QÜESTIIÓ, vol. 23, 3, p. 491-505, 1999

UTILISATION D'INFORMATIONS AUXILIAIRES DANS LES ENQUÊTES PAR SONDAGE

Y. TILLÉ
Laboratoire de Statistique d'Enquête*

La notion de représentativité est apparue dès la naissance de la théorie des sondages à la fin du dix-neuvième siècle. Pourtant, ce concept qui est appliqué autant aux plans par quotas qu'aux plans probabilistes est largement galvaudé. Après avoir rappelé quelques éléments de l'histoire de la théorie des sondages, nous rappelons quelques techniques de base de plans aléatoires et à choix raisonnés. Nous montrons ensuite que le concept de plan équilibré permet de lever les ambiguïtés fondamentales de la notion de représentativité.

Use of Auxiliary Information in Survey Sampling

Mots clés: Sondage, plans équilibrés, représentativité

AMS Classification (MSC 2000): 62D05

* Laboratoire de Statistique d'Enquête. CREST - ENSAI. École Nationale de la Statistique and de l'Analyse de l'Information rue Blaise Pascal, Campus de Ker Lann 35170 Bruz, France, email: tille@ensai.fr

– Reçu en juillet de 1999.
– Accepté en octobre de 1999.

1. SONDAGE ET REPRÉSENTATIVITÉ

La théorie des sondages est un ensemble d'outils statistiques permettant l'étude d'une population au moyen de l'examen d'une partie de celle-ci. Le sondage s'oppose au recensement qui est l'étude exhaustive de la population. La théorie des sondages vise à justifier ce processus d'extrapolation (illustré en figure 1.). Nous verrons cependant que l'extrapolation de la partie au tout est une démarche qui a été rejetée par les statisticiens jusqu'au début du vingtième siècle. Les arguments visant à valider cette extrapolation ne sont pas encore toujours clairs. La justification la plus couramment utilisée est la «représentativité» de l'échantillon.

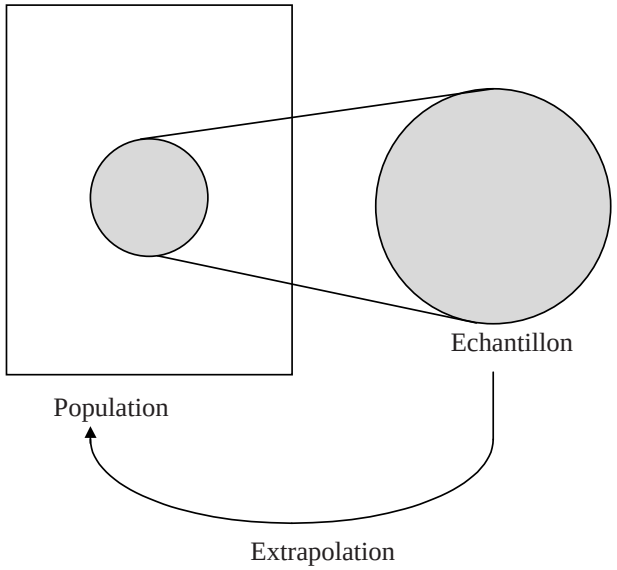

Figure 1. Extrapolation de l'échantillon à la population

On dit souvent qu'un échantillon est représentatif de la population s'il en constitue un modèle réduit. Un bon échantillon devrait «ressembler» autant que possible à la population à étudier de sorte que certaines catégories apparaissent en même proportion dans l'échantillon et dans la population. Pourtant, cette théorie couramment véhiculée par les médias et même par certains ouvrages de méthodologie est incorrecte: un échantillon pour être valide ne doit pas être représentatif (ou sens où nous venons de le définir).

Il est en effet souvent souhaitable d'effectuer des tirages à probabilités inégales ou de surreprésenter certaines parties de la population. Pour estimer de manière précise

une fonction d'intérêt, il faut aller chercher l'information de manière judicieuse plutôt que d'accorder la même importance à chaque unité. Prenons un exemple, si on veut estimer la production de fer d'un pays et qu'on sait que cette production est assurée, d'une part, par deux entreprises sidérurgiques gigantesques qui occupent des milliers de travailleurs et, d'autre part, par plusieurs centaines des petites entreprises artisanales de moins de cinquante travailleurs, va-t-on sélectionner chaque unité avec une même probabilité? Non, bien sûr. On va commencer par s'enquérir de la production des deux grandes entreprises (qui seront donc sélectionnées d'office dans l'échantillon). Ensuite, on sélectionnera les petites de manière aléatoire selon un plan de sondage à déterminer. Cet exemple simple va à l'encontre de l'idée de représentativité et montre bien qu'il faut aller chercher l'information là où elle se trouve, et que le concept de représentativité n'est pas pertinent.

2. ORIGINE DE LA THÉORIE DES SONDAGES

Le développement de la méthodologie statistique (Étymologiquement, science de l'État) est indissociable de l'émergence des États modernes au dix-neuvième siècle. Une des personnalités les plus marquantes de la statistique officielle du dix-neuvième siècle est le belge Adolphe Quetelet (1796-1874) qui fut d'abord attiré par l'idée d'utiliser des données partielles, mais s'est rapidement rallié à l'idée selon laquelle l'utilisation de données partielles est incompatible avec la déontologie statistique. Depuis lors, Quetelet a toujours considéré l'exactitude comme un principe de base de la science statistique. Celui-ci eut une grande influence dans le développement de la statistique officielle. Il organisa le premier Congrès International de la Statistique à Bruxelles en 1853. Il a vraisemblablement contribué à faire admettre par toute la communauté scientifique que l'utilisation de sondages n'est pas une méthode statistique valide.

Au dix-neuvième siècle, l'établissement d'un appareil statistique fut une nécessité dans l'édification des grands États modernes. À cette époque, l'objectif du statisticien était surtout de réaliser des énumérations. La préoccupation majeure était d'inventorier les ressources des nations. Dans ce contexte, le recours à l'échantillonnage fut unanimement rejeté comme une procédure inexacte et donc foncièrement anti-scientifique. Tout au long de ce siècle, les discussions des statisticiens portent essentiellement sur la méthode à appliquer pour obtenir des données fiables et sur la présentation, l'interprétation et éventuellement la modélisation (par un ajustement) de ces données.

En 1895, le Norvégien A.N. Kiaer, directeur du Bureau Central de la Statistique de Norvège, présente au Congrès de l'Institut International de Statistique (IIS) à Berne un travail intitulé «Observations et expériences concernant des dénombrements représentatifs» relatif à un sondage réalisé en Norvège. Kiaer sélectionne d'abord un échantillon de villes et de communes. Ensuite, dans chacune de ces communes, il ne sélectionne

qu'une partie des individus selon la première lettre de leurs noms de famille. Il applique donc un plan à deux degrés mais le choix des unités n'est pas aléatoire. Kiaer défend l'intérêt de l'utilisation de données partielles pour peu qu'elles soient produites au moyen d'une «méthode représentative». Selon cette méthode, l'échantillon doit être une représentation de la population à taille réduite. La notion de représentativité de Kiaer est donc liée à la méthode des quotas. L'intervention de Kiaer est suivie d'un débat houleux, les actes du congrès de l'IIS rendent compte d'une longue polémique. Examinons de plus près l'argumentation de deux des opposants à la méthode de Kiaer (voir Procès-verbal de l'Assemblée Générale de l'IIS, 1896).

> M.V. Mayr [...] C'est surtout dangereux de se déclarer pour ce système des investigations représentatives au sein d'une assemblée de statisticiens. On comprend que pour des buts législatifs ou administratifs un tel dénombrement restreint peut être utile - mais alors il ne faut pas oublier qu'il ne peut jamais remplacer l'observation statistique complète. Il est d'autant plus nécessaire d'appuyer là-dessus, qu'il y a parmi nous dans ces jours un courant au sein des mathématiciens qui, dans beaucoup de directions, voudraient plutôt calculer qu'observer. Mais il faut rester ferme et dire: pas de calcul là où l'observation peut être faite.
>
> M. Milliet. Je crois qu'il n'est pas juste de donner par un voeu du congrès à la méthode représentative (qui enfin ne peut être qu'un expédient) une importance que la statistique sérieuse ne reconnaîtra jamais. Sans doute, la statistique faite avec cette méthode ou, comme je pourrais l'appeler, la statistique, Pars pro toto, nous a donné ça et là des renseignements intéressants ; mais son principe est tellement en contradiction avec les exigences que doit avoir la méthode statistique, que, comme statisticiens, nous ne devons pas accorder aux choses imparfaites le même droit de bourgeoisie, pour ainsi dire, que nous accordons à l'idéal que scientifiquement nous nous proposons d'atteindre.

Le contenu de ces réactions peut se résumer ainsi: comme la statistique est par définition exhaustive, renoncer au dénombrement complet c'est nier la mission même de la science statistique. La discussion ne porte donc pas sur la méthode proposée par Kiaer mais sur la définition de la science statistique. Kiaer ne désarme pourtant pas et continue à défendre la méthode représentative en 1897 au congrès de l'IIS à Saint-Petersbourg, en 1901 à Budapest et en 1903 à Berlin. Après cette date, la question ne sera plus mentionnée au congrès de l'IIS. Kiaer obtient cependant l'appui d'Arthur Bowley (1869-1957) qui jouera ensuite un rôle déterminant dans le développement des sondages. Bowley (1906) présente une vérification empirique pour l'application du théorème central limite à l'échantillonnage. Celui-ci fut le véritable promoteur des techniques de sondage aléatoire, il développe les plans stratifiés avec allocations proportionnelles et utilise la formule de décomposition de la variance.

En 1924, une commission (composée de Arthur Bowley, Corrado Gini, Adolphe Jensen, Lucien March, Verrijn Stuart, et Frantz Zizek) est créée afin d'évaluer la pertinence de l'utilisation de la méthode représentative. Les résultats de cette commission intitulés «Reports on the representative method in statistics» sont présentés au congrès

de l'IIS de 1925 à Rome. La commission accepte le principe du sondage pour autant que la méthodologie soit respectée. Plus de trente ans après la communication de Kiaer, l'idée de l'échantillonnage est donc officiellement acceptée. La commission jettera les bases des recherches futures: deux méthodes sont clairement distinguées «la sélection aléatoire» et la «sélection raisonnée». Ces deux méthodes correspondent à deux démarches scientifiques fondamentalement différentes. D'une part, la validation des méthodes aléatoires est basée sur le calcul des probabilités qui permet de construire des intervalles de confiance pour certains paramètres. D'autre part, la validation des méthodes par sélection raisonnée ne peut être donnée que par l'expérimentation en comparant les estimations obtenues à des résultats de recensement. Les méthodes aléatoires sont donc validées par un argument strictement mathématique tandis que les méthodes par choix raisonnés sont validées par une démarche expérimentale.

Depuis la publication de ce rapport, l'opposition entre ces deux types de plans de sondage est restée pleinement d'actualité. Dans un article récent, Brewer (1999) oppose encore les plans probabilistes stratifiés aux plans obtenus par une méthode de choix raisonnés. Les méthodes probabilistes sont plus largement utilisées en statistique officielle, tandis que les méthodes à choix raisonnés (et plus particulièrement la méthode des quotas), sont largement utilisées (en Europe) dans les instituts privés de statistique. Une des ambiguïtés majeures du terme représentativité est qu'il est appliqué indifféremment aux plans probabilistes et aux plans à choix raisonnés.

3. INFORMATION AUXILIAIRE

La notion d'information auxiliaire regroupe toute information extérieure à l'enquête proprement dite permettant d'augmenter la précision des résultats d'un sondage. De manière générale, on appelle information auxiliaire toute information connue sur la population. Cette information peut être la connaissance des valeurs d'une ou de plusieurs variables sur toutes les unités de la population ou simplement d'une fonction de ces valeurs. Pour la plupart des enquêtes, une information auxiliaire est disponible. Elle peut être donnée par un recensement ou tout simplement par la base de sondage. On peut citer comme exemple d'information auxiliaire: le total d'un caractère sur la population, des sous-totaux selon des sous-populations, des moyennes, des proportions, des variances, les valeurs d'un caractère sur toutes les unités de la base de sondage. La notion d'information auxiliaire englobe donc toute donnée issue de recensement.

Les variables dont au moins une fonction des valeurs est connue sont alors appelées variables auxiliaires. L'objectif principal consiste donc à mettre à profit toutes ces informations pour obtenir des résultats précis. L'information auxiliaire peut être utilisée à deux moments: à l'étape de la conception du plan de sondage et à l'étape de l'estimation des paramètres (voir figure 2.).

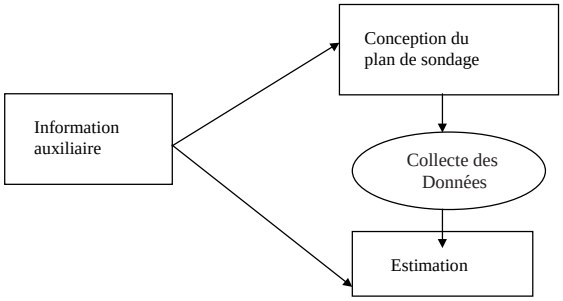

Figure 2. Les deux étapes de l'utilisation de l'information auxiliaire

Quand l'information auxiliaire est mise à profit pour concevoir le plan de sondage, on cherche un plan qui fournit des estimateurs précis pour un prix donné ou qui est peu coûteux pour des critères de précision donnés. Pour ces raisons, on utilisera des plans à probabilités inégales, par grappes ou à plusieurs degrés. Quand l'information est utilisée à l'étape de l'estimation, elle sert à «recaler» les résultats du sondage sur l'information auxiliaire du recensement. Les estimateurs sont alors basés sur deux sources d'informations: l'information auxiliaire connue sur toute la population, et l'information concernant les variables d'intérêt connue uniquement sur les unités sélectionnées dans l'échantillon (voir figure 3.). La méthode générale de calage (en anglais: *calibration*) de Deville et Särndal (1992) permet d'utiliser des informations auxiliaires en modifiant les poids affectés aux unités de manière à ce que les estimateurs de totaux calculés dans l'échantillon soient égaux aux totaux de la population pour toutes les variables auxiliaires connues. Nous limiterons cependant, par la suite, à l'utilisation de l'information auxiliaire à l'étape de la planification.

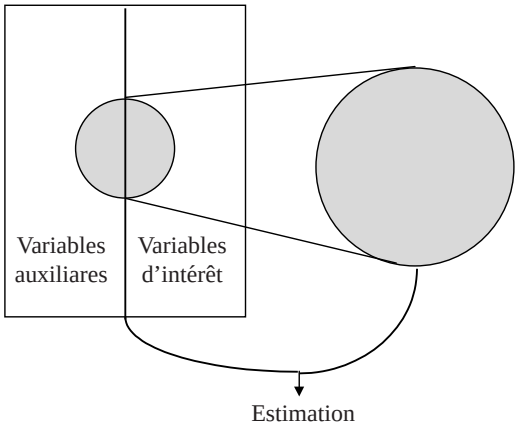

Figure 3. Estimation avec information auxiliaire

4. PLAN DE SONDAGE ET OBJECTIF D'ESTIMATION

Considérons une population de taille N, et supposons que les unités d'observation peuvent être désignées par un numéro d'ordre $k \in \{1, \ldots, k, \ldots, N\} = U$. On s'intéresse à une variable d'intérêt y dont la valeur prise sur l'unité k est notée y_k, pour tout $k \in U$. L'objectif est d'estimer le total de ces valeurs

$$Y = \sum_{k \in U} y_k,$$

au moyen d'un échantillon de cette population. La taille de la population est un total particulier qui s'obtient quand $y_k = 1, k \in U$. Dans ce cas,

$$N = \sum_{k \in U} 1.$$

La moyenne de la variable y dans la population peut alors s'écrire comme un rapport de deux totaux

$$\overline{Y} = \frac{Y}{N},$$

qui seront estimés séparément.

Un échantillon est un sous-ensemble non-vide de U et un plan de sondage est une loi de probabilité $p(.)$ sur tous les échantillons $s \subset U, \#s = n$, telle que

$$p(s) \geqslant 0, \text{ pour tout } s \subset U, \text{ tel que } \sum_{s \subset U} p(s) = 1.$$

Si S est l'échantillon aléatoire tel que $Pr(S = s) = p(s)$, on note I_k la variable aléatoire indicatrice qui prend la valeur 1 si $k \in S$ et 0 sinon, pour tout $k \in U$. De plus, on note $\pi_k = E(I_k) = Pr(k \in S)$, la probabilité d'inclusion d'ordre un, c'est-à-dire la probabilité que l'unité k soit sélectionnée dans l'échantillon. Enfin, on note $\pi_{l\ell} = E(I_k I_\ell) = P(k \in S \text{ et } \ell \in S), k \neq \ell$, la probabilité d'inclusion d'ordre deux, c'est-à-dire la probabilité que deux unités distinctes k et ℓ soient sélectionnées conjointement dans l'échantillon.

Le total Y peut s'estimer sans biais par l'estimateur d'Horvitz-Thompson.

$$\hat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

La variance de l'estimateur de Horvitz-Thompson est donnée par

$$\text{Var}\left[\hat{Y}_\pi\right] = \sum_{k \in S} \sum_{k \in S} \frac{y_k\, y_\ell}{\pi_k\, \pi_\ell} \Delta_{k\ell},$$

où

$$\Delta_{k\ell} = \begin{cases} \pi_k(1-\pi_k) & \text{si} \quad k = \ell \\ \pi_{k\ell} - \pi_k\,\pi_\ell & \text{si} \quad k \neq \ell \end{cases}$$

La précision de l'estimateur de Horvitz-Thompson ne dépend donc du plan qu'au travers des probabilités d'inclusion à l'ordre un et deux.

5. INFORMATION AUXILIAIRE DANS LES PLANS PROBABILISTES

L'introduction d'information auxiliaire peut avoir deux objectifs: l'amélioration de la précision pour un coût donné, ou l'amélioration de l'organisation de l'enquête. L'impact de l'introduction de l'information auxiliaire dans le plan sur la précision des estimateurs se fera, soit sur les probabilités d'inclusion d'ordre un, soit sur les probabilités d'inclusion d'ordre deux, soit sur les deux en même temps.

5.1. Plans à probabilités inégales

Les plans à probabilités inégales consistent à introduire un «effet» sur les probabilités d'inclusion d'ordre un. Ces plans s'avèrent particulièrement intéressants quand les variables sont liées par un effet de taille. Par exemple, pour des entreprises, des variables comme le chiffre d'affaires, le nombre de travailleurs, sont liées par un tel effet. Si une variable auxiliaire x permet de mesurer approximativement cet effet, il est particulièrement intéressant de sélectionner les unités d'observation avec des probabilités d'inclusion proportionnelles à cette variable auxiliaire. Le gain de précision sera alors très important. L'idée même du tirage à probabilités inégales va à l'encontre de la notion de représentativité telle que nous l'avons définie précédemment.

5.2. Plans stratifiés

La technique classique de stratification permet presque toujours d'améliorer la précision d'un estimateur. La stratification consiste à partitionner la population en strates, puis à sélectionner un plan aléatoire simple de taille fixe dans chaque strate. Pour pouvoir réaliser un tel tirage, il est nécessaire de disposer d'une information auxiliaire qui permet d'affecter chaque unité à une strate. L'estimateur d'Horvitz-Thompson présente l'intéressante propriété d'être naturellement «calé» sur les tailles des strates. En effet, si on estime les tailles des strates à partir de l'échantillon, on estime ces tailles sans biais avec une variance nulle.

Rappelons également qu'en stratification, les probabilités d'inclusion ne doivent pas nécessairement être égales d'une strate à l'autre. La stratification optimale de Neyman consiste d'ailleurs à surreprésenter les unités dans les strates où la dispersion est plus importante. Il est intéressant de constater que la stratification optimale de Neyman infirme l'idée de la représentativité. Il n'est pas du tout nécessaire que les unités soient «représentées» dans l'échantillon de manière proportionnelle aux effectifs des strates dans la population.

5.3. Plans à plusieurs degrés

Les plans à plusieurs degrés visent plutôt à une économie de moyens. Un premier échantillonnage est appliqué sur des unités primaires (par exemple des communes). On sélectionne ensuite des unités secondaires (par exemple des ménages) dans les unités primaires sélectionnées. Un plan classique consiste à sélectionner les unités primaires à probabilités inégales proportionnelles aux nombres d'unités secondaires (nombre de ménages dans la commune). Ensuite, on sélectionne un nombre fixe d'unités secondaires dans les unités primaires sélectionnées. Un tel plan présente l'intérêt d'être facile à gérer en terme de répartition de travail entre les enquêteurs. En établissant le formulaire, on constate que le premier degré d'échantillonnage contribue beaucoup plus à la variance des estimateurs que la seconde. Il est donc important de «soigner» le tirage des unités primaires.

6. INFORMATION AUXILIAIRE DANS LES PLANS À CHOIX RAISONNÉS

La méthode des quotas est la méthode empirique la plus utilisée. Le principe est le suivant: on divise la population en un certain nombre de sous-populations selon une ou plusieurs variables catégorielles. Ensuite, on demande aux enquêteurs d'interroger un nombre d'individus proportionnel à chacune de ces sous-populations. Les enquêteurs sont libres de choisir les personnes à interroger. Ce sont donc les enquêteurs qui construisent le plan de sondage. Le plan de sondage et les probabilités d'inclusion sont inconnus. Les avantages de cette méthode sont nombreux: il n'est pas nécessaire de disposer de la base de sondage. Les seules informations utiles sont les effectifs de certaines catégories de la population. De plus, le problème des refus de réponse ne se pose pas puisque l'enquêteur peut choisir lui-même les individus à interroger.

La technique des quotas marginaux consiste à demander aux enquêteurs de sélectionner un certain nombre d'unités de manière à vérifier conjointement les effectifs de plusieurs variables catégorielles, classe d'âge, profession, niveau d'études, etc. L'enquêteur reçoit une «feuille» de quotas indiquant l'effectif à atteindre pour chaque modalité de chaque variable. L'enquêteur peut choisir assez librement les premières personnes à interroger,

mais verra ses choix de plus en plus contraints au fur et à mesure que sa feuille de quotas se remplit.

Si on peut considérer les plans stratifiés comme la version probabiliste des plans par quotas sur une variable catégorielle, il n'existait pas de version probabiliste des plans par quotas marginaux. Un des intérêts de la technique des quotas est qu'elle élude le problème de la non-réponse. Cependant comme un remplacement est organisé d'office par les enquêteurs, le biais dû aux non réponses reste présent dans l'enquête. Comme le gestionnaire d'enquête ne sait pas qui a refusé de répondre, il est impossible de réaliser une correction de ce biais de non-réponse.

7. PLANS ÉQUILIBRÉS: LA SYNTHÈSE

Les plans équilibrés présentent à la fois les avantages des plans par quotas et des plans probabilistes. De manière générale, on se réfère à la définition suivante.

Définition 1. *Un plan de sondage $p(s)$ est dit équilibré pour les variables auxiliaires $x_1, \ldots, x_p$, si et seulement si il vérifie les équations d'équilibrage:*

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj}.$$

où x_{kj} est la valeur prise par la variable j sur l'unité k.

Un plan équilibré estime exactement les totaux des variables auxiliaires avec l'estimateur naturel d'Horvitz-Thompson. Il peut être à probabilités inégales et les variables $x_1, \ldots, x_p$, peuvent être catégorielles ou quantitatives. Si les plans par quotas étaient probabilistes, ils seraient donc équilibrés pour les variables de quotas. Examinons quelques cas particuliers:

Exemple 1. *Un plan de taille fixe est équilibré sur la variable auxiliaire $x_k = \pi_k, k \in U$. En effet,*

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} 1 = \sum_{k \in U} \pi_k = n.$$

Exemple 2. *Supposons que le plan soit stratifié et que dans chaque strate $U_h, h = 1, \ldots, H$, de taille N_h on sélectionne un plan simple sans remise de taille fixe n_h, alors le plan est équilibré sur les variables δ_{kh} de valeurs*

$$\delta_{h\ell} = \begin{cases} 1 & si \quad k \in U_h \\ 0 & si \quad k \notin U_h \end{cases}$$

En effet,

$$\sum_{k \in S} \frac{\delta_{hk}}{\pi_k} = \sum_{k \in S} \frac{N_H \, \delta_{hk}}{n_h} = N_H,$$

pour $h = 1, \ldots, H$.

La notion de plan équilibré éclaircit les ambiguïtés soulevées par le concept galvaudé de représentativité. Un plan équilibré peut être à probabilités inégales. De plus, il n'est jamais équilibré en soi mais pour un ensemble quelconque de variables particulières. La définition formelle énoncée ci-dessus correspond donc à l'exigence de rigueur du statisticien. Le concept de plan équilibré est ancien. Il est déjà présent dans Thionet (1953). Il a été longuement discuté dans le cadre de l'inférence basée sur un modèle par Royall et Herson (1973) par exemple. Plus récemment, une procédure de tirage équilibré a été appliquée pour le tirage de l'échantillon-maître par Ardilly (1991). Cependant jusqu'à présent, aucune procédure simple ne permettait de sélectionner un échantillon équilibré pour un ensemble de variables. Récemment, Deville et Tillé (1999) ont proposé une méthode permettant de sélectionner des échantillons équilibrés sur un ensemble de variables auxiliaires. Le tirage à probabilités inégales de taille fixe, la stratification en sont des cas particuliers. La méthode permet également d'utiliser plusieurs critères de stratification dans le même plan. La méthode a été implémentée sous SAS, et servira probablement à la sélection des unités primaires dans de nombreux plans de sondages.

RÉFÉRENCES

Ardilly, P. (1991). «Echantillonnage représentatif optimum à probabilités inégales», 23, 91-113.

Bowley, A.L. (1906). «Address to the economic and statistics section for the British Association of Advancement of Sciences», *Journal of the Royal Statistical Society*, 69, 540-558.

Brewer, K.R.W. (1999). «Design-based or prediction based inference? Stratified random vs stratfied balanced sampling», *International Statistical Review*, 67, 35-47.

Deville J.-C. et Särndal, C.-E. (1992). «Calibration estimators in survey sampling», *Journal of the American Statistical Association*, 87, 376-382.

Deville J.-C. et Tillé, Y. (1999). *Balanced sampling by means of the cube method*. Manuscrit non-publié, ENSAI, Paris.

Horvàth, R.A. (1974). «Les idées de Quetelet sur la formation d'une discipline moderne et sur le rôle de la théorie des probabilités», in *Mémorial Adolphe Quetelet*, N°3, Académie Royale des Sciences de Belgique.

Royall, R. et Herson, J. (1973). «Robust estimation in finite populations I», *Journal of the American Statistical Association*, 68, 880-889.

Stigler, S.M. (1986). *The History of Statistics*, Cambridge-London, Harvard University Press.

Thionet, P. (1953). «La théorie des sondages», *Etudes théoriques*, N°5, Paris, INSEE.

—— «Procès-verbal de l'Assemblée Générale de l'Institut International de Statistique, N°13», *Séance du vendredi matin 30 août, Bulletin de l'Institut International de Statistique*, Berne, 9, livre 1, 1896, pp. LXXXVIII-XCVII.

ENGLISH SUMMARY

USE OF AUXILIARY INFORMATION IN SURVEY SAMPLING

Y. TILLÉ
Laboratoire de Statistique d'Enquête*

The concept of representativeness appears with the creation of the theory of survey sampling at the end of the 19th century. Nevertheless, this concept which is applied at the same time for quota sampling and for random sampling is dramatically overworked. After a brief presentation of the history of survey sampling, we give a short overview of the basic techniques of planning for purposive selection and for random sampling. Furthermore it is shown that the concept of balanced sampling allows to remove the ambiguity of the notion of representativeness.

Keywords: Sampling, balanced sampling, representativeness

AMS Classification (MSC 2000): 62D05

* Laboratoire de Statistique d'Enquête. CREST - ENSAI. École Nationale de la Statistique and de l'Analyse de l'Information rue Blaise Pascal, Campus de Ker Lann 35170 Bruz, France, email: tille@ensai.fr

– Received July 1999.
– Accepted October 1999.

The sampling theory allows to study a population by means of subset of this population called sample. The sampling theory aims at justifying this extrapolation process (see Figure 1). The idea of extrapolation from the sample to the population was rejected till the beginning of the 20th century. The arguments used were not always very clear? The most common justification was the representativeness of the sample.

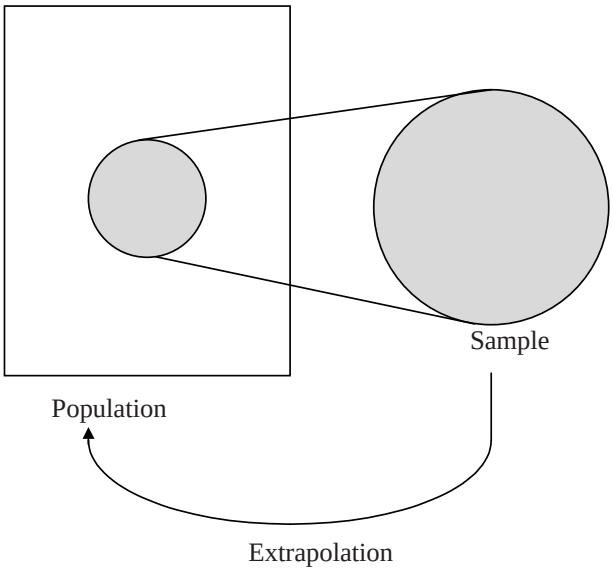

Figure 1. Extrapolation from the sample to the population

Usually a sample is said to be representative when it is a «small-scale» model of the population. A «good» sample should be very similar to the target population. That is some categories appears with the same proportion in the sample and in the population. Kiaer already advocated for the use of representative samples in 1895 at the congress of the International Statistical Institute (Berne). Nevertheless a rapid examination of the modern sampling theory shows that it is often more efficient to select units with unequal probabilities and that the intuitive idea of representativeness is actually false.
Consider a population $U = \{1, \ldots, k, \ldots, N\}$. We are interested to estimate the total of the values y_k, for all $k \in U$. Thus the objective is to estimate

$$Y = \sum_{k \in U} y_k.$$

Suppose also that a random sample S is selected. Let us define by $\pi_k = E(I_k) = Pr(k \in S)$ the first order inclusion probabilities. An unbiased estimator of Y is given

by the Horvitz-Thompson estimator

$$\hat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Moreover suppose that the values of p auxiliary variables $x_1, \ldots, x_p$, are known on all the units of the population.

Definition 1. *A sampling design is said to be balanced on the auxiliary variables $x_1, \ldots, x_p$, if and only if it satisfies*

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

for $j = 1, \ldots, p$, where x_{kj} is the value taken of variable j for unit k.

The concept of balanced sampling generalises most of the sampling techniques that allows to use auxiliary information at the estimation stage. Moreover it allows us to use unequal inclusion probabilities, and it removes the ambiguity of the concept of representativeness. Recently Deville and Tillé (1999) have proposed a general method to select a random sample balanced on a large number of auxiliary variables.