Sondage à plusieurs degrés

Myriam Maumy[1]

[1]IRMA, Université Louis Pasteur
Strasbourg, France

Master 2ème Année 27-11-2005

Ce chapitre s'appuie essentiellement sur 2 ouvrages :

- "Manuel de Sondages"

de Rémy Clairin et Philippe Brion, téléchargeable sur

http ://ceped.cirad.fr/activite/publi/integral/html/manuels/pdf/

puis "manuels cpd 03.pdf"

- "Méthodes statistiques des sondages"

de Jean-Marie Grosbras

On utilise une succession de regroupements des unités statistiques pour tirer l'échantillon.

Exemple :

- On tire un échantillon de villes.
- Puis on tire, parmi les villes tirées un échantillon de ménages.

On parle alors d'un tirage à 2 degrés.

Dans cet exemple, on a un tirage à 2 degrés. On peut généraliser à 3 degrés, à 4 degrés, . . .

À chacun des degrés, les méthodes présentées aux chapitres précédents peuvent être utilisées.

Exemple :

- On fait un tirage proportionnel à la taille au premier degré, donc à probabilités inégales (cf chapitre 4).
- Puis au deuxième degré, on réalise un sondage aléatoire simple (cf chapitre 2).

- Il faut dire un mot sur l'utilisation de **"sondage par grappes"**. Cette expression doit être réservée au cas particulier du **sondage à plusieurs degrés où l'ensemble des unités au dernier degré de tirage est enquêté**.
- **Dans l'exemple ci-dessus :** Ce serait l'ensemble des ménages des villes sélectionnées (soit une "grappe" de ménages) qui serait enquêté.
- C'est dans ce sens que sera utilisé **"sondage par grappes"** dans ce cours, bien que, dans certains manuels, cette expression soit utilisée pour parler d'un sondage à 2 degrés, de façon générale.

Exemple :

• On veut étudier 2 000 ménages dans un pays qui en compte 500 000 répartis dans 6 000 villes.

• On dispose d'une liste des villes avec une estimation de leur population.

• Élaborer une liste des ménages au niveau national en visitant chaque ville serait une tâche gigantesque.

• En outre, les ménages de l'échantillon seraient géographiquement extrêmement dispersés. D'où un temps énorme perdu en déplacements.

$\Rightarrow$ Le coût de l'enquête serait prohibitif.

Le **sondage à plusieurs degrés** permet de résoudre les 2 problèmes suivants.

- En l'absence d'une base de sondage, on se contente d'un travail partiel d'établissement de cette base. Seule la connaissance exhaustive des unités primaires est nécessaire. On se limite à recenser, dans **l'exemple précédent**, les ménages des villes tirés au premier degré.
- Globalement, on réalise
  - des économies de temps
  - de frais de déplacement (au niveau du travail des enquêteurs).

• Par contre le **sondage à plusieurs degrés** est, en général, **moins précis** que le sondage à un seul degré, pour une taille donnée de l'échantillon. Ceci est dû aux "effets de grappe".
• Les unités statistiques regroupées dans une même unité primaire (ou dans une même unité secondaire si on a 3 degrés de tirage) ont souvent tendance à se ressembler, à avoir des caractéristiques communes.

• Le fait de concentrer l'échantillon sur un échantillon d'unités primaires conduit à une certaine "redondance" de l'information sur ces unités et un certain "manque de représentativité" de l'ensemble.
• On montre que la majeure partie de la variance des estimateurs dans le cas des **tirages à plusieurs degrés** provient souvent du premier degré de tirage.

La pratique des **sondages par grappes** ou des **sondages à plusieurs degrés** est très largement répandue.
Elle est motivée par

- la nature des données à recueillir,
- des considérations de coût ou de faisabilité,
- la mauvaise qualité,
- l'inexistence des bases de sondage.

Voici 3 exemples mais il y en a plein d'autres...

- **Le premier exemple : Contrôle par lot.**

Il s'agit de contrôler des livraisons de produits fabriqués en grande série, ou de produits alimentaires.
Physiquement, les objets à contrôler sont conditionnés par caissses.
La nature même des données à recueillir impose que l'échantillon soit organisé par caisses, c'est-à-dire par grappes.

• **Le deuxième exemple : Études médicales.**
Certaines études sont réalisées auprès d'échantillons de médecins qui sont considérés, pour l'enquête, comme des grappes de patients ou de prescriptions.
Autre exemple : Des recherches effectuées pour analyser l'évolution du SIDA et, plus généralement, des MST, ont été basées sur des laboratoires d'analyses médicales, grappes d'actes et analyses.

- **Le troisième exemple : Sondages électoraux.**

On connaît les estimations établies par les instituts de sondage, les soirs de consultations électorales.
Il s'agit généralement de sondages "sortie des urnes" réalisés auprès d'électeurs à la sortie de bureaux de vote.
Il est clair qu'il s'agit de sondages à deux degrés, le premier degré consistant à choisir les bureaux de vote où opéreront les enquêteurs.

Dans ce chapitre, on se place essentiellement dans le cas du **sondage à** 2 **degrés**. On utilisera les notations suivantes :

- unités primaires :

$$\begin{array}{ll} M & \text{dans la population } (\alpha = 1, \ldots, M) \\ m & \text{tirées dans l'échantillon } (i = 1, \ldots, m). \end{array}$$

- unités secondaires :

$$\begin{array}{ll} N_\alpha & \text{dans l'unité primaire } \alpha \ (\beta = 1, \ldots, N_\alpha) \\ n_\alpha & \text{dans l'échantillon pour l'unité primaire } \alpha \ (j = 1, \ldots, n_\alpha). \end{array}$$

• Dans l'unité primaire α, le total T_α par unité secondaire est égal à

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

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

• Le total sur l'ensemble de la population est égal à

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

• Dans l'unité primaire α, la moyenne $\overline{\overline{Y}}_\alpha$ par unité secondaire est égale à

$$\overline{\overline{Y}}_\alpha = \frac{1}{N_\alpha} \sum_{\beta=1}^{N_\alpha} Y_{\alpha\beta},$$

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

• La moyenne par unités secondaires est égale à

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

- On note S_c^2 un indicateur de la variance des totaux

$$S_c^2 = \frac{1}{M-1} \sum_{\alpha=1}^{M} \left(T_\alpha - \overline{T}\right)^2,$$

où

$$\overline{T} = \frac{1}{M} \sum_{\alpha=1}^{M} T_\alpha = \frac{T}{M},$$

désigne le total moyen par unités primaires.

- On se gardera de confondre $\overline{T}$ et $\overline{\overline{Y}}$.

- Dans la suite de ce chapitre, on cherche à trouver des estimateurs de T, $\overline{T}$ ou encore de $\overline{\overline{Y}}$.
- On les notera $\widehat{T}$, $\widehat{\overline{T}}$ ou encore de $\widehat{\overline{\overline{Y}}}$.
- On suppose généralement que les N_α sont connus, ce qui est vrai ou à peu près vrai dans la plupart des situations pratiques.
- Les cas où les N_α sont inconnus posent problème pour l'estimation de la moyenne, qu'on traite alors comme l'estimation d'un ratio, selon les méthodes rencontrées au paragraphe 8 du Chapitre 6. On reviendra sur cette remarque par la suite.

On tire sans remise au premier degré, qui est *a priori* préférable pour la précision.

Remarque importante : Nous allons commencer par donner dans ce paragraphe l'estimateur d'un total puis l'estimateur d'une moyenne. Ceci n'est pas la logique habituelle de ce cours, mais inévitable dans le cas d'un **sondage à plusieurs degrés**. Vous comprendrez en lisant la suite du cours pourquoi nous sommes obligés d'adopter cette démarche.

$$\widehat{T} = \frac{M}{m} \sum_{i=1}^{m} \widehat{T}_i$$

estime T où $\widehat{T}_i$ est l'estimateur du total T_i à partir du plan de sondage choisi au second degré de tirage.
Cet estimateur est sans biais.

Remarque : On retrouve dans cette formule l'estimation du total aux 2 degrés de tirage.

Un estimateur de la moyenne par unité secondaire s'en déduit immédiatement en divisant par N, (**ce qui signifie que nous connaissons N**) l'estimateur d'un total :

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

Cet estimateur est sans biais.

Remarque : Le seul problème que pose cette formule est qu'il faut connaître N. On reviendra sur ce problème par la suite (cf section 7 de ce paragraphe).

$$\text{Var}\left[\widehat{T}\right] = M^2 \frac{M-m}{Mm} S_c^2 + \frac{M}{m} \sum_{\alpha=1}^{M} Z_\alpha$$

où

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

- Z_α est la variance de l'estimateur $\widehat{T}_\alpha$ du total T_α dans l'unité primaire α consécutive au plan de sondage choisi au deuxième degré.

On calcule la variance de l'estimateur d'une moyenne en divisant cette dernière égalité par N^2.

$$\text{Var}\left[\widehat{\overline{Y}}\right] = \frac{1}{N^2} M^2 \frac{M-m}{Mm} S_c^2 + \frac{1}{N^2} \frac{M}{m} \sum_{\alpha=1}^{M} Z_\alpha$$

où

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

- Z_α est la variance de l'estimateur $\widehat{T}_\alpha$ du total T_α dans l'unité primaire α consécutive au plan de sondage choisi au deuxième degré.

À partir de l'échantillon (d'unités primaires et d'unités secondaires), la variance de l'estimateur d'un total est estimée par :

$$\widehat{\mathrm{Var}}\left[\widehat{T}\right] = M^2 \frac{M-m}{Mm} s_c^2 + \frac{M}{m} \sum_{i=1}^{m} \widehat{Z}_i$$

où

- $s_c^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left(\widehat{T}_i - \frac{\widehat{T}}{m} \right)^2$

- $\widehat{Z}_i$ est l'estimateur de la variance de l'estimation $\widehat{T}_i$ selon le plan de sondage au deuxième degré.

• Ces formules sont relativement complexes et entraînent des calculs d'estimation assez lourds.
Aussi les statisticiens cherchent en jouant, par exemple sur les taux de sondage f_i au second degré, à obtenir des estimateurs plus simples.

• Dans la formule de $\mathrm{Var}\left[\widehat{T}\right]$, le premier terme est en général le plus important. Les 2 termes de cette formule sont relatifs aux 2 degrés de tirage et permettent de décomposer la variance pour observer la part de chacun de ces 2 degrés.

Suite des Remarques :

• Si on augmente m dans la formule de $\mathrm{Var}\left[\widehat{T}\right]$, on voit que les 2 termes diminuent.

Si on augmente les nombres n_α d'unités enquêtées au second degré, seul le deuxième terme diminue (par l'intermédiaire des Z_α).

On a donc intérêt à avoir plutôt un grand nombre d'unités primaires tirées.

• Dans la formule de $\widehat{\text{Var}}\left[\widehat{T}\right]$, on a également 2 termes qui semblent correspondre à la décomposition selon les 2 degrés de tirage.
En fait ce n'est pas le cas, contrairement à ce qui pu être dit précédemment pour $\text{Var}\left[\widehat{T}\right]$.

• Pour estimer la moyenne à partir du total, parfois on ne connaît pas le nombre total N d'unités statistiques. En fait, on n'a pas de base de sondage au niveau des unités secondaires mais plutôt seulement la liste des unités primaires.
• On estime N à partir de l'échantillon d'unités primaires :

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

où $\widehat{\overline{N}}$ est l'effectif moyen observé pour les unités primaires de l'échantillon. Puis, on estime la moyenne par

$$\widehat{\overline{Y}} = \frac{\widehat{T}}{\widehat{N}}.$$

• La taille d'échantillon est aléatoire. Elle dépend des unités tirées.

• La qualité de l'estimation dépend **avant tout** de la variance des T_i. Du point de vue de la théorie, il est important que les grappes soient le plus ressemblantes possible.
En effet, la qualité essentielle d'un échantillon est de représenter la dispersion de la variable d'intérêt dans la population. Or le **sondage à plusieurs degrés** concentre l'échantillon par "paquets" d'observations, ce qui est **a priori** nuisible à la dispersion et donc à la représentativité, **sauf si chaque unité est elle-même une image fidèle de l'ensemble**.

Pour reprendre les exemples du paragraphe 3, cela signifie que
- chaque caisse de fruits devrait être représentative de l'ensemble des caisses,
- chaque médecin devrait avoir une clientèle semblable aux autres,
- chaque bureau de vote devrait être à l'image de l'ensemble du corps électoral.

• Dans la pratique, cette dernière condition pose un problème. En effet le regroupement des individus dans les grappes correspond souvent à des caractérisitiques particulières : **"qui se ressemble s'assemble !"**

C'est ainsi que
• les médecins ont des clientèles typées selon leur lieu d'exercice,
• les bureaux de vote de la banlieue parisienne ne se ressemblent pas, qu'ils sont différents des bureaux en zones rurales,
• etc...

• Ce phénomène, appelé **effet de grappe**, fait qu'à nombre d'individus égal, un **sondage à plusieurs degrés** est moins précis qu'un sondage aléatoire simple.

• Il reste que la concentration des observations est un facteur de réduction des coûts.
Exemple : Le déplacement des enquêteurs

• Enfin, il n'y a souvent pas d'autre solution lorsque la base de sondage est défaillante.

C'est la situation la plus fréquente dans le domaine des études auprès des ménages ou d'individus :

- études de comportement,
- études d'opinion,
- études de marché,
- mesures d'audience,
- etc...

- Pour alléger les formules, on se place ici dans le cas du **tirage des unités primaires à probabilités inégales avec remise**.
- Les formules correspondant au cas du **tirage des unités primaires à probabilités inégales sans remise** sont assez lourdes.

- D'ailleurs dans la pratique, les tailles d'échantillon sont suffisamment élevées pour que les approximations par le formulaire des tirages avec remise soient convenables.
- P_α est la probabilité, pour qu'à chaque tirage, l'unité primaire α soit tirée. On a alors

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

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

estime T où $\widehat{T}_i$ est l'estimateur du total pour l'unité primaire i. $\widehat{T}_i$ tient compte de la méthode de sondage utilisée au second degré de tirage.
$\widehat{T}$ **est un estimateur sans biais du total sur la population.**

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

$$\text{Var}\left[\widehat{T}(Y)\right] = \frac{1}{m}\sum_{\alpha=1}^{M} P_\alpha \left(\frac{T_\alpha}{P_\alpha} - T\right)^2 + \frac{1}{m}\sum_{\alpha=1}^{M} \frac{Z_\alpha}{P_\alpha}$$

où

• Z_α est la variance de l'estimateur $\widehat{T}_\alpha$ du total T_α dans l'unité primaire, tenant compte du plan de sondage au deuxième degré.

Remarque : Comme au paragraphe précédent, on remarque que m est au dénominateur des 2 termes. Il est donc important d'avoir beaucoup d'unités primaires. Augmenter les taux de sondage au second degré ne peut améliorer que le second terme.

À partir de l'échantillon, l'estimateur de la variance de l'estimateur d'un total est :

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

Ce résultat est remarquable par sa simplicité. Les conditions d'obtention sont :

- tirage PIAR au premier degré
- indépendance au second degré des tirages de chaque unité primaire
- méthode de tirage du second degré quelconques, pourvu qu'elles fournissent des estimateurs sans biais $\widehat{T}_i$ de T_i.

• Le problème du choix des P_α n'a pas encore été abordé. Souvent on décide de tirer les unités avec une probabilité proportionnelle à leur taille :

$$P_\alpha = \frac{N_\alpha}{N}.$$

• Dans ce cas, il est intéressant de procéder, au deuxième degré, à un tirage aléatoire simple avec le même nombre n_0 d'unités secondaires dans chaque unité primaire tirée (quelle que soit sa taille).

- La formule d'estimation devient :

$$\begin{aligned}\widehat{T} &= \frac{1}{m}\sum_{i=1}^{m}\frac{N}{N_i}\left(\frac{N_i}{n_0}\sum_{j=1}^{n_0}y_{ij}\right)\\ &= \frac{N}{mn_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0}y_{ij}\end{aligned}$$

avec

$$\forall i \in \{1,...,m\}, \quad n_i = n_0.$$

Chaque unité enquêtée a le même coefficient d'extrapolation.
On a un **sondage dit "autopondéré"**.

En pratique, on se trouve rarement exactement dans cete situation.

• On tire proportionnellement à une taille qui est connue grâce à des données qui, même si elles sont récentes, ont pu évoluer. La taille de l'unité primaire effectivement constatée lors du dénombrement réalisé pendant l'enquête sera, en général, légèrement différente.

• Il faudra recalculer les pondérations exactes à l'aide de la page 34. Si le nombre d'unités contenues dans l'unité primaire i est, au moment de l'enquête, N_i', la pondération de l'unité j dans l'unité primaire i vaudra alors :

$$\frac{N}{mN_i}\frac{N_i'}{n_0}.$$

Pour estimer une moyenne par unité secondaire sur la population, il faudra souvent estimer le nombre total d'unités secondaires qui est inconnu.