Techniques d'enquête, décembre 1991
Vol. 17, n° 2, pp. 177–195
Statistique Canada

Une théorie des enquêtes par quotas

JEAN-CLAUDE DEVILLE[1]

RÉSUMÉ

Les enquêtes par quotas simples ou marginaux sont analysées par deux méthodes: (1) modélisation des comportements (modèle de superpopulation) et estimation par prédiction et (2) modélisation de l'échantillonnage (sondage aléatoire simple sous contraintes) et estimation dérivée de la distribution échantillonnale. Dans les deux cas on précise les limites de la théorie, à l'intérieur de laquelle on établit des formules de variance et d'estimation de variance quand on mesure des totaux. Une extension de la méthode des quotas (quotas non-proportionnels) est, au passage, décrite et analysée. Elle autorise, dans certains cas, une très nette amélioration de la précision des enquêtes. Les mérites de la méthode des quotas sont comparés à ceux de l'échantillonnage aléatoire. Ce dernier reste indispensable dans le cas d'enquêtes de grande taille dans le cadre de la Statistique officielle.

MOTS CLÉS: Enquêtes par quotas; modèles de superpopulation; échantillonnage contraint; estimation par régression.

1. INTRODUCTION

L'échantillonnage par quotas est la méthode la plus fréquemment utilisée en France par les Instituts de sondage privés. Facile de mise en oeuvre, peu coûteuse, elle possède de nombreux avantages pratiques. Ses défauts, toutefois, sont aussi assez bien connus: possibilités de biais, impossibilité de traiter les non-réponses, nécessité d'une information externe pour fixer les quotas. Dans la littérature anglo-saxonne (Cochran 1977 ou Madow et coll. 1983 par exemple) les quotas ont fort mauvaise réputation à cause de l'absence d'une théorie fiable sur laquelle une inférence statistique puisse être fondée. Les "défenseurs" de la méthode (Smith 1983 en particulier) se basent sur des principes d'inférence conditionnelle à l'échantillon où le plan de sondage peut, généralement, être oublié.

Cet article propose une théorie des enquêtes par quotas basée sur deux types de modélisation: modélisation du comportement de la population (qui est l'optique de Smith ou des idées exprimées dans Gouriéroux 1981), modélisation du mode de recueil de l'échantillon, ce qui correspond, peut-être, à une idée plus réaliste.

Dans chaque cas, on obtient les variances des estimateurs en se ramenant des variantes d'estimateurs par régression.

L'article commence par une description de la méthode des quotas et des résultats de théorie des sondages utiles pour la suite. Les parties 2 et 3 développent des modèles de comportement des individus de la population ou des enquêteurs, qui justifient la méthode. La dernière partie évoque des problèmes ouverts et montre en quoi la méthode des quotas complète les méthodes probabilistes classiques plus qu'elle ne les concurrence.

[1] Jean-Claude Deville, Institut National de la Statistique et des Études Économiques, 18, Boulevard Adolphe Pinard, 75675, Paris Cedex 14, France.

2. UNE REVUE RAPIDE DE LA MÉTHODE DES QUOTAS ET DE LA THÉORIE DES SONDAGES

2.1 Quotas sur des cellules; quotas sur les marges d'une table de contingence - quelques aspects pratiques de la méthode

Au niveau le plus simple, la méthode des quotas ressemble à l'échantillonnage stratifié. On connaît la répartition dans la population d'un caractère discret h que N_h individus possèdent ($h = 1$ à H). L'échantillon comporte n_h individus de catégorie h dont le choix est laissé aux enquêteurs. Le taux de sondage $f_h = n_h/N_h$ peut, éventuellement, varier de catégorie en catégorie.

En pratique, on préfère, généralement, contrôler plusieurs critères qu'on notera $i, j, \ldots, h$ ($i = 1$ à $I, j = 1$ à $J, \ldots, h = 1$ à H). Idéalement, la connaissance des effectifs $N_{ij\ldots h}$ du tableau de contingence à entrées multiples permettrait de se ramener à la méthode précédente pour la définition d'effectifs $n_{ij\ldots h}$ composant l'échantillon selon des taux $f_{ij\ldots h}$. Sauf dans des cas très particuliers (peu de critères ayant chacun peu de modalités) cette méthode est irréaliste car elle conduit à la recherche d'individus extrêmement difficiles à détecter.

On préfère fixer des **quotas marginaux** en calibrant l'échantillon de façon à ce que sa répartition selon le premier critère conduise à des effectifs $n_{i+\ldots+}$ donnés, de même en ce qui concerne les autres critères. La seule contrainte sur ces effectifs marginaux est de s'additionner à n, effectif global de l'échantillon. Pratiquement, on adopte presque toujours un taux de sondage f unique pour chaque batterie de quotas: $n_{i+\ldots+} = fN_{i+\ldots+}$, $n_{+j\ldots+} = fN_{+j\ldots+}$ et $n_{++\ldots h} = fN_{++\ldots h}$ avec des notations évidentes (+ à la place d'un indice indique la sommation sur toutes les modalités de la catégorie notée par l'indice).

Outre son avantage évident de collecte, cette technique est le plus souvent imposée par les données externes sur lesquelles on asseoit les quotas. Celles-ci proviennent, par exemple, de sources différentes interdisant tout croisement. Une autre situation se présente quand les quotas sont établis à partir d'une grosse enquête (par exemple sur l'emploi): chaque répartition selon un critère (âge, catégorie socio-professionnelle *etc.*) peut être considérée comme fiable. En revanche, les croisements sont entachés d'une erreur aléatoire importante et ne peuvent pas être utilisés pour fixer des quotas.

En pratique, la méthode des quotas est le plus souvent utilisée en complément de méthodes plus traditionnelles comme ultime technique d'échantillonnage dans une enquête stratifiée à plusieurs degrés sur des bases géographiques (région, taille des agglomérations). Chaque unité primaire est confiée à un enquêteur à qui sont fixés des quotas. Celui-ci reçoit, de plus, des consignes destinées à disperser son échantillon de façon à rapprocher le recueil des données de ce qu'aurait donné le hasard.

2.2 La théorie traditionnelle des sondages

On désire mesurer le total Y d'une variable dont la valeur Y_k pour l'individu k n'a rien d'aléatoire. Seul l'échantillon s est aléatoire et sa loi de probabilité est connue car contrôlée par le statisticien. Par suite, la probabilité π_k qu'a chaque individu d'apparaître dans s est aussi connue. Sans autre information, l'estimateur naturel (sans biais) à retenir est l'estimateur par les valeurs dilatées:

$$\hat{Y} = \sum_{k \in s} Y_k/\pi_k = \sum_s d_k Y_k \quad \text{avec} \quad d_k = 1/\pi_k.$$

Dans le cas où les π_k sont tous égaux à n/N, le taux de sondage, on a:

$$\hat{Y} = N/n \sum_s Y_k = N\bar{y},$$

où $\bar{y}$ désigne la moyenne de Y sur l'échantillon.

Cet estimateur a une variance connue qui est une forme quadratique $V(Y_U)$ sur le vecteur des Y_k dans la population:

$$\text{Var}(\hat{Y}) = V(Y_U) = \sum_k Y_k(d_k - 1) + \sum_{kl} Y_k Y_l d_k d_l(\pi_{kl} - \pi_k \pi_l), \quad (2.2.1)$$

où π_{kl} est la probabilité d'avoir simultanément k et l dans s.

De même on peut estimer sur les données la variance de $\hat{Y}$ par une forme quadratique formée sur le vecteur Y_s des Y_k de l'échantillon:

$$\hat{V}(Y_s) = \sum_{kl\epsilon s} \Delta_{kl} Y_k Y_l,$$

avec

$$\Delta_{kl} = (1 - \pi_k)/\pi_k^2 \quad \text{si} \quad k = l$$

$$= (\pi_{kl} - \pi_k \pi_l)/(\pi_{kl}\pi_k\pi_l) \quad \text{si} \quad k \neq l.$$

Selon les plans de sondage ces expressions prennent des formes particulières qu'on trouve dans les manuels (Desabie 1965, Cochran 1977, Wolter 1985).

Toute information externe peut améliorer la qualité de l'estimation. Celle-ci, souvent, se présente sous la forme d'un vecteur X dont chacune des p composantes est le total d'une variable mesurable dans chacun des échantillons possibles. On peut alors améliorer l'estimation de Y en utilisant l'estimation par régression:

$$\hat{Y}_{\text{Reg}} = \hat{Y} + (X - \hat{X})'\hat{B},$$

où B est le vecteur des coefficients de la régression des Y_k sur les X_k estimé par:

$$\hat{B} = \sum_s (d_k X_k X_k')^{-1} \sum_s d_k X_k Y_k.$$

Dans le cas où la constante fait partie des régresseurs où si elle est combinaison linéaire des régresseurs et que l'échantillon est à probabilités égales la formule se simplifie en:

$$\hat{Y}_{\text{Reg}} = X'\hat{B}.$$

La variance de $\hat{Y}_{\text{Reg}}$ s'exprime simplement en introduisant les résidus de la régression $E_k = Y_k - X_k'B$ dans la population. Il est connu qu'on a:

$$\text{Var}(\hat{Y}_{\text{Reg}}) = V(E_U)$$

où on porte donc dans la formule (2.2.1) le vecteur E_U des résidus E_k. De même on estime approximativement cette variance par $\hat{V}(e_s)$ où e_s est le vecteur des $e_k = Y_k - X_k'\hat{B}$, résidus estimés de la régression.

Sous divers plans de sondage ces expressions prennent des allures particulières. En règle générale, V et $\hat{V}$ étant des formes quadratiques positives et les E_k ou e_k des quantités plus petites que les Y_k, l'estimateur par régression conduit à des améliorations substantielles par rapport aux valeurs dilatées.

Un cas particulier important que nous utiliserons dans la suite est celui où X est un vecteur de totaux de variables de comptage (effectifs à partir desquels on construit des quotas). Typiquement l'information auxiliaire est le vecteur de dimension $I + (J - 1) + \ldots + (H - 1)$

formé par les quantités: $N_{i++\ldots+}$, $N_{+j+\ldots+}$, $N_{+\ldots+h}$ pour $i = 1$ à I, $j = 1$ à $J - 1$, et $h = 1$ à $H - 1$ (de façon à ne conserver que des variables linéairement indépendantes). Les régresseurs sont alors les variables indicatrices des catégories $i (i = 1 \text{ à } I), j (j = 1 \text{ à } J - 1)$ et $h = 1$ à $(H - 1)$. Comme la constante est combinaison linéaire des régresseurs (c'est la somme de I premiers d'entre eux) l'estimateur par régression prendra la forme:

$$\hat{Y}_{\text{Reg}} = \sum_i N_{i++\ldots+} \hat{A}_i + \sum_j N_{+j\ldots+} \hat{B}_j + \ldots + \sum_h N_{++\ldots h} \hat{C}_h, \qquad (2.2.2)$$

où $\hat{A}_i$ (par exemple) est le coefficient de l'indicatrice d'appartenance à la catégorie i.

Si on ne travaille qu'avec une seule catégorisation les régresseurs sont 2 à 2 orthogonaux et on a:

$$\hat{Y}_{\text{Reg}} = \sum_i N_i \hat{\bar{Y}}_i$$

où $\hat{\bar{Y}}_i$ est l'estimateur de la moyenne de Y dans la catégorie i. $\hat{Y}_{\text{Reg}}$ n'est alors pas autre chose que l'estimateur poststratifié.

2.3 Théorie des sondages basée sur des modèles

Dans cette optique, on considère les Y_k comme des variables aléatoires régies par un modèle de superpopulation. Celui-ci comporte des paramètres qu'on estime à partir de l'échantillon. On peut alors calculer l'espérance, sous le modèle estimé, des valeurs non observées de Y, soit $\hat{Y}_k$. L'estimateur par prédiction, somme des valeurs observées et des valeurs prévues est donné par:

$$\hat{Y}_{\text{Pred}} = \sum_s Y_k + \sum_{U-s} \hat{Y}_k.$$

Si, par exemple, dans un sondage à probabilités égales, le modèle est une régression $Y_k = X_k' \cdot \beta + \epsilon_k$, ϵ_k indépendantes, centrées, d'égales variances et que la constante figure dans la régression (ou qu'une combinaison linéaire de X_k soit constante) alors on a $\sum_s Y_k = \sum_s X_\kappa' \beta$ et l'estimateur par la prédiction se confond avec l'estimateur par la régression.

On dira que $\hat{Y}$ est sans biais sous le modèle si, pour tout s, $\mathcal{E}(\hat{Y} - Y) = 0$ (on note $\mathcal{E}$ et $\mathcal{V}$ l'espérance et la variance sous le modèle, conditionnellement à l'échantillon). Pour l'estimateur par prédiction il suffit qu'on ait pour tout k la condition naturelle $\mathcal{E}\hat{Y}_k = \mathcal{E}Y_k$ pour que cela soit vrai. Sous le modèle on peut évaluer également l'écart quadratique moyen: $\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2$ sachant que les deux termes $\hat{Y}_{\text{Pred}}$ et Y sont aléatoires et que $\hat{Y}_{\text{Pred}}$ dépend de l'échantillon s. L'espérance ci-dessus est donc une espérance conditionnelle à l'échantillon s. Celui-ci a une certaine loi de probabilité déjà introduite au paragraphe précédent. On peut mesurer la précision de cet estimateur en calculant:

$$\mathcal{V}(\hat{Y}_{\text{Pred}}) = E\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2.$$

Si la loi de s et celle des Y_k sont indépendantes (échantillonnage dit non informatif) alors cette quantité est égale à:

$$\mathcal{E}(E(\hat{Y}_{\text{Pred}} - Y)^2),$$

où l'espérance intérieure est prise conditionnelle aux Y_k. Si $\hat{Y}_{\text{Pred}}$ s'identifie à $\hat{Y}_{\text{Reg}}$, et que la condition d'indépendance est réalisée, on aura donc:

$$\mathcal{V}(\hat{Y}_{\text{Pred}}) = \mathcal{E}(\text{Var}(\hat{Y}_{\text{Reg}})).$$

2.4 Remarques sur les deux optiques appliquées à la méthode des quotas

a) Dans les deux cas, l'estimation sera efficace si la variable d'intérêt est bien expliquée par les indicatrices des catégories sur lesquelles on fonde les quotas, grosso modo, parce que les résidus d'ajustement de la régression seront petits.

b) Dans une enquête par quotas le "plan de sondage" est inconnu du statisticien. Celui-ci ne peut donc faire d'inférence sans recourir à un modèle. Ce peut être un modèle de comportement de la population (optique "modèle") qui l'oblige à prendre des responsabilités vis-à-vis de la nature de ce qu'il observe. Ce point de vue sera développé dans la troisième partie de ce papier. Ce peut être aussi une modélisation du plan de sondage, ce qui veut dire une prise de responsabilité vis-à-vis du fonctionnement du processus de collecte. Ce point de vue sera développé dans la quatrième partie de l'article.

Dans tous les cas la spéculation modélisatrice doit être mobilisée pour valider une forme d'inférence. La question est de savoir s'il est plus facile et plausible de modéliser le comportement des individus qu'on sonde que de modéliser le processus de recueil de l'échantillon (y compris dans ses aspects de contact entre enquêteur et enquêté).

c) À cet égard l'hypothèse faite en 2.3 d'indépendance entre des aléas dans la population et des aléas dans le processus de collecte est **cruciale**. Si l'échantillonnage est contrôlé par les statisticiens, cette hypothèse est garantie, sauf effet des non-réponses. Dans le cas de la méthode des quotas on n'a aucune garantie. Supposons par exemple que l'on désire mesurer des revenus Y_k; la probabilité π_k de trouver k dans l'échantillon peut être très diminuée si Y_k est grand. Autrement dit l'appartenance à l'échantillon (variable qui vaut 1 si k est dans s et 0 sinon) et le résidu du modèle de superpopulation ϵ_k sont corrélés négativement. Cet exemple illustre bien le principal danger de la méthode des quotas; la théorie qui suit n'en tient pas compte.

3. THÉORIE DES QUOTAS AVEC MODÈLE DE SUPERPOPULATION

3.1 Quotas par cellule

On a une seule catégorisation en cellules $i = 1$ à I d'effectifs connus N_i. Le modèle qu'on peut imaginer est le suivant:

$$Y_k = m_i + \epsilon_k, \tag{3.1.1}$$

les ϵ_k sont centrées indépendantes de variance σ_i^2 et i est la cellule à laquelle appartient k.

Les estimateurs de Gauss-Markov des m_i sont les moyennes observées dans les différentes cellules $\bar{y}_i$. L'estimateur par prédiction vaut alors:

$$\hat{Y}_{\text{Pred}} = \sum_i (N_i - n_i)\,\bar{y}_i + \sum_i n_i \bar{y}_i = \sum_i N_i \bar{y}_i. \tag{3.1.2}$$

Il a la forme de l'estimateur poststratifié. On obtient de plus, immédiatement que:

$$\text{Var}(\hat{Y}_{\text{Pred}} - Y)^2 = \sum_i \sigma_i^2 N_i (N_i - n_i)/n_i. \tag{3.1.3}$$

Cette quantité ne dépend pas de l'échantillon s puisque celui-ci comporte toujours (avec probabilité 1 !) n_i individus de la cellule i.

L'estimation de $E\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2$ se fait en remplaçant σ_i^2 par son estimateur habituel $s_i^2 = (n_i - 1)^{-1} \sum_{k \in s_i} (Y_k - \bar{y}_i)^2$ avec s_i partie de s dans la cellule i.

Ces résultats sont dus à Gouriéroux (1981) et constituent, dans une certaine mesure, une justification de la méthode des quotas simples.

3.2 Quotas marginaux - Cas "représentatif"

Dans ce paragraphe et dans toute la suite, nous nous bornerons au cas de quotas croisant 2 critères i et j. La généralisation à plus de 2 critères ne pose aucun problème particulier mais génère des notations très lourdes auxquelles on a préféré renoncer (voir l'annexe).

La situation est donc la suivante: les effectifs N_{i+} et N_{+j} des deux ventilations de l'univers sont connues. L'échantillonnage n'autorise que des échantillons de taille fixe $n = fN$ comportant pour chaque i $n_{i+} = fN_{i+}$ individus et pour chaque j $n_{+j} = fN_{+j}$ individus.

Nous postulons dans la population, un modèle de type analyse de la variance formulé de la façon suivante:

Si k appartient à la cellule (i,j):

$$Y_k = \alpha_i + \beta_j + \epsilon_k. \tag{3.2.1}$$

Les ϵ_k sont centrées, indépendantes et on a Var $\epsilon_k = \sigma_i^2 + \gamma_j^2$.

Pour des raisons d'identification du modèle on pose $\beta_J = 0$.

Ceci équivaut à poser $Y_k = (\alpha_i + u_{ik}) + (\beta_j + v_{jk})$ où les u_{ik} et les v_{jk} son indépendants et de variance respectives σ_i^2 et τ_j^2.

On estime les α_i et β_j par la méthode des moindres carrés ordinaires (MCO) car on ignore les valeurs des éléments de la variance; les $\hat{\alpha}_i$ et $\hat{\beta}_j$ sont solutions du système:

$$\begin{aligned} \sum_j n_{ij}\bar{y}_{ij} &= n_{i+}\hat{\alpha}_i + \sum_j n_{ij}\hat{\beta}_j \quad (i = 1 \text{ à } I) \\ \sum_i n_{ij}\bar{y}_{ij} &= n_{+j}\hat{\beta}_j + \sum_i n_{ij}\hat{\alpha}_i \quad (j = 1 \text{ à } J-1), \end{aligned} \tag{3.2.2}$$

avec $\bar{y}_{ij}$ moyenne des Y_k sur s_{ij} partie de l'échantillon dans la cellule (i,j). L'estimateur par prédiction s'écrit alors:

$$\hat{Y}_{\text{Pred}} = \sum_{ij} (N_{ij} - n_{ij})(\hat{\alpha}_i + \hat{\beta}_j) + \sum_{ij} n_{ij}\bar{y}_{ij}.$$

Résultat 1: Sous le modèle (3.2.1), l'estimateur par prédiction utilisant les MCO est $N\bar{y}$. On vérifie qu'il est sans biais pour le modèle c'est à dire que $\mathcal{E}(N\bar{y} - Y) = 0$.

Preuve: immédiate à partir de (3.2.2) et du fait que les quotas sont proportionnels aux effectifs dans la population.

Résultat 2: On a:

$$\mathcal{E}(N\bar{y} - Y)^2 = (N^2/n)(1-f)n^{-1}\left(\sum_i n_{i+}\sigma_i^2 + \sum_j n_{+j}\tau_j^2\right).$$

Cette quantité ne dépend pas de l'échantillon (puisque qu'elle ne dépend que des quotas). On a donc là, dans une certaine mesure, une justification de la méthode des quotas marginaux.

Preuve: Avec $m_k = \mathcal{E}Y_k$ on a, en utilisant le caractère non biaisé de l'estimateur:

$$\begin{aligned} \mathcal{E}(N\bar{y} - Y)^2 &= \mathcal{E}\left((N/n)\sum_s (Y_k - m_k) - \sum_U (Y_l - m_l)\right)^2 \\ &= \mathcal{E}\left((N/n)\sum_s \epsilon_k - \sum_U \epsilon_l\right)^2 \\ &= (N/n)^2 \sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2) - 2(N/n)\sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2) + \sum_{ij} N_{ij}(\sigma_i^2 + \tau_j^2). \end{aligned}$$

Mais

$$\sum_{ij} N_{ij}(\sigma_i^2 + \tau_j^2) = \sum_i N_{i+}\sigma_i^2 + \sum_j N_{+j}\tau_j^2$$

$$= (N/n)\left(\sum_i n_{i+}\sigma_i^2 + \sum_j n_{+j}\tau_j^2 \right)$$

d'où:

$$\mathcal{E}(N\bar{y} - Y)^2 = (N^2/n)(1 - f)n^{-1}\left(\sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2) \right)$$

$$= (N^2/n)(1 - f)\left(\sum_i p_{i+}\sigma_i^2 + \sum_j p_{+j}\tau_j^2 \right)$$

$$\text{avec} \quad p_{i+} = N_{i+}/N \quad \text{et} \quad p_{+j} = N_{+j}/N.$$

L'estimation de la précision de $E\,(N\bar{y} - Y)^2$ en découle. En effet s_{ij}^2 a pour espérance sous modèle $\sigma_i^2 + \tau_j^2$. On obtient donc un estimateur sans biais de la précision par

$$(N/n)^2\,(1 - f) \sum_{ij} n_{ij} s_{ij}^2$$

si tous les n_{ij} sont supérieurs ou égaux à 2.

Cet estimateur est, formellement, identique à celui qu'on utiliserait dans une poststratification complète sur les cellules (i,j). On peut aussi utiliser $(N/n)^2\,(1 - f)\,\sum_s e_k^2$ où les e_k sont les résidus estimés du modèle.

3.3 Et si le modèle est faux?

3.3.1 Une première façon de voir la question est de plonger le modèle (3.2.1) dans le modèle général où la moyenne de Y_k dépend du couple (i,j). On peut écrire cela sous la forme:

$$Y_k = \alpha_i + \beta_j + \gamma_{ij} + \epsilon_k, \tag{3.3.1.1}$$

avec les hypothèses habituelles sur les ϵ_k et les termes d'interaction γ_{ij} qui vérifient des contraintes d'identifiabilité:

$$\sum_j N_{ij}\,\gamma_{ij} = 0 \quad \text{et} \quad \sum_i N_{ij}\,\gamma_{ij} = 0. \tag{3.3.1.2}$$

On a alors, de façon immédiate:

$$\mathcal{E}(N\bar{y} - Y) = \sum_{ij} (Nn_{ij}/n - N_{ij})\gamma_{ij}, \tag{3.3.1.3}$$

de sorte que l'estimateur est biaisé pour le modèle sauf si $n_{ij} = fN_{ij}$ ce qui n'a aucune raison d'être réalisé.

Ceci dit les termes de la somme (3.3.1.3) peuvent très bien se compenser, avec un peu de chance, car leurs signes sont *a priori* indéterminés.

D'autre part, si de "bonnes" précautions d'échantillonnage sont prises, $Nn_{ij}/n - N_{ij}$ devrait être voisin de 0 assez souvent.

Il est clair, en tout cas, que mieux le modèle additif "colle" (γ_{ij} petits) et plus le plan de sondage se rapproche de l'aléatoire, plus le biais a des chances de se réduire.

3.3.2 Une autre façon d'envisager la fausseté du modèle, déjà signalée, est de ne plus admettre l'indépendance entre l'aléa d'échantillonnage et l'aléa du modèle additif. Ceci revient à dire que des modèles distincts doivent être développés pour des vecteurs $(Y_k,\ k \in s)$ et $(Y_l,\ l \notin s)$. Cette façon de voir les choses a été souvent employée dans la littérature économétrique à laquelle nous renvoyons le lecteur. Il est clair que la prise de risque vis à vis des données devient alors énorme et, souvent, incompatible avec un travail objectif de statisticien.

3.4 Quotas marginaux à taux inégaux

Dans le cas de quotas par cellules on peut fixer arbitrairement les quotas de chaque cellule. Jusqu'ici, dans le cas de quotas marginaux, nous n'avons envisagé que le cas où les quotas étaient proportionnels aux effectifs de la population.

Dans de nombreux cas, toutefois, on est tenté de surreprésenter certaines catégories. Si on désire étudier, par exemple, les patrimoines des ménages, on désirera fixer des quotas plus importants pour les ménages âgés d'une part (quotas par groupes d'âge) et pour ceux dont le chef est travailleur indépendant (quotas par catégories sociales).

Formellement, on impose donc à l'échantillon de respecter des effectifs n_{i+} et n_{+j} *a priori* quelconques (avec toutefois la somme des n_{i+} égale à la somme des n_{+j}).

Dans ce cas, en utilisant toujours les MCO comme technique d'estimation, on trouve facilement que l'estimateur par prédiction du total est:

$$\hat{Y}_{\text{Pred}} = \sum_i N_{i+}\hat{\alpha}_i + \sum_j N_{+j}\hat{\beta}_j, \tag{3.4.1}$$

les $\hat{\alpha}_i$ et $\hat{\beta}_j$ vérifiant toujours les équations estimantes (3.2.2). Il est facile de voir que cet estimateur peut se mettre sous la forme:

$$\hat{Y}_{\text{Pred}} = \sum_{ij} (w_i^{(1)} + w_j^{(2)})\, n_{ij}\bar{y}_{ij} = \sum_{ij} \hat{N}_{ij}\bar{y}_{ij}.$$

Les quantités $(w_i^{(1)} + w_j^{(2)})\, n_{ij}$ apparaissent donc comme des estimations des effectifs des cellules (i,j), idée qui sera largement exploitée dans la suite.

En revanche, la variance sous modèle de cet estimateur dépend de l'ensemble des n_{ij}, comme le montre un calcul un peu laborieux. La justification de la méthode des quotas évoquée précédemment ne fonctionne plus.

4. MODÈLES POUR LE PLAN DE SONDAGE

4.1 Un modèle de plan de sondage

L'idée est celle d'un sondage aléatoire simple (S.A.S) contraint par les quotas imposés. L'algorithme de tirage, tout à fait utopique, serait de tirer une suite d'échantillons aléatoires simples jusqu'à ce qu'on en rencontre un qui vérifie les quotas. Ainsi, chaque échantillon vérifiant les quotas à la même probabilité positive d'être tiré, les échantillons ne vérifiant pas les quotas ayant une probabilité nulle.

Cette vue de l'esprit cherche à modéliser le fait qu'un enquêteur va suivre correctement les consignes de dispersion des unités sondées données par son encadrement.

4.2 Quotas par cellule

Ce modèle d'échantillonnage se ramène à la stratification à *priori*. Son avantage pratique est de ne pas nécessiter une base de sondage où sont présentes les variables de stratification. Il est mis en oeuvre rigoureusement dans certains cas, par exemple celui d'un sondage par téléphone où on part d'une liste aléatoire de numéros non informatifs et où on réalise des enquêtes jusqu'à ce que les quotas soient satisfaits.

Les formules donnant estimateurs, variances et estimations de précision sont donc celles qu'on trouve dans tous les manuels. Elles présentent une analogie certaine avec celles données en 3.1 (voir Gouriéroux 1981).

4.3 Cas des quotas marginaux: généralités-estimateurs

Le modèle d'échantillonnage est celui du sondage aléatoire simple contraint par les quotas marginaux. Le S.A.S fournit des échantillons comportant des effectifs n_{ij} dans les différentes cellules qu'on peut voir comme un vecteur aléatoire (à valeurs entières) dans R^{IJ}. La contrainte des quotas signifie qu'on se limite au vecteur aléatoire conditionné par:

$$\sum_j n_{ij} = n_{i+} \ (i = 1 \text{ à } I) \quad \text{et} \quad \sum_i n_{ij} = n_{+j} \ (j = 1 \text{ à } J - 1),$$

c'est-à-dire variant dans un sous espace de dimension $IJ - I - J + 1$. Nous nous plaçons dans le cas où le taux de sondage global est négligeable et où la loi des n_{ij} peut être assimilée à une loi multinomiale $(n, p_{ij} = N_{ij}/N)$.

Conditionnellement aux n_{ij}, les $\bar{y}_{ij}$ estiment sans biais les $\bar{Y}_{ij}$. L'idée est maintenant de construire un estimateur du total de Y en pondérant les $\bar{y}_{ij}$ par des estimateurs des N_{ij}, c'est-à-dire des p_{ij}. Si on choisit de maximiser la vraisemblance, celle-ci est proportionnelle à:

$$\prod_{ij} p_{ij}^{n_{ij}}. \tag{4.3.1}$$

On doit donc maximiser

$$\sum_{ij} n_{ij} \operatorname{Log} p_{ij} \tag{4.3.2}$$

sous les contraintes

$$\sum_j p_{ij} = p_{i+} \ (i = 1 \text{ à } I) \quad \text{et} \quad \sum_i p_{ij} = p_{+j} \ (j = 1 \text{ à } J - 1) \tag{4.3.3}$$

ce qui amène à résoudre le système en a_i, b_j (les $p_{i+} = N_{i+}/N$ et $p_{+j} = N_{+j}/N$ sont connus):

$$\begin{aligned} &\sum_j \hat{p}_{ij}^{\circ} \, (a_i + b_j)^{-1} = p_{i+} \quad (i = 1 \text{ à } I) \\ &\sum_i \hat{p}_{ij}^{\circ} \, (a_i + b_j)^{-1} = p_{+j} \quad (j = 1 \text{ à } J - 1; \, b_J = 0), \end{aligned} \tag{4.3.4}$$

avec $\hat{p}_{ij}^{\circ} = n_{ij}/n$ fréquence observée sur l'échantillon.

Les estimateurs des p_{ij} sont alors les $\hat{p}^{\circ}_{ij}\ (a_i + b_j)^{-1}$ et l'estimateur cherché s'écrit:

$$\hat{Y}_Q = (N/n) \sum_{ij} n_{ij}(a_i + b_j)^{-1} \bar{y}_{ij} = (N/n) \sum_{s} w_k Y_k, \tag{4.3.5}$$

où $w_k = (a_i + b_j)^{-1}$ est la pondération à appliquer à Y_k dans le cas ou k appartient à la cellule (i,j). Cet estimateur est asymptotiquement sans biais sous le modèle de S.A.S. dans U comme le sont les estimateurs du maximum de vraisemblance. Les quotas n'interviennent pas de façon explicite dans 3.3.4 mais ils influent sur les valeurs des a_i et b_j.

Dans le cas habituel où les quotas marginaux sont "proportionnels" avec une fraction de sondage fixe f, la solution des équations 4.3.4 est évidente: $a_i = 1$ pour tout i et $b_j = 0$ pour tout j. L'estimateur du total vaut $N\bar{y}$, comme on pouvait s'y attendre et à la même expression que pour un sondage probabiliste à probabilités égales.

Remarque. L'utilisation du maximum de vraisemblance pour estimer les proportions est assez arbitraire. Un critère du type chi-2 (minimiser $\sum_{ij} (p_{ij} - \hat{p}^{\circ}_{ij})^2/\hat{p}^{\circ}_{ij}$) rendrait linéaire le système (4.3.4).

4.4 La variance de l'estimateur et son estimation

4.4.1 Pour établir une formule de variance nous utiliserons la paramétrisation de la variable Y utilisée dans Deville et Särndal (1990) que nous énonçons sous forme d'un:

Lemme: Pour toute variable $Y = (Y_k; k \in U)$ on peut choisir une paramétrisation définie de façon unique

$$Y_k = \bar{Y}_{ij} + R_k \quad \text{si } k \text{ est dans la cellule } (i,j) \quad (k \in U_{ij}) \quad \text{avec} \sum_{k \in U_{ij}} R_k = 0,$$

$$\bar{Y}_{ij} = A_i + B_j + E_{ij} \quad \text{avec} \quad B_J = 0$$

$$\sum_{j} N_{ij} E_{ij} = 0 \quad i = 1 \text{ à } I$$

$$\sum_{i} N_{ij} E_{ij} = 0 \quad j = 1 \text{ à } J - I.$$

De fait les A_i et B_j sont les nombres qui minimisent la quantité $\sum_U (Y_k - A_i - B_j)^2$ où, de façon équivalente, $\sum_{ij} N_{ij} (\bar{Y}_{ij} - A_i - B_j)^2$.

On peut alors écrire:

$$\hat{Y}_Q = (N/n) \sum_{ij} n_{ij} (a_i + b_j)^{-1} (A_i + B_j + E_{ij} + \bar{R}_{ij}) \quad \text{où} \quad \bar{R}_{ij} = \sum_{s_{ij}} R_k/n_{ij}.$$

Compte tenu des équations 4.3.4 et du lemme on en déduit:

$$\hat{Y}_Q - Y = \sum_{ij} \hat{N}_{ij} (E_{ij} + \bar{R}_{ij}) \quad \text{avec} \quad \hat{N}_{ij} = (N/n)\, n_{ij}\, (a_i + b_j)^{-1}, \tag{4.4.1}$$

qui est l'expression de base pour le calcul de variance.

Conditionnellement aux n_{ij}, les $\hat{N}_{ij}$ sont constants et les sous-échantillons s_{ij} des sondages aléatoires simples indépendants. On a donc:

$$\text{Biais cond}(\hat{Y}_Q) = \sum_{ij} \hat{N}_{ij} E_{ij} = N \sum_{ij} \hat{p}_{ij} E_{ij}$$

$$\text{Var cond}(\hat{Y}_Q) = \sum_{ij} \hat{N}_{ij}^2 V_{ij}/n_{ij} \quad \text{où} \quad V_{ij} = (1/N_{ij}) \sum_{U_{ij}} R_k^2.$$

Or (démonstration en annexe) on a le:

Résultat 1:

$$\text{Var}\left(\sum_{ij} \hat{p}_{ij} E_{ij} \right) = 1/n \sum_{ij} p_{ij} E_{ij}^2 .$$

Par ailleurs, l'espérance de $\hat{p}_{ij}^{\circ}(a_i + b_j)^{-1}$ vaut, (à des termes en $1/n$ près) $p_{ij}(a_i^{\circ} + b_j^{\circ})^{-1}$ où a_i° et b_j° sont solutions des équations (4.3.4) dans lesquelles on remplacerait les $\hat{p}_{ij}^{\circ}$ par les p_{ij} exacts.

D'où le résultat:

Résultat 2: La variance de l'estimateur des quotas $\hat{Y}_Q$ est donnée par:

$$\text{Var}(\hat{Y}_Q) = (N^2/n) \sum_{ij} p_{ij}(E_{ij}^2 + (a_i^{\circ} + b_j^{\circ})^{-1} V_{ij}).$$

Si les quotas sont proportionnels aux effectifs dans la population, on aura:

$$\text{Var}(\hat{Y}_Q) = (N^2/n) \sum_{ij} p_{ij}(E_{ij}^2 + V_{ij}).$$

4.4.2 Estimation de la variance

La variance conditionnelle de $\hat{Y}_Q$ s'estime immédiatement par:

$$\sum_{ij} \hat{N}_{ij}^2 s_{ij}^2/n_{ij} = (N^2/n) \sum_{ij} \hat{p}_{ij}(a_i + b_j)^{-1} s_{ij}^2,$$

où s_{ij}^2 est l'estimateur sans biais habituel de V_{ij}. L'espérance du carré du biais conditionnel vaut $(N^2/n) \sum_{ij} p_{ij} E_{ij}^2$ et s'estime par $(N^2/n) \sum_{ij} \hat{p}_{ij} \hat{E}_{ij}^2$ où $\hat{E}_{ij} = \bar{y}_{ij} - \hat{A}_i - \hat{B}_j$ avec $\hat{A}_i$ et $\hat{B}_j$ solutions de:

$$\begin{aligned} &\sum_j \hat{p}_{ij} (\hat{A}_i + \hat{B}_j) = \sum_j \hat{p}_{ij} \bar{y}_{ij} \quad (i = 1 \text{ à } I), \\ &\sum_j \hat{p}_{ij} (\hat{A}_i + \hat{B}_j) = \sum_j \hat{p}_{ij} \bar{y}_{ij} \quad (j = 1 \text{ à } J - I) \quad \text{avec} \quad B_J = 0. \end{aligned} \tag{4.4.2}$$

Autrement dit on obtient l'estimation des E_{ij} en ajustant aux données un modèle ANOVA additif sans interaction, le critère d'ajustement étant celui des moindres carrés pondérés par les poids $(a_i + b_j)^{-1}$.

L'estimateur de la variance est alors:

$$\widehat{\mathrm{Var}}(\hat{Y}_Q) = (N^2/n) \sum_{ij} \hat{p}_{ij} (\hat{E}_{ij}^2 + (a_i + b_j)^{-1} s_{ij}^2). \tag{4.4.3}$$

Dans le cas des quotas proportionnels aux effectifs cette expression se simplifie en:

$$(N^2/n) \sum_{ij} n_{ij} (\hat{E}_{ij}^2 + s_{ij}^2)/n. \tag{4.4.4}$$

Si les n_{ij} sont tous suffisamment grands pour qu'on puisse prendre $n_{ij}/(n_{ij} - 1) = 1$ on voit que la somme de la formule n'est autre que la somme des carrés des résidus estimés dans l'ajustement par les MCO du modèle $Y_k = A_i + B_j +$ résidu. La procédure d'estimation est alors simple:

- ajuster par les MCO le modèle additif sur données individuelles
- créer la variable e_k des résidus estimés
- $\widehat{\mathrm{Var}}(\hat{Y}_Q) = (N^2/n) \cdot (1/n) \sum_s e_k^2$.

Cette formule est exactement celle qui avait été proposée à la section 3 à partir du modèle de superpopulation, situation assez sympathique!

4.4.3 Discussion des résultats

La variance se décompose en deux parts, l'une vue comme l'espérance du carré du biais conditionnel, la seconde comme l'espérance de la variance conditionnelle.

Le premier terme ne dépend pas des quotas imposés à l'échantillon, mais seulement de la qualité de l'ajustement d'un modèle additif sur la variable d'intérêt. On diminue cette partie de la variance en choisissant des critères de quotas qui expliquent au mieux ce qu'on veut mesurer.

Le second terme, en revanche, dépend de la variabilité restante $(N_{ij}^2 V_{ij}/n_{ij})$ et du nombre d'observations recueillies dans chaque cellule. La taille de l'échantillon étant fixe on doit donc chercher à rendre les n_{ij} les plus proches possible de l'allocation de Neyman: $n_{ij} \propto N_{ij} V_{ij}^{1/2}$. Ceci peut s'obtenir approximativement en surchargeant les quotas n_{i+} et n_{+j} qui correspondent à de grandes valeurs de V_{ij}. On peut ainsi, dans certains cas, améliorer sensiblement la précision des enquêtes par quotas.

4.5 Combinaison de la méthode des quotas avec les échantillonnages stratifiés où à plusieurs degrés

4.5.1 Cas d'un sondage stratifié avec quotas dans chaque strate

Si les effectifs des critères servant à fabriquer les quotas sont connus dans chaque strate, la méthode qui vient d'être décrite permet de construire un estimateur sans biais sous l'hypothèse que l'échantillonnage fonctionne comme un SAS contraint dans **chaque strate**. Si l'allocation des quotas est proportionnelle aux effectifs de chaque strate, l'estimateur est l'estimateur naturel du sondage stratifié. Si on applique des quotas "nationaux" à chaque strate, une correction doit être faite par repondération.

En revanche, si les effectifs des variables de quotas sont inconnus au niveau des strates, on ne dispose d'aucun moyen de corriger les estimateurs des "effets de structures" relatifs à la stratification. Comme, de plus, la stratification a pour but de construire des sous populations dissemblables, ces corrections seraient généralement fortes. La méthode des quotas n'est alors pas à recommander (sauf, *cf* section 3, si la validité d'un modèle additif s'impose d'elle-même).

4.5.2 Cas d'un sondage à deux degrés

Supposons un sondage à deux degrés (éventuellement à l'intérieur d'une strate où les effectifs des variables de quotas sont connus). Si les effectifs des variables de quotas sont connus au niveau de chaque unité primaire il n'y a pas de problème. La théorie en 4.4 permet de former un estimateur du total de Y dans chaque unité primaire, ainsi que de calculer sa variance et un estimateur de celle-ci. Ces quantités peuvent donc être utilisées pour former un estimateur de Y ainsi qu'un estimateur de précision (*cf* Rao 1975).

Si les effectifs des critères de quotas sont inconnus au niveau des UP mais connus seulement au niveau de la strate, on a de nouveau un problème de correction impossible. Toutefois, le mal doit généralement être limité si les UP sont relativement semblables entre elles: la structure de chaque UP est proche de celle de la strate toute entière et les corrections à faire pour chaque UP sont voisines de celles qu'on doit mettre en oeuvre au niveau de la strate.

4.5.3 En conclusion

Dans le cas d'un sondage complexe stratifié à plusieurs degrés, la méthode des quotas peut être utilisée comme ultime méthode d'échantillonnage si la stratification a été réalisée de façon efficace en regroupant des unités primaires assez semblables entre elles et si on applique dans chaque UP des quotas dérivés de données relatives à sa strate.

Dans la mesure ou l'hypothèse d'un échantillonnage aléatoire simple contraint dans chaque UP peut sembler assez satisfaisante, la méthode des quotas reçoit une justification indépendante de tout modèle de superpopulation.

5. CONCLUSIONS ET PROBLÈMES OUVERTS

5.1 Comment prendre en compte la non-réponse?

Comme nous l'avons déjà signalé, cette question est la limitation la plus importante de notre théorie. Quand on échantillonne par la méthode des quotas, on n'a, en principe, aucune information sur la population qui refuse de répondre à l'enquête et on se trouve démuni de l'information individuelle au sujet des non-répondants. La situation n'est, cependant, peut être pas si désespérée qu'on pourrait le croire. Illustrons cette intuition par un exemple très simplifié.

On a réalisé une enquête par quotas simples chargeant l'échantillon de n_i individus de la catégorie i d'effectif N_i. Un modèle (admis) de non-réponse postule une probabilité r_c de réponse si un individu appartient à une catégorie c d'effectif N_c. L'effectif (inconnu) du croisement entre la catégorie i de quota et la classe c du modèle de non-réponse est noté N_i^c. L'effectif susceptible de répondre dans la catégorie i vaut donc $N_{ri} = \sum_c N_i^c r_c$. En fixant un quota n_i dans cette catégorie, dans le cadre du modèle (4.1), on obtient une probabilité d'inclusion dans l'échantillon égale à $w_i^{-1} = n_i/N_{ri}$. Dans l'échantillon, on recueille n_i^c individus appartenant au croisement (i,c) des deux catégorisations. Cette quantité est aléatoire et son espérance vaut $N_i^c r_c w_i^{-1}$. Si on cherche à estimer les N_i^c, on résoudra les équations estimantes déduites des relations:

$$N_i^c = n_i^c w_i r_c^{-1},$$

$$\sum_c N_i^c = N_i,$$

$$\sum_i N_i^c = N^c.$$

Une technique de raking ratio permet donc de calculer des estimations $\hat{r}_c$ et $\hat{w}_i$ des r_c et w_i. On en déduit des estimateurs $\hat{N}_i^c = n_i^c \hat{w}_i \hat{r}_c^{-1}$ des effectifs du croisement (i,c). On en déduit aussi un estimateur du total de Y:

$$\hat{Y}_{NR} = \sum_{ic} N_i^c \bar{y}_i^c = \sum_{ic} r_c^{-1} w_i n_i^c \bar{y}_i^c,$$

où $\bar{y}_i^c$ est la moyenne des Y_k de l'échantillon classés dans la catégorie (i,c). Ainsi, les techniques d'estimation par calage devraient permettre un traitement honorable de la non-réponse y compris dans des enquêtes par quotas.

5.2 Quelques points de comparaison avec des sondages probabilistes

La méthode des quotas, quelque soit la façon dont on essaie de l'envisager, réclame la formulation d'un modèle hypothétique qu'on plaque sur les données. À l'inverse, un sondage probabiliste ne dépend d'aucun modèle, en principe. En pratique, l'échantillonnage d'un sondage probabiliste **est** un modèle auquel la réalité de la collecte des données essaie de se conformer. On sait bien, en effet, que dans tout sondage probabiliste, quelques accommodements de détail doivent être pris avec le modèle (exclusion d'office de certaines unités, remplacement de certaines après tirage mais avant collecte, *etc.*). On peut, cependant, affirmer sans risque que les biais statistiques sont toujours beaucoup plus faibles dans les tirages probabilistes qu'avec la méthode des quotas. En revanche, les quotas permettent d'utiliser au stade de l'échantillonnage une information auxiliaire qui n'est pas mobilisable dans un tirage probabiliste. Il en résulte que la variance d'un échantillonnage par quotas, est du genre de celle d'une estimation par régression et qu'elle est donc plus faible en règle générale que celle qui résulte d'un sondage probabiliste associé à son estimation de valeurs dilatées standard. Biais dû au modèle associé à une faible variance, contre absence de biais, tel est le bilan. On peut tirer de cette approche deux types de conclusions:

5.2.1 La précision dépend avant tout de la taille des échantillons. Dans le cas de faibles échantillons, le sondage probabiliste va donner de piètres résultats en moyenne et le biais d'un sondage par quotas sera plus tolérable que l'imprécision du sondage probabiliste. Pour de gros échantillons au contraire, la méthode des quotas aura un biais évident incompatible avec l'intervalle de confiance sans biais du sondage probabiliste.

Où fixer la limite entre les deux méthodes? La théorie peut difficilement être affirmative. En revanche, la pratique des instituts français propose une solution à cette question: la plupart des enquêtes nationales par quotas sont réalisées sur des échantillons de 1,000 à 2,000 individus. En revanche, aucune enquête probabiliste nationale ne mobilisera moins de 5,000 unités. Il paraît légitime de dire qu'un effectif de 2,500 à 3,000 enquêtes fixe une limite pratique entre les deux familles de méthodes.

5.2.2 Statistique officielle ou marketing

Tout modèle spéculatif constitue, dans une enquête, une prise de risque méthodologique. Ce risque peut être parfaitement légitime si les utilisateurs en sont conscients, s'ils ont ratifié la spéculation qui a conduit à la spécification d'un modèle. C'est typiquement ce qui se produit, au moins de façon implicite, dans les enquêtes de marketing: un organisme, société, administration ou association, passe commande d'une enquête par sondage avec une société d'études. Un contrat marque l'accord entre les deux parties sur la réalisation de l'enquête, son prix, les délais de livraison des résultats et **la méthodologie employée**. Dans cette méthodologie il y a les modèles utilisés pour formaliser l'échantillonnage ou le comportement de la population. La méthode des quotas peut donc être, de ce point de vue, tout à fait légitime.

La statistique officielle, à l'opposé, est chargée d'élaborer des données utilisables par l'ensemble du corps social, susceptibles, en particulier, de servir d'éléments pour l'arbitrage de conflits entre divers groupes, divers partis, voire diverses classes sociales. Le recours à des modèles statistiques, économétriques en particulier, décrivant le comportement des agents économiques, peut se révéler assez dangereux, partial, influencé par une théorie économique contestable, ou contestée. La statistique officielle ne doit tolérer aucun biais incontrôlable dans sa production. Elle se doit de réaliser des enquêtes par sondage par des méthodes probabilistes.

Il n'y a pas réellement opposition entre les enquêtes par quotas et les techniques ayant recours à un aléatoire contrôlé, mais, bien au contraire complémentarité. À preuve, les statistiques qui servent à construire les quotas sont elles-mêmes très souvent tirées de grosses enquêtes réalisées par les Instituts Nationaux de Statistique. Les techniciens des enquêtes par quotas admettraient mal que ces données soient élaborées autrement que par des méthodes probabilistes confirmées et bien théorisées.

REMERCIEMENTS

Je remercie bien sincèrement l'arbitre et le rédacteur associé du travail positif qu'ils ont réalisé et qui a contribué à l'amélioration de cette article.

ANNEXE

Démonstration des résultats en 4.4

1. Notations et résultats

Pour traiter la question de façon générale nous aurons besoin de certaines notations commodes. On dispose de Q variables qualitatives dont les modalités sont indicées de 1 à I_q pour $q = 1$ à Q. On note c une "cellule", c'est-à-dire une suite de Q indices, le $q^{\text{ième}}$ pouvant valoir de 1 à I_q, et q_c la valeur du $q^{\text{ième}}$ indice ($q^{\text{ième}}$ projection de c); dans une population finie U d'effectif N, U_c est la population des individus classés dans la cellule c d'effectif N_c. La quantité $N_i^{+q} = \sum_{q_c} = iN_c$ est la marge de la table de contingence Q-dimensionnelle dont les cellules sont les c, pour la $i^{\text{ième}}$ modalité de la $q^{\text{ième}}$ variable. On pose

$$\bar{Y}_c = \frac{1}{N_c} \sum_{k \in U_c} Y_k.$$

On a le:

Résultat 1: La variable $Y_k (k \in U)$ peut être paramétré par les nombres $A^q_{q_c}$, E_c et R_k par:

$$\bar{Y}_k = \bar{Y}_c + R_k \quad \text{si} \quad k \in U_c. \quad \text{On a} \quad \sum_{U_c} R_k = 0 \quad \text{pour tout} \quad c.$$

$$\bar{Y}_c = \sum_{q=1}^{Q} A^q_{q_c} + E_c \quad \text{avec} \quad A^q_{I_q} = 0 \quad \text{pour} \quad q = 2 \text{ à } Q \quad \text{et}$$

$$\sum_{q_c = i} N_c E_c = 0 \quad \text{pour} \quad q = i \text{ à } Q \quad \text{et} \quad i = 1 \text{ à } I_q.$$

Ces nombres proviennent de la minimisation de:

$$\sum_{U} \left(Y_k - \sum_{q=1}^{Q} A^q_{q_c(k)} \right)^2 = \sum_{c} N_c \left(\bar{Y}_c - \sum_{q=1}^{Q} A^q_{q_c} \right)^2.$$

Soit maintenant un échantillon s. On note avec des n les quantités analogues dans l'échantillon à ce qu'on a déjà indiqué dans la population avec des N.

On suppose s tiré par sondage aléatoire simple (avec ou sans remise) selon un schéma à probabilités égales contraint par des marges n_i^{+q} ($q = 1$ à Q, $i = 1$ à I_q), les quotas.

Le but de cette annexe est de montrer le:

Résultat 2: La variance de $\sum_c \hat{N}_c E_c$ est approximativement égale à $1/n \sum_c N_c E_c^2$ quand n et N/n deviennent arbitrairement grands.

La suite donnera une formulation précise à ce résultat.

2. Schéma d'échantillonnage et réduction asymptotique

Considérons les deux modèles d'échantillonnages SR et AR suivant:

SR: Échantillonnage bernouillien. Chaque unité parmi les N appartient à s avec probabilité f, les N tirages étant indépendants.

AR: Chaque unité est tirée un nombre ν_k de fois; ν_k suit une loi de Poisson de paramètre f. Les ν_k sont des variables indépendantes.

Un sondage aléatoire simple sans remise (SASSR) de taille fixe n est un échantillonnage SR conditionnellement au fait que la taille totale de l'échantillonnage est n.

Un sondage aléatoire simple avec remise (SASAR) de taille fixe n est un échantillonnage AR conditionnellement au fait qu'on a fait n observations, c'est-à-dire que $\sum_k \nu_k = n$.

Dans le cas SR, la loi du vecteur n_c est donnée par:

$$\Pr(\{n_c\}) = \prod_c \binom{N_c}{n_c} f^{n_c} (1 - f)^{N_c - n_c}.$$

Dans le cas AR, on a:

$$\Pr(\{n_c\}) = \prod_c \frac{(N_c f)^{n_c}}{n_c !} \exp(-fN_c).$$

Dans les deux cas les variables n_c sont indépendantes.

Dans le cas SR contraint par $\sum n_c = n$, la loi des n_c est une hypergéométrique:

$$\Pr(\{n_c\}) = \prod_c \binom{N_c}{n_c} \binom{N}{n}^{-1}.$$

Dans le cas AR contraint, c'est une multinomiale:

$$\Pr(\{n_c\}) = \prod_c p_c^{n_c}/n_c !.$$

Le modèle de plan de sondage retenu pour la méthode des quotas à la section 4 correspond à des contraintes sur ces deux schémas, ou, ce qui est équivalent,à des contraintes sur les schémas SR et AR.

Si on suppose que N tend vers l'infini, que f tend vers zéro, et que $n^* = fN$ tend vers l'infini, alors, dans les deux schémas, la loi des $u_c = n^{*-1/2} (n_c - fN_c) = n^{*1/2} (p_c^* - p_c)$, avec $p_c^* = n_c/n^*$, tend vers une loi normale multidimensionnelle avec les u_c indépendantes, d'espérances nulles et de variances égales aux p_c.

3. Cas des sondages "proportionnels"

On a alors $\hat{N}_c = N/n\, n_c$ de sorte que la quantité dont on cherche la variance est:

$$\frac{N}{n^{*1/2}} \sum_c u_c E_c,$$

où le vecteur des u_c suit une loi normale centrée de matrice des covariances diagonale $\Delta = \text{diag}(p_c)$, contrainte par les relations qui expriment les quotas:

$$\sum_{q_c = i} u_c = 0 \text{ pour } q = 1 \text{ à } Q, \quad i = 1 \text{ à } I_q \text{ si } q = 1, \quad i = 1 \text{ à } I_q - 1 \text{ si } q = 2 \text{ à } Q.$$

Ces relations s'écrivent, en notant U le vecteur des u_c:

$$AU = 0,$$

avec A matrice à $l = \sum_q I_q - (Q - 1)$ lignes et $k = \Pi_q I_q$ colonnes, composée de 1 et de 0 pour traduire les contraintes. Ceci exprime aussi le fait que U varie dans le noyau L de l'opérateur défini par la matrice A. La loi (asymptotique) de U est donc celle d'un vecteur gaussien W centré de matrice des covariances égales à Δ, conditionnellement à $AW = 0$. On est ramené à évaluer la variance du produit scalaire $U'\underline{E}$ où $\underline{E}$ est le vecteur des E_c.

Remarquons les deux points suivants:

- Les contraintes sur les E_c données au résultat 1 se traduisent matriciellement par $A\Delta\underline{E} = 0$. Autrement dit $\Delta\underline{E}$ est un vecteur de $L = \mathrm{Ker}A$, ou encore $\underline{E}$ est un vecteur de $\mathrm{Ker}(A\Delta)$.
- Soit P le projecteur de $\mathfrak{R}^k$ sur L orthogonal dans la métrique Δ^{-1}. P vérifie les relations:
  - $\forall x \in L,\ Px = x;\ Im\, P = L$
  - $Py = 0 \Leftrightarrow \forall x \in L,\ x'\Delta^{-1} y = 0;\ \mathrm{Ker}P = \Delta(L^{\perp})$,

où $L^{\perp}$ est le supplémentaire orthogonal de L dans la métrique naturelle.

Les vecteurs gaussiens PW et $(1 - P)W$ varient respectivement dans L et dans $\Delta(L^{\perp})$, leur somme est égale à W. De plus ils sont indépendants; en effet leur matrice de covariance vaut $E(PW)((1 - P)W)' = P\Delta(1 - P')$. Or P' est le projecteur de noyau $L^{\perp}$ et d'image $\Delta(L^{\perp})^{\perp}$. L'image du projecteur $(1 - P')$ est donc $L^{\perp}$, celle de $\Delta(1 - P')$ est $\Delta(L^{\perp})$, c'est-à-dire le noyau de P, c.q.f.d.

Il faut maintenant évaluer la variance de $\sum_c u_c E_c = U'\underline{E}$. Or, d'après ce qui précède, on peut écrire $W = U + V$, avec U et V indépendants. La loi de W conditionnellement à $W \in L$ n'est autre que la loi de W conditionnellement à $V = 0$.

Par ailleurs on a:

$$V'\underline{E} = (\Delta^{-1} V)'(\Delta E).$$

Comme ΔE est dans L et que V varie dans $\Delta(L^{\perp})$, le produit scalaire ci-dessus est nul. On en déduit que:

$$\mathrm{Var}(U'\underline{E}) = \mathrm{Var}(W'\underline{E}) = \underline{E}'\Delta\underline{E} = \sum_c p_c E_c^2.$$

La variance asymptotique de $N/n^* \sum_c n_c E_c$ vaut donc

$$\frac{N^2}{n} \sum_c p_c E_c^2 = \frac{N}{n} \sum_c N_c E_c^2.$$

4. Cas des sondages par quotas "non-proportionnels"

Complétons la réduction asymptotique précédente. Maintenant le vecteur $\hat{p}^{\circ}$ des n_c/n^* est contraint par

$$A\hat{p}^{\circ} = Ap + n^{*-1/2} AV_0,$$

où Ap est le vecteur (à l dimensions) des "quotas proportionnels" et V_0 l'unique vecteur (à k dimensions) de $\Delta(L^{\perp})$ tel que $A(p + n^{*-1/2} V_0)$ soit le vecteur des quotas imposés. De ce fait $U. = n^{*1/2}(\hat{p}^{\circ} - p)$ peut-être, comme au paragraphe précédent, analysé comme un vecteur gaussien $W = U + V$ conditionnel à $V = V_0$. Par suite $EU_0 = V_0$ et la matrice des covariances de U_0 est la même que celle de U.

On passe, par ailleurs, de $\hat{p}^\circ$ à $\hat{p}$ par une estimation du maximum de vraisemblance. Dans les conditions asymptotiques gaussiennes, celle-ci s'identifie à la minimisation de la forme quadratique $(\hat{p}^\circ - \hat{p})' \Delta^{-1} (\hat{p}^\circ - \hat{p})$ sous les contraintes $A\hat{p} = Ap$. Comme $\hat{p}^\circ$ varie dans le sous-espace affine $L + V_0$ parallèle à L, que la minimisation en question revient à projeter $\hat{p}^\circ$ sur L orthogonalement pour la métrique Δ^{-1} c'est-à-dire le long de $\Delta(L^\perp)$, il suit qu'on a $\hat{p} = \hat{p}^\circ - n^{*-1/2} V_0$ dans les conditions asymptotiques. Le vecteur aléatoire $\hat{p}$ est donc translaté de $\hat{p}^\circ$, sans biais et de même matrice de covariance que $\hat{p}^\circ$ et donc que $n^{*-1/2} U$.

Finalement, on a:

$$E\left(\sum_c \hat{p}_c E_c \right)^2 = E(\hat{p}' \underline{E})^2 = \frac{1}{n^*} \sum_c p_c E_c^2$$

comme dans le cas précédent.

BIBLIOGRAPHIE

CASSEL, C.M., SÄRNDAL, C.E., et WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley & Sons.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley & Sons.

DESABIE, J. (1965). *Théorie et pratique des sondages*. Paris: Dunod.

DEVILLE, J.C., et SÄRNDAL, C.E. (1990). Calibration estimators and generalized raking techniques. Manuscrit soumis pour publication.

GOURIÉROUX, C. (1981). *Théorie des sondages*. Paris: Economica.

MADOW, W.G., OLKIN, I., et RUBIN, D.B., (éds.) (1983). *Incomplete Data in Sample Surveys*. New York: Academic Press.

RAO, J.N.K. (1976). Unbiased variance estimation for multistage designs. *Sankhyā*, Série C, 37, 133-139.

SMITH, T.M.F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society*, A, 146, 394-403.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.