Sondage par grappes

Myriam Maumy[1]

[1]IRMA, Université Louis Pasteur
Strasbourg, France

Master 1ère Année 24-04-2006

Ce chapitre s'appuie essentiellement sur :

le livre de Jean-Marie Grosbras,

« Méthodes statistiques des sondages »,

Economica.

Sommaire

Résumons :

Auparavant, on a vu que la méthode de base est le sondage aéatoire simple à probabilités égales, qui n'intègre aucune connaissance a priori des phénomènes étudiés.

L'apport de renseignements liés à ces phénomènes et fournis par des variables auxiliaires est susceptible d'améliorer la précision des résultats.

Méthodes

Les méthodes sont :

- la stratification a priori ou a posteriori (Chapitre 3 et Chapitre 5)
- les tirages à probabilités inégales (Chapitre 4)
- les redressements par le quotient (Chapitre 6).

Le choix dépend de la nature des informations auxiliaires et de leur degré de disponibilité.

Ces méthodes supposent que la population des unités concernées est représentée par **une base de sondage complète** et qu'on peut distinguer une unité des autres pour alimenter un échantillon.

Dans la pratique, il arrive qu'il n'y a **pas de base de sondage complète et disponible**.

Il arrive que **les unités soient groupées en "paquets"** et qu'il soit plus économique d'accéder à ces données par paquets, c'est-à-dire par grappes.

Sommaire

Les N unités de la population sont réparties en M sous-ensembles, appelés **grappes** ou **unités primaires**.

La grappe α ($\alpha = 1, \ldots, M$) contient N_α unités de la population, appelées **unités secondaires**.

On prend un échantillon de m grappes. La grappe i ($i = 1, \ldots, m$) de l'échantillon est explorée complétement (on examine tous les grains).

Sommaire

Il y a **deux niveaux d'observation** :

- le **niveau primaire**, c'est-à-dire la grappe, indicée par α ($\alpha = 1, \ldots, M$) dans la population, et par i ($i = 1, \ldots, m$) dans l'échantillon
- le **niveau secondaire**, c'est-à-dire l'unité statistique visée.

Notations

- $N_\alpha =$ taille (nombre d'unités secondaires) de la grappe α.
- $Y_{\alpha\beta} =$ valeur de la variable étudiée pour l'unité secondaire β de la grappe α.
- $Y_\alpha = \sum_{\beta=1}^{N} Y_{\alpha\beta} =$ total de la variable dans la grappe α.

Suite des notations

- $\overline{N} = \frac{1}{M} \sum_{\alpha=1}^{M} N_\alpha =$ taille moyenne des grappes.
- $\overline{Y} = \frac{1}{M} \sum_{\alpha=1}^{M} Y_\alpha =$ total moyen par grappe.
- $\overline{\overline{Y}}_\alpha = \frac{1}{N_\alpha} \sum_{\alpha=1}^{N_\alpha} Y_{\alpha\beta} = \frac{Y_\alpha}{N_\alpha} =$ moyenne de la variable à l'intérieur de la grappe α.

Remarques

La **simple barre** indique une moyenne par unité primaire.

La **double barre** indique une moyenne au niveau unité secondaire.

Suite des notations

Les relations avec les moyennes ou les totaux d'ensemble sont :

- $N = \sum_{\alpha=1}^{M} N_\alpha$
- $Y = \sum_{\alpha=1}^{M} Y_\alpha$
- $\overline{\overline{Y}} = \frac{1}{N} \sum_{\alpha=1}^{M} \sum_{\beta=1}^{N_\alpha} Y_{\alpha\beta} = \sum_{\alpha=1}^{M} \frac{N_\alpha}{N} \overline{Y}_\alpha = \frac{Y}{N}.$

Sommaire

Grappes de tailles égales

Soit N_0 la taille commune des grappes.

Par conséquent, on a pour tout $\alpha = 1, \ldots, M$,

$$N_\alpha = N_0 = \overline{N}$$

et

$$N = MN_0.$$

On obtient alors :

$$\begin{aligned}\overline{\overline{Y}} &= \sum_{\alpha=1}^{M} \frac{N_0}{N} \overline{\overline{Y}}_\alpha \\ &= \frac{1}{M} \sum_{\alpha=1}^{M} \overline{\overline{Y}}_\alpha .\end{aligned}$$

On peut donc dire que la moyenne générale $\overline{\overline{Y}}$ est la moyenne arithmétique des moyennes par grappe.

Problème de l'estimation de la moyenne :

L'estimation de la moyenne générale $\overline{\overline{Y}}$ est donc un problème d'estimation dans un sondage aléatoire simple dont

- la population de référence est constituée des $\overline{\overline{Y}}_\alpha$,
- les échantillons sont constitués de quantités calculées $\overline{\overline{Y}}_i$.

Solution :

Supposons que les m grappes soient tirées à probabilités égales sans remise, alors les résultats du Chapitre 2 montrent que l'estimateur par grappes :

$$\overline{\overline{y}}_G = \frac{1}{m}\sum_{i=1}^{m}\overline{\overline{Y}}_i$$

estime sans biais la moyenne générale $\overline{\overline{Y}}$.

Calcul des variances :

D'autre part, on a

$$\mathrm{Var}\left[\overline{\overline{y}}_G\right] = \frac{M-m}{Mm}\frac{1}{M-1}\sum_{\alpha=1}^{M}(\overline{\overline{Y}}_\alpha - \overline{\overline{Y}})^2$$

$$\widehat{\mathrm{Var}}\left[\overline{\overline{y}}_G\right] = \frac{M-m}{Mm}\frac{1}{m-1}\sum_{i=1}^{m}(\overline{\overline{Y}}_i - \overline{\overline{y}}_G)^2.$$

Comparaison avec le SAS

Supposons que l'on puisse réaliser un échantillon de même taille n, sans tenir compte du groupement par grappes. On a

$$n = \sum_{i=1}^{m} N_i = mN_0, \quad \text{puisque pour tout } i, \quad N_i = N_0.$$

La moyenne générale $\overline{\overline{Y}}$ est alors estimée par :

$$\overline{\overline{y}} = \frac{1}{n} \sum_{j=1}^{n} y_j.$$

Calcul de variance :

La variance de $\overline{\overline{y}}$ est égale à :

$$\mathrm{Var}\left[\overline{\overline{y}}\right] = \frac{N-n}{Nn} S_c^2,$$

où

$$S_c^2 = \frac{1}{N-1} \sum_{\alpha=1}^{M} \sum_{\beta=1}^{N_\alpha} \left(Y_{\alpha\beta} - \overline{\overline{Y}}\right)^2.$$

Comme $N = MN_0$ et $n = mN_0$, on a alors :

$$\mathrm{Var}\left[\overline{\overline{y}}\right] = \frac{1}{N_0} \frac{M-m}{Mm} S_c^2.$$

Comparons $\mathrm{Var}\left[\overline{\overline{y}}\right]$ à $\mathrm{Var}\left[\overline{\overline{y}}_G\right]$:

Pour cela, on rappelle (slide 12) que

$$\begin{aligned}\mathrm{Var}\left[\overline{\overline{y}}_G\right] &\simeq \frac{M-m}{Mm}\frac{1}{M}\sum_{\alpha=1}^{M}\left(\overline{\overline{Y}}_\alpha - \overline{\overline{Y}}\right)^2 \\ &\simeq \frac{M-m}{Mm}\eta^2 S_c^2,\end{aligned}$$

où η^2 est le **rapport de corrélation inter-grappes**.

On rappelle la définition du rapport de corrélation inter-grappes η^2 :

$$\eta^2 = \frac{\sum_{\alpha=1}^{M} N_0 \left(\overline{\overline{Y}}_\alpha - \overline{\overline{Y}}\right)^2}{\sum_{\alpha=1}^{M}\sum_{\beta}^{N_\alpha} \left(Y_{\alpha\beta} - \overline{\overline{Y}}\right)^2}.$$

Conclusion :

Par conséquent

$$\text{Var}\left[\overline{\overline{y}}_G\right] < \text{Var}\left[\overline{\overline{y}}\right] \Leftrightarrow \eta^2 < \frac{1}{N_0}.$$

Le **sondage par grappes** est meilleur que le sondage aléatoire simple si le rapport de corrélation inter-grappes est inférieur à $1/N_0$.

La conclusion est double :

- Il est souhaitable que les moyennes des grappes $\overline{\overline{Y}}_\alpha$ soient le plus semblables possible
- Il n'est pas souhaitable que la taille N_0 des grappes soit trop élevée.

Le **sondage par grappes** est intéressant s'il y a beaucoup de petites grappes, les plus ressemblantes possibles.

Sommaire

Tailles inégales des grappes

On rappelle que

$$\overline{\overline{Y}} = \sum_{\alpha=1}^{M} \frac{N_\alpha}{N} \overline{\overline{Y}}_\alpha.$$

Dans ce paragraphe, on va maintenant considérer que

les tailles N_α des grappes ne sont plus égales.

Comme

$$\overline{N} = \frac{N}{M},$$

on peut alors écrire

$$\overline{\overline{Y}} = \frac{1}{M} \sum_{\alpha=1}^{M} \frac{N_\alpha}{\overline{N}} \overline{\overline{Y}}_\alpha.$$

On est encore ramené au problème de l'estimation d'une moyenne simple, d'éléments de la forme $\frac{N_\alpha}{\overline{N}} \overline{\overline{Y}}_\alpha$.

Supposons que les m grappes soient tirées à probabilités égales sans remise, alors les résultats du Chapitre 2 montrent que l'estimateur par grappes :

$$\overline{\overline{y}}_G = \frac{1}{m} \sum_{i=1}^{m} \frac{N_i}{\overline{N}} \overline{\overline{Y}}_i$$

estime sans biais la moyenne générale $\overline{\overline{Y}}$.

Calcul des variances :

D'autre part, on a

$$\mathrm{Var}\left[\overline{\overline{y}}_G\right] = \frac{M-m}{Mm}\frac{1}{M-1}\sum_{\alpha=1}^{M}\left(\frac{N_\alpha}{\overline{\overline{N}}}\overline{\overline{Y}}_\alpha - \overline{\overline{Y}}\right)^2$$

$$\widehat{\mathrm{Var}}\left[\overline{\overline{y}}_G\right] = \frac{M-m}{Mm}\frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{N_i}{\overline{\overline{N}}}\overline{\overline{Y}}_i - \overline{\overline{y}}_G\right)^2$$

Remarques

Le nombre d'unités statistiques de l'échantillon est aléatoire, car il dépend des grappes choisies. En effet, on a

$$n = \sum_{i=1}^{m} N_i.$$

Si on calcule l'espérance mathématique de n, on trouve que

$$\mathbb{E}[n] = m\overline{N}.$$

Remarques :

On pourrait comme dans le paragraphe précédent, comparer le sondage par grappes de tailles inégales avec un sondage aléatoire simple de taille $m\overline{N}$.

La conclusion est la suivante : Il est préférable d'avoir beaucoup de grappes,

- dont la taille moyenne $\overline{N}$ soit faible
- et dont les moyennes ne soient pas trop dissemblables.

Sommaire

Tirages à probabilités inégales

Considérons des tirages de m grappes à probabilités inégales avec remise. À chaque tirage, la grappe α est retenue avec la probabilité P_α, avec :

$$\sum_{\alpha=1}^{M} P_\alpha = 1.$$

D'après le chapitre 3, on rappelle que l'estimateur du total T est égal à :

$$\widehat{T} = \frac{1}{m} \sum_{i=1}^{m} \frac{Y_i}{P_i}.$$

Calcul des variances :

La variance de l'estimateur du total $\widehat{T}$ est égale à :

$$\mathrm{Var}\left[\widehat{T}\right] = \frac{1}{m}\sum_{\alpha=1}^{M} P_\alpha \left(\frac{Y_\alpha}{P_\alpha} - Y\right)^2 .$$

La variance estimée de l'estimateur du total $\widehat{T}$ est égale à :

$$\widehat{\mathrm{Var}}\left[\widehat{T}\right] = \frac{1}{m(m-1)}\sum_{i=1}^{m} \left(\frac{Y_i}{P_i} - \widehat{Y}\right)^2 .$$

Remarque :

L'estimateur d'une moyenne $\widehat{\mu}$ s'en déduit en divisant l'estimateur du total $\widehat{T}$ par N, et les variances par N^2.

Par conséquent, l'estimateur de la moyenne $\widehat{\mu}$ est égal à :

$$\widehat{\mu} = \frac{1}{N}\frac{1}{m}\sum_{i=1}^{m}\frac{Y_i}{P_i}.$$

Calcul des variances :

La variance de l'estimateur $\widehat{\mu}$ est égale à :

$$\mathrm{Var}\left[\widehat{\mu}\right] = \frac{1}{N^2}\frac{1}{m}\sum_{\alpha=1}^{M} P_\alpha \left(\frac{Y_\alpha}{P_\alpha} - Y\right)^2 .$$

La variance estimée de l'estimateur $\widehat{\mu}$ est égale à :

$$\widehat{\mathrm{Var}}\left[\widehat{\mu}\right] = \frac{1}{N^2}\frac{1}{m(m-1)}\sum_{i=1}^{m} \left(\frac{Y_i}{P_i} - \widehat{Y}\right)^2 .$$

Probabilités proportionnelles aux tailles

Si les totaux des grappes sont corrélés avec le nombre d'unités qu'elles contiennent, il est naturel de choisir les probabilités P_α proportionnelles aux N_α.
D'où :

$$\forall \alpha = 1, \ldots, M \quad P_\alpha = \frac{N_\alpha}{N}.$$

Définition de l'estimateur du total

En portant ces probabilités dans les formules du paragraphe 6.1., il vient :

$$\widehat{T} = \frac{N}{m} \sum_{i=1}^{m} \frac{Y_i}{N_i} = \frac{N}{m} \sum_{i=1}^{m} \overline{\overline{Y}}_i$$

Calcul de variance :

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

$$\mathrm{Var}\left[\widehat{T}\right] = \frac{N^2}{m}\sum_{\alpha=1}^{M}\frac{N_\alpha}{N}\left(\overline{\overline{Y}}_\alpha - \overline{\overline{Y}}\right)^2.$$

Calcul de variance :

La variance estimée de $\widehat{T}$ est égale à

$$\widehat{\mathrm{Var}}\left[\widehat{T}\right] = \frac{N^2}{m(m-1)} \sum_{i=1}^{m} \left(\overline{\overline{Y}}_i - \overline{\overline{y}}_G\right)^2,$$

où

$$\overline{\overline{y}}_G = \frac{1}{m} \sum_{i=1}^{m} \overline{\overline{y}}_i.$$

Calcul des variances :

Les variances théoriques et estimées de $\overline{\overline{y}}_G$ se déduisent de $\mathrm{Var}\left[\widehat{T}\right]$ et de $\widehat{\mathrm{Var}}\left[\widehat{T}\right]$ en les divisant par N^2 :

$$\mathrm{Var}\left[\overline{\overline{y}}_G\right] = \frac{1}{m}\sum_{\alpha=1}^{M}\frac{N_\alpha}{N}\left(\overline{\overline{Y_\alpha}} - \overline{\overline{Y}}\right)^2$$

$$\widehat{\mathrm{Var}}\left[\overline{\overline{y}}_G\right] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\overline{\overline{Y_i}} - \overline{\overline{y}}_G\right)^2.$$

Remarques :

1. L'examen des variances montre que l'estimation d'un total ou d'une moyenne est d'autant plus précise que les grappes sont de tailles faibles et de moyennes semblables.

2. Le nombre d'unités secondaires dans l'échantillon est aléatoire, puisque dependant des grappes retenues.

Sommaire

Taille des grappes inconnues a priori

Il est clair qu'on ne pourra connaître que les tailles des grappes retenues dans l'échantillon.

A défaut de renseignements complémentaires, les tirages de grappes se font à probabilités égales sans remise.

Taille globale connue

Il se peut que N soit connu, sans qu'on sache la répartition des unités dans les grappes.
Dans ce cas, la méthode est simple. Il suffit de se reporter à l'estimation d'un total dans un sondage à probabilités égales sans remise (cf Chapitre 2).

Rappel : L'estimateur du total est égal à

$$\widehat{T} = \frac{M}{m} \sum_{i=1}^{m} Y_i = M\widehat{\overline{Y}}.$$

Calcul de variance :

La variance de $\widehat{T}$ est égale respectivement à

$$\mathrm{Var}\left[\widehat{T}\right] = M^2 \frac{M-m}{Mm} S_c^2$$

où

$$S_c^2 = \frac{1}{M-1} \sum_{\alpha=1}^{M} \left(Y_\alpha - \overline{Y}\right)^2 .$$

Calcul de variance :

La variance estimée de $\widehat{T}$ est égale respectivement à

$$\widehat{\mathrm{Var}}\left[\widehat{T}\right] = M^2 \frac{M-m}{Mm} s_c^2$$

où

$$s_c^2 = \frac{1}{m-1} \sum_{\alpha=1}^{m} \left(Y_i - \widehat{\overline{Y}}\right)^2.$$

D'où

$$\overline{\overline{y}}_G = \widehat{T}/N = \widehat{T}/\overline{N}.$$

Taille globale inconnue

Le paragraphe précédent montre comment estimer un total. Le problème dans l'estimation d'une moyenne est qu'il faut estimer la taille N de la population comme on vient de le voir avec la dernière formule du transparent précédent.

$$\overline{\overline{Y}} = \frac{T}{N} \quad \Rightarrow \quad \overline{\overline{y}}_G = \frac{\widehat{T}}{\widehat{N}}.$$

D'une façon équivalente, si on considère le total moyen et l'effectif moyen par grappe on a :

$$\overline{\overline{Y}} = \overline{Y}/\overline{N} \quad \Rightarrow \quad \overline{\overline{y}}_G = \widehat{\overline{Y}}/\widehat{\overline{N}},$$

où

$$\widehat{\overline{Y}} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

et

$$\widehat{\overline{N}} = \frac{1}{m}\sum_{i=1}^{m} N_i.$$

On se retrouve donc dans la situation de l'estimation d'un ratio, telle que nous l'avons étudiée au paragraphe 8 du Chapitre 6.

Sommaire

Pour tous les modèles étudiés dans ce chapitre les conclusions sont convergentes : il faut des grappes de taille réduite, et de moyennes semblables.

Il est clair que cette condition est très forte et n'est en général pas respectée dans la pratique.

Par contre, elle peut-être approximativement respectée pour des sous-ensembles de grappes.

Il est donc **essentiel** de réfléchir avant tout à une **bonne stratification de grappes**, c'est-à-dire telle que dans chaque strate les grappes soient le plus homogènes possible.