2

SONDAGES ET QUALITE DES DONNEES

Anne-Marie Dussaix

Ecole Supérieure des Sciences Economiques et Commerciales
Cergy-Pontoise

Jean-Marie Grosbras

Ecole Nationale de la Statistique et de l'Administration Economique
Paris

2.1 Introduction

Une enquête est une activité de collecte d'informations sur une population ; elle est basée sur des concepts, des méthodes et des procédures clairement définis. Conceptuellement on s'intéresse à des variables **Y**, définies pour chaque individu de la population, et on souhaite estimer des fonctions des valeurs individuelles Y_i : par exemple des totaux Y, des moyennes $\overline{Y}$ (une proportion est un cas particulier de moyenne), des variances, médianes, etc. On supposera toujours que la population est de taille finie.

Un sondage est une enquête menée auprès d'une fraction, plus ou moins importante, de la population. Il comprend donc :

- un plan de sondage ou d'échantillonnage : procédures de sélection des unités
- des modalités de collecte
- l'estimation des caractéristiques recherchées.

Les problèmes de collecte sont traités dans d'autres parties de cet ouvrage. Il sera question ici des seuls plans d'échantillonnage (les principales méthodes et leurs qualités comparées) et les comparaisons porteront essentiellement sur deux indicateurs de qualité :

1) *L'absence de biais* (ou : vise-t-on la bonne cible ?). En effet, un estimateur issu d'un échantillon est une variable aléatoire, il est dit sans biais si son espérance mathématique est bien le paramètre visé. En d'autres termes, si P est le paramètre

(inconnu) à estimer, et $\hat{P}$ l'estimateur fourni par un échantillon conçu selon une méthode donnée, on veut que l'ensemble des $\hat{P}$ fournis par tous les échantillons possibles selon cette méthode donne en moyenne P, ce qu'on écrit :

$$E(\hat{P}) = P$$

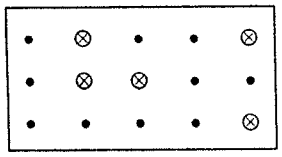
2) *le risque* (ou erreur quadratique moyenne). Il s'agit de caractériser la dispersion de l'ensemble des $\hat{P}$ autour de la cible P. On définit le risque par $E(\hat{P} - P)^2$. La notion de risque est donc liée à une mesure de la distance moyenne de l'estimation à sa cible. Dans le cas d'un estimateur sans biais, le risque est égal à sa variance. L'objectif du sondeur est de choisir une méthode qui garantisse un risque minimum à coût donné ou, problème dual, qui minimise le coût d'échantillonnage à risque consenti.

Si la population de référence est accessible à partir d'une base de sondage, c'est-à-dire par une liste exhaustive et sans doubles comptes d'identifiants des individus qui la composent, on utilisera des méthodes dites "aléatoires" (ou probabilistes). Des outils classiques du calcul des probabilités permettent de produire des espérances mathématiques et des variances des estimateurs. L'échantillonnage sera d'autant plus efficace qu'on saura mobiliser les informations auxiliaires contenues dans la base de sondage pour affiner les probabilités individuelles de sélection.

S'il n'y pas de base de sondage, mais qu'on dispose de données de cadrage de la population, on applique des méthodes dites empiriques ou "à choix raisonné" dont la plus répandue, et de très loin, est *la méthode des quotas*. La qualité d'un échantillon dépend alors essentiellement de la fiabilité et de la pertinence des données de cadrage.

2.2 Méthodes de sondage aléatoires

2.2.1 Sondage aléatoire simple à probabilités égales



Toutes les combinaisons de n parmi N sont autorisées, avec la même probabilité (cas du tirage sans remise)

Tous les individus de la population ont, a priori, la probabilité $f = \frac{n}{N}$ de figurer dans l'échantillon ; f est aussi appelé taux de sondage.

L'échantillon est "auto-pondéré". Pour estimer sans biais une moyenne $\overline{Y}$, on utilise la moyenne simple de l'échantillon :

$$\hat{\overline{Y}} = \overline{y} = \sum_{i \in Ech} Y_i / n$$

La variance de $\hat{\overline{Y}}$ dépend du taux de sondage, de la taille de l'échantillon, de la variance σ^2 de Y dans la population :

$$V(\hat{\overline{Y}}) = \frac{N}{N-1}(1-f)\frac{\sigma^2}{n}$$

Dans la pratique, on a le plus souvent :

$$V(\hat{\overline{Y}}) \cong \frac{\sigma^2}{n}$$

Les intervalles de confiance (ou marges d'erreurs, ou "fourchettes") sont fondés sur une forme de la loi des grands nombres (théorème central limite). Ainsi, pour des échantillons de taille suffisante, l'intervalle de confiance à 95 % pour $\overline{Y}$ est :

$$IC = \hat{\overline{Y}} \pm 1{,}96\sqrt{\hat{V}\left(\hat{\overline{Y}}\right)}$$

où

$$\hat{V}\left(\hat{\overline{Y}}\right) = (1-f)\frac{\hat{\sigma}^2}{n} \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i \in Ech}\left(Y_i - \hat{\overline{Y}}\right)^2$$

Dans le cas d'une proportion P, on a :

$$\sigma^2 = P(1-P)$$

Le tableau 1 suivant donne en fonction :

- de la taille d'échantillon n,
- de la proportion observée $\hat{P}$ dans l'échantillon,

le demi-intervalle de confiance estimé, en pourcentage, pour la proportion P dans la population (au degré de confiance 95 %) lorsque n est très petit devant N.

Tableau 1

Précision de l'estimation d'une proportion calculée à partir d'un échantillon

n (taille d'échantillon)	Proportion observée $\hat{P}$									
	5 % ou 95 %	8 % ou 92 %	10 % ou 90 %	15 % ou 85 %	20 % ou 80 %	25 % ou 75 %	30 % ou 70 %	35 % ou 65 %	40 % ou 60 %	50 %
100					8	8,6	9,2	9,6	9,8	10
150				5,7	6,4	6,9	7,3	7,6	7,8	8
200			4,3	5,1	5,7	6,1	6,5	6,8	6,9	7,1
250	2,8	3,4	3,8	4,5	5	5,4	5,8	6	6,2	6,3
300	2,5	3,1	3,5	4,2	4,6	5	5,3	5,6	5,7	5,8
350	2,3	2,9	3,2	3,8	4,2	4,6	4,9	5,1	5,2	5,3
400	2,2	2,7	3	,36	4	4,3	4,6	4,8	4,9	5
500	2	2,4	2,7	3,2	3,6	3,9	4,1	4,3	4,4	5
600	1,8	2,2	2,4	3	3,3	3,5	3,8	3,9	4	4,1
700	1,7	2,1	2,3	2,7	3	3,2	3,5	3,5	3,7	3,8
800	1,5	1,9	2,1	2,5	2,8	3	3,2	3,3	3,4	3,5
900	1,5	1,8	2	2,4	2,7	2,9	3	3,1	3,2	3,3
1 000	1,4	1,7	1,8	2,3	2,5	2,7	2,9	3	3	3,1
1 500	1,2	1,4	1,5	1,9	2,1	2,4	2,4	2,5	2,6	2,6
2 000	1	1,2	1,3	1,6	1,8	2,1	2,1	2,2	2,2	2,3
3 000	0,8	1	1,1	1,3	1,4	1,6	1,6	1,7	1,8	1,8
5 000	0,6	0,8	0,8	1	1,1	1,3	1,3	1,4	1,4	1,4
10 000	0,4	0,5	0,6	0,7	0,8	0,9	0,9	1	1	1

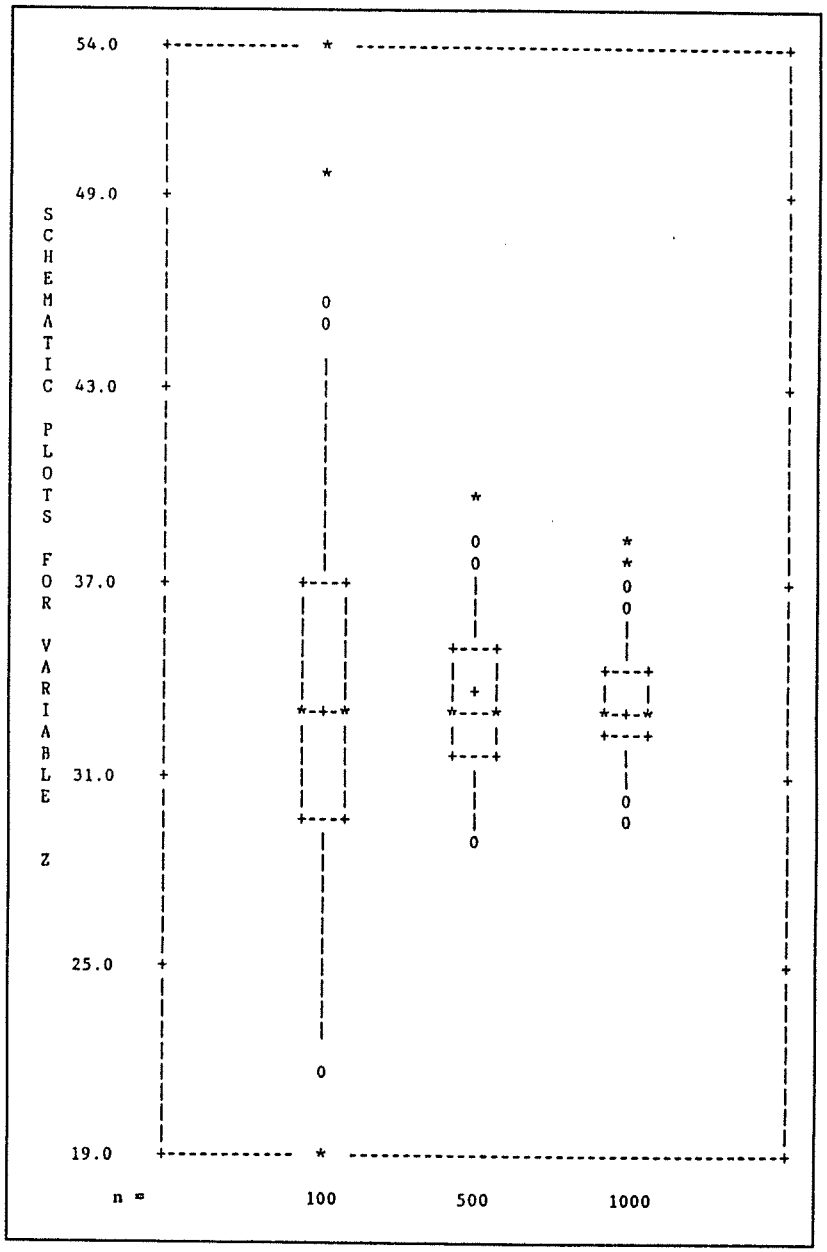
Exemple :

n = 400 et $\hat{P}$ = 25 %.

L'intervalle de confiance, pour P, au degré de confiance 95 % est :

0,25 - 0,043 < P < 0,25 + 0,043 soit 20,7 % < P < 29,3 %

Le graphique 1 montre les résultats de simulations d'échantillons de différentes tailles pour $P = \frac{1}{3}$.



Graphique 1

2.2.2 *Sondage aléatoire simple à probabilités inégales*

Le sondage aléatoire n'exige pas que toutes les unités aient la même probabilité d'être choisies, mais que chacune ait une probabilité d'inclusion π_i non nulle et connue. On peut donc choisir des unités avec des probabilités inégales, c'est-à-dire hiérarchiser leur importance.

Le cas le plus fréquent est celui où l'on dispose d'une variable caractérisant la "taille" des unités. Si X_i est la mesure relative à l'unité i, le plan de sondage lui attribue la probabilité d'inclusion $\pi_i = n\, X_i/X$, où X est le total des X_i dans la population.

Exemples :

- Exploitations agricoles tirées d'après leur surface,
- Entreprises tirées d'après leur chiffre d'affaires,
- Communes,
- Etc.

En ce cas les estimations se font en redressant les valeurs observées par les inverses des probabilités d'inclusion. L'estimateur sans biais du total Y est ainsi :

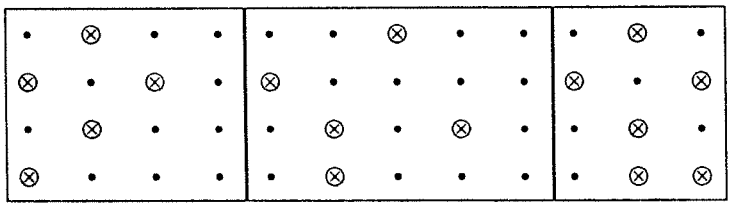
$$\widehat{Y} = \sum_{i \in Ech} \frac{Y_i}{\pi_i}$$

Exemples :

- Si $\pi_i = 1$, l'unité i ne "représente" qu'elle-même,
- Si $\pi_i = 0{,}01$, l'unité i de l'échantillon "représente" 100 unités de la population.

La précision dépend essentiellement de la qualité de la corrélation entre la variable auxiliaire définissant les π_i et les variables d'intérêt du sondage.

2.2.3 *Sondage aléatoire stratifié*



La stratification est un regroupement des unités de la base de sondage en sections relativement homogènes appelées strates.

L'échantillon final est composé d'échantillons prélevés indépendamment dans *chacune* des strates.

Les estimateurs tiennent compte des poids des strates :

$$\widehat{\widehat{Y}} = \sum_h \frac{N_h}{N} \widehat{\overline{Y}_h}$$

où $\widehat{\overline{Y}_h}$ est l'estimateur de la moyenne propre à la strate h.

On a recours à la stratification afin de réduire l'erreur d'échantillonnage (nécessité de strates homogènes), pour permettre un meilleur contrôle des coûts, pour avoir des estimateurs par strate, etc.

Le problème de la stratification est la richesse de l'information disponible pour constituer des strates homogènes.

Si l'échantillon se répartit au prorata des effectifs de la population :

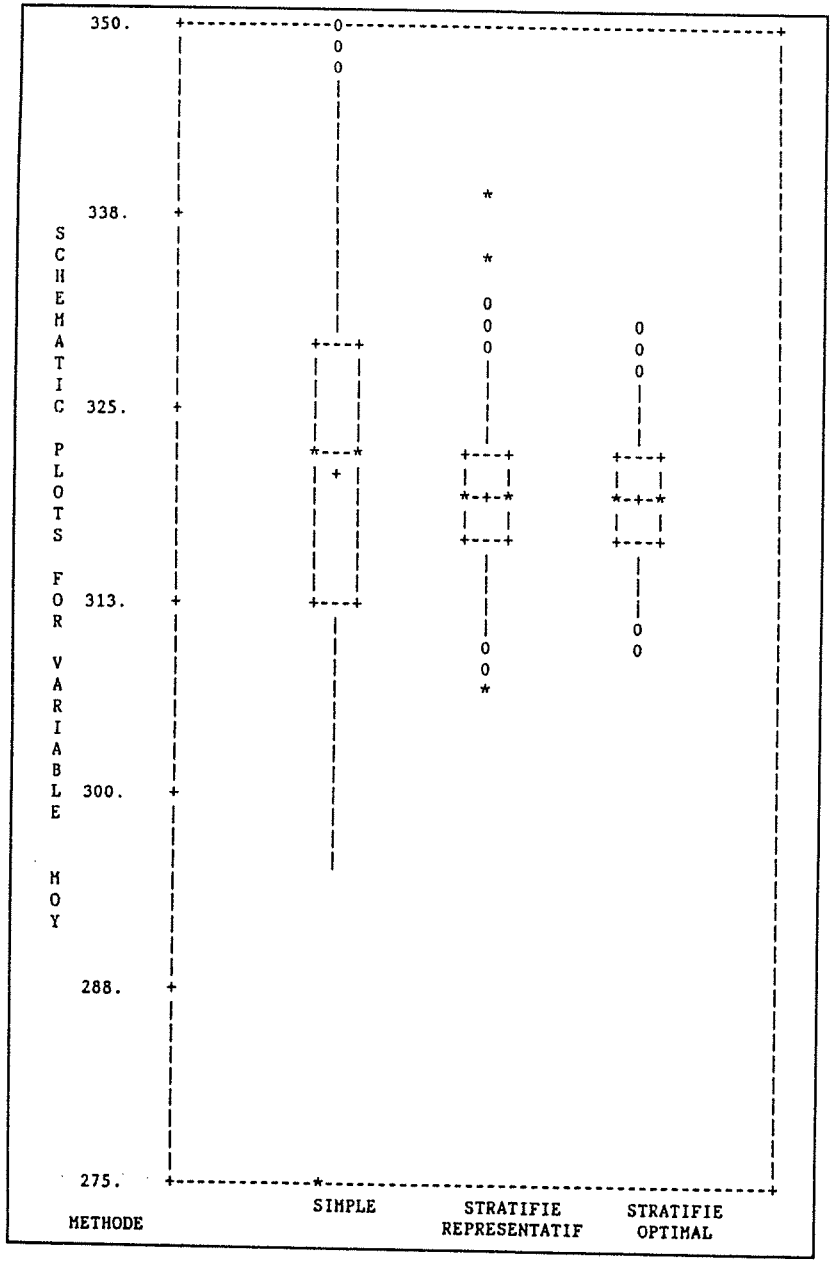
$$\frac{n_h}{n} = \frac{N_h}{N},$$

on parle d'échantillon "stratifié représentatif" ou échantillon "stratifié proportionnel".

Mais on a la faculté d'améliorer le dosage de l'échantillon en sur-représentant les strates les plus hétérogènes et en sous-représentant les plus homogènes.

L'échantillon "optimal" de Neyman est construit d'après la connaissance a priori des variances intra-strates.

Le graphique 2 et le tableau 2 montrent des résultats de simulations illustrant l'intérêt de la stratification par rapport au sondage simple et de la stratification optimale par rapport à la stratification représentative.



Graphique 2

Tableau 2

A — Etude par simulations

Caractéristiques de la base de sondage :

Strate	Effectif	Moyenne	Ecart-type
1	10 000	151.4	89.2
2	10 000	219.9	50.5
3	30 000	410.0	20.0
Ensemble	50 000	320.3	122.0

Comparaison de trois plans de sondage :

Echantillons de taille $n = 100$

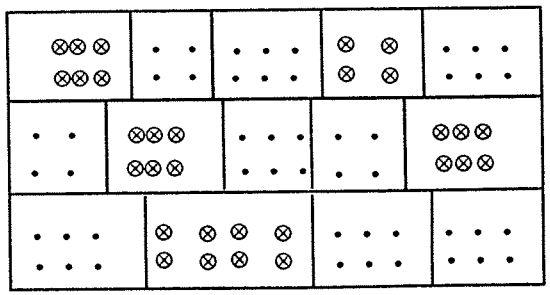
1) Simple
2) Stratifié représentatif ($n_1 = 20$, $n_2 = 20$, $n_3 = 60$)
3) Stratifié optimal ($n_1 = 45$, $n_2 = 25$, $n_3 = 30$)

200 simulations pour chaque cas.

B — Résultats des simulations

Méthode	Simple	Représentatif	Optimal
Moyenne	321.1	319.7	319.6
Minimum	275.6	306.7	309.1
Maximum	349.6	338.7	330.7
Ecart-type empirique	12.6	5.2	4.1
Ecart-type exact	12.2	4.8	4.0

2.2.4 Sondage par grappes



L'échantillonnage par grappes est un processus en deux étapes :

- Définition et choix des grappes ;
- Sélection de toutes les unités appartenant aux grappes choisies.

Dans la pratique, un très grand nombre de sondages se réalisent par grappes. Ainsi :

- En contrôle de qualité, les contrôles à réception se font sur des échantillons de lots ;
- Les enquêtes d'opinion auprès des utilisateurs des compagnies aériennes portent sur des échantillons de passagers regroupés sur des vols donnés ;
- Certaines études médicales sont réalisées à partir des malades d'un échantillon de médecins ;
- Les études sur l'emploi sont conduites sur des échantillons de logements regroupés en aires, etc.

Les méthodes vues précédemment peuvent s'appliquer au choix des grappes : échantillonnage aléatoire simple, avec probabilités égales ou inégales, stratification.

Avantages :

- Utilisable même en l'absence de liste complète de la population ;
- Réduction des coûts par concentration de l'échantillon.

Inconvénients :

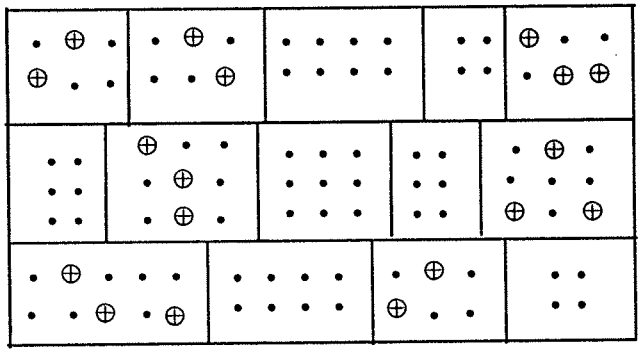
- Perte d'efficacité quand les unités d'une grappe se ressemblent trop (effets de grappe) ;
- Estimation plus difficile (quand les grappes sont de tailles inégales).

Idéalement, la répartition en grappes devrait être telle que chacune devrait constituer un portrait fidèle de la population. Dans la pratique, on ne peut approcher cet idéal que par une *stratification préalable* des grappes.

L'utilisation de grappes est particulièrement adaptée à l'échantillonnage à partir de bases de sondage de type aréolaire.

2.2.5 Echantillonnage à deux ou plusieurs degrés

La situation est analogue à celle du sondage en grappes, mais chaque grappe désignée n'est pas explorée exhaustivement, mais par un sondage qui lui est propre.



Premier degré :

- Définition et choix de grappes (unités primaires).

Deuxième degré :

- Sélection d'unités au sein des grappes choisies au premier degré.

Avantages :

- La base de sondage est construite pour bien représenter les unités de premier degré (unités primaires).
- La base de sondage pour les unités secondaires peut être construite durant la "visite" des unités primaires choisies.

Inconvénients :

- Comme pour les grappes, il y a perte d'efficacité quand les unités d'une unité primaire se ressemblent trop (effets de grappe) ;
- Les estimateurs ont également une forme plus complexe ;
- Il est nécessaire de bien stratifier les unités primaires.

Cas particuliers :

- Plans de sondage conduisant à des échantillons auto-pondérés :

a) - Choix des unités primaires avec des probabilités égales ;

- Choix des unités secondaires avec des probabilités égales, et avec le même taux de sondage à l'intérieur de chaque unité primaire tirée.

b) - Choix des unités primaires avec des probabilités inégales, proportionnelles à leur taille exprimée en nombre d'unités secondaires.

- Choix des unités secondaires avec des probabilités égales, le *même nombre* dans chaque unité primaire.

2.3 Méthodes de sondage non aléatoires

2.3.1 Méthodes des quotas

L'expression "non aléatoire" ne veut pas dire qu'il n'y a pas de "hasard" ; elle signifie surtout que faute d'avoir à disposition une base de sondage opérationnelle, on ne peut pas appliquer *directement*, comme dans les méthodes dites probabilistes, les outils du calcul des probabilités.

De très loin la plus utilisée, la méthode des quotas tire partie du fait qu'il existe des données de cadrage pour la population de référence (répartition par sexe, âge, C.S., etc.) et que ces données sont liées à l'objet du sondage.

On détermine des tailles d'échantillons (quotas) à partir des distributions connues. A la différence de l'échantillonnage stratifié, on choisit des unités jusqu'à ce que les quotas soient remplis (c'est-à-dire que les non-répondants soient remplacés).

Concrètement, on demande aux enquêteurs de constituer leur partie d'échantillon en respectant certaines contraintes, comme l'illustre l'exemple qui suit.

Exemple :

Feuille de quotas d'un enquêteur

Enquêteur : Jules-Smith ENFACE
Interviews à réaliser : 25

Electeurs inscrits	25	1	2	3	4	5	6	7	8	9	10	11	12	13
		14	15	16	17	18	19	20	21	22	23	24	25	
SEXE														
Hommes	13	1	2	3	4	5	6	7	8	9	10	11	12	13
Femmes	12	1	2	3	4	5	6	7	8	9	10	11	12	
AGE														
18-34	12	1	2	3	4	5	6	7	8	9	10	11	12	
35-49	8	1	2	3	4	5	6	7	8					
50-64	3	1	2	3										
65 et +	2	1	2											
C.S.														
Artisan Commerçant	1	1												
Ind. lib. Cad. sup.	2	1	2											
Cad. moy. Emp.	10	1	2	3	4	5	6	7	8	9	10			
Ouvrier	10	1	2	3	4	5	6	7	8	9	10			
Inactif	2	1	2											

L'important n'est pas que chaque enquêteur respecte exactement l'intégralité de sa feuille de quotas (problèmes de fin de feuille de quotas), mais que la réunion des résultats respecte les distributions voulues.

Cette méthode est attrayante car elle est relativement peu coûteuse à mettre en oeuvre et à gérer. Elle permet de pallier le problème de l'absence de base de sondage. Bien qu'elle semble assurer une certaine "représentativité", elle peut masquer les différents degrés de refus et les autres biais dus à la collecte.

Elle ne permet pas de produire des intervalles de confiance, sauf à faire appel à des modèles rendant compte des relations entre les variables fondant les quotas et les variables d'intérêt du sondage. Il est clair que l'efficacité de la méthode des quotas dépend essentiellement de la qualité de ces relations.

2.3.2 Autres méthodes non aléatoires

Ces autres méthodes sont surtout employées pour obtenir un éclairage rapide et sommaire afin d'initier des recherches plus approfondies.

- *Unités-types* : désignation d'unités supposées "représenter" des tendances caractéristiques ou moyennes de la population ;
- *Volontariat* : C'est, par exemple, la seule possibilité pour certaines études médicales. La méthode est cependant très sensible aux biais de sélection.

Ce panorama général sur les méthodes de sondage peut donner l'impression que, dans une enquête par sondage, on choisit une seule de ces méthodes. Il n'en est rien. Il arrive souvent, en effet, que ces différentes méthodes soient combinées dans le plan d'échantillonnage adopté. Par exemple, un schéma classique (présenté avec quelques simplifications) dans les enquêtes par sondage en France réalisées par enquêteur à domicile auprès d'individus ou de ménages, consiste à :

- stratifier la France selon la double classification région et catégorie de commune (par exemple : moins de 2 000 habitants, 2 000 à 50 000 habitants, plus de 50 000 habitants),
- tirer dans chaque strate un échantillon d'unités primaires, qui sont des communes rurales ou des quartiers ou communes dans les agglomérations, à probabilités inégales proportionnelles à leur taille et avec remise,
- tirer, dans chaque unité secondaire sélectionnée, un nombre fixe de ménages ou d'individus. A ce dernier stade, dans les instituts de sondage privés, l'enquête se fait généralement par quotas, i.e. selon une méthode de sondage empirique.

2.4 Amélioration des estimations : les redressements

Un échantillon ayant été sélectionné et interrogé, il est possible au moment du dépouillement des résultats d'améliorer les résultats du sondage si l'on constate *a posteriori* que l'échantillon diffère de la population sur une (ou des) caractéristique(s) dont on *connaît* la distribution dans la population.

Toutes les méthodes de redressement s'inspirent du même principe : on dispose d'une (ou de plusieurs) variable(s) de contrôle X (ou variable auxiliaire) dont on connaît :

- soit la valeur moyenne ou le total dans la population (cas de variables de contrôle quantitatives),
- soit la distribution dans la population (cas de variables qualitatives).

Ces renseignements sont insuffisants pour être utilisés dans le plan d'échantillonnage ; par exemple, une méthode de stratification demande de connaître les valeurs prises par la variable de stratification pour chaque individu de la population ; ou trop nombreux pour être tous pris en compte dans une stratification ou comme variables de quotas.

On constate a posteriori une différence entre échantillon et population sur ces caractéristiques (nous supposerons dans ce paragraphe que les écarts constatés ne proviennent que des fluctuations d'échantillonnage ; nous verrons au paragraphe 2.5 que des techniques similaires peuvent être utilisées lorsque les écarts proviennent de non-réponses). On essaie donc de tenir compte de cet écart pour mieux se rapprocher de la cible. Nous distinguerons le cas d'une variable de contrôle unique du cas de variables de contrôle multiples.

2.4.1 Redressement sur une variable

2.4.1.1 Variable auxiliaire quantitative

Lorsque la variable auxiliaire X est quantitative, il existe de nombreux estimateurs dont le choix dépend :

- de la forme de relation existant entre la variable étudiée et la variable auxiliaire et,
- du plan de sondage qui a été adopté.

Pour simplifier, nous supposerons le plan de sondage aléatoire simple à probabilités égales et ne détaillerons qu'un estimateur : *l'estimateur par le quotient.*

Nous renvoyons le lecteur à Cochran (1977), Grosbras (1987)... pour l'étude d'autres estimateurs possibles (ex : estimateur par différence, par régression).

L'estimateur par le quotient est utilisé lorsqu'il existe une relation de proportionnalité entre les variables Y et X.

Si l'on connaît la moyenne de X dans la population (soit $\overline{X}$) et dans l'échantillon (soit $\overline{x}$), l'estimateur par le quotient $\overline{Y}$ est :

$$\widehat{\overline{Y}}_q = \overline{y}\,\frac{\overline{X}}{\overline{x}}$$

où $\overline{y}$ est la moyenne simple d'échantillon. Le principe est donc celui de la règle de 3. Utiliser cet estimateur revient à supposer que le coefficient de pondération qui ramènerait la valeur observée $\overline{x}$ à la valeur réelle $\overline{X}$ est également celui qui ramènerait $\overline{y}$ à $\overline{Y}$.

Les propriétés de cet estimateur sont intéressantes car :

- il est biaisé mais son biais tend vers zéro quand n croît,
- son risque (ou erreur quadratique moyenne) est inférieur à la variance de l'estimateur classique $\overline{y}$ si le coefficient de corrélation linéaire ρ entre Y et X est supérieur à la moitié du rapport entre les coefficients de variation de Y et X soit :

$$\rho > \frac{1}{2}\,\frac{\sigma_X / \overline{X}}{\sigma_y / \overline{Y}}$$

ce qui est le cas si les X_α et les Y_α sont approximativement proportionnels.

Exemples de redressement par le quotient :

- Enquêtes agricoles où l'on utilise la superficie totale cultivée de la population des exploitations pour améliorer les estimations de production de certains produits agricoles ;
- Panels de détaillants pour estimer les ventes de certaines catégories de produits : le redressement utilise le chiffre d'affaires total X des magasins de la population.

L'estimateur par le quotient et les estimateurs analogues (différence et régression) se généralisent au cas de plusieurs variables auxiliaires (cf. Särndal et al, 1992).

2.4.1.2 Variable auxiliaire qualitative : stratification a posteriori

Une variable liée au sujet de l'étude et dont on connaît la distribution dans la population (ex : nombre de personnes au foyer) n'a pas été prise en compte au moment du tirage de l'échantillon pour des raisons déjà évoquées (elle est cependant connue pour chaque individu de l'échantillon).

On va alors stratifier *a posteriori* selon cette variable qui, si elle a L modalités, définit L poststrates.

L'estimateur poststratifié pour $\overline{Y}$ est :

$$\widehat{\overline{Y}}_{pos} = \sum_h \frac{N_h}{N}\,\widehat{\overline{Y}}_h$$

où $\widehat{\overline{Y}}_h$ est l'estimateur de $\overline{Y_h}$ à l'aide des n_h individus de l'échantillon figurant dans la poststrate h (les n_h sont aléatoires).

Dans le cas où $\widehat{\overline{Y}}_h$ est la moyenne simple d'échantillon $\overline{y_h}$, on peut montrer facilement que $\widehat{\overline{Y}}_{pos}$ s'écrit :

$$\widehat{\overline{Y}}_{pos} = \frac{1}{n}\sum_h \sum_{i=1}^{n_h} P_{hi} Y_{hi} \qquad \text{où} \qquad P_{hi} = \frac{N_h/N}{n_h/n},\ \forall i = 1,\ldots,n_h$$

L'estimateur poststratifié est donc une moyenne pondérée où chaque individu de l'échantillon est affecté d'un poids

- supérieur à 1 si la poststrate à laquelle il appartient est sous-représentée dans l'échantillon : $n_h / n < N_h / N$,
- inférieur à 1 dans le cas contraire.

Dans le cas où la poststratification est utilisée pour corriger les fluctuations d'échantillonnage, l'estimateur $\widehat{\overline{Y}}_{pos}$ est non biaisé et a une variance approximativement égale à celle d'un échantillon stratifié proportionnel (mais évidemment pas d'un échantillon stratifié optimal). Un estimateur semblable est utilisé pour améliorer l'effet des non-réponses (mais les propriétés sont différentes, cf. § 2.5).

Exemple :

Echantillon de 1 000 individus de 15 ans et plus interrogés sur la lecture du magazine ROND-POINT au cours des six mois précédant l'enquête. Le taux de lecture observé dans l'échantillon est de 12 %.

On constate par ailleurs a posteriori que, dans l'échantillon, 29 % des individus interrogés habitent en zone rurale alors que cette proportion n'est que de 25 % dans la population.

La ventilation des résultats de l'échantillon selon la variable "habite en zone rurale - habite en zone urbaine" donne le tableau suivant :

Effectifs	Zone rurale	Zone urbaine	Total
Ont lu au moins une fois ROND-POINT	10	110	120
N'ont pas lu ROND-POINT	280	600	880
Total	290	710	1000

L'estimation sans stratification a posteriori est $\hat{P} = 0,12$.

L'estimation par poststratification est :

$$\hat{P}_{pos} = \frac{N_{rur}}{N} \hat{P}_{rur} + \frac{N_{urb}}{N} \hat{P}_{urb}$$

soit

$$\hat{P}_{pos} = 0,25 \times \frac{10}{290} + 0,75 \times \frac{110}{710} = 0,1248$$

2.4.2 Redressement sur critères multiples

2.4.2.1 La méthode RAS

Dans la plupart des enquêtes, on peut disposer de plusieurs variables de contrôle permettant de réaliser une stratification *a posteriori* ; il est évidemment tentant d'en réaliser simultanément plusieurs.

Cette méthode ne pose pas de problème si l'on connaît la répartition des individus de la population selon les croisements des modalités de ces différentes variables : on se retrouve alors dans le cas de la stratification a posteriori.

Mais, dans la majorité des cas, on ne connaît que les distributions marginales. On se trouve alors confronté au problème dit de "l'équilibrage d'un tableau dont on connaît les marges" ou au problème du choix des coefficients de redressement par case (les p_{hi} de la poststratification) respectant les conditions à la marge.

La méthode la plus utilisée est la méthode RAS (Ranking Adjusted Statistics) que nous décrirons brièvement dans le cas de deux critères :

Soit $A = (a_{ij})$ le tableau d'effectifs observés à ajuster
Le total de ligne observé est $a_{i.}$; le total théorique est r_i.
Le total de colonne observé est $a_{.j}$; le total théorique est s_j.

On commence par ajuster les totaux en ligne :

$$a_{ij} \rightarrow a_{ij} \times r_i./a_{i.}$$

puis les totaux en colonne

$$a_{ij} \rightarrow a_{ij} \times s_j./a_{.j}$$

Mais l'ajustement en colonne détruit l'ajustement en ligne ; on itère donc le processus jusqu'à convergence.

Après redressement, on aboutit au tableau redressé (x_{ij}) ce qui revient à appliquer à chaque individu relatif à la case (i, j) le poids $p_{ij} = x_{ij}/a_{ij}$ (dans le cas de 2 critères).

Cette technique est généralisable à des redressements portant sur un nombre plus élevé de critères.

Il ne faut cependant pas multiplier exagérément les critères, ce qui aurait pour effet de multiplier les cases du tableau multiple. Dans ce cas, des cases seront vides par nature (exemple : retraité x moins de 25 ans) mais d'autres le seront parce que l'échantillon est de taille trop faible pour représenter correctement la population dans un découpage aussi fin. Pour ces dernières, étant donné que l'algorithme consiste à effectuer des multiplications case à case, on ne parviendra pas à les remplir mais on risque de créer des déviations dans les cases avoisinantes.

Une règle empirique est de concevoir des redressements qui soient tels que la taille totale de l'échantillon divisé par le nombre total de modalités des variables de redressement soit supérieur à 50.

Dans la pratique, enfin, on a coutume de limiter arbitrairement la variation possible des poids de redressement obtenus afin de ne pas risquer de perturber exagérément les résultats si un individu de comportement marginal se trouvait affecté d'un poids de redressement trop important.

Le lecteur trouvera dans Grosbras (1987) d'autres méthodes de redressement qui, par ailleurs, font l'objet de nombreuses recherches actuellement (cf. Deville et Särndal, 1992 ; Sautory, 1991).

En conclusion, les méthodes de redressement sont justifiées et efficaces à certaines conditions :

- La ou les variables de contrôle doivent être bien corrélées avec les variables qui font l'objet de l'enquête ;
- Les statistiques pour le calage sont pertinentes et fiables (elles concernent la même population, elles proviennent d'un recensement ou d'une très grosse enquête, elles ne sont pas périmées...) ;
- L'échantillon est de taille suffisante. Il est illusoire de vouloir corriger un échantillon de 100 individus ;
- Les redressements doivent être conçus comme un "lissage" des résultats pour en affiner la présentation ; ils n'ont donc pour effet que des modifications "à la marge". Une modification brutale serait le signe probable d'un mauvais échantillon initial.

2.4.2.2 Exemple de redressement

A titre d'illustration, nous présentons le redressement de l'enquête 75 000 Radio de Médiamétrie.

A) Le plan de sondage

L'enquête "75 000 Radio", dont l'objectif essentiel est de mesurer l'audience de la radio et des stations, porte sur un échantillon de 75 250 personnes physiques réparties sur 10 mois à raison de 250 interviews par jour. La population de référence est constituée par l'ensemble des individus âgés de 15 ans et plus résidant en France métropolitaine et appartenant à un ménage ordinaire de nationalité française ou étrangère.

Principe :

1. Les 75 000 appels sont stratifiés par département, au prorata de la population départementale.
2. A l'intérieur de chaque département, on procède à une sélection de communes par tirage avec probabilités inégales sans remise qui permet :
   * de prendre en compte les plus petites communes,

   * d'assurer une dispersion maximale des interviews.

   Pour chaque commune échantillonnée, on détermine ainsi le nombre d'interviews théoriques à réaliser.
3. Tirage des numéros de téléphone dans l'annuaire.
4. Au niveau individuel, réalisation de 250 interviews quotidiennes selon des quotas par sexe, âge et activité.

B) Le programme de redressement

C'est une méthode de type RAS, fondée sur les sept variables suivantes :

JOUR	Jour d'audience
PCSI 7 classes :	Profession de l'individu - Agriculteurs - Petits patrons - Cadres sup., Prof. lib. - Prof. intermédiaires - Employés - Ouvriers - Inactifs
PCSC	Profession du chef de ménage (idem que PCSI)
HAB8 8 classes :	Type d'habitat - Communes rurales hors ZPIU - Communes rurales dans ZPIU - Communes < 20 000 - De 20 à 50 000 - De 50 à 100 000 - De 100 à 200 000 - + de 200 000 (hors Paris) - Paris
SXAG 12 classes :	Sexe x Tranche d'âge H/F croisé par : - 15/19 ans - 20/24 ans - 25/34 ans - 35/49 ans - 50/64 ans - 65 ans et +

ACT4	Sexe x Activité
4 classes :	H/F x Actifs/Inactifs
SXRA	Sexe x Région UDA
18 classes :	H/F croisé par :
	- Région Parisienne
	- Nord
	- Est
	- Bassin Parisien Est
	- Bassin Parisien Ouest
	- Ouest
	- Sud Ouest
	- Sud Est
	- Méditerranée

Les valeurs objectifs, prises par les variables de redressement, proviennent de l'INSEE.

C) Les résultats (tableau 3)

Les résultats suivants proviennent du redressement de l'échantillon entre Janvier et Mars 1992 (base Lundi - Vendredi).
La taille de l'échantillon était 16 271.

2.4.2.3 Les logiciels de redressement

La plupart des logiciels d'analyse statistique permettent la prise en compte de poids dans l'analyse des résultats. L'utilisateur a donc la possibilité d'affecter à chaque individu i de l'échantillon un poids de redressement p_i qui sera alors introduit dans le calcul des moyennes, proportions...

Certains logiciels plus orientés vers le dépouillement d'enquêtes proposent une méthode de redressement de type RAS.

Les instituts de sondage, gérant de grosses enquêtes ou panels, ont généralement développé leurs propres programmes de redressement.

2.5 La qualité d'un sondage

Dans tout ce qui précède, nous avons présenté différentes méthodes de sondage et d'estimation, et indiqué que l'objectif majeur du sondeur était d'obtenir un risque minimum pour un coût d'enquête donné. Pour simplifier l'exposé, nous avons supposé que l'écart constaté entre les estimations obtenues dans l'échantillon et les paramètres de la population ne provenait que des fluctuations dues à l'échantillonnage (ou *erreurs d'échantillonnage*).

Tableau 3

NOMBRE D'ITERATIONS = 50

MOYENNE POIDS FINAL = 2.7853 ECART-TYPE .7488

POIDS MINIMUM = 1.3525 POIDS MAXIMUM = 8.8902

GROUPES DE POIDS FINAL	0.0	1.04	1.54	2.04	2.54	3.03	3.53	4.03	4.53
	.0	.2	11.4	27.3	32.4	16.9	6.9	2.1	2.7

ECHANTILLON REEL

CRITERE	CODE	OBJECTIF THEORIQUE (%)	DONNEES BRUTES (%)	DONNEES REDRESSEES (%)
JOU5	1	200	200	200,01
JOU5	2	200	200	199,99
JOU5	3	200	200	200,00
JOU5	4	200	200	200,00
JOU5	5	200	199	200,01
PCSI	1	30	18	30,01
PCSI	2	38	34	38,01
PCSI	3	56	75	56,02
PCSI	4	104	142	104,05
PCSI	5	151	150	151,05
PCSI	6	162	121	162,05
PCSI	7	459	461	458,80
PCSC	1	53	27	53,01
PCSC	2	59	59	59,01
PCSC	3	94	123	94,02
PCSC	4	142	170	142,03
PCSC	5	100	116	100,02
PCSC	6	246	201	246,04
PCSC	7	306	305	305,88
HAB8	1	107	98	106,99
HAB8	2	163	161	163,00
HAB8	3	158	172	158,00
HAB8	4	65	70	65,01
HAB8	5	67	69	67,00
HAB8	6	75	75	74,99
HAB8	7	204	198	203,99
HAB8	8	161	157	161,02
SXAG	1	46	47	46,07
SXAG	2	47	44	47,09
SXAG	3	93	95	93,22
SXAG	4	125	122	125,30
SXAG	5	97	100	97,20
SXAG	6	70	71	70,11
SXAG	7	45	44	44,90
SXAG	8	48	48	47,91
SXAG	9	96	96	95,85
SXAG	10	124	121	123,79
SXAG	11	104	106	103,80
SXAG	12	105	106	104,77
ACT4	1	307	307	307,65
ACT4	2	171	173	171,34
ACT4	3	234	233	233,56
ACT4	4	288	288	287,45
SXRA	1	88	89	88,00
SXRA	2	33	32	33,00
SXRA	3	43	43	43,00
SXRA	4	41	41	41,00
SXRA	5	47	47	47,00
SXRA	6	63	64	63,00
SXRA	7	52	52	52,00
SXRA	8	56	57	56,00
SXRA	9	56	56	56,00
SXRA	10	99	97	99,01
SXRA	11	36	34	36,00
SXRA	12	46	48	46,00
SXRA	13	43	43	43,00
SXRA	14	50	48	50,00
SXRA	15	68	69	68,00
SXRA	16	56	56	56,00
SXRA	17	61	63	61,00
SXRA	18	62	61	62,00

Or, la production des résultats d'une enquête par sondage est un processus complexe incluant de nombreuses étapes comme :

- L'identification des objectifs de l'enquête,
- Le choix d'une base de sondage,
- Le choix du mode d'administration du questionnaire,
- Le plan d'échantillonnage (méthode de sondage, taille d'échantillon...),
- La construction du questionnaire,
- Le terrain incluant l'information et le contrôle des enquêteurs,
- La codification et la saisie,
- Le dépouillement des résultats incluant les méthodes d'estimation appropriées et le traitement des non-réponses.

A chacune de ces étapes, on peut effectuer des choix (ex : ordre des questions du questionnaire) ou commettre des erreurs (ex : erreurs de saisie) qui peuvent avoir un impact sur la *qualité* des résultats de l'enquête.

2.5.1 Les sources d'erreurs dans les enquêtes par sondage

Les principales *sources d'erreurs* dans les enquêtes par sondage sont :

- *L'erreur de couverture* qui provient de l'écart entre la valeur des paramètres étudiés pour la population visée par l'étude et celle qui apparaît dans la base de sondage. L'exemple typique est celui de l'existence des listes rouges et oranges dans les sondages par téléphone. De telles erreurs interviennent également dans l'utilisation de bases de logements, lorsque l'objet du sondage est fortement lié à la possession de logements neufs, etc.
- *L'erreur de non-réponse* c'est à dire l'erreur due à l'absence totale ou partielle d'informations concernant des individus de l'échantillon.
- *L'erreur de mesure* qui provient de l'inexactitude des réponses enregistrées à cause :
  * de l'effet induit par l'enquêteur sur les réponses des interviewés
  * des insuffisances (ou choix) dans la rédaction du questionnaire
  * des répondants (incapacité de répondre aux questions, désir de valorisation...)

Cet aspect est traité dans une autre intervention. Nous n'y reviendrons pas, mais examinerons le problème des non-réponses et quelques moyens d'y remédier.

2.5.2 L'erreur de non-réponse.

L'erreur de non-réponse provient donc de l'absence totale ou partielle d'informations concernant des individus de l'échantillon.

- La *non-réponse totale* provient de diverses causes :
  * unité non contactée (5 à 8 % des enquêtes ménages)
  * refus d'emblée (6 à 15 % des enquêtes ménages)
  * abandon en cours d'enquête
  * incapacité de répondre
  * perte ou effacement de documents
  * etc.
- La *non-réponse partielle* est due à :
  * refus devant certaines questions sensibles
  * incompréhension
  * donnée incohérente
  * etc.

On montre que les réponses manquantes dans une enquête créent un biais qui dépend de deux facteurs :

- Le taux de non-réponses,
- L'écart entre les comportements des répondants et non-répondants en ce qui concerne la ou les variables étudiées.

Autrement dit, le taux de non-réponses ne mesure pas à lui tout seul l'erreur de non-réponse. Un taux de non-réponses faible peut être très pernicieux si répondants et non-répondants ont des comportements très différents en ce qui concerne les thèmes de l'enquête. Et inversement, un taux de non-réponses élevé n'est pas trop grave si ces comportements sont très voisins.

Le problème dû aux *non-réponses partielles* est généralement moins aigu : les réponses aux autres questions du questionnaire donnent souvent des pistes d'explication exploitées par les méthodes *d'imputation* que nous décrirons plus loin.

Le problème des *non-réponses totales* est plus difficile et se pose par ailleurs très différemment selon le type d'enquêtes et de méthodes de sondage utilisés :

- Dans les méthodes empiriques (méthode des quotas par exemple lorsque l'enquête est réalisée en face à face) l'utilisateur de résultats n'a généralement pas connaissance du problème des non-réponses, des refus qu'a subi l'enquêteur, etc. Il peut éventuellement en avoir une appréciation sommaire, en examinant les consignes données aux enquêteurs, le déroulement du terrain... ;
- Dans les méthodes aléatoires par contre, les individus composant l'échantillon sont nommément désignés. On connaîtra donc après enquête le nombre de non-réponses et leurs causes ;
- Enfin, dans les panels, les non-répondants à une vague de panel sont bien identifiés. On possède généralement beaucoup d'informations les concernant (via le questionnaire de recrutement et leurs réponses à d'autres vagues du panel). On peut alors utiliser ces informations pour corriger l'effet des non-réponses ou pour estimer les réponses manquantes.

De façon générale, deux principes essentiels doivent être rappelés :

Principe n° 1 :

Eviter d'avoir des non-réponses, en soignant très consciencieusement les techniques de collecte (questionnaires, enquêteurs, rappels...)

Principe n° 2 :

Procéder à une analyse statistique poussée par les techniques de l'analyse descriptive et plus particulièrement l'analyse des correspondances multiples, mais aussi des techniques telles que l'analyse LOGIT qui peuvent éclairer utilement sur les facteurs influençant la non-réponse et la nature des biais encourus.

Les techniques de redressement d'échantillon sont fondées sur des hypothèses formulées sur la population (remarque : ne rien faire, c'est appliquer un modèle bien particulier !).

Pour la non réponse totale, la seule stratégie envisageable est celle de la *repondération* de l'échantillon des répondants : on réajuste sur des

distributions connues. Une méthode très utilisée est celle du calage proportionnel sur marges (méthode RAS). Il est clair que l'efficacité des redressements réside dans la qualité de la corrélation entre les caractéristiques supports des calages et les variables d'intérêt du sondage. L'exemple suivant montre le fonctionnement du redressement.

Exemple :

Panel postal mensuel de 10 000 individus

Age	Fréquences relatives (population)	Taux de réponses par classe	Effectifs des répondants	Proportion de répondants
15-30 ans	25 %	60 %	2 500 x 0,6 = 1 500	22,1 %
30-50 ans	35 %	60 %	2 100	30,9 %
+ de 50 ans	40 %	80 %	3 200	47,0 %
TOTAL	100 %		6 800	100 %

Le redressement consiste à pondérer :

- les questionnaires des répondants des classes 15-30 et 30-50 par :

$$\frac{25\ \%}{22{,}1\ \%} \approx \frac{35\ \%}{30{,}9\ \%} = 1{,}13$$

- les questionnaires des répondants de la classe d'âge +50 ans par :

$$\frac{40\ \%}{47\ \%} = 0{,}85$$

Supposons que les moyennes observées dans les trois classes d'âge soient :

$$\bar{y}_1 = 3 \quad \bar{y}_2 = 2 \quad \bar{y}_3 = 1$$

L'estimation *sans* pondération est :

$$\bar{y} = \frac{1\,500 \times 3 + 2\,100 \times 2 + 3\,200 \times 1}{6\,800} = 1{,}75$$

soit, en désignant par n_h (h = 1, 2, 3) les effectifs de répondants dans chaque classe

$$\bar{y} = \frac{\sum_{h=1}^{3} n_h \bar{y}_h}{n} \quad \text{avec} \quad n = \sum_h n_h$$

L'estimation *avec* pondération s'écrit :

$$\bar{y}_p = \sum_h W_h \bar{y}_h \quad \text{où} \quad W_h = \frac{N_h}{N}$$

$$= 0{,}25 \times 3 + 0{,}35 \times 2 + 0{,}40 \times 1$$

$$= 1{,}85$$

Pour la non réponse partielle, on utilise des méthodes par *imputation* : on remplace chaque donnée manquante par une donnée "prédite" en fonction des renseignements obtenus pour le même individu et pour des individus proches. Les principales méthodes sont les suivantes :

a) *Déductive* : imputation par règle déterministe souvent utilisée pour corriger des données incohérentes ou invalides

Ex : Age ≤ 14 ans ⟶ activité professionnelle = inactif

b) *"Cold-deck"* : utilisation d'une information extérieure relative à la même unité.

Ex : valeur observée à une date antérieure.

c) *"Hot-deck"* : on remplace la valeur manquante par la valeur observée chez un répondant proche : le donneur

* hot-deck d'ensemble : le donneur est choisi au hasard parmi les répondants

* hot-deck par classe : le donneur est choisi au hasard dans une classe à laquelle appartient le receveur

* hot-deck séquentiel : On défile le fichier à corriger, si une unité est défaillante, on lui impute la valeur renseignée par l'individu le plus "récent" appartenant à la même classe

* hot-deck hiérarchisé : on utilise une suite de critères $C_1, ..., C_k$. On remplace l'unité défaillante par une unité ayant les mêmes valeurs $C_1, ..., C_k$ s'il en existe, sinon par une unité ayant les mêmes valeurs pour $C_1, ..., C_{k-1}$, etc.

* hot-deck métrique : méthode du plus proche voisin. On construit une distance d(i,j) entre unités en fonction de variables bien renseignées qu'elles ont en commun. Si l'unité k est défaillante, on lui impute la valeur observée chez son plus proche voisin.

d) *Imputation par prédicteur :* On impute la moyenne des répondants, ou d'une classe particulière de répondants, aux unités défaillantes. Certaines techniques sophistiquées font appel à des modèles économétriques plus généraux.

2.5.3 La qualité d'un sondage

Jusqu'à une date récente, la qualité des résultats était surtout appréhendée à travers la construction (rare) d'intervalles de confiance dont on a souligné qu'ils ne tiennent compte que de l'erreur d'échantillonnage. Ceci était partiellement justifié par le fait qu'il est difficile de mesurer l'impact sur les résultats des autres types d'erreurs : erreurs de mesures et de non-réponses essentiellement.

Or, le chiffre publié n'indique rien sur sa qualité ou sur les conditions dans lesquelles il a été obtenu. Et pourtant, de tels chiffres, s'ils sont faux, peuvent conduire à des décisions catastrophiques.

Les organismes professionnels ont, depuis longtemps, établi des directives sur la façon dont doivent être détaillées les différentes phases d'une enquête (cf. Directives CCI/ESOMAR pour parvenir à un accord sur une étude de marché). Un colloque récent (La qualité de l'information dans les enquêtes, ASU, 1992) a réuni des spécialistes d'enquêtes sur le thème de la qualité.

Une tendance actuelle est d'appliquer et d'adapter à la production de résultats d'enquêtes les concepts et techniques de la Qualité Totale (la qualité étant définie comme "l'aptitude d'un produit ou d'un service à satisfaire au moindre coût et dans les moindres délais les besoins des utilisateurs").

En ce qui concerne les enquêtes par sondage, ceci implique en particulier :

- Le contrôle des phases sensibles d'une enquête en prenant en compte la connaissance de toutes les sources d'erreurs dans les enquêtes ;

- La mise à disposition des utilisateurs de l'information qui leur est nécessaire pour apprécier la qualité des données ;

- La mise au point d'indicateurs pour mesurer la qualité en prenant en compte et en développant les éléments suivants :

  a) erreur de couverture (différences entre la base de sondage utilisée et la population étudiée).

b) erreur d'échantillonnage dans le cas d'enquêtes aléatoires. Dans le cas d'échantillon non-aléatoire, essayer de dégager les conséquences sur la généralisation des résultats,

c) biais de non-réponse, (en prenant en compte les taux de réponses, les différences constatées entre les caractéristiques des répondants et des non-répondants, les méthodes d'imputation et d'estimation utilisées pour tenir compte des non-réponses, l'histogramme des poids de redressement...),

d) biais de réponse, c'est à dire mise en évidence de problèmes pouvant provenir du questionnaire et de son administration,

e) comparabilité des résultats avec des résultats antérieurs dans le cas d'enquêtes répétées dans le temps et facteurs pouvant affecter cette comparabilité,

f) comparabilité avec d'autres sources de données,

g) effet des opérations de saisie et d'imputation.

L'identification de tous ces indicateurs constitue une voie de recherche essentielle, la difficulté étant de les adapter au type d'enquêtes :

- Les enquêtes des Instituts Nationaux de Statistique, les grandes enquêtes quantitatives des instituts de sondage privés, les sondages d'opinion ne doivent pas avoir le même type d'indicateurs ;
- De la même façon, les enquêtes ponctuelles et les enquêtes répétées dans le temps doivent faire l'objet d'un traitement différent.

2.6 Conclusion : qu'est-ce qu'un échantillon représentatif ?

On trouve dans l'histoire des sondages (cf. Droesbeke et Tassi, 1990) et dans les débats des pères fondateurs de la théorie des sondages les interrogations et éclaircissements concernant la notion de représentativité.

Citons par exemple :

- La question de Kiaer en 1895 :
  "... De quelle manière un dénombrement représentatif doit-il être exécuté pour qu'il puisse présenter une miniature aussi correcte que possible de la société entière ?..."

- Le débat entre les tenants du "random sampling" (échantillonnage au hasard) et ceux de la "purposive selection" (choix raisonné)
- L'introduction par A.I. Tchuprow de l'allocation optimale dans un échantillon stratifié au prorata des effectifs de chaque strate mais aussi des écarts-types de la variable étudiée dans chaque strate, ce qui est en contradiction avec une des idées de la représentativité où l'échantillon est un "modèle réduit" de la population.

Dans une série de quatre articles, Kruskal et Mosteller (1979a, 1979b, 1979c et 1980) exposent les différents sens du concept de représentativité tels qu'on peut les trouver dans la littérature scientifique et non-scientifique. Dans la littérature statistique en particulier, ils relèvent neuf significations pour ce concept parmi lesquels nous citerons les six sens les plus courants :

1) Absence ou présence de biais de sélection dans le processus d'échantillonnage

   On considère souvent que pour éviter tout biais de sélection, chaque individu de la population doit avoir la même probabilité d'appartenir à l'échantillon (tirage à probabilités égales). Dans la première partie de l'exposé, on a vu que le tirage à probabilités inégales peut être une meilleure stratégie de sondage à condition d'utiliser le bon estimateur dans l'analyse des résultats.

   Une idée très répandue est que les biais de sélection peuvent être évités en construisant soigneusement un échantillon par choix raisonné comme dans la méthode des quotas. Mais il ne faut pas oublier qu'on ne peut multiplier les variables de quotas sous peine de rendre impossible le travail de l'enquêteur ; rien ne garantit donc qu'il n'y ait pas de biais de sélection sur les variables non contrôlées.

2) Miniature, modèle réduit de la population

   C'est l'idée de base de la méthode des quotas de construire un modèle réduit de la population selon quelques variables choisies comme quotas (ex : sexe, âge, CSP du chef de ménage dans une enquête sur individus en France). Mais rien ne garantit, comme nous l'avons déjà indiqué, une bonne représentativité sur les caractéristiques qui n'ont pas été prises comme quotas (ex : distance au chef-lieu le plus proche dans une enquête par enquêteur à domicile).

3) Cas typique ou idéal

Dans cette acception du terme, les individus de l'échantillon sont sélectionnés par choix raisonné et doivent avoir les valeurs les plus proches possible des valeurs moyennes de la population sur certaines caractéristiques. Ce type d'échantillon peut convenir dans le cas de taille d'échantillon faible mais ne permettra pas d'étudier la population dans sa diversité. D'autre part, rien ne garantit l'absence de biais importants sur des caractéristiques non contrôlées. Cette signification correspond à la méthode empirique des unités-types déjà citée. Elle a également été développée dans les théories d'échantillon équilibré et par les tenants de modèles dans les sondages.

4) Prise en compte de l'hétérogénéité de la population

Ce sens implique que chaque classe d'une partition appropriée de l'échantillon soit représentée par au moins un individu dans la population mais n'implique pas qu'elle soit représentée au prorata de sa taille. Ce type d'échantillon est couramment utilisé dans les études qualitatives.

5) Echantillon obtenu par une méthode aléatoire de sondage

Pour certains, un échantillon est représentatif s'il est obtenu par une des méthodes aléatoires de sondage qui ont été décrites au début du paragraphe 2.2. Un sens encore plus fréquent l'assimile à un échantillon aléatoire stratifié proportionnel (appelé souvent représentatif). En fait, le sens est trop restrictif à cause des problèmes de non-réponse, de questionnaire, qui peuvent affecter la qualité des résultats.

6) Echantillon permettant une bonne estimation

Cette utilisation du terme recouvre l'absence de biais, une erreur d'échantillonnage faible. C'est le sens que nous adoptons. En effet, l'objectif essentiel de tout sondage est de permettre d'étudier une population et ses caractéristiques. Un échantillon est représentatif s'il permet d'estimer les paramètres étudiés de la population avec des marges d'erreurs acceptables, étant donnés les objectifs de l'enquête.

La qualité de représentativité d'un échantillon dépasse donc le problème de la composition de l'échantillon et du choix de la méthode de sondage, mais dépend bien de toutes les phases de l'enquête (conception, réalisation, traitement).

3

LA MISE AU POINT D'UN QUESTIONNAIRE

Michel Lejeune

Ecole Nationale de la Statistique et de l'Administration Economique
Paris

3.1 Introduction

Le savoir-faire en matière de conception de questionnaires découle essentiellement de l'expérience et doit être considéré au moins autant comme un art qu'une science. Par ailleurs, contrairement à ce que l'on serait tenté de croire de prime abord, le simple bon sens, s'il est évidemment nécessaire, ne suffit pas. Il existe des règles, des principes et des techniques qui résultent d'observations et d'expérimentations nombreuses effectuées au cours des cinquante dernières années.

Les publications sont légions sur ce sujet. L'ouvrage le plus ancien que nous citons en référence bibliographique est le livre de S. L. Payne dont la première édition remonte à 1951 et qui fait encore autorité pour "l'art de poser des questions". Avec l'apparition de moyens techniques nouveaux s'appuyant sur l'informatique et les télécommunications, la mise à jour des connaissances dans le domaine est permanente. Pour la France le recueil d'articles "La qualité de l'information dans les enquêtes" édité en 1992, donne un panorama très actuel de l'état de l'art.

Dans cet exposé on verra que le ton général est certes normatif, mais en même temps teinté de nuances et de précautions. En effet aucune norme ne saurait être absolue tant il est vrai qu'il existe une variété infinie de situations, toutes particulières, selon le thème abordé par l'enquête, la population étudiée, les contraintes matérielles, le mode de recueil, les composantes culturelles et autres circonstances spécifiques.

3.2 Forme, contenu et mode d'enquête

Schématiquement nous pouvons distinguer les aspects relevant plutôt de la forme de ceux relevant plutôt du contenu. Pour la forme se posent les questions suivantes :

- Quelle architecture générale donner au questionnaire ?