Sondage à probabilités inégales

Myriam Maumy[1]

[1]IRMA, Université Louis Pasteur
Strasbourg, France

Master 1ère Année 13-03-2006

Ce chapitre s'appuie essentiellement sur deux ouvrages :

1. "Manuel de Sondages"
   de Rémy Clairin et Philippe Brion,
   téléchargeable sur
   http ://ce-ped.cirad.fr/activite/publi/integral/html/manuels/pdf/
   puis "manuels cpd 03.pdf".

2. "Méthodes statistiques des sondages"
   de Jean-Marie Grosbras,
   aux éditions Economica.

Sommaire

Principe :

Dans certains cas, on peut 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, on peut 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 (voir chapitre 6) :*
   - *tirage de communes avec probabilité proportionnelle à leur population*
   - *puis tirage de ménages ou d'individus au deuxième degré.*

Remarque

Sur les 2 exemples précédents, on remarque 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

Remarque

Dans ce chapitre et plus particulièrement dans ce paragraphe, on traitera les **sondages à probabilités inégales avec remise** (PIAR).

Remarque

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, l'échantillon est de taille n et on a :

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

Remarque

P_i est souvent proportionnel à une mesure de la taille de l'unité i. Si Y_i est sa taille, alors on choisit

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

Définition

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

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

où x_i est la valeur de la variable X pour l'unité sélectionnée au ième tirage et P_i sa probabilité d'être sélectionnée à chaque tirage.

Propriété

On montre, par calcul, que **cet estimateur sans biais**, *i.e.*

$$\mathbb{E}\left[\widehat{\mu}_{PIAR}\right] = \frac{1}{N}\sum_{i=1}^{N} X_i = \mu.$$

Définition

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

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

Remarque

On tient compte des probabilités de tirage différentes pour produire l'estimation du total.

Propriété

On montre, par calcul, que **cet estimateur sans biais**, *i.e.*

$$\mathbb{E}\left[\widehat{T}_{PIAR}\right] = \sum_{i=1}^{N} X_i = T.$$

Remarque

Démontrer ce résultat.

Propriété

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

$$\begin{aligned} \mathrm{Var}\left[\widehat{\mu}_{PIAR}\right] &= \frac{1}{nN^2}\sum_{i=1}^{N} P_i\left(\frac{X_i}{P_i} - \left(\sum_{i=1}^{N} X_i\right)\right)^2 \\ &= \frac{1}{nN^2}\left(\sum_{i=1}^{n}\frac{x_i^2}{P_i} - T^2\right). \end{aligned}$$

Propriété

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

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

Propriété

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

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

Propriété

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

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

Choix optimal des P_i :

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

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

Remarque

Bien entendu, on ne connait pas les X_i (sinon il n'y aurait pas besoin de sondage), mais ce résultat montre qu'on peut avoir des résultats très précis si les P_i sont choisis en fonction d'une variable liée aux X_i.

Exemple

Taille des entreprises pour une production.

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

$$\mathrm{Var}\left[\widehat{T}_{pear}\right] = N^2\frac{\sigma^2}{n} = \frac{1}{n}\left(\sum_{i=1}^{N} NX_i^2 - X^2\right).$$

Comparaison avec les sondages PEAR

$$\mathrm{Var}\left[\widehat{T}_{pear}\right] > \mathrm{Var}\left[\widehat{T}_{PIAR}\right] \quad \Leftrightarrow \quad \sum_{i=1}^{N} NX_i^2 > \sum_{i=1}^{N} X_i^2/P_i$$

$$\Leftrightarrow \quad \sum_{i=1}^{N} X_i^2\left(1/P_i - 1/(1/N)\right) < 0.$$

L'inégalité sera d'autant plus vraie si :

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

c-à-d si X_i^2, donc X_i est corrélé positivement avec P_i.

Remarque

Ce résultat est conforme à l'intuition.

Sommaire

Généralités

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

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

On fait appel à une autre approche que l'on va présenter rapidement : **les estimateurs de Horvitz-Thompson**.

Le point de départ de cette approche développée pour les sondages SR est la probabilité d'inclusion :

π_i : probabilité que l'unité i appartienne à l'échantillon
ou encore probabilité d'inclusion d'ordre 1,

π_{ij} : probabilité que les unités i et j appartiennent simultanément à l'échantillon
ou encore probabilité d'inclusion d'ordre 2.

Comment calculer ces probabilités d'inclusion π_i ?

Si l'on dispose d'une variable auxiliaire $y_i > 0$, $i \in U$, "suffisamment" proportionnelle à la variable x_i, il est souvent intéressant de sélectionner les unités à probabilités inégales proportionnelles aux y_i.

Pour ce faire, on calcule d'abord les probabilités d'inclusion suivant la formule suivante :

$$\pi_i = n \frac{y_i}{\sum_{l \in U} y_l}.$$

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).
On recalcule ensuite les π_i selon la formule ci-dessus sur les unités restantes.

Remarques

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

$$\sum_{i=1}^{N} \pi_i = n.$$

$$\sum_{j \neq k} \pi_{jk} = n(n-1).$$

Définition

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

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

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

Propriété

On montre, par calcul, que **cet estimateur sans biais**, *i.e.*

$$\mathbb{E}\left[\widehat{\mu}_{PISR}\right] = \mu.$$

Définition

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

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

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

Propriété

On montre, par calcul, que **cet estimateur sans biais**, *i.e.*

$$\mathbb{E}\left[\widehat{T}_{PISR}\right] = T.$$

Remarque

Démontrer ce résultat.

Propriété

Si l'échantillon est de taille fixe n, alors la variance de $\widehat{\mu}_{PISR}$ est égale à :

$$\mathrm{Var}\left[\widehat{\mu}_{PISR}\right] = \frac{1}{2N^2} \sum_{j \neq k} \sum_{j=1,\ldots,N, k=1,\ldots,N} \left(\pi_j \pi_k - \pi_{jk}\right) \left(\frac{y_j}{\pi_j} - \frac{x_k}{\pi_k}\right)^2 .$$

Propriété

Si l'échantillon est de taille fixe n, alors la variance de $\widehat{T}_{PISR}$ est égale à :

$$\mathrm{Var}\left[\widehat{T}_{PISR}\right] = \frac{1}{2}\sum_{j\neq k}\sum_{j=1,\ldots,N,k=1,\ldots,N}\left(\pi_j\pi_k - \pi_{jk}\right)\left(\frac{y_j}{\pi_j} - \frac{x_k}{\pi_k}\right)^2.$$

Propriété

Un estimateur de la variance de l'estimateur de Horvitz-Thompson de $\widehat{\mu}_{PISR}$ se définit par :

$$\widehat{\mathrm{Var}\,[\widehat{\mu}_{PISR}]} = \frac{1}{2N^2} \sum_{i \neq j} \sum_{i \in S, j \in S} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j} \right)^2 .$$

Dans la pratique d'un tel sondage à probabilités inégales sans remise, on se fixe un "jeu" de π_i et un algorithme respectant ce jeu de probabilités.

Propriété

Un estimateur de la variance de l'estimateur de Horvitz-Thompson de $\widehat{T}_{PISR}$ se définit par :

$$\widehat{\mathrm{Var}\left[\widehat{T}_{PISR}\right]} = \frac{1}{2}\sum_{i\neq j}\sum_{i\in S, j\in S}\frac{\left(\pi_i\pi_j-\pi_{ij}\right)}{\pi_{ij}}\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j}\right)^2.$$

On calcule alors les π_{ij} ou on les détermine de manière approximative (car dans certains cas, le calcul rigoureux est impossible).

On peut ainsi calculer la précision (par la variance) de l'estimateur de Horvitz-Thompson (qui lui ne fait appel qu'aux π_i).

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. On suppose que l'on connaît les

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

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

Tirages systématiques : On définit

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

On génère ensuite une v.a.u. dans $[0, 1]$, u, qui donnera le "départ aléatoire"

La première unité sélectionnée i_1 est telle que

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

La k-ième unité sélectionnée i_k est telle que

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

La n-ième unité sélectionnée i_n est telle que

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

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

On prend

$$u = 0,36.$$

Par conséquent,

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

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.

Remarque

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 $\widehat{\mathrm{Var}}\left[\widehat{T}_{pi}\right]$.

Remarque

Pour de plus amples renseignements et pour obtenir en détails les formules, nous renvoyons le lecteur au livre "Méthodes statistiques des sondages", de Jean-Marie 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 Rémy Clairin et Philippe Brion.

- Supposons que l'on ait une liste de 207 villes avec une estimation de leur population.
- On veut enquêter 21 villes.Par conséquent $n = 21$.
- On calcule d'abord la population cumulée correspondant à chaque ville (cf le tableau du slide 48).
- Pour la dernière ville, elle vaut 58 626.

- On tire 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, on peut imaginer qu'on a 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 tirage à probabilités inégales : 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

- Supposons, par exemple, que l'on ait tiré entre autres : 937, 58 302. Ces deux nombres désignent respectivement les villes 3 et 206.
- Supposons que l'on tire ensuite 727, la ville 3 est à nouveau sélectionnée.
- Ceci induit les probabilités inégales P_i associées à chacune des villes.
- On peut améliorer la procédure en rangeant par taille les unités, et en procédant à un tirage systématique dans les chiffres cumulés.
- On obtient ainsi une répartition "satisfaisante" de l'échantillon par rapport au critère de tri choisi.

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.