Centre français sur la population
et le développement

Institut national de la statistique
et des études économiques

MANUEL DE

SONDAGES

Applications aux pays en développement

2e édition

Rémy Clairin
Philippe Brion

Paris, novembre 1997

Documents et Manuels du CEPED n° 3

MANUEL DE SONDAGES

Applications aux pays en développement

Déjà parus dans la collection "Documents et Manuels du CEPED" :

- n° 1 : *La démographie de 30 États d'Afrique et de l'Océan Indien*, par Valérie GUÉRIN (éd.) (1994), 352 p.
- n° 2 : *Clins d'œil de démographes à l'Afrique et à Michel François*, par Jacques VALLIN (éd.) (1995), 244 p.
- n° 3 : *Manuel de sondages. Applications aux pays en développement*, par Rémy CLAIRIN et Philippe BRION (1997), 108 p. (2e édition).
- n° 4 : *L'analyse des enquêtes biographiques à l'aide du logiciel STATA*, par Philippe BOCQUIER (1996), 208 p. + disquette.
- n° 5 : *Genre et développement : des pistes à suivre*, par Thérèse LOCOH, Annie LABOURIE-RACAPÉ et Christine TICHIT (éds) (1996), 154 p.
- n° 6 : *Les migrations internes au Kenya 1979-1989*, par Valérie GOLAZ (1997), 126 p.

Directeur de la publication : Alain LERY
Responsable scientifique : Christophe LEFRANC
Composition et mise en page : Valérie GUÉRIN-MARY et Sabine JOAO

Couverture : Poteau sculpté. Ngounié, Sud Gabon.
Détail relevé par Michel FRANÇOIS.

Le CEPED, *Centre français sur la population et le développement*, est un "Groupement d'intérêt scientifique" (GIS) créé en 1988 par l'INED, l'INSEE, l'ORSTOM, l'université Paris VI et l'École des hautes études en sciences sociales, pour conjuguer leurs efforts en matière de recherche, de formation et de coopération avec les pays du Sud dans le domaine de la population et de ses relations avec le développement. Ses activités de recherche portent essentiellement sur les facteurs de la dynamique des populations (santé, famille, fécondité, migrations), leurs relations avec les divers aspects du développement économique et social (éducation, emploi, activité économique, structures sociales...) ainsi que les méthodes d'observation et d'analyse appropriées. Ses travaux sont définis et conduits en étroite relation avec les organismes partenaires du tiers monde (offices statistiques, centres de recherche, universités). Le CEPED accueille régulièrement à Paris des chercheurs de ces pays, et met à la disposition du public un important centre de documentation sur les thèmes de sa compétence. Pour toutes ces tâches, le CEPED reçoit un large concours du Secrétariat d'État à la Coopération.

Le Département des relations internationales et de la coopération de l'INSEE a pour mission de coordonner les relations de l'INSEE avec les services homologues de tous les pays étrangers et avec les organisations internationales gouvernementales ou non gouvernementales, qu'il s'agisse d'actions de coopération, avec les pays en développement et les pays en transition, ou de relations techniques n'entrant pas dans le champ des actions de coopération.
Au sein de ce département, la Division des études et méthodes statistiques pour le développement est chargée de développer et promouvoir des outils méthodologiques adaptés au contexte des pays en développement et en transition ; elle édite une revue sur ce sujet, *Statéco*, et gère également un fonds documentaire sur la statistique du développement.

Documents et Manuels du CEPED n° 3

Rémy CLAIRIN
Philippe BRION

MANUEL DE SONDAGES

Applications aux pays en développement

2e édition

Centre français sur la population et le développement
(EHESS-INED-INSEE-ORSTOM-Université Paris VI)

Novembre 1997

Éléments de catalogage :

Manuel de sondages. Applications aux pays en développement, par Rémy CLAIRIN et Philippe BRION. – Paris, Centre français sur la population et le développement, 1997, 108 p ; 24 cm. (2e édition).

ISBN : 2-87762-109-X 2e édition revue et corrigée
ISBN : 2-87762-082-4 1re édition
ISSN : 1264-2487

Centre français sur la population et le développement
15, rue de l'École de médecine - 75270 PARIS Cedex 06 - FRANCE
Téléphone : 01 44 41 82 30 - Fax : 01 44 41 82 31
Téléphone international : 33 1 44 41 82 30 – Fax international : 33 1 44 41 82 31

MANUEL DE SONDAGES – 2E ÉDITION

Documents et Manuels du CEPED, n° 3

ERRATUM

Page 47 – Dernière formule, lire :

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{T}_i(Y) - \frac{\hat{T}(Y)}{M}\right)^2$$

Page 90 – Milieu de l'encadré, lire :

$$Z = \sum_{i=1}^{M} \frac{dF}{dX_i}\left(\hat{X}\right)\hat{X}_i$$

Page 90 – Avant-dernière ligne, lire :

"La variance de l'estimation du total de Z..."

SOMMAIRE

PRÉFACE

Rémy Clairin nous a quittés le 12 octobre 1987, trois mois avant que le CEPED, dont il faisait partie, ne voie officiellement le jour.

Administrateur de l'INSEE (1954), Rémy Clairin commence sa carrière en Guinée où il participe à la première grande enquête démographique lancée par son Institut en Afrique. Auteur ou collaborateur de nombreux travaux et ouvrages de référence, en particulier sur la collecte des données et les techniques d'ajustement, homme de terrain, analyste, mais aussi brillant théoricien, il a toujours eu le souci de former les jeunes statisticiens et démographes, directement, par ses manuels, ses publications et ses articles nombreux. Spécialiste reconnu en matière d'application de la théorie des sondages, il a laissé au CEPED le manuscrit inachevé d'un manuel de sondages à l'usage des jeunes cadres africains.

Mon prédécesseur, Francis Gendreau, et Michel François, statisticien-démographe de l'INSEE, ont avec patience et obstination fait tout ce qu'il fallait pour rendre un dernier hommage à Rémy Clairin en faisant aboutir ce projet de manuel.

Une première version du manuscrit a été mise en forme au CEPED par Valérie Delaunay et Elisabeth Omoluabi, démographes, pour être soumise aux observations de statisticiens et démographes ayant une certaine pratique des enquêtes démographiques par sondage en Afrique. Mais le temps a passé, et le manuscrit, inachevé et déjà ancien, demandait à être complété au fond.

C'est finalement Philippe Brion, chef de la division des "études et méthodes statistiques pour le développement" de l'INSEE, qui a accepté d'élaborer une version complète d'un manuel de sondages en y intégrant le travail de Rémy Clairin.

Le manuel initial a été repris, tout en conservant sa vocation première d'ouvrage à destination des praticiens. Point n'est besoin de rappeler l'importance de la méthode des sondages dans le domaine de la démographie appliquée aux pays en développement : de nombreuses enquêtes viennent compléter les données des recensements de la population pour contribuer à enrichir la connaissance de ces pays.

Quelques exemples tirés d'enquêtes autres que les enquêtes démographiques servent à illustrer le manuel : enquêtes agricoles, enquêtes budget-consommation.

Celui-ci reste cependant centré sur les sondages appliqués aux enquêtes auprès des ménages des pays en développement et on n'y trouvera pas, par exemple, d'application des sondages au domaine des entreprises.

L'ensemble de l'ouvrage a reçu les avis et critiques particulièrement utiles et constructifs de Christopher Scott et Louis Lohlé-Tart.

Enfin MM. Ardilly, Blaizeau, Brilleau, Deville, Gendreau, Lefranc et Waltisperger ont également apporté, par leurs remarques sur différents points, leur contribution à la version finale de l'ouvrage.

Je tiens, au nom du CEPED et de tous ceux qui en son temps ont apprécié le talent et l'humanisme de Rémy Clairin, à remercier ici chacun des multiples auteurs de cet hommage à sa mémoire et à son œuvre, et à les féliciter, Philippe Brion en premier bien sûr, pour la qualité du résultat.

Jacques VALLIN
Directeur de recherche à l'INED
Ancien directeur du CEPED

La première édition du "Manuel de sondages. Applications aux pays en développement" *de Rémy Clairin et Philippe Brion ayant été rapidement épuisée, il est apparu nécessaire de le rééditer. Cette deuxième édition incorpore, de plus, quelques compléments que Philippe Brion a souhaité apporter au texte initial.*

Alain LERY
Directeur du CEPED

RÉSUMÉ

La méthode des sondages est utilisée dans les pays en développement pour produire de l'information sur différents domaines : enquêtes démographiques, enquêtes socio-économiques comme celles sur le budget et la consommation des ménages, ou encore enquêtes agricoles. Son principe est de remplacer le tout par une partie, l'échantillon, dont l'observation sert de base à l'extrapolation à l'ensemble. Lors de l'application de cette méthode, un certain nombre de contraintes techniques et organisationnelles sont à prendre en compte, et interfèrent dans les choix théoriques.

L'objet de ce manuel, qui résulte d'un projet initié par Rémy Clairin et, à l'origine, consacré aux enquêtes démographiques dans les pays africains, est de présenter les bases théoriques de manière simple, et de faire le lien avec la pratique, en utilisant en particulier certains exemples s'appuyant sur des enquêtes réalisées par sondage dans divers pays en développement.

SUMMARY

Sampling methods are used in developing countries to produce information for different fields : demographic surveys, socioeconomic surveys - for example studying household budget, or consumption -, or agricultural surveys. The principle is to replace the whole by a part, the sample, which will be used to extrapolate to the whole. While using the method, some technical or organisational constraints must be considered, jointly with the theoretical aspects.

The object of this manual, which is resulting from a project initiated by Rémy Clairin and, originally, dedicated to demographic surveys in african countries, is to present the theoretical principles in a simple manner, and to link them to the practical problems, particularly using examples based on sample surveys conducted in developing countries.

INTRODUCTION

Dans l'arrière-boutique, le négociant plonge la main dans le sac de café, afin d'évaluer la qualité de la marchandise qu'il va acheter... Cette image résume bien la philosophie de la méthode des sondages : observation d'une partie d'un domaine d'études, l'échantillon, afin de produire de l'information sur l'ensemble.

Cette méthode a fait ses preuves depuis plusieurs dizaines d'années dans le contexte des pays en développement, dans le cadre de l'étude d'un certain nombre de sujets pour lesquels il est impossible d'enquêter l'ensemble des unités : enquêtes démographiques, dont une partie résulte de grands programmes lancés par des organismes internationaux, enquêtes socio-économiques, comme celles sur le budget des ménages, ou encore enquêtes agricoles.

Lors de l'application de la méthode, un certain nombre de contraintes pratiques, liées à l'organisation de l'enquête, à son budget..., viennent interférer dans les choix qui pourraient résulter de considérations uniquement théoriques. L'objet de ce manuel est de présenter la théorie de manière simple, en ne proposant par exemple pas de démonstrations mathématiques des formules énoncées, et d'essayer de faire le lien avec la pratique, en s'appuyant sur un certain nombre d'exemples d'enquêtes réalisées dans les pays en développement, et parfois également dans les pays développés.

Ces exemples peuvent être approfondis en se référant aux sources citées en bibliographie ; certaines imprécisions ou erreurs peuvent subsister concernant ces exemples - et nous nous en excusons auprès des concepteurs des enquêtes -, en particulier pour les enquêtes réalisées dans un grand nombre de pays et pour lesquelles des ajustements par pays existent quant à la méthode de sondage utilisée. La théorie pourra, elle aussi, être approfondie en se référant aux ouvrages indiqués dans la bibliographie.

Le manuel débute par une introduction à la théorie des sondages (chapitres 1 et 2), il passe ensuite en revue les différentes méthodes utilisées : stratification, sondages à probabilités inégales, sondages à plusieurs degrés (chapitres 3 à 5). Ces

méthodes ne sont pas exclusives l'une de l'autre, et sont souvent combinées dans les plans de sondage mis en place pour les enquêtes réalisées sur différents sujets.

C'est l'objet du chapitre 8 que de mettre en perspective l'ensemble des méthodes présentées, et d'insister sur un certain nombre de points pratiques qu'on ne peut négliger quand on procède à la conception ou au traitement d'une enquête par sondage. Un de ces points est relatif au problème des redressements, en particulier en raison des non-réponses, abordé au chapitre 6. Enfin, le chapitre 7 présente la méthode des quotas, très utilisée dans certains pays développés, peu actuellement dans les pays en développement, mais qui pourrait y connaître des champs d'application.

CHAPITRE 1

PRINCIPES

La technique des sondages permet de produire de l'information sur un domaine donné à partir de l'observation d'une partie de ce domaine. Elle s'applique particulièrement à l'étude des populations nombreuses. Avant d'aborder les différentes méthodes de sondage, il est nécessaire de présenter un certain nombre de notions qui seront utilisées tout au long de cet ouvrage.

1. Quelques définitions : univers, unités statistiques, échantillon, variables

a) Univers

Le domaine étudié est souvent qualifié d'univers ou de population. Il s'agit d'une population au sens statistique du terme, c'est-à-dire qu'on parlera de population d'individus, mais aussi de population de villages, de champs ou d'événements (naissances, décès, migrations...).

L'univers étudié doit être défini de manière précise, que ce soit du point de vue des unités élémentaires le composant (voir paragraphe suivant) ou du point de vue de ses limites : si, par exemple, on étudie le domaine de l'agriculture, décide-t-on d'inclure ou non les jardins familiaux dans l'univers ? Cette définition des limites de celui-ci conditionne la portée des résultats qu'on tirera du sondage.

b) Unités statistiques

Les unités statistiques sont les éléments composant l'univers. Elles peuvent être de différents types :

- individus au sens courant du terme,

- villages,
- hameaux, quartiers, îlots, etc.,
- ménages,
- parcelles cultivées,
- etc.

Un même univers peut être décomposé selon différents types d'unités élémentaires (par exemple en ménages ou en individus).

On peut aussi être amené à considérer une décomposition de l'univers en unités à plusieurs degrés[1], chaque unité d'un degré donné étant elle-même composée d'unités du degré suivant. Par exemple, du point de vue démographique, une zone rurale peut être décomposée en villages, unités du premier degré (unités primaires) composées :

- de ménages, unités du second degré (unités secondaires), eux-mêmes composés :
- d'individus, unités du troisième degré (unités tertiaires).

c) Échantillon

On appelle échantillon un sous-ensemble d'unités statistiques prélevé dans l'univers, dont on veut connaître certaines caractéristiques. C'est à partir des résultats observés sur l'échantillon qu'on va "extrapoler" pour produire des estimations sur cet univers (figure 1).

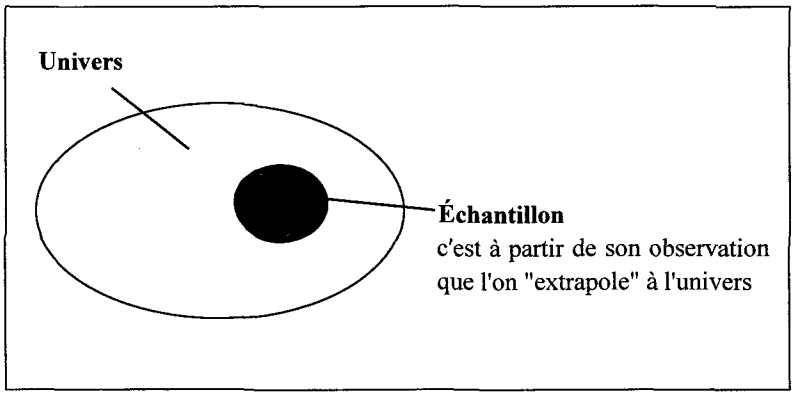


Figure 1. Principe de la méthode des sondages

[1] Les sondages à plusieurs degrés seront abordés plus loin (chapitre 5).

d) Variables

Les études statistiques portent sur les valeurs prises par certaines variables (caractères) pour chacune des unités statistiques. Ces variables peuvent être quantitatives continues (taille, revenu), discontinues (nombre d'enfants) ou qualitatives (situation matrimoniale, nationalité).

Parmi les variables qualitatives, on distingue les variables dichotomiques qui ne présentent que deux modalités, par exemple le sexe ou le fait d'être en vie ou décédé.

Les variables quantitatives peuvent se compter ou se mesurer, les variables qualitatives ne peuvent en principe qu'être "qualifiées". Toutefois, dans le cas des variables dichotomiques, on attribuera souvent la valeur 0 à l'une des modalités, et 1 à l'autre. Par exemple :

0 pour une personne du sexe masculin,
1 pour une personne du sexe féminin.

Ainsi codée, cette variable permet d'obtenir la proportion de femmes dans la population en calculant la moyenne des valeurs observées. On se trouve alors dans une situation comparable à celle des variables quantitatives discontinues.

2. Estimateur, variable aléatoire, variance, biais

a) Estimateur

Ayant utilisé un procédé de sondage déterminé (on reviendra tout au long des chapitres suivants sur la manière de faire), on va chercher à produire des estimations pour une ou plusieurs variables dites d'intérêt. Un estimateur associé à un procédé de sondage utilisé est une "formule mathématique" qui permet de calculer l'estimation d'une grandeur à partir des données observées sur l'échantillon tiré. On voit que, pour un procédé de sondage déterminé, le "hasard" peut conduire à différents échantillons, donc à différentes estimations (calculées à partir de l'estimateur).

En outre, pour une méthode de sondage déterminée, on aura souvent, en fait, un choix d'estimateurs. Les chapitres 2 à 5 présentent les estimateurs les plus "naturels" associés aux différentes méthodes - et on parlera dans ce cas de

"l'estimateur" comme s'il était unique - mais on verra au chapitre 6 d'autres estimateurs plus complexes.

b) Variable aléatoire

Une variable aléatoire est une variable qui peut prendre "un certain nombre de valeurs" avec, pour chaque valeur, une probabilité correspondante : on a donc une "distribution" de la variable aléatoire.

Si l'on s'intéresse au domaine des sondages, on a vu qu'un échantillon fournit une estimation de la grandeur qu'on cherche à estimer ; mais si l'on tire un autre échantillon selon les mêmes règles de sélection, on aura sans doute un autre résultat pour l'estimation de la grandeur étudiée. **L'estimateur est donc une variable aléatoire**.

La distribution de l'estimateur est fournie par l'ensemble des résultats obtenus à partir de l'ensemble des échantillons possibles (selon le procédé de sondage qu'on s'est fixé) ; le caractère aléatoire provient du "tirage au sort" de l'échantillon.

c) Moyenne, variance

La variance d'une variable Y donne une idée de la dispersion de Y autour de sa moyenne. Elle vaut :

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

où $\overline{Y}$ est la moyenne de Y sur l'univers : $\overline{Y} = \frac{1}{N}\sum_{\alpha=1}^{N} Y_\alpha$

Y_α est la valeur de Y pour l'unité statistique α,
et N le nombre d'unités statistiques de l'univers.

On définit aussi la covariance de deux variables sur l'univers, qui fournit une mesure de la manière dont deux variables Y et Z varient simultanément :

$$cov\,(Y,Z) = \frac{1}{N}\sum_{\alpha=1}^{N}\left(Y_\alpha - \overline{Y}\right)\left(Z_\alpha - \overline{Z}\right),$$

et le coefficient de corrélation linéaire entre Y et Z, qui mesure la "solidité" de la relation linéaire entre Y et Z :

$$\rho = \frac{cov(Y,Z)}{\sqrt{V(Y)V(Z)}}$$

Cette grandeur est comprise entre -1 et +1.

Pour une variable aléatoire A, on parle d'espérance et de variance.

L'espérance est définie par : $E(A) = \sum_i P_i Y_i$

où les Y_i sont tous les résultats possibles[2],

et où P_i, pour i donné, est la probabilité que A prenne la valeur Y_i.

De manière intuitive, l'espérance est la valeur que prend "en moyenne" la variable aléatoire A.

Quant à la variance, elle se définit comme suit : $V(A) = \sum_i P_i (Y_i - E(A))^2$

On voit qu'on utilise, dans le cadre des enquêtes par sondage, le terme de variance à la fois pour les variables dans l'univers étudié (donc mesurables sur chaque unité statistique) et pour la variable aléatoire "estimateur résultant du plan de sondage".

Cette double utilisation du terme variance peut conduire à certaines confusions. Il sera nécessaire de toujours préciser de quelle variance il s'agit quand on traite de problèmes de sondages.

d) Écart-type, coefficient de variation

L'écart-type d'une variable Y est la racine carrée de la variance :

$$\sigma_Y = \sqrt{V(Y)}$$

Cette notion s'applique aussi bien à la variance d'une variable qu'à la variance d'une variable aléatoire. Par ailleurs, l'écart-type s'exprime dans la même unité que la variable, alors que la variance s'exprime en cette unité "au carré".

Le coefficient de variation est :

$$\text{c.v.} = \frac{\sigma_Y}{\bar{Y}} \text{ pour une variable } Y$$

[2] Dans l'expression définissant *E(A)*, la somme symbolisée par le Σ ne s'applique pas nécessairement à un ensemble fini de valeurs possibles.

$$\text{c.v.} = \frac{\sigma_A}{E(A)} \text{ pour une variable aléatoire } A.$$

L'intérêt de cette grandeur est qu'on trouve au numérateur et au dénominateur des valeurs exprimées dans les mêmes unités ; elle est donc "sans unité" et donne une idée de l'importance de l'écart-type par rapport à la moyenne (ou l'espérance), donc une idée de la plus ou moins grande dispersion de la distribution.

e) Biais

On dit qu'un estimateur A d'une grandeur G est sans biais si $E(A) = G$, c'est-à-dire si "en moyenne" les résultats fournis par cet estimateur sont égaux à la grandeur qu'on cherche à estimer. Dans le cas contraire, on a un estimateur biaisé, qui peut néanmoins dans certaines conditions être acceptable (partie 4).

3. Sondage aléatoire, sondage non aléatoire, base de sondage

a) Sondage aléatoire

Un sondage est dit aléatoire, ou probabiliste, si toute unité statistique a une probabilité non nulle et connue d'être sélectionnée dans l'échantillon. La méthode des sondages aléatoires est fondée sur le principe que l'échantillon doit être déterminé d'une façon objective, dans laquelle aucun "facteur personnel" n'intervient, de façon que tout élément de l'ensemble à étudier ait des chances d'être choisi et que ces chances puissent être déterminées avec certitude. Ce qui veut dire que, pour le choix d'un échantillon, on fait appel au hasard (au sens probabiliste du terme qu'il faut bien distinguer du sens qu'il a souvent dans le langage courant), et qu'on pourra mettre en œuvre une formalisation mathématique pour étudier les propriétés de cet échantillon. Certaines méthodes, ne respectant pas la condition du sondage aléatoire, sont présentées au chapitre 7.

b) Base de sondage

Le sondage aléatoire nécessite de donner à toute unité statistique de l'univers une probabilité non nulle d'être sélectionnée : d'où la nécessité de disposer d'une base de sondage, afin de pouvoir accéder à l'ensemble des unités statistiques. Une

base de sondage est une liste complète et à jour des unités de l'univers sans omission ni double compte, et telle que l'identification de chaque unité se fasse sans ambiguïté.

Le terme de "liste" doit être entendu au sens large : s'il s'agit en général d'un fichier (manuel ou informatique) issu d'un recensement ou d'une source administrative, il peut aussi s'agir d'une couverture photographique aérienne qu'on va découper en zones élémentaires quand on utilise une méthode de sondage aréolaire, ou d'un autre moyen d'accéder aux unités statistiques.

Il est intéressant de disposer, dans la base de sondage, d'informations concernant les unités statistiques (en plus, bien sûr, de leur localisation) utilisables pour le sondage. Par exemple, dans une enquête démographique, la base de sondage peut indiquer pour chaque village de l'univers une estimation de la population : population au dernier recensement, nombre de personnes imposables, nombre de ménages, etc. Ces renseignements sont appelés variables auxiliaires. Ces variables auxiliaires peuvent être utilisées, soit pour améliorer la technique de tirage, soit pour calculer une estimation plus efficace (estimateurs plus complexes proposés au chapitre 6).

Qu'utilise-t-on comme base de sondage ? On peut fournir plusieurs types d'exemples :

- des documents administratifs existants, par exemple des listes fiscales (qu'il faudra compléter ou corriger éventuellement) ;
- le fichier des clients d'une société, ou des anciens élèves d'une école ;
- une liste venant d'une enquête précédente, en particulier d'un recensement ;
- une liste qui est dressée à l'occasion de l'enquête : on peut, à l'occasion d'une enquête sur un centre urbain, procéder à un dénombrement des ménages à partir duquel on tirera l'échantillon ;
- une liste d'unités aréolaires (en particulier les zones de dénombrement utilisées pour le recensement de la population) dans laquelle on tirera un échantillon d'unités pour lesquelles on procédera à un dénombrement des ménages (chapitres 5 et 8).

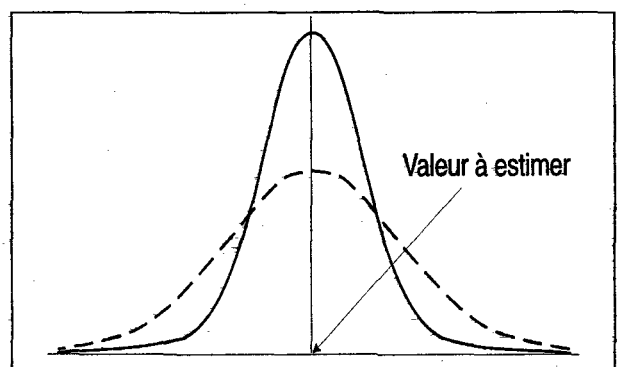
Si l'on revient aux "qualités" de la base de sondage (exhaustivité et absence de doubles comptes), on se trouve rarement dans la situation idéale où on a une base de sondage parfaite ; on fera cependant avec la base dont on dispose. On reviendra, au chapitre 8, sur les problèmes de mise à jour de la base de sondage.

4. "Qualités" d'une enquête par sondage

a) La recherche de précision

On cherchera à ce que l'estimation obtenue à partir de l'échantillon soit, en moyenne, la plus proche possible de la grandeur (inconnue) qu'on veut estimer.

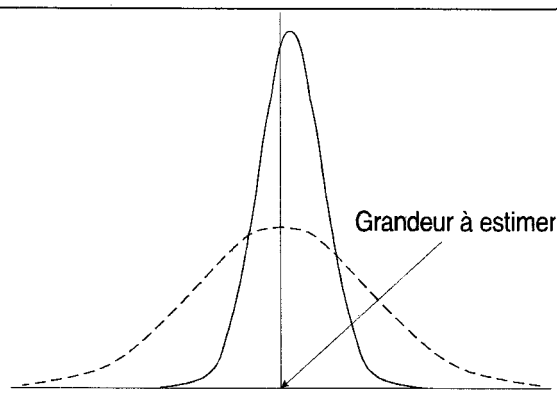
Souvent, les estimateurs étudiés seront sans biais et on cherchera, dans ce cas (figure 2), à avoir la variance de l'estimateur la plus petite possible (soit encore l'écart-type le plus petit possible, ou encore le coefficient de variation le plus petit possible).



L'estimateur dont la distribution est représentée par la courbe en plein est préférable à celui associé à la courbe en pointillé, car sa distribution est plus concentrée autour de la valeur à estimer.

Figure 2. Comparaison des distributions de deux estimateurs sans biais

Parfois, on acceptera des estimateurs biaisés à condition que l'écart-type qui leur est associé soit d'un ordre de grandeur "dominant" par rapport à celui du biais (figure 3).



Bien que l'estimateur dont la distribution est représentée par la courbe en plein soit biaisé, on le préférera toutefois à celui associé à la courbe en pointillé, car les résultats qu'il fournit sont globalement plus proches de la valeur à estimer.

Figure 3. Comparaison entre un estimateur sans biais et un estimateur biaisé

b) La notion de "représentativité"

La notion de "représentativité" est une notion à manipuler avec précaution : elle sous-entend que l'échantillon donne une image réduite, mais fidèle, de l'ensemble sur lequel il est prélevé, au moins en ce qui concerne les caractéristiques que l'on cherche à évaluer. On verra dans les chapitres suivants (en particulier celui sur la stratification) qu'il n'est pas nécessaire que l'échantillon soit une maquette "exacte" de l'univers : certaines parties de celui-ci peuvent être surreprésentées dans l'échantillon (à condition bien sûr d'en tenir compte dans l'estimateur).

Il ne faut pas non plus perdre de vue que des choix au niveau de la technique de sondage peuvent donner un résultat très satisfaisant en ce qui concerne une variable (par exemple, le taux de natalité) mais des résultats beaucoup moins précis en ce qui concerne d'autres caractéristiques (par exemple l'emploi ou les migrations). Le problème devient complexe lorsqu'on entreprend une enquête à objectifs multiples : par exemple démographie, habitat, consommation, nutrition, budget de famille, etc. Le choix d'une méthode implique alors des arbitrages entre des impératifs qui peuvent être très différents, voire contradictoires (qualité, coût, etc).

c) *Affiner le mode de tirage* a priori, *mais aussi améliorer la précision* a posteriori

Les chapitres suivants présentent les différentes options qu'on peut utiliser pour "améliorer" la qualité des estimateurs en tenant compte des variables connues *a priori* avant le tirage afin de sélectionner "au mieux" l'échantillon, mais aussi après l'observation de l'échantillon en tenant compte de certaines informations auxiliaires (chapitre 6 sur les redressements).

d) Les erreurs d'observation

Aux erreurs d'échantillonnage (dues au fait qu'on n'observe qu'une partie de l'univers) viennent "s'additionner" les erreurs d'observation. Ces erreurs peuvent avoir un côté "aléatoire" et se compenser ; il existe cependant certaines erreurs systématiques qui introduisent un biais et faussent les résultats de façon irrémédiable (par exemple mauvaise compréhension d'une question).

L'erreur d'observation est difficilement évaluable ; sa maîtrise sera bien sûr liée au soin apporté à l'ensemble des étapes de l'enquête, depuis la conception du questionnaire jusqu'au traitement des données en passant par la phase de terrain (collecte) et la saisie-codification. Pour la mise au point des questionnaires, et plus généralement les problèmes d'organisation d'enquêtes, on pourra se référer à d'autres ouvrages (Jacquart, 1988 ; Blaizeau, Dubois, 1989).

L'erreur d'observation doit être maîtrisée "le plus possible", car son ordre de grandeur peut être nettement supérieur à celui de l'erreur d'échantillonnage.

La mauvaise couverture du "champ" de l'enquête influe également sur la qualité des résultats : on reviendra sur ces problèmes (liés à la base de sondage utilisée) au chapitre 8.

5. Notations

Quelques généralités relatives aux notations, qui seront affinées au fur et à mesure des chapitres :

a) Sur l'univers

Unités statistiques $\alpha = 1, \ldots\ldots, N$

Moyenne de la variable Y : $\overline{Y} = \frac{1}{N}\sum_{\alpha=1}^{N} Y_\alpha$ où Y_α est la valeur de Y pour l'unité α.

Variance de Y : $V(Y) = \frac{1}{N}\sum_{\alpha=1}^{N}\left(Y_\alpha - \overline{Y}\right)^2$

On utilise souvent la notation σ^2 pour la variance : $V(Y) = \sigma^2$.

Par ailleurs, on définit aussi : $S^2 = \frac{1}{N-1}\sum_{\alpha=1}^{N}(Y_\alpha - \overline{Y})^2$

b) Sur l'échantillon

Unités statistiques $i = 1, \ldots, n$

$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ moyenne de la variable Y calculée sur l'échantillon : $\overline{y}$ est une variable aléatoire.

$E(\overline{y})$ espérance de la variable aléatoire $\overline{y}$.

$V(\overline{y})$ variance de la variable aléatoire $\overline{y}$ (qui, rappelons-le, est la variance de l'estimateur $\overline{y}$ et non la variance de la variable Y calculée sur les unités de l'échantillon).

c) Taux de sondage

$$f = \frac{n}{N}$$

d) L'utilisation de la notation ^

On utilisera parfois la notation ^ pour les estimateurs produits à partir de l'échantillon. Par exemple, pour estimer un total $T(Y)$, on utilisera un estimateur qu'on notera :

$$\hat{T}(Y).$$

CHAPITRE 2

SONDAGES ALÉATOIRES SIMPLES

1. Principe

De l'univers on extrait un échantillon de taille n, en accordant à chaque unité statistique la même probabilité d'être tirée. L'échantillon peut être tiré :

- avec remise : une unité ayant été sélectionnée à un des tirages est "remise dans l'urne de tirage" et participe aux tirages suivants ; elle peut donc être tirée deux fois, ou plus ;
- sans remise : une fois une unité tirée, elle n'est plus prise en compte pour les tirages suivants (c'est le mode de tirage qui semble le plus "naturel").

Le sondage aléatoire simple est la base de la méthode des sondages, à partir de laquelle sont développées les autres méthodes présentées dans ce manuel.

2. Estimation d'une moyenne

Pour estimer la moyenne $\overline{Y}$ d'une variable Y sur l'univers ($\overline{Y}$ est bien sûr inconnue) il semble naturel d'utiliser l'estimateur :

$$\boxed{\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i} \quad (1)$$ moyenne calculée sur les unités de l'échantillon[3]

[3] Dans le cas d'un tirage avec remise, si une unité est tirée plusieurs fois dans l'échantillon, sa valeur sera comptée dans cette formule pour autant de fois qu'elle a été tirée ; la valeur n correspondra alors au nombre de tirages, et non au nombre d'unités différentes tirées.

a) $\bar{y}$ est un estimateur sans biais de la grandeur $\bar{Y}$

Cette propriété, qui signifie "qu'en moyenne" les valeurs fournies par l'estimateur $\bar{y}$ tombent de part et d'autre de $\bar{Y}$, peut s'écrire :

$$E(\bar{y}) = \bar{Y}$$

b) Variance de $\bar{y}$

Dans le cas avec remise : $V(\bar{y}) = \frac{V(Y)}{n}$ **(2)**

Dans le cas sans remise : $V(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{N}{N-1}V(Y)$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n}S^2$$

$$= \left(\frac{N-n}{N-1}\right)\frac{V(Y)}{n} \quad \textbf{(3)}$$

Ceci veut dire que la variance de l'estimateur sera d'autant plus faible que :

- $V(Y)$ sera faible ;
- la taille de l'échantillon sera importante.

Par ailleurs, comme N-n/N-1 est toujours inférieur à 1, la variance de l'estimateur sans remise est plus faible que celle de l'estimateur avec remise pour une même variable étudiée. Cependant, quand N est grand, le coefficient N-n/N-1 est souvent proche de 1 ; les deux variances sont alors équivalentes.

c) Pour un échantillon suffisamment grand, $\bar{y}$ suit une loi normale

À partir d'une certaine taille d'échantillon (disons au moins 30), la distribution de la variable aléatoire $\bar{y}$ (c'est-à-dire l'ensemble des estimations fournies par tous les échantillons obtenus avec le tirage équiprobable de taille n) s'ajuste sur une loi normale (courbe "en cloche" de Gauss) dont les caractéristiques sont liées aux valeurs $E(\bar{y})$ et $V(\bar{y})$ étudiées ci-dessus.

Ce résultat fondamental vient du "théorème central limite" et, ceci doit être souligné, il est indépendant de la forme de la distribution de la variable Y dans l'univers.

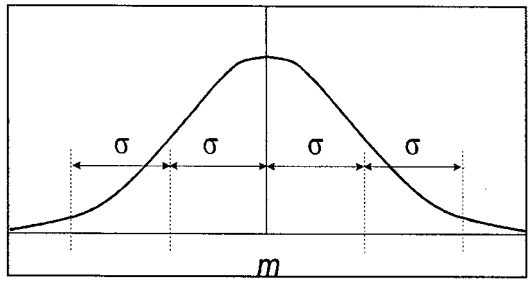


Figure 4. Distribution de la loi normale

On sait que la distribution d'une loi normale (figure 4) de moyenne m et d'écart-type σ comprend environ les 2/3 des valeurs dans l'intervalle $[m-\sigma ; m+\sigma]$, et 95 % des valeurs[4] dans l'intervalle $[m-2\sigma ; m+2\sigma]$.

Ici, on peut donc dire que 95 % des valeurs de $\bar{y}$ sont situées dans l'intervalle $\left[\bar{Y}-2\sqrt{V(\bar{y})},\bar{Y}+2\sqrt{V(\bar{y})}\right]$, où $V(\bar{y})$ est donnée par les formules **(2)** ou **(3)** ci-dessus.

d) La variance $V(\bar{y})$ peut être estimée à partir de l'échantillon

Dans les formules **(2)** et **(3)** ci-dessus, la quantité $V(Y)$ est inconnue. Celle-ci va être estimée à partir des données observées sur l'échantillon. Si l'on note s^2 la grandeur calculée sur l'échantillon :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

alors on a le résultat suivant :

. dans le cas sans remise, s^2 est un estimateur sans biais de $\frac{N}{N-1}V(Y)$ (ou encore S^2),

. dans le cas avec remise, s^2 estime sans biais $V(Y)$.

$V(\bar{y})$ est donc estimée sans biais par :

$$\hat{V}(\bar{y}) = \frac{s^2}{n} \text{ dans le cas du tirage avec remise}$$

[4] En toute rigueur c'est dans l'intervalle $[m-1{,}96\sigma ; m+1{,}96\sigma]$; mais on se contentera d'arrondir la valeur 1,96 à 2.

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$ dans le cas du tirage sans remise

Remarque importante : on retrouve ici la distinction entre $V(\bar{y})$, variance de l'estimateur $\bar{y}$ (donc d'une variable aléatoire) et $V(Y)$, variance de la variable Y qui est estimée à partir de l'échantillon.

Une erreur, malheureusement trop souvent rencontrée, consiste à utiliser comme estimation de $V(\bar{y})$ l'estimation de $V(Y)$; en fait on voit qu'à quelques approximations près dans le cas sans remise, on passe de $V(Y)$ à $V(\bar{y})$ en divisant la première grandeur par n.

e) Dans la pratique on n'a tiré qu'un échantillon

Ce qui vient d'être dit aux paragraphes précédents doit être replacé dans cette perspective. Les résultats ci-dessus présentent la manière dont l'ensemble des valeurs calculées sur tous les échantillons possibles se répartissent par rapport à la grandeur recherchée $\bar{Y}$.

En pratique, le seul résultat dont on dispose est la moyenne $\bar{y}$ calculée sur un échantillon, et $\bar{Y}$ est inconnue. On tient un raisonnement analogue au précédent, mais à partir de $\bar{y}$ (encadré 1), pour fournir un "intervalle de confiance" pour $\bar{Y}$:

- on dispose de la valeur $\bar{y}$;
- on estime $V(\bar{y})$ à partir de l'échantillon, on obtient donc une estimation $\hat{\sigma}(\bar{y})$ (racine carrée de la variance estimée $\hat{V}(\bar{y})$)[5] ;
- on peut donc fournir un intervalle de confiance : à 95 chances sur 100[6], la grandeur $\bar{Y}$ est dans l'intervalle $\left[\bar{y} - 2\hat{\sigma}(\bar{y}); \bar{y} + 2\hat{\sigma}(\bar{y})\right]$. Ceci donne une idée de la précision du sondage.

---

[5] Cette estimation de la variance est elle-même sujette à une erreur de sondage qu'il est d'ailleurs possible d'estimer, et ainsi de suite. En fait, en général, on n'effectue pas ce calcul et on "fait comme si" l'estimation de la variance était la vraie valeur ; ceci incite à une certaine prudence dans l'interprétation des chiffres...

[6] Le sondeur reconnaît donc avoir quelques chances (5 sur 100) de se tromper. On peut, en utilisant une table relative à la loi normale, fournir des intervalles de confiance à d'autres valeurs, par exemple 99 % ou encore 90 %...

Encadré 1

Récapitulation sur l'estimation d'une moyenne dans le cas d'un sondage aléatoire simple sans remise

La moyenne $\overline{Y}$ est estimée à partir de l'échantillon par : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Cet estimateur est sans biais et sa variance est estimée à partir de l'échantillon par :

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}s^2$$

où $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

La "fourchette d'estimation" à 95 % est alors :

$$\left[\bar{y} - 2\sqrt{\hat{V}(\bar{y})}; \bar{y} + 2\sqrt{\hat{V}(\bar{y})}\right]$$

f) Remarques

La précision est essentiellement liée au nombre d'unités enquêtées

La précision, en termes de variance de $\bar{y}$, est essentiellement liée au nombre d'unités enquêtées n, et relativement peu[7] au taux de sondage n/N (pas du tout dans le cas avec remise). Ceci est un point fondamental. On peut l'illustrer de deux façons :

- deux pays de tailles différentes menant des enquêtes, à partir d'échantillons de même taille issus de sondages aléatoires simples, et sur des variables présentant la même dispersion $V(Y)$ (ce qui est souvent approximativement le cas si la variable Y étudiée est la même dans les deux pays), obtiendront des résultats équivalents en précision, bien que les taux de sondage soient différents (figure 5) ;

- quand on publie des résultats d'une enquête sur une partie de l'univers étudié (par exemple une région si on a réalisé une enquête sur un pays, ou encore si l'on s'intéresse aux résultats dans une "case" d'un tableau croisant deux modalités, par exemple l'âge du chef de famille et sa catégorie sociale), le taux de sondage est le même (pour un sondage équiprobable) au niveau de l'ensemble de l'univers (par

[7] Le taux de sondage $f = n/N$ intervient dans le cas du tirage sans remise par le coefficient $(1 - f)$. Il affecte la variance de manière sensible, et donc la précision, s'il est proche de 1, ce qui est rarement le cas en pratique.

exemple le pays) ou de la partie étudiée (par exemple la région) ; et pourtant, le nombre d'unités enquêtées relatif à cette partie est nettement plus faible, et la précision des résultats moindre. Ceci incite à être vigilant sur la qualité des résultats publiés dans les cases d'un tableau qui en comprend beaucoup.

Deux enquêtes tirées de façon aléatoire simple, utilisant 10 000 questionnaires sur le même thème d'étude, l'une menée sur l'ensemble du continent africain, l'autre sur un seul pays, auront une précision à peu près équivalente.

Celle-ci dépend donc peu de la taille du pays où a lieu l'enquête.

Figure 5. Précision du sondage et taille du pays dans lequel on réalise l'enquête

La variance est inversement proportionnelle au nombre d'unités enquêtées

La variance est proportionnelle à $1/n$; l'écart-type, qui permet d'établir l'intervalle de confiance, est proportionnel à $1/\sqrt{n}$. Ceci veut dire que pour réduire de moitié l'intervalle de confiance, il faut quatre fois plus de questionnaires.

Tirage avec remise, tirage sans remise

La différence de précision entre le tirage avec remise et le tirage sans remise est en faveur du tirage sans remise, mais est en général peu marquée, sauf dans le cas où le taux de sondage $f = n/N$ s'approche de 1 ; on utilisera quasiment toujours le cas sans remise.

Il est à noter que si, pour le cas sans remise, la variance est nulle quand $n = N$ (c'est-à-dire qu'il n'y a plus d'erreur d'échantillonnage), il subsiste une variance pour le tirage avec remise quand $n = N$.

Évaluation "chiffrée" de la précision

On utilise souvent, pour chiffrer la précision d'une enquête par sondage, le coefficient de variation. Dire qu'une enquête est précise à 2 % près pour l'estimation d'une moyenne $\overline{Y}$ veut dire que le coefficient de variation $\sigma(\bar{y})/\overline{Y}$ est de 2 % ; l'intervalle de confiance associé à 95 chances sur 100 - celui qui est en général utilisé - vaut dans ce cas $\left[\overline{Y}-2\sigma(\bar{y});\overline{Y}+2\sigma(\bar{y})\right]$, soit encore $\left[\overline{Y}-0{,}04\,\overline{Y};\overline{Y}+0{,}04\,\overline{Y}\right]$.

Combien d'unités enquêter dans l'échantillon ?

On voit que la réponse à cette question se trouve dans les formules **(2)** et **(3)** : si l'on fixe un niveau de précision (largeur de l'intervalle de confiance, ou coefficient de variation, ce qui signifie une valeur "maximum" de la variance), on en déduit le nombre d'unités minimum nécessaire. Le problème est que la quantité $V(Y)$ est inconnue avant l'enquête, dans les formules **(2)** et **(3)** ; la seule solution est de l'estimer, soit de manière intuitive, soit à partir de données observées sur le passé (l'estimation proposée de $V(Y)$ au paragraphe d) se situe une fois l'enquête réalisée, et on se place ici dans une phase de réflexion en amont de la réalisation sur le terrain).

3. Estimation d'un total

Les estimations de totaux sont en général des estimations d'inventaire (effectifs de migrants, de classes d'âges, ...). Le total d'une variable Y est estimé, à partir de l'échantillon, par l'estimateur de sa moyenne multiplié par l'effectif de l'univers :

$$\hat{T}(Y)=N\,\bar{y} \qquad \textbf{(4)}$$

On voit apparaître dans $N\,\bar{y} = N\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{N}{n}\sum_{i=1}^{n} y_i$ la "pondération" de chaque unité de l'échantillon *N/n*, encore appelée coefficient d'extrapolation (qui permet "d'étendre à l'univers" la donnée observée sur cette unité).

La variance[8] de cet estimateur vaut :

$$V\left(\hat{T}(Y)\right)=N^2V(\bar{y}).$$

[8] On notera au passage que le fait de "sortir *N* de la parenthèse" amène à multiplier $V(\bar{y})$ par le carré de *N* (une erreur répandue est de multiplier $V(\bar{y})$ par seulement *N* pour calculer $V(N\,\bar{y})$).

$V(N\bar{y})$ peut donc être estimée à partir de l'échantillon :

- dans le cas avec remise par $\frac{N^2}{n}s^2$

- dans le cas sans remise par $N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}$

4. Estimation d'une proportion

a) Principe

Une proportion sur l'univers (par exemple le pourcentage de femmes) est la moyenne d'une variable indicatrice :

$Y_\alpha = 1$ si l'unité α a la caractéristique correspondant à la proportion (être une femme par exemple) ;

$Y_\alpha = 0$ sinon.

L'estimation d'une proportion est donc l'estimation de la moyenne de cette variable.

Une des caractéristiques d'une telle variable est que sa variance s'écrit de manière simple :

si P est la proportion recherchée, $P = \frac{1}{N}\sum_{\alpha=1}^{N} Y_\alpha = \frac{N_\alpha}{N}$,

où N_α est le nombre d'unités correspondant à la caractéristique étudiée.

Le calcul montre[9] que : $V(Y) = P(1-P)$.

Dans ce cas, si l'on a une idée de l'ordre de grandeur de P (mais sa véritable valeur est inconnue et c'est l'enquête par sondage qui cherche à l'estimer de manière

---

[9] En effet, $\bar{Y} = P$ et $V(Y) = \frac{1}{N}\sum_{\alpha=1}^{N}\left(Y_\alpha - \bar{Y}\right)^2$

$$= \frac{1}{N}\left[N_\alpha(1-P)^2 + (N-N_\alpha)P^2\right] = \frac{1}{N}\left[N_\alpha - 2N_\alpha P + NP^2\right]$$

$$= \frac{1}{N}\left[N_\alpha - 2N_\alpha P + N_\alpha P\right] = \frac{N_\alpha}{N}\left[1-P\right]$$

précise), on pourra anticiper la précision en fonction du nombre de questionnaires puisqu'on aura *a priori* un ordre de grandeur de la variance de Y.

b) Exemple

Le tableau 1, tiré de ORSTOM-INSEE-INED (1971), fournit des éléments de précision pour l'estimation des taux[10] bruts de natalité et de mortalité à partir d'un sondage aléatoire simple et de différentes tailles d'échantillon.

Tableau 1. Précision de l'estimation par sondage aléatoire simple des taux de natalité et de mortalité

Taille de l'échantillon	Taux de natalité (45 ‰)		Taux de mortalité (20 ‰)	
	2σ	c.v.	2σ	c.v.
25 000	2,6 ‰	2,9 %	1,8 ‰	4,4 %
50 000	1,9 ‰	2,1 %	1,3 ‰	3,1 %
100 000	1,3 ‰	1,5 %	0,9 ‰	2,2 %

Interprétation : les taux de natalité et de mortalité à estimer sont situés vers les valeurs 45 ‰ et 20 ‰. Si l'on néglige le taux de sondage, on peut dire que :

$$V(\bar{y}) = \frac{V(Y)}{n}$$

Or $V(Y) = P(1-P)$.

Dans le cas $n = 25\,000$ et pour le taux de natalité ($P = 45$ ‰ $= 0{,}045$), on trouve que $V(\bar{y}) = (1{,}3\text{ ‰})^2$ d'où la largeur de l'intervalle de confiance à 95 chances sur 100 et la valeur du coefficient de variation *c.v.* (égal à σ/45 ‰ dans ce cas).

[10] On assimile ici les taux à des proportions, ce qui n'est pas tout à fait rigoureux puisqu'on divise par exemple le nombre de décès d'une année par l'effectif moyen de la population ; pour être rigoureux, il faudrait se placer dans le cadre d'observations longitudinales (suivi de cohortes).

5. Estimation d'un ratio

L'estimation d'un ratio peut être délicate, et révéler des pièges. Prenons un exemple : supposons que l'univers soit un univers de ménages (la base de sondage est une liste de ménages), et que certaines caractéristiques comme le nombre d'enfants de moins de cinq ans ne soient pas connues.

Comment estimer le poids corporel moyen des enfants de moins de cinq ans à partir d'un échantillon de ménages tiré de façon aléatoire simple ? Remarquons que l'unité statistique utilisée pour le sondage est le ménage et non l'individu. On procède ainsi :

- on estime le nombre total d'enfants de moins de cinq ans ;
- on estime ensuite le poids corporel total des enfants de moins de cinq ans de l'univers ;
- le ratio (ou quotient) de ces deux masses est l'estimation du poids moyen des enfants de moins de cinq ans.

Ceci revient en fait à estimer (même si la première estimation $\bar{y}$ peut paraître "artificielle") :

- $\bar{y}$ poids corporel moyen "cumulé" par ménage des enfants de moins de cinq ans y vivant ;
- $\bar{x}$ nombre moyen d'enfants de moins de cinq ans par ménage.

L'estimateur final est $\bar{y}/\bar{x}$, estimateur du poids moyen des enfants de moins de cinq ans[11].

Cet estimateur n'est plus, contrairement aux estimateurs proposés précédemment, sans biais, car, en général, $E(\bar{y}/\bar{x})$ n'est pas égal à $E(\bar{y})/E(\bar{x})$.

[11] Une erreur à ne pas commettre est de calculer la moyenne simple des poids moyens des enfants de moins de cinq ans des ménages de l'échantillon ; cette estimation ne tiendrait pas compte du fait que tous les ménages n'ont pas le même nombre d'enfants de moins de cinq ans.

On peut considérer, quand l'échantillon est de taille suffisamment importante, que ce biais est "négligeable" par rapport à l'erreur aléatoire[12].

Pour estimer un ratio, on passe en général par l'estimation de deux masses.

6. Méthodes de tirage

a) Méthode simple

L'idée est de numéroter les unités statistiques, et de procéder à un tirage au hasard de numéros (entre 1 et N). Pour ce faire, on peut utiliser :

- une table de nombres au hasard (qu'on parcourt dans un sens bien défini au départ, par exemple ligne par ligne) ;
- un algorithme informatique de tirage (par exemple par génération au hasard d'un nombre réel entre 1 et N).

b) Tirage systématique

Une autre méthode est celle du tirage systématique : on procède par "sauts" dans la liste des unités statistiques.

Exemple

On doit tirer 10 personnes parmi 153 personnes (numérotées).
Le "pas de tirage" sera de 153/10 = 15,3.
On tire un premier nombre au hasard entre 1 et 15 : 3

---

[12] Pour un calcul de l'erreur aléatoire, voir Desabie (1971), chapitre 9 ou Ardilly (1994), chapitre 3.3. On montre que si $\hat{r}$ est l'estimation du ratio $\bar{y} / \bar{x}$, alors une estimation de la variance de $\hat{r}$ est donnée, dans le cas d'un sondage aléatoire simple sans remise, par :

$$\hat{V}(\hat{r}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{\bar{x}^2} \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{r} x_i)^2$$

On enquête les personnes :

- 3	→	n° 3
- (3 + 15,3)	→	n° 18
- (3 + 2 x 15,3)	→	n° 34
...		...
- (3 + 9 x 15,3)	→	n° 141

Cette méthode équivaut à la méthode élémentaire si les unités de la base de sondage sont réparties absolument au hasard ; elle risque d'être peu représentative dans le cas, très rare en pratique, où les unités présenteraient pour les caractères étudiés une périodicité du même ordre que l'inverse de la fraction de sondage ; elle est plus avantageuse que la méthode élémentaire si les unités ont été classées suivant un critère (variable auxiliaire) en corrélation avec les caractères à estimer et si l'on suppose qu'il existe un "effet" dans l'ordre de classement qui va donner plus de "représentativité" à l'échantillon tiré.

Par exemple, si on a classé des ménages selon leur taille, on sera assuré par ce mode de tirage d'avoir à la fois des ménages de faible taille et des ménages de taille importante dans l'échantillon. On peut donc dire que, même si on a présenté cette méthode de tirage dans le chapitre sur les sondages aléatoires simples, elle correspond en fait déjà à une version plus élaborée de tirage.

CHAPITRE 3

SONDAGES STRATIFIÉS

1. Principe, objectifs

Stratifier un univers consiste à le répartir avant le tirage de l'échantillon en sous-ensembles homogènes (par rapport à certains caractères connus *a priori*), appelés strates. Le tirage s'effectue de manière indépendante à l'intérieur de chaque strate (figure 6).

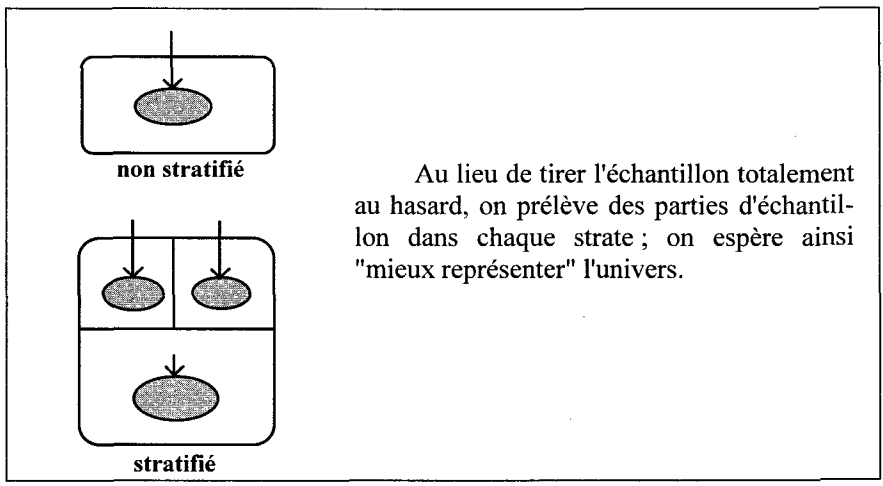


Figure 6. Principe du sondage stratifié

La stratification peut avoir comme objectif principal :

- soit d'augmenter la précision d'ensemble ;

- soit d'obtenir une précision suffisante au niveau de chacune des strates[12].

Ces deux objectifs ne doivent pas être confondus. Une fois que l'on aura réparti la base de sondage entre les strates, il y aura un choix à faire en ce qui concerne la répartition de l'échantillon entre ces strates. Ce choix dépendra de l'objectif que l'on juge prioritaire.

On peut également être amené à stratifier la population pour des raisons techniques : application de méthodes de tirage différentes suivant la strate (par exemple en milieu sédentaire et en milieu nomade, etc.).

2. Formules d'estimation

On se place ici dans le cas d'un tirage, à l'intérieur de chaque strate, aléatoire simple sans remise[13].

a) Notations

- On a k strates ($h = 1, 2, ..., k$)

- Pour la strate h, l'effectif total est N_h ($N = \sum_{h=1}^{k} N_h$)

la moyenne de Y est $\overline{Y}_h$

$$S_h^2 = \frac{1}{N_h - 1} \sum_{\alpha_h=1}^{N_h} \left(Y_{\alpha_h} - \overline{Y}_h \right)^2$$

le nombre d'unités tirées est n_h

l'indice des unités de l'échantillon est i_h ($i_h = 1, ..., n_h$)

$$\overline{y}_h = \frac{1}{n_h} \sum_{i_h=1}^{n_h} y_{i_h}$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i_h=1}^{n_h} \left(y_{i_h} - \overline{y}_h \right)^2$$

---

12 Par exemple, on cherche à comparer le niveau de la fécondité en milieux urbain et rural. On stratifiera la population suivant ce critère et l'on s'efforcera d'obtenir à peu près le même écart-type pour les estimations dans chacune de ces strates.

13 De manière générale, on se placera dans le cadre de sondages sans remise dans la suite de ce manuel. Cependant, pour des raisons exposées par la suite, on verra l'utilisation de tirages avec remise pour les sondages à probabilités inégales.

b) Estimation du total de Y sur l'univers à partir du sondage stratifié

Pour la strate h le total de Y est estimé par $N_h\bar{y}_h$, l'estimation du total de Y sur l'univers est donc :

$$\hat{T}(Y) = \sum_{h=1}^{k} N_h \bar{y}_h \qquad \textbf{(1)}$$

Remarque : cette formule peut aussi s'écrire :

$$\hat{T}(Y) = \sum_{h=1}^{k}\left[N_h \frac{1}{n_h} \sum_{i_h=1}^{n_h} y_{i_h} \right] = \sum_{h=1}^{k}\left[\sum_{i_h=1}^{n_h} \frac{N_h}{n_h} y_{i_h} \right]$$

Toute unité observée de l'échantillon est pondérée par le coefficient N_h/n_h (dont la valeur dépend de la strate), afin d'extrapoler (ou "d'étendre") les résultats à l'univers : ce coefficient est souvent appelé coefficient d'extrapolation.

c) Estimation de la moyenne de Y sur l'univers à partir du sondage stratifié

Pour cela on utilise l'estimation du total de Y divisée par le nombre total d'unités de l'univers N (N est connu).

L'estimateur est :

$$\hat{\bar{Y}} = \frac{1}{N} \sum_{h=1}^{k} N_h \bar{y}_h \qquad \textbf{(2)}$$

puisque

$$\hat{\bar{Y}} = \frac{\hat{T}(Y)}{N}$$

L'estimation d'une proportion (proportion de femmes par exemple) se fera, comme présenté au chapitre 2, par l'estimation de la moyenne d'une variable qui vaut 1 si l'unité a la caractéristique étudiée et 0 sinon.

d) Les estimateurs $\hat{T}(Y)$ et $\hat{\bar{Y}}$ sont des estimateurs sans biais du total et de la moyenne de Y

$$E\left(\hat{\bar{Y}}\right) = \bar{Y} \qquad \text{et} \qquad E\left(\hat{T}(Y)\right) = T(Y)$$

e) Variance de l'estimateur du total et de l'estimateur de la moyenne

$$V\left(\hat{T}(Y)\right) = V\left(\sum_{h=1}^{k} N_h \bar{y}_h\right) = \sum_{h=1}^{k} N_h^2 V\left(\bar{y}_h\right)$$

$$\text{avec } V\left(\bar{y}_h\right) = \left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} S_h^2$$

(résultat du sondage aléatoire simple sans remise, chapitre 2).

Ceci peut être écrit car les tirages dans chaque strate se font de manière indépendante, et donc les variables aléatoires $\bar{y}_h$ sont indépendantes.

Alors
$$V\left(\hat{T}(Y)\right) = \sum_{h=1}^{k} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} S_h^2 \qquad \textbf{(3)}$$

et
$$V\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{k} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} S_h^2 \qquad \textbf{(4)}$$

f) Estimation de ces variances d'estimation à partir de l'échantillon

$$\hat{V}\left(\hat{T}(Y)\right) = \sum_{h=1}^{k} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} s_h^2$$

$$\hat{V}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{k} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} s_h^2$$

Ces deux estimations de la variance des estimateurs du total et de la moyenne vont permettre de calculer l'écart-type de ces estimateurs, et donc, comme au chapitre 2, de proposer des intervalles de confiance pour ces estimateurs.

g) Cas particulier : le taux de sondage est le même pour toutes les strates

Les formules présentées ci-dessus sont valables quels que soient les nombres d'unités tirées par strate ; le taux de sondage n_h/N_h peut donc être variable d'une strate à une autre.

Quand on impose un taux de sondage identique pour toutes les strates, on qualifie alors le sondage de "stratifié représentatif", ou "stratifié proportionnel".

L'estimation de la moyenne vaut alors :

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{h=1}^{k}\left[N_h \frac{1}{n_h}\sum_{i_h=1}^{n_h} y_{i_h} \right] = \frac{1}{n}\sum_{h=1}^{k}\left[\sum_{i_h=1}^{n_h} y_{i_h} \right]$$

puisque $n_h/N_h = n/N$ (où n est le nombre total de questionnaires) ; c'est donc la moyenne simple calculée sur l'échantillon qui permet d'estimer la moyenne sur l'univers ; on a un sondage dit "autopondéré".

La variance de l'estimateur $\hat{\bar{Y}}$ vaut $V\left(\hat{\bar{Y}}\right) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{k}\frac{N_h}{N}S_h^2$

On montre que, dans ce cas, cette variance est liée à la variance de l'estimateur $\bar{y}$ issu du sondage aléatoire simple obtenu à partir du même nombre d'unités tirées, par la relation :

$$V(\bar{y}) = V\left(\hat{\bar{Y}}\right) + \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{k}\frac{N_h}{N}\left(\bar{Y}_h - \bar{Y}\right)^2$$

Ceci veut dire que le sondage stratifié représentatif a une variance d'estimateur toujours inférieure ou égale à celle du sondage "simple", et d'autant plus inférieure que les strates ont des moyennes différentes de la moyenne générale. On perçoit intuitivement ce résultat en se souvenant que le tirage stratifié a consisté à forcer le hasard "général" et à imposer à l'échantillon de "représenter" l'univers strate par strate, donc à concentrer les valeurs observées autour des moyennes de chaque strate.

3. Choix des strates

L'idée est de déterminer des strates les plus homogènes possibles, relativement au sujet qu'on étudie. Deux types de considérations vont conduire au choix des critères de stratification :

- disponibilité des critères dans la base de sondage ;
- pertinence des différents critères pour créer des strates homogènes, ce qui nécessite une connaissance soit intuitive, soit venant d'études réalisées antérieurement.

On prendra généralement comme critères :

- des critères relevant d'une typologie (par exemple la catégorie sociale) ;

- des critères de taille (prenant par exemple en compte le nombre de personnes du ménage) ;

souvent en les croisant ensemble.

Au niveau des unités de sondage "géographiques", par exemple les villages, on pourra stratifier selon la région, l'activité dominante des localités, le caractère sédentaire ou nomade ; on séparera souvent milieu rural et milieu urbain.

Au niveau des ménages ou des individus, on utilisera les critères qu'on pense être en corrélation avec le sujet d'étude de l'enquête : par exemple la catégorie sociale, le niveau d'instruction, la taille du ménage, le type d'habitat, etc.

Une stratification peut fort bien être très efficace pour l'étude d'un phénomène, par exemple la mortalité, et l'être très peu pour l'étude d'autres phénomènes, par exemple l'activité économique ou les mouvements migratoires. Cette situation se présente avec une acuité particulière lorsqu'un échantillon est destiné à des études à objectifs multiples, par exemple "démographie" et "agriculture".

Lorsque l'on multiplie le nombre de strates, le gain marginal d'efficacité devient rapidement faible et les résultats calculés au niveau de chaque strate ne sont guère significatifs en raison de la petite taille de l'échantillon (mais au niveau global, les résultats sont significatifs).

4. Répartition de l'échantillon entre les strates

a) Répartition représentative, répartition de Neyman

La répartition représentative a déjà été présentée, elle consiste à utiliser le même taux de sondage pour toutes les strates ; mais d'autres répartitions sont possibles.

La répartition de Neyman consiste à respecter l'égalité :

$$\frac{n_h}{N_h S_h} = \text{constante} = \frac{n}{\sum_{h=1}^{k} N_h S_h}$$

Elle utilise un taux de sondage proportionnel à la dispersion S_h de la variable Y étudiée dans chaque strate : plus une strate est hétérogène vis-à-vis de cette variable, plus on y utilise un taux de sondage important. La théorie montre que cette répartition est celle qui fournit la variance la plus faible (c'est-à-dire la meilleure précision au niveau de l'estimation globale sur l'univers) une fois les strates déterminées.

L'application de la formule ci-dessus pour calculer la répartition de Neyman suppose connues *a priori* les valeurs S_h. Ce peut être le cas à partir d'études antérieures au sondage, mais en général il n'en est pas ainsi. Lorsque le critère de stratification est la taille des unités (strates définies par des "tranches" de taille), on constate que l'écart-type est sensiblement proportionnel à la taille moyenne des unités de la strate. C'est un ordre de grandeur de cette taille moyenne (plus facile à estimer que S_h) qu'on utilise pour calculer la répartition des questionnaires entre les strates.

En pratique, on utilise la répartition de Neyman quand le phénomène qu'on étudie a une distribution très dissymétrique (par exemple, pour des sondages auprès des entreprises ou auprès d'exploitations agricoles, s'il existe à la fois de petites exploitations et quelques très grandes exploitations concentrant une partie importante de la production). Par contre, si ce phénomène a une distribution symétrique par rapport à sa moyenne, un sondage stratifié proportionnel (ou "représentatif") fournira des résultats d'une qualité suffisante.

Enfin, si les coûts d'enquête sont différents d'une strate à l'autre, on en tiendra compte dans la répartition de Neyman, qui respecter a la condition n_h / N_h proportionnel à $S_h / \sqrt{C_h}$ où (C_h est le coût unitaire d'enquête dans la strate h).

b) Exemple

Un pays est divisé en deux régions (strates) présentant les caractéristiques du tableau 2. On veut estimer la population totale à partir d'un sondage de villages au 1/50ème. L'univers est donc l'ensemble des villages, la grandeur S_h est relative à la dispersion de la population des villages à l'intérieur de chaque strate et la variable $\overline{Y_h}$ est la taille moyenne des villages pour chaque strate.

Tableau 2. Exemple de stratification : caractéristiques des strates

Strate h	Nombre de villages (N_h)	Population totale	S_h	$\overline{Y_h}$
1	3 000	956 800	100	319
2	1 000	605 000	200	605
Total	4 000	1 561 800		390

On va tirer un échantillon de 80 villages. On peut choisir :

- une répartition proportionnelle ;
- la répartition de Neyman.

Remarquons que l'échantillon de Neyman dépend du caractère que l'on veut estimer en priorité (ici, ce sera par exemple la population totale) ; c'est pour ce caractère que l'on prendra la variance en considération. En général, celle-ci ne sera pas connue *a priori* ; elle pourra être estimée à partir d'une enquête antérieure ou d'études limitées (c'est la colonne S_h du tableau précédent).

Les deux répartitions, proportionnelle et de Neyman, sont indiquées dans le tableau 3.

Tableau 3. Exemple de stratification : répartition de l'échantillon

Strate	Proportionnelle	Neyman
1	60	48
2	20	32
Total	80	80

La répartition de Neyman a réaffecté des questionnaires vers la strate 2 (par rapport à la répartition proportionnelle), où la dispersion est plus forte. On peut "raffiner" la démarche précédente en introduisant des éléments de coût : coûts d'accès aux différentes strates qui peuvent être, là aussi, variables. On est ramené à un problème d'optimisation (minimisation de la variance de l'estimateur) sous contraintes (taille de l'échantillon fixée). La résolution d'un problème de ce type est présentée au chapitre 5 (partie 2).

c) Recherche de précision au niveau de chaque strate

On se trouve devant un problème complètement différent quand on veut obtenir des renseignements significatifs pour chaque strate, par exemple si l'on veut estimer la fécondité ou la mortalité pour la population urbaine et la population rurale, ou pour différentes régions d'un pays, ou encore pour des populations sédentaires et nomades. Ici, il faudra avantager relativement les strates les moins peuplées, généralement au détriment de la précision globale.

Si l'on souhaite la même précision au niveau de chaque strate et si l'on estime que les strates présentent la même hétérogénéité pour le caractère étudié, il faudra prendre à peu près la même taille d'échantillon dans chacune.

Retour sur l'exemple précédent

On peut vouloir obtenir la même largeur de l'intervalle de confiance pour l'estimation de la taille moyenne du village pour chacune des deux strates.

Donc $$V(\bar{y}_1)=V(\bar{y}_2)$$

Soit $$\left(1-\frac{n_1}{N_1}\right)\frac{S_1^2}{n_1}=\left(1-\frac{n_2}{N_2}\right)\frac{S_2^2}{n_2}$$

En "négligeant" les taux de sondage $\frac{n_h}{N_h}$ pour simplifier :

$$\frac{S_1^2}{n_1}=\frac{S_2^2}{n_2}=\frac{S_1^2+S_2^2}{n} \quad \text{et} \quad n=80$$

D'où $n_1 = 16$ et $n_2 = 64$. Ici, pour obtenir une estimation précise sur la strate 2 (qui comporte pourtant moins de villages que la strate 1 mais pour laquelle la dispersion de la taille des villages est plus forte), on sera conduit à privilégier l'affectation des unités enquêtées vers cette strate.

d) Conclusion

Il faut accorder à la stratification un préjugé favorable, mais une stratification peut être efficace pour un caractère et pas du tout pour un autre.

L'échantillon autopondéré simplifie le dépouillement, conduit à des calculs faciles et ne réserve pas de mauvaise surprise en cas d'erreur. Mais il risque d'aboutir à des résultats peu significatifs pour les petites strates et il ne donne pas - sauf exception- la précision globale la meilleure. Pour les enquêtes démographiques, c'est cependant la méthode la plus utilisée (pour ce domaine les variances des variables ne sont pas suffisamment différenciées d'une strate à l'autre pour justifier des différences de taux de sondage).

L'échantillon de Neyman est d'une application difficile, il suppose une bonne information préalable et des erreurs dans celle-ci peuvent avoir des conséquences graves. Il désavantage les petites strates mais donne de bons résultats globaux. C'est une méthode utilisée dans le cas d'enquêtes sur la production, destinées à fournir des estimations de totaux (cas des enquêtes sur les entreprises, lorsqu'une base de sondage est disponible sous forme, par exemple, d'un registre). Enfin, si l'on augmente la précision au niveau de la strate, on risque une perte d'efficacité à l'échelon global.

Encadré 2

Un exemple de sondage stratifié

L'enquête budget – consommation du Gabon de 1993

Cette enquête s'est déroulée dans deux centres urbains du Gabon : Libreville et Port Gentil. Un dénombrement exhaustif a permis de dresser la liste des ménages de chaque centre (52 800 ménages à Libreville par exemple) en posant un nombre limité de questions.

Le sondage utilisé a été un sondage stratifié selon trois critères supposés en relation avec le niveau de vie :

- nationalité du chef de ménage (Gabonais, autre africain) ;
- type d'habitat (5 modalités de "précaire" à "luxe") ;
- statut d'occupation (propriétaire, autre).

Soit en tout 20 strates pour chaque centre urbain, avec tirage d'un échantillon (de 2 300 ménages pour Libreville par exemple), en utilisant le même taux de sondage pour chaque strate (échantillon "représentatif"). De plus, à l'intérieur de chaque strate, le fichier a été trié selon la taille du ménage avant le tirage, et un tirage "systématique"[14] a assuré une bonne représentativité de l'échantillon selon ce dernier critère.

[14] Voir chapitre 2, partie 6.b.

CHAPITRE 4

SONDAGES À PROBABILITÉS INÉGALES

1. Principe

On peut, dans certains cas, décider d'accorder à certaines unités une probabilité plus forte d'être sélectionnées qu'à d'autres.

Par exemple :

- pour des enquêtes auprès des entreprises, on peut tirer les unités avec une probabilité proportionnelle à leur nombre de salariés ;
- dans certaines enquêtes, on tire des "aires" (ou "segments") à partir d'une grille de points placée sur une carte où on a découpé le territoire ; les aires sélectionnées sont celles sur lesquelles se trouve un point, elles ont donc une probabilité d'être tirée proportionnelle à leur superficie ;
- enfin, le sondage à probabilités inégales est souvent utilisé au premier degré d'un tirage à plusieurs degrés (chapitre 5) : tirage de villages, de communes ou d'autres unités aréolaires, avec probabilité proportionnelle à leur population (puis tirage de ménages ou d'individus au deuxième degré).

On voit sur les exemples précédents que la probabilité de tirage d'une unité est, en général, proportionnelle à une mesure de taille, l'idée étant que plus une unité est "grande", plus elle apporte de l'information et plus il est important de la sélectionner.

2. Formules d'estimation dans le cas avec remise

Bien qu'on se soit jusqu'à présent placé, dans ce manuel, dans le cadre de tirages sans remise, on commencera ici par parler de tirages à probabilités inégales

avec remise, pour des raisons de complexité de la formalisation mathématique du cas sans remise. Celui-ci sera abordé dans la quatrième partie de ce chapitre.

Chaque unité α de l'univers a la probabilité A_α[15] d'être tirée à chacun des tirages et on tire un échantillon de taille n.

On a $\sum_{\alpha=1}^{N} A_\alpha = 1$ (donc chaque A_α est inférieur à 1 et souvent de valeur très faible).

a) Estimation d'un total

L'estimateur du total[16] de la variable Y (sur l'univers) proposé à partir de l'échantillon tiré est :

$$\hat{T}(Y) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{A_i} \qquad \textbf{(1)}$$

où y_i est la valeur de la variable Y pour l'unité sélectionnée au i$^{\text{ème}}$ tirage et A_i sa probabilité d'être sélectionnée à chaque tirage : on tient donc compte des probabilités de tirage différentes pour produire l'estimation du total.

Cet estimateur est sans biais : $E\left(\hat{T}(Y)\right) = \sum_{\alpha=1}^{N} Y_\alpha$

Sa variance vaut : $V\left(\hat{T}(Y)\right) = \frac{1}{n}\sum_{\alpha=1}^{N} A_\alpha \left[\frac{Y_\alpha}{A_\alpha} - \left(\sum_{\alpha=1}^{N} Y_\alpha\right)\right]^2$

Elle peut être estimée sans biais à partir de l'échantillon par :

$$\hat{V}\left(\hat{T}(Y)\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{A_i} - \hat{T}(Y)\right)^2$$

---

[15] Souvent A_α est proportionnel à une mesure de la taille de l'unité : si Z_α est sa taille, on prendra $A_\alpha = Z_\alpha \Big/ \left(\sum_{\alpha=1}^{N} Z_\alpha\right)$

[16] Cette formule peut, à première vue, paraître étrange pour l'estimation d'un total et sembler plus appropriée pour estimer une moyenne ; on doit cependant se rappeler que les A_i au dénominateur ont des valeurs très faibles.

Remarque : dans la formule relative à $V\left(\hat{T}(Y)\right)$, on voit que $V\left(\hat{T}(Y)\right)=0$ si $A_\alpha = \frac{Y_\alpha}{\sum_{\alpha=1}^{N} Y_\alpha}$ pour toutes les unités statistiques de l'univers.

Ceci veut dire que, si l'on utilise ce jeu de probabilités A_α, le sondage est "parfait" (variance nulle). Mais ceci est irréalisable en pratique, car nécessitant de connaître à l'avance le résultat recherché, puisque demandant de connaître l'ensemble des Y_α ; cependant, l'utilisation d'un critère variant de manière approximativement proportionnelle à Y pour établir les probabilités de tirage (par exemple les superficies des exploitations agricoles pour estimer leur production) pourra permettre de "s'approcher" de cette situation, et donc d'avoir une variance assez réduite. C'est la raison pour laquelle un critère de taille est souvent utilisé pour des estimations relatives à des totaux (production, effectifs).

b) Estimation d'une moyenne, d'un ratio

Pour estimer la moyenne $\overline{Y}$, on utilise l'estimateur $\frac{\hat{T}(Y)}{N}$.

Sa variance est : $V\left(\frac{\hat{T}(Y)}{N}\right) = \frac{1}{N^2} V\left(\hat{T}(Y)\right)$.

Un ratio est estimé comme le rapport de l'estimation de deux masses.

3. Méthodes de tirage

a) Méthode des chiffres cumulés

Supposons que l'on ait une liste de 207 villages avec une estimation de leur population. On veut enquêter 21 villages, $n = 21$. On calcule d'abord la population cumulée correspondant à chaque village (tableau 4). Pour le dernier village, 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'un village est tiré si un habitant de ce village a un billet gagnant).

Tableau 4. Exemple de tirage à probabilités inégales : chiffres cumulés

Village	Population par village	Population cumulée
1	531	531
2	177	708
3	348	1 056
4	235	1 291
5	290	1 581
6	124	1 705
----	-----	-----
205	425	58 254
206	219	58 473
207	153	58 626

Supposons, par exemple, que l'on ait tiré entre autres 937 et 58 302 ; ces chiffres désignent respectivement les villages n° 3 et n° 206. Supposons que l'on tire ensuite 727 ; le village n° 3 est à nouveau sélectionné.

On peut améliorer la procédure en rangeant par taille les unités, et en procédant à un tirage systématique (présenté au chapitre 2) dans les chiffres cumulés. On obtient ainsi une répartition "satisfaisante" de l'échantillon par rapport au critère de tri choisi.

b) Méthodes aréolaires utilisant des grilles de points

Ces méthodes consistent à placer une grille de points sur une carte préalablement découpée, ou sur une photographie aérienne (figure 7).

Les "zones élémentaires" de la carte (parfois appelées segments) ou les parcelles sur la photographie sont sélectionnées si elles contiennent un point de la grille, avec une probabilité proportionnelle à leur surface (en toute rigueur, ce mode de tirage est plus complexe puisque s'apparentant à un tirage systématique ; pour le formaliser de manière complète, il faudrait établir une modélisation de la "structure" du paysage sur lequel on procède au sondage).

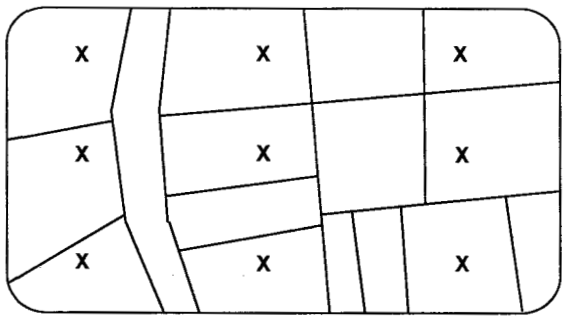


Figure 7. Méthode de tirage aréolaire à partir d'une grille de points

4. Aperçu sur le sondage à probabilités inégales sans remise

Le modèle qui a été appliqué précédemment pour produire un estimateur est beaucoup plus difficile à utiliser : en effet, les probabilités de tirage se déforment au fur et à mesure qu'on réalise les tirages.

Au premier tirage $A_i^1 = A_i$;

Au deuxième tirage $A_j^2 = \dfrac{A_j^1}{1 - A_i^1}$ sachant que c'est i qui a été tiré au premier tirage ; etc.

L'estimateur de Horvitz-Thompson

On fait donc appel à une autre approche, que nous présenterons rapidement : celle de Horvitz-Thompson. Le point de départ de cette approche développée pour les tirages sans remise est la probabilité d'inclusion :

Π_i probabilité que i appartienne à l'échantillon,
Π_{ij} probabilité que i et j) soient simultanément dans l'échantillon.

Remarquons que si l'échantillon s est de taille fixe n, alors :

$$\sum_{\alpha=1}^{N} \Pi_\alpha = n$$

L'estimateur de Horvitz-Thompson du total $\hat{T}(Y)$ est :

$$\hat{T}(Y) = \sum_{i \in s} \frac{y_i}{\Pi_i} \qquad \textbf{(2)}$$

Si l'échantillon est de taille fixe, alors :

$$V\left(\hat{T}(Y)\right) = \frac{1}{2} \sum_{\substack{\alpha \neq \beta \\ \alpha=1,\ldots,N \\ \beta=1,\ldots,N}} \left(\Pi_\alpha \Pi_\beta - \Pi_{\alpha\beta}\right)\left(\frac{Y_\alpha}{\Pi_\alpha} - \frac{Y_\beta}{\Pi_\beta}\right)^2$$

La variance de l'estimateur de Horvitz-Thompson peut être estimée par :

$$\hat{V}\left(\hat{T}(Y)\right) = \frac{1}{2} \sum_{\substack{i \neq j \\ i \in s \\ j \in s}} \frac{\left(\Pi_i \Pi_j - \Pi_{ij}\right)}{\Pi_{ij}}\left(\frac{y_i}{\Pi_i} - \frac{y_j}{\Pi_j}\right)^2$$

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 (Ardilly, 1994, chapitre II.4.3.).

Alors on calcule les Π_{ij} (ou on les détermine de manière approximative car, dans certains cas, le calcul rigoureux est impossible) et on peut ainsi calculer la précision (par la variance) de l'estimateur de Horvitz-Thompson (qui, lui, ne fait appel qu'aux Π_i). Certains auteurs ont, par ailleurs, proposé des formules d'approximation de la variance de l'estimateur de Horvitz-Thompson ne faisant intervenir que les Π_i.

Cette approche est une approche générale, pas seulement limitée aux sondages à probabilités inégales ; elle est présentée dans ce chapitre car étant la seule utilisable quand on tire à probabilités inégales sans remise.

CHAPITRE 5

SONDAGES À PLUSIEURS DEGRÉS

1. Principe, notations

a) Principe

On utilise une succession de regroupements des unités statistiques pour tirer l'échantillon. Par exemple (figure 8), on tire un échantillon de villages (unités primaires), puis on tire, parmi les villages tirés, un échantillon de ménages (unités secondaires).

On a dans ce cas un tirage à deux degrés (villages puis ménages). On peut généraliser à trois degrés, quatre... À chacun des degrés, les méthodes présentées aux chapitres précédents peuvent être utilisées (par exemple tirage proportionnel à la taille au premier degré, donc à probabilités inégales, tirage aléatoire simple au deuxième degré).

Il faut dire un mot sur l'utilisation de l'appellation "sondage en grappes". Celle-ci doit être réservée au cas particulier du sondage à plusieurs degrés (souvent deux degrés) où l'ensemble des unités au dernier degré de tirage est enquêté : ce serait, dans l'exemple présenté ci-dessus, l'ensemble des ménages des villages sélectionnés (soit une "grappe" de ménages) qui serait enquêté. C'est dans ce sens que sera utilisée l'expression "sondage en grappes" dans ce manuel (bien que, dans certains ouvrages, elle soit utilisée pour parler d'un sondage à deux degrés, de façon générale).

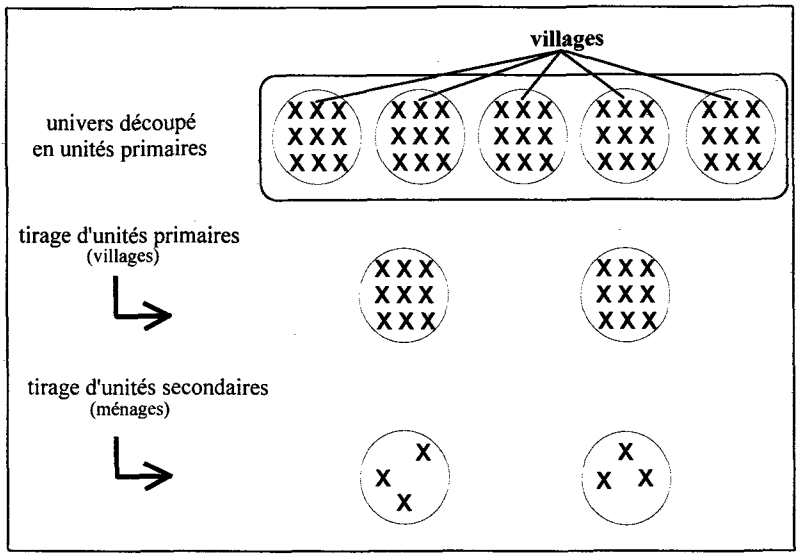


Figure 8. Exemple de tirage à deux degrés

b) Justification, caractéristiques

Prenons un exemple : on veut étudier 2 000 ménages dans un pays qui en compte environ 500 000 répartis dans 6 000 villages. On dispose seulement d'une liste des villages avec une estimation de leur population. Élaborer une liste des ménages au niveau national en visitant chaque village serait une tâche gigantesque. En outre, les ménages de l'échantillon seraient géographiquement extrêmement dispersés, d'où un temps énorme perdu en déplacements. Le coût de l'opération serait prohibitif.

Le sondage à plusieurs degrés permet donc de résoudre les deux problèmes suivants :

- en l'absence d'une base de sondage, on peut se contenter d'un travail partiel d'établissement de cette base de sondage : seule la connaissance exhaustive des unités primaires est nécessaire ; on peut se limiter à recenser, dans l'exemple précédent, les ménages des villages tirés au premier degré ;
- globalement, on va réaliser des économies de temps et 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 (en nombre d'unités statistiques au dernier degré de tirage). Ceci est dû aux "effets de grappe"[17].

Les unités statistiques regroupées dans une même unité primaire (ou dans une même unité secondaire si on a trois 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 peut conduire à une certaine "redondance" de l'information sur ces unités et un certain "manque de représentativité" de l'ensemble. On peut établir que la majeure partie de la variance des estimateurs dans le cas de tirages à plusieurs degrés provient souvent du premier degré de tirage. L'effet de grappe est abordé de manière plus approfondie dans la cinquième partie de ce chapitre.

c) Notations

Dans ce chapitre, on se placera essentiellement dans le cas du sondage à deux degrés et on utilisera les notations suivantes :

- unités primaires : M dans l'univers ($\alpha = 1, ..., M$)
  m tirées dans l'échantillon ($i = 1,..., m$)
- unités secondaires : N_α dans l'unité primaire α ($\beta = 1, ..., N_\alpha$)
  n_i dans l'échantillon pour l'unité primaire i ($j = 1, ..., n_i$)

- $T_\alpha(Y)$ total de Y sur l'unité primaire α

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

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

$$S_1^2 = \frac{1}{M-1}\sum_{\alpha=1}^{M}\left(T_\alpha(Y) - \overline{T}\right)^2$$

$$\text{où } \overline{T} = \frac{1}{M}\sum_{\alpha=1}^{M} T_\alpha(Y)$$

- $T(Y)$ total de Y sur l'univers : $T(Y) = \sum_{\alpha=1}^{M} T_\alpha(Y)$

---

[17] Cette expression est sans doute à l'origine de l'utilisation de l'appellation "sondages en grappes" pour "sondage à plusieurs degrés".

2. Tirage des unités primaires à probabilités égales (tirage à deux degrés)

On se placera dans le cas d'un tirage sans remise au premier degré, qui est *a priori* préférable pour la précision.

a) Estimation du total de Y

La formule :

$$\hat{T}(Y) = \frac{M}{m} \sum_{i=1}^{m} \hat{T}_i(Y) \qquad \textbf{(1)}$$

estime le total $T(Y)$ où $\hat{T}_i(Y)$ est l'estimateur du total $T_i(Y)$ à partir du plan de sondage choisi au deuxième degré de tirage. Cet estimateur est sans biais. On retrouve dans cette formule l'estimation du total aux deux degrés de tirage.

Par exemple, si au deuxième degré on a tiré de façon aléatoire simple, la formule **(4)** du chapitre 2 donne :

$$\hat{T}_i(Y) = \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Cas particulier : sondage autopondéré

Si on tire à probabilités égales les unités primaires et si, de plus, le taux de sondage est le même pour le deuxième degré de tirage (toujours à probabilités égales) à l'intérieur de toutes les unités primaires tirées :

alors, puisque

$$\hat{T}(Y) = \frac{M}{m} \sum_{i=1}^{m} \underbrace{\frac{N_i}{n_i}}_{\text{constante}} \sum_{j=1}^{n_i} y_{ij}$$

la pondération utilisée est la même pour toutes les unités statistiques de l'échantillon (en l'occurrence les unités secondaires) ; le sondage est dit autopondéré. Dans ce

cas, la moyenne simple calculée sur l'ensemble des unités tirées est utilisée comme estimateur de la moyenne sur l'univers (ce qui n'est pas le cas si l'on tire au deuxième degré avec des taux de sondage différents selon les unités primaires).

b) Variance de l'estimateur du total de Y

$$V\left(\hat{T}(Y)\right) = \frac{M^2}{m}\left(1-\frac{m}{M}\right)S_1^2 + \frac{M}{m}\sum_{\alpha=1}^{M} Z_\alpha \qquad \textbf{(2)}$$

où Z_α est la variance de l'estimateur $\hat{T}_\alpha(Y)$ du total $T_\alpha(Y)$ dans l'unité primaire α consécutive au plan de sondage choisi au deuxième degré.
Par exemple, si au deuxième degré, on a tiré, dans chaque unité primaire α, n_α unités à probabilités égales sans remise, les formules du chapitre 2 permettent le calcul de Z_α :

$$Z_\alpha = \frac{N_\alpha{}^2}{n_\alpha}\left(1-\frac{n_\alpha}{N_\alpha}\right)\frac{1}{N_\alpha - 1}\sum_{\beta=1}^{N_\alpha}\left(Y_{\alpha\beta} - \overline{Y}_\alpha\right)^2$$

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

c) Estimation de la variance de l'estimateur du total de Y

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

$$\hat{V}\left(\hat{T}(Y)\right) = \frac{M^2}{m}\left(1-\frac{m}{M}\right)s_1^2 + \frac{M}{m}\sum_{i=1}^{m}\hat{Z}_i$$

$$\text{où } s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{T}i(Y) - \frac{\hat{T}(Y)}{m}\right)^2$$

et $\hat{Z}_i$ est l'estimateur de la variance de l'estimation $\hat{T}_i(Y)$ selon le plan de sondage au deuxième degré.

Par exemple, si au deuxième degré de tirage on a tiré à probabilités égales sans remise,

$$\hat{Z}_i = \frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 \quad \text{où } \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

d) Remarques

Dans la formule **(2)** ci-dessus, le premier terme est en général le plus important. Les deux termes de cette formule sont d'ailleurs relatifs aux deux degrés de tirage et permettent de décomposer la variance pour observer la part de chacun de ces deux degrés.

Si on augmente m dans cette formule, on voit que les deux 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 l'estimateur de la variance de l'estimateur du total $\left(\hat{V}\left(\hat{T}(Y)\right)\right)$, on a également deux termes qui semblent correspondre à la décomposition selon les deux degrés de tirage : en fait ce n'est pas le cas, contrairement à ce qui a pu être dit précédemment pour $V\left(\hat{T}(Y)\right)$. Dans le cas d'un sondage aléatoire simple au deuxième degré, par exemple, le premier terme de la formule **(2)** serait estimé par : $\frac{M^2}{m}\left(1-\frac{m}{M}\right)\left[s_1^2-\frac{1}{m}\sum_{i=1}^{m}\hat{Z}_i\right]$ et le second terme par : $\frac{M^2}{m^2}\sum_{i=1}^{m}\hat{Z}_i$.

e) Estimation d'une moyenne, d'un ratio

Pour estimer la moyenne de Y par unité statistique sur l'univers à partir du total, on ne connaît pas en général le nombre total d'unités statistiques de celui-ci (en général on n'a pas de base de sondage au niveau des unités secondaires mais plutôt seulement la liste des unités primaires). On est donc obligé d'estimer ce nombre total à partir de l'échantillon d'unités primaires, soit $\hat{N}$; on estime la moyenne par $\hat{T}(Y)/\hat{N}$.

Un ratio sera estimé comme le rapport de deux masses estimées.

f) Application pratique au cas d'une enquête agricole

Pour écrire la formule donnant $V\left(\hat{T}(Y)\right)$, dans le cas où le tirage au deuxième degré se fait de façon aléatoire simple, on procède aux approximations suivantes :

- on néglige les taux de sondage aux deux degrés et on "assimile" les coefficients $\left(1-\frac{m}{M}\right)$ et $\left(1-\frac{n_\alpha}{N_\alpha}\right)$ à 1 ;
- le nombre d'unités secondaires tirées par unité primaire est supposé à peu près constant et égal à n_0.

Alors $V\left(\hat{T}(Y)\right)$ s'écrit sous la forme : $V\left(\hat{T}(Y)\right)=\frac{A}{m}+\frac{B}{mn_0}$, où A et B sont des constantes.

Cette formule va permettre d'étudier la meilleure répartition possible de l'échantillon.

Par exemple, pour une enquête agricole par sondage réalisée au Mali en 1956-1957 (tirage à deux degrés : villages puis rizières), on est arrivé à une formule de ce type pour la variance de l'estimation du rendement moyen (qui n'est pas un total mais, en fait, un ratio ; les calculs après approximations mènent cependant à une formule du même type que celle présentée pour la variance de l'estimation d'un total) :

$$V\left(\hat{R}\right)=\frac{110\,000}{m}+\frac{253\,000}{mn_0}$$

$\hat{R}$, le rendement moyen du riz à l'hectare valait 1 027 kg.

Si l'on cherche à obtenir une précision de 5 % (c'est-à-dire un coefficient de variation[18] de 5 %), l'écart type vaut 51,35, la variance 2 636,8.

On a donc $m=\frac{110\,000n_0+253\,000}{2\,636{,}8n_0}$.

En fonction du nombre de rizières observées dans chaque village, on obtient le nombre d'unités primaires m (villages) à enquêter pour assurer une précision de 5 % (tableau 5).

---

[18] $c.v.=\frac{\sigma\left(\hat{R}\right)}{\hat{R}}=0{,}05$

Tableau 5. Exemple d'une enquête agricole : répartition de l'échantillon entre les deux degrés de tirage

n_0	m	n_0m
2	90	180
5	61	305
8	54	432
10	51	510

On observe que, lorsque le nombre d'unités secondaires enquêtées par unité primaire augmente, le nombre d'unités primaires nécessaire pour la précision voulue diminue, mais le nombre total d'unités enquêtées (colonne n_0m) augmente : on aura donc plutôt intérêt à tirer beaucoup d'unités primaires et à y mettre relativement peu d'unités secondaires à enquêter.

Introduction d'un modèle de coût

Les considérations précédentes peuvent être modulées par l'introduction d'éléments de coût ; le coût de l'enquête peut être décomposé en $C = C_0 + C_1m + C_2mn_0$, où :

- C_0 est un coût fixe ;
- C_1 est le coût "d'accès" à l'unité primaire (le village) ;
- C_2 est le "coût de l'unité secondaire enquêtée".

C_1 représente le coût de déplacement et d'établissement de la liste des unités secondaires, pour une unité primaire donnée, le coût est important et doit être isolé.

On doit donc minimiser $\frac{A}{m} + \frac{B}{mn_0}$ sous la contrainte : $C = C_0 + C_1m + C_2\, mn_0$.

On résout ce problème à l'aide des multiplicateurs de Lagrange :

$$L = \frac{A}{m} + \frac{B}{mn_0} + \lambda\left(C_0 + C_1m + C_2mn_0\right)$$

$$\frac{dL}{dm} = 0 \qquad -\frac{A}{m^2} - \frac{B}{m^2n_0} + \lambda C_1 + \lambda C_2n_0 = 0$$

$$\frac{dL}{dn_0} = 0 \qquad -\frac{B}{mn_0^{\,2}} + \lambda C_2m = 0$$

d'où $\lambda = \frac{B}{C_2m^2n_0^{\,2}}$ et $n_0 = \sqrt{\frac{BC_1}{AC_2}}$;

on en déduit m en fonction de la contrainte de coût.

Par exemple, pour l'enquête sur le riz au Mali, les coûts C_1 et C_2 avaient été chiffrés à :

$C_1 = 3$ (un village nécessite 3 jours de travail : accès au village, dénombrement des rizières) ;
$C_2 = 0{,}5$ (une rizière nécessite une demi-journée d'enquête).

Alors $n_0 = 3{,}7$. On enquête donc 4 rizières par village (ce résultat est à rapprocher de la pratique d'un certain nombre d'enquêtes agricoles à plusieurs degrés dans les pays africains où on tire entre cinq et dix exploitations agricoles par village sélectionné ; voir par exemple Brilleau, 1993).

3. Tirage des unités primaires à probabilités inégales (tirage à deux degrés)

On se placera ici, pour simplifier, dans le cadre du sondage avec remise. A_α est la probabilité de l'unité primaire α d'être tirée à chacun[19] des tirages d'unités primaires.

a) Estimateur du total de Y

$$\hat{T}(Y) = \frac{1}{m}\sum_{i=1}^{m}\frac{\hat{T}_i(Y)}{A_i} \qquad (3)$$

$\hat{T}(Y)$ est un estimateur sans biais du total de Y sur l'univers ; on voit qu'on passe par l'estimateur $\hat{T}_i(Y)$ du total de Y pour l'unité primaire i, puis qu'on utilise la formule **(1)** du chapitre 4. $\hat{T}_i(Y)$ tient compte de la méthode de sondage utilisée au deuxième degré de tirage.

[19] Remarquons qu'il s'agit de la probabilité d'être tirée à chaque tirage et non de la probabilité ("globale") d'être dans l'échantillon.

b) Variance de l'estimateur du total, estimateur de cette variance

La variance de $\hat{T}(Y)$ vaut :

$$V\left(\hat{T}(Y)\right) = \frac{1}{m}\sum_{\alpha=1}^{M} A_\alpha \left(\frac{T_\alpha(Y)}{A_\alpha} - T(Y)\right)^2 + \frac{1}{m}\sum_{\alpha=1}^{M}\frac{Z_\alpha}{A_\alpha} \qquad \textbf{(4)}$$

où Z_α est la variance de l'estimateur de $T_\alpha(Y)$ tenant compte du plan de sondage au deuxième degré.

L'estimateur de cette variance à partir de l'échantillon est :

$$\hat{V}\left(\hat{T}(Y)\right) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{\hat{T}_i(Y)}{A_i} - \hat{T}(Y)\right)^2$$

c) Cas particulier important

On n'a pas abordé pour l'instant le problème du choix des A_α. Souvent on décide de tirer les unités avec une probabilité proportionnelle à leur taille :

$$A_\alpha = \frac{N_\alpha}{N} \text{ (où } N = \sum_{\alpha=1}^{M} N_\alpha \text{)}$$

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 :

$$\hat{T}(Y) = \frac{1}{m}\sum_{i=1}^{m}\frac{N}{N_i}\left(\frac{N_i}{n_0}\sum_{j=1}^{n_0} y_{ij}\right) = \frac{N}{mn_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} y_{ij} \text{ avec } \forall i,\, n_i = n_0$$

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

En pratique on se trouve rarement exactement dans cette 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 formule **(3)**. 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 :

$$\frac{N}{mN_i}\frac{N'_i}{n_0}$$

d) Estimation d'une moyenne, d'un ratio

Pour estimer une moyenne par unité secondaire sur l'univers, il faudra souvent estimer le nombre total d'unités secondaires qui est inconnu. Un ratio sera estimé comme le rapport de deux masses estimées.

e) Retour sur le choix avec remise - sans remise

Dans cette partie consacrée au tirage à deux degrés avec sélection des unités primaires à probabilités inégales, on s'est placé dans le cas où celles-ci étaient tirées avec remise ; ceci, en fait, pour des raisons de difficultés à appréhender correctement, sur le plan de la formalisation, le cas sans remise.

En pratique, on procèdera très souvent à des tirages des unités primaires à probabilités inégales sans remise : on utilisera les formules précédentes ("comme si" on avait tiré avec remise), en sachant que les estimations de précision obtenues (variance d'estimateur) majoreront la véritable précision.

On réalisera parfois le tirage en rangeant les unités selon un certain critère (par exemple, la taille de la localité) et en procédant à un tirage systématique dans le cumul des tailles (chapitre 4).

4. Sondage en grappes

a) Principe

C'est le cas particulier du sondage à plusieurs degrés où l'ensemble des unités du "dernier degré" est enquêté : par exemple on tire un échantillon de villages à l'intérieur desquels on va enquêter tous les ménages, ou tous les individus.

Là encore, l'intérêt de ce type de sondage réside en des coûts de déplacements moindres (si on utilise des unités primaires correspondant à des regroupements géographiques) et en la non-obligation de disposer d'une base de sondage complète.

b) Estimation d'un total dans le cas d'un tirage des grappes à probabilités égales

Si $T_i(Y)$ est le total de Y observé sur la grappe i (ou unité primaire) sans erreur aléatoire (puisqu'on a enquêté exhaustivement la grappe) :

$$\hat{T}(Y) = \frac{M}{m}\sum_{i=1}^{m} T_i(Y) \quad \textbf{(5)}$$ est l'estimateur du total de Y sur l'univers.

On est donc ramené à l'estimateur classique proposé au chapitre 2. Sa variance peut être estimée à partir de l'échantillon par :

$$\hat{V}\left(\hat{T}(Y)\right) = M^2\left(1-\frac{m}{M}\right)\frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}\left(T_i(Y)-\overline{T}(Y)\right)^2 \text{ où } \overline{T}(Y) = \frac{1}{m}\sum_{i=1}^{m} T_i(Y)$$

c) Estimation d'une moyenne (par unité statistique élémentaire, par exemple unité secondaire) dans le cas d'un tirage des grappes à probabilités égales

Si on connaît le nombre total N d'unités statistiques sur l'univers, on estime la moyenne par $\frac{1}{N}\hat{T}(Y)$.

Le problème est plus délicat quand, et c'est fréquemment le cas, on ne connaît pas N.

On est conduit à estimer N par : $\hat{N} = \frac{M}{m}\sum_{i=1}^{m} N_i$. L'estimateur de la moyenne est $\frac{\hat{T}(Y)}{\hat{N}}$. Sa variance, plus complexe à calculer, est celle d'un ratio (chapitre 2).

d) Estimation d'un total dans le cas d'un tirage des grappes à probabilités inégales

Si $T_i(Y)$ est le total de Y observé sur la grappe i,

$$\hat{T}(Y) = \frac{1}{m}\sum_{i=1}^{m}\frac{T_i(Y)}{A_i} \quad \textbf{(6)}$$

est l'estimateur du total de Y sur l'univers (A_i est la probabilité de la grappe i d'être tirée à chaque tirage).

5. L'effet de grappe

a) Principe

Comme on l'a vu précédemment, le fait de tirer à deux degrés, ou de tirer des grappes, induit souvent une perte de précision (par rapport à un sondage simple à partir du même nombre d'unités enquêtées) due au fait que les unités situées à l'intérieur d'une même unité primaire ont souvent tendance à se ressembler. On se placera ici dans le cas de tirages à deux degrés.

b) Le coefficient de corrélation intragrappe

Il est défini comme :

$$\delta = \frac{\sum_{\alpha=1}^{M} \sum_{\beta=1}^{N_\alpha} \sum_{\substack{\gamma=1 \\ \gamma \neq \beta}}^{N_\alpha} \left(Y_{\alpha\beta} - \overline{Y}\right)\left(Y_{\alpha\gamma} - \overline{Y}\right)}{\sum_{\alpha=1}^{M} \sum_{\beta=1}^{N_\alpha} \left(Y_{\alpha\beta} - \overline{Y}\right)^2} * \frac{1}{\overline{N} - 1}$$

où $\overline{Y}$ est la moyenne de Y par unité secondaire (donc calculée sur l'ensemble des unités statistiques - ici, unités secondaires - de l'univers),

$\overline{N}$ est la taille moyenne des unités primaires (donc, le nombre moyen d'unités secondaires par unité primaire).

Ce coefficient, qui ressemble à un coefficient de corrélation linéaire (mais ne fait intervenir qu'une variable), peut être positif (cas général quand il y a similitude des unités secondaires à l'intérieur des unités primaires) ou, parfois, négatif. Il est appelé "*RHO*" dans certains manuels.

c) Conséquences sur la précision du sondage

On montre que, si l'on procède à un tirage à deux degrés ou en grappes sans stratification ni tirage à probabilités inégales des unités primaires, si toutes les unités

primaires sont de même taille $\overline{N}$ et que la taille de l'échantillon d'unités secondaires par unité primaire est constante et égale à $\overline{n}$:

$$V\left(\hat{T}(Y)\right) = \left(1 + \delta(\overline{n} - 1)\right) V_{sas}\left(\hat{T}(Y)\right)$$

où $V\left(\hat{T}(Y)\right)$ est la variance à partir de l'estimation du total de Y du plan de sondage à deux degrés ou en grappes, et $V_{sas}\left(\hat{T}(Y)\right)$ la variance de l'estimation du même total à partir d'un plan de sondage aléatoire simple (à probabilités égales).

La grandeur *DEFF*[20] qui est le rapport des deux variances d'estimation permet d'estimer la perte de précision obtenue lors du passage d'un plan de sondage à l'autre : on l'appelle "effet de sondage" (en anglais "*Design Effect*", d'où l'appellation *DEFF*) :

$$DEFF = 1 + \delta(\overline{n} - 1)$$

Cet effet de sondage est en fait une notion plus générale qui mesure le rapport de la variance d'un sondage pratiqué à la variance du sondage aléatoire simple utilisant la même taille d'échantillon. Si le sondage effectivement appliqué avait, en plus des deux degrés, utilisé une stratification des unités primaires ou un tirage à probabilités inégales de celles-ci, on aurait un effet de sondage plus complexe. Dans ce cas, on peut introduire une grandeur, que certains auteurs appellent *ROH* (par analogie à *RHO*) qui est définie par l'équation :

$$DEFF = 1 + ROH(\overline{n} - 1).$$

C'est le calcul effectif de la variance du sondage pratiqué et de celle du sondage aléatoire simple qui permet d'obtenir des valeurs de *DEFF* et, par la suite, de *ROH* pour certains paramètres (dans ce cas, *ROH* n'est plus le coefficient de corrélation intragrappe puisqu'il prend en compte, par exemple, la stratification des unités primaires s'il y en a une). Les valeurs de *DEFF* et de *ROH* sont donc alors obtenues par une démarche "expérimentale" plus que théorique.

d) Valeurs numériques de δ, utilisation de ces valeurs

L'expérience montre que δ est souvent compris entre 0 et 0,2 pour un certain nombre de variables.

Si $\delta = 0$, la caractéristique étudiée est répartie aléatoirement entre les unités primaires et le sondage aléatoire simple et le sondage à deux degrés sont équivalents

---

[20] On trouvera parfois dans la littérature anglo-saxonne la notion $D^2 eff$ plutôt que *DEFF* pour le rapport des deux variances d'estimation.

en précision. Quand δ augmente (si par exemple δ est proche de 1, on a une variabilité essentiellement entre les grappes et quasi nulle à l'intérieur des grappes), le plan de sondage à plusieurs degrés va voir sa précision se dégrader.

Dans certains cas, on peut avoir une valeur négative de δ : par exemple si l'on veut estimer la proportion d'hommes à partir d'un sondage d'individus en grappes de ménages, chaque ménage étant systématiquement composé d'un père, d'une mère, d'un garçon et d'une fille (mais ceci est un "cas d'école" et en pratique on rencontre très rarement des δ négatifs), on voit que le sondage en grappes sera nettement plus précis qu'un sondage aléatoire simple d'individus.

En pratique, comment utilise-t-on ce coefficient ?

On connaît des ordres de grandeur de δ pour certains types de variables et pour les enquêtes "classiques" menées dans les pays en développement (à partir de données anciennes, citées par exemple dans différentes sources mentionnées en bibliographie) :

$\delta \approx 0{,}002$ pour le taux de natalité ;
$\delta \approx 0{,}003$ pour le taux de mortalité ;
$\delta \approx 0{,}02$ pour le taux de mortalité infantile ;
$\delta \approx 0{,}05$ pour les questions concernant la contraception ;
$\delta \approx 0{,}1$ pour le taux d'activité masculine.

Ces ordres de grandeur seront à confirmer bien entendu en fonction du contexte de l'enquête qu'on va réaliser : l'effet de grappe est par exemple, pour certaines variables comme l'usage de la contraception, plus marqué en milieu rural qu'en milieu urbain..

Ces valeurs de δ sont beaucoup plus faibles pour les variables ayant trait à la mortalité et la fécondité que pour celles relatives à l'emploi, aux migrations ou aux phénomènes socio-économiques. Les sondages en grappes ou à deux degrés seront donc plus appropriés au premier type de variable.

On peut, à partir de différentes tailles $\bar{n}$ (nombre d'unités secondaires enquêtées par unité primaire tirée) et différentes valeurs de δ dresser un tableau donnant les valeurs correspondantes de *DEFF* (tableau 6).

Tableau 6. Valeurs de *DEFF* pour différents paramètres

δ	$\bar{n}$		
	100	300	500
0,002	1,2	1,6	2,0
0,003	1,3	1,9	2,5
0,05	6,0	16,0	26,0

On suppose alors qu'il existe une "portabilité" de l'effet de sondage, c'est-à-dire que la valeur attendue de *DEFF* (obtenue à partir de la formule utilisant les valeurs de $\overline{n}$ et δ, qui s'applique "bien" quand les tailles des unités primaires N_i sont voisines ; ceci incite à éviter d'avoir des valeurs de ces tailles trop dispersées) pour l'enquête sera du même ordre de grandeur que celle mesurée dans une enquête ancienne réalisée dans les mêmes conditions. Dire que *DEFF* vaut 2 revient à dire que, pour obtenir la même précision avec un échantillon tiré de façon aléatoire simple (à probabilités égales sans différents degrés de tirages), il faudra deux fois plus d'unités enquêtées pour le sondage à deux degrés ou en grappes.

Les valeurs du tableau précédent peuvent être combinées aux valeurs du tableau 1 (valeurs obtenues à partir des formules du sondage aléatoire simple) pour avoir une idée de la précision attendue : par exemple, pour un taux de natalité de l'ordre de 45 ‰ , avec un échantillon aléatoire simple de 25 000 questionnaires, on a un écart type de 1,3 ‰. Si on procède à un tirage à deux degrés avec enquête de 300 individus par unité primaire (et toujours 25 000 questionnaires au total), et si on utilise la valeur $\delta = 0{,}002$, la variance sera environ multipliée par 1,6 ; l'écart type sera donc multiplié par 1,26 et deviendra 1,6 ‰.

Un retour sur les valeurs numériques relatives aux effets de grappe citées dans certains articles...

Ces valeurs doivent bien entendu être considérées comme des ordres de grandeur, plus que comme des valeurs "portables telles quelles" à d'autres contextes, d'autant plus qu'elles sont elles-mêmes souvent l'objet d'estimations à partir d'un échantillon.

Un point important à signaler et relatif aux enquêtes démographiques concerne le fait que, pour ces dernières, la valeur de δ (ou *ROH*) calculée à partir de la formule :

$DEFF = 1 + \delta(\overline{n} - 1)$ (*DEFF* étant elle-même estimée à partir de l'échantillon) utilise des valeurs de $\overline{n}$ qui peuvent être exprimées en nombre de ménages parfois, et en nombre d'individus dans d'autres cas.

Le choix de l'unité utilisée pour la valeur de $\overline{n}$ conditionne donc fortement la valeur obtenue pour δ. Ceci explique que, sur des domaines apparemment comparables, on ait pu trouver dans différents articles des valeurs de δ variant parfois dans un rapport de 1 à 5, puisqu'établies pour des échantillons d'individus d'une part, de ménages d'autre part. Les ordres de grandeur cités dans ce paragraphe sont, eux, relatifs à des valeurs calculées sur des échantillons d'individus.

6. Considérations pratiques

a) Quand utiliser des sondages à plusieurs degrés ?

Ce type de méthode est efficace pour certains domaines d'études, pas du tout pour d'autres (pour lesquels il existe une certaine "concentration" du phénomène étudié sur quelques unités primaires, par exemple). L'efficacité dépendra de la capacité à construire des unités primaires hétérogènes[21] (à l'intérieur d'elles-mêmes et vis-à-vis de la variable étudiée, c'est-à-dire contenant des individus plutôt différents les uns des autres).

On a évoqué précédemment deux méthodes de tirages autopondérés souvent utilisées selon que le tirage des unités primaires se fait à probabilités égales ou inégales (en général proportionnelles à leur taille). Si l'on dispose de l'information pour tirer à probabilités inégales, c'est cette méthode que l'on utilisera en général[22].

De plus, on a intérêt à stratifier au niveau des unités primaires[23] : par exemple, si celles-ci sont des regroupements locaux, on créera des strates en utilisant des critères administratifs ou agro-écologiques. C'est à ce niveau qu'on gagne de la précision. La stratification des unités primaires sera parfois liée aussi à la volonté de produire des résultats au niveau des strates. Par contre, la stratification des unités secondaires s'avère beaucoup moins rentable (de plus, il y aurait un risque à transformer le recensement au sein des unités primaires en une opération trop lourde destinée à cette stratification des unités secondaires).

Par ailleurs, on aura intérêt à forcer sur le nombre d'unités primaires enquêtées plus que sur le nombre d'unités secondaires enquêtées au deuxième degré (voir par exemple la formule $DEFF = 1+\delta(\overline{n}-1)$ où l'augmentation de $\overline{n}$ entraîne une perte de précision) : il vaut mieux beaucoup d'unités primaires avec peu de questionnaires dans chacune, plutôt que peu d'unités primaires enquêtées (même

---

[21] Cette construction va donc dans un sens contraire à la stratification, qui, elle, cherche à créer des regroupements d'individus semblables.

[22] On doit remarquer qu'avec cette méthode on connaît à l'avance le nombre total de questionnaires à réaliser (on fixe le nombre d'unités secondaires enquêtées par unité primaire tirée). Par contre, pour le tirage des unités primaires à probabilités égales avec taux de sondage au deuxième degré fixé *a priori* ou le tirage par grappes, on ne connaît pas *a priori* le nombre d'unités secondaires par unité primaire (on le découvre par comptage sur le terrain) et le nombre final de questionnaires est inconnu.

[23] Rappelons que dans le cas d'un sondage stratifié, on calcule la variance d'estimation strate par strate et que la variance de l'estimateur du total est la somme de ces variances sur l'ensemble des strates.

avec beaucoup d'unités secondaires). Comme on l'a vu dans l'exemple de l'enquête agricole présenté plus haut, ceci doit bien sûr tenir compte des différents facteurs de coût.

Enfin, la pertinence du plan de sondage dépendra du sujet étudié. Prenons le cas du ménage, grappe d'individus, composé en général de personnes de sexe et d'âge différents et comprenant souvent des actifs, des inactifs, des écoliers, etc. Dans une enquête socio-économique, le ménage sera une grappe très efficace pour estimer les variables sexe, activité, âge, etc. Mais ce sera une grappe moins efficace pour étudier le niveau d'instruction, l'origine ethnique, etc.

b) Pour les enquêtes démographiques dans les pays en développement

Un certain nombre de "valeurs numériques" ont été présentées précédemment pour l'effet de grappe. On pourra compléter ces valeurs par d'autres fournies dans certains articles[24].

La constitution d'unités primaires fait souvent appel au travail réalisé à l'occasion du recensement de la population : un travail de découpage a déjà été opéré, aboutissant à une cartographie des zones de dénombrement (ou districts). L'analyse des enquêtes passées dans les pays africains semble indiquer que la taille "optimale" des unités primaires (ou des unités à l'intérieur desquelles on procède au tirage des ménages, dans le cas de tirages à plus de deux degrés) se situerait "vers" 300 à 500 individus[25]. Éventuellement, on subdivise ou on regroupe les zones de dénombrement pour se rapprocher de cette taille.

Ensuite, pour l'enquête, il est nécessaire de procéder pour chaque unité primaire tirée à un dénombrement complet des ménages : il n'est pas possible de se limiter à une liste issue du recensement, puisque des modifications se sont produites depuis, en particulier la construction de nouvelles habitations.

Combien d'unités enquêter par unité primaire ? Pour certains sujets (mortalité, natalité), on peut enquêter la grappe complète (par exemple 300 personnes). Pour d'autres (par exemple la fécondité), il y a un effet de grappe plus marqué ; on se limite à moins d'unités enquêtées (par exemple 50) ; ce choix dépend également de considérations pratiques, en particulier la longueur du questionnaire à remplir. La taille totale de l'échantillon des enquêtes démographiques est variable selon les thèmes étudiés, allant par exemple de 5 000 à 100 000 questionnaires, voire plus

---

[24] On pourra par exemple se référer à Verma, Scott et O'Muircheartaigh (1980) ou à Verma et Lê (1996).

[25] Au-delà de cette taille on risque une dégradation du travail de dénombrement du district de recensement ; d'autre part on a vu qu'il est préférable d'avoir des unités primaires, dans les sondages à plusieurs degrés, "pas trop grandes".

pour certaines enquêtes, ceci pour des estimations "globales" ; si l'on veut des estimations plus fines, par exemple sur les taux de mortalité par groupes d'âges quinquennaux, il faudra augmenter cette taille[26].

7. Aperçu sur le tirage à trois degrés

Supposons qu'on tire des communes, puis des districts, puis des ménages. Le principe d'extrapolation d'un total est simple : on estime le total pour l'unité à l'intérieur de laquelle on a procédé au troisième degré de tirage (district), puis on tient compte du mode de tirage au deuxième degré pour estimer le total au niveau de la commune tirée, et on extrapole ensuite à l'univers. Cette succession d'estimations doit bien sûr être adaptée aux différentes options qu'on a choisies pour les trois degrés de sondage.

Pour l'estimation d'une moyenne par ménage, la méthode est différente selon que l'on connaît le nombre total de ménages (c'est alors direct à partir du total estimé sur l'univers) ou non (on doit alors estimer le nombre total de ménages à partir de l'échantillon).

[26] Dans ce cas, il faut cependant mentionner que l'usage de certains modèles démographiques (tables-type de mortalité par exemple) permet d'éviter de "gonfler" l'échantillon dans une proportion égale à celle qu'on devrait avoir si on estimait chaque taux quinquennal comme un élément indépendant des autres taux quinquennaux, puisqu'on part de cette table-type et de certains paramètres estimés pour décomposer la mortalité selon les tranches d'âges (CEPED, 1988, chapitre 20).

Encadré 3

Un exemple de sondage à plusieurs degrés

Les enquêtes démographiques et de santé EDS (Scott, 1987)

Ce programme d'enquêtes a été lancé depuis 1986 dans un nombre important de pays. Bien que les conditions de réalisation ne soient pas exactement semblables dans l'ensemble des pays, on peut donner une idée des principes de réalisation de ces enquêtes.

1. Principe de sélection de l'échantillon

Le sondage est à deux ou trois degrés, pour aboutir à un échantillon de femmes âgées de 15 à 49 ans.

Au premier degré, on tire des districts de recensement : ceux-ci sont, en général, trop grands pour constituer les unités qu'on va dénombrer (la taille "idéale" serait de 500 personnes ; il n'est pas rare d'avoir des districts de recensement de 1 000 à 2 000 personnes). On divise donc, pour chaque district, sa taille (connue au recensement) par 500 et, en arrondissant à l'entier le plus proche, on détermine le nombre de "segments théoriques" qu'il contient.

On tire ensuite un échantillon de districts avec une probabilité proportionnelle à ce nombre de segments théoriques.../...

Au second degré, pour chaque district tiré au premier degré, on découpe le district en autant de "segments théoriques" qu'il en contient et on sélectionne un de ces segments. La probabilité de tirage, *in fine*, de chaque segment de l'univers est la même.

Au troisième degré, après avoir dénombré les ménages des segments tirés, on sélectionne de façon aléatoire simple un échantillon de ménages dont on enquête toutes les femmes de 15 à 49 ans (en général pour arriver à un nombre de questionnaires d'environ 20 en milieu urbain et 40 en milieu rural ; on procède parfois à un niveau supplémentaire de tirage, en choisissant une femme de 15 à 49 ans dans chaque ménage sélectionné).

2. Remarques

- Le tirage n'est pas un tirage aléatoire simple de segments puisque deux segments appartenant au même district ne seront jamais tirés ensemble.

- Il n'est pas nécessaire de procéder au découpage des segments pour l'ensemble des districts ; on limite cette opération aux districts tirés au premier degré.

- Des districts trop petits peuvent être regroupés avec le suivant (ou le précédent).

- Le sondage est bâti pour être à peu près autopondéré : on devrait pouvoir se contenter de calculer des moyennes simples sur l'échantillon. Cependant, des perturbations (non réponses, non identification...) peuvent remettre en cause le caractère autopondéré du sondage. Il faut recalculer, *in fine*, les pondérations.

- On peut opérer, au niveau des unités primaires, une stratification géographique.

CHAPITRE 6

UTILISATION D'INFORMATION AUXILIAIRE, REDRESSEMENTS

Les chapitres précédents se plaçaient dans une perspective "avant l'enquête". Une fois l'enquête réalisée, on doit intégrer deux types de considérations :

- un certain nombre d'événements se sont produits, qui ont perturbé le schéma "idéal" prévu (refus de répondre, pertes de questionnaires...) ;
- des variables ont été collectées par l'enquête et fournissent une information sur l'échantillon : on a tiré par exemple un échantillon de ménages dans une base de sondage où on ne disposait pas d'information sur la taille des ménages, et, *a posteriori* (à partir des questionnaires), on observe la manière dont l'échantillon tiré se comporte par rapport à des statistiques (connues par ailleurs) sur la distribution des ménages. On utilise donc dans ce cas une information auxiliaire qui est "extérieure" au sondage (figure 9) et on cherche à "caler" l'échantillon de façon à respecter cette information (en général une distribution) connue.

On peut alors (et on doit dans le cas de non-réponses par exemple) proposer des estimateurs tenant compte des informations dont on dispose (sur le déroulement de l'enquête, ou informations "extérieures").

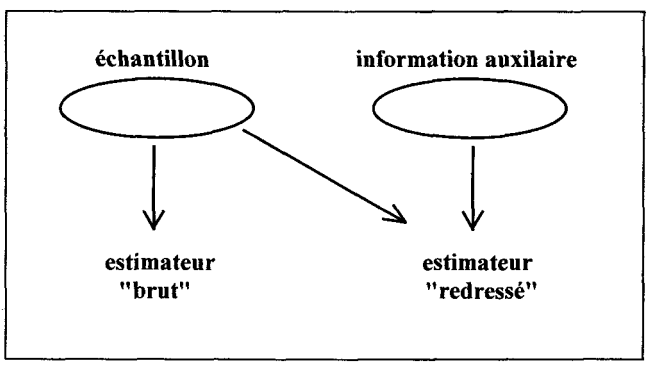


Figure 9. Utilisation d'information auxiliaire

Deux types de méthodes sont présentés dans les paragraphes suivants, avant un retour sur les non-réponses.

1. Stratification *a posteriori*

a) Principe

C'est le même principe que celui présenté au chapitre 3. On découpe l'univers en strates et on effectue des estimations par strates avant de concaténer le tout pour obtenir une estimation globale. Par exemple, dans le cas où on a tiré un échantillon de manière aléatoire simple (c'est-à-dire à probabilités égales) :

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ est l'estimateur "brut" (avant redressement) de la moyenne de la variable Y.

Si l'on découpe l'univers en strates $h = 1, \ldots, k$ et si l'on connaît les effectifs N_h des strates, alors :

$\bar{y}_{sp} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h$ est l'estimateur stratifié *a posteriori* de la moyenne de Y ($\bar{y}_h$ étant la moyenne simple calculée sur la partie de l'échantillon se trouvant dans la strate h).

On voit qu'on modifie les pondérations des questionnaires par rapport à l'estimateur "brut".

b) Quelle est la différence avec la stratification a priori *?*

On ne maîtrise pas, dans la stratification *a posteriori*, la répartition des unités enquêtées entre les strates ; un cas extrême est celui où on ne trouverait pas d'unité de l'échantillon dans une strate définie *a posteriori*. La stratification *a posteriori*, si elle "recale" l'échantillon en le pondérant pour l'ajuster sur une distribution connue, est en général moins efficace qu'une stratification *a priori* bien choisie.

c) Exemple

On tire, parmi un univers de 2 536 villages, un échantillon de 127 villages dont on veut estimer la population par enquête. L'échantillon est tiré à probabilités égales et on observe une taille moyenne $\bar{y}$ de 377,2 habitants sur l'échantillon des 127 villages.

L'estimation "brute" du total de la population à l'enquête est donc :

$$N\bar{y} = 2\,536 \times 377{,}2 = 956\,600 \text{ habitants.}$$

On s'aperçoit que l'échantillon, relativement à sa répartition géographique, a plutôt sur-représenté les villages de la zone Sud (tableau 7).

Tableau 7. Stratification *a posteriori* : exemple

Zone	Villages de l'univers	Villages de l'échantillon	Taille moyenne des villages de l'échantillon
Nord	1 421	65	402,8
Sud	1 115	62	350,4
Total	2 536	127	377,2

Or ces villages (du Sud) ont en moyenne une taille plus faible. Comment prendre en compte cette information ? On utilise l'estimateur de la moyenne stratifié *a posteriori* :

$$\bar{y}_{sp} = \frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2$$

et l'estimateur du total de la population vaut donc :

$$N\bar{y}_{sp} = N_1\bar{y}_1 + N_2\bar{y}_2 = (1\,421 \times 402{,}8) + (1\,115 \times 350{,}4)$$
$$= 963\,100 \text{ habitants.}$$

Soit une estimation légèrement supérieure à l'estimation "brute".

d) La pratique

Le critère choisi pour stratifier *a posteriori* doit être corrélé avec la variable d'intérêt (ou les variables d'intérêt) pour que la technique soit efficace.

Par ailleurs, il est essentiel que les effectifs des strates (les N_h) soient connus de manière précise et surtout récente : une stratification *a posteriori* ajustant un échantillon sur une distribution ancienne (et susceptible de s'être déformée) sera à déconseiller.

Enfin, il est préférable de ne pas avoir de corrections des pondérations trop importantes : une règle empirique indique d'éviter d'avoir des taux de correction plus de cinq fois supérieurs au taux de correction le plus faible. On déconseille aussi de stratifier *a posteriori* sur des strates trop peu nombreuses (éviter des strates telles que $N_h/N < 10$ %).

e) Que faire si le plan de sondage est plus complexe qu'un sondage aléatoire simple ?

On peut donner quelques pistes (pour plus de détails, voir Ardilly, 1994) :

- si le plan de sondage est complexe sans stratification et si les probabilités d'inclusion des unités (c'est-à-dire les probabilités d'appartenance à l'échantillon) sont égales (par exemple pour certains tirages à deux degrés), alors on utilise l'estimateur de la moyenne de Y :

$$\bar{y}_{sp} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h$$

où $\bar{y}_h$ est la moyenne simple calculée sur les questionnaires de la strate h ;

- si le plan de sondage est sans stratification et à probabilités inégales, on estime la moyenne de Y par :

$$\bar{y}_{sp} = \sum_{h=1}^{k} \frac{N_h}{N} \frac{\hat{T}_h}{\hat{N}_h}$$

où $\hat{T}_h$ estime le total de Y et $\hat{N}_h$ est l'estimateur de la taille pour la strate h (bien que cette taille soit connue de manière exacte) ;

- si le plan de sondage est stratifié, il faut essayer de poststratifier à l'intérieur des strates initiales.

f) La méthode du raking ratio

Cette méthode est employée quand on essaye de "caler" l'échantillon sur plusieurs critères sans qu'on connaisse le croisement des distributions associées : par exemple on se cale sur la distribution des ménages selon leur taille, et aussi sur leur distribution selon le niveau d'instruction du chef de ménage.

On cale d'abord l'échantillon sur une distribution en modifiant les pondérations des questionnaires (c'est-à-dire qu'on multiplie la pondération de toutes les unités d'une même "tranche" de la distribution par le même coefficient pour qu'après extrapolation on retrouve le nombre d'unités de cette tranche connu par la distribution "extérieure"). Dans un deuxième temps, on remodifie les pondérations pour se caler sur l'autre distribution. Puis on recommence avec la première distribution et après un certain nombre d'itérations, on obtient les pondérations définitives.

Cette méthode peut être adoptée dans le cas d'un sondage aléatoire simple, ou d'un plan plus complexe où les probabilités d'inclusion finales des unités sont égales et l'échantillon de taille fixe[26].

2. Estimation par le quotient

a) Principe

Il est différent du principe relatif à la stratification *a posteriori*. Pour celle-ci, on se "calait" sur des effectifs (distribution selon un certain critère) ; ici, on va se caler par rapport à une valeur moyenne.

On a tiré un échantillon pour lequel on étudie une variable Y, mais on observe aussi une variable X. Pour cette variable X, on connaît la moyenne $\overline{X}$ de manière exacte sur l'univers. On peut "observer" le résultat $\bar{x}$ obtenu sur l'échantillon et le comparer à $\overline{X}$. L'idée est la suivante : pour les variables qui varient "proportionnellement" à la variable X, on tient compte du résultat $\bar{x}$ et on propose l'estimateur par le quotient (de la moyenne de Y) :

$$\bar{y}_q = \bar{y}\frac{\overline{X}}{\bar{x}}$$

[26] Pour plus de détails sur la méthode, voir Ardilly, 1994.

Donc si l'échantillon fournit un résultat $\bar{x}$ inférieur à $\overline{X}$, on pense qu'il est intéressant de "donner un coup de pouce" à l'estimation brute $\bar{y}$ (et inversement si $\bar{x} > \overline{X}$).

L'estimateur par le quotient est biaisé, mais si la variable auxiliaire X et la variable Y étudiée sont approximativement proportionnelles, sa variance est inférieure à celle de l'estimateur simple[27] ; le biais étant d'un ordre de grandeur "dominé" par celui de l'écart-type, l'estimateur par le quotient est alors intéressant. Mais ceci, répétons-le, ne s'applique qu'au cas où il existe une relation de proportionnalité présumée entre les deux variables X et Y.

b) Exemple

Reprenons l'exemple de la première partie de ce chapitre. On tire, parmi un univers de 2 536 villages, un échantillon de 127 villages pour estimer la population par enquête.

Pour l'ensemble des 2 536 villages, on a l'information "population au dernier recensement". La taille moyenne des villages au dernier recensement est de 345,1 habitants ($\overline{X}$ = 345,1). On a constaté, sur l'échantillon, que $\bar{x}$ = 341,7.

D'autre part la population moyenne, à la date de l'enquête, des villages de l'échantillon est de 377,2 habitants. On peut donc proposer deux estimations de la population de l'univers étudié :

- l'estimation "brute" $N\bar{y}$ = 2 536 x 377,2 = 956 600 habitants ;

- l'estimation par le quotient $N\bar{y}_q = N\bar{y}\dfrac{\overline{X}}{\bar{x}}$ = 966 100 habitants.

Cette deuxième estimation repose sur la constatation que l'estimation issue de l'échantillon tiré était, au recensement de la population, un peu en dessous de la valeur moyenne, et qu'il existe certainement une relation de proportionnalité entre population au recensement et population actuelle : le redressement consiste à réévaluer à la hausse "l'estimation brute".

---

[27] On montre que, dans le cas d'un sondage aléatoire simple, la variance de l'estimateur par le quotient peut être estimée à partir de l'échantillon par :

$$\hat{V}\left(\bar{y}_q\right) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\left(s_y^2 + \hat{r}^2 s_x^2 - 2\hat{r}\hat{\rho}s_x s_y\right)$$

où s_y^2 et s_x^2 sont calculés sur l'échantillon (formule habituelle),

$\hat{r} = \dfrac{\bar{y}}{\bar{x}}$,

$\hat{\rho}$ est le coefficient de corrélation linéaire entre X et Y estimé sur l'échantillon.

Pour revenir à la validité de l'application de la méthode, la relation de proportionnalité supposée entre les deux variables, à savoir la population au recensement et la population au moment de l'enquête, n'est pas toujours vérifiée. On a pu constater par exemple, dans certains pays africains, des évolutions de population dues à des raisons climatiques qui ont conduit certains villages à doubler leur population en quelques années alors que d'autres, pendant la même période, voyaient leur population réduite de moitié. Dans ce cas, on voit les risques qu'il y a à appliquer brutalement une telle méthode de redressement...

c) L'estimateur par la régression

Cette méthode suppose une relation de type affine entre Y, la variable d'intérêt, et X, la variable auxiliaire, qui n'est plus une relation de simple proportionnalité (comme pour l'estimateur par le quotient) : $Y = a + bX$.

L'idée va être d'estimer le paramètre b, puis d'utiliser la grandeur $\overline{X}$ (valeur moyenne de X sur l'univers, connue) pour redresser et fournir l'estimateur par la régression de la moyenne : $\bar{y}_{reg} = \bar{y} + \hat{b}(\overline{X} - \bar{x})$ où $\hat{b}$ est l'estimation de b par la méthode des moindres carrés ordinaires appliquée à l'échantillon.

Cette méthode suppose des calculs complexes et est peu utilisée en pratique (voir cependant l'encadré 4 qui présente un exemple d'application). On utilise parfois une "variante", l'estimation par la différence, où la valeur de b est choisie *a priori* égale à 1 :

$\bar{y}_{diff} = \bar{y} + \left(\overline{X} - \bar{x}\right)$ (on ajoute à $\bar{y}$ la différence constatée entre $\overline{X}$ et $\bar{x}$).

3. Les non-réponses

a) Non-réponses partielles et totales

Il n'existe pas de remède miracle à ce problème. Le premier conseil qu'on peut donner est d'essayer de les éviter au maximum. Cependant, les non-réponses existent et il faut les prendre en compte. On doit tout d'abord les séparer en deux catégories :

- non-réponses partielles (on a une partie du questionnaire, il manque[28] certains renseignements) ;

[28] Il peut également s'agir de données qui ont été jugées incohérentes lors de la phase d'apurement du fichier.

- non-réponses totales (pour des raisons de refus, d'impossibilité de joindre l'unité à enquêter,...) : il faut, déjà à ce niveau, vérifier que ces non-réponses totales (particulièrement quand on n'arrive pas à joindre l'unité à enquêter) ne correspondent pas en fait à des unités "hors champ de l'enquête".

Pour traiter ces problèmes, on utilise une approche "modèle", c'est-à-dire qu'on "sous-entend" des hypothèses de comportement pour prendre en compte les valeurs manquantes.

Encadré 4

Un exemple d'estimation par la régression

L'estimation de superficies agricoles à partir d'images satellites

Chaque année, une enquête est menée par le service de statistique agricole français (SCEES) dans un certain nombre de départements, à partir d'un échantillon de segments (carrés de 50 hectares) visités par des enquêteurs qui procèdent à des relevés permettant de déterminer la superficie de différentes catégories d'utilisation du sol. Cette enquête fournit des estimations "brutes", par exemple pour la superficie boisée du département.

Par ailleurs, l'image satellite délivre une information exhaustive sur la zone, mais avec une certaine erreur d'observation : il existe des confusions dans l'affectation des points de l'image aux différentes catégories d'utilisation du sol.

On peut, sur l'image satellite, repérer les segments de l'enquête de terrain et disposer pour chacun de ceux-ci de :

y superficie boisée à l'enquête de terrain

x superficie boisée sur l'image satellite

Le fait de disposer, sur l'image satellite, de la valeur $\overline{X}$, moyenne sur l'ensemble des segments, va permettre de fournir un estimateur redressé de la superficie moyenne (par segment) en bois :

$$\bar{y}_{reg} = \bar{y} + \hat{b}\left(\overline{X} - \bar{x}\right)$$

où $\bar{x}$, $\bar{y}$ sont les moyennes sur l'échantillon de segments ($\bar{y}$ sert d'ailleurs à produire l'estimation brute de la superficie en bois du département), et $\hat{b}$ est la pente de la droite de régression entre x et y, estimée par l'échantillon. L'image satellite est donc l'information auxiliaire qui "redresse" les résultats de l'enquête de terrain (et non l'inverse). Pour plus de détails, voir Pastorelli (1992).

b) Comment traiter les non-réponses totales ?

Celles-ci sont susceptibles d'introduire un biais dans l'estimation : les non-répondants peuvent avoir des caractéristiques spécifiques relativement aux variables étudiées par l'enquête.

Il est donc important de disposer d'information sur les non-répondants (variables se trouvant dans la base de sondage), et de mener des analyses sur ceux-ci afin de déterminer leur "profil" : sont-ce plutôt des jeunes, des ruraux... ?

La méthode la plus généralement proposée consiste à utiliser un estimateur (du total, par exemple) du type : $\hat{T} = \sum_{i \in r} \frac{W_i}{R_i} Y_i$

où
- w_i est la pondération initiale de l'unité i ;
- R_i est la probabilité de réponse de l'unité i ;
- la sommation est faite sur r, l'échantillon des répondants.

Le problème est d'estimer R_i. Ceci est impossible au niveau individuel ; on propose de déterminer des grandes catégories pour lesquelles on va avoir la même valeur R_i, souvent en calculant, pour chacune, la proportion d'unités ayant répondu à l'enquête (ceci nécessite d'avoir l'indication dans la base de sondage de la catégorie à laquelle appartient chaque unité).

Au niveau du choix des critères définissant les catégories sur lesquelles on va redresser, on essaiera de définir des catégories homogènes vis-à-vis des thèmes étudiés par l'enquête, de taille suffisante et pour lesquelles le "mécanisme" de non-réponse est différencié d'une catégorie à l'autre : ce sera le cas, par exemple, pour certaines enquêtes, de la variable "âge". L'estimateur proposé ci-dessus repose alors sur l'hypothèse que les répondants et les non-répondants de chaque catégorie ont le même comportement vis-à-vis des variables étudiées par l'enquête.

Parfois, on utilise une stratification *a posteriori* (à condition que les strates définies aient suffisamment d'unités) : on va voir sur un exemple simple que l'estimateur post-stratifié est différent de celui proposé précédemment. En supposant qu'on redresse un échantillon aléatoire simple de n unités en utilisant deux catégories, l'estimateur utilisant les R_i est :

$$\hat{T} = \sum_{i \in r} \frac{y_i}{\frac{n}{N} R_i} = \sum_{i \in r_1} \frac{N}{nR_1} y_i + \sum_{j \in r_2} \frac{N}{nR_2} y_j$$

où r_1 est l'échantillon des répondants de la catégorie 1 ;
et r_2 est l'échantillon des répondants de la catégorie 2.

R_1 est égal à n_{1r}/n_1, si n_{1r} est le nombre de répondants de la catégorie 1, et n_1 le nombre de questionnaires initialement tirés et se trouvant dans cette catégorie. De même R_2 est égal à n_{2r}/n_2.

Donc $$\hat{T}=\frac{Nn_1}{nn_{1r}}\sum_{i\in r_1} y_i+\frac{Nn_2}{nn_{2r}}\sum_{j\in r_2} y_j$$

L'estimateur utilisant la post-stratification vaut :

$$\hat{T}_{post}=\frac{N_1}{n_{1r}}\sum_{i\in r_1} y_i+\frac{N_2}{n_{2r}}\sum_{j\in r_2} y_j$$

Ces deux estimations n'ont la même valeur que si $n_1=(N_1n)/N$ et $n_2=(N_2n)/N$, c'est-à-dire si le nombre de questionnaires tirés dans la catégorie 1 est calé sur la proportion de cette catégorie dans l'univers (ce qui, *a priori*, n'a rien d'évident).

L'estimateur post-stratifié demande en fait plus d'informations (les effectifs N_h) que l'estimateur construit à partir des probabilités de réponse estimées R_i ; par contre il permet d'ajuster les effectifs extrapolés sur la distribution N_h.

Enfin, la résolution du problème des non-réponses totales peut aussi passer par une "relance" des non-répondants, afin d'estimer cette population de manière directe (cette méthode est cependant rarement utilisée en pratique) ; elle peut aussi passer par l'utilisation d'une réserve d'unités remplaçantes (tirées en "supplément") auxquelles on fait appel quand on a des non-réponses (ce qui suppose que les non-répondants n'ont rien de spécifique ; cette méthode sera appliquée, avec beaucoup de précautions, quand les non-réponses sont en nombre très limité).

c) Comment traiter les non-réponses partielles ?

On remplace les données manquantes d'une unité à l'aide d'une des méthodes d'"imputation" suivantes (qui supposent, là aussi, l'utilisation d'hypothèses de comportement) :

- déductive, à l'aide d'une règle déterministe (par exemple, un individu de moins de 14 ans est inactif), ou prédictive (en fonction des caractéristiques observées de l'unité, on fait appel aux données constatées sur les autres unités répondantes et semblables) ;
- "*hot-deck*" : on prend, pour la variable manquante, la valeur de l'unité précédente dans le fichier, ou celle de la dernière unité rencontrée et dont on pense qu'elle est suffisamment semblable à l'unité pour laquelle l'information manque ;

- "*cold-deck*" : on utilise une information "extérieure " à l'enquête relative à l'unité pour laquelle l'information manque, par exemple la valeur de la variable à une date antérieure, ou la valeur située dans un autre fichier.

Pour conclure, il faut signaler que ces méthodes d'"imputation" peuvent être utilisées pour traiter la non-réponse totale : si un questionnaire manque, on peut être tenté de "doubler" un questionnaire d'une unité analogue (c'est par exemple le cas des sondages à deux degrés où on peut doubler un répondant appartenant à la même unité primaire que le non-répondant). Cette manière de procéder peut conduire à prendre plus de risques que la méthode de repondération proposée pour les non-réponses totales, qui agit plus en fonction d'un "questionnaire moyen". Et on peut fournir une règle générale concernant les traitements des non-réponses : ne pas aboutir à une dispersion trop importante des pondérations finales.

CHAPITRE 7

LA MÉTHODE DES QUOTAS

Bien qu'encore peu utilisée dans les pays en développement, cette méthode est présentée dans ce manuel en raison de son caractère spécifique ; elle pourrait trouver des champs d'application dans ces pays.

1. Principe

Cette méthode consiste à imposer à l'échantillon de respecter certains quotas (c'est-à-dire des répartitions selon certains critères) afin de "représenter" au mieux l'univers.

Exemple

On veut faire une enquête socio-économique sur la population active d'une ville. Un recensement récent a fourni les répartitions globales suivantes d'après trois critères (tableau 8 : remarquons qu'il s'agit de distributions marginales, c'est-à-dire qu'on n'a pas fait de ventilation suivant deux critères, par exemple un tableau à double entrée suivant l'âge et le secteur d'activité).

On a décidé d'interroger 5 000 personnes avec dix enquêteurs travaillant 10 jours. On dira à chaque enquêteur : vous interrogez en tout 500 personnes dont 240 hommes et 260 femmes, 70 jeunes de 16 à 24 ans, 185 personnes âgées de 25 à 44 ans, 80 cadres ou patrons du secteur formel... Dans cet exemple, on a utilisé des quotas marginaux (plus utilisés en pratique) mais on peut aussi se servir de quotas croisés (nécessitant de l'information sur la répartition de la population pour chaque case résultant du croisement des critères).

Tableau 8. Exemple de quotas pour une enquête socio-économique

Sexe		Âge		Secteur d'activité	
Hommes	48 %	16-24 ans	14 %	*Secteur formel*	
Femmes	52 %	25-44 ans	37 %	Cadres, patrons	16 %
		45-64 ans	35 %	Employés, ouvriers	24 %
	100 %	65 ans et +	14 %		
				Secteur informel	
			100 %	Cadres, patrons	3 %
				Travailleur indépend.	36 %
				Employés, aides familiaux, etc	21 %
					100 %

2. La méthode est non aléatoire

La probabilité d'être enquêtée de chaque "unité statistique" n'est pas connue *a priori* : elle dépend en fait de l'enquêteur. Certains enquêteurs iront plus facilement interviewer certaines personnes que d'autres, pour des raisons d'affinités socio-culturelles.

De plus, certaines unités peuvent avoir une probabilité nulle d'être enquêtées : personnes grabataires, malades, étrangers maîtrisant mal la langue du pays, ou simplement personnages grincheux. On voit que ceci peut introduire un biais d'observation.

3. La pratique

L'utilisation de quotas cherche à réduire les inconvénients présentés au paragraphe précédent : elle suppose que les variables retenues pour la détermination des quotas ont une influence prépondérante pour les caractéristiques que l'on veut estimer.

La qualité des résultats dépend de deux paramètres :

- des enquêteurs bien formés, n'introduisant pas de biais de sélection systématique, et habitués à "gérer" leurs quotas (par exemple en

s'arrangeant pour ne pas se trouver, en fin d'enquête, avec l'obligation d'enquêter un retraité de 25 à 44 ans...) ;

- des informations sur les quotas fiables, c'est-à-dire récentes et exactes. La possibilité d'utiliser la méthode des quotas nécessite donc l'existence d'une "infrastructure" statistique, souvent provenant des producteurs "primaires" d'information (en l'occurence les services statistiques de l'administration).

Plus encore que pour les méthodes aléatoires, il est nécessaire de contrôler l'échantillon obtenu au regard de certains critères pour lesquels on connaît la distribution (en plus des critères utilisés pour les quotas) ; une stratification *a posteriori* sera opérée le cas échéant pour redresser les résultats (ceux-ci, *a priori*, sont calculés en "faisant comme si" on avait tiré l'échantillon de manière aléatoire simple).

Sous toutes ces conditions, la méthode peut fournir de bons résultats, particulièrement pour les échantillons de faible taille. Remarquons que sa mise en œuvre est beaucoup plus légère (pas besoin de base de sondage, pas de tirage de l'échantillon à réaliser, déplacements des enquêteurs moins contraints). Parfois, la méthode des quotas est utilisée au dernier degré d'un tirage où le premier degré (ou les premiers degrés) est un tirage aléatoire de zones géographiques.

En conclusion, elle présente un intérêt certain de par sa facilité de mise en œuvre mais son application repose d'abord sur la disponibilité de statistiques fiables et récentes pour établir les quotas (ce point constitue sans doute la raison pour laquelle cette méthode est très peu appliquée actuellement dans les pays en développement) ainsi que sur un certain savoir-faire du gestionnaire de l'enquête et des enquêteurs (afin d'éviter les biais de sélection des enquêtés).

Remarquons enfin que son application n'est pas limitée à des quotas sur les personnes et qu'on pourrait envisager d'autres champs d'application (par exemple, pour des enquêtes agricoles, quotas sur les concessions).

4. La méthode des itinéraires

Cette méthode est parfois présentée comme une variante de la méthode des quotas : elle ne nécessite pas de base de sondage, et on fournit à l'enquêteur des indications sur les unités à enquêter. Plus précisément, on lui impose un itinéraire fixé sur une carte, avec un point de départ et des points d'enquête déterminés le long de celui-ci.

La méthode s'approche en fait d'une méthode aléatoire (choix aléatoire de l'itinéraire) : l'enquêteur a une latitude beaucoup moins importante qu'avec la

méthode des quotas. Par contre, le coût de préparation (définition de l'itinéraire) est plus élevé.

La mise en œuvre pratique de la méthode des itinéraires pose cependant des problèmes, en particulier dans les pays en développement, relativement à la structure de l'habitat dans certaines zones : le travail de détermination de tous les itinéraires possibles permettant d'accéder à l'ensemble des habitants de la zone est souvent difficile à "boucler" de manière parfaite, les habitations n'étant pas réparties selon des "axes" simples. Là aussi, il est nécessaire d'opérer un contrôle important du travail des enquêteurs : la difficulté à "boucler" le travail de repérage des habitations a en particulier des conséquences au niveau de l'estimation de totaux (bien souvent, le nombre total de logements est inconnu).

CHAPITRE 8

EN GUISE DE SYNTHÈSE

Ce chapitre a pour but de mettre en perspective ce qui a été présenté dans les chapitres précédents, tout en insistant sur un certain nombre de points pratiques.

1. Quelle méthode dans quel contexte ?

Les méthodes présentées dans les chapitres précédents sont loin d'être antinomiques : on a même vu qu'elles s'utilisent souvent en complément les unes des autres.

Le choix de la procédure définitivement adoptée va dépendre des paramètres suivants :

- raisons pratiques d'organisation de l'enquête (éléments de coût, de rémunération des enquêteurs, contraintes d'accessibilité de l'ensemble des zones du territoire, organisation du travail des enquêteurs) ;
- disponibilité d'une base de sondage fiable (complète, exacte et à jour), ou d'une liste pour constituer les unités primaires d'un sondage à plusieurs degrés ;
- caractéristiques profondes du phénomène étudié : semble-t-il *a priori* réparti à peu près uniformément sur l'univers, ou y a-t-il concentration forte sur certaines parties de celui-ci (zones géographiques ou catégories d'unités) ?

a) Un ou plusieurs degrés ?

Un des choix fondamentaux sera relatif à l'utilisation d'un plan de sondage à un degré ou à plusieurs degrés : celui-ci, plus pratique à mettre en œuvre en particulier

quand on organise une enquête de grande envergure (à l'échelle d'un pays), ne sera efficace que si le phénomène étudié ne présente pas de concentration sur certaines zones correspondant à des unités primaires.

La technique de la stratification sera, elle, quasiment systématiquement utilisée, très souvent conjointement avec d'autres méthodes.

Pour beaucoup d'enquêtes dans les pays en développement, on peut recommander la procédure suivante :

- stratification préalable des unités primaires (en utilisant des critères pertinents, intégrant des variables géographiques le plus souvent) ;
- tirage à plusieurs degrés[28], avec un premier degré aréolaire constitué d'unités primaires de tailles aussi voisines que possible. Le tirage des unités primaires se fera à probabilités égales si celles-ci sont de tailles à peu près équivalentes (ce qui peut résulter en particulier d'une stratification judicieuse), ou à probabilités inégales proportionnelles à leurs tailles (si on dispose de cette information).

Parfois, il sera possible d'utiliser une méthode à un seul degré (encadré 2 ou, quand un registre est disponible, enquêtes auprès des entreprises).

Une fois le mode de tirage déterminé (avec tous ses "raffinements"), se pose le problème de l'extrapolation des valeurs observées sur l'échantillon. Pour l'estimation d'un total, on utilise le principe du "jeu de construction" : estimation strate par strate s'il y a stratification puis sommation générale, remontée par chacun des degrés de tirage si on est dans le cas de plusieurs degrés (par exemple estimation du total au niveau unité primaire, puis estimation au niveau global). Un ratio sera estimé comme le rapport de deux masses ; quant à l'estimation d'une moyenne, elle proviendra, si le nombre d'unités "de base" est inconnu, d'un ratio où le dénominateur (le N) sera estimé à partir de l'échantillon.

b) Les panels

Un cas particulier est relatif aux panels, pour lesquels on enquête le même échantillon à plusieurs dates : l'intérêt est qu'ils permettent, en général, des

[28] La taille du chapitre 5 de ce manuel montre l'importance des sondages à plusieurs degrés pour les enquêtes démographiques... Un certain nombre de publications ont présenté des plans de sondage d'enquêtes démographiques de manière détaillée : on pourra par exemple se référer à Rémy Clairin (1978) et Christopher Scott (1987).

estimations d'évolution (entre les différentes dates) plus précises[29] que celles qu'on aurait obtenues en tirant à chaque date un nouvel échantillon ; de plus, l'intérêt des panels est de pouvoir produire des matrices de passage entre deux dates d'enquêtes : si par exemple on suit des personnes du point de vue de leur emploi, on pourra étudier comment se font les passages d'une catégorie d'emploi à une autre (combien du secteur public vers l'informel...) alors que la même enquête pratiquée sur deux échantillons indépendants (aux deux dates) ne permettrait que d'avoir deux "photographies" de l'emploi aux deux dates.

Par contre, la gestion d'un panel s'avère complexe dans le cas d'unités à "géométrie variable", en particulier dans le cas de ménages : il peut y avoir des arrivées, des départs, des fusions... qui rendent le suivi des unités difficile.

c) Les sondages en deux phases

Cette technique est parfois utilisée quand on dispose de très peu d'informations dans la base de sondage : elle consiste à tirer un échantillon nombreux sur lequel on pose peu de questions (première phase) avant de sélectionner un sous-échantillon en deuxième phase, lequel sera "ciblé" en fonction des objectifs poursuivis (par exemple étude de populations rares) et à partir des informations collectées lors de la première phase.

On utilisera, par exemple, pour les résultats de la deuxième phase un estimateur stratifié du type :

$$\bar{y}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h$$

mais ici les N_h/N seront des résultats d'estimations provenant de la première phase : si on a tiré, lors de la première phase, n unités et que n_1 ont été "constatées" comme appartenant à la strate 1, n_1/n sera l'estimateur de N_1/N.

---

[29] Si on veut estimer l'évolution de la moyenne d'une variable Y entre deux dates t_1 et t_2, on montre que la variance de l'estimation obtenue à partir d'un panel tiré de façon aléatoire simple ($\hat{\Delta Y}_p$) est, par rapport à l'estimation $\hat{\Delta Y}$ obtenue à partir de deux échantillons indépendants de même taille tirés selon la même méthode aux deux dates, égale à :

$$V\left(\hat{\Delta Y} p\right) = V\left(\hat{\Delta Y}\right)(1-\rho)$$

où ρ est le coefficient de corrélation linéaire entre la variable Y à la date t_1 et la même variable Y à la date t_2. Comme pour beaucoup de variables étudiées dans les enquêtes, ρ est positif (remarquons que sa valeur est souvent liée à la longueur de l'intervalle entre t_1 et t_2), l'estimation de l'évolution par l'échantillon "permanent" (panel) est dans ce cas préférable.

La moyenne sur la strate h, $\bar{y}_h$, sera, elle, calculée sur le sous-échantillon tiré dans cette strate pour la deuxième phase. La variance de l'estimateur $\bar{y}_{st}$ sera plus complexe à calculer, puisque résultant de deux niveaux d'estimations aléatoires : $\bar{y}_h$ d'une part, mais aussi l'estimation de N_h/N.

Cette technique a, par exemple, été utilisée au Cameroun pour étudier le secteur informel de Yaoundé (DIAL-DSCN, 1994). Le schéma est assez complexe puisqu'en première phase, une enquête "emploi" permet d'identifier des personnes travaillant dans des unités de production informelles ; la deuxième phase consiste en une enquête auprès de ces unités de production (les unités d'enquête sont donc différentes d'une phase à l'autre). L'intérêt de cette méthode est de permettre une bonne couverture du secteur informel, alors que les études menées à partir de recensements directs des unités de production informelles semblent "laisser de côté" une partie de ces unités.

d) Les estimations pour de petits domaines

Un problème courant se pose au moment de la production de résultats d'enquêtes : celui des estimations locales, pour de "petits domaines", par exemple une subdivision administrative d'un pays. On peut, bien entendu, fournir une estimation directe à partir des unités de l'échantillon qui sont dans le domaine en question : $\bar{y}_d$, la moyenne d'une variable calculée sur ces unités, est par exemple l'estimateur de la moyenne de cette variable sur ce domaine, et si l'on connait N_d, l'effectif du domaine, $N_d \bar{y}_d$ est l'estimateur du total de la variable. Le problème est que cet estimateur est souvent très imprécis, du fait du faible nombre d'unités enquêtées à l'intérieur du domaine.

L'idée est alors d'utiliser une "approche modèle" qui passe par des hypothèses de comportement : on utilise de l'information connue et "extérieure" au domaine (en fait plus globale), qu'on réinjecte dans l'estimateur (Aliaga, 1995). Par exemple on utilise un estimateur $\bar{y}$ calculé sur l'ensemble de l'échantillon, ou sur une partie plus grande que le domaine considéré, et on estime le total sur le domaine par $N_d \bar{y}$. Ceci suppose donc que le comportement des unités à l'intérieur du domaine puisse être représenté par celui observé sur un ensemble plus vaste : c'est une hypothèse forte que l'on fait en procédant de la sorte, et qui incite à être prudent dans l'utilisation de ce genre de méthode.

On peut raffiner la technique, en utilisant des estimations $\bar{y}_h$ sur des sous-groupes h (par exemple des estimations établies, au niveau global et non au niveau du domaine, pour des tranches d'âges) et en tenant compte de la répartition connue de ces sous-groupes dans le domaine (valeurs $N_{d,h}$ et N_d connues), pour fournir un estimateur de la moyenne :

$$\hat{\bar{y}}_d = \sum_h \frac{N_{d,h}}{N_d} \bar{y}_h$$

On utilise également, quand on dispose d'une information exhaustive X_d sur le petit domaine, l'estimateur "synthétique par le ratio", obtenu par :

$$\hat{y}_d = X_d \hat{R}$$

où $\hat{R}$ est estimé sur un ensemble plus vaste que le petit domaine (on fait donc l'hypothèse que la proportionnalité entre le total de X et le total de y est la même sur le petit domaine et sur l'ensemble plus vaste).

On peut combiner les deux types d'estimateurs précédents ("direct" obtenu à partir des unités de l'échantillon contenues dans le domaine, et "synthétique" obtenu à partir d'un domaine plus vaste) en une moyenne pondérée des deux (Destandau, 1996).

e) La méthode des segments

Cette méthode aréolaire utilise un découpage du territoire (une "segmentation", d'où le nom de la méthode) en segments qui sont déterminés à partir de limites naturelles (cours d'eau, routes,...) ou construits "artificiellement" sur des cartes, souvent sous forme de carrés. Elle est utilisée dans les pays en développement en particulier pour les enquêtes agricoles, et il est important de bien distinguer les différentes catégories d'applications présentées ci-après.

La méthode du **segment fermé** consiste à relever l'ensemble des types d'utilisation du territoire à l'intérieur du segment, afin de produire des estimations de superficie (pour un exemple d'utilisation, voir Fournier, 1986). Cette méthode est limitée aux aspects physiques de composition du segment. L'unité d'observation étant le segment, les estimations sont donc produites en fonction de la manière dont l'échantillon de segments a été tiré. La méthode n'est par contre pas applicable si l'on s'intéresse aux caractéristiques de nature socio-économique des exploitations agricoles (démographie des chefs d'exploitation, type de matériel agricole, variables financières,...).

Pour ce genre d'objectifs, on peut utiliser la méthode du **segment ouvert**, qui consiste à enquêter toutes les exploitations agricoles dont le "siège" (encore faut-il avoir défini ce concept) est à l'intérieur d'un des segments tirés dans l'échantillon. On a donc un tirage en grappes et il faut noter que le questionnaire de l'enquête concerne les terres des exploitations agricoles qui sont situées hors du segment aussi bien que celles qui sont à l'intérieur (d'où le nom de segment "ouvert").

Cette méthode présente cependant, souvent, le défaut d'un manque de couverture : si l'on réalise sur la même zone une enquête par les deux méthodes (segment ouvert, segment fermé), on constate en général des estimations de superficie, pour la même catégorie, plus faibles à partir du segment ouvert. La raison de ceci réside dans la difficulté à retrouver l'ensemble des exploitations agricoles situées, de par leur siège, dans un segment ; cet effet est particulièrement flagrant dans les zones péri-urbaines, où l'on constate une sous-estimation forte des superficies maraîchères ("ceintures maraîchères" des grandes villes) à partir de la méthode du segment ouvert.

Une troisième méthode, plus complexe d'application, la méthode du **segment pondéré**, permet de remédier à cette difficulté. Pour l'ensemble des segments tirés dans l'échantillon, on réalise un questionnaire pour chaque exploitation agricole ayant des terres sur le segment : on va donc avoir un nombre total de questionnaires plus important que par la méthode du segment ouvert. L'intérêt de l'approche "par les terres" est qu'elle permet de repérer de manière plus facile les exploitations agricoles concernées.

Pour extrapoler les résultats observés sur les exploitations agricoles enquêtées, on crée au niveau de chaque segment de l'échantillon une variable "artificielle" :

$$Z_i = \sum_{k=1}^{f_i} P_{ik} x_k$$

où pour le segment i, on a f_i exploitations agricoles enquêtées,
x_k est la valeur d'une variable (par exemple superficie en céréales) pour l'exploitation k,
$P_{ik} = \frac{S_{ik}}{S_k}$ est la part de la superficie totale de l'exploitation k qui est comprise dans le segment i.

Si l'on prend comme variable X la superficie en céréales, on calcule donc une "pseudo-superficie" en céréales pour le segment i, qui n'a aucune raison d'être la superficie en céréales effective du segment ; cette "pseudo-superficie" est obtenue en pondérant chaque questionnaire rempli auprès d'une exploitation agricole.

On peut alors, à partir du plan de sondage qui a abouti au choix de l'échantillon de segments, l'extrapoler à l'univers pour estimer le total de la variable Z (par exemple si l'échantillon de segments est aléatoire simple, on calculera $\frac{N}{n}\sum_{i=1}^{n} Z_i$).

Le total de Z étant égal au total de X[30], on a donc une estimation du total de X. Ce raisonnement s'applique à tous les types de variable, et on peut donc là aussi obtenir des résultats sur les caractéristiques socio-économiques des exploitations agricoles. Cette méthode est notamment utilisée au Canada (Julien et Maranda, 1989).

f) La question de la taille de l'échantillon

C'est une question fondamentale à résoudre concernant la mise au point du plan de sondage : cette taille résultera bien sûr d'un niveau de précision souhaité, mais aussi, souvent, de contraintes de budget. Dans la mesure du possible, on essaiera de tenir compte de l'expérience des enquêtes passées pour estimer *a priori* la taille de l'échantillon nécessaire à l'obtention de la précision voulue. Le statisticien travaille, sur ce point, plus par approximations successives que par application "brute" d'une théorie ; de plus, des objectifs multiples pour une enquête peuvent conduire à chercher une précision minimum pour chaque variable objectif, et donc à déterminer la taille d'échantillon comme la valeur supérieure des tailles minimum correspondantes.

2. Retour sur les problèmes liés à la base de sondage

a) Cas des enquêtes démographiques dans les pays en développement

On l'a vu, la base de sondage est souvent, pour ces pays, aréolaire au niveau du premier degré de tirage. Il faut, pour les unités primaires tirées, procéder à un dénombrement exhaustif des ménages y résidant : ceci veut dire que les limites de chaque unité primaire sont faciles à identifier sur le terrain car, même si on dispose d'une liste des habitations incluses dans l'unité primaire au dernier recensement, il a pu s'y construire de nouvelles habitations depuis (en particulier dans les zones d'habitat spontané). Par ailleurs, utiliser la liste des ménages recensés (au dernier recensement de la population) dans l'unité primaire ne se justifie que si le recensement est très récent (un à trois mois).

Par ailleurs, on pourrait imaginer d'établir la liste, d'échantillonner et d'interroger les unités en une seule opération de terrain : ceci est à éviter,

---

[30] En effet, $\sum_{\alpha=1}^{N} Z_\alpha = \sum_\alpha \sum_k \frac{S_{\alpha k}}{S_k} X_k = \sum_k X_k \left(\sum_\alpha \frac{S_{\alpha k}}{S_k} \right) = \sum_k X_k$

l'expérience montre qu'il vaut mieux séparer l'opération "dénombrement" et l'opération "enquête".

Au niveau de la constitution des unités primaires (ou, pour un tirage à plus de deux degrés, au niveau des unités à l'intérieur desquelles on procède au tirage de ménages), on a vu au chapitre 5 qu'il existe une taille "conseillée" pour ces unités : 300 à 500 personnes. On peut partir de zones de dénombrement du recensement et s'il le faut, les découper en segments pour arriver à cette taille[31] (encadré 3).

Quand on travaille sur une zone rurale, on peut utiliser les villages comme unités primaires : dans ce cas on sera conduit à regrouper certains villages (si possible voisins géographiquement), ou à en découper d'autres pour s'approcher de cette taille "optimale". Mais on aura souvent intérêt, en zone rurale, à utiliser plutôt les zones de dénombrement du recensement de la population, qui sont cartographiées de manière précise et homogènes en taille (pour les villages, on peut avoir des problèmes de délimitation exacte).

b) La mise à jour de la base de sondage

Une base de sondage vieillit, et doit être mise à jour (Brion, 1995) : s'il s'agit d'une base de sondage aréolaire, de nouvelles constructions peuvent exister, certaines peuvent avoir été détruites... Certaines limites d'unités aréolaires qui ne posaient pas de problème auparavant peuvent demander à être précisées lors de la mise à jour, quand la localisation d'une nouvelle habitation est ambiguë par rapport à la délimitation de l'unité. L'entretien d'une base de sondage doit être un service régulier d'un institut de statistique (suivi des limites des unités, suivi des effectifs si possible), afin de ne pas être obligé de travailler dans l'urgence lors de la mise en place d'une nouvelle enquête.

Si la base de sondage est une liste, on aura intérêt à y intégrer des informations "exogènes" concernant les unités (créations, disparitions, modifications ; par exemple, dans certains pays, la base de sondage des enquêtes ménages est la liste des logements issue du recensement de la population, complétée chaque année par la liste des logements neufs). Il arrive que, dans une base de sondage, on trouve des informations datées d'époques différentes (certaines datent par exemple du dernier recensement, d'autres sont très récentes) : ceci n'est pas gênant à partir du moment où la base de sondage est un outil pour tirer des échantillons (par exemple en stratifiant), et non un fichier qu'on utilise pour sortir des statistiques.

---

[31] Parfois, quand les zones de dénombrement du recensement sont voisines de cette taille, on découpe les plus grosses et on regroupe les plus petites pour arriver à une taille à peu près homogène.

On voit, à partir des considérations précédentes, l'importance que peut avoir un recensement (de la population en particulier) en tant que base de sondage pour un certain nombre d'enquêtes, et la complémentarité entre recensement et enquêtes par sondage.

c) La nécessité d'adapter la base de sondage au domaine d'étude

Pour l'étude de certains domaines, il faut parfois faire preuve d'imagination afin de mettre en place une base de sondage d'un type différent de celles utilisées habituellement. Ainsi, pour enquêter les populations d'éleveurs nomades, une base de sondage proposée est la liste des points d'eau du pays. On tire un échantillon de points d'eau où on procédera à l'enquête en tenant compte des deux remarques suivantes :

- la période d'enquête doit être suffisamment longue pour qu'on soit assuré que, durant celle-ci, les hommes et les animaux auront visité au moins un point d'eau ;
- il peut y avoir des doubles comptages : une solution peut être de ne compter que les "premières visites" à l'un des points d'eau échantillonnés durant la période d'enquête.

Cette méthode n'est pas la seule applicable au cas des enquêtes de populations nomades : certaines enquêtes font appel par exemple à un échantillon de marchés (Kidane, 1995 ; CEPED, 1988).

Par ailleurs, il ne faut pas hésiter à utiliser une base de sondage composite (Brion, 1995). On peut, pour certaines enquêtes sur l'élevage par exemple, utiliser une base aréolaire (tirage de zones puis d'unités d'élevage dans ces zones) complétée par un fichier de gros éleveurs (disponible par ailleurs) : les gros éleveurs devront, dans ce cas, être exclus des zones tirées au premier degré.

d) L'utilisation de la télédétection

On peut mentionner l'utilisation d'images satellites pour élaborer le plan de sondage d'enquêtes démographiques en milieu urbain. L'image satellite, base de sondage, possède deux avantages majeurs : récente, elle permet de délimiter correctement la zone urbaine étudiée (ce qui est très important dans le cas de villes à rythme élevé de croissance démographique) ; ensuite, on peut s'en servir pour stratifier la zone (l'information morphologique contenue sur l'image étant liée aux caractéristiques démographiques des différents quartiers) et procéder à un tirage

d'îlots à l'intérieur des strates définies. On voit que c'est donc au niveau de la "cartographie" que se situe l'apport de la télédétection.

Des expériences récentes ont montré les possibilités d'application de cette technique pour les enquêtes socio-démographiques en milieu urbain (Barbary, Dureau, 1991 ; Cogneau, Roubaud, 1992). Pour les enquêtes agricoles, la télédétection peut être utilisée comme information auxiliaire exhaustive pour produire des estimations de superficies (encadré 4), ou comme support pour établir une stratification du territoire en zones agro-écologiques homogènes ; à l'intérieur de chaque strate, on procédera alors à un sondage classique à deux degrés.

3. La nécessité d'un travail soigné à tous les niveaux

a) Au niveau de la collecte

Les erreurs autres que d'échantillonnage ont été évoquées au chapitre 1. Elles recouvrent différents types de problèmes :

- erreurs de couverture liées à un "décalage" entre la base de sondage et le champ de l'enquête ;
- non-réponses (chapitre 6) ;
- erreurs d'observation liées à la mauvaise compréhension du questionnaire (en particulier si celui-ci a été mal construit ou mal formulé) ;
- erreurs au niveau de la saisie ou de la codification des données.

On rappelle ici, car l'enjeu est important, la nécessité d'une formation sérieuse des enquêteurs, de consignes claires (en plus bien sûr d'un questionnaire, lui aussi, ne donnant pas lieu à des possibilités d'interprétation plus ou moins variables), d'un suivi du travail de collecte par une équipe de contrôleurs expérimentés (la "règle" d'un contrôleur pour quatre à cinq enquêteurs est une garantie d'un travail correctement suivi).

b) Au niveau du traitement des données

Un autre élément de la "chaîne" montre souvent des signes de défaillance. Il s'agit de la phase de traitement des données. Là aussi, une utilisation maîtrisée des

logiciels de dépouillement s'impose, en particulier relativement à l'application des formules utilisant les pondérations[32] (coefficients d'extrapolation).

Ces pondérations, rappelons-le, résultent du plan de sondage initial, puis des traitements opérés pour tenir compte de la non-réponse et des autres redressements intégrant de l'information auxiliaire. Pour la non-réponse, il faut bien distinguer les méthodes de repondération proposées pour la non-réponse globale des méthodes d'imputation proposées pour la non-réponse partielle : l'application de la repondération n'est pas possible en pratique pour la non-réponse partielle, car elle conduit à des jeux de pondération différents selon les variables, et à une impossibilité de sortir des tableaux croisant ces variables.

Un problème fondamental, au niveau du traitement des données, est relatif au suivi des unités, en particulier pour les situations suivantes :

- les changements de strate : si une unité est constatée, à l'enquête, comme ayant des caractéristiques la "rattachant" à une strate différente de celle dans laquelle elle avait été classée au moment du tirage, elle ne doit pas être pourvue du coefficient de pondération de cette nouvelle strate ; en effet, le coefficient de pondération est lié à la manière dont s'est déroulé le tirage de l'échantillon, et donc pas à la nouvelle strate. Par contre, au niveau de la publication des résultats, si ceux-ci utilisent les strates comme critères de ventilation, l'unité sera, bien sûr, classée selon les caractéristiques observées à l'enquête ;
- les déménagements des ménages : souvent, on utilise l'unité d'habitation comme référence ; on ne cherche donc pas, dans le cas où le ménage qui habitait le logement tiré dans la base de sondage a déménagé, à "retrouver" ce ménage ; on enquête le nouveau ménage occupant ;
- les logements vides : si un logement est constaté vide à l'enquête, on ne le remplace pas *a priori* pour remplir un questionnaire "à tout prix" ; en effet, l'unité "logement vide" est elle-même porteuse d'une information, et représentative d'un certain nombre de logements se trouvant dans le même cas.

c) Au niveau de la documentation des différentes phases de l'enquête

Nombre d'erreurs auraient pu être évitées, ou corrigées, si l'on avait pu reconstituer la manière dont s'est déroulé l'ensemble des phases de l'enquête. L'expérience montre malheureusement que la documentation correspondante est trop souvent absente : ceci se révèle particulièrement pénalisant au moment de la phase de redressement avant la sortie des résultats.

---

[32] Il n'est, malheureusement, pas rare de rencontrer des résultats d'enquête qui ont été calculés en "oubliant" les pondérations...

d) Au niveau de la publication des résultats

Publier la méthodologie de l'enquête est d'une grande aide pour l'utilisateur des résultats, mais aussi pour les personnes qui, quelques années plus tard, mettront en place une opération analogue.

Les variances d'estimateurs relatives aux grandeurs les plus importantes de l'enquête doivent être calculées, dans la mesure du possible (encadré 5), et publiées. Par ailleurs, on évitera de publier des tableaux où chaque case a été estimée à partir d'un nombre d'unités enquêtées trop faible (par exemple, moins de 20 unités) ; si, cependant, on publie des tableaux comportant de telles cases, on s'efforcera de placer un signe d'avertissement dans les cases concernées.

Encadré 5

Une méthode d'estimation de la variance d'estimation dans le cas de statistiques complexes : la linéarisation
(Deville et Roth, 1986)

On a vu dans ce manuel comment estimer la variance d'estimation d'un total, en fonction de la méthode de sondage utilisée. Cependant certaines statistiques sont plus complexes, et se présentent sous la forme d'une fonction de plusieurs totaux : $F(X_1,\ldots X_M)$.

La méthode proposée repose sur le développement de Taylor et consiste à passer par une variable linéarisée :

$$Z=\sum_{i=1}^{n}\frac{dF}{dX_i}(\hat{X})\hat{X}_i$$

On estime la variance de l'estimation de la statistique complexe par la variance de la variable linéarisée Z.

Exemple : cas d'un ratio

$F(X_1,X_2)=\frac{X_1}{X_2}$ où X_1 et X_2 sont deux totaux sur l'univers

$$\frac{dF}{dX_1}=\frac{1}{X_2}\quad\frac{dF}{dX_2}=-\frac{X_1}{X_2^2}$$

Pour chaque unité j de l'échantillon, on calcule la variable $Z_j=\frac{1}{\hat{X}_2}X_1j-\frac{\hat{X}_1}{\hat{X}_2^2}X_2j$, où $\hat{X}_1$ et $\hat{X}_2$ sont les estimations à partir de l'échantillon des totaux X_1 et X_2.

La variance de l'estimation du total de 2 (que l'on sait estimer à partir du plan de sondage) fournit une estimation de la variance de l'estimation du ratio.

Encadré 6

Quelques rappels ou conseils fondamentaux
(la liste n'est sans doute pas exhaustive)

- utiliser les pondérations pour calculer les estimations ;

- contrôler, dès les premières sorties de résultats, que les estimations des grandes masses (effectif total de la population, ...) sont conformes à ce qu'on connaît ;

- contrôler sur l'échantillon les distributions des variables d'étude (afin, en particulier, de repérer les erreurs de saisie) ;

- calculer les variances d'estimation pour les principales variables d'étude, afin d'avoir une idée de la précision des résultats (à ce sujet, ne pas confondre variance de l'estimation et variance de la variable ...) ;

- utiliser le plus possible les sources complémentaires d'information, que ce soit pour le tirage, le contrôle ou le redressement de l'échantillon ;

- éviter de donner l'illusion d'une précision parfaite en fournissant des résultats "à la décimale près" ; arrondir plutôt les résultats.

ANNEXE 1

BIOGRAPHIE DE RÉMY CLAIRIN

Rémy Clairin est né le 13 décembre 1923 à Tourcoing (Nord). En 1940, il entre à l'École supérieure des industries chimiques de Nancy puis, en 1943, rejoint le Maroc après avoir été interné en Espagne, s'engage dans l'armée de l'air et sert au Maroc, en Algérie, aux États-Unis et en Italie. De 1946 à 1952, il travaille au sein des services français d'occupation en Allemagne, à Berlin, Mayence puis Francfort.

C'est en 1954 que commence sa carrière de statisticien-démographe à laquelle il se consacrera totalement, après deux années de formation comme élève-administrateur de l'INSEE. C'est d'abord la Guinée (1954-1956) où il participe à la première grande enquête démographique lancée par l'INSEE en Afrique, puis le Mali (alors Soudan français), de 1956 à 1958, où il dirige la Mission socio-économique du Soudan dans le delta central du Niger.

De 1958 à 1959, il revient à Paris au ministère de la Communauté puis de la Coopération, et est envoyé ensuite (1959-1961) au Burkina (alors Haute-Volta) pour y créer le service statistique et concevoir et réaliser l'enquête démographique nationale, dans laquelle il met l'accent sur l'étude des migrations : cette enquête fait encore référence dans le domaine.

En 1961, il rejoint le service Coopération de l'INSEE où il reste jusqu'en 1966, à l'exception d'un stage à l'*Office of Population Research* de l'université de Princeton en 1962. À l'INSEE il travaille sur l'analyse de l'enquête de Haute-Volta, puis sur la préparation de celles du Niger, de Mauritanie et du Cameroun occidental.

Il passe ensuite un an, pour le compte des Nations unies, au *Cairo Demographic Center* (1966-1967) et regagne l'INSEE pour un an à la division de la démographie (1967-1968) avant d'être affecté au ministère des Départements et territoires d'outre-mer, où il prépare et organise divers recensements : Polynésie française (1972), Port-Vila et Santo (anciennes Nouvelles-Hébrides) (1972), puis l'ensemble des départements d'outre-mer (1973-1974).

Durant cette période, il assure une charge d'enseignement à l'Institut de démographie de l'université de Paris (IDUP), notamment le cours sur les méthodes d'ajustement des statistiques démographiques imparfaites, ce qui lui permettra de rédiger un manuel devenu célèbre, publié par le Groupe de démographie africaine auquel il participe de façon suivie.

Il retourne à nouveau à l'INSEE en 1974 où il prend part à de nombreux travaux réalisés dans des cadres divers :

- diverses études de synthèse et travaux méthodologiques (Groupe de démographie africaine) ;
- élaboration des nouvelles tables-type de mortalité (OCDE) ;
- étude des migrations (Banque mondiale, OCDE, UIESP) ;
- nombreuses missions sur le terrain, d'appui technique ou d'évaluation de programmes (Côte d'Ivoire, Cameroun,...) ;
- direction du stage de recyclage pour les démographes africains sur l'observation démographique (Bordeaux, 1982).

Enfin, depuis le début 1987, il était affecté par l'INSEE au CEPED alors en voie de constitution, où il prenait une part active à son lancement, tant sur les plans matériel et organisationnel que sur le plan scientifique. Ses travaux reconnus de chercheur ne l'ont jamais détourné des réalités du terrain, ce qui n'était pas la moindre de ses qualités.

Rémy Clairin nous a quittés en pleine activité le 12 octobre 1987.

Il est difficile de faire une sélection parmi les nombreuses contributions de Rémy Clairin, risquons-nous toutefois à conseiller au lecteur trois ouvrages particulièrement utiles :

- *Ajustement des données imparfaites*, Paris, GDA, 1973, 184 p.
- *Manuel sur les méthodes d'estimation des statistiques démographiques imparfaites dans les pays en développement*, OCDE, Centre de développement, 1986, 265 p. (en collaboration avec Julien Condé).
- *De l'homme au chiffre*, Paris, CEPED, 1988, 329 p. (Les Études du CEPED, n° 1). (Rémy Clairin et Louis Lohlé-Tart (éds.), avec la collaboration de Michel François et Francis Gendreau).

ANNEXE 2

DÉVELOPPEMENT DES SIGLES UTILISÉS

AEF	Afrique équatoriale française
AFSA	Association africaine de statistique
AISO	Association internationale pour les statistiques officielles
AOF	Afrique occidentale française
CEA	Commission économique pour l'Afrique (commission régionale des Nations unies), Addis Abeba
CEPED	Centre français sur la population et le développement, Paris
DEFF	*effet de sondage* (*design effect*, en anglais)
DIAL	Développement des investigations sur l'ajustement à long terme (groupement d'intérêt scientifique), Paris
DSA	dimension sociale de l'ajustement en Afrique sub-saharienne (programme de la Banque Mondiale)
DSCN	Direction de la statistique et de la comptabilité nationale, Yaoundé
EDS	enquête démographique et de santé (DHS, en anglais)
EHESS	École des hautes études en sciences sociales, Paris
FAO	Organisation des Nations unies pour l'alimentation et l'agriculture, Rome
FNUAP	Fonds des Nations Unies pour la population
GDA	Groupe de démographie africaine, Paris
GDD	Groupe de démographie du développement, Paris
IDUP	Institut de démographie de l'université de Paris
IIS	Institut international de statistique, Voorburg (Pays-Bas) (ISI, en anglais)
INED	Institut national d'études démographiques, Paris

INSEE	Institut national de la statistique et des études économiques, Paris
ISI	*International Statistical Institute*, Voorburg (Pays-Bas) (IIS, en français)
OCDE	Organisation de coopération et de développement économique, Paris
ORSTOM	Institut français de recherche scientifique pour le développement en coopération, Paris
PUF	Presses universitaires de France, Paris
SCEES	Service central des enquêtes et études statistiques (ministère français de l'Agriculture), Paris
SEAE	secrétariat d'État aux Affaires étrangères chargé de la coopération, Paris
UIESP	Union internationale pour l'étude scientifique de la population, Liège
UNESCO	Organisation des Nations unies pour l'éducation, la science et la culture, Paris

LISTE DES TABLEAUX ET ENCADRÉS

LISTE DES FIGURES

RÉFÉRENCES BIBLIOGRAPHIQUES

1) Ouvrages généraux

ARDILLY Pascal, 1994. – *Les techniques de sondage.* – Paris, Éditions Technip, 393 p.

ASSELIN Louis Marie, 1984. – *Techniques de sondage avec application à l'Afrique.* – Québec, Gaétan Morin.

COCHRAN William, 1977. – *Sampling techniques.* – New York, Wiley, 428 p.

DESABIE Jacques, 1971. – *Théorie et pratique des sondages.* – Paris, Dunod, 483 p.

DUSSAIX Anne-Marie et GROSBRAS Jean-Marie, 1992. – *Exercices de sondage, avec aide-mémoire et solutions.* – Paris, Économica, 169 p.

DUSSAIX Anne-Marie et GROSBRAS Jean-Marie, 1993. – *Les sondages : principes et méthodes.* – Paris, PUF, 122 p. (Coll. Que sais-je ?), n° 701.

GROSBRAS Jean-Marie, 1987. – *Méthodes statistiques des sondages.* – Paris, Économica, 331 p.

KISH Leslie, 1965. – *Survey sampling.* – New-York, Wiley, 643 p.

SÄRNDAL Carl-Erik, SWENSSON Beugt et WRETNAM Jan, 1992. – *Model assisted survey sampling.* – New York, Springer-Verlag, 694 p.

2) Articles ou ouvrages relatifs aux enquêtes démographiques dans les pays en développement

CEA-UNESCO, 1974. – *Manuel des enquêtes démographiques par sondage en Afrique.* – Addis Abeba, 279 p.

CEPED, 1988. – *De l'homme au chiffre. Réflexions sur l'observation démographique en Afrique.* – Paris, CEPED, 329 p. (Les Études du CEPED, n° 1).

CLAIRIN Rémy, 1978. – "Plan de sondage de l'enquête démographique à passages répétés en Côte d'Ivoire", *STATÉCO*, n° 16, p. 63-103.

CLAIRIN Rémy, 1983. – *L'estimation de la précision de certains paramètres démographiques obtenus à partir d'une enquête par sondage.* – Paris, Groupe de démographie du développement, 62 p. (Études et documents du GDD, n° 11).

CLELAND John et SCOTT Christopher (éd.), 1987. – *The world fertility survey : an assessment.* – New-York, Oxford University Press, 1049 p.

EL GHAZALI Abdelaziz, 1989. – "L'enquête démographique à passages répétés du Maroc (1986-1988), méthodologie et organisation", *STATÉCO*, n° 57, p. 85-106.

INED-INSEE-ORSTOM-SEAE, 1973. – *Sources et analyse des données démographiques : application à l'Afrique d'expression française et à Madagascar.* – Paris, 3 vol., 414 p. + 183 p. + 475 p.

NATIONS UNIES, 1971. – "Methodology of demographic sample surveys", *Statistical Papers*, Série M, n° 51. (Communication présentée à *l'International workshop on methodology of sample surveys*, Copenhague, 24 septembre – 3 octobre 1969).

NATIONS UNIES, 1962. – *Manuel sommaire des méthodes de sondage, volume 1 : éléments de la théorie des enquêtes par sondage.* – New York, 235 p. (Études méthodologiques, Série F, n° 9, rév. 1).

NATIONS UNIES, 1992. – *Les enquêtes de suivi pour la mesure de la fécondité, de la mortalité et de la migration.* – New York, 168 p. (Études méthodologiques, Série F, n° 41).

ORSTOM-INSEE-INED, 1971. – *Les enquêtes démographiques à passages répétés : application à l'Afrique d'expression française et à Madagascar.* – Paris, 290 p.

SCOTT Christopher, 1967. – "Sampling for demographic and morbidity surveys in Africa", *Review of the ISI*, vol. 35, n° 2, p. 155-171.

SCOTT Christopher, 1987. – *Demographic and health surveys : sampling manual.* – Columbia, États-Unis, Institute for Resource Development, 68 p.

VERMA Vijay et LÊ Thanh, 1996 – "An analysis of sampling errors for the demographic and health surveys". *International Statistical Review*, vol. 64, n° 3, p. 265-294.

VERMA Vijay, SCOTT Christopher et O'MUIRCHEARTAIGH Colm, 1980. – "Sample designs and sampling errors for the world fertility survey", *Journal of the Royal Statistical Society*, vol. 143, n° 4, Série A, p. 431-473.

3) Articles ou ouvrages relatifs à d'autres enquêtes

ABZAHD Mohamed, 1982. – "Le dispositif d'enquêtes périodiques sur l'emploi au Maroc", *STATÉCO*, n° 29, p. 59-80.

ADJIKOUN Justin, BABUT Eric et WADAGNI Nestor, 1986. – "Méthodologie de l'enquête budget-consommation de la République Populaire du Bénin", *STATÉCO*, n° 47, p. 41-64.

BAHILI Jean et BAKARY Djaby, 1993. – "L'enquête nationale sur les effectifs du cheptel du Burkina Faso", *STATÉCO*, n° 73, p. 49-62.

BANQUE MONDIALE, 1991. – *The social dimensions of adjustment priority survey*, Washington, 180 p. (Working Paper DSA n° 12).

BANQUE MONDIALE, 1992. – *The social dimensions of adjustment integrated survey*, Washington, 207 p. (Working Paper DSA n° 14).

BRILLEAU Alain, 1993. – "Les enquêtes agricoles dans les pays sahéliens", *STATÉCO*, n° 73, p. 5-24.

DAHO Bakary, CHIA BLE Kouakou et OUATTARA Idrissa, 1985. – "L'enquête permanente auprès des ménages de Côte d'Ivoire, présentation générale", *STATÉCO*, n° 44, p. 39-60.

DIAL-DSCN, 1994. – "L'enquête 1-2-3 sur l'emploi et le secteur informel à Yaoundé", *STATÉCO*, n° 78, p. 1-141.

FOURNIER Philippe, 1986. – *Enquête sur l'utilisation du territoire effectuée en 1985 par la méthode des segments.* – Paris, SCEES, 63 p. (Série S, n° 13).

JULIEN C. et MARANDA F., 1989 – *Le remaniement du plan de sondage de l'enquête nationale sur les fermes de 1988.* – Ottawa, Statistique Canada, 22 p. (Cahier de Travail n° BSMD-89-012F).

KISH Leslie, 1994. – *Méthodes de sondage pour les enquêtes statistiques agricoles.* – Rome, FAO, 388 p. (Coll. Développement statistique).

NGASSAM André, 1986. – "L'enquête budget-consommation du Cameroun de 1983-1984. Présentation générale et description de la technique de mise à jour de la base de sondage", *STATÉCO*, n° 47, p. 23-40.

ROY Gildas, 1984. – "Enquête nationale budget-consommation du Rwanda, plan de sondage, estimation, erreur d'échantillonage", *STATÉCO*, n° 38, p. 58-88.

4) Sur la pratique des sondages (organisation, questionnaires, base de sondage, etc)

ALIAGA Alfredo, 1995. – *Developments in small area estimation and major issues confronted.* – Addis Abeba, 16 p. (Communication présentée à la conférence conjointe AISO-AFSA, 22-24 mai 1995).

AYED Mohamed, 1982. – "Pour la constitution d'une base de sondage permanente pour les enquêtes auprès des ménages ; le cas de la Tunisie", *STATÉCO*, n° 29, p. 81-98.

BLAIZEAU Didier et DUBOIS Jean-Luc, 1989. – *Connaître les conditions de vie des ménages dans les pays en développement.* – Paris, Ministère de la Coopération et du développement, 3 vol., 165 p. + 312 p. + 175 p.

BRION Philippe, 1995. – "Base de sondage : entre rigueur et bricolage", *in* : VALLIN Jacques (éd.), *Clins d'œil de démographes à l'Afrique et à Michel François*, p. 117-124, – Paris, CEPED, 244 p. (Documents et manuels du CEPED, n° 2).

DESTANDAU Sophie, 1996. – *Estimation sur des petits domaines. Application à l'enquête éducation 92.* – Paris, INSEE, 48 p. (Communication présentée aux Journées de méthodologie statistique de l'INSEE, Paris, 11-12 décembre 1996).

DEVILLE Jean-Claude et ROTH Nicole, 1986. – "La précision des enquêtes sur l'emploi", *Économie et Statistique*, n° 193-194, p. 127-134.

JACQUART Hugues, 1988. – *Qui ? Quoi ? Comment ? ou la pratique des sondages.* – Paris, Éditions Eyrolles, 307 p.

KIDANE Asmeron, 1995. – *Need for data that ignore boundaries, development in small area estimation and the problems of protecting confidentiality.* – Addis Abeba, 9 p. (Communication présentée à la conférence conjointe AISO-AFSA, 22-24 mai 1995).

5) Sur l'utilisation de la télédétection

BARBARY Olivier et DUREAU Françoise, 1991. – "L'enquête par sondage sur image satellite : une solution pour améliorer l'observation des populations citadines", *STATÉCO*, n° 67, p. 63-100.

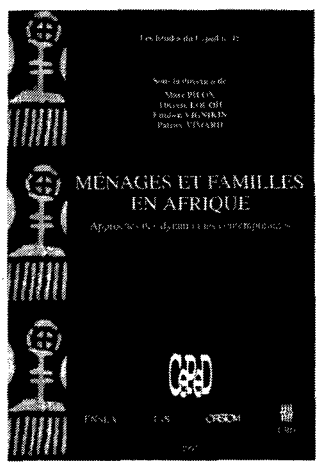
COGNEAU Denis et ROUBAUD François, 1992. – "Utilisation de la télédétection pour l'élaboration du plan de sondage d'une enquête sur le secteur informel : le cas de Yaoundé", *STATÉCO*, n° 71, p. 5-26.

DUREAU Françoise, BARBARY Olivier, MICHEL Alain et LORTIC Bernard, 1989. –*Sondages aréolaires sur image satellite pour des enquêtes socio-démographiques en milieu urbain.* – Paris, Éditions de l'ORSTOM, 8 + 15 p. (Manuel de formation).

PASTORELLI Robert, 1992. – "Superficies agricoles à partir d'images satellite", *Courrier des Statistiques*, n° 61-62, p. 47-52.

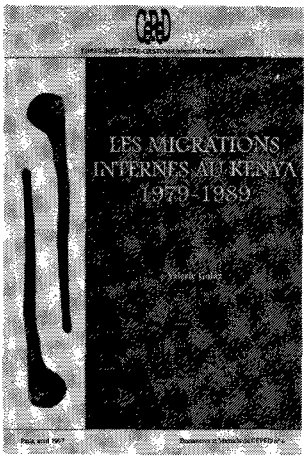
LES PUBLICATIONS DU CEPED

Collection *Les Études du CEPED*



- n°15 : *Ménages et familles en Afrique. Approche des dynamiques contemporaines*, sous la direction de Marc PILON, Thérèse LOCOH, Émilien VIGNIKIN et Patrice VIMARD (éds.) (1997), 424 p. (*151,66 F HT*, **160 F TTC,** frais de port 27 F).
- n°14 : *Permanences et changements de l'Afrique rurale. Dynamiques familiales chez les Bwa du Mali*, par Véronique HERTRICH (1996), 570 p. (*170,62 F HT*, **180 F TTC**, frais de port 36 F).
- n°13 : *Crise et population en Afrique. Crises économiques, politiques d'ajustement et dynamiques démographiques*, par Jean COUSSY et Jacques VALLIN (dir.) (1996), 580 p. (*170,62 F HT*, **180 F TTC**, frais de port 36 F).
- n°12 : *Sauver les enfants : le rôle des vaccinations*, par Annabel DESGRÉES DU LOÛ (1996), avec la collaboration du Muséum national d'histoire naturelle, 261 p. (*94,79 F HT*, **100 F TTC**, frais de port 27 F).
- n°11 : *L'économie algérienne à l'épreuve de la démographie*, par Lhaocine AOURAGH (1996), 337 p. (*94,79 F HT*, **100 F TTC**, frais de port 27 F).
- n°10 : *Les conséquences démographiques du sida en Abidjan : 1986-1992,* par Michel GARENNE, Maria MADISON, Daniel TARANTOLA, Benjamin ZANOU, Joseph AKA et Raymond DOGORÉ (1995), 198 p. (*94,79 F HT*, **100 F TTC**, frais de port 16 F).
- n° 9 : *La maternité chez les Bijago de Guinée Bissau,* par Alexandra DE SOUSA et Dominique WALTISPERGER (collab.) (1995), 114 p. (*94,79 F HT*, **100 F TTC**, frais de port 16 F).
- n° 8 : *La crise de l'asile politique en France*, par Luc LEGOUX (1995), 344 p. (*94,79 F HT*, **100 F TTC**, frais de port 27 F).
- n° 7 : *L'entrée en vie féconde. Expression démographique des mutations socio-économiques d'un milieu rural sénégalais*, par Valérie DELAUNAY (1994), 326 p. (*85,31 F HT*, **90 F TTC**, frais de port 27 F).
- n° 6 : *La traite des esclaves au Gabon du XVII^e^ au XIX^e^ siècle, essai de quantification pour le XVIII^e^ siècle*, par Nathalie PICARD-TORTORICI et Michel FRANÇOIS (1993), 156 p. (épuisé).
- n° 5 : *Croissance urbaine, migrations et population au Bénin*, par Julien GUINGNIDO GAYE (1992), 114 p. (*94,79 F HT*, **100 F TTC**, frais de port 16 F).
- n° 4 : *Un siècle de démographie tamoule*, par Christophe GUILMOTO (1992), 175 p. (*113,74 F HT*, **120 F TTC**, frais de port 16 F).
- n° 3 : *Mobilité spatiale et mobilité professionnelle dans la région nord-andine de l'Équateur*, par Jean PAPAIL (1991), 87 p. (*75,83 F HT*, **80 F TTC**, frais de port 16 F).

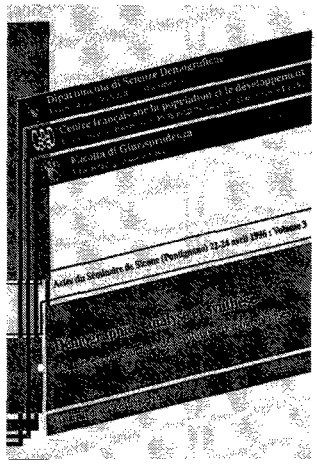
- n° 2 : *Mortal, logiciel d'analyse de la mortalité*, par Jean-Michel COSTES et Dominique WALTISPERGER (1988), 99 p. + disquette (épuisé).
- n° 1 : *De l'homme au chiffre, réflexions sur l'observation démographique en Afrique*, édité par Louis LOHLÉ-TART et Rémy CLAIRIN (1988), 329 p. (*142,18 F HT*, **150 F TTC**, frais de port 27 F).



Collection *Documents et Manuels du CEPED*

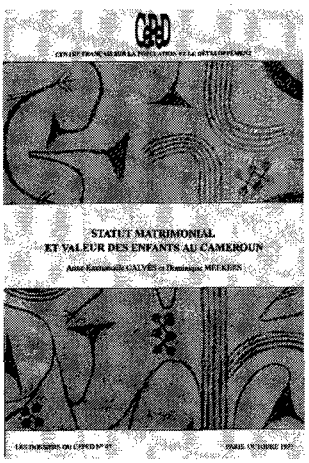
- n° 6 : *Les migrations internes au Kenya 1979-1989*, par Valérie GOLAZ (1997), 126 p. (*94,79 F HT*, **100 F TTC**, frais de port 16 F).
- n° 5 : *Genre et développement : des pistes à suivre*, édité par Thérèse LOCOH, Annie LABOURIE-RACAPÉ et Christine TICHIT (1996), 154 p. (*94,79 F HT*, **100 F TTC**, frais de port 16 F).
- n° 4 : *L'analyse des enquêtes biographiques à l'aide du logiciel STATA*, par Philippe BOCQUIER (1996), 208 p. (*113,74 F HT*, **120 F TTC**, frais de port 16 F) + disquette.
- n° 3 : *Manuel de sondages. Applications aux pays en développement*, par Rémy CLAIRIN et Philippe BRION (1997), 108 p. (*75,83 F HT*, **80 F TTC**, frais de port 16 F). (2e édition).
- n° 2 : *Clins d'œil de démographes à l'Afrique et à Michel François*, édité par Jacques VALLIN (1995), 244 p. (épuisé).
- n° 1 : *La démographie de 30 États d'Afrique et de l'Océan Indien*, CEPED (1994), 352 p. (épuisé).

Coéditions



- *Démographie : analyse et synthèse. Causes et conséquences des évolutions démographiques,* CEPED/DSD/FACOLTÀ DI GIURISPRUDENZA, 276 p. (Actes du colloque de Sienne, 22-24 avril 1996, vol. 3) (*151,66 F HT*, **160 F TTC**, frais de port 27 F).
- *Démographie : analyse et synthèse. Causes et conséquences des évolutions démographiques,* CEPED/DSD, 408 p. (Actes du colloque de Sienne, 22-24 avril 1996, vol. 2) (*170,62 F HT*, **180 F TTC**, frais de port 36 F).
- *Populations et environnement dans les pays du Sud*, sous la direction de Francis GENDREAU, Patrick GUBRY et Jacques VÉRON, Karthala/CEPED, 308 p. (*151,66 F HT*, **160 F TTC**, frais de port 36 F).
- *Le retour au village. Une solution à la crise économique au Cameroun ?*, par Patrick GUBRY *et al.* (1996), CEPED/IFORD/MINREST/L'Harmattan, 206 p. (*113,74 F HT*, **120 F TTC**, frais de port 16 F).
- *Populations africaines et sida*, sous la direction de Jacques VALLIN (1994), CEPED/La Découverte, 218 p. (*141,23 F HT*, **149 F TTC**, frais de port 27 F).

- *La population de l'Afrique. Manuel de démographie*, par Francis GENDREAU (1993), CEPED/Karthala, 463 p. (*170,62 F HT*, **180 F TTC**, frais de port 36 F).
- *Politiques de développement et croissance démographique rapide en Afrique*, édité par Jean-Claude CHASTELAND, Jacques VÉRON et Magali BARBIERI (1993), 314 p. (INED/CEPED/PUF) (*170,62 F HT*, **180 F TTC**, frais de port 27 F).
- *Les spectres de Malthus, déséquilibres alimentaires, déséquilibres démographiques*, édité par Francis GENDREAU, Claude MEILLASSOUX, Bernard SCHLEMMER et Martin VERLET (1991), CEPED/EDI/ORSTOM, 444 p. (*218,01 F HT*, **230 F TTC**, frais de port 27 F).



Collection *Les Dossiers du CEPED* (*28,44 F HT*, **30 F TTC/ numéro,** frais de port 5 F) (gratuit pour les pays du Sud).

- n° 47 : *État matrimonial et valeur des enfants au Cameroun*, par Anne-Emmanuèle CALVÈS et Dominique MEEKERS (1997), 35 p. (traduction du *CEPED Series* n° 3).
- n° 46 : *Migrations et institutions au Sénégal : effets d'échelle et déterminants*, par Christophe Z. GUILMOTO (1997), 42 p.
- n° 45 : *L'émergence des migrations spontanées au Viêt-Nam. Le cas de Vung Tau et de Dong Nai*, par Mau Diep DOAN, Patrick GUBRY, Jerrold W. HUGUET et Khac Tham TRINH (1996), 48 p.
- n° 44 : *Politiques de population et baisse de la fécondité en Afrique sub-saharienne*, par Thérèse LOCOH et Yara MAKDESSI (1996), 47 p.
- n° 43 : *Essai d'utilisation des statistiques d'état civil et sanitaires dans l'analyse de la mortalité à Yaoundé*, par Samuel KÉLODJOUÉ (1996) , 43 p.
- n° 42 : *La polyandrie chez les Bashilele du Kasaï occidental (Zaïre) : fonctionnement et rôles*, par Séraphin NGONDO A PITSHANDENGE (1996), 20 p.
- n° 41 : *La régulation des naissances se généralise*, par Henri LERIDON et Laurent TOULEMON (1996), 19 p.
- n° 40 : *Ho Chi Minh Ville : de la migration à l'emploi*, par Truong SI ANH, Patrick GUBRY, Vu Ti HONG et Jerrold W. HUGUET (1996), 52 p.
- n° 39 : *La population de Cuba : principales caractéristiques et tendances démographiques*, par Sonia I. CATASUS CERVERA (1996), 35 p.
- n° 38 : *Effets de la guerre civile au Centre-Mozambique et évaluation d'une intervention de la Croix rouge*, par Michel GARENNE, Rudi CONINX et Chantal DUPUY (1996), 25 p.
- n° 37 : *Ressources économiques et comportements démographiques des ménages agricoles : le cas des Éwé du Sud-Togo*, par Kokou VIGNIKIN (1996), 35 p.
- n° 36 : *Structure de production et comportement procréateur en Côte d'Ivoire*, par Aka KOUAMÉ et Mburano RWENGE (1996), 31 p.
- n° 35 : *Les migrations comoriennes en France : histoire de migrations coutumières*, par Géraldine VIVIER (1996), 38 p.
- n° 34 : *La transition démographique. Trente ans de bouleversements (1965-1995)*, par Jean-Claude CHESNAIS (1994), 25 p. (2e tirage).

- n° 33 : *Pluralisme thérapeutique et stratégies de santé chez les Évhé du sud-est Togo*, par Nadia LOVELL (1995), 20 p.
- n° 32 : *Peut-on échapper à la polygamie à Dakar ?*, par Philippe ANTOINE et Jeanne NANITELAMIO (1995), 31 p. (2e tirage).
- n° 31 : *Familles africaines, population et qualité de la vie*, par Thérèse LOCOH (1995), 48 p. (3e tirage).
- n° 30 : *La mortalité dans le monde : tendances et perspectives*, par France MESLÉ et Jacques VALLIN (1995), 25 p. (3e tirage).
- n° 29 : *Planification sanitaire et ajustement structurel au Cameroun*, par Antoine KAMDOUM (1994), 40 p. (épuisé).
- n° 28 : *Migration et sida en Afrique de l'Ouest, un état des connaissances*, par Richard LALOU et Victor PICHÉ (1994), 52 p. (3e tirage).
- n° 27 : *Éducation de la mère et soins aux enfants à Ouagadougou*, par Christine OUEDRAOGO (1994), 37 p.
- n° 26 : *Réflexions sur l'avenir de la population mondiale*, par Jacques VALLIN (1994), 24 p. (4e tirage).
- n° 25 : *Facteurs de fécondité en milieu rural forestier ivoirien*, par KOFFI N'GUESSAN (1993), 40 p.
- n° 24 : *Les disparités régionales de la mortalité au Bénin*, par Martin LAOUROU (1993), 36 p.
- n° 23 : *Contribution à l'étude de l'évolution de la population de l'Afrique occidentale française 1904-1960*, par Raymond R. GERVAIS (1993), 50 p.
- n° 22 : *Solidarité dans la crise ou crise des solidarités familiales au Cameroun ?*, par Parfait Martial ÉLOUNDOU-ÉNYÉGUÉ (1992), 40 p. (épuisé).
- n° 21 : *La mortalité des enfants à Luanda*, par Maria Julia VAZ-GRAVE (1992), 39 p.
- n° 20 : *Mortalité maternelle : deux études communautaires en Guinée*, par Pierre CANTRELLE, Patrick THONNEAU et Boubacar TOURE (1992), 43 p.
- n° 19 : *Vingt ans de planification familiale en Afrique sub-saharienne*, par Thérèse LOCOH (1992), 27 p. (épuisé).
- n° 18 : *Les déterminants de la mortalité des enfants dans le tiers-monde*, par Magali BARBIERI (1991), 33 p. (épuisé).
- n° 17 : *La fécondité en Mauritanie*, par KEUMAYE IGNEGONGBA (1991), 39 p. (épuisé).
- n° 16 : *Dix problèmes de population en perspective - Hommage à Jean Bourgeois-Pichat et à Alfred Sauvy*, par Léon TABAH (1991), 31 p. (épuisé).
- n° 15 : *La mesure de l'infécondité et de la sous-fécondité*, par EVINA AKAM (1990), 39 p. (épuisé).
- n° 14 : *Statut de la femme, structure familiale, fécondité : transitions dans le golfe du Bénin*, par Laurent Mensan ASSOGBA (1988), 28 p. (épuisé).
- n° 13 : *Estimer la mortalité maternelle à l'aide de la méthode des sœurs*, par Véronique FILIPPI et Wendy GRAHAM (1990), 29 p. (épuisé).
- n° 12 : *La montée du célibat féminin dans les villes africaines. Trois cas : Pikine, Abidjan et Brazzaville*, par Philippe ANTOINE et Jeanne NANITELAMIO (1990), 27 p. (épuisé).
- n° 11 : *Deux études sur l'emploi dans le monde arabe*, par Jacques CHARMES (1990), 37 p. (épuisé).

- n° 10 : *Facteurs culturels et sociaux de la santé en Afrique de l'Ouest*, par Pierre CANTRELLE et Thérèse LOCOH (1990), 36 p. (épuisé).
- n° 9 : *Éléments du débat population-développement*, par Jacques VÉRON (1989), 48 p. (2e tirage). (épuisé).
- n° 8 : *Transformations agraires et mobilités de la main d'œuvre dans la région nord andine de l'Équateur*, par LE CHAU et Jean PAPAIL (1989), 18 p.
- n° 7 : *Prospective des déséquilibres mondiaux – Démographie et santé*, par Pierre CANTRELLE et Francis GENDREAU (1989), 33 p. (épuisé).
- n° 6 : *Les politiques de population en matière de fécondité dans les pays francophones : l'exemple du Togo*, par Thérèse LOCOH (1989), 20 p. (épuisé).
- n° 5 : *Rétention de la population et développement en milieu rural : à l'écoute des paysans Mafa des Monts Mandara (Cameroun)*, par Patrick GUBRY (1988), 24 p. (épuisé).
- n° 4 : *État et besoins de la recherche démographique dans la perspective des recommandations de la conférence de Mexico et de ses réunions préparatoires*, par Jean-Claude CHASTELAND (1988), 23 p. (épuisé).
- n° 3 : *La fécondité en Afrique noire : un progrès rapide des connaissances mais un avenir encore difficile à discerner*, par Thérèse LOCOH (1988), 26 p. (épuisé).
- n° 2 : *Politiques africaines en matière de fécondité : de nouvelles tendances*, par Patrick GUBRY et Mpembele SALA-DIAKANDA (1988), 50 p. (épuisé).
- n° 1 : *La connaissance des effectifs de population en Afrique : bilan et évaluation - Hommage à Rémy Clairin*, par Rémy CLAIRIN et Francis GENDREAU (1988), 35 p. (épuisé).

Collection *The CEPED Series* (*35,07 F HT*, **37 F TTC/numéro,** frais de port 5 F)

- n° 3 : *The advantages of having many children for women in formal and informal unions in Cameroon*, by Anne-Emmanuèle CALVÈS and Dominique MEEKERS, 38 p.
- n° 2 : *Population policies and fertility decline in Sub-saharan Africa*, by Thérèse LOCOH and Yara MAKDESSI, 43 p. (Translated from french by Fallon M. MOURSUND).
- n° 1 : *Mortality in the world : trends and prospects*, by France MESLÉ et Jacques VALLIN, 24 p. (Translated from french by Isabelle WALLERSTEIN).

Collection *Los Documentos del CEPED* (*35,07 F HT*, **37 F TTC/numéro,** frais de port 5 F)

- n° 1 : *La mortalidad en el mundo : tendencias y perspectivas*, para France MESLÉ y Jacques VALLIN, 24 p. (Traducido del francès para Maria Celina AÑAÑOS). (épuisé).

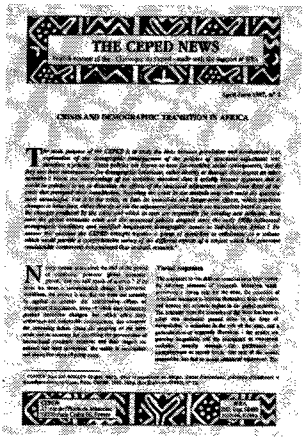
Collection *Données de base sur la population* (gratuit)
(dossiers réalisés par Nuria LOPEZ-ESCARTIN)

31 brochures réalisées.
Restent disponibles : Burkina Faso, Burundi, Cap Vert, Côte d'Ivoire, Gabon, Guinée, Guinée-Bissau, Mozambique, Niger, Nigéria, Sao Tome e Principe, Seychelles, Togo, Zaïre, Viêt-Nam.

La Chronique du CEPED, bulletin trimestriel de liaison du CEPED (26 numéros parus).
9,48 F HT ou **10 F TTC** le numéro. Abonnement : *28,44 F HT* ou **30 F TTC** par an. Diffusion gratuite dans les pays du Sud. (Pas de frais de port).

n° 26

Changements matrimoniaux en Afrique



n° 2

Crisis and demographic transition in Africa

The CEPED News, english version of the *Chronique du CEPED* made with the support of IFRA (10 F/number or subscription 30 F/year).

Imprimé en France par INSTAPRINT S.A.
1-2-3, levée de la Loire – LA RICHE – B.P. 5927 – 37059 TOURS Cedex 1
Tél. 02 47 38 16 04

Dépôt légal 4ème trimestre 1997

Rémy CLAIRIN (1923-1987) a commencé sa carrière de statisticien-démographe en 1954 après deux années de formation comme élève administrateur de l'INSEE. De 1954 à 1987, il a effectué de nombreux séjours en Afrique et dans les départements et territoires d'Outre-Mer avec pour mission de concevoir et réaliser divers recensements et enquêtes démographiques. Ses travaux méthodologiques, ses études de synthèse et autres missions d'évaluation ont marqué l'évolution de la démographie du développement. Depuis le début 1987, il était affecté par l'INSEE au CEPED, alors en voie de constitution.
Il est décédé le 12 octobre 1987.

Philippe BRION, ancien élève de l'École Polytechnique et de l'ENSAE (École nationale de la statistique et de l'administration économique), était le chef de la division « Études et méthodes statistiques pour le développement » de l'INSEE au moment de la rédaction de cet ouvrage. Il est à présent chef de la division « Agriculture ».

Le principe de la technique des sondages est le suivant : remplacement du tout par une partie, l'échantillon, pour produire de l'information sur un domaine étudié. Cet ouvrage présente les principes de base de la méthode, ainsi que quelques exemples de plans de sondages d'enquêtes réalisées dans les pays en développement.

Résultant d'un projet de manuel initié par Rémy Clairin et consacré aux enquêtes démographiques dans les pays africains, il aborde les considérations pratiques à prendre en compte lors de la mise en place et du traitement d'enquêtes par sondage dans les pays en développement, et se veut essentiellement à destination des praticiens travaillant dans ces pays.

CEPED
15, rue de l'École de Médecine
75270 PARIS Cedex 06
Téléphone : (33) 1 44 41 82 30
Télécopie : (33) 1 44 41 82 31

INSEE
Département des Relations
Internationales et de la Coopération
18, boulevard Adolphe Pinard
75675 PARIS Cedex 14
Téléphone : (33) 1 41 17 53 13
Télécopie : (33) 1 41 17 66 52

Couverture : poteau sculpté
Ngounié, Sud-Gabon
Détail relevé par Michel François
Maquette : CEPED

PRIX : 80 FF TTC