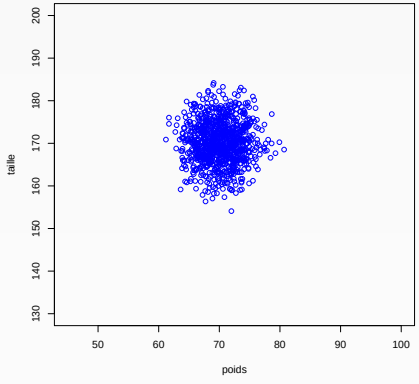
Dans cette population, un échantillon de taille $n = 50$ est sélectionné et enquêté. Pour chaque individu enquêté, on obtient son poids (en kg), sa taille (en cm) et son âge.

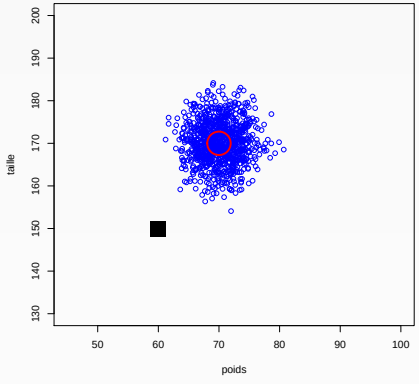
On s'intéresse à l'estimation du poids moyen et de la taille moyenne (carré noir). Chaque échantillon permet d'obtenir une estimation (points bleus) de ces paramètres. La moyenne des estimations est représentée par le point rouge.

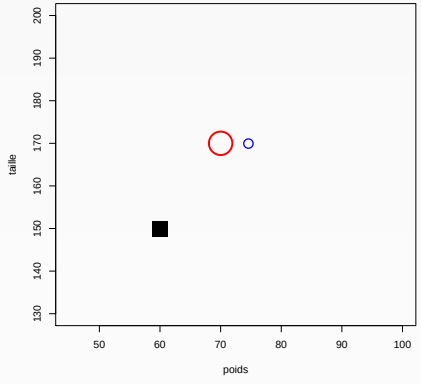

taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100

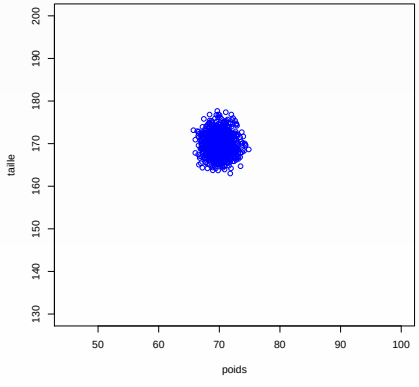

taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100

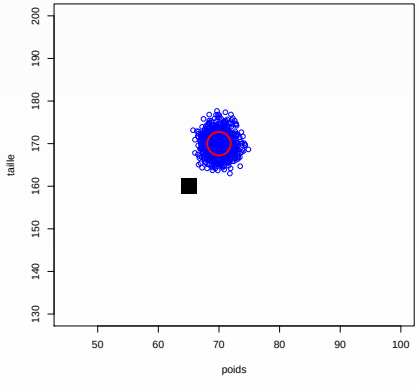

taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100

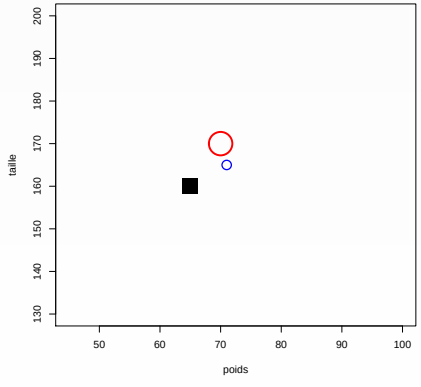

taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100

Probabilités d'inclusion d'ordre 1

On note π_k la *probabilité d'inclusion* de l'unité k, c'est à dire la probabilité que l'unité k soit retenue dans l'échantillon :

$$\pi_k = \mathbb{P}(k \in S) = \sum_{s/k \in s} p(s)$$

La somme des probabilités d'inclusion donne la taille moyenne de l'échantillon sélectionné :

$$\sum_{k \in U} \pi_k = E_p \left[n(S) \right].$$

En pratique, les probabilités d'inclusion π_k sont fixées avant le tirage à l'aide d'une **information auxiliaire**. On utilise ensuite un plan de sondage qui respecte ces probabilités d'inclusion.

Probabilités d'inclusion d'ordre 2

On note π_{kl} la probabilité que deux unités distinctes k et l soient sélectionnées conjointement dans l'échantillon :

$$\pi_{kl} = \mathbb{P}(k, l \in S) = \sum_{s/k,l \in s} p(s)$$

Ces probabilités doubles interviennent notamment dans la variance des estimateurs. Il est souvent difficile de les calculer exactement, sauf pour des plans de sondage particuliers.

Application

Soit la population $U = \{1, 2, 3, 4\}$, et $p(\cdot)$ le plan de sondage défini par :

$$\begin{array}{lllllll} p(\{1,2\}) & = & 0.2 & p(\{1,4\}) & = & 0.1 & p(\{3,4\}) = 0.3 \\ p(\{1,2,3\}) & = & 0.3 & p(\{2,3,4\}) & = & 0.1 & \end{array}$$

Calculer les probabilités d'inclusion d'ordre 1, et donner la taille moyenne d'échantillon obtenue à l'aide de ce plan de sondage.

Donner les probabilités d'inclusion d'ordre 2 :

- des unités 1 et 2,
- des unités 1 et 4,
- des unités 2 et 4.

Variables indicatrices

L'utilisation de la variable $I_k = 1(k \in S)$, indiquant l'appartenance à l'échantillon de l'unité k, permet souvent de simplifier les calculs.

Pour deux unités k et l distinctes, on a notamment les propriétés suivantes :

$$\begin{aligned} E_p(I_k) &= \pi_k, \\ V_p(I_k) &= \pi_k(1-\pi_k), \\ Cov_p(I_k, I_l) &= \pi_{kl} - \pi_k\pi_l \\ &\equiv \Delta_{kl}. \end{aligned}$$

On note $\Delta = [\Delta_{kl}]_{k,l \in U}$ la matrice de variance-covariance du plan de sondage $p(\cdot)$.

En résumé

Un plan de sondage est une loi de probabilité sur les parties de U. L'alea porte sur le sous-ensemble S d'individus observés.

Les notions d'espérance et de de variance d'un estimateur $\hat{\theta}(S)$ s'adaptent de façon naturelle :

$$\begin{aligned} B_p\left[\hat{\theta}(S)\right] &= \sum_{s\subset U} p(s)\left[\hat{\theta}(s)-\theta\right], \\ V_p\left[\hat{\theta}(S)\right] &= \sum_{s\subset U} p(s)\left\{\hat{\theta}(s)-E_p[\hat{\theta}(S)]\right\}^2. \end{aligned}$$

On appelle probabilités d'inclusion d'ordre 1 et 2 :

$$\begin{aligned} \pi_k &= \mathbb{P}(k\in S), \\ \pi_{kl} &= \mathbb{P}(k,l\in S). \end{aligned}$$

Estimation de Horvitz-Thompson

Objectif

Nous nous intéressons essentiellement, dans la suite de ce cours, à l'estimation du total

$$t_y = \sum_{k \in U} y_k$$

de la variable y, et d'une moyenne

$$\mu_y = \frac{1}{N} \sum_{k \in U} y_k.$$

La π-estimation

La connaissance des probabilités π_k permet une estimation sans biais d'un total sous le plan de sondage, i.e. sous le mécanisme aléatoire associé au plan de sondage. Le total t_y est estimé sans biais par

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k \tag{1}$$

si tous les π_k sont > 0. On parle d'estimateur de Horvitz-Thompson ou encore de π-estimateur.

C'est un estimateur pondéré, où les *poids de sondage* $d_k = 1/\pi_k$ ne dépendent pas de la variable d'intérêt.

Principe : un individu k de l'échantillon représente $d_k = 1/\pi_k$ individus de la population.

Biais de couverture

Si certaines probabilités d'inclusion sont nulles, le π-estimateur est biaisé :

$$\begin{aligned} E\left[\hat{t}_{y\pi}\right] &= \sum_{\substack{k\in U \\ \pi_k>0}} y_k \\ &= t_y - \sum_{\substack{k\in U \\ \pi_k=0}} y_k. \end{aligned}$$

Ce problème peut notamment se poser :

- en cas de défaut de couverture de la base de sondage (liste des individus pas à jour, ou individus impossibles à joindre),
- quand on choisit de laisser de côté une partie de la population (cut-off sampling, parfois utilisé dans les enquêtes-entreprise).

Enquête "Sans-Domicile 2001" (De Peretti et al., 2006)

Sans-domicile : personne qui dort dans un lieu non prévu pour l'habitation ou prise en charge par un organisme fournissant un hébergement gratuit ou à faible participation.

Méthode d'échantillonnage indirect : sélection d'un échantillon de jours $\times$ services d'aide (hébergement, restauration).
Champ de l'enquête : sans-domicile ayant fréquenté, au moins une fois dans la semaine d'enquête, soit un service d'hébergement, soit une distribution de repas chauds.

Exclut les personnes :

- qui dorment dans la rue pour une période de temps courte et ne font pas appel à un centre ou à une distribution de repas,
- qui ne font pas (ou ne peuvent pas faire) appel au circuit d'assistance.

Variance

La variance du π-estimateur est donnée par

$$V_p\left[\hat{t}_{y\pi}\right] = \sum_{k,l\in U} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\Delta_{kl}. \tag{2}$$

Cette variance peut être estimée sans biais par

$$v_{HT}\left[\hat{t}_{y\pi}\right] = \sum_{k,l\in S} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}} \tag{3}$$

si tous les π_{kl} sont strictement positifs. On parle de l'estimateur de variance de Horvitz-Thompson.

Principe : un couple (k,l) d'individus de l'échantillon représente $1/\pi_{kl}$ couples de la population.

Définitions

Définition

*Un plan de sondage $p(\cdot)$ est dit **de taille fixe**, égale à n, si seuls les échantillons de taille n ont une probabilité non nulle d'être tirés :*

$$Card(s) \neq n \Rightarrow p(s) = 0.$$

Définition

*Un plan de sondage $p(\cdot)$ est dit **simple** si deux échantillons de même taille ont la même probabilité d'être sélectionnés :*

$$Card(s_1) = Card(s_2) \Rightarrow p(s_1) = p(s_2).$$

Exemples

Soit la population $U = \{1, 2, 3, 4\}$.

Exemple 1 :

$$p(\{1,2\}) = 0.2 \quad p(\{1,4\}) = 0.1 \quad p(\{3,4\}) = 0.3$$
$$p(\{1,2,3\}) = 0.3 \quad p(\{2,3,4\}) = 0.1$$

Exemple 2 :

$$p(\{1,2\}) = 1/3 \quad p(\{1,4\}) = 1/3 \quad p(\{3,4\}) = 1/3$$

Exemple 3 :

$$p(\{1,2,3\}) = 1/4 \quad p(\{1,2,4\}) = 1/4 \quad p(\{1,3,4\}) = 1/4$$
$$p(\{2,3,4\}) = 1/4$$

Variance

Pour un plan de taille fixe, la variance du π-estimateur peut se réécrire sous la forme

$$V_p\left[\hat{t}_{y\pi}\right] = -\frac{1}{2}\sum_{k\neq l\in U}\left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right]^2 \Delta_{kl}. \tag{4}$$

Cette variance peut être estimée sans biais par

$$v_{YG}\left[\hat{t}_{y\pi}\right] = -\frac{1}{2}\sum_{k\neq l\in S}\left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right]^2 \frac{\Delta_{kl}}{\pi_{kl}} \tag{5}$$

si tous les π_{kl} sont strictement positifs. On parle de l'estimateur de variance de Yates-Grundy.

Si le plan de sondage vérifie les **conditions de Yates-Grundy** :
$\forall k \neq l \in U \quad \Delta_{kl} \leq 0$, cet estimateur de variance est toujours positif.

Biais de l'estimateur de variance

Proposition

Pour un plan de sondage quelconque, on a :

$$E_p\left\{v_{HT}\left[\hat{t}_{y\pi}\right]\right\} = V_p\left[\hat{t}_{y\pi}\right] + \sum_{\substack{k,l\in U\\ \pi_{kl}=0}} y_k y_l.$$

Pour un plan de sondage de taille fixe, on a :

$$E_p\left\{v_{YG}\left[\hat{t}_{y\pi}\right]\right\} = V_p\left[\hat{t}_{y\pi}\right] - \frac{1}{2}\sum_{\substack{k,l\in U\\ \pi_{kl}=0}} \pi_k\pi_l\left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right]^2.$$

Si y est à valeurs positives, les deux estimateurs de variance sont respectivement biaisés positivement et négativement.

Choix des probabilités d'inclusion

D'après la formule de Yates-Grundy, la variance est nulle si les probabilités d'inclusion sont proportionnelles à la variable d'intérêt. En pratique, ce choix n'est pas possible car :

- une enquête comporte généralement de nombreuses variables d'intérêt,
- ces variables sont inconnues au stade de l'échantillonnage.

On peut définir ces probabilités d'inclusion proportionnellement à une mesure de taille.

Interprétation : si les individus peuvent être de tailles très différentes, on utilise les probabilités d'inclusion pour lisser les rapports y_k/π_k.

Probabilités proportionnelles à la taille

La taille moyenne d'échantillon sélectionné est donnée par

$$E_p\left[n(S)\right] = \sum_{k \in U} \pi_k.$$

Si n désigne la taille d'échantillon souhaitée, les probabilités d'inclusion proportionnelles à une variable auxiliaire positive x sont données par

$$\pi_k = n \frac{x_k}{\sum_{l \in U} x_l}.$$

La variable x_k doit être connue avant le tirage pour chaque individu k de U.

Recalcul des probabilités d'inclusion

Si certaines unités sont particulièrement grosses (au sens de la variable auxiliaire x), on peut obtenir des probabilités d'inclusion supérieures à 1. Dans ce cas, on sélectionne d'office les unités correspondantes, et on recalcule les probabilités d'inclusion des autres unités.

Exemple : population de $N = 6$ entreprises dont on connait le nombre d'employés

Unité	1	2	3	4	5	6
x	200	80	50	50	10	10

Donner les probabilités d'inclusion correspondant à un tirage de taille 4, à probabilités proportionnelles au nombre d'employés.

Intervalle de confiance

Intervalle de confiance

On suppose que $\hat{t}_{y\pi}$ estime sans biais t_y. Alors un intervalle de confiance pour t_y de niveau approximatif $1-\alpha$ est donné par :

$$IC_{1-\alpha}\left[t_y\right] = \left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{V_p\left[\hat{t}_{y\pi}\right]}\right],$$

avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi normale centrée réduite $\mathcal{N}(0,1)$.

Rappel :

- $\alpha = 0.05 \Rightarrow z_{0.975} = 1.96$
- $\alpha = 0.10 \Rightarrow z_{0.95} = 1.64$

Interprétation (pour $\alpha = 0.05$) : le vrai total t_y est contenu dans l'intervalle de confiance pour (approximativement) 95% des échantillons.

Intervalle de confiance

Comme la vraie variance $V_p\left[\hat{t}_{y\pi}\right]$ est généralement inconnue, on la remplace par un estimateur noté $v\left[\hat{t}_{y\pi}\right]$.

On obtient l'intervalle de confiance estimé :

$$\widehat{IC}_{1-\alpha}\left[t_y\right] = \left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}}\sqrt{v\left[\hat{t}_{y\pi}\right]}\right].$$

L'intervalle de confiance est (approximativement) valide :

- si l'estimateur $\hat{t}_{y\pi}$ suit approximativement une loi $\mathcal{N}\left[t_y, V_p\left(\hat{t}_{y\pi}\right)\right]$,
- si l'estimateur de variance $v\left(\hat{t}_{y\pi}\right)$ est faiblement consistant.

Coefficient de variation

La précision de l'estimation du total peut également être donnée sous la forme du coefficient de variation

$$CV_p\left[\hat{t}_{y\pi}\right] = \frac{\sqrt{V_p\left(\hat{t}_{y\pi}\right)}}{t_y} \quad \text{estimé par} \quad \hat{CV}\left[\hat{t}_{y\pi}\right] = \frac{\sqrt{v\left(\hat{t}_{y\pi}\right)}}{\hat{t}_{y\pi}}.$$

Il s'agit d'une grandeur sans dimension, plus facile à comparer et à interpréter que la variance. Avec un niveau de confiance de 0.95, l'intervalle de confiance du total est donné par

$$\begin{aligned} \widehat{IC}_{0.95}\left[t_y\right] &= \left[\hat{t}_{y\pi} \pm 1.96 \sqrt{v\left[\hat{t}_{y\pi}\right]}\right] \\ &= \hat{t}_{y\pi}\left[1 \pm 1.96\, \hat{CV}\left[\hat{t}_{y\pi}\right]\right]. \end{aligned}$$

Interprétation : un CV de $x\%$ correspond à un total connu à plus ou moins $2\ x\%$, avec un niveau de confiance de 0.95.

En résumé

La connaissance des probabilités d'inclusion d'ordre 1 permet de calculer l'estimateur de Horvitz-Thompson du total

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Pour un plan de sondage quelconque, sa variance est estimée sans biais par

$$v_{HT}\left[\hat{t}_{y\pi}\right] = \sum_{k,l \in S} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}}$$

si tous les π_{kl} sont strictement positifs.

En utilisant une approximation normale pour $\hat{t}_{y\pi}$, on obtient l'intervalle de confiance

$$\left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{v\left[\hat{t}_{y\pi}\right]}\right] \quad \text{où} \quad v(\hat{t}_{y\pi}) \equiv \begin{cases} v_{HT}(\hat{t}_{y\pi}) & \text{pds quelconque,} \\ v_{YG}(\hat{t}_{y\pi}) & \text{pds de taille fixe.} \end{cases}$$

Estimation d'une fonction de totaux

Estimateur par substitution

On s'intéresse à un paramètre de la forme $\theta = f(t_{\mathbf{y}})$ avec $\mathbf{y}_k = (y_{1k}, \ldots, y_{qk})^T$ un q-vecteur de variables d'intérêt, et $f : \mathbf{R}^q \to \mathbf{R}$.

Il est naturel d'estimer θ en remplaçant le total $t_{\mathbf{y}}$ inconnu par son π-estimateur. On obtient l'**estimateur par substitution** :

$$\hat{\theta}_\pi = f(\hat{t}_{\mathbf{y}\pi}).$$

Si la fonction $f(\cdot)$ est différentiable au voisinage de $t_{\mathbf{y}}$, on obtient :

$$\begin{aligned} \hat{\theta}_\pi - \theta &\simeq \left[f'(t_{\mathbf{y}})\right]^T \left[\hat{t}_{\mathbf{y}\pi} - t_{\mathbf{y}}\right] \\ &= \hat{t}_{u\pi} - t_u, \end{aligned} \tag{6}$$

en notant $u_k = \left[f'(t_{\mathbf{y}})\right]^T [\mathbf{y}_k]$.

Technique de linéarisation

Sous l'approximation (6), on a

$$\begin{aligned} E_p\left[\hat{\theta}_\pi - \theta\right] &\simeq 0, \\ V_p\left[\hat{\theta}_\pi - \theta\right] &\simeq V_p\left[\hat{t}_{u\pi}\right]. \end{aligned}$$

On parle de l'approximation de variance par linéarisation pour $\hat{\theta}_\pi$, avec

$$u_k \equiv u_k(\theta) = \left[f'(t_{\mathbf{y}})\right]^T \left[\mathbf{y}_k\right]$$

la **variable linéarisée** du paramètre θ.

Estimation de variance

Pour un plan de sondage quelconque, on obtient :

$$\begin{aligned} V_p\left[\hat{\theta}_\pi\right] &\simeq V_p\left[\hat{t}_{u\pi}\right] \\ &= \sum_{k,l\in U} \frac{u_k}{\pi_k}\frac{u_l}{\pi_l}\Delta_{kl}. \end{aligned}$$

Pour passer à un *estimateur de variance* :

1. on remplace la formule de variance par l'estimateur de variance correspondant au pds utilisé,
2. on remplace dans u_k les paramètres inconnus par des estimateurs $\Rightarrow$ variable linéarisée estimée $\hat{u}_k$.

On obtient finalement :

$$v\left[\hat{\theta}_\pi\right] = \sum_{k,l\in S} \frac{\hat{u}_k}{\pi_k}\frac{\hat{u}_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}.$$

Application : estimation d'un ratio

Paramètre $R = \frac{t_y}{t_x}$, estimé par substitution par $\hat{R}_\pi = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}$.

Calcul de la variable linéarisée :

$$f(x_1, x_2) = \frac{x_1}{x_2} \qquad f'(x_1, x_2) = \left(\frac{1}{x_2}, -\frac{x_1}{(x_2)^2} \right)$$

$$u_k(R) = \frac{1}{t_x} y_k - \frac{t_y}{(t_x)^2} x_k = \frac{1}{t_x}(y_k - R x_k)$$

$$\hat{u}_k(R) = \frac{1}{\hat{t}_{x\pi}}(y_k - \hat{R}_\pi x_k)$$

Calcul de variance :

$$V_p\left[\hat{\theta}_\pi\right] \simeq \sum_{k,l \in U} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l} \Delta_{kl},$$

$$v\left[\hat{\theta}_\pi\right] = \sum_{k,l \in S} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}}.$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SRS. Estimation des totaux t_x et t_y.

k	x_k	y_k	
1	5	1	
2	1	3	
3	4	2	
4	8	10	
	$\bar{x} = 4.5$ $s_x^2 = 8.3$	$\bar{y} = 4$ $s_y^2 = 16.7$	

$$\hat{t}_{x\pi} = N\bar{x} = 45 \qquad v\left[\hat{t}_{x\pi}\right] = N^2 \frac{1-f}{n} s_x^2 = 125$$
$$\hat{t}_{y\pi} = N\bar{y} = 40 \qquad v\left[\hat{t}_{y\pi}\right] = N^2 \frac{1-f}{n} s_y^2 = 250$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SRS. Estimation du ratio t_y/t_x.

k	x_k	y_k	$\hat{u}_k = \frac{1}{\hat{t}_{x\pi}}(y_k - \hat{R}x_k)$
1	5	1	-0.08
2	1	3	0.05
3	4	2	-0.03
4	8	10	0.06
	$\bar{x} = 4.5$	$\bar{y} = 4$	$\bar{\hat{u}} = 0$
	$s_x^2 = 8.3$	$s_y^2 = 16.7$	$s_{\hat{u}}^2 = 4.4\ 10^{-3}$

$$\hat{t}_{x\pi} = N\bar{x} = 45 \qquad v\left[\hat{t}_{x\pi}\right] = N^2\frac{1-f}{n}s_x^2 = 125$$
$$\hat{t}_{y\pi} = N\bar{y} = 40 \qquad v\left[\hat{t}_{y\pi}\right] = N^2\frac{1-f}{n}s_y^2 = 250$$
$$\hat{R} = \frac{\bar{y}}{\bar{x}} = 0.89 \qquad v\left[\hat{R}\right] = N^2\frac{1-f}{n}s_{\hat{u}}^2 = 0.07$$

Méthodes d'échantillonnage

Le tirage de Bernoulli

Principe

On se donne une même probabilité d'inclusion $\pi_k \equiv \pi$ pour chaque unité de la population. Le choix se fait indépendamment d'une unité à l'autre :

- Etape 1 : on génère $u_1 \sim U[0,1]$. Si $u_1 \leq \pi$, l'unité 1 est retenue dans l'échantillon.
- Etape 2 : on génère $u_2 \sim U[0,1]$ indépendamment de u_1. Si $u_2 \leq \pi$, l'unité 2 est retenue dans l'échantillon.
- ...
- Etape N : on génère $u_N \sim U[0,1]$ indépendamment de $u_1, \ldots, u_{N-1}$. Si $u_N \leq \pi$, l'unité N est retenue dans l'échantillon.

C'est un principe de piles ou faces indépendants, avec une même pièce mais un lancer différent pour chaque unité.

Estimateur de Horvitz-Thompson

En utilisant les propriétés d'une loi $U([0,1])$, on a :

$$\mathbb{P}(k \in S) = \mathbb{P}(u_k \leq \pi) = F_U(\pi) = \pi.$$

Les probabilités d'inclusion souhaitées sont donc bien respectées, et le total t_y est estimé sans biais par

$$\hat{t}_{y\pi} = \frac{1}{\pi} \sum_{k \in S} y_k.$$

Du fait de l'indépendance dans la sélection des unités :

$$\begin{aligned} \pi_{kl} &= \pi^2 \text{ pour } k \neq l, \\ V_p\left(\hat{t}_{y\pi}\right) &= \frac{1-\pi}{\pi} \sum_{k \in U} y_k^2. \end{aligned}$$

D'autre part, la taille d'échantillon $n(S)$ est aléatoire et suit une loi $B(N, \pi)$.

Application

Dans la population ci-dessous, utiliser les nombres aléatoires pour sélectionner un échantillon selon un tirage de Bernoulli, et en déduire une estimation de t_y.

Unité	1	2	3	4	5	6	7	8
π_k	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
u_k	0.07	0.44	0.52	0.19	0.95	0.24	0.54	0.07
y_k	0	0	1	2	2	2	2	4
Tirage								

Quelle est la taille moyenne d'échantillon attendue ? Quelle est la taille d'échantillon effectivement obtenue ?
Comparer $\hat{t}_{y\pi}$ et t_y, et commenter la différence observée.

Estimateur d'une moyenne

Si la taille de la population N est connue, on peut choisir entre les estimateurs

$$\begin{aligned}\hat{\mu}_{y\pi} &= \frac{\hat{t}_{y\pi}}{N} = \frac{1}{E[n(S)]}\sum_{k\in S} y_k,\\ \tilde{\mu}_y &= \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} = \frac{1}{n(S)}\sum_{k\in S} y_k.\end{aligned}$$

On peut montrer que ces deux estimateurs sont non biaisés pour t_y, mais que l'**estimateur par substitution** $\tilde{\mu}_y$ est généralement préférable en termes de variance :

$$\begin{aligned}V_p\left(\hat{\mu}_{y\pi}\right) &= \left(\frac{1}{n}-\frac{1}{N}\right)\times\frac{1}{N}\sum_{k\in U} y_k^2,\\ V_p\left(\tilde{\mu}_y\right) &\simeq \left(\frac{1}{n}-\frac{1}{N}\right)\times\frac{1}{N}\sum_{k\in U}(y_k-\mu_y)^2.\end{aligned}$$

Sondage aléatoire simple sans remise

Sondage aléatoire simple sans remise

Définition-propriété

Il existe un unique plan de sondage $p(\cdot)$ vérifiant les propriétés :

1. *$p(\cdot)$ est un plan simple,*
2. *$p(\cdot)$ est un plan de taille fixe n.*

On l'appelle ***plan de sondage aléatoire simple sans remise*** *SRS de taille n dans $U \equiv SRS(U; n)$.*

Il s'agit donc du plan qui donne la même probabilité à tous les échantillons de taille n d'être sélectionnés. On a :

$$p(s) = \begin{cases} 1/C_N^n & \text{si } n(s) = n, \\ 0 & \text{sinon.} \end{cases}$$

Estimateur de Horvitz-Thompson

Proposition

Soient k et l deux unités distinctes quelconques. Alors :

$$\pi_k = \frac{n}{N}, \qquad \pi_{kl} = \frac{n(n-1)}{N(N-1)}.$$

Le π-estimateur du total peut donc se réécrire sous la forme

$$\begin{aligned} \hat{t}_{y\pi} &= \frac{N}{n} \sum_{k \in S} y_k \\ &= N \ \bar{y}. \end{aligned}$$

Variance du π-estimateur

La variance du π-estimateur s'obtient à partir de la formule de Sen-Yates-Grundy :

$$V_p[\hat{t}_{y\pi}] = N^2 \frac{1-f}{n} S_y^2 \quad \text{avec} \quad S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \mu_y)^2.$$

On l'estime sans biais par

$$v_{SRS}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 \quad \text{avec} \quad s_y^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - \bar{y})^2.$$

On note $f = n/N$ le **taux de sondage**.

Variance de la moyenne estimée

Par linéarité, la moyenne μ_y peut être estimée sans biais par

$$\bar{y} = \frac{1}{n} \sum_{k \in S} y_k.$$

La variance de cet estimateur est donnée par

$$V_p[\bar{y}] = \frac{1-f}{n} S_y^2. \tag{7}$$

Remarques :

- Le facteur $(1-f)$ donne le gain de variance dû au tirage sans remise. On l'appelle *correction de population finie*. Ce gain peut être très important (cas des enquêtes-entreprise).
- Si le taux de sondage est faible, la variance ne dépend que de la taille d'échantillon n.

A retenir

Dans un sondage aléatoire simple sans remise (SRS) :

1. la moyenne simple dans l'échantillon estime sans biais la moyenne simple dans la population,
2. la dispersion calculée sur l'échantillon estime sans biais la dispersion dans la population.

Dans une enquête avec un faible taux de sondage, la variance est (approximativement) inversement proportionnelle à la taille d'échantillon.

Cas d'une proportion

Dans le cas particulier où le paramètre d'intérêt est une proportion notée P, la variable d'intérêt y est une variable indicatrice dont on cherche à estimer la moyenne.

Exemple : proportion d'étudiants portant des lunettes dans la promotion,

$$y_k = \begin{cases} 1 & \text{si l'étudiant } k \text{ porte des lunettes,} \\ 0 & \text{sinon.} \end{cases}$$

En particulier, le paramètre peut s'écrire sous la forme

$$P = \frac{1}{N} \sum_{k \in U} y_k,$$

et être estimé par

$$\hat{P} = \frac{1}{n} \sum_{k \in S} y_k.$$

Proposition

Dans le cas d'une variable indicatrice (0/1) y, on a :

$$\begin{aligned} S_y^2 &= \frac{N}{N-1} P(1-P), \\ s_y^2 &= \frac{n}{n-1} \hat{P}(1-\hat{P}). \end{aligned}$$

La variance de l'estimateur de la moyenne $\hat{P}$ peut alors se réécrire

$$V_p[\hat{P}] = \frac{1-f}{n} \frac{N}{N-1} P(1-P),$$

et être estimée sans biais par

$$v[\hat{P}] = \frac{1-f}{n-1} \hat{P}(1-\hat{P}).$$

Application : détermination de taille d'échantillon

On cherche une taille d'échantillon minimale permettant de respecter avec un niveau de confiance fixé (par exemple de 95 %) une contrainte de précision en termes :

1. soit d'*erreur absolue* :
   P connu à plus ou moins $0.02 \Leftrightarrow |\hat{P} - P| \leq 0.02$.
2. soit d'*erreur relative* :
   P connu à 8 % près $\Leftrightarrow |\frac{\hat{P}-P}{P}| \leq 0.08$.

Application : détermination de taille d'échantillon

Erreur absolue

Avec un niveau de confiance de 95 % la contrainte de précision peut se réécrire :

$$
\begin{aligned}
|\hat{P} - P| \leq \beta \quad &\Leftrightarrow \quad 1.96\sqrt{V_p(\hat{P})} \leq \beta \\
&\Leftrightarrow \quad 1.96\sqrt{\left[\frac{1}{n} - \frac{1}{N}\right]\frac{N}{N-1}P(1-P)} \leq \beta \\
&\Leftrightarrow \quad n \geq \frac{1}{\frac{1}{N} + \frac{N-1}{N}\left[\frac{\beta}{1.96}\right]^2 \frac{1}{P(1-P)}}.
\end{aligned}
$$

On peut toujours se placer dans le pire des cas en prenant $P = 0.5$, mais il est préférable de disposer d'un a priori (même vague) sur le paramètre P.

Application : détermination de taille d'échantillon

Erreur relative

Avec un niveau de confiance de 95 % la contrainte de précision peut se réécrire :

$$\left|\frac{\hat{P}-P}{P}\right| \leq \gamma \quad \Leftrightarrow \quad 1.96\ CV_p(\hat{P}) \leq \gamma$$

$$\Leftrightarrow \quad 1.96\sqrt{\left[\frac{1}{n}-\frac{1}{N}\right]\frac{N}{N-1}\frac{1-P}{P}} \leq \gamma$$

$$\Leftrightarrow \quad n \geq \frac{1}{\frac{1}{N}+\frac{N-1}{N}\left[\frac{\gamma}{1.96}\right]^2\frac{P}{1-P}}.$$

Calculer cette borne nécessite de disposer d'un a priori sur le paramètre P, ou au moins d'un majorant pour ce paramètre.

Application

Parmi les 350 étudiants de l'Ensai, on veut estimer la proportion qui portent des lunettes. Quelle taille d'échantillon faut-il sélectionner pour que cette proportion soit estimée à 10% près, avec un niveau de confiance de 0.95 :

1. en utilisant l'information suivante : 50% des personnes de la population française portent des lunettes ;
2. en utilisant maintenant l'information suivante : 20% des $15-25$ ans portent des lunettes.

Algorithmes de sélection

Algorithme 1 Méthode de sélection draw by draw

1. Pour $k = 1, \ldots, n$, sélectionner une unité dans U à probabilités égales parmi les unités qui n'ont pas déjà été tirées.

Inconvénient : méthode lente, qui nécessite n lectures de fichier.

Algorithme 2 Méthode du tri aléatoire

1. On attribue un nombre aléatoire $u_k \sim U[0,1]$ à chaque unité $k \in U$.
2. On trie la population selon les u_k croissants (ou décroissants).
3. L'échantillon est constitué des n premiers individus de la population triée.

Inconvénient : nécessite un tri du fichier.

Algorithmes de sélection

Algorithme 3 Méthode de sélection-rejet

1. On initialise $j = 0$.
2. Pour $k = 1, \ldots, N$, faire :
   - Avec une probabilité $\dfrac{n-j}{N-(k-1)}$, on sélectionne l'unité k et $j = j+1$.

Avantage : nécessite une seule lecture de fichier.

Algorithme 4 Méthode du réservoir

1. Les n premières unités sont tirées dans l'échantillon.
2. Pour $k = n+1, \ldots, N$, faire :
   - Avec une probabilité $\dfrac{n}{k}$, on sélectionne l'unité k.
   - On tire à probabilités égales une unité dans l'échantillon, qui est remplacée par k.

Avantage : la taille de la population peut être inconnue au départ.

Le sondage aléatoire simple stratifié

Information auxiliaire

On parle d'*information auxiliaire* lorsqu'une information est connue sur l'ensemble de la population, sous forme détaillée ou synthétique.

Il est fréquent de disposer d'une information auxiliaire sur la population, qui va permettre de partitionner la population et d'obtenir un plan de sondage plus efficace que le SRS.

Exemples d'information auxiliaire :

- le sexe et l'âge, pour une enquête auprès d'individus physiques,
- la taille (nombre d'employés) pour les enquêtes-entreprise.

Motivations pour la stratification (Cochran, 1977)

- Précision maîtrisée pour des sous-populations,
- simplicité administrative (enquêtes conduites par différentes agences),
- plans de sondage adaptés aux sous-populations,
- gain global de précision.

Principales questions :

1. Comment construire les strates ?
2. Quelle taille d'échantillon sélectionner dans chaque strate ?
3. Quel plan de sondage utiliser dans chaque strate ?

Notation et sondage stratifié

Définition

La population U est dite **stratifiée** quand les unités peuvent être partitionnées en H sous-populations disjointes $U_1, \ldots, U_H$ appelées **strates**.

Le plan de sondage est dit stratifié quand des **échantillons indépendants** sont sélectionnés dans chaque strate.

On parle de **Sondage aléatoire simple stratifié** (STSRS) si des échantillons aléatoires simples sont sélectionnés dans chaque strate.

Décomposition

On note N_h la taille de la strate U_h. Un total t_y peut se décomposer sous la forme

$$t_y = \sum_{h=1}^{H} t_{yh},$$

avec $t_{yh} = \sum_{k \in U_h} y_k$ le total de la variable y dans U_h. La moyenne μ_y peut se décomposer sous la forme d'une moyenne pondérée

$$\mu_y = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_{yh},$$

avec $\mu_{yh} = t_{yh}/N_h$ la moyenne dans la strate U_h.

Estimation d'un total

Dans chaque strate U_h, on sélectionne un échantillon S_h de taille n_h selon un SRS(U_h, n_h). Pour deux unités quelconques $k \neq l \in U_h$, on a donc les probabilités d'inclusion

$$\pi_k = \frac{n_h}{N_h} \qquad \pi_{kl} = \frac{n_h(n_h-1)}{N_h(N_h-1)}.$$

Pour deux unités quelconques k et l appartenant à deux strates distinctes U_h et $U_{h'}$, respectivement :

$$\pi_{kl} = \frac{n_h}{N_h}\ \frac{n_{h'}}{N_{h'}}.$$

Estimation d'un total

Par linéarité, le total t_y peut être estimé sans biais par

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} \hat{t}_{y_h\pi} = \sum_{h=1}^{H} N_h \ \bar{y}_h$$

avec $\bar{y}_h$ la moyenne simple dans S_h.

La variance s'obtient par sommation (les tirages sont indépendants dans les strates) :

$$V_p\left[\hat{t}_{y\pi}\right] \quad = \quad \sum_{h=1}^{H} V_p\left[\hat{t}_{y_h\pi}\right].$$

Estimation d'un total

Dans le cas d'un SRS stratifié, on obtient

$$V_p\left[\hat{t}_{y\pi}\right] = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{yh}^2 \quad \text{avec} \quad S_{yh}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \mu_{yh})^2,$$

que l'on estime par

$$v_{ST}\left[\hat{t}_{y\pi}\right] = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} s_{yh}^2 \quad \text{avec} \quad s_{yh}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_h)^2,$$

avec $f_h = n_h/N_h$ le taux de sondage dans la strate U_h.

Estimation d'une moyenne

De façon analogue, la moyenne μ_y peut être estimée sans biais par

$$\hat{\mu}_{y\pi} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h.$$

Sa variance est donnée par

$$V_p\left[\hat{\mu}_{y\pi}\right] = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h} S_{yh}^2,$$

et peut être estimée sans biais par

$$v_{ST}\left[\hat{\mu}_{y\pi}\right] = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h} s_{yh}^2.$$

Allocations pour le tirage stratifié

Allocation d'échantillon entre les strates

On suppose que la taille globale d'échantillon n est fixée, et que les strates ont été définies.

On doit choisir les tailles $n_1, \ldots, n_H$ des sous-échantillons à sélectionner dans chaque strate.

Nous revenons sur quelques allocations classiques pour le sondage aléatoire simple stratifié :

- Allocation Proportionnelle,
- Allocation Optimale,
- Allocation de compromis.

Allocation Proportionnelle

Allocation Proportionnelle

Avec une allocation proportionnelle, le taux de sondage est le même dans chaque strate :

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f.$$

On peut le réécrire sous la forme

$$n_h = n\frac{N_h}{N}.$$

Autrement dit, plus la strate est grande, plus l'échantillon sélectionné dedans est grand.

Allocation Proportionnelle

Chaque unité de la population possède la même probabilité d'inclusion $\pi_k = n/N$, et l'estimateur stratifié de la moyenne est identique à la moyenne simple sur l'échantillon :

$$\hat{\mu}_{y\pi} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y}.$$

Cette allocation conduit à un plan de sondage *auto-pondéré* où tous les individus possèdent le même poids $d_k = N/n$.

La variance de l'estimateur stratifié du total est donnée par

$$V_p\left[\hat{t}_{y\pi}\right] = N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_{yh}^2.$$

Equation de décomposition de la variance

La dispersion de la variable y dans la population U peut se décomposer sous la forme

$$S_y^2 = \underbrace{\sum_{h=1}^{H} \frac{N_h - 1}{N-1} S_{yh}^2}_{S_{y,intra}^2} + \underbrace{\sum_{h=1}^{H} \frac{N_h}{N-1} (\mu_{yh} - \mu_y)^2}_{S_{y,inter}^2}$$

$$\simeq \sum_{h=1}^{H} \frac{N_h}{N} S_{yh}^2 + \sum_{h=1}^{H} \frac{N_h}{N} (\mu_{yh} - \mu_y)^2$$

Le premier terme mesure la dispersion à l'intérieur des strates, alors que le second terme mesure la dispersion entre les strates.

Equation de décomposition de la variance

Notons que la dispersion globale S_y^2 est fixée. Le poids de chacune des deux composantes dépend de la variable de stratification choisie.

Exemple

k	1	2	3	4	5	6	7	8
y_k	1	1	1	1	5	5	5	5
x_{1k}	0	0	0	0	1	1	1	1
x_{2k}	0	0	1	1	1	1	0	0

Décomposition de la variance pour S_y^2 :

- si x_{1k} est la variable de stratification,
- si x_{2k} est la variable de stratification.

Retour vers l'allocation proportionnelle

La variance de l'estimateur stratifié avec allocation proportionnelle est approximativement donnée par

$$V_p\left[\hat{t}_{y\pi}\right] \simeq N^2 \frac{1-f}{n} S^2_{y,intra},$$

de sorte que :

- le SRS stratifié à allocation proportionnelle est (presque) toujours plus efficace que le SRS,
- la stratification devrait être choisie de façon à ce que la **dispersion à l'intérieur des strates** soit minimisée.

Allocation de Neyman

Principe

L'allocation de Neyman donne, pour une stratification donnée et une variable d'intérêt donnée, l'allocation d'échantillon pour laquelle la variance du π-estimateur est minimisée.

On cherche à résoudre un problème de minimisation (de la variance) sous contraintes (taille globale d'échantillon fixée) :

$$\min_{n_h} V\left[\hat{t}_{y\pi}\right] \quad \text{t.q.} \quad \sum_{h=1}^{H} n_h = n$$

Principe

Avec un sondage aléatoire simple stratifié, ce problème de minimisation peut se réécrire :

$$\min_{n_h} \sum_{h=1}^{H} \left[\frac{1}{n_h} - \frac{1}{N_h} \right] N_h^2 S_{yh}^2 \quad \text{t.q.} \quad \sum_{h=1}^{H} n_h = n.$$

En utilisant une technique de Lagrangien, on obtient :

$$n_h = n \frac{N_h S_{yh}}{\sum_{j=1}^{H} N_j S_{yj}}.$$

Notons en particulier que le calcul de cette allocation optimale nécessite la connaissance des dispersions dans les strates.

Principe

L'allocation de Neyman indique qu'il faut sélectionner un échantillon plus grand :

- dans les grandes strates,
- dans les strates présentant une forte dispersion.

L'allocation n'est optimale que pour la variable d'intérêt particulière y : pour une autre variable d'intérêt, elle peut conduire à des résultats plus imprécis que l'allocation proportionnelle (voire que le sondage aléatoire simple).

Calcul de l'allocation

L'allocation de Neyman peut conduire à des tailles d'échantillon supérieures aux tailles de strates, si ces dernières présentent une forte dispersion et/ou sont de grande taille.

Dans ce cas :

1. on effectue un recensement dans les strates concernées (on fixe $n_h = N_h$),
2. on recalcule l'allocation d'échantillon dans les autres strates.

Mise en oeuvre pratique

En pratique, on peut dériver une allocation proche de l'allocation de Neyman :

- si on possède un a priori sur la dispersion dans les strates (approche "métier"),
- ou si on peut estimer cette dispersion à l'aide d'une enquête antérieure.

Un problème alternatif peut être d'optimiser la précision sous une contrainte de coût global C fixé

$$C_0 + \sum_{h=1}^{H} C_h n_h = C,$$

où C_0 donne le coût fixe de l'enquête, et C_h le coût associé à une unité de U_h.

Principe

On résout le problème d'optimisation :

$$\min_{n_h} \sum_{h=1}^{H} \left[\frac{1}{n_h} - \frac{1}{N_h} \right] N_h^2 S_{yh}^2 \quad \text{t.q.} \quad C_0 + \sum_{h=1}^{H} C_h n_h = C.$$

En utilisant une technique de Lagrangien, on obtient :

$$n_h = [C - C_0] \frac{\left[N_h S_{yh} / \sqrt{C_h} \right]}{\sum_{j=1}^{H} \sqrt{C_j} N_j S_{yj}}.$$

Allocation de compromis

Allocation de compromis (1)

Imaginons que l'on souhaite obtenir la même précision dans chaque strate, par exemple si les strates sont des domaines d'estimation.

On veut obtenir

$$V(\bar{y}_h) = \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{yh}^2 = Cste.$$

Si les strates sont suffisamment grandes, on obtient :

$$n_h = n \frac{S_{yh}^2}{\sum_{j=1}^{H} S_{yj}^2}.$$

Allocation de compromis (2)

Supposons maintenant que l'on souhaite obtenir une allocation optimale, sous des contraintes

- de taille globale d'échantillon fixée,
- de précision dans les strates supérieure à un seuil fixé.

Il s'agit d'un problème réaliste (contraintes imposées par Eurostat dans les enquêtes).

Il est généralement difficile d'obtenir une solution explicite, mais ce type de problème peut être résolu (par exemple) à l'aide de la proc NLP de SAS.

Principales questions

1. Comment construire les strates ?
   $\Rightarrow$ de façon à ce que la dispersion intra soit minimisée
2. Quelle taille d'échantillon sélectionner dans chaque strate ?
   $\Rightarrow$ tirer de plus gros échantillons dans les strates avec une grande dispersion
3. Quel plan de sondage utiliser dans chaque strate ?
   $\Rightarrow$ le SRS par strate est une bonne stratégie si les unités présentes dans une même strate sont proches (au sens de la variable d'intérêt)

Introduction

Nous avons vu précedemment que la stratification était une méthode simple permettant de réduire la variance des estimateurs. Si les strates sont homogènes relativement à la variable d'intérêt (dispersion intra faible), le sondage aléatoire simple stratifié constitue une stratégie efficace d'échantillonnage.

En pratique, il peut subsister une forte hétérogénéité dans les strates. Dans ce cas, on peut rechercher une stratégie d'échantillonnage plus efficace en individualisant les probabilités de sélection π_k de chacun des individus.

On doit ensuite faire le choix d'un *algorithme de tirage*, i.e. d'une méthode pratique de sélection respectant les probabilités d'inclusion choisies.

Algorithmes de tirage

Il existe en pratique des dizaines d'algorithmes de tirage permettant de respecter un jeu de probabilités d'inclusion fixé (voir Tillé, 2006). Nous détaillerons deux de ces algorithmes :

- le tirage poissonien,
- le tirage systématique.

Les différents algorithmes se distinguent par les probabilités d'inclusion d'ordre 2 obtenues, i.e. par la variance des estimateurs. Cependant, il n'existe pas d'algorithme uniformément préférable en termes de variance.

Le choix de la méthode à utiliser dépend de la connaissance que l'on a de la base de sondage mais aussi des contraintes pratiques sur l'échantillonnage.

Propriétés d'un algorithme de tirage

Pour un algorithme de tirage, on se posera généralement les questions suivantes :

1. Est-ce que l'algorithme est **exact**, i.e. permet de respecter exactement un jeu de probabilités d'inclusion $(\pi_k)_{k \in U}$ fixé ?
2. Est-ce que l'algorithme est **de taille fixe**, i.e. ne sélectionne que des échantillons de la taille (moyenne) voulue ?
3. Est-ce que les **probabilités** π_{kl} sont **calculables**, et que vaut la variance du π-estimateur avec cet algorithme ?
4. Est-ce que ces probabilités π_{kl} :
   - **sont** > 0 : assure qu'on dispose d'un estimateur sans biais de variance.
   - vérifient les **conditions de Yates-Grundy** : assure qu'on dispose d'un estimateur toujours positif de variance (pour un plan de taille fixe).

Le tirage de Poisson

Principe

C'est une généralisation du tirage de Bernoulli au cas des probabilités inégales :

- Etape 1 : on génère $u_1 \sim U[0,1]$. Si $u_1 \leq \pi_1$, l'unité 1 est retenue dans l'échantillon.
- Etape 2 : on génère $u_2 \sim U[0,1]$ indépendamment de u_1. Si $u_2 \leq \pi_2$, l'unité 2 est retenue dans l'échantillon.
- ...
- Etape N : on génère $u_N \sim U[0,1]$ indépendamment de $u_1, \ldots, u_{N-1}$. Si $u_N \leq \pi_N$, l'unité N est retenue dans l'échantillon.

C'est un principe de piles ou faces indépendants, avec une pièce et un lancer différents pour chaque unité.

Estimation de Horvitz-Thompson

En utilisant les propriétés d'une loi $U[0,1]$, on a :

$$\mathbb{P}(k \in S) = \mathbb{P}(u_k \leq \pi_k) = F_U(\pi_k) = \pi_k.$$

Du fait de l'indépendance dans la sélection des unités :

$$\pi_{kl} \quad = \quad \pi_k \ \pi_l \ \text{si} \ k \neq l.$$

Le plan de sondage peut être entièrement spécifié. Pour une partie quelconque $s = \{i_1, \ldots, i_p\}$ de U, on a :

$$\mathbb{P}(S = s) \quad = \quad \prod_{k \in s} \pi_k \prod_{k \in u \backslash s} (1 - \pi_k).$$

Application

Dans la population ci-dessous, utiliser les nombres aléatoires pour sélectionner un échantillon selon un tirage de Poisson, et en déduire une estimation de t_y.

Unité	1	2	3	4	5	6	7	8
π_k	0.1	0.1	0.1	0.1	0.4	0.4	0.4	0.4
u_k	0.07	0.44	0.52	0.19	0.95	0.24	0.54	0.07
y_k	0	0	1	2	2	2	2	4
Tirage								

Quelle est la taille moyenne d'échantillon attendue ? Quelle est la taille d'échantillon effectivement obtenue ?
Comparer $\hat{t}_{y\pi}$ et t_y, et commenter la différence observée. Comparer avec le tirage de Bernoulli.

Estimateur de Horvitz-Thompson

La variance s'obtient à partir de l'expression générale de Horvitz-Thompson :

$$V_{pois}\left[\hat{t}_{y\pi}\right] = \sum_{k\in U}\left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1-\pi_k), \tag{8}$$

que l'on estime sans biais par

$$v\left[\hat{t}_{y\pi}\right] = \sum_{k\in S}\left(\frac{y_k}{\pi_k}\right)^2 (1-\pi_k).$$

En particulier, cela implique que la taille d'échantillon est aléatoire :

$$V_{pois}\left[n(S)\right] = \sum_{k\in U}\pi_k(1-\pi_k).$$

Estimateur par le ratio

Si la taille de la population N est connue, on préfère à l'estimateur de Horvitz-Thompson l'**estimateur par le ratio**

$$\hat{t}_{yR} = \frac{N}{\hat{N}_\pi}\hat{t}_{y\pi}.$$

Sa variance est approximativement donnée par

$$V_{pois}\left[\hat{t}_{yR}\right] \simeq \sum_{k\in U}\left(\frac{E_k}{\pi_k}\right)^2 \pi_k(1-\pi_k) \quad (9)$$

avec $E_k = y_k - \mu_y$. On peut utiliser l'estimateur de variance

$$v\left[\hat{t}_{yR}\right] = \sum_{k\in S}\left(\frac{e_k}{\pi_k}\right)^2 (1-\pi_k) \quad (10)$$

avec $e_k = y_k - \frac{\hat{t}_{y\pi}}{\hat{N}_\pi}$.

Estimateur par le ratio (2)

Pour l'estimation de la moyenne μ_y, on peut utiliser l'estimateur par substitution $\tilde{\mu}_y = \frac{\hat{t}_{y\pi}}{\hat{N}_\pi}$, de variance (approximative)

$$V_{pois}\left[\tilde{\mu}_y\right] \simeq \frac{1}{N^2}\sum_{k\in U}\left(\frac{E_k}{\pi_k}\right)^2 \pi_k(1-\pi_k). \quad (11)$$

On peut utiliser l'estimateur de variance

$$v\left[\tilde{\mu}_y\right] = \frac{1}{\hat{N}_\pi^2}\sum_{k\in S}\left(\frac{e_k}{\pi_k}\right)^2 (1-\pi_k). \quad (12)$$

L'estimateur par substitution $\tilde{\mu}_y$ peut être calculé même si la taille de la population est inconnue.

Utilisation

Le tirage poissonien est rarement utilisé pour un tirage d'échantillon, en raison de sa grande variance. On trouve cependant des applications dans le cas d'échantillonnage forestier (Schreuder et al., 1993).

Le tirage poissonien est également utilisé dans un contexte de *non-réponse*. On parle de *non-réponse totale* quand certains individus échantillonnés ne peuvent finalement pas être enquêtés. Pour exploiter l'échantillon de répondants, noté S_r, il est généralement nécessaire de modéliser le mécanisme de réponse.

Une modélisation classique consiste à supposer que l'échantillon S_r est obtenu par sous-échantillonnage dans l'échantillon S d'origine. Plus de détails dans le cours sur les Données Manquantes (Attachés).

Le tirage systématique

Principe

C'est une méthode simple et très rapide permettant de sélectionner un échantillon à probabilités inégales et de taille fixe.

Principe :

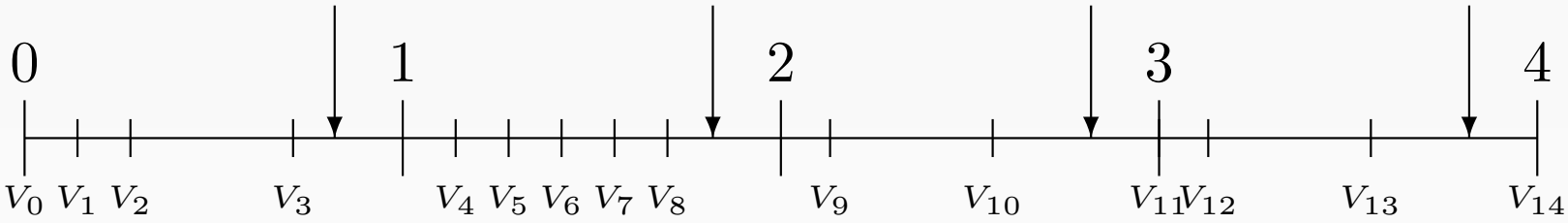
- On pose $V_k = \sum_{l=1}^{k} \pi_l$ pour $k \in U$, avec la convention $V_0 = 0$.
- On tire une variable aléatoire u selon une loi uniforme $U[0, 1]$.
- On sélectionne toutes les unités k telles que, pour un entier $i \in \{1, \ldots, n\}$:

$$V_{k-1} \leq u + (i - 1) < V_k.$$

Exemple

Population U de taille $N = 14$ avec $n = 4$:

- $\pi_1 = \pi_2 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = \pi_{12} = 1/7$,
- $\pi_3 = \pi_4 = \pi_9 = \pi_{10} = \pi_{11} = \pi_{13} = \pi_{14} = 3/7$.



$$
\begin{aligned}
u = 0.82 \in [V_3, V_4] &\Rightarrow \text{l'unité 4 est sélectionnée,} \\
1 + u = 1.82 \in [V_8, V_9] &\Rightarrow \text{l'unité 9 est sélectionnée,} \\
2 + u = 2.82 \in [V_{10}, V_{11}] &\Rightarrow \text{l'unité 11 est sélectionnée,} \\
3 + u = 3.82 \in [V_{13}, V_{14}] &\Rightarrow \text{l'unité 14 est sélectionnée.}
\end{aligned}
$$

Probabilités d'inclusion

Les probabilités π_k sont exactement respectées. En effet :

$$\begin{aligned} \mathbb{P}(k \in S) &= \mathbb{P}(V_{k-1} \leq u + (i-1) < V_k) \\ &= V_k - V_{k-1} = \pi_k. \end{aligned}$$

Les probabilités d'inclusion d'ordre deux sont plus difficiles à calculer (Tillé, 2006, p. 126).

Beaucoup d'unités présentent des probabilités d'inclusion doubles égales à 0 (méthode très peu aléatoire)
$\Rightarrow$ il n'existe pas d'estimateur sans biais de variance pour l'estimateur de Horvitz-Thompson.

Applications du tirage systématique

Exemple 1 : sélection pour contrôle d'un sous-échantillon de questionnaires, arrivant à flux tendu.

Exemple 2 : enquête auprès des clients entrant dans un magasin.

Exemple 3 : tirage de logements dans un pâté de maison lors d'une enquête ménage.

Cas des probabilités égales

On suppose dans la suite de cette section que les probabilités d'inclusion sont égales ($\pi_k = n/N$), et que le *pas de tirage* $p = N/n$ est entier.

Dans ce cas, l'algorithme peut être simplifié de la façon suivante :

- On tire un individu i au hasard parmi les p premiers.
- On sélectionne les individus $i, i+p, \ldots, i+(n-1)p$.

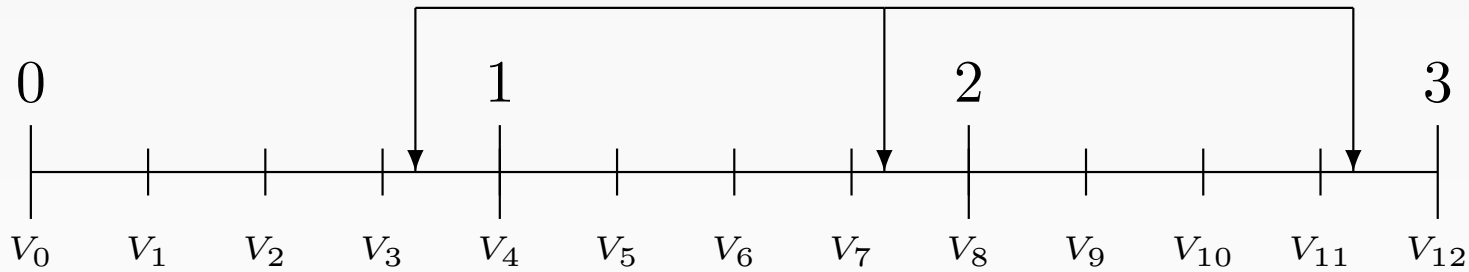
On a en particulier

$$\pi_{kl} = \begin{cases} n/N & \text{si } k \equiv l \quad [p], \\ 0 & \text{sinon.} \end{cases}$$

Il n'existe pas d'estimateur sans biais de variance, et les conditions de Yates-Grundy ne sont pas vérifiées.

Exemple

Sélection d'un échantillon de taille 3 dans une population de taille 12 selon un tirage systématique à probabilités égales ($\pi_k = \frac{1}{4}$).



$$
\begin{aligned}
u = 0.82 \in [V_3, V_4] &\Rightarrow \text{l'unité } 4 \text{ est sélectionnée,} \\
1 + u = 1.82 \in [V_7, V_8] &\Rightarrow \text{l'unité } 8 \text{ est sélectionnée,} \\
2 + u = 2.82 \in [V_{11}, V_{12}] &\Rightarrow \text{l'unité } 12 \text{ est sélectionnée.}
\end{aligned}
$$

Seuls 4 échantillons sont sélectionnables, chacun avec une probabilité de 0.25 :

$$\{1, 5, 9\} \qquad \{2, 6, 10\} \qquad \{3, 7, 11\} \qquad \{4, 8, 12\}$$

Précision du tirage systématique

Dans le cas précédent (probabilités égales, pas de tirage entier), le tirage est équivalent à un *tirage par grappes* de taille $m = 1$ dans la population $U_g = \{u_1, \ldots, u_p\}$, avec

$$u_i = \{i, i + p, \ldots, i + (n - 1)p\}.$$

Principe du tirage par grappes :

1. on tire un échantillon S_I de grappes (ici, de taille $m = 1$),
2. tous les individus contenus dans les grappes de s_I sont retenus dans l'échantillon s finalement enquêté.

Pour plus de détails : cours de Méthodologie d'Enquête (Attachés).

Précision du tirage systématique (2)

Le π-estimateur peut se réécrire sous la forme

$$\hat{t}_{y\pi} = \frac{N}{n} \sum_{u_i \in S_I} Y_i,$$

avec $Y_i = \sum_{k \in u_i} y_k$ le total sur la grappe u_i. En utilisant les résultats du SRS, on obtient

$$V_{sys}\left[\hat{t}_{y\pi}\right] = N^2 \frac{1-f}{n} \frac{S_Y^2}{n}$$

avec

$$S_Y^2 = \frac{1}{p-1} \sum_{u_i \in U_g} \left[Y_i - \frac{t_y}{p}\right]^2.$$

Comparaison avec le SRS

On appelle *design-effect* (ou *effet de plan*)

$$\mathrm{DEFF}_p(y) = \frac{V_p\left[\hat{t}_{y\pi}\right]}{V_{SRS}\left[\hat{t}_{y\pi}\right]}$$

le rapport entre la variance associée à un plan de sondage, et la variance associée au SRS de même taille. On a ici :

$$\mathrm{DEFF}_{sys}(y) = \frac{S_Y^2/n}{S_y^2}.$$

D'autre part, en utilisant une décomposition de la variance :

$$S_y^2 = \frac{n-1}{N-1}\sum_{i=1}^{p} S_{yi}^2 + \frac{(p-1)}{n(N-1)} S_Y^2 \tag{13}$$

$$\simeq \frac{1}{p}\sum_{i=1}^{p} S_{yi}^2 + \frac{1}{n}\left(S_Y^2/n\right).$$

Comparaison avec le SRS (2)

Le tirage systématique sera donc efficace par rapport au SRS si dans l'équation (13), le terme de dispersion intra est grand, autrement dit si les grappes sont **hétérogènes en intra**. Ce sera par exemple le cas si la population est triée avant le tirage selon une variable auxiliaire x_k corrélée avec la variable d'intérêt.

Le tirage systématique peut au contraire être très inefficace si les grappes sont **homogènes en intra** : c'est une difficulté et une situation habituelle dans le cas d'un tirage par grappes.

Le cas le plus défavorable est celui où la variable d'intérêt présente une périodicité + le pas de tirage $p = N/n$ est proportionnel à cette périodicité.

Exemple

Considérons une population U de taille 12 sur laquelle on relève trois caractéristiques y_1, y_2, y_3 (valeur moyenne 40) :

Unité	1	2	3	4	5	6	7	8	9	10	11	12
y_1	10	10	10	15	45	45	50	50	60	60	60	65
y_2	10	45	60	15	50	65	10	50	60	10	45	60
y_3	15	45	10	60	60	50	45	65	10	50	10	60

On sélectionne un échantillon de taille 2 selon un tirage systématique
$\Rightarrow$ 6 échantillons possibles.

Exemple

On obtient comme valeurs échantillonnées possibles pour y_1

$$\{10,50\} \quad \{10,50\} \quad \{10,60\} \quad \{15,60\} \quad \{45,60\} \quad \{45,65\},$$

pour y_2

$$\{10,10\} \quad \{45,50\} \quad \{60,60\} \quad \{15,10\} \quad \{50,45\} \quad \{65,60\},$$

et pour y_3

$$\{15,45\} \quad \{45,65\} \quad \{10,10\} \quad \{60,50\} \quad \{60,10\} \quad \{50,60\}.$$

On obtient également :

	y_1	y_2	y_3
$\mathrm{DEFF}_p(y)$	0.50	2.18	1.39

Méthodes de redressement

Principe

Nous revenons au cas de l'estimation d'un total. On suppose qu'un échantillon S a été sélectionné selon un plan de sondage $p(\cdot)$. Un estimateur direct est donné par :

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k y_k.$$

Dans cet estimateur, les **poids de sondage** d_k dépendent de l'information auxiliaire mobilisée au moment de l'échantillonnage :

$$\begin{aligned} \text{SRS} &\Rightarrow d_k = N/n \\ \text{SRS stratifié} &\Rightarrow d_k = N_h/n_h \text{ pour } k \in U_h \end{aligned}$$

Principe

Il se peut qu'une partie de l'information auxiliaire n'ait pas été utilisée au moment de la sélection de l'échantillon, ou qu'elle n'ait pas été disponible.

Si cette information est explicative de la variable d'intérêt, il va être néanmoins intéressant de l'utiliser. On peut le faire au stade de l'estimation, en **redressant** l'estimateur de Horvitz-Thompson.

$$\text{Poids de sondage } d_k \Rightarrow \text{Poids redressés } w_k$$

Principe

On dit que l'on **redresse** l'échantillon lorsque l'on modifie le système de pondérations associé à S afin de respecter un certain nombre d'**informations auxiliaires**.

On parle d'information auxiliaire lorsque l'on dispose d'une information connue **sur l'ensemble de la population**.

Exemples :

- Chiffre d'affaire total des entreprises d'un secteur d'activité,
- Répartition par sexe et par âge d'une population d'individus.

Estimateur par calage

Principe

On suppose ici que l'on dispose d'un vecteur $\mathbf{x}_k = [x_{1k}, \ldots, x_{pk}]^\top$ de variables auxiliaires, dont les totaux $t_\mathbf{x} = [t_{x1}, \ldots, t_{xp}]^\top$ sur la population sont connus.

Avant calage, on a pour toute variable y l'estimateur sans biais du total :

$$\begin{aligned} \hat{t}_{y\pi} &= \sum_{k \in S} d_k y_k, \\ E_p\left[\hat{t}_{y\pi}\right] &= t_y, \end{aligned}$$

et en particulier pour les variables de calage :

$$E_p\left[\hat{t}_{\mathbf{x}\pi}\right] = t_\mathbf{x}.$$

Modification des poids

On cherche de nouveaux poids w_k qui

1. **restent proches** des poids de départ d_k,
2. **vérifient les équations de calage**

$$\sum_{k \in S} w_k \mathbf{x}_k = t_{\mathbf{x}}.$$

Plus formellement, on résoud le problème suivant :

$$\min_{w_k} \sum_{k \in s} d_k G\left(\frac{w_k}{d_k}\right) \text{ s.c. } \sum_{k \in s} w_k \mathbf{x}_k = t_{\mathbf{x}}$$

où G désigne une **fonction de distance**.

Modification des poids

On cherche à **réduire la variance** de l'estimation à l'aide du calage sur les totaux connus. La variance est nulle pour les variables auxiliaires ; elle sera faible pour les variables d'intérêt bien expliquées par les variables auxiliaires.

Pour respecter les totaux de variables auxiliaires, on accepte de **biaiser** légèrement l'estimation. Ce biais sera généralement négligeable car on assure que les poids calés restent proches des poids d'origine.

Solution théorique

On choisit une fonction de distance G telle que $G(w_k/d_k)$ mesure la distance entre le poids initial d_k et le poids final w_k. Nous supposons que

- $G(1) = 0$,
- G est positive et convexe (i.e, plus w_k/d_k s'éloigne de 1, plus $G(w_k/d_k)$ est grand)

Le Lagrangien s'écrit

$$L = \sum_{k \in s} d_k G(w_k/d_k) - \lambda^\top \left(\sum_{k \in s} w_k \mathbf{x}_k - t_{\mathbf{x}} \right)$$

où $\lambda = [\lambda_1, \ldots, \lambda_p]^\top$ est un vecteur de multiplicateurs de Lagrange.

Solution théorique (2)

La résolution du problème d'optimisation conduit à :

$$w_k \quad = \quad d_k F[\lambda^\top \mathbf{x}_k]$$

avec F la fonction inverse de G'.

Le vecteur λ peut être déterminé en résolvant le système (non-linéaire) constitué par les équations de calage

$$\sum_{k \in s} d_k F[\lambda^\top \mathbf{x}_k]\mathbf{x}_k \quad = \quad t_{\mathbf{x}},$$

par exemple à l'aide de la méthode itérative de Newton-Raphson.

Fonctions de distance usuelles : la méthode linéaire

$G(r) = \frac{1}{2}(r-1)^2$ et $F(u) = 1 + u$. La convergence est obtenue à l'étape 2 de l'algorithme de Newton, et on obtient l'**estimateur par la régression généralisée**

$$\begin{aligned} \hat{t}_{y,greg} &= \sum_{k \in s} w_k y_k \\ &= \hat{t}_{y\pi} + \hat{\mathbf{b}}_\pi^\top \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi} \right] \end{aligned}$$

avec

$$\hat{\mathbf{b}}_\pi = \left[\sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k} \right]^{-1} \sum_{k \in S} \frac{\mathbf{x}_k y_k}{\pi_k}.$$

Cette méthode de calage peut conduire à des poids finaux w_k négatifs.

Les fonctions de distance usuelles (2)

La méthode raking ratio

$G(r) = r\log(r) - r + 1$ et $F(u) = \exp(u)$. Cette méthode permet d'assurer que les poids finaux w_k sont > 0.

Les méthodes bornées

Elles peuvent être vues comme des versions "tronquées" des deux méthodes précédentes. Ces deux méthodes permettent de spécifier explicitement des bornes LO et UP pour les rapports de poids, i.e. d'assurer que pour tout individu $k \in S$

$$LO \leq \frac{w_k}{d_k} \leq UP.$$

Estimation après calage

Après calage, on a pour toute variable y l'estimateur calé :

$$\hat{t}_{yw} = \sum_{k \in s} w_k y_k.$$

L'estimation est **exacte** pour les totaux de variables auxiliaires :

$$\hat{t}_{\mathbf{x}w} = t_{\mathbf{x}}.$$

Elle est **approximativement sans biais** pour les autres variables d'intérêt :

$$E_p\left[\hat{t}_{yw}\right] \simeq t_y.$$

Estimateur par la régression généralisée

Motivation de l'estimateur par la régression

L'estimateur par la régression généralisée est obtenu par calage avec la méthode linéaire. Il est motivé par le modèle

$$y_k = \beta^\top \mathbf{x}_k + \epsilon_k \quad \text{avec} \quad V_m[\epsilon_k] = \sigma_k^2. \tag{14}$$

Si on dispose des données sur toute la population, le meilleur estimateur de β s'obtient par les MCG :

$$\mathbf{b} = \left[\sum_{k\in U} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2}\right]^{-1} \sum_{k\in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2} \quad \text{et} \quad E_k = y_k - \mathbf{b}^\top \mathbf{x}_k.$$

On les remplace par leurs estimateurs $\hat{\mathbf{b}}_\pi$ et $e_k = y_k - \hat{\mathbf{b}}_\pi^\top \mathbf{x}_k$ pour obtenir l'estimateur par la régression :

$$\hat{t}_{y,greg} = \underbrace{\hat{\mathbf{b}}_\pi^\top t_{\mathbf{x}}}_{\text{prédiction du total}} + \underbrace{\hat{t}_{e\pi}}_{\text{estimation de l'erreur totale}}.$$

Fonctions de distance usuelles : la méthode linéaire

On peut réécrire l'estimateur GREG sous la forme :

$$\hat{t}_{y,greg} = \hat{t}_{y\pi} + \hat{\mathbf{b}}_{\pi}^{\top} \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi} \right].$$

En utilisant l'approximation $\hat{t}_{y,greg} \simeq \hat{t}_{y\pi} + \mathbf{b}^{\top} \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi} \right]$, on obtient :

$$\begin{aligned} E_p \left[\hat{t}_{y,greg} \right] &\simeq E_p \left[\hat{t}_{y\pi} + \mathbf{b}^{\top} \{ t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi} \} \right] \\ &= t_y, \\ V_p \left[\hat{t}_{y,greg} \right] &\simeq V_p \left[\hat{t}_{y\pi} - \mathbf{b}^{\top} \hat{t}_{\mathbf{x}\pi} \right] \\ &= V_p \left[\hat{t}_{E\pi} \right]. \end{aligned}$$

L'estimateur GREG est donc **approximativement sans biais**, et sa variance est approximativement donnée par les **résidus de la régression** de la variable y_k sur les variables auxiliaires $\mathbf{x}_k$.

Fonctions de distance usuelles : la méthode linéaire

Application : régression simple

On se place dans le cas d'un SRS(n). On suppose que l'on utilise les variables auxiliaires $\mathbf{x}_k = [1, x_k]^\top$, de totaux connus. Le modèle de régression sous-jacent est :

$$y_k = a + b\ x_k + E_k.$$

On obtient

$$b = \frac{\sum_{k\in U}(x_k - \mu_x)(y_k - \mu_y)}{\sum_{k\in U}(x_k - \mu_x)^2} = \frac{S_{xy}}{S_x^2} \qquad a = \mu_y - b\ \mu_x$$

$$\hat{b} = \frac{\sum_{k\in S}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k\in S}(x_k - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \qquad \hat{a} = \bar{y} - \hat{b}\ \bar{x}$$

En notant $\rho = \frac{S_{xy}}{S_x S_y}$ le coefficient de corrélation linéaire, on a :

$$V_{srs}\left[\hat{t}_{y,greg}\right] \simeq N^2 \frac{1-f}{n} S_y^2 (1-\rho^2).$$

Variance d'un estimateur calé

Variance d'un estimateur calé

Quelle que soit la fonction de distance utilisée, la variance de l'estimateur calé $\hat{t}_{yw}$ est **approximativement celle de l'estimateur par la régression**.

La variance de l'estimateur calé $\hat{t}_{yw}$ est donc approximativement égale à

$$V_p\left[\hat{t}_{yw}\right] \simeq \sum_{k,l \in U} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \Delta_{kl}$$

où $E_k = y_k - \mathbf{b}^\top \mathbf{x}_k$ donne les résidus de la régression de y sur le vecteur de variables auxiliaires $\mathbf{x}_k$ dans la population U.

Estimation de variance

Deux estimateurs de variance peuvent être utilisés :

$$
\begin{aligned}
v_1\left[\hat{t}_{yw}\right] &= \sum_{k,l\in S} \frac{e_k}{\pi_k}\frac{e_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}} \\
v_2\left[\hat{t}_{yw}\right] &= \sum_{k,l\in S} \frac{g_k\ e_k}{\pi_k}\frac{g_l\ e_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}},
\end{aligned}
$$

où $g_k = w_k/d_k$, et $e_k = y_k - \hat{\mathbf{b}}_\pi^T \mathbf{x}_k$ donne les résidus estimés.

Le second estimateur est généralement (légèrement) préférable.

Estimation de variance

Un logiciel classique d'estimation de variance pour l'estimation de totaux $\hat{t}_{y\pi}$ peut être utilisé pour l'estimation de variance d'estimateurs calés $\hat{t}_{yw}$ de la façon suivante :

- Effectuer sur l'échantillon S la régression pondérée (par les poids d_k) de la variable y sur les variables auxiliaires $x_1, \ldots, x_p$,
- Prendre les résidus e_k de la régression et calculer les $g_k = w_k/d_k$,
- Utiliser le logiciel en remplaçant les y_k par les e_k (estimateur de variance v_1) ou par les $g_k\ e_k$ (estimateur de variance v_2).

Exemple

Echantillon de taille $n = 5$ tiré selon un SRS dans une population de taille $N = 100$. On suppose connu le total $t_x = 320$.

x_{0k}	x_{1k}	y_k	
1	1	3	
1	3	1	
1	2	8	
1	5	15	
1	4	3	

$$\hat{t}_{x\pi} = 300 \quad \hat{t}_{y\pi} = 600 \quad v(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 = 6.08\ 10^4$$

Exemple

Echantillon de taille $n = 5$ tiré selon un SRS dans une population de taille $N = 100$. On suppose connu le total $t_x = 320$.

x_{0k}	x_{1k}	y_k	$e_k = y_k - \hat{a} - \hat{b}\, x_{1k}$
1	1	3	0.8
1	3	1	-5
1	2	8	3.9
1	5	15	5.2
1	4	3	-4.9

$$\hat{t}_{x\pi} = 300 \quad \hat{t}_{y\pi} = 600 \quad v(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 = 6.08\ 10^4$$

$$\hat{a} = 0.3 \quad \hat{b} = 1.9$$

$$\hat{t}_{yw} = 638 \quad v(\hat{t}_{yw}) = N^2 \frac{1-f}{n} s_e^2 = 4.365\ 10^4$$

Application des méthodes de redressement

Estimateur par le ratio

L'estimateur par le ratio

On suppose connu le total t_x d'une seule variable auxiliaire (positive) x_k. L'estimateur par le ratio est défini par

$$\hat{t}_{yR} = \hat{t}_{y\pi} \times \frac{t_x}{\hat{t}_{x\pi}} = \sum_{k \in S} w_k y_k$$

avec $w_k = d_k \times \frac{t_x}{\hat{t}_{x\pi}}$.

Exemple : enquête auprès d'entreprises, avec redressement sur la variable d'effectif salarié.

Motivation

L'estimateur par le ratio est motivé par le modèle

$$y_k = \beta\ x_k + \epsilon_k \quad \text{avec} \quad V_m[\epsilon_k] = \sigma^2\ x_k.$$

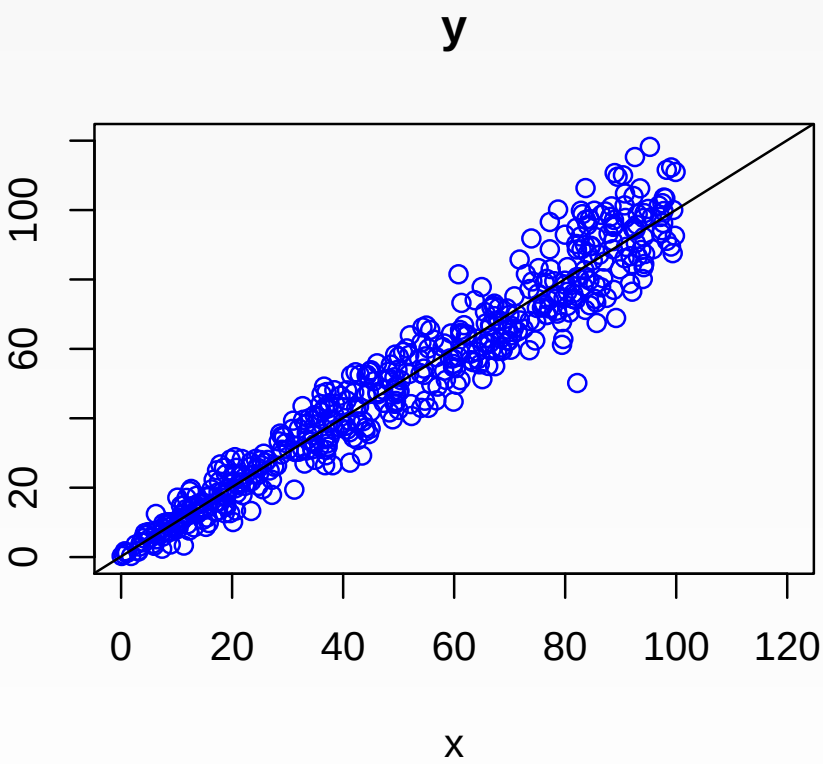
C'est un cas particulier de l'estimateur par la régression généralisée, obtenu avec $\mathbf{x}_k = x_k$ et $\sigma_k^2 = \sigma^2\ x_k$. On a :

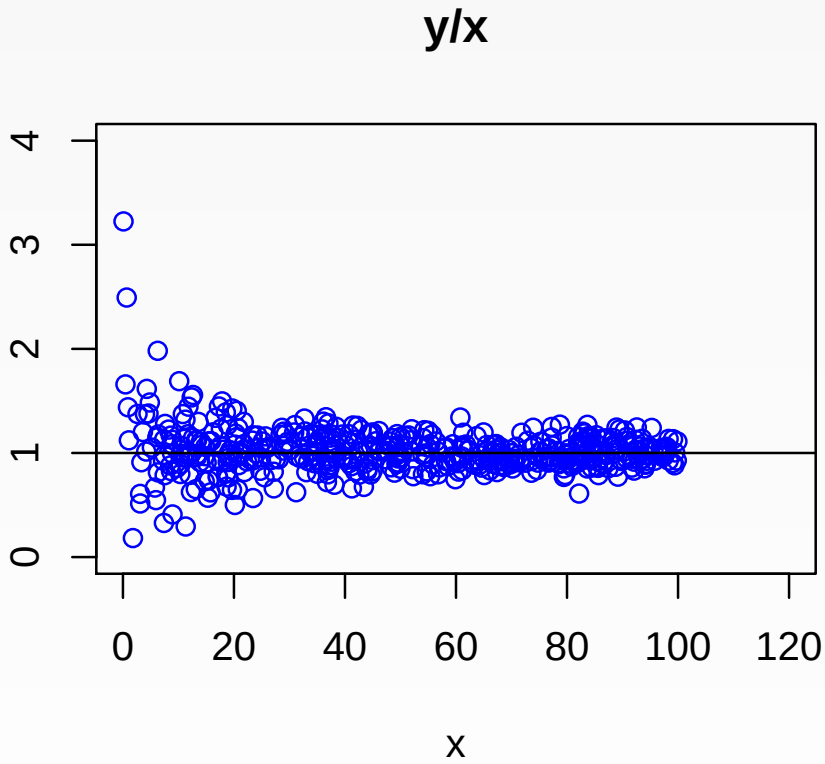
$$\begin{aligned} \hat{\mathbf{b}}_\pi &= \left[\sum_{k\in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2\ \pi_k}\right]^{-1} \sum_{k\in S} \frac{\mathbf{x}_k y_k}{\sigma_k^2\ \pi_k} \\ &\equiv \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}} = \hat{R}_\pi, \end{aligned}$$

et

$$\hat{t}_{e\pi} = \sum_{k\in S} \frac{y_k - \hat{\mathbf{b}}_\pi^\top \mathbf{x}_k}{\pi_k} \equiv \sum_{k\in S} \frac{y_k - \hat{R}_\pi x_k}{\pi_k} = 0.$$

Exemple de données





Propriétés de l'estimateur par le ratio

L'estimateur par le ratio est approximativement non biaisé pour le total t_y. On a

$$V_p\left[\hat{t}_{yR}\right] \simeq V_p\left[\hat{t}_{E\pi}\right]$$

avec $E_k = y_k - \mathbf{b}^\top \mathbf{x}_k \equiv y_k - R\, x_k$. La variance est donc réduite si les variables y_k et x_k sont approximativement proportionnelles.

L'estimateur par le ratio est **calé** sur le total t_x :

$$\hat{t}_{xR} = t_x.$$

Cas du sondage aléatoire simple

Application au sondage aléatoire simple

Dans le cas d'un SRS(n), on obtient :

$$V_p\left[\hat{t}_{yR}\right] \simeq N^2 \frac{1-f}{n} S_E^2.$$

On peut l'estimer par

$$\tilde{v}\left[\hat{t}_{yR}\right] = N^2 \frac{1-f}{n} s_E^2,$$

mais $E_k = y_k - R\, x_k$ n'est pas calculable sur l'échantillon. On la remplace par la variable donnant les résidus estimés $e_k = y_k - \hat{R}_\pi\, x_k$ pour obtenir l'estimateur de variance final :

$$v\left[\hat{t}_{yR}\right] = N^2 \frac{1-f}{n} s_e^2.$$

Exemple

Echantillon de taille $n = 5$ tiré selon un SRS dans une population de taille $N = 100$. On suppose connu le total $t_x = 320$.

x_k	y_k	
1	3	
3	1	
2	8	
5	15	
4	3	

$$\hat{t}_{x\pi} = 300$$
$$\hat{t}_{y\pi} = 600 \quad v(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 = 6.08\ 10^4$$

Exemple

Echantillon de taille $n = 5$ tiré selon un SRS dans une population de taille $N = 100$. On suppose connu le total $t_x = 320$.

x_k	y_k	$e_k = y_k - \hat{R}_\pi \, x_k$
1	3	1
3	1	-5
2	8	4
5	15	5
4	3	-5

$$\hat{t}_{x\pi} = 300$$
$$\hat{t}_{y\pi} = 600 \quad v(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 = 6.08 \; 10^4$$

$$\hat{R}_\pi = 2$$
$$\hat{t}_{yR} = 640 \quad v(\hat{t}_{yR}) = N^2 \frac{1-f}{n} s_e^2 = 4.37 \; 10^4$$

Efficacité de l'estimateur par le ratio

Dans le cas du SRS, l'estimateur par le ratio est préférable à l'estimateur direct si

$$\frac{V_p\left[\hat{t}_{yR}\right]}{V_p\left[\hat{t}_{y\pi}\right]} \leq 1 \quad \Leftrightarrow \quad \frac{S_y^2 - 2RS_{xy} + S_x^2}{S_y^2} \leq 1$$

$$\Leftrightarrow \quad \rho \geq \frac{1}{2}\frac{cv_x}{cv_y}$$

avec $cv_x = \sqrt{S_x^2}/\mu_x$ et $cv_y = \sqrt{S_y^2}/\mu_y$.

Même si les variables x et y sont corrélées positivement, l'estimateur direct peut être plus efficace. Si la corrélation est négative, l'estimateur direct est toujours plus efficace.

Efficacité de l'estimateur par le ratio (2)

On peut également montrer que

$$V_{srs}\left[\hat{t}_{yR}\right] - V_{srs}\left[\hat{t}_{y,greg}\right] \simeq N^2 \frac{1-f}{n} S_y^2 \left(\rho - R\frac{S_x}{S_y}\right)^2,$$

avec $\hat{t}_{y,greg}$ l'estimateur par la régression simple obtenu avec le vecteur $\mathbf{x}_k = (1, x_k)^\top$.

L'estimateur par la régression simple est donc toujours meilleur (asymptotiquement) que l'estimateur par le ratio dans le cas d'un sondage aléatoire simple.

Cas du sondage aléatoire simple stratifié

Application au sondage aléatoire simple stratifié

Si le total t_x sur l'ensemble de la population est connu, on obtient l'**estimateur par le ratio combiné**

$$\hat{t}_{y,RC} = t_x \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}} = t_x \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{\sum_{h=1}^{H} N_h \bar{x}_h}$$

et sa variance est approximativement égale à

$$\begin{aligned} V_p\left[\hat{t}_{y,RC}\right] &\simeq V_p\left[\hat{t}_{E\pi}\right] \\ &= \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{Eh}^2. \end{aligned}$$

avec $E_k = y_k - R\, x_k$. La variance est donc réduite si les variables y et x sont **approximativement proportionnelles sur l'ensemble de la population**.

Application au sondage aléatoire simple stratifié (2)

Cette variance peut être estimée par

$$v\left[\hat{t}_{y,RC}\right] = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} s_{eh}^2,$$

avec $e_k = y_k - \hat{R}_\pi \; x_k$.

D'un autre côté, si les totaux par strate t_{xh} sont connus, on peut appliquer un redressement par le ratio strate par strate.

On obtient l'**estimateur par le ratio séparé**

$$\hat{t}_{y,RS} = \sum_{h=1}^{H} t_{xh} \frac{\hat{t}_{y_h}}{\hat{t}_{x_h}} = \sum_{h=1}^{H} t_{xh} \frac{\bar{y}_h}{\bar{x}_h}.$$

Cas d'un sondage aléatoire simple stratifié

Sa variance est approximativement donnée par

$$V_p\left[\hat{t}_{y,RS}\right] \simeq \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{Eh}^2$$

avec $E_k = y_k - R_h\, x_k$ pour $k \in U_h$, et $R_h = t_{yh}/t_{xh}$. La variance est donc réduite si les variables y et x sont **approximativement proportionnelles dans les strates**.

Cette variance peut être estimée par

$$v\left[\hat{t}_{y,RS}\right] = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} s_{eh}^2,$$

avec $e_k = y_k - \hat{R}_h\, x_k$ et $\hat{R}_h = \hat{t}_{yh}/\hat{t}_{xh}$.

Estimateur post-stratifié

Principe

On suppose que l'on connaît **après le tirage de l'échantillon** une partition de la population en H groupes notés $U_1, \ldots, U_H$. On parle de **post-stratification**.

Les effectifs des post-strates, notés $N_1, \ldots, N_H$, sont supposés connus.

Le π-estimateur peut se réécrire

$$\begin{aligned} \hat{t}_{y\pi} &= \sum_{h=1}^{H} \sum_{k \in S_h} \frac{y_k}{\pi_k} \\ &= \sum_{h=1}^{H} \hat{t}_{yh} \end{aligned}$$

avec S_h l'intersection de S et de U_h.

Principe de post-stratification

L'estimateur post-stratifié est défini par

$$\hat{t}_{ypost} = \sum_{h=1}^{H} N_h \tilde{\mu}_{yh},$$

avec

$$\tilde{\mu}_{yh} = \frac{\sum_{k \in S_h} \frac{y_k}{\pi_k}}{\sum_{k \in S_h} \frac{1}{\pi_k}} = \frac{\hat{t}_{y_h}}{\hat{N}_h}$$

l'estimateur par substitution de la moyenne μ_{yh} dans la post-strate U_h.

Chaque post-strate peut être vue comme un domaine, non pris en compte lors de l'échantillonnage. L'estimateur post-stratifié s'obtient à l'aide d'un redressement par le ratio dans chaque post-strate.

Motivation

L'estimateur post-stratifié est motivé par le modèle

$$y_k = \beta_h + \epsilon_k \quad \text{et} \quad V_m(\epsilon_k) = \sigma_h^2 \text{ dans chaque strate } U_h.$$

C'est un cas particulier de l'estimateur par la régression généralisée, obtenu avec $\mathbf{x}_k = [1(k \in U_1), \ldots, 1(k \in U_H)]^T$ et $\sigma_k^2 = \sigma_h^2$ pour $k \in U_h$. On a :

$$\begin{aligned} \hat{\mathbf{b}}_\pi &= \left[\sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2 \, \pi_k}\right]^{-1} \sum_{k \in S} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \, \pi_k} \\ &\equiv [\tilde{\mu}_{y1}, \ldots, \tilde{\mu}_{yH}]^T, \end{aligned}$$

et

$$e_k = y_k - \hat{\mathbf{b}}_\pi^\top \mathbf{x}_k \equiv y_k - \tilde{\mu}_{yh} \text{ pour } k \in U_h.$$

Propriétés de l'estimateur post-stratifié

L'estimateur par le ratio est approximativement non biaisé pour le total t_y. On a

$$V_p\left[\hat{t}_{y,post}\right] \simeq V_p\left[\hat{t}_{E\pi}\right]$$

avec $E_k = y_k - \mathbf{b}^\top \mathbf{x}_k \equiv y_k - \mu_{yh}$. La variance est donc réduite si la variable y est peu dispersée dans chaque post-strate.

L'estimateur post-stratifié est **calé** sur les effectifs des post-strates :

$$\hat{N}_{hpost} = N_h \qquad \forall h = 1, \ldots, H.$$

On obtient un estimateur de variance :

- en prenant l'estimateur $v[\hat{t}_{E\pi}]$ associé au pds $p(\cdot)$,
- en remplaçant les résidus inconnus E_k par les résidus estimés $e_k = y_k - \tilde{\mu}_{yh}$ pour $k \in S_h$.

Cas du sondage aléatoire simple

L'estimateur post-stratifié se réécrit :

$$\hat{t}_{ypost} = \sum_{h=1}^{H} N_h \times \bar{y}_h \quad \text{avec} \quad \bar{y}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k.$$

Sa variance vaut approximativement

$$V_p\left[\hat{t}_{ypost}\right] \simeq N^2 \frac{1-f}{n} S_E^2 = N^2 \frac{1-f}{n} \underbrace{\sum_{h=1}^{H} \frac{N_h - 1}{N-1} S_{yh}^2}_{S_{y,intra^2}},$$

et peut être estimée par

$$v\left[\hat{t}_{ypost}\right] = N^2 \frac{1-f}{n} S_e^2 = N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{n_h - 1}{n-1} s_{yh}^2.$$

Mise en oeuvre pratique d'un calage

Comment choisir les variables de calage ?

- Les variables auxiliaires les plus explicatives doivent être utilisées pour le calage (sélection avec une PROC GLM, par exemple).
- Les variables utilisées pour concevoir le plan de sondage doivent être utilisées pour le calage (ex : variables de stratification).
- Si le calage est utilisé pour compenser de la non-réponse, les variables explicatives de la probabilité de réponse devraient être incluses dans le calage.

Est-ce que toute l'information auxiliaire doit être utilisée dans le calage ?

En principe, plus on utilise de variables de calage, plus les résidus sont faibles et donc plus la variance de l'estimateur calé diminue. En pratique :

- le nombre de variables de calage doit rester faible devant la taille de l'échantillon,
- les variables les plus explicatives sont généralement suffisantes pour obtenir une forte diminution de la variance.

Peut-on utiliser plusieurs niveaux d'information auxiliaire ?

C'est possible avec la macro CALMAR 2, qui permet d'utiliser jusqu'à trois niveaux d'information auxiliaire. Par exemple, pour une enquête auprès des ménages :

- information auxiliaire sur les unités primaires (ex : les communes),
- information auxiliaire sur les ménages,
- information auxiliaire sur les individus.

La macro SAS CALMAR2

Paramètres pour les tables SAS en entrée

DATAMEN = nom de la table contenant les données de l'échantillon

- Observations : unités échantillonnées,
- Variables : variables de calage, variable identifiante, poids initial.

MARMEN = nom de la table contenant l'information auxiliaire

- Observations : variables de calage,
- Variables : nom de variable, nombre de modalités, marges associées.

Paramètres pour les tables SAS en entrée

POIDS = variable

Variable numérique donnant les poids initiaux des individus de l'échantillon.

PONDQK = variable

Variable numérique de pondération pour les individus de l'échantillon, différente de **POIDS** (utilisée si le modèle (14) est supposé hétéroscédastique).

IDENT = variable

Variable identifiante pour les unités échantillonnées.

Paramètres pour les tables SAS en entrée

PCT = OUI or NON

Si **PCT**=OUI, les marges pour les variables catégorielles de la table **DATAMAR** sont données en pourcentage.

EFFTOT = valeur

Nombre total d'unités dans la population (à renseigner si **PCT**=OUI).

Paramètres pour la méthode de calage

M = 1,2,3 or 4

Fonction de distance :

1. Méthode linéaire
2. Méthode Raking Ratio
3. Méthode Logit
4. méthode linéaire tronquée

LO = valeur

Borne Inférieure pour les rapports de poids (à spécifier si **M**=3 ou 4).

Paramètres pour la méthode de calage

UP = valeur

Borne Supérieure pour les rapports de poids (à spécifier si **M**=3 ou 4).

SEUIL = valeur

Seuil déterminant l'arrêt de l'algorithme de Newton (optionnel).

MAXITER = valeur entière

Nombre maximum d'itérations de l'algorithme de Newton (optionnel).

Paramètres pour les tables SAS en sortie

DATAPOI = nom de la table SAS contenant les poids finaux

- observations : unités échantillonnées non supprimées,
- variables : variable identifiante, poids final.

MISAJOUR = OUI ou NON

Spécifie le traitement de variables en sortie :

- Si **MISAJOUR**=OUI, la variable donnant les poids calés est ajoutée à la table DATAPOI,
- Si **MISAJOUR**=NON, une nouvelle table SAS est créée. L'ancienne table SAS est détruite.

Paramètres pour les tables SAS en sortie

POIDSFIN = variable
Nom de la variable donnant les poids calés.

LABELPOI = label
Label associé à la variable donnant les poids calés.

OBSELI = OUI ou NON
Si **OBSELI**=OUI, crée une table SAS OBSELI avec, pour chaque unité supprimée de l'échantillon d'origine, la variable identifiante, les variables de calage et les poids initiaux.

Un petit exemple

```
data don; input nom $ x $ y $ z pond;cards;
A 1 2 1 10
B 1 1 2 0
C 1 1 3 .
D 2 2 1 11
E 2 2 3 13
F 2 1 2 7
G 2 1 2 8
H 1 1 2 8
I 2 2 2 9
J . 1 2 10
K 2 1 2 14
;run;

data marges; input var $ n mar1 mar2;cards;
x 2 20 60
y 2 30 50
z 0 140 .
;run;
```

```
title « A short example of calibration »;

%CALMAR2(DATAMEN=don,POIDS=pond,IDENT=nom,
         MARMEN=marges,M=2,EDITPOI=oui,OBSELI=oui,
        DATAPOI=sortie,POIDSFIN=pondfin,
         LABELPOI=weighting raking ratio);
```

```
                    ***********************************
                    ***   PARAMETERS OF THE MACRO   ***
                    ***********************************


INPUT TABLE(S) :
 TABLE OF DATAS OF LEVEL 1                DATAMEN   =  DON
    IDENTIFIER OF LEVEL 1                 IDENT     =  NOM
 TABLE OF DATAS OF LEVEL 2                DATAIND   =
    IDENTIFIER OF LEVEL 2                 IDENT2    =
 TABLE OF KISH INDIVIDUALS                DATAKISH  =
 INITIAL WEIGHTING                        POIDS     =  POND
 SCALE FACTOR                             ECHELLE   =  1
 WEIGHTING QK                             PONDQK    =  __UN
 KISH WEIGHTING                           POIDKISH  =
 NUMBER OF KISH INDIVIDUALS BY HOUSEHOLD  NKISH     =
 CONSTANT WEIGHTS BY HOUSEHOLD            EGALPOI   =  NON

TABLE(S) OF MARGINS :
 OF LEVEL 1                               MARMEN    =  MARGES
 OF LEVEL 2                               MARIND    =
 OF KISH LEVEL                            MARKISH   =
 MARGINS IN PERCENT                       PCT       =  NON
 TOTAL NUMBER IN THE POPULATION :
  OF ELEMENTS OF LEVEL 1                  POPMEN    =
  OF ELEMENTS OF LEVEL 2                  POPIND    =
  OF KISH ELEMENTS                        POPKISH   =
```

```
TABLE(S) WITH FINAL WEIGHTS
 OF LEVEL 1                              DATAPOI       =  SORTIE
 OF LEVEL 2                              DATAPOI2      =
 OF KISH LEVEL                           DATAPOI3      =
 UPDATING OF TABLE(S) DATAPOI(2)(3)      MISAJOUR      =  OUI
 FINAL WEIGHTS                           POIDSFIN      =  PONDFIN
 LABEL OF FINAL WEIGHTS                  LABELPOI      =  RAKING RATIO WEIGHTINGIO
 FINAL WEIGHTS FOR KISH INDIVIDUALS      POIDSKISHFIN  =
 LABEL OF KISH INDIVIDUALS               LABELPOIKISH  =
 CONTENT OF TABLE(S) DATAPOI(2)(3)       CONTPOI       =  OUI

ÉDITING RESULTS                          EDITION       =  3
ÉDITING WEIGHTS                          EDITPOI       =  OUI
STATISTICS ON WEIGHTS                    STAT          =  OUI

CONTROLS                                 CONT          =  OUI
TABLE OF REMOVED OBSERVATIONS            OBSELI        =  OUI
SAS NOTES                                NOTES         =  NON
```

COMPARISON BETWEEN MARGINS ESTIMATED FROM THE SAMPLE (INITIAL WEIGHTS)
AND REAL MARGINS IN THE POPULATION (CALIBRATION MARGINS)

VARIABLE	MODALITY	SAMPLE MARGIN	POPULATION MARGIN	SAMPLE PERCENT	POPULATION PERCENT
X	1	18	20	22.50	25.00
	2	62	60	77.50	75.00
Y	1	37	30	46.25	37.50
	2	43	50	53.75	62.50
Z		152	140	.	.

METHOD : RAKING RATIO

FIRST SUMMARY OF ALGORITHM :

VALUE OF THE TERMINATION CRITERION AND NUMBER OF NEGATIVE WEIGHTS AFTER EACH ITERATION

ITERATION	TERMINATION CRITERION	NEGATIVE WEIGHTS
1	0.58354	0
2	0.09811	0
3	0.00310	0
4	0.00000	0

METHODE : RAKING RATIO

SECOND SUMMARY OF ALGORITHM :

COEFFICIENTS OF VECTOR LAMBDA OF LAGRANGE MULTIPLIERS

AFTER EACH ITERATION

VARIABLE	MODALITY	LAMBDA1	LAMBDA2	LAMBDA3	LAMBDA4
X	1	0.73880	0.64924	0.64601	0.64601
X	2	0.80537	0.71694	0.71381	0.71380
Y	1	-0.28875	-0.26971	-0.26887	-0.26887
Y	2	.	.	.	.
Z		-0.34571	-0.32124	-0.32019	-0.32019

MÉTHODE : RAKING RATIO

COMPARISON BETWEEN FINAL MARGINS IN THE SAMPLE (WITH FINAL WEIGHTS) AND MARGINS IN THE POPULATION (CALIBRATION MARGINS)

VARIABLE	MODALITY	SAMPLE MARGIN	POPULATION MARGIN	SAMPLE PERCENT	POPULATION PERCENT
X	1	20	20	25.00	25.00
	2	60	60	75.00	75.00
Y	1	30	30	37.50	37.50
	2	50	50	62.50	62.50
Z		140	140	.	.

METHOD : RAKING RATIO

RATIOS OF WEIGHTS (FINAL WEIGHTS / INITIAL WEIGHTS)

FOR EACH COMBINATION OF VALUES OF VARIABLES

Obs	x	y	z	NUMBER COMBINATION	RATIO OF WEIGHTS
1	1	1	2	1	0.76855
2	1	2	1	1	1.38516
3	2	1	2	3	0.82247
4	2	2	1	1	1.48233
5	2	2	2	1	1.07618
6	2	2	3	1	0.78132

METHOD : RAKING RATIO

STATISTICS FOR RATIOS OF WEIGHTS (= FINAL WEIGHTS / INITIAL WEIGHTS)
AND FINAL WEIGHTS

PROC UNIVARIATE

Variable : _F_ (RATIO OF WEIGHTS)

Statistical indicators

Position		Variability	
Mean	0.995118	Standard Deviation	0.28848
Median	0.822468	Variance	0.08322
Mode	0.822468	Range	0.71377
		Interquartile Interval	0.42878

Quantile	Estimation
100% Max	1.482327
99%	1.482327
95%	1.482327
90%	1.482327
75% Q3	1.230671
50% Median	0.822468
25% Q1	0.801894
10%	0.768553
5%	0.768553
1%	0.768553
0% Min	0.768553

MÉTHOD : RAKING RATIO

STATISTICS FOR RATIOS OF WEIGHTS (= FINAL WEIGHTS / INITIAL WEIGHTS)
AND FINAL WEIGHTS

PROC UNIVARIATE

Variable : __WFIN (FINAL WEIGHT)

Statistical Indicators

Position		Variability	
Mean	10.00000	Standard Deviation	3.80875
Median	9.92141	Variance	14.50657
Mode	.	Range	10.54832
		Interquartile Interval	6.31898

Quantile	Estimation
100% Max	16.30560
99%	16.30560
95%	16.30560
90%	16.30560
75% Q3	12.68306
50% Médiane	9.92141
25% Q1	6.36409
10%	5.75728
5%	5.75728
1%	5.75728
0% Min	5.75728

METHOD : RAKING RATIO

MEAN RATIOS OF WEIGHTS (FINAL WEIGHTS / INITIAL WEIGHTS) FOR EACH VALUE OF VARIABLES

VARIABLE	MODALITY	NUMBER OF OBSERVATIONS OF LEVEL 1	RATIO OF WEIGHTS
X	1	2	1.07686
X	2	6	0.96787
Y	1	4	0.80899
Y	2	4	1.18125
TOTAL		8	0.99512

```
                    ************************
                    ***      RESULTS     ***
                    ************************
*

*   ***********************************
*   INPUT TABLE : DON
*   ***********************************
 *  NUMBER OF OBSERVATIONS IN INPUT TABLE          :            11
*   NUMBER OF REMOVED OBSERVATIONS                 :            3
*   NUMBER OF KEPT OBSERVATIONS                    :            8
*
*   VARIABLE OF WEIGHTS : POND
*
*   NUMBER OF CATEGORICAL VARIABLES : 2
*   LIST OF CATEROGICAL VARIABLES AND NUMBER OF MODALITIES :
      x (2) y (2)
*
*   SUM OF INITIAL WEIGHTS                                    : 80
*   SIZE OF THE POPULATION                                    : 80
*
*   NUMBER OF NUMERICAL VARIABLES: 1
*   LIST OF NUMERICAL VARIABLES : z
*
*   USED METHOD : RAKING RATIO
*   CALIBRATION REACHED INTO 4 ITERATIONS
*   WEIGHTS STOCKED INTO VARIABLE PONDFIN OF TABLE SORTIE
```

Bibliographie

- Ardilly, P. (2005). *Panorama des principales méthodes d'estimation sur petits domaines*. Actes des Journées de Méthodologie Statistique, Insee.
- Ardilly, P. (2006), *Les Techniques de Sondage*, Technip, Paris.
- Ardilly, P., et Tillé, Y. (2003), *Exercices corrigés de méthodes de sondage Sondage*, Technip, Paris.
- Cochran, W.G (1977), *Sampling Techniques*, Wiley, New-York.
- De Peretti, P. et al (2006). *L'enquête sans-domicile 2001*. Insee Méthodes, 116, Paris.
- Deville, J-C. (1991). *Une théorie des enquêtes par quotas*. Techniques d'Enquête, 17, 177-195.
- Hajek, J. (1964). *Asymptotic theory of rejective sampling with varying probabilities from a finite population*. Annals of Mathematical Statistics, 35, 1491-1523.

Bibliographie

- Loonis, V. (2009). *La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation*. Actes des Journées de Méthodologie Statistique, Paris.
- Rao, J.N.K (2003). *Small Area Estimation*. New-York, Wiley.
- Särndal, C.-E., and Swensson, B., and Wretman, J.H. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New-York.
- Sautory, O., et Le Guennec, J. (2003). *La macro CALMAR2 : Redressement d'un échantillon par calage sur marges*, Insee.
- Schreuder, H.T., and Gregoire, T.G., and Wood, G.B. (1993). *Sampling Methods for Multiresource Forest Inventry*, Wiley, New-York.
- Tillé, Y. (2006). *Sampling algorithms*, Springer, New-York.