Sondage à probabilités inégales

Myriam Maumy-Bertrand[1]

[1] IRMA, Université de Strasbourg
Strasbourg, France

Master 1ère Année 06-11-2014

Références

Ce chapitre s'appuie essentiellement sur deux ouvrages :

1. « Manuel de Sondages »
   de R. Clairin et P. Brion,
   téléchargeable à
   `http://www.ceped.org/cdrom/integral_publication_1988_2002/manuels/pdf/manuels_cpd_03.pdf`
2. « Méthodes statistiques des sondages »
   de J.-M. Grosbras,
   Economica, 1987.

Sommaire

Principe

Dans certains cas, nous pouvons décider d'accorder à certaines unités une probabilité plus forte d'être sélectionnées que d'autres.

Remarque

L'usage de **sondages à probabilités inégales** est particulièrement intéressant lorsque la plupart des variables sont liées par un effet de taille.

Sommaire

Exemples

1. Pour des enquêtes auprès des entreprises, nous pouvons tirer les unités avec une probabilité proportionnelle, par exemple, à leur nombre de salariés, à leur chiffre d'affaires...
2. Le sondage à probabilités inégales est souvent utilisé au premier degré d'un tirage à plusieurs degrés :
   - tirage de communes avec probabilité proportionnelle à leur population
   - puis tirage de ménages ou d'individus au deuxième degré.

Remarque

Sur les deux exemples précédents, nous remarquons que, la probabilité de tirage d'une unité est, en général, proportionnelle à une mesure de taille.

L'idée est simple !

Plus une unité est « grande », plus elle apporte de l'information. Par conséquent il est important de la sélectionner.

Sommaire

Remarques

- Dans ce chapitre et plus particulièrement dans ce paragraphe, nous traiterons les **sondages à probabilités inégales avec remise** (PIAR).
- Le cas des **sondages à probabilités inégales sans remise** (PISR) sera traité au paragraphe suivant de ce chapitre.

Définition

Un sondage est dit à PIAR si chaque unité i de la population U a la probabilité P_i d'être tirée à chacun des tirages.
De plus, souvent l'échantillon est de taille fixe n et nous avons :

$$\sum_{i=1}^{N} P_i = 1.$$

Remarque

P_i est souvent proportionnelle à une mesure de la taille de l'unité i. Si X_i est sa taille, alors nous choisissons :

$$P_i = \frac{X_i}{\left(\sum_{i=1}^{N} X_i\right)} \quad \text{et} \quad \sum_{i=1}^{N} P_i = 1.$$

Définition

L'estimateur de la moyenne μ dans un sondage à probabilités inégales avec remise se définit par :

$$\widehat{\mu}_{PIAR} = \frac{1}{nN} \sum_{i=1}^{n} \frac{Y_i}{P_i},$$

où Y_i est la variable aléatoire pour l'unité qui sera sélectionnée au i-ème tirage et P_i sa probabilité d'être sélectionnée à chaque tirage.

Propriété

Nous montrons, par calcul, que **cet estimateur sans biais**, *i.e. :*

$$\mathbb{E}\left(\widehat{\mu}_{PIAR}\right) = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N} Y_k = \mu.$$

Remarque

Pour démontrer ce résultat, il suffit de noter que :

$$\mathbb{E}\left(\frac{Y_i}{P_i}\right) = \sum_{k}^{N} Y_k.$$

Définition

L'estimateur du total T dans un sondage à probabilités inégales avec remise se définit par :

$$\widehat{T}_{PIAR} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{P_i},$$

où Y_i est la variable aléatoire pour l'unité qui sera sélectionnée au i-ème tirage et P_i sa probabilité d'être sélectionnée à chaque tirage.

Propriété

Nous montrons, par calcul, que **cet estimateur sans biais**, *i.e. :*

$$\mathbb{E}\left(\widehat{T}_{PIAR}\right) = \sum_{k=1}^{N} Y_k = T.$$

Remarque

Savoir démontrer ce résultat.

Propriété

La variance de $\widehat{\mu}_{PIAR}$ est égale à :

$$\begin{aligned} \mathrm{Var}\left(\widehat{\mu}_{PIAR}\right) &= \frac{1}{nN^2}\sum_{k=1}^{N} P_k\left(\frac{Y_k}{P_k} - \left(\sum_{k=1}^{N} Y_k\right)\right)^2 \\ &= \frac{1}{nN^2}\left(\sum_{k=1}^{N}\frac{Y_k^2}{P_k} - T^2\right). \end{aligned}$$

Remarque

Pour établir ce résultat, il suffit de noter que :

$$\mathrm{Var}\left(\frac{Y_i}{P_i}\right) = \sum_{k=1}^{N} P_k \left(\frac{Y_k}{P_k} - \left(\sum_{k=1}^{N} Y_k\right)\right)^2 .$$

Propriété

La variance de $\widehat{T}_{PIAR}$ est égale à :

$$\begin{aligned} \mathrm{Var}\left(\widehat{T}_{PIAR}\right) &= \frac{1}{n}\sum_{k=1}^{N} P_k \left(\frac{Y_k}{P_k} - \left(\sum_{k=1}^{N} Y_k\right)\right)^2 \\ &= \frac{1}{n}\left(\sum_{k=1}^{N} \frac{Y_k^2}{P_k} - T^2\right). \end{aligned}$$

Propriété

La variance de $\widehat{\mu}_{PIAR}$ peut être estimée sans biais à partir de l'échantillon par :

$$\widehat{\mathrm{Var}\,(\widehat{\mu}_{PIAR})} = \frac{1}{N^2 n(n-1)} \sum_{i=1}^{n} \left(\frac{Y_i}{P_i} - \widehat{T}_{PIAR} \right)^2 .$$

Propriété

La variance de $\widehat{T}_{PIAR}$ peut être estimée sans biais à partir de l'échantillon par :

$$\widehat{\mathrm{Var}\left(\widehat{T}_{PIAR}\right)} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{Y_i}{P_i} - \widehat{T}_{PIAR} \right)^2 .$$

Choix optimal des P_k :

D'après la formule de $\mathrm{Var}\left(\widehat{T}_{PIAR}\right)$, il est évident que la variance est nulle, c'est-à-dire minimale, si :

$$P_k = \frac{Y_k}{\sum_{k=1}^{N} Y_k}.$$

Remarques

- Bien entendu, nous ne connaissons pas toutes les variables Y_i (sinon il n'y aurait pas besoin de sondage), mais ce résultat montre que nous pouvons avoir des résultats très précis si les P_k sont choisis en fonction d'une variable liée à la variable Y.
- D'ailleurs G. Saporta dit dans son cours : « Tirage à probabilités inégales : une manière d'utiliser de l'information auxiliaire ».

Exemple

Taille des entreprises pour une production.

Rappel de la variance du total dans le cas d'un tirage à PEAR :

$$\mathrm{Var}\left(\widehat{T}_{PEAR}\right) = N^2\frac{\sigma^2}{n} = \frac{1}{n}\left(\sum_{k=1}^{N} NY_k^2 - T^2\right).$$

Effet de sondage : comparaison avec le sondage à PEAR

$$\mathrm{Var}\left(\widehat{T}_{PEAR}\right) > \mathrm{Var}\left(\widehat{T}_{PIAR}\right) \quad \Leftrightarrow \quad \sum_{k=1}^{N} NY_k^2 > \sum_{k=1}^{N} Y_k^2/P_k$$

$$\Leftrightarrow \quad \sum_{k=1}^{N} Y_k^2\left(1/P_k - 1/(1/N)\right) < 0.$$

L'inégalité sera d'autant plus vraie si :

$$P_k > 1/N \quad \text{si } Y_k^2 \text{ est grand}$$
$$P_k < 1/N \quad \text{si } Y_k^2 \text{ est petit.}$$

c'est-à-dire si Y_k^2, donc si la variable Y_k est corrélée positivement avec la probabilité P_k.

Remarque

Ce résultat est conforme à l'intuition.

Sommaire

Généralités

Le problème se complique du fait que chaque tirage modifie les conditions des tirages suivants.

Ainsi, en plus des probabilités de sortie au premier tirage, il faut connaître les probabilités de sortie des unités U_j au deuxième tirage, sachant que l'unité U_i est sortie au premier tirage, et ainsi de suite.

Définition

Nous faisons appel à un autre estimateur : **l'estimateur de Horvitz-Thompson**, introduit en 1952. Nous l'appelons aussi l'**estimateur par valeurs dilatées** ou le π**-estimateur**.
Le point de départ de cette approche développée pour les sondages sans remise (cf T.D. n°1, Exercice 8) est la probabilité d'inclusion :

- π_i : probabilité que l'unité i appartienne à l'échantillon $\mathcal{S}$ ou encore probabilité d'inclusion d'ordre 1,
- π_{kl} : probabilité que les unités k et l appartiennent simultanément à l'échantillon $\mathcal{S}$ ou encore probabilité d'inclusion d'ordre 2.

Remarques sur les probabilités d'inclusion

- Ces probabilités d'inclusion sont comprises entre 0 et 1.
- Comme la taille de l'échantillon n est fixée, ces probabilités respectent les deux égalités suivantes :

$$\sum_{k=1}^{N} \pi_k = n.$$

$$\sum_{k \neq l} \sum_{k,l=1}^{N} \pi_{kl} = n(n-1).$$

Définition

L'estimateur de Horvitz-Thompson de la moyenne μ dans un sondage à probabilités inégales sans remise se définit par :

$$\widehat{\mu}_{PISR} = \frac{1}{N} \sum_{i=1}^{n} \frac{Y_i}{\pi_i},$$

où π_i désigne la probabilité d'inclusion d'ordre 1.

Propriété

Nous montrons, que si $\pi_k > 0$ *pour tout* $k \in U$, *alors* **cet estimateur sans biais**, *i.e. :*

$$\mathbb{E}\left(\widehat{\mu}_{PISR}\right) = \mu_Y.$$

Remarque

Savoir démontrer ce résultat.

Définition

L'estimateur de Horvitz-Thompson du total T dans un sondage à probabilités inégales sans remise se définit par :

$$\widehat{T}_{PISR} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i},$$

où π_i désigne la probabilité d'inclusion d'ordre 1.

Propriété

Nous montrons, que si $\pi_k > 0$ *pour tout* $k \in U$, *alors* **cet estimateur sans biais**, *i.e. :*

$$\mathbb{E}\left(\widehat{T}_{PISR}\right) = T_Y.$$

Remarque

Savoir démontrer ce résultat.

Remarques sur l'estimateur de Horvitz-Thompson

- C'est un estimateur linéaire.
- Les poids de sondage ne dépendent pas de l'échantillon.
- Il permet d'estimer la taille N de la population.
- Il est valable quel que soit le plan de sondage.
- Il généralise les résultats du sondage aléatoire simple sans remise de taille fixe n.
- Si certaines probabilités d'inclusion sont nulles, alors l'estimateur de Horvitz-Thompson est biaisé. Ce biais ne dépend que des unités qui n'ont aucune chance d'être échantillonnées : nous parlons de problème de couverture.

Propriété

Dans le cas général et si $\pi_k > 0$ pour tout $k \in U$, alors la variance de l'estimateur d'Horvitz-Thompson de la moyenne $\widehat{\mu}_{PISR}$ est égale à :

$$\mathrm{Var}\left(\widehat{\mu}_{PISR}\right) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \left(\pi_{kl} - \pi_k \pi_l\right) \frac{Y_k}{\pi_k} \frac{Y_l}{\pi_l}.$$

Propriété

Si $\pi_k > 0$ pour tout $k \in U$ et si le plan est de taille fixe, alors Sen, Yates et Grundy ont montré que la variance de l'estimateur d'Horvitz-Thompson de la moyenne peut aussi s'écrire :

$$\mathrm{Var}\left(\widehat{\mu}_{PISR}\right) = -\frac{1}{2N^2} \sum_{k \neq l} \sum_{k,l=1}^{N} \left(\pi_{kl} - \pi_k \pi_l\right) \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2 .$$

Propriété

Dans le cas général et si $\pi_k > 0$ pour tout $k \in U$, alors la variance de l'estimateur d'Horvitz-Thompson du total $\widehat{T}_{PISR}$ est égale à :

$$\mathrm{Var}\left(\widehat{T}_{PISR}\right) = \sum_{k \in U} \sum_{l \in U} \left(\pi_{kl} - \pi_k \pi_l\right) \frac{Y_k}{\pi_k} \frac{Y_l}{\pi_l}.$$

Propriété

Si $\pi_k > 0$ pour tout $k \in U$ et si le plan est de taille fixe, alors Sen, Yates et Grundy ont montré que la variance de l'estimateur d'Horvitz-Thompson du total peut aussi s'écrire :

$$\mathrm{Var}\left(\widehat{T}_{PISR}\right) = -\frac{1}{2}\sum_{k\neq l}^{N}\sum_{k,l=1}^{N}(\pi_{kl}-\pi_k\pi_l)\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2.$$

Propriété

Dans le cas général *et si* $\pi_{kl} > 0$ *pour tous* k *et* l *de* U*, alors l'estimateur de la variance de l'estimateur de Horvitz-Thompson de la moyenne se définit par :*

$$\widehat{\mathrm{Var}\left(\widehat{\mu}_{PISR}\right)} = \frac{1}{N^2} \sum_{k \in \mathcal{S}} \sum_{l \in \mathcal{S}} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{Y_k}{\pi_k} \frac{Y_l}{\pi_l}.$$

Propriété

Si $\pi_{kl} > 0$ pour tous k et l de U et **si le plan est de taille fixe**, *alors l'estimateur de la variance de l'estimateur de Horvitz-Thompson de la moyenne se définit par :*

$$\widehat{\mathrm{Var}\,(\widehat{\mu}_{PISR})} = -\frac{1}{2N^2}\sum_{k\neq l}\sum_{k\in S, l\in S}\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2.$$

Propriété

Dans le cas général *et si* $\pi_{kl} > 0$ *pour tous* k *et* l *de* U*, alors l'estimateur de la variance de l'estimateur de Horvitz-Thompson du total se définit par :*

$$\widehat{\mathrm{Var}\left(\widehat{T}_{PISR}\right)} = \sum_{k\in\mathcal{S}}\sum_{l\in\mathcal{S}} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \frac{Y_k}{\pi_k} \frac{Y_l}{\pi_l}.$$

Propriété

Si $\pi_{kl} > 0$ pour tous k et l de U et **si le plan est de taille fixe**, *alors l'estimateur de la variance de l'estimateur de Horvitz-Thompson du total se définit par :*

$$\widehat{\mathrm{Var}\left(\widehat{T}_{PISR}\right)} = -\frac{1}{2}\sum_{k\neq l}\sum_{k\in S, l\in S}\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2.$$

Comparaison de ces deux estimateurs

- Dans le cas où est le plan est de taille fixe, nous disposons généralement de deux estimateurs concurrents et différents.
- Ils sont tous les deux sans biais dès que tous les $\pi_{kl} > 0$ quels que soient les individus k et l de la population.
- Les deux estimateurs peuvent prendre des valeurs négatives, mais il existe une condition suffisante pour que le second estimateur soit positif. Cette condition, dite condition de Sen-Yates-Grundy, est :

$$\forall k \neq l \in U, \pi_{kl} - \pi_k \pi_l \leqslant 0.$$

Définition

Lorsque la taille de la population N est inconnue, nous pouvons estimer la moyenne de Y avec l'estimateur de Hájek défini par :

$$\widehat{\mu}_H = \frac{\widehat{T}_\pi}{\widehat{N}_\pi} = \frac{1}{\widehat{N}_\pi} \sum_{i \in \mathcal{S}} \frac{Y_i}{\pi_i} = \frac{\sum_{i \in \mathcal{S}} \frac{Y_i}{\pi_i}}{\sum_{i \in \mathcal{S}} \frac{1}{\pi_i}}.$$

Remarques

- L'estimateur de Hájek est un ratio de deux π-estimateurs.
- Il est biaisé : nous montrons que le biais est asymptotiquement nul et nous le considérons négligeable lorsque la taille de l'échantillon est grande.
- Sa précision peut s'avérer supérieur à celle de l'estimateur de Horvitz-Thompson.

Comment faisons-nous dans la pratique ?

Dans la pratique d'un sondage à probabilités inégales sans remise, nous nous fixons un « jeu » de π_i et un algorithme respectant ce jeu de probabilités.
En ce qui concerne l'algorithme, nous renvoyons le lecteur au dernier paragraphe de ce chapitre intitulé « Méthodes de tirage ».

Comment calculons-nous ce « jeu » de π_i

Si nous disposons d'une variable auxiliaire $X_i > 0$, $i \in U$, « suffisamment » proportionnelle à la variable d'intérêt Y_i, il est souvent intéressant de sélectionner les unités à probabilités inégales proportionnelles aux X_i.

Pour ce faire, nous calculons d'abord les probabilités d'inclusion d'ordre 1 suivant la formule suivante :

$$\pi_i = n \frac{X_i}{\sum_{k \in U} X_k}.$$

Remarque

Si l'expression ci-dessus fournit des $\pi_i > 1$, les unités correspondantes sont sélectionnées d'office dans l'échantillon (avec une probabilité d'inclusion égale à 1).
Nous recalculons ensuite les π_i selon la formule ci-dessus sur les unités restantes.

Suite : calcul des π_{ij}

Nous calculons alors les π_{ij} ou nous les déterminons de manière approximative (car dans certains cas, le calcul rigoureux est impossible). Nous pouvons ainsi calculer la précision (en calculant la variance) des différents estimateurs de Horvitz-Thompson.

Remarque

Cette approche est une approche générale, pas seulement limitée aux sondages à probabilités inégales. Cette approche est présentée dans ce chapitre car étant la seule utilisable dans un sondage PISR.

Sommaire

Méthodes de tirage

Comment confectionner un échantillon dont les unités qui vont le caractériser n'ont pas la même probabilité, c'est-à-dire ont des probabilités inégales ?

Tirages systématiques

C'est de loin la méthode la plus économique et la plus simple à utiliser.
Cette procédure est de taille fixe n. Nous supposons que nous connaîssons les

$$0 < \pi_i < 1, \quad i = 1, \ldots, N \quad \text{avec} \quad \sum_{i=1}^{N} \pi_i = n.$$

Nous voulons sélectionner un échantillon de taille fixe n avec des probabilités d'inclusion proportionnelles aux π_i.

Suite

Nous définissons

$$\begin{aligned} C_0 &= 0 \\ C_1 &= \pi_1 \\ C_2 &= C_1 + \pi_2 \\ \dots &= \dots \\ C_N &= C_{N-1} + \pi_N. \end{aligned}$$

Nous générons ensuite une v.a.u. dans $[0, 1]$, u, qui donne le « départ aléatoire ».

Suite et fin

La première unité sélectionnée i_1 est telle que :

$$C_{i_1-1} \leqslant u < C_{i_1}.$$

La k-ième unité sélectionnée i_k est telle que :

$$C_{i_k-1} \leqslant u < C_{i_k}.$$

La n-ième unité sélectionnée i_n est telle que :

$$C_{i_n-1} \leqslant u < C_{i_n}.$$

Exemple du tirage systématique (d'après le livre « Méthodes statistiques des sondages » de Jean-Marie Grosbras) :

$N = 8$ et $n = 3$.

U_i	$100\,\pi_i$	$100\,C_i$
1	15	15
2	81	96
3	26	122
4	42	164
5	20	184
6	16	200
7	45	245
8	55	300

Suite de l'exemple

Nous prenons :

$$u = 0,36.$$

Par conséquent, nous avons :

$$\begin{aligned} 100\,(u+0) &= 36 \\ 100\,(u+1) &= 136 \\ 100\,(u+2) &= 236. \end{aligned}$$

36 se situe entre 15 et 96 donc on choisit U_2.
136 se situe entre 122 et 164 donc on choisit U_4.
236 se situe entre 200 et 245 donc on choisit U_7.

Remarques

- Hartley et Rao (1962) ont montré que cette procédure respecte bien les π_i voulues et ont fourni, après des calculs laborieux, des approximations de $\mathrm{Var}\left(\widehat{T}_{PI}\right)$ et de $\widehat{\mathrm{Var}\left(\widehat{T}_{PI}\right)}$.
- Pour plus de renseignements nous renvoyons au livre « Méthodes statistiques des sondages », de J.M. Grosbras.

Conclusion sur la méthode des tirages systématiques

Cette procédure de sélection est très simple et des variances approximatives assez faciles à calculer.

Méthode des chiffres cumulés

Cette méthode provient du livre « Manuel de sondages » de R. Clairin et P. Brion.

- Supposons que nous ayons une liste de 207 villes avec une estimation de leur population.
- Nous voulons enquêter 21 villes. Par conséquent $n = 21$.
- Nous calculons d'abord la population cumulée correspondant à chaque ville (cf le tableau du slide 57).
- Pour la dernière ville, elle vaut 58 626.

- Nous tirons au hasard 21 nombres à 5 chiffres inférieurs ou égaux à 58 626.
- Ceci permet de sélectionner les unités pour lesquelles ces nombres appartiennent à la « portion de population cumulée » correspondante, donc avec une probabilité proportionnelle à leur population.
- Pour visualiser ceci, nous pouvons imaginer que nous avons distribué à chaque habitant un billet de loterie numéroté et qu'une ville est tirée si un habitant de cette ville a un billet gagnant.

Exemple de la méthode des chiffres cumulés

Ville	Population par ville	Population cumulée
1	531	531
2	177	708
3	348	1056
4	235	1291
5	290	1581
6	124	1705
...	...	...
205	425	58254
206	219	58473
207	153	58626

Suite de l'exemple

- Supposons que nous ayons tiré : 937, 58 302. Ces deux nombres désignent respectivement les villes 3 et 206.
- Supposons que nous tirons ensuite 727, la ville 3 est à nouveau sélectionnée.
- Ceci induit les probabilités inégales P_i associées à chacune des villes.
- Nous pouvons améliorer la procédure en rangeant par taille les unités, et en procédant à un tirage systématique dans les chiffres cumulés.
- Nous obtenons ainsi une répartition « satisfaisante » de l'échantillon par rapport au critère de tri choisi.

Remrque

Pour d'autres méthodes de tirage dans le cas de sondage à probabilités inégales, nous renvoyons aux deux livres suivants :

1. « Méthodes statistiques des sondages »,
   de Jean-Marie Grosbras,
   aux éditions Economica.
2. « Théorie des sondages »,
   de Yves Tillé,
   aux éditions Dunod.