Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 197-206
Statistique Canada, N° 12-001-XPB au catalogue

Extensions de la méthode d'échantillonnage indirect et son application aux enquêtes dans le tourisme

Jean-Claude Deville et Myriam Maumy-Bertrand [1]

Résumé

On doit procéder à une enquête portant sur la fréquentation touristique d'origine intra ou extra-régionale en Bretagne. Pour des raisons matérielles concrètes, les « enquêtes aux frontières » ne peuvent plus s'organiser. Le problème majeur est l'absence de base de sondage permettant d'atteindre directement les touristes. Pour contourner ce problème, on applique la *méthode d'échantillonnage indirect* dont la pondération est obtenue par la *méthode généralisée de partage des poids* développée récemment par Lavallée (1995), Lavallée (2002), Deville (1999) et présentée également dans Lavallée et Caron (2001). Cet article montre comment adapter cette méthode à l'enquête. Certaines extensions s'avèrent nécessaires. On développera l'une d'elle destinée à estimer le total d'une population dont on a tiré un échantillon bernoullien.

Mots clés : Méthode généralisée de partage des poids ; base incomplète et bases multiples.

1. Introduction

Une « enquête aux frontières » portant sur la fréquentation touristique extra-régionale en Bretagne (hormis celle des Bretons) a été réalisée sur la période d'avril à septembre 1997. L'Observatoire Régional du Tourisme de Bretagne et les Comités Départementaux de Tourisme aimeraient recommencer ce type d'enquête. Malheureusement ils n'ont plus la possibilité de recueillir une certaine masse d'informations récoltées aux frontières régionales ou intra-régionales, car les forces de police ne désirent plus collaborer à la réalisation d'enquêtes au bord des routes.

C'est pourquoi l'Observatoire Régional du Tourisme de Bretagne avec l'aide d'un comité technique constitué de méthodologues et d'opérateurs de terrain ont décidé de mettre en place une nouvelle méthodologie d'enquête en remplacement de la méthodologie des « enquêtes aux frontières ». De plus, l'évaluation de la part du tourisme intra-régional (des Bretons prenant des vacances en Bretagne, par exemple) est indispensable pour définir les facteurs de développement.

Un des problèmes majeurs est l'absence d'une base de sondage permettant d'interroger directement les touristes. Pour contourner ce problème, l'idée principale, déjà utilisée par la région des Asturies en Espagne (Valdés, De La Ballina, Aza, Loredo, Torres, Estébanez, Domínguez et Del Valle (2001) et Torres Manzanera, Sustacha Melijosa, Menéndez Estébanez et Valdés Pelaáez (2002)), est d'échantillonner des services destinés principalement aux touristes et de les interroger sur les différents lieux de ces nombreuses prestations touristiques. Il est bien évident qu'un touriste peut utiliser une ou plusieurs fois un ou plusieurs services de la base de sondage pendant la période d'enquête considérée. Pour pouvoir estimer des paramètres d'intérêts relatifs aux touristes, il faut avoir la possibilité d'échantillonner de façon rigoureuse certains services puis relier le jeu de poids des services échantillonnés au jeu de poids des touristes qui ont fréquenté ces services. Le but de cet article est de présenter une méthode qui permet de faire ce calcul. Cette méthode va s'appuyer principalement sur la *méthode généralisée de partage des poids* (MGPP) mise au point par Lavallée (1995), Lavallée (2002) et Deville (1999).

2. La méthode généralisée de partage des poids

On va rappeler très brièvement le principe de la *méthode généralisée de partage des poids* (MGPP). Pour de plus amples informations, on renvoie à Lavallée (1995), Lavallée (2002) et Deville (1999).

Soient U^A une population finie contenant N^A unités, où chaque unité est désignée par j et U^B une population finie contenant N^B unités, où chaque unité est désignée par i. La correspondance entre U^A et U^B peut être représentée par une matrice de liens $\Theta_{AB} = [\theta_{ji}^{AB}]$, de taille $N^A \times N^B$ où chaque élément $\theta_{ji}^{AB} \geq 0$. Autrement dit, l'unité j de U^A est reliée à l'unité i de U^B à condition que $\theta_{ji}^{AB} > 0$; sinon, il n'existe aucun lien entre les deux unités.

Dans le cas du sondage indirect, on sélectionne l'échantillon s^A de n^A unités à partir de U^A selon un plan d'échantillonnage donné. Soit $\pi_j^A > 0$, la probabilité de sélection de l'unité j. Pour chaque unité j sélectionnée dans s^A, on identifie les unités i de U^B pour lesquelles $\theta_{ji}^{AB} > 0$. Soit s^B, l'ensemble des n^B unités de U^B identifiées au moyen des unités $j \in s^A$, c'est-à-dire

$$s^B = \{i \in U^B ; \exists j \in s^A \text{ et } \theta_{ji}^{AB} > 0\}.$$

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquêtes, ENSAI/CREST, Campus de Ker-Lann, 35170 BRUZ (France). Courriel : deville@ensai.fr; Myriam Maumy-Bertrand, Laboratoire de Statistique de l'Université Louis Pasteur, 7, rue René Descartes 67084 STRASBOURG Cedex (France). Courriel : mmaumy@math.u-strasbg.fr.

Pour chaque unité i de s^B, une variable d'intérêt y_i est mesurée.

On suppose que, pour toute unité j de s^A, on peut obtenir les valeurs de θ_{ji}^{AB} pour $i = 1, ..., N^B$ par entrevue directe ou à partir d'une source administrative. Pour toute unité i identifiée de U^B (ou seulement de s^B), on suppose que l'on peut obtenir les valeurs de θ_{ji}^{AB} pour $j = 1, ..., N^A$. Par conséquent, il n'est pas nécessaire de connaître les valeurs de θ_{ji}^{AB} pour la totalité de la matrice de liens Θ_{AB}. En fait, on ne doit connaître les valeurs de θ_{ji}^{AB} que pour les lignes j de Θ_{AB}, où $j \in s^A$, ainsi que pour les colonnes i de Θ_{AB} où $i \in s^B$.

Par exemple si le but est d'estimer une variable d'intérêt Y^B de la population cible U^B, où

$$Y^B = \sum_{i=1}^{N^B} y_i, \qquad (2.1)$$

avec y_i mesurée d'après l'ensemble U^B. On utilise alors un estimateur de la forme

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i, \qquad (2.1)$$

où w_i est le poids d'estimation de l'unité i de s^B, avec $w_i = 0$ pour $i \notin s^B$. Pour obtenir une estimation sans biais d'une variable d'intérêt Y^B, il suffirait d'utiliser comme poids w_i l'inverse de la probabilité de sélection π_i^B de l'unité i. Comme il est mentionné dans Lavallée (1995) et Lavallée (2002), il est généralement difficile, voire impossible, d'obtenir ces probabilités. On a alors recours à la MGPP. Dans celle-ci les poids sont donnés par

$$w_i = \sum_{j \in s^A} \frac{\tilde{\theta}_{ji}^{AB}}{\pi_j^A},$$

où $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \sum_{j=1}^{N^A} \theta_{ji}^{AB}$. De cette construction, l'estimateur $\hat{Y}^B$ est sans biais. De même, la variance de cet estimateur peut-être calculée et estimée car elle est identique à celle de

$$\sum_{j \in s^A} \frac{z_j}{\pi_j^A},$$

avec $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$.

3. L'enquête tourisme en milieu ouvert

3.1 Objectifs de l'enquête

Le principe de l'enquête est le suivant :

« *atteindre les touristes* (*étrangers ou français habitant la Bretagne ou pas*) *par le biais de services destinés à satisfaire leurs besoins élémentaires ou spécifiques* »

comme l'hébergement, la nourriture, les activités de loisirs, les transports.

3.2 La population d'intérêt

Soit G *un champ géographique* (les quatre départements bretons) et P *une période de référence* (pour nous celle qui s'étend du mois de février 2005 au mois de décembre 2005).

Un touriste est une personne ayant passé au moins une nuit dans G hors de sa résidence principale (nuitée).

Pour un touriste, *un séjour* est un intervalle sej de P de durée le cardinal de sej noté $|sej|$, au cours duquel le touriste passe toutes ses nuits dans G hors résidence principale et, les nuits immédiatement avant ou après le séjour sej étant passées hors de G (ou à la résidence principale).

Un voyage est un ensemble de touristes (ménage touristique) partageant le même séjour et avec le même hébergement au cours du séjour. On utilisera aussi le terme de ménage touristique par un léger abus de langage (un même ménage touristique peut faire plusieurs voyages au cours d'une période, mais nous n'avons aucun moyen de les distinguer).

L'unité statistique i de l'enquête est le voyage.

Les sous unités d'enquête sont les séjours, les touristes et les nuitées. Un voyage i comporte n_i touristes pendant le séjour de durée $|sej|$ et donc $n_i \times |sej|$ nuitées. Ici la population U^B est donc l'ensemble des voyages dans G au cours de P. ($sej \cap P \neq \varnothing$).

3.3 Le plan de sondage de l'enquête

Pour utiliser la MGPP, la population théorique U^A est constituée par un ensemble de « services ». Dans cette enquête, ceux-ci sont constitués par :

- les achats en boulangerie, constituant une première strate de U^A.
- les visites d'un ensemble de sites culturels ou de loisirs ou familiaux très connus. En pratique, pour chacun d'eux, un « point de passage obligé » a été défini. C'est l'ensemble des passages par ce point qui est la seconde strate de U^A.
- les passages sortant de la Bretagne au péage autoroutier de La Gravelle qui regroupe environ 80 % des sorties des touristes de la Bretagne en voiture. Ce mode de transport caractérise lui-même 80 % des séjours de non-résidents bretons. Ce passage constitue la troisième strate de U^A.

En d'autres termes, *la base de sondage* est donc formellement constituée de trois strates :

1. les achats en boulangerie;
2. les visites d'un ensemble de sites emblématiques de la Bretagne;
3. le passage au péage autoroutier de La Gravelle.

Dans *la première strate*, on réalise un échantillon à trois degrés :

- un échantillon de boulangeries;
- un échantillon de jours d'enquête;
- un échantillon de clients dans la boulangerie à un jour donné.

Dans *la deuxième strate*, on réalise un échantillon à deux degrés :

- un échantillon de jours d'enquête;
- un échantillon de personnes qui passent sur un des 16 sites référés à un jour donné.

Enfin dans *la troisième strate*, on réalise un échantillon à deux degrés :

- un échantillon de jours d'enquête;
- un échantillon de personnes qui passent au péage autoroutier de La Gravelle à un jour donné.

On admet que tout ménage touristique consomme au moins un des « services » (achats en boulangeries, visites de sites emblématiques de la Bretagne, passage au péage autoroutier de La Gravelle), ou tout du moins, que très peu de ménages ne consomment aucun d'entre eux.

Chaque échantillonnage (boulangerie, jours, « service ») requiert des techniques particulières et il serait très long de détailler chacune d'elle. On donnera néanmoins les quelques indications techniques suivantes :

- les boulangeries sont échantillonnées selon un plan classique stratifié géographiquement (cinq strates : partie « littorale » des quatre départements bretons, intérieur de la Bretagne). Dans chaque strate les boulangeries sont échantillonnées avec des probabilités proportionnelles à leur « potentiel touristique » construit à partir de leur chiffre d'affaire et de la capacité d'hébergement touristique et du nombre de résidences principales de la commune à laquelle elles appartiennent. Théoriquement du moins, car pratiquement ce tirage a été un peu « forcé » par des circonstances fortuites (refus de boulangers, fermetures durant certaines périodes par exemple).
- Les sites ne sont pas échantillonnés mais choisis pour leur notoriété et pour la possibilité technique d'y définir un « point de passage obligé » (parfois approximativement).
- Pour chaque boulangerie, chaque site et pour le péage autoroutier de La Gravelle, on a défini des « grappes de jours » complétement homogènes de chaque période P. Une grappe a été attribuée aléatoirement à chaque boulangerie, site ainsi qu'au péage autoroutier de La Gravelle. Pratiquement cela signifie qu'un enquêteur employé à plein temps est mobilisé sur plusieurs grappes.
- Pour chacun des « services » les utilisateurs sont échantillonnés selon des techniques habituelles de sélection aléatoire à mesure des arrivées : échantillon pseudo-systématique car pendant que l'enquêteur fait accepter un questionnaire, des gens passent sans qu'il puisse compter. Le nombre total de visiteurs ne peut donc pas être estimé directement. Si un site est accessible par une billeterie (musée ou château par exemple) l'échantillonnage s'appuie sur elle. Au final, l'échantillon d'utilisateurs d'un « service » à un jour donné est considéré comme un échantillon bernoullien, c'est-à-dire un sondage aléatoire simple si on connait la taille de la population c'est-à-dire le nombre de visiteurs du jour donné.

Remarques 3.1. La définition même du *touriste* est liée à l'hébergement, et il parait naturel d'utiliser une base directement liée à ce service. La pratique montre que c'est difficilement réalisable.

On n'a, d'abord, aucune base de sondage correcte pour l'hébergement non marchand (parents, amis, résidence secondaire) ni pour les locations meublées saisonnières.

Pour l'hébergement en hôtels, campings et gîtes familiaux, les tests de l'été 2004 ont montré l'existence de biais catastrophiques liés à l'intervention des hôteliers dans le processus de sélection des enquêtés. Ceux-ci ne respectent absolument pas les consignes d'échantillon aléatoire et distribuent « essentiellement » les questionnaires à leurs bons clients. Cette partie du dispositif de l'enquête a dû être abandonnée et remplacée par le passage au péage autoroutier de La Gravelle, qui est régulièrement l'objet d'enquêtes de qualité honnête faites par divers organismes.

Par ailleurs, les questionnaires collectés dans les boulangeries et sur les sites emblématiques de la Bretagne pendant l'été 2004, rendent apparemment (qualitativement et quantitativement) bien compte des différents modes d'hébergement.

De même, l'alimentation eut sans doute mieux été capturée par des questionnaires à la sortie des supermarchés. Mais là, le problème réside dans l'hétérogénéité de ces établissements et dans la lutte au couteau que se livrent les enseignes, le groupe C... accepte les enquêtes dans ses établisements uniquement si le groupe I ... en est exclu ! En revanche, l'adhésion des artisans boulangers au concept de l'enquête a été excellente.

Remarques 3.2. Par définition même de la méthode utilisée, on se place formellement dans le contexte de l'échantillonnage à partir de bases multiples. Le problème a donné lieu à une abondante littérature (Hartley (1962), Lund (1968) et Hartley (1974) pour le moins). La MGPP s'applique à ce problème en considérant tout simplement

chaque base de sondage comme une strate à la condition de pouvoir identifier, pour chaque unité échantillonnée, l'ensemble des bases dans laquelle elle figure. Elle fournit alors une solution rigoureuse, efficace et uniquement basée sur le plan à ce problème. Cette remarque pourrait fonder un article autonome, mais les auteurs savent que cela n'en vaut pas la peine : une idée qui s'exprime en dix lignes n'a pas besoin d'un article ou d'un livre pour sa survie.

4. Les paramètres d'intérêt

On définit l'application F, qui à tout service j durant la période de référence P dans les trois types d'établissements du champ de l'enquête, associe le voyage i utilisateur de ce service.

$$\begin{aligned} F : \text{services} &\rightarrow \text{voyage} \\ j &\rightarrow F(j) = i. \end{aligned}$$

Soit U^B, la population des voyages i de la période de référence P. Cette population d'intérêt U^B est l'image par F de l'ensemble des services durant la période de référence P dans les trois types d'établissements du champ de l'enquête. La population U^A est l'image par F^{-1} de l'ensemble des voyages durant la période de référence P. Pour tout $i \in U^B$, on définit $R_i(B) = \text{card}(F^{-1}(i))$, le nombre d'antécédents de i au cours de la période d'enquête, c'est-à-dire, le nombre de services j utilisés par le ménage touristique i donné.

Les paramètres d'intérêt peuvent être des totaux, des effectifs ou des ratios. Supposons par exemple, que l'on s'intéresse à l'estimation d'un total relatif à une variable y définie sur la population U^B,

$$Y^B = \sum_{i \in U^B} y_i. \qquad (4.1)$$

Un cas particulier de ces totaux est l'effectif de U^B, noté N^B et défini par

$$N^B = \text{card}(U^B) = \sum_{i \in U^B} 1.$$

Par exemple, Y^B peut-être le nombre de personnes ayant pratiqué une certaine activité, le budget total dépensé par le ménage touristique à l'intérieur de la Bretagne, la provenance géographique des ménages touristiques, le nombre de jours que le ménage touristique passe en Bretagne. Il faut noter que pour beaucoup de variables, le total Y^B dépend de la taille du ménage touristique, c'est-à-dire le nombre de personnes qui forment ce groupe et de la longueur du séjour (uniquement les jours passés en Bretagne).

Désormais, on peut écrire

$$Y^B = \sum_{i \in U^B} y_i = \sum_{l=1}^{3} \sum_{a_l \in A_l} \sum_{d_l \in D_l} \sum_{j \in C_{d_l}} z_j, \qquad (4.2)$$

où

$$z_j = \frac{y_i}{R_i(B)}, \quad \text{pour} \quad j \in F^{-1}(i),$$

où

- A_1 : l'ensemble des boulangeries du champ de l'enquête repéré par l'indice a_1
- A_2 : les 16 lieux de passage du champ de l'enquête repérés par l'indice a_2
- A_3 : le péage autoroutier de La Gravelle repéré par l'indice a_3
- D_l : l'ensemble des jours d'enquête, repérés par l'indice d_l dans un établissement a_l de A_l, pour l variant de 1 à 3
- C_{d_l} : l'ensemble des services dans un établissement a_l de A_l de la journée d_l de D_l repérés par l'indice j.

5. Estimation sans biais d'un total

Dans le paragraphe précédent, on a montré que le total d'intérêt s'écrit comme un total sur l'ensemble des services du champ. Supposons que l'on dispose d'un échantillon de services répondants j, auxquels on peut associer des poids de sondage δ_j. Ces poids sont supposés sans biais car l'échantillon de services suit les canons d'un échantillon à plusieurs degrés, chaque sondage élémentaire étant sans biais.

Pour alléger les notations, on ne fait pas apparaître, dans ce qui suit, tous les degrés de tirage de l'échantillon en fonction de l'établissement a_l. Soient

- s^B : l'ensemble des ménages touristiques i correspondant à l'ensemble des services échantillonnés au cours de la période d'enquête
- s_{A_l} : l'ensemble des établissements échantillonnés
- s_{D_l} : l'ensemble des jours échantillonnés dans l'établissement a_l
- s_{d_l} : le sous-échantillon de services j correspondant au jour de l'établissement a_l.

Disposant d'un jeu de poids de sondage δ_j pour les services répondants, et si on connaît les $R_i(B)$, on estime alors le total Y^B sans biais par

$$\hat{Y}^B = \sum_{i \in s^B} w_i \, y_i \qquad (5.1)$$

où

$$w_i = \frac{\sum_{l=1}^{3} \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{d_l}} \delta_j}{R_i(B)}.$$

On est ramené à une estimation sur la population des ménages touristiques. Cette formule n'est autre que celle donnée par la MGPP évoquée dans la section 2. Notons que $U^A = U^{A_1} \cup U^{A_2} \cup U^{A_3} = \bigcup_{l=1}^{3} U^{A_l}$, $\theta_{ji}^{AB} = 1$ si le service j a été utilisé par le voyage i et enfin $\delta_j = 1/\pi_j^A$.

L'estimation de la variance est possible selon les mêmes principes (*cf.* Lavallée (2002)). Elle ne sera pas détaillée ici car elle n'est qu'une application assez lourde en calcul des principes généraux.

De même, l'utilisation des informations auxiliaires sous forme de totaux, que ce soit dans les populations U^{A_l} ou dans la population U^B, ne pose pas de problème particulier que ce soit pour l'estimation ponctuelle ou pour l'estimation de la variance (*cf.* Lavallée (2002)).

Remarques 5.1. La procédure qui vient d'être décrite pour partager les poids peut-être considérée comme naïve. De fait, on sait optimiser la matrice de liens Θ_{AB} comme il est montré dans Deville et Lavallée (2006). L'application de l'enquête bretonne est décrite dans Deville, Lavallée et Maumy (2005).

6. Un exemple de problème particulier : Les points de visite en rase campagne

Comme on l'a déjà signalé, la mise en place de l'enquête sur le tourisme en Bretagne a nécessité d'assez nombreuses recherches complémentaires. On a déjà signalé ce qui concerne l'optimisation du partage des poids. L'utilisation d'informations auxiliaires relatives aux diverses bases et aux divers degrés de sondage est un autre chantier. On voudrait insister ici sur celui de l'estimation de certaines de ces informations auxiliaires, en particulier pour ce qui concerne les visites des sites touristiques en rase campagne.

Dans certains cas, on ne connaît malheureusement pas le nombre total de personnes, noté $T_P^{A_2}$, venant sur le site à un jour donné. En effet, dans l'ensemble A_2, on ne connaît pas tous les services (ici le nombre de visites) de la population. On ne peut donc pas avoir directement $\pi_j^{A_2}$ et donc δ_j pour $j \in A_2$. Pour contourner ce problème, on estime alors le nombre de visiteurs journaliers afin de déduire $\tilde{\pi}_j^{A_2} = n_{A_2} / \hat{T}_P^{A_2}$.

Dans la suite, on va développer deux approches d'estimation du nombre de visiteurs journaliers pour des sites accessibles en voitures uniquement (ou presque !). La première se base sur un système d'échantillonnage de voitures destiné à estimer le nombre de visiteurs sur le site. La seconde approche utilise un échantillon de visiteurs et est destinée à estimer la même quantité à partir de l'individu interrogé qui donne le nombre de personnes qui voyagent avec lui dans la voiture. Ces deux approches sont développées dans les sections 7 et 8 suivantes.

7. Construction d'un estimateur du nombre de visiteurs à partir d'un échantillonnage de voitures

Dans ce paragraphe, on est dans le cas où un enquêteur relève en « bâtonnant » (c'est le terme utilisé par les praticiens du tourisme) le nombre d'occupants des voitures, c'est-à-dire, relève le nombre de personnes dans une voiture qui franchissent l'endroit où un oeil électronique ou un système équivalent a été placé pour compter les voitures dont le nombre total noté T_V est connu avec une erreur de mesure négligeable près.

7.1 Définition et variance de $\hat{T}_P^{A_2}$

Le nombre total de voitures vaut

$$T_V = \sum_{\kappa=1,\ \ldots} t_\kappa = \sum_{l \in U_v} 1, \qquad (7.1)$$

où t_κ représente le nombre de voitures transportant κ personnes et U_V l'univers des voitures.

Remarques 7.1. Dans un souci d'allégement des notations, on utilisera ici et jusqu'à la fin de cet article, T_P pour $T_P^{A_2}$.

Le nombre total de personnes visitant le site vaut

$$T_P = \sum_{\kappa=1,\ \ldots} \kappa t_\kappa = \sum_{k \in U_P} 1, \qquad (7.2)$$

où U_P désigne l'univers des personnes. On a aussi l'égalité

$$T_P = \sum_{l \in U_v} v_l, \qquad (7.3)$$

où v_l est le nombre de personnes dans la voiture l.

Comme on l'a mentionné dans la section précédente, le nombre total de personnes T_P est inconnu. Par conséquent construisons un estimateur de T_P. Soit $\hat{T}_P$ le $\pi-$estimateur fondé sur s_V un échantillon aléatoire simple de voitures de taille n et de probabilité d'inclusion n/T_V

$$\hat{T}_P = \frac{T_V}{n} \sum_{l \in s_V} v_l = T_V \bar{v}, \qquad (7.4)$$

en posant

$$\bar{v} = \frac{1}{n}\left(\sum_{l \in s_V} v_l \right).$$

Il est clair que $\hat{T}_P$ est un estimateur sans biais du nombre total de personnes T_P et que $\bar{v}$ estime sans biais le nombre moyen $\bar{V}$ de personnes dans une voiture.

La variance de $\hat{T}_P$ est donc égale à

$$\begin{aligned}\mathrm{Var}[\hat{T}_P] &= T_V^2\left(\frac{1}{n} - \frac{1}{T_V}\right)S_V^2 \\ &= \frac{1}{n}\, T_V^2\, S_V^2 - T_V\, S_V^2, \end{aligned} \tag{7.5}$$

où S_V^2 désigne la variance corrigée de la population U_V.

7.2 Construction d'un estimateur d'une variable d'intérêt dans le cas d'un échantillonnage de voitures

On veut estimer une variable d'intérêt Y de la population U_P qui s'écrit sous la forme

$$Y = \sum_{k\in U_P} y_k, \tag{7.6}$$

où y_k est la variable d'intérêt qu'on mesure dans le questionnaire final. Soit $\hat{Y}$ le π – estimateur défini par

$$\hat{Y} = \sum_{k\in s_P} w_k^P\, y_k, \tag{7.7}$$

où le poids w_k^P est égal à $\hat{T}_P / m$. Par conséquent l'estimateur $\hat{Y}$ peut s'écrire

$$\hat{Y} = \frac{\hat{T}_P}{m}\sum_{k\in s_P} y_k = \hat{T}_P\,\bar{y} \tag{7.8}$$

en posant

$$\bar{y} = \frac{1}{m}\left(\sum_{k\in s_P} y_k\right).$$

Par la suite, les variables $\hat{T}_P$ et $\bar{y}$ seront supposées indépendantes. L'hypothèse est réaliste, car sur le terrain nous avons recours à deux enquêteurs indépendants.

7.2.1 Calcul de la variance de l'estimateur $\hat{Y}$

D'après le théorème de Huygens (1673), en conditionnant selon l'échantillon s_V, on obtient

$$\begin{aligned} V_Y &= \mathrm{Var}[\hat{Y}] \\ &= \bar{Y}^2\,\mathrm{Var}[\hat{T}_P] + T_P^2\,\mathrm{Var}[\bar{y}] \\ &\quad + \mathrm{Var}[\hat{T}_P]\,\mathrm{Var}[\bar{y}]. \end{aligned} \tag{7.9}$$

Dans le cas présent, on assimile l'échantillon à un sondage aléatoire simple sans remise. L'égalité (7.9) devient alors

$$\begin{aligned} V_Y &= \bar{Y}^2\left(\frac{1}{n}\,T_V^2\,S_V^2 - T_V\,S_V^2\right) \\ &\quad + T_P^2\left(\frac{1}{m}S_Y^2 - \frac{S_Y^2}{T_P}\right) \\ &\quad + \left(\frac{1}{n}T_V^2\,S_V^2 - T_V\,S_V^2\right)\left(\frac{1}{m}S_Y^2 - \frac{S_Y^2}{T_P}\right), \end{aligned}$$

avec $S_Y^2 = 1/(T_P - 1)\sum_{k\in U_P}(y_k - \bar{Y})^2$. En réorganisant les termes, on obtient

$$\begin{aligned} V_Y &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right)T_V^2\,S_V^2\,\frac{1}{n} \\ &\quad + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ &\quad + T_V^2\,S_V^2\,S_Y^2\,\frac{1}{nm} + \frac{T_V}{T_P}\,S_V^2\,S_Y^2 \\ &\quad - \bar{Y}^2\,T_V\,S_V^2 - T_P\,S_Y^2. \end{aligned}$$

On cherche maintenant l'allocation des tailles des échantillons s_P et s_V qui minimise la variance de l'estimateur $\hat{Y}$ pour des tailles de population T_P et T_V fixées.

On doit donc minimiser l'égalité (7.10) en n, m sous la contrainte

$$C_V\, n + C_P\, m = C,$$

où C_V désigne le coût (en temps par exemple) des questionnaires posés autour des voitures, C_P le coût (en temps) des questionnaires posés aux personnes et C le coût total.

On peut écrire l'équation lagrangienne

$$\begin{aligned} L(n,\, m,\, \lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right)T_V^2\,S_V^2\,\frac{1}{n} \\ &\quad + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ &\quad + T_V^2\,S_V^2\,S_Y^2\,\frac{1}{nm} + \frac{T_V}{T_P}\,S_V^2\,S_Y^2 \\ &\quad - \bar{Y}^2\,T_V\,S_V^2 - T_P\,S_Y^2 \\ &\quad + \lambda(C_V\, n + C_P\, m - C). \end{aligned} \tag{7.11}$$

En annulant les dérivées partielles par rapport aux variables n, m, λ, on obtient

$$\begin{aligned} \frac{\partial L}{\partial n}(n,\, m,\, \lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right)T_V^2\,S_V^2\left(-\frac{1}{n^2}\right) \\ &\quad + T_V^2\,S_V^2\,S_Y^2\left(-\frac{1}{mn^2}\right) \\ &\quad + \lambda\, C_V = 0, \end{aligned}$$

$$\frac{\partial L}{\partial m}(n,\ m,\ \lambda) = (T_P^2 - T_V\ S_V^2)\, S_Y^2 \left(-\frac{1}{m^2}\right) + T_V^2\ S_V^2\ S_Y^2 \left(-\frac{1}{nm^2}\right) + \lambda\, C_P = 0,$$

$$\frac{\partial L}{\partial \lambda}(n,\ m,\ \lambda) = C_V\ n + C_P\, m - C = 0.$$

Après calculs, on obtient une équation du troisième degré en n qui s'écrit

$$\lambda C_V^2\, n^3 - \lambda C_V C n^2 - C_V T_V^2\ S_V^2 \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) n + T_V^2\ S_V^2 \left(C\left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) + C_P\, S_Y^2\right) = 0.$$

Cette équation du troisième degré en n admet une solution réelle que l'on peut déterminer avec des méthodes numériques.

En faisant le même raisonnement, on obtient une équation du troisième degré en m

$$\lambda C_P^2\, m^3 - \lambda C_P C m^2 - C_P\, S_Y^2\, (T_P^2 - T_V\ S_V^2)\, m + S_Y^2\, (C(T_P^2 + T_V\ S_V^2) + C_V\, T_V^2\ S_V^2) = 0.$$

7.2.2 Cas simplifié

Pour simplifier le calcul de la variance de l'estimateur $\hat{Y}$, nous pouvons faire une approximation dans l'égalité (7.10). En effet, nous pouvons supposer que le terme $1/nm$ est négligeable devant les termes $1/n$ et $1/m$.

On obtient alors la transformation suivante de l'égalité (7.10)

$$V_Y = \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\ S_V^2\ \frac{1}{n} + (T_P^2 - T_V\ S_V^2)\ S_Y^2\ \frac{1}{m} + \frac{T_V}{T_P} S_V^2\ S_Y^2 - \overline{Y}^2\, T_V\ S_V^2 - T_P\ S_Y^2. \qquad (7.12)$$

On cherche maintenant l'allocation des tailles des échantillons s_P et s_V qui minimise la variance de l'estimateur $\hat{Y}$ pour des tailles de population T_P et T_V fixées.

On doit donc minimiser l'égalité (7.12) en n, m sous la contrainte

$$C_V\ n + C_P\, m = C.$$

On peut écrire l'équation lagrangienne

$$L(n,\ m,\ \lambda) = \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\ S_V^2\, \frac{1}{n} + (T_P^2 - T_V\ S_V^2)\, S_Y^2\, \frac{1}{m} + \frac{T_V}{T_P} S_V^2\ S_Y^2 - \overline{Y}^2\, T_V\ S_V^2 - T_P\ S_Y^2 + \lambda\, (C_V\, n + C_P\, m - C). \qquad (7.13)$$

En annulant les dérivées partielles par rapport aux variables n, m, λ, on obtient

$$\frac{\partial L}{\partial n}(n,\ m,\ \lambda) = \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\ S_V^2 \left(-\frac{1}{n^2}\right) + \lambda\, C_V = 0,$$

$$\frac{\partial L}{\partial m}(n,\ m,\ \lambda) = (T_P^2 - T_V\ S_V^2)\, S_Y^2 \left(-\frac{1}{m^2}\right) + \lambda\, C_P = 0,$$

$$\frac{\partial L}{\partial \lambda}(n,\ m,\ \lambda) = C_V\ n + C_P\, m - C = 0.$$

Après calculs, on obtient

$$n_{\text{opt}} = \frac{C}{\left(C_V + \sqrt{C_P\, C_V\, \dfrac{T_P\, S_Y^2\, (T_P^2 - T_V\, S_V^2)}{T_V^2\, S_V^2\, (T_P\, \overline{Y}^2 - S_Y^2)}}\right)},$$

$$m_{\text{opt}} = \frac{C}{\left(C_V + \sqrt{C_P\, C_V\, \dfrac{T_V^2\, S_V^2\, (T_P\, \overline{Y}^2 - S_Y^2)}{T_P\, S_Y^2\, (T_P^2 - T_V\, S_V^2)}}\right)}.$$

8. Construction d'un estimateur du nombre de visiteurs à partir d'un échantillonnage de visiteurs

La méthode précédente peut s'avérer compliquée et coûteuse à réaliser sur certains sites. On peut obtenir une collecte plus simple en demandant à la personne k le nombre u_k de passagers de la voiture i qui l'a transportée. Ce nombre u_k est ici égal à v_l pour la voiture l qui a transporté la personne k. Cette méthode a en outre l'avantage d'obtenir avec précision le nombre de passagers au sens de l'enquête (compte-t-on les bébés ?).

8.1 Définition de $\hat{\hat{T}}_P$

Rappelons l'égalité suivante

$$T_P = \sum_{l \in U_V} v_l,$$

où v_l désigne le nombre de passagers de la voiture l. Rappelons également

$$T_P = \sum_{l \in U_P} 1.$$

Le nombre moyen de passagers dans une voiture $\overline{V}$ peut s'exprimer sous la forme

$$\overline{V} = \frac{\sum_{l \in U_V} v_l}{\sum_{l \in U_V} 1} = \frac{\sum_{\kappa=1,\ \ldots} \kappa t_\kappa}{\sum_{\kappa=1,\ \ldots} t_\kappa} = \frac{\sum_{\kappa=1,\ \ldots} m_\kappa}{\sum_{\kappa=1,\ \ldots} M_\kappa / \kappa}, \tag{8.1}$$

où t_κ est le nombre de voitures à κ passagers et M_κ le nombre de personnes venues dans une voiture à κ passagers.

Cette dernière relation permet de donner une dernière écriture de T_P

$$T_P = T_V \overline{V}. \tag{8.2}$$

Par conséquent un estimateur de T_P s'écrit sous la forme suivante

$$\hat{\hat{T}}_P = T_V \hat{\overline{V}}, \tag{8.3}$$

où le nombre total de voitures T_V est parfaitement connu. En observant cette expression, on constate que pour connaître l'estimateur $\hat{\hat{T}}_P$, il suffit de déterminer la quantité $\hat{\overline{V}}$. Introduisons alors l'estimateur suivant de $\overline{V}$

$$\hat{\overline{V}} = \frac{\sum_{\kappa \in s_P} m_\kappa}{\sum_{\kappa \in s_P} m\kappa / \kappa},$$

où m_κ est le nombre de personnes de l'échantillon voyageant dans une voiture à κ passagers. L'estimateur $\hat{\overline{V}}$ peut s'écrire également de la façon suivante

$$\hat{\overline{V}} = \frac{\sum_{k \in s_P} 1}{\sum_{k \in s_P} 1/u_k}$$

ou encore

$$\hat{\overline{V}} = \frac{m}{\sum_{k \in s_P} 1/u_k}. \tag{8.4}$$

Cette dernière égalité nous permet d'écrire l'égalité suivante

$$\frac{1}{\hat{\overline{V}}} = \frac{1}{m} \sum_{k \in s_P} \frac{1}{u_k}. \tag{8.5}$$

Cette dernière quantité représente la moyenne empirique des $1/u_k$ et $\hat{\overline{V}}$ est la moyenne harmonique des u_k. On peut d'ailleurs calculer sa variance qui est égale à

$$\mathrm{Var}\left[\frac{1}{\hat{\overline{V}}}\right] = \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \tag{8.6}$$

8.2 Calcul de la variance de l'estimateur de $\hat{\hat{T}}_P$ sans échantillonnage de voitures

Reste à calculer la variance de l'estimateur $\hat{\overline{V}}$ sachant (8.6). Pour cela, remarquons que l'on peut écrire

$$\begin{aligned}\frac{1}{\hat{\overline{V}}} &= \frac{1}{\overline{V}\left(\frac{\hat{\overline{V}}}{\overline{V}} - 1 + 1\right)} \\ &= \frac{1}{\overline{V}} \times \frac{1}{1 + \frac{\hat{\overline{V}} - \overline{V}}{\overline{V}}} \\ &= \frac{1}{\overline{V}}\left(1 - \frac{\hat{\overline{V}} - \overline{V}}{\overline{V}} + o\left(\frac{\hat{\overline{V}} - \overline{V}}{\overline{V}}\right)\right).\end{aligned}$$

Par conséquent, on obtient

$$\mathrm{Var}\left[\frac{1}{\hat{\overline{V}}}\right] \simeq \left(\frac{1}{\overline{V}}\right)^2 \times \frac{\mathrm{Var}[\hat{\overline{V}}]}{\overline{V}^2}.$$

Finalement, on a

$$\mathrm{Var}[\hat{\overline{V}}] \simeq \overline{V}^4 \times \mathrm{Var}\left[\frac{1}{\hat{\overline{V}}}\right],$$

ou encore, avec (8.6)

$$\mathrm{Var}[\hat{\overline{V}}] \simeq \overline{V}^4 \times \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \tag{8.7}$$

Or par définition, la variance $S^2_{1/u}$ est égale à

$$S^2_{1/u} = \frac{1}{T_P - 1} \sum_{k \in U_P} \left(\frac{1}{u_k} - \frac{1}{\overline{V}}\right)^2. \tag{8.8}$$

Comme la quantité T_P est inconnue, cette relation peut être estimée par

$$\frac{1}{m - 1} \sum_{k \in s_P} \left(\frac{1}{u_k} - \frac{1}{\overline{V}}\right)^2. \tag{8.9}$$

Grâce à (8.7) et à (8.9) on peut donc connaître facilement la variance de l'estimateur $\hat{\overline{V}}$ et par conséquent celle de l'estimateur $\hat{\hat{T}}_P$ et finalement celle de la variable d'intérêt $\hat{Y}$.

Remarques 8.1. L'estimateur $\hat{\hat{T}}_P$ est biaisé et asymptotiquement sans biais.

Remarque 8.2. Si les variables $\hat{\hat{T}}_P$ et $\overline{y}$ ne sont pas indépendantes alors on aurait

$$\begin{aligned}\mathrm{Var}\left[\hat{\hat{T}}_P\,\bar{y}\right] &= \bar{Y}^2\,\mathrm{Var}\left[\hat{\hat{T}}_P\right] + T_P^2\,\mathrm{Var}[\bar{y}] \\ &\quad + \mathrm{Var}\left[\hat{\hat{T}}_P\,\bar{y}\right]\mathrm{Var}[\bar{y}] \\ &\quad + \text{termes liées à la non indépendance éventuelle des variables } \hat{\hat{T}}_P \text{ et } \bar{y}.\end{aligned}$$

9. Illustration numérique

Un compteur mécanique d'un site en rase campagne donne $T_V = 100$ voitures. On supose qu'il y a 20 % de voitures à une personne, 20 % de voitures à deux personnes, 20 % de voitures à trois personnes, 20 % de voitures à quatre personnes, 20 % de voitures à cinq personnes. Ainsi, on a 300 visiteurs sur ce site. La variance S_V^2 est égale à deux en négligeant les corrections de population finie. Le nombre moyen de passagers $\bar{V}$ est de trois. En effet, on a :

$$\begin{aligned}\frac{1}{V} &= \frac{1}{1}\times\frac{20}{300} + \frac{1}{2}\times\frac{40}{300} + \frac{1}{3}\times\frac{60}{300} \\ &\quad + \frac{1}{4}\times\frac{80}{300} + \frac{1}{5}\times\frac{100}{300} = \frac{1}{3}.\end{aligned}$$

D'où $\bar{V} = 3$.

Calculons maintenant une estimation de $S_{1/u}^2$. Après simplifications de (8.8) et en supposant que T_P est suffisamment grand devant un, on a

$$S_{1/u}^2 \simeq \frac{1}{T_P}\sum_{k\in U_P}\frac{1}{u_k^2} - \left(\frac{1}{\bar{V}}\right)^2.$$

Ainsi, on a

$$\begin{aligned}S_{1/u}^2 &= \frac{1}{30}\left(2 + 1 + \frac{2}{3} + \frac{1}{2} + \frac{2}{5}\right) - \frac{1}{3^2} \\ &= \frac{1}{30}\left(\frac{60 + 30 + 20 + 15 + 12}{30}\right) - \frac{1}{3^2} \\ &= \frac{137}{30^2} - \frac{1}{3^2} = \frac{37}{30^2}.\end{aligned}$$

Pusique nous connaissons $S_{1/u}^2$, nous pouvons calculer la variance de l'estimateur $\bar{V}$. Ainsi on a

$$\mathrm{Var}[\hat{\bar{V}}] \simeq 3^4 \times \frac{37}{30^2}\times\frac{1}{m}.$$

Enfin, on peut calculer la variance de l'estimateur $\hat{\hat{T}}_P$

$$\begin{aligned}\mathrm{Var}\left[\hat{\hat{T}}_P\right] &= T_V^2\,\mathrm{Var}\left[\hat{\bar{V}}\right] \\ &\simeq 10^4 \times 3^4 \times \frac{37}{30^2}\times\frac{1}{m}.\end{aligned}$$

La première approche donne une variance de l'estimateur $\hat{T}_P$ égale à

$$\mathrm{Var}\left[\hat{T}_P\right] = 10^4 \times 2 \times \frac{1}{n}.$$

Donc, afin que l'estimateur $\hat{\hat{T}}_P$ ait la même variance que l'estimateur $\hat{T}_P$, il suffit que la taille m de l'échantillon s_P soit égale à

$$m \simeq 1{,}66\,n.$$

En première conclusion, on peut dire que la seconde approche rend les opérations de terrain plus simples et moins coûteuses en termes de personnel car elle ne nécessite qu'un seul enquêteur. Elle est plus précise qu'un comptage sans contact pour obtenir la composition du ménage touristique. Elle ne nécessite qu'un échantillon environ une fois et demie plus gros que la première approche pour apporter la même précision, ce qui est tout à fait tolérable vu la simplification de la collecte qui en résulte. En pratique donc, sur tous les sites on appliquera de préférence la seconde méthode.

Conclusion

Cet article a présenté les grandes lignes d'une nouvelle méthode applicable à la statistique du tourisme. Elle consiste à saisir les touristes à partir de la consommation de certains services sur lesquels on sait construire des échantillons probabilistes. La méthode de partage des poids permet de passer de l'exactitude statistique sur les services à l'exactitude sur les unités statistiques pertinentes en tourisme : le voyage, le séjour, le ménage touristique, le touriste ou la nuitée. Cependant la méthode requiert de nombreuses adaptations et compléments au partage des poids. On s'est attardé à l'une d'elles qui est l'estimation du nombre de visiteurs d'un site en rase campagne. Deux méthodes pouvaient être mises en concurrence. L'une, plus précise en terme de taille d'échantillon, demande en fait une organisation relativement lourde et fait courir le risque d'erreurs de mesures désagréables. Au prix d'une collecte de données un peu plus abondante, on préfère donc la seconde méthode.

D'autres études de ce genre ont été faites avant et pendant la réalisation de l'enquête, de sorte que la méthodologie complète est difficile à résumer en un seul article.

Remerciements

Les auteurs remercient chaleureusement les deux arbitres et l'éditeur associé qui ont grandement contribué à améliorer la lisibilité de ce texte.

Bibliographie

Deville, J.-C. (1999). Les enquêtes par panel : En quoi différent-elles des autres enquêtes ? suivi de : Comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistiques*, *INSEE Méthodes*, 84-85-86, 63-82.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : Les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*. 32, 2, 185-196.

Deville, J.-C., Lavallée, P. et Maumy, M. (2005). Composition, factorisation et conditions d'optimalité (faible, forte) dans la méthode de partage des poids. Application à l'enquête sur le tourisme en Bretagne. *Actes des journées de méthodologie statistiques*, *INSEE Méthodes*.

Hartley, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected apllications. *Sankhyā*, Series C, 36, 99-118.

Huygens, C. (1673). Horologium Oscillatorium sive de motu pendulorum.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

Lavallée, P. (2002). Le sondage indirect, ou la méthode généralisée du partage des poids. Éditions de l'Université de Bruxelles, éditions Ellipses, Bruxelles.

Lavallée, P., et Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements. *Techniques d'enquête*, 27, 171-187.

Lund, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 282-288.

Torres Manzanera, E., Sustacha Melijosa, I., Menéndez Estébanez, J.M. et Valdés Pelaáez, L. (2002). A solution to problems and disadvantages in statistical operations of surveys of visitors at accommodation establishments and at popular visitors places. (Éd. Ákos Probáld). *Proceedings Of The Sixth International Forum On Tourisme Statistics. Hungarian Central Statistical Office*, *Budapest*.

Valdés, L., De La Ballina, J., Aza, R., Loredo, E., Torres, E., Estébanez, J.M., Domínguez, J.S. et Del Valle, E. (2001). A methodology to measure tourism expenditure and total tourism production at the regional level. Dans *Tourism Statistics* : *International Perspectives and Current Issues*, (Éd. Lennon, J.J.), Continuum, Grande Bretagne, 317-334.