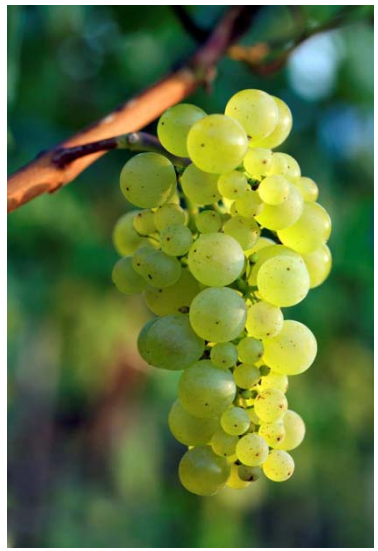
le cnam

STA108 Enquêtes et sondages

Sondages à plusieurs degrés et par grappes



Philippe Périé, novembre 2011

Sondages à plusieurs degrés et par grappes

- Introduction
- Sondages à plusieurs degrés
  - Tirage des unités primaires à probabilités égales sans remise (PESR)
  - Tirage des unités primaires à probabilités inégales avec remise (PIAR)
- Sondage par grappes
  - Taille des grappes connues a priori
  - Taille des grappes non connues a priori
- Grappes et stratification - mise en œuvre efficace

Sondages à plusieurs degrés et par grappes

INTRODUCTION

Sondages à plusieurs degrés

Les sondages à plusieurs degrés utilisent une succession de regroupements des unités statistiques pour tirer l'échantillon.

Par exemple :

- Tirer un échantillon de villes,
- Tirer un échantillon d'ilots
- Tirer un échantillon de ménages (logements) dans ces ilots

On a ici un exemple de sondage à 3 degrés, mais on peut généraliser à 2,3,4 degrés …

Pour chaque degré, les méthodes déjà présentées (probabilités égales ou inégales, stratification, …) peuvent s'appliquer.

Dans quels cas ?

Les plans de sondage en plusieurs degrés visent le plus souvent à améliorer l'organisation de l'enquête ou sa réalisation économique

Dans la pratique, il arrive de ne pas avoir à disposition une base de sondage complète et disponible au moment où l'on planifie l'étude. Dans la plupart des cas, on peut n'avoir simplement qu'un degré (les villes, les quartiers, …).

Ainsi, souvent ce type de sondage se rencontre quand les degrés constituent des unités géographiques et dans les enquêtes dont le mode de collecte est le face à face, car il y a un intérêt économique à limiter les déplacements autour de 'points de chute' définis.

On réalise dans ce dernier cas des économies de temps et de frais de déplacement. C'est moins vrai au téléphone, en online ou en postal : la dispersion des unités ne crée pas vraiment de coût

Exemple

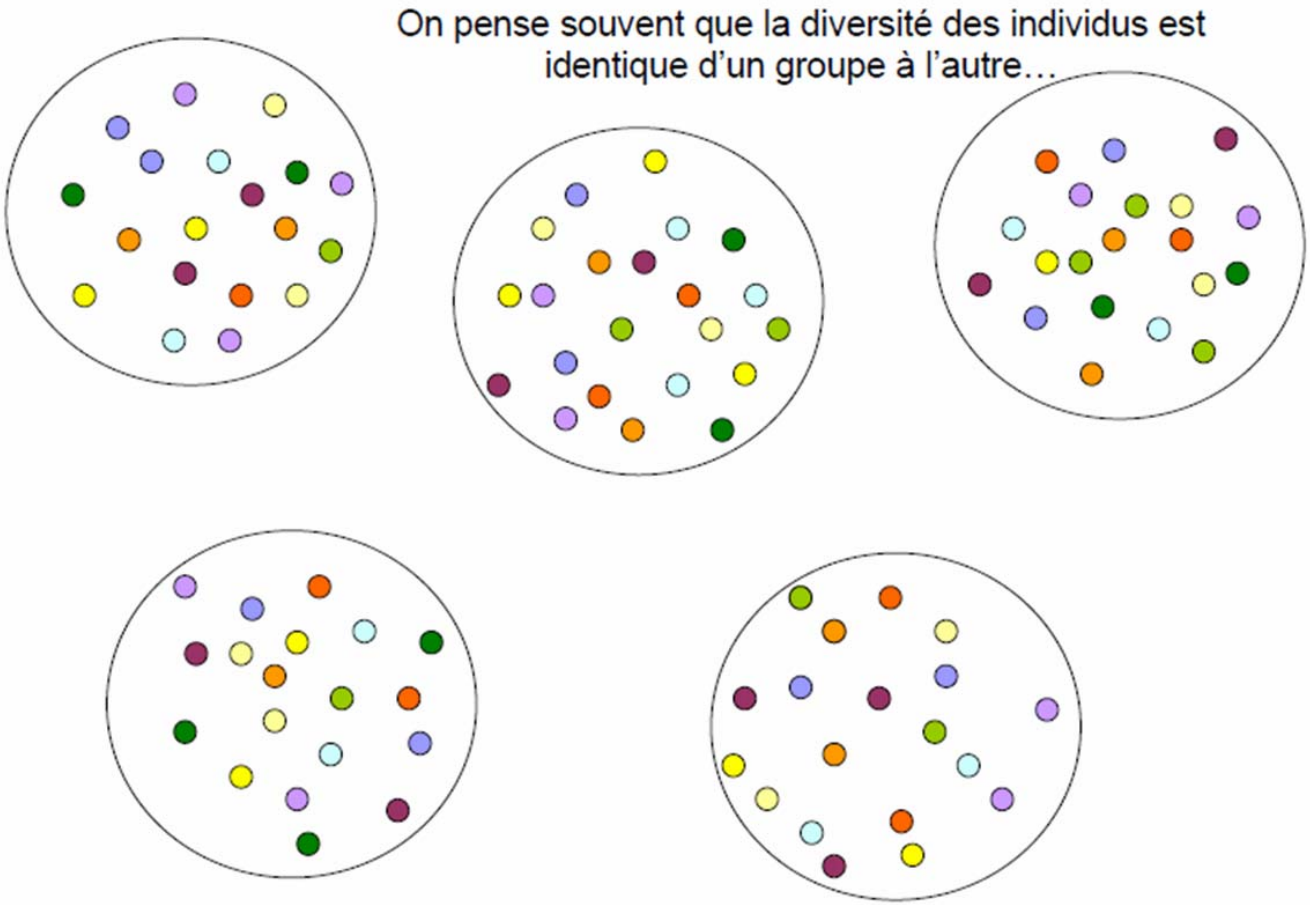
On veut interroger 5000 ménages en France métropolitaine qui en comporte 27 millions répartis sur plus de 36000 communes.

La liste des communes et des ilots, avec leurs caractéristiques est disponible à partir des enquêtes de recensement. Par contre il serait prohibitif en termes de temps et de coûts de vouloir constituer une liste exhaustive de ménages avant de lancer l'enquête, et d'y envoyer les enquêteurs au hasard.

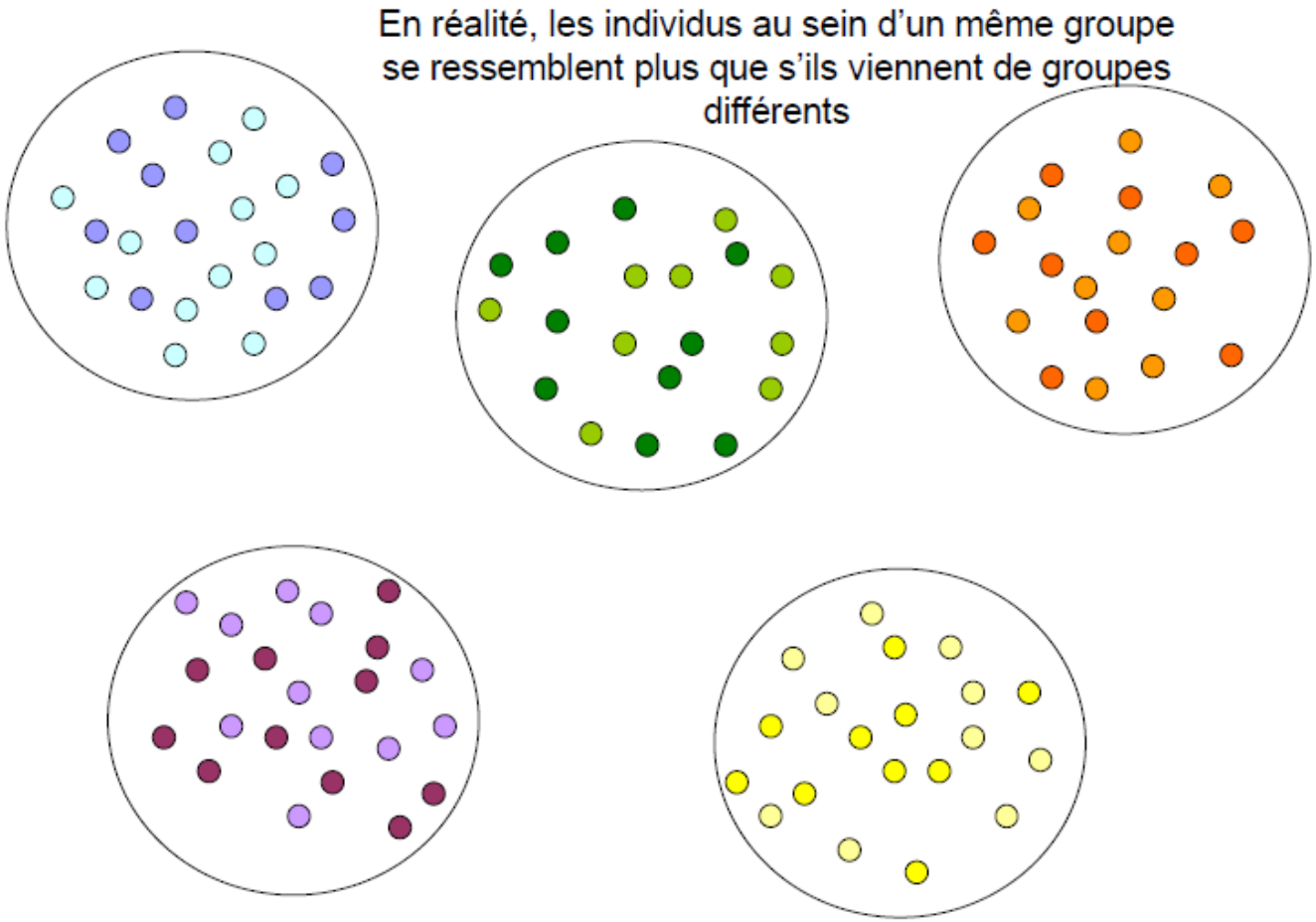
Un sondage à plusieurs degrés permet de réaliser cette enquête avec une base de sondage exhaustive seulement au premier degré (ville) et au deuxième degré (ilot).

Une fois tiré les ilots, on envoie l'enquêteur interroger tout ou partie des logements. Si on interroge tous les ménages de chaque ilot tiré et non pas une sélection, on parle de grappes de ménages.

Les limites …


On pense souvent que la diversité des individus est identique d'un groupe à l'autre…

Les limites …


En réalité, les individus au sein d'un même groupe se ressemblent plus que s'ils viennent de groupes différents

Limites …

Un sondage à plusieurs degrés sera par contre moins précis qu'un sondage aléatoire simple par exemple pour une même taille d'échantillon : on parle d'effet de grappe (cluster effect)

Une idée intuitive est que l'on ***'disperse moins'*** l'échantillon : les unités regroupées dans un même groupe (une grappe) ont une certaine tendance à se ressembler (penser aux habitant d'un immeuble par exemple). Il y a donc une certaine redondance d'information : chaque unité supplémentaire d'une grappe apporte 'moins' qu'une unité tirée au hasard dans l''ensemble de la population

La plus grande partie de la variance dans le cas des tirages à plusieurs degrés vient souvent des premiers degrés. A la limite, si toutes les unités se ressemblaient parfaitement dans une grappe, alors c'est comme si l'on avait interrogé un échantillon non pas de d'individus mais de grappes.

Sondages en grappes vs sondages à plusieurs degrés

Les N unités de la population sont réparties en M sous ensembles, appelées unités primaires (ou grappes)

La grappe $\alpha(\alpha = 1, \dots M)$ contient $N_\alpha$unités de la population appelées unités secondaires

Lors d'un sondage en grappes, on prend un échantillon de m grappes, la grappe $i(i = 1, \dots m)$ de l'échantillon est complètement enquêtée

Sondages en grappes vs sondages à plusieurs degrés

Le sondage par grappes est un cas particulier de sondage à plusieurs degrés dans lequel l'ensemble des unités du dernier degré est enquêtée

Exemples

Etudes médicales ‘cas patients’ : un échantillon de médecins qui donnent tout ou partie de leur patientèle, un effet de grappe médecin.

Etudes pour suivre certaines épidémies : grappes de laboratoires

INSEE : enquête emploi en continu
http://www.insee.fr/fr/methodes/default.asp?page=sources/ope-enq-emploi-continu.htm

« L'échantillon est aréolaire. Les grappes ont été constituées à partir des informations collectées à l'occasion de la campagne 2006 de la taxe d'habitation. Chaque année, cette base de tirage est complétée par les logements nouveaux repérés dans les fichiers de la taxe d'habitation. La taille moyenne des grappes est de 20 logements. Au moment du tirage, on a utilisé une stratification par région et degré d'urbanisation. Chaque trimestre, environ 67 000 logements sont identifiés comme résidences principales et enquêtés. Ils sont renouvelés par sixième chaque trimestre.
Au final, les fichiers d'enquête comptent environ 108 000 personnes de 15 ans ou plus répondantes chaque trimestre, réparties dans 57 000 ménages. »

Sondages à plusieurs degrés et par grappes

SONDAGE À PLUSIEURS DEGRÉS

Notations

Pour simplifier les notations, nous développons le sondage à 2 degrés, mais toutes les notions sont généralisables à 3,4 .. Degrés

Unités primaires:

M unités primaires dans la population $\alpha = (1, \ldots . M)$

m unités tirées dans l'échantillon $i = (1, \ldots m)$

Unités secondaires :

N_α dans l'unité primaire α $(unités\ secondaires\ \beta = 1, \ldots N_\alpha)$

n_α dans l'échantillon pour chaque unité primaire α $(j = 1, \ldots n_i)$

Dans chaque unité primaire α, **le total** $\boldsymbol{T_\alpha}$ par unité secondaire est :

$$T_\alpha = \sum_{\beta=1}^{N_\alpha} Y_{\alpha\beta}$$

où $Y_{\alpha\beta}$ est la valeur de la variable Y pour l'unité secondaire β de l'unité primaire α

Notations

Le total sur l'ensemble de la population est donné par :

$$\sum_{\alpha=1}^{M} \sum_{\beta=1}^{N_\alpha} Y_{\alpha,\beta} = \sum_{\alpha=1}^{M} T_\alpha$$

Dans chaque unité primaire α, **la moyenne μ_α par unité secondaire** est :

$$\mu_\alpha = \frac{1}{N_\alpha} \sum_{\beta=1}^{N_\alpha} Y_{\alpha,\beta}$$

où $Y_{\alpha,\beta}$ est la valeur de la variable Y pour l'unité secondaire β de l'unité primaire α

La moyenne sur toutes les unités secondaires est :

$$\mu = \frac{T}{N} \sum_{\alpha=1}^{M} \frac{N_\alpha}{N} \mu_\alpha$$

Un estimateur de la variance des totaux est :

$$\sigma^2(T) = \frac{1}{M-1} \sum_{\alpha=1}^{M} (T_\alpha - \bar{T})^2$$

Avec $\bar{T} = \frac{1}{M} \sum_{\alpha=1}^{M} T_\alpha = \frac{T}{M}$ le total moyen sur les unités primaires

Sondages à plusieurs degrés et par grappes

TIRAGE À PROBABILITÉS ÉGALES SANS REMISE (PESR) AUX DEUX DEGRÉS

Tirage à probabilités égales

Un plan assez classique consiste à sélectionner les unités primaires et les unités secondaires selon un plan aléatoire simple sans remise.

Les probabilités d'inclusion d'ordre 1 et 2 pour le premier tirage sont donc :

$$\pi_{1i} = \frac{m}{M} \text{ et } \pi_{1ij} = \frac{m}{M}\frac{(m-1)}{(M-1)}$$

Pour le second tirage, la taille des échantillons au sein des unités primaires est n_i, la probabilité d'inclusion pour l'ensemble du plan de sondage vaut donc :

$$\pi_k = \frac{mn_i}{MN_i}$$

➔ *Ce plan présente des inconvénients importants : la taille de l'échantillon est aléatoire -puisque cela dépend des unités tirées- et donc son coût*

Estimateurs du total et de la moyenne

Un estimateur du total T d'une population U est :

$$\hat{T} = \frac{M}{m}\sum_{i=1}^{m}\hat{T}_i = \frac{M}{m}\sum_{i=1}^{m}\left(\frac{N_i}{n_i}\sum_{j=1}^{n_i}Y_{ij}\right)$$

Où $\hat{T}_i$ est l'estimateur du total d'une unité primaire $i(i = 1, \dots m)$

C'est un estimateur sans biais formé simplement par les totaux aux deux degrés de tirage

Un estimateur de la moyenne par unité secondaire :

$$\hat{\mu} = \frac{1}{N}\hat{T}$$

$$\hat{\mu} = \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m}\hat{T}_i = \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m}\left(\frac{N_i}{n_i}\sum_{j=1}^{n_i}Y_{ij}\right)$$

Cet estimateur est aussi sans biais, mais suppose que N est connu.

Plus tard nous proposons une solution dans le cas ou N n'est pas connu

Estimateurs du total et de la moyenne

La variance de l'estimateur du total T est donnée par :

$$Var(\hat{T}) = M^2 \left(1 - \frac{m}{M}\right) \frac{{\sigma^2}_1}{m} + \frac{M}{m} \sum_{\alpha=1}^{M} {N^2}_\alpha \left(1 - \frac{n_\alpha}{N_\alpha}\right) \frac{{\sigma^2}_2}{n_\alpha}$$

Avec ${\sigma^2}_1 = \frac{1}{M-1} \sum_{\alpha=1}^{M} (T_\alpha - T)^2$

Et ${\sigma^2}_2 = \frac{1}{N_\alpha - 1} \sum_{\beta=1}^{N_\alpha} (Y_{\alpha,\beta} - \mu_\alpha)^2$

… **La variance de l'estimateur de la moyenne** est donnée par :

$$Var(\hat{\mu}) = M^2 \frac{1}{N^2} \left(1 - \frac{m}{M}\right) \frac{{\sigma^2}_1}{m} + \frac{M}{m} \frac{1}{N^2} \sum_{\alpha=1}^{M} {N^2}_\alpha \left(1 - \frac{n_\alpha}{N_\alpha}\right) \frac{{\sigma^2}_2}{n_\alpha}$$

Dans la formule, les deux termes correspondent aux deux degrés de tirage, et permettent donc de décomposer la variance à chacune des deux étapes. Si on augmente le nombre d'unités primaires m, les deux termes diminuent : on a intérêt au nom de la dispersion de les maximiser. Si on augmente le nombre d'unités secondaires, seul le deuxième terme diminue

Estimateurs du total et de la moyenne

Un estimateur sans biais de la variance de l'estimateur du total est :

$$\widehat{Var}(\hat{T}) = M^2\left(1-\frac{m}{M}\right)\frac{\widehat{\sigma_1}^2}{m} + \frac{M}{m}\sum_{i=1}^{m} N^2{}_i\left(1-\frac{n_i}{N_i}\right)\frac{\widehat{\sigma_2}^2}{n_i}$$

Avec $\widehat{\sigma_1}^2=\frac{1}{M-1}\sum_{\alpha=1}^{m}\left(\widehat{T_i}-\frac{\hat{T}}{M}\right)^2$

Et $\sigma^2{}_2=\frac{1}{n_i-1}\sum_{\beta=1}^{n_i}(Y_{ij}-\widehat{\mu_i})^2$

Dans le cas ou N n'est pas connue

Nous avons vu comment estimer la moyenne à partir du total, mais parfois la taille de la population N est inconnue. Cela correspond à des cas pour lesquels nous avons la liste des unité primaires, mais pas celles des unités secondaires

Dans ce cas on l'estime par :

$$\widehat{N} = M\widehat{\bar{N}} = \frac{M}{m}\left(\sum_{i=1}^{m} N_i\right)$$

Où $\widehat{\bar{N}}$ est l'effectif moyen observé pour les unités primaires de l'échantillon

Un estimateur de la moyenne μ est alors : $\hat{\mu} = \frac{\hat{T}}{\widehat{N}}$

Sondages à plusieurs degrés et par grappes

TIRAGE DES UNITÉS PRIMAIRES À PROBABILITÉS INÉGALES AVEC REMISE (PIAR)

Avec ou sans remise ?

Comme on a pu la voir les formules sans remise sont assez lourdes, celles à probabilités inégales le sont encore plus

En général les tailles d'échantillon sont assez importantes pour que l'on puisse considérer les approximations faites en l'approchant par un tirage avec remise comme acceptables

Tirages des unités primaires avec remise

Si $\pi_{1\alpha}$ est la probabilité de tirage de l'unité primaire α, alors, **un estimateur du total est** : $\hat{T} = \frac{1}{m}\sum_{i=1}^{m}\frac{\hat{T}_i}{\pi_{1\alpha}}$ où est $\hat{T}_i$ l'estimateur du total pour l'unité primaire.

$\hat{T}$ est sans biais, $\hat{T}_i$ est lié à la méthode de sondage utilisée au second degré de tirage

La variance de l'estimateur du total T est donnée par :

$$Var(\hat{T}) = \frac{1}{m}\sum_{\alpha=1}^{M}\pi_{1\alpha}\left(\frac{T_\alpha}{\pi_{1\alpha}} - T\right)^2 + \frac{1}{m}\sum_{\alpha=1}^{M}\frac{Var_\alpha}{\pi_{1\alpha}}$$

Où Var_α est la variance de l'estimateur $\hat{T}_\alpha$ du total T_α dans l'unité primaire α, qui est liée au plan de sondage du deuxième degré

Interprétation

Un estimateur de la variance de l'estimateur du total T est donnée par :

$$\widehat{Var}(\hat{T}) = \frac{1}{m}\sum_{i=1}^{m} \pi_{1i}\left(\frac{\hat{T}_i}{\pi_{1i}} - \hat{T}\right)^2 + \frac{1}{m}\sum_{i=1}^{m}\frac{\widehat{Var_i}}{\pi_{1i}}$$

Où $\widehat{Var_i}$ est l'estimateur de la variance de l'estimateur $\hat{T}_i$ du total T_i dans l'unité primaire i, qui est liée au plan de sondage du deuxième degré

En augmentant le nombre d'unités primaires on diminue la variance des deux termes (deux termes en $\frac{1}{m}$). En travaillant le plan de sondage au deuxième degré, on diminue seulement le deuxième terme. Pour cela il faut que la méthode de tirage au premier degré soit avec remise et qu'il y ait indépendance entre les degrés de tirage. La méthode de tirage au second degré est quelconque, la condition nécessaire est qu'elle permettent d'obtenir des estimateurs sans biais des totaux des unités primaires

Remarque : le sondage à deux degrés autopondéré

Le sondage autopondéré résout le problème de taille aléatoire rencontré dans la méthode PESR. A la première étape, on sélectionne les unités primaires avec des probabilité d'inclusion proportionnelles à la taille de ces unités primaires

Les probabilités de sélection des unités primaires sont donc : $\pi_{1i} = \frac{N_i m}{N}$

A la deuxième étape, on sélectionne des unités secondaires selon un plan aléatoire simple sans remise avec une taille d'échantillon $n_i = n_0$ constante (quelle que soit la taille de l'unité primaire), on a donc pour chaque unité primaire : $\pi_{k|i} = \frac{n_0}{N_i}$

La probabilité d'inclusion d'ordre 1 vaut donc :$\pi_k = \pi_{1i}\,\pi_{k|i} = \frac{N_i m}{N}\frac{n_0}{N_i} = \frac{m n_0}{N}$

- **Les probabilité d'inclusion sont donc toutes constantes pour tous les individus de la population**
- **Le plan est de taille fixe, la taille de l'échantillon vaut toujours** $n = m n_0$

Tirage PESR autopondéré

La formule de l'estimateur du total devient dans ce cas

$$\hat{T}=\frac{1}{m}\sum_{i=1}^{m}\frac{N}{N_i}\left(\frac{N_i}{n_0}\sum_{j=0}^{n_0}y_{ij}\right)=\frac{N}{mn_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0}y_{ij}$$

Chaque unité a le même coefficient d'extrapolation : le sondage est bien 'autopondéré'

Sondages à plusieurs degrés et par grappes

SONDAGE PAR GRAPPES

Sondages en grappes vs sondages à plusieurs degrés

RAPPEL : Le sondage par grappes est un cas particulier de sondage à plusieurs degrés dans lequel l'ensemble des unités du dernier degré est enquêtée

➔ En conséquence l'estimation de la moyenne générale sera simplement un problème d'estimation à partir d'une population de grappes, les échantillons seront constitués des quantités calculées des moyennes dans les grappes

➔ Dans les formules de variance, il n'y aura plus d'aléa au deuxième niveau, puisque l'on tire tous les individus dans une grappe (on effectue un 'recensement' dans chaque grappe tirée)

Notations

Unités primaires:

M unités primaires dans la population $\alpha = (1, \ldots . M)$

m unités tirées dans l'échantillon $i = (1, \ldots m)$

Unités secondaires :

N_α taille de la grappe α $(unités\ secondaires\ \beta = 1, \ldots N_\alpha)$

$Y_{\alpha\beta}$ est la valeur de la variable étudiée pour l'unité secondaire β de la grappe α

Par grappe :

Dans chaque grappe α, le total T_α est : $T_\alpha = \sum_{\beta=1}^{N_\alpha} Y_{\alpha\beta}$

Taille moyenne des grappes : $\overline{N} = \frac{1}{M} \sum_{\alpha=1}^{M} N_\alpha$

Total moyen par grappe : $\overline{T} = \frac{1}{M} \sum_{\alpha=1}^{M} T_\alpha$

Moyenne a l'intérieur de chaque grappe α : $\overline{Y_\alpha} = \frac{1}{N_\alpha} \sum_{\alpha=1}^{N_\alpha} Y_{\alpha\beta} = \frac{T_\alpha}{N_\alpha}$

Notations

Pour l'ensemble de la population

La taille de la population : $N = \sum_{\alpha=1}^{M} N_\alpha$

Le total général : $T = \sum_{\alpha=1}^{M} Y_\alpha$

La moyenne générale : $\bar{Y} = \frac{1}{N} \sum_{\alpha=1}^{M} \sum_{\beta=1}^{N_\alpha} Y_{\alpha\beta} = \sum_{\alpha=1}^{M} \frac{N_\alpha}{N} \overline{Y_\alpha} = \frac{T}{N}$

Sondages à plusieurs degrés et par grappes

TAILLES DES GRAPPES CONNUES A PRIORI - TIRAGE DES GRAPPES PESR, GRAPPES DE TAILLES ÉGALES

Grappes de tailles égales, probabilités égales, estimation d'une moyenne

On réalise un sondage aléatoire simple sans remise dans une population de grappes, les échantillons seront constitués des quantités calculées dans les grappes, chaque grappe apporte le même nombre d'individus

La taille de l'échantillon est donc fixe : nombre de grappes x nombre d'individus tirés dans chaque grappe

L'estimateur de la moyenne découle de la définition d'un SAS : moyenne arithmétique des moyennes calculées dans les grappes, la variance découle des écarts entre la moyenne globale et les moyennes calculées dans les strates

Grappes de tailles égales, probabilités égales, estimation d'une moyenne

On note N_0 la taille des grappes ($N_\alpha = N_0 = \bar{N}\ \forall\alpha$) , on a donc $N = MN_0$

Dans ce cas simple, la moyenne générale est simplement la moyenne arithmétique des moyennes par grappe :

$$\overline{Yg} = \sum_{\alpha=1}^{M} \frac{N_0}{N} \overline{Y_\alpha} = \frac{1}{M} \sum_{\alpha=1}^{M} \overline{Y_\alpha}$$

Si les m grappes sont tirées à probabilités égales alors un estimateur sans biais de la moyenne générale est :

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

Grappes de tailles égales, probabilités égales, estimation d'une moyenne

La variance de l'estimateur :

$$Var(\widehat{\overline{Yg}}) = \frac{M-m}{Mm}\frac{1}{M-1}\sum_{\alpha=1}^{M}(\overline{Y_\alpha} - \bar{Y})^2$$

Son estimation :

$$\widehat{Var}\left(\hat{\bar{Y}}g\right) = \frac{M-m}{Mm}\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\bar{Y}}_i - \hat{\bar{Y}}\right)^2$$

➔ **A ce stade, une première conclusion est qu'un sondage par grappes sera d'autant plus précis qu'il y a beaucoup de grappes (m est grand) qui se ressemblent en moyenne.**

Grappes de tailles égales, probabilités égales, estimation d'une moyenne

Nous allons commencer à partir de ce cas simple à étudier les conditions qui vont rendre un sondage par grappe intéressant du point de vue de la précision

Nous allons faire la comparaison avec un plan de sondage de référence, le SAS. Pour cela il nous faudra établir une mesure du degré de similarité entre les grappes.

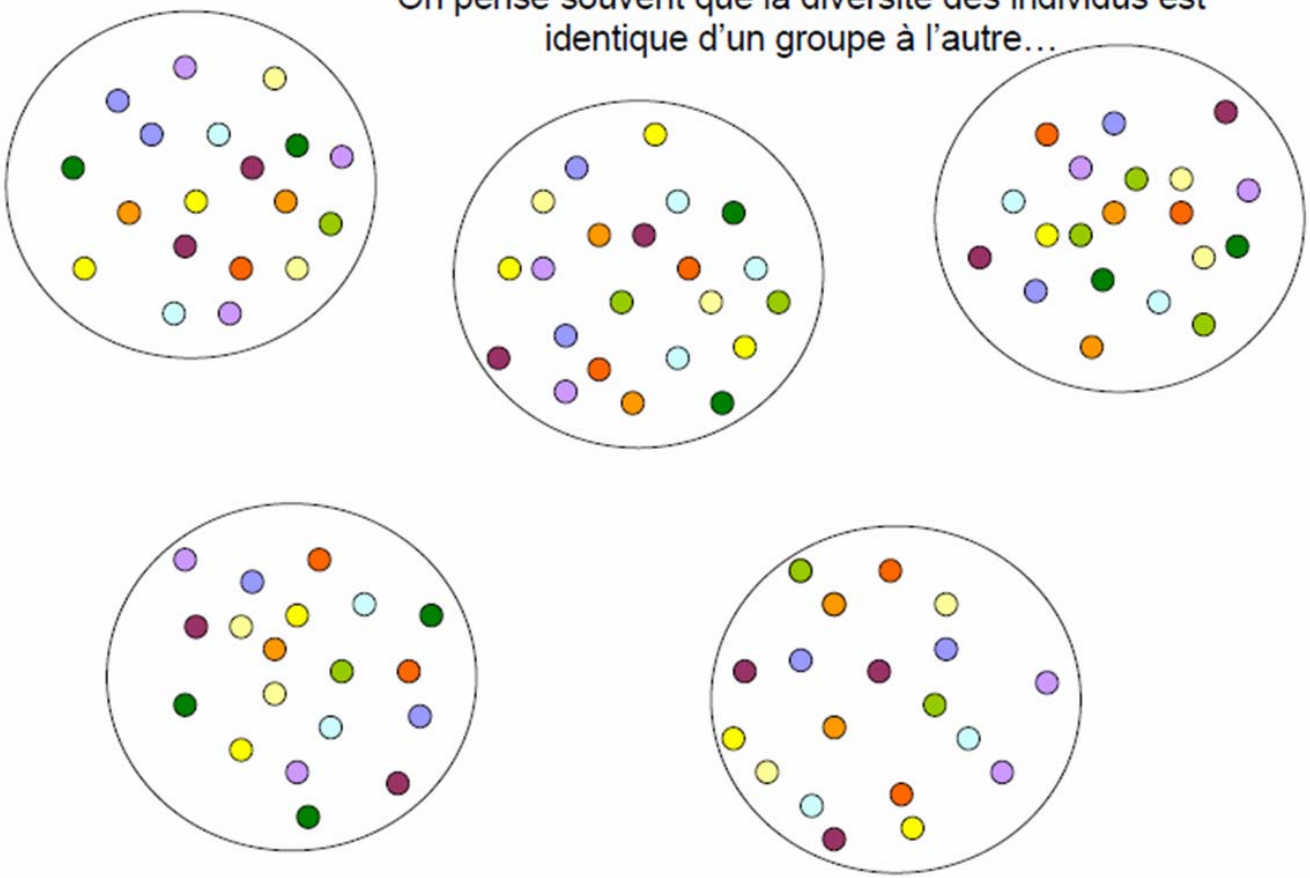
Notion de rapport de corrélation inter-grappes

La rapport de corrélation inter-grappes est le rapport de la variance inter grappes (entre les différentes grappes) sur la variance totale

$$\rho^2 = \frac{\sum_{\alpha=1}^{M} N_0 (\overline{Y_\alpha} - \bar{Y})^2}{\sum_{\alpha=1}^{M} \sum_{\beta=1}^{N_0} \left(Y_{\alpha\beta} - \bar{Y}\right)^2}$$

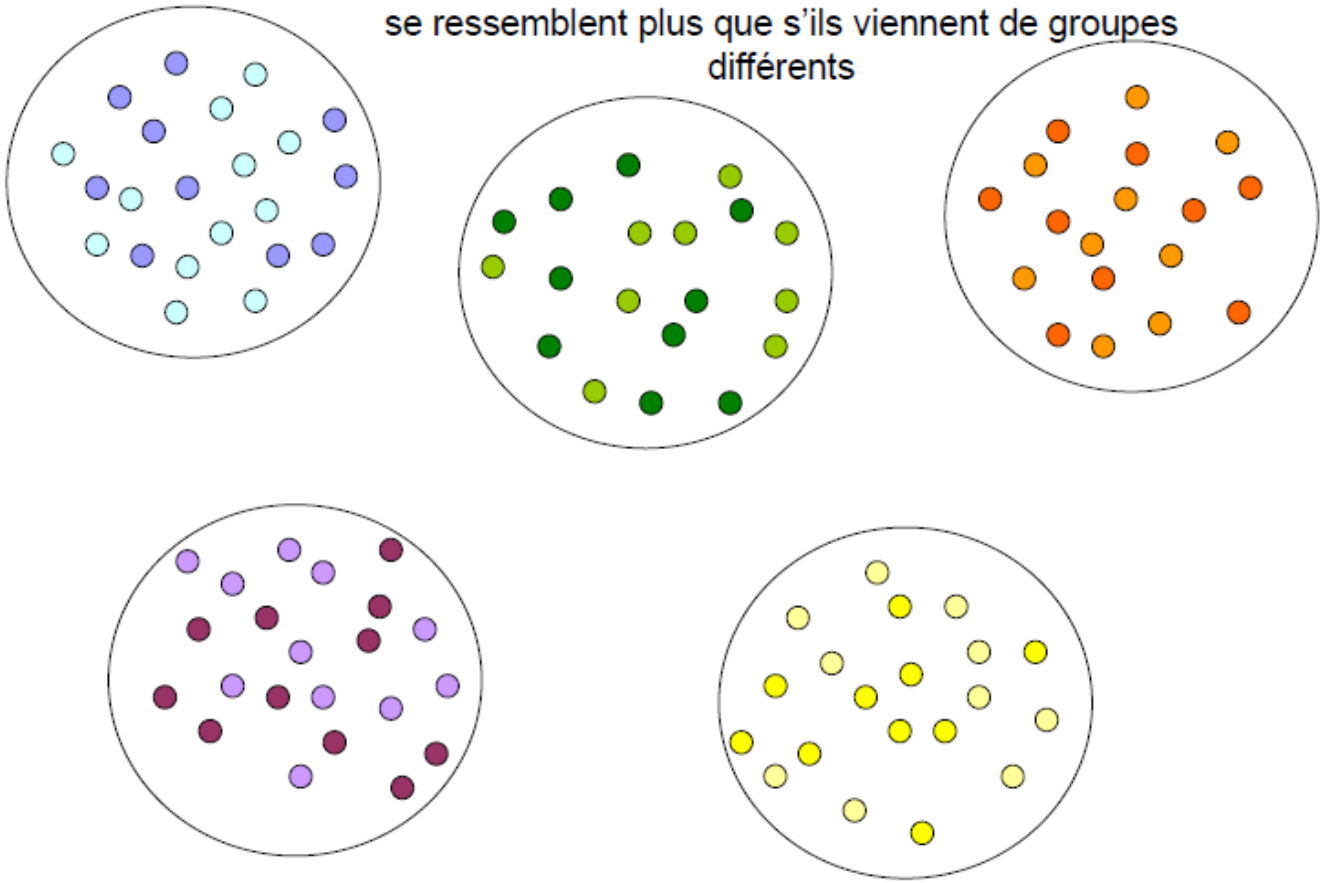
Cas 1 : ρ^2 sera faible

On pense souvent que la diversité des individus est identique d'un groupe à l'autre…



Cas 2 : ρ^2 sera fort

En réalité, les individus au sein d'un même groupe se ressemblent plus que s'ils viennent de groupes différents



Comparaison avec un SAS de même taille

Comparons avec la réalisation d'un échantillon de même taille, en ignorant les grappes. On a donc : $n = \sum_{i=1}^{m} N_i = mN_0$ puisque $(N_i = N_0 = \bar{N}\ \forall i)$

La moyenne générale est estimée par : $\hat{\bar{Y}} = \frac{1}{n}\sum_{j=1}^{n} y_j$

La variance de $\hat{\bar{Y}}$ est : $Var\left(\hat{\bar{Y}}\right) = \frac{N-n}{Nn} {S^2}_c$

Avec ${S^2}_c = \frac{1}{N-1}\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N_0}\left(Y_{\alpha\beta} - \hat{\bar{Y}}\right)^2$

Puisque $N = MN_0 et\ n = mN_0$, alors :

$$Var\left(\hat{\bar{Y}}\right) = \frac{1}{N_0}\frac{M-n}{Mn}{S^2}_c$$

Comparaison avec un SAS de même taille

Variance dans le cas des grappes :

$$Var\left(\hat{\bar{Y}}g\right) = \frac{M-m}{Mm}\frac{1}{M-1}\sum_{\alpha=1}^{M}(\overline{Y_\alpha} - \bar{Y})^2$$

Si N et M sont grands, N_0 petit (grande population constituée d'un grand nombre de grappes), alors l'approximation suivante est acceptable :

$$Var\left(\hat{\bar{Y}}g\right) \approx \frac{M-m}{Mm}\rho^2 S^2{}_c$$

Dans le cas d'un SAS :

$$Var\left(\hat{\bar{Y}}\right) = \frac{1}{N_0}\frac{M-n}{Mn}S^2{}_c$$

➔ **Le sondage par grappes est intéressant si le rapport de corrélation inter grappes est faible**, inférieur à $\frac{1}{N_0}$ à l'approximation près

Conclusions

Il est souhaitable que les moyennes des grappes soient les plus semblables possible. Il ne faut pas que la taille des grappes soit trop élevée

La sondage par grappes dans une population de grappes de tailles égales est d'autant plus efficace que la dispersion totale est essentiellement constituée par l'hétérogénéité des individus au sein des classes.

➔ Le sondage par grappes est efficace s'il y a beaucoup de petites grappes, les plus ressemblantes possibles

Sondages à plusieurs degrés et par grappes

TAILLES DES GRAPPES CONNUES A PRIORI - TIRAGE DES GRAPPES PESR, GRAPPES DE TAILLES INÉGALES

Grappes de tailles inégales, probabilités égales, estimation d'une moyenne

On réalise un sondage aléatoire simple sans remise dans une population de grappes, les échantillons seront constitués des quantités calculées des moyennes dans les grappes, chaque grappe apporte un nombre différent d'individus. La taille de l'échantillon n'est plus fixe : même si on décide à priori un nombre à tirer dans chaque strate de la population, elle dépendra des grappes choisies finalement

L'estimateur est maintenant la moyenne pondérée par les tailles relatives des grappes des moyennes calculées dans les grappes, la variance des écarts entre moyenne globale et moyennes calculées dans les strates

Grappes de tailles inégales, probabilités égales, estimation d'une moyenne

➔ **Les formules sont analogues à celles du cas simple introductif des tailles égales, on intègre le facteur des tailles relatives de grappes dans l''échantillon**

Dans ce cas , la moyenne générale est simplement la moyenne pondérée des moyennes par grappe :

$$\overline{Yg} = \frac{1}{M}\sum_{\alpha=1}^{M} \frac{N_\alpha}{\overline{N}} \overline{Y_\alpha} \text{ avec } N = \frac{N}{\overline{M}}$$

Si les m grappes sont tirées à probabilités égales alors un estimateur sans biais de la moyenne générale est :

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

Grappes de tailles inégales, probabilités égales, estimation d'une moyenne

La variances de l'estimateur :

$$Var(\widehat{\overline{Yg}}) = \frac{M-m}{Mm}\frac{1}{M-1}\sum_{\alpha=1}^{M}\left(\overline{Y_\alpha}\frac{N_\alpha}{\overline{N}} - \overline{Y}\right)^2$$

Son estimation :

$$\widehat{Var}\left(\hat{\overline{Y}}g\right) = \frac{M-m}{Mm}\frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{\overline{Y_i}}\frac{N_i}{\overline{N}} - \hat{\overline{Y}}\right)^2$$

➔ **A ce stade, une première conclusion est qu'un sondage par grappes sera d'autant plus précis qu'il y a beaucoup de grappes (m est grand) qui se ressemblent en moyenne.**

Comparaison avec un SAS

Le nombre d'unités statistiques de l'échantillon est aléatoire car il dépend des grappes choisies. On réalise donc la comparaison avec un SAS de taille $m\bar{N}$ l'espérance mathématique de la taille de l'échantillon

Les conclusions sont identiques à ce que l'on a vu précédemment : il est préférable d'avoir beaucoup de grappes, dont la taille moyenne est faible et dont les moyennes ne soient pas trop dissemblables

Sondages à plusieurs degrés et par grappes

TAILLES DES GRAPPES CONNUES A PRIORI - TIRAGE DES GRAPPES PIAR

Tirage des grappes à probabilités inégales avec remise, estimation d'une moyenne

On réalise un sondage à probabilités inégales avec remise dans une population de grappes, les échantillons seront constitués des quantités calculées des moyennes dans les grappes.

A chaque tirage, la grappe α est retenue avec la probabilité P_α

L'estimateur du total T est maintenant :

$$\hat{T} = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_i}{\pi_i}$$

L'estimateur de la moyenne est :

$$\hat{Y} = \frac{1}{\cdot\cdot}\frac{1}{m}\sum^{m}\frac{Y_i}{}$$

Tirage des grappes à probabilités inégales avec remise, estimation d'une moyenne

La variance de l'estimateur de la moyenne est :

$$Var\left(\hat{\bar{Y}}\right) = \frac{1}{N^2} Var\left(\hat{\bar{T}}\right) = \frac{1}{N^2}\frac{1}{M}\sum_{\alpha=1}^{M} \pi_\alpha \left(\frac{\widehat{\overline{T_\alpha}}}{\pi_\alpha} - \hat{T}\right)$$

Son estimation :

$$\widehat{Var}\left(\hat{\bar{Y}}\right) = \frac{1}{N^2} \widehat{Var}\left(\hat{\bar{T}}\right) = \frac{1}{N^2}\frac{1}{M}\sum_{i=1}^{m} \pi_i \left(\frac{\hat{\bar{T}}_i}{\pi_i} - \hat{T}\right)$$

Cas particulier : probabilités proportionnelles aux tailles.

Si les moyennes ou les totaux par strates sont corrélés avec le nombre d'unités qu'elles contiennent, on sera efficace de choisir les probabilités proportionnelles à ces tailles : $\pi_\alpha = \frac{N_\alpha}{N}$

Sondages à plusieurs degrés et par grappes

TAILLES DES GRAPPES INCONNUES A PRIORI - TIRAGE DES GRAPPES PESR, GRAPPES DE TAILLES INÉGALES

Tailles des grappes inconnues a priori, mais population totale connue

A défaut d'information complémentaire, les tirages se feront PESR. **Si la taille globale de la population est connue**, mais pas celles des grappes, alors on est dans le cas d'un sondage à probabilités inégales sans remise.

L'estimateur du total T est donc : $\hat{T} = \frac{M}{m}\sum_{i=1}^{m} T_i = M\hat{Y}$

Sa variance est :

$Var(\hat{T}) = M^2 \frac{M-m}{Mm} {S^2}_c$ avec ${S^2}_c = \frac{1}{M-1}\sum_{\alpha=1}^{M}(\overline{T_\alpha} - \bar{T})^2$

Un estimateur de sa variance est:

$\widehat{Var}(\hat{T}) = M^2 \frac{M-m}{Mm} \widehat{{S^2}_c}$ avec $\widehat{{S^2}_c} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{\overline{T_i}} - \hat{\bar{T}}\right)^2$

Pour la moyenne, on a : $\hat{\bar{Y}} = \frac{1}{N}\hat{\bar{T}}$ et $\widehat{Var}\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}\widehat{Var}\left(\hat{\bar{T}}\right)$

Tailles des grappes inconnues a priori, mais population totale inconnue

Si la taille globale N de la population est inconnue, il faut l'estimer, et on se retrouve dans le cas de l'estimation d'un ratio (quantités aléatoires au numérateur et au dénominateur)

Si on note $\widehat{N}$ l'estimation de N alors $\hat{\bar{Y}} = \frac{1}{\widehat{N}} \hat{\bar{T}}$

Sondages à plusieurs degrés et par grappes

GRAPPES ET STRATIFICATION, MISE EN ŒUVRE EFFICACE

Mise en œuvre pratique

Pour avoir un rapport de corrélation inter grappes les plus petit possible, nous avons vu qu'il faut un grand nombres de grappes dont les moyennes sont peut différentes les unes des autres, ce qui est n'est pas réalisé dans les conditions concrètes (on voudrait que chaque grappe constitue une 'mini population', on contredit la notion même de grappe …)

Par contre, cette condition peut être approchée si l'on constitue des sous ensemble de grappes : des strates

C'est ce que l'on fait en pratique pour conjuguer les effets bénéfiques de la stratification sur la précision et des grappes sur l'économie des moyens

Mise en œuvre pratique

Le lien avec le principe de la stratification est facile. : les strates doivent être les plus contrastées possible pour bien prendre en compte la variabilité du phénomène étudié. Mais à l'intérieur d'une strate, les grappes doivent se ressemble le plus possible

La répartition de l'échantillon dans les strates doit aussi intégrer la variabilité interne aux strates : si dans une strate, les grappes sont très ressemblantes, on pourra en sélectionner moins que dans les strates où les grappes sont plus différentes les unes des autres (application du principe de l'allocation optimale de Neyman)

Quelques cas

Etudes de satisfaction des passagers de compagnies aérienne : stratification selon le type de vol (les périodes, les horaires sont plus ou moins loisir vs business) et les faisceaux (Asie, Europe, …)

Une fois cette stratification opérée, les vols sont des grappes de passagers.

Etudes de marché : en général, stratification région x catégorie d'agglomération puis tirage des unités secondaires (iris/ilot, …) proportionnel à la taille. Les instituts privés font à la différence de l'INSEE (du fait de l'absence de base de sondage) la dernière étape par quotas : de 10 personnes par 'point de chute'

à partir d'une feuille de quotas.

Sondages à plusieurs degrés et par grappes