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

Jean-Claude Deville, Myriam Maumy

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

JEAN-CLAUDE DEVILLE * et **MYRIAM MAUMY** †

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* ou encore appelée la *méthode généralisée de partage des poids* développée récemment par Lavallée (1995) et Lavallée (2002) et présentée également dans Lavallée et Caron (2001). Cet article montre comment adapter cette méthode à l'enquête. Nous développerons des extensions nécessaires dans ce sens.

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, lévaluation 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 (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 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) et Lavallée (2002).

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

*Laboratoire de Statistique d'Enquête, ENSAI/crest, Campus de Ker-Lann , 35170 BRUZ (France), deville@ensai.fr

†Laboratoire de Statistique de l'Université de Rennes 2, Place du recteur Henri Le Moal, CS 24307, 35043 RENNES cedex (France), myriam.maumy@uhb.fr

renvoit 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 $\mathbf{\Theta}_{AB} = [\theta_{ji}^{AB}]$, 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 2 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

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

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

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

$$Y^B = \sum_{i=1}^{N^B} y_i, \tag{2.1}$$

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

$$\widehat{Y}^B = \sum_{i=1}^{N^B} w_i y_i, \tag{2.2}$$

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

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

où $\widetilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \sum_{j=1}^{N^A} \theta_{ji}^{AB}$. De cette construction, l'estimateur $\widehat{Y}^B$ 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 $z_j = \sum_{i \in N^B} \widetilde{\theta}_{ji}^{AB} y_i$.

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

Pour un touriste t, **un séjour** est un intervalle s de P de durée le cardinal de s noté $|s|$, au cours duquel le touriste passe toutes ses nuits dans G hors résidence principale et, les nuits immédiatement avant ou après s 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.

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

Les sous unités d'enquête sont les séjours, les touristes et les nuitées. Un voyage v comporte n_V touristes pendant le séjour de durée $|s|$ et donc $n_v \times |s|$ nuitées. La population U^B est donc l'ensemble des voyages dans G au cours de P. $(s \cap P \neq \emptyset)$.
La population d'intérêt est constituée des personnes qui ont fréquenté au moins un service destiné en principe aux touristes du champ de l'enquête pendant la période de référence.

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 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 3 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 3 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 2 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 2 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

4 Les paramètres d'intérêt

Introduisons les notations dont nous aurons besoin dans la suite de cet article. Soient
- 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 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.

On définit l'application F , qui à tout service j durant la période de référence D dans les 3 types d'établissements du champ de l'enquête, associe le ménage touristique i utilisateur de ce service.

$$\begin{array}{ccc} F : \text{services} & \rightarrow & \text{ménage touristique} \\ j & \rightarrow & F(j) = i. \end{array}$$

Soit U^B, la population des ménages touristiques i de la période de référence D. Cette population d'intérêt U^B est l'image par F de l'ensemble des services durant la période de référence D dans les 3 types d'établissements du champ de l'enquête. Pour tout $i \in U^B$, on définit $R_i(B) = card(F^{-1}(i))$, 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,

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

Un cas particulier de ces totaux est l'effectif de U^B, $N_B = card(U^B) = \sum_{i \in U^B} \mathbf{1}$.

Par exemple, T^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 T^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 :

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

où

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

5 Estimation sans biais d'un total

Dans le paragraphe précédent, nous avons 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 comme on l'a démontré dans la section 2.

Pour alléger les notations, on ne fait pas apparaître 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 T^B sans biais par :

$$\widehat{T}^B = \sum_{i \in s^B} w_i y_i \quad (5.1)$$

où

$$w_i = \frac{\sum_{l=1}^{4} \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{d_l}} \delta_j}{R_i(B)}.$$

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 ménage touristique i et enfin $\delta_j = 1/\pi_j^A$.

6 Cas particulier de certains sites : les points de visite en rase campagne

Dans certains sites, on ne connaît malheureusement pas le nombre total de personnes venant sur le site. En effet, dans l'ensemble A_4, on ne connaît pas tous les services (ici le nombre de visites) de la population. On ne peut donc pas avoir directement $\pi_j^{A_4}$ et donc δ_j pour $j \in A_4$. Pour contourner ce problème, on estime alors le nombre de visiteurs journaliers afin de déduire $\widetilde{\pi}_j^{A_4} = n_{A_4}/\widehat{T}_P^{A_4}$.

Dans la suite, nous allons développer 2 approches d'estimation du nombre de visiteurs journaliers. 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 interogé qui donne le nombre de personnes qui voyagent avec lui dans la voiture.

6.1 Construction d'un estimateur du nombre de visiteurs à partir d'un échantillonage de voitures

Dans ce paragraphe, nous sommes dans le cas où un enquêteur relève en "bâtonnant" 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 est connu aux erreurs de mesure près.

6.1.1 Définition de $\widehat{T}_P$

Soit T_V le nombre total de voitures défini par

$$T_V = \sum_{k=1,\ldots} t_k, \tag{6.1}$$

où t_k représente le nombre de voitures transportant k personnes. On peut également définir T_V par l'égalité suivante

$$T_V = \sum_{k \in U_V} \mathbf{1}, \tag{6.2}$$

où U_V désigne l'univers des voitures.

Remarque 6.1. Le nombre total de voitures T_V est considéré comme connu parcequ'il est donné par un distributeur mécanique.

Soit T_P le nombre total de personnes visitant le site défini par

$$T_P = \sum_{k=1,\ldots} k t_k. \tag{6.3}$$

Comme dans (6.2), on peut remarquer que le nombre total des personnes T_P est donné par :

$$T_P = \sum_{l \in U_P} \mathbf{1}, \tag{6.4}$$

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

$$T_P = \sum_{l \in U_V} v_l \tag{6.5}$$

où v_l est le nombre de personnes dans la voiture l. Comme nous l'avons mentionné en début de section, le nombre total de personnes T_P est inconnu. Par conséquent construisons un estimateur de T_P. Soit $\widehat{T}_P$ le π-estimateur défini par

$$\widehat{T}_P = \sum_{l \in s_V} w_l v_l, \tag{6.6}$$

où s_V est un échantillon de voitures de taille n et le poids w_l est égal à T_V/n, ce qui permet d'écrire l'estimateur $\widehat{T}_P$ sous la forme suivante

$$\widehat{T}_P = \frac{T_V}{n} \sum_{l \in s_V} v_l = T_V \overline{v}, \tag{6.7}$$

en posant $\overline{v} = \left(\sum_{l \in s_V} v_l \right) \Big/ n.$

Il est clair que $\widehat{T}_P$ est un estimateur sans biais du nombre total de personnes T_P.

6.1.2 Calcul de la variance de l'estimateur $\widehat{T}_P$ dans le cas d'un échantillonnage de voitures

On veut calculer la variance de l'estimateur $\widehat{T}_P$. Dans le cas présent, on assimile l'échantillon s_V à un sondage aléatoire simple sans remise. Par conséquent, on a

$$\begin{aligned} \mathrm{Var}[\widehat{T}_P] &= T_V^2 \left(\frac{1}{n} - \frac{1}{T_V} \right) S_V^2 \\ &= \frac{1}{n} T_V^2 S_V^2 - T_V S_V^2, \end{aligned} \tag{6.8}$$

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

6.1.3 Construction d'un estimateur d'une variable d'intérêt dans le cas d'un échantillonnage de voiture

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

$$Y = \sum_{i \in U_P} y_i, \tag{6.9}$$

où y_i est la variable d'intérêt qu'on mesure dans le questionnaire final du ménage touristique i. Soit $\widehat{Y}$ le π-estimateur défini par :

$$\widehat{Y} = \sum_{i \in s_P} w_i y_i, \tag{6.10}$$

où le poids w_i est égal à $\widehat{T}_P/m$. Par conséquent l'estimateur $\widehat{Y}$ peut s'écrire :

$$\widehat{Y} = \frac{\widehat{T}_P}{m} \sum_{i \in s_P} y_i = \widehat{T}_P \overline{y} \tag{6.11}$$

en posant $\overline{y} = \left(\sum_{i \in s_P} y_i \right) \Big/ m.$

6.1.4 Calcul de la variance de l'estimateur $\widehat{Y}$ dans le cas d'un échantillonnage de voitures

Il faut noter que les calculs développés par la suite, sont réalisés sous l'hypothèse que les variables $\widehat{T}_P$ et $\overline{y}$ sont indépendantes. L'hypothèse est réalisable. En effet, sur le terrain, c'est exactement cette situation qui se déroulera puisque

nous avons recours à 2 enquêteurs indépendants.

6.1.4.a Cas général

Calcul de la variance de l'estimateur $\widehat{Y}$:

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

$$\begin{aligned} V_Y &= \mathrm{Var}\left[\widehat{Y}\right] \\ &= \overline{Y}^2 \mathrm{Var}\left[\widehat{T}_P\right] + T_P^2 \mathrm{Var}\left[\overline{y}\right] \\ &\quad + \mathrm{Var}\left[\widehat{T}_P\right] \mathrm{Var}\left[\overline{y}\right]. \end{aligned} \tag{6.12}$$

Dans le cas présent, on assimile l'échantillon à un sondage aléatoire simple sans remise. L'égalité (6.12) devient alors

$$\begin{aligned} V_Y &= \overline{Y}^2 \left(\frac{T_V^2}{n} S_V^2 - T_V S_V^2\right) \\ &\quad + T_P^2 \left(\frac{1}{m} S_{\overline{y}}^2 - \frac{1}{T_P} S_{\overline{y}}^2\right) \\ &\quad + \left(\frac{T_V^2}{n} S_V^2 - T_V S_V^2\right) \left(\frac{1}{m} S_{\overline{y}}^2 - \frac{1}{T_P} S_{\overline{y}}^2\right) \\ &= \left(\overline{Y}^2 - \frac{1}{T_P} S_{\overline{y}}^2\right) T_V^2 S_V^2 \frac{1}{n} \\ &\quad + \left(T_P^2 - T_V S_V^2\right) S_{\overline{y}}^2 \frac{1}{m} \\ &\quad + T_V^2 S_V^2 S_{\overline{y}}^2 \frac{1}{nm} + \frac{T_V}{T_P} S_V^2 S_{\overline{y}}^2 \\ &\quad - \overline{Y}^2 T_V^2 S_V^2 - T_P S_{\overline{y}}^2. \end{aligned} \tag{6.13}$$

Optimisation de la variance de l'estimateur $\widehat{Y}$ dans le cas général :

Maintenant, l'étape est de chercher l'allocation des tailles des échantillons s_P et s_V qui minimise la variance de l'estimateur $\widehat{Y}$ pour des tailles de population T_P et T_V fixées.

On doit donc minimiser la quantité suivante

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

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

$$\begin{aligned} \mathcal{L}(n, m, \lambda) &= \left(\overline{Y}^2 - \frac{1}{T_P} S_{\overline{y}}^2\right) T_V^2 S_V^2 \frac{1}{n} \\ &\quad + \left(T_P^2 - T_V S_V^2\right) S_{\overline{y}}^2 \frac{1}{m} \\ &\quad + T_V^2 S_V^2 S_{\overline{y}}^2 \frac{1}{nm} + \frac{T_V}{T_P} S_V^2 S_{\overline{y}}^2 \\ &\quad - \overline{Y}^2 T_V^2 S_V^2 - T_P S_{\overline{y}}^2 \\ &\quad + \lambda \left(C_V n + C_P m - C\right). \end{aligned} \tag{6.14}$$

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

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial n}(n, m, \lambda) &= \left(\overline{Y}^2 - \frac{S_{\overline{y}}^2}{T_P}\right) T_V^2 S_V^2 \left(-\frac{1}{n^2}\right) \\ &\quad + T_V^2 S_V^2 S_{\overline{y}^2}^2 \left(-\frac{1}{mn^2}\right) \\ &\quad + \lambda C_V = 0, \\ \frac{\partial \mathcal{L}}{\partial m}(n, m, \lambda) &= \left(T_P^2 - T_V S_V^2\right) S_{\overline{y}}^2 \left(-\frac{1}{m^2}\right) \\ &\quad + T_V^2 S_V^2 S_{\overline{y}}^2 \left(-\frac{1}{nm^2}\right) \\ &\quad + \lambda C_P = 0, \\ \frac{\partial \mathcal{L}}{\partial \lambda}(n, m, \lambda) &= C_V\, n + C_P\, m - C = 0. \end{aligned}$$

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

$$\begin{aligned} &\lambda C_V^2 n^3 - \lambda C_V C n^2 \\ &- C_V T_V^2 S_V^2 \left(\overline{Y}^2 - \frac{1}{T_P} S_{\overline{y}^2}^2\right) n \\ &+ T_V^2 S_V^2 \left(C \left(\overline{Y}^2 - \frac{1}{T_P} S_{\overline{y}^2}^2\right) + C_P S_{\overline{y}}^2\right) = 0. \end{aligned}$$

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 :

$$\begin{aligned} &\lambda C_P^2 m^3 - \lambda C_P C m^2 \\ &- S_{\overline{y}}^2 C_P \left(T_P^2 - T_V S_V^2\right) m \\ &+ S_{\overline{y}}^2 \left(C(T_P^2 + T_V S_V^2) + C_V T_V^2 S_V^2\right) = 0. \end{aligned}$$

Remarque 6.2. Un autre cas : on assimile l'échantillonnage à un sondage aléatoire simple avec remise. Par conséquent l'égalité (6.12) devient alors

$$V_Y = \overline{Y}^2 T_V^2 \frac{\sigma_V^2}{n} + T_V^2 \frac{\sigma_{\overline{y}}^2}{m} + T_P^2 \frac{\sigma_V^2}{n}\frac{\sigma_{\overline{y}}^2}{m}.$$

On peut procéder à la même démarche que précedemment, c'est-à-dire rechercher l'allocation des tailles des échantillons en minimisant la variance de $\widehat{Y}$, mais la conclusion est la même, une équation du troisième degré à résolution numérique.

6.4.1.b Cas simplifié

Pour remédier au problème, nous pouvons faire une approximation dans l'égalité (6.13). En effet, nous pouvons supposer que le terme $1/nm$ est négligeable devant les termes $1/n$ et $1/m$. Cette hypothèse nést pas absurde puisque n et m peuvent prendre des grandes valeurs.

Calcul de la variance de l'estimateur $\widehat{Y}$:

Par conséquent, nous obtenons alors la transformation suivante de l'égalité (6.13)

$$\begin{aligned} V_Y &= \left(\overline{Y}^2 - \frac{1}{T_P} S_{\overline{y}}^2\right) T_V^2 S_V^2 \frac{1}{n} \\ &\quad + \left(T_P^2 - T_V S_V^2\right) S_{\overline{y}}^2 \frac{1}{m} \\ &\quad + \frac{T_V}{T_P} S_V^2 S_{\overline{y}}^2 - \overline{Y}^2 T_V^2 S_V^2 \\ &\quad - T_P S_{\overline{y}}^2. \end{aligned} \tag{6.15}$$

Optimisation de la variance de l'estimateur $\widehat{Y}$ dans le cas simplifié

Maintenant l'étape est de chercher l'allocation des tailles des échantillons s_P et s_V qui minimise la variance de l'estimateur $\widehat{Y}$ pour des tailles de population T_P et T_V fixées.

On doit donc minimiser

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

en n, m sous la contrainte

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

On peut écrire l'équation lagrangienne

$$\begin{aligned} \mathcal{L}(n,m,\lambda) &= \left(\overline{Y}^2 - \frac{1}{T_P} S_{\overline{y}}^2\right) T_V^2 S_V^2 \frac{1}{n} \\ &\quad + \left(T_P^2 - T_V S_V^2\right) S_{\overline{y}}^2 \frac{1}{m} \\ &\quad + \frac{T_V}{T_P} S_V^2 S_{\overline{y}}^2 - \overline{Y}^2 T_V^2 S_V^2 \\ &\quad - T_P S_{\overline{y}}^2 \\ &\quad + \lambda \left(C_V n + C_P m - C\right) \end{aligned} \tag{6.16}$$

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

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial n}(n,m,\lambda) &= \left(\overline{Y}^2 - \frac{S_{\overline{y}}^2}{T_P}\right) T_V^2 S_V^2 \left(-\frac{1}{n^2}\right) \\ &\quad + \lambda C_V = 0, \\ \frac{\partial \mathcal{L}}{\partial m}(n,m,\lambda) &= \left(T_P^2 - T_V S_V^2\right) S_{\overline{y}}^2 \left(-\frac{1}{m^2}\right) \\ &\quad + \lambda C_P = 0, \\ \frac{\partial \mathcal{L}}{\partial \lambda}(n,m,\lambda) &= C_V\, n + C_P\, m - C = 0. \end{aligned}$$

Après calculs, on obtient

$$\begin{aligned} n &= \frac{C}{\left(C_V + \sqrt{C_P C_V \frac{T_P S_{\overline{y}^2}\left(T_P^2 - T_V S_V^2\right)}{T_V^2 S_V^2 \left(T_P \overline{Y}^2 - S_{\overline{y}}\right)}}\right)}, \\ m &= \frac{C}{\left(C_P + \sqrt{C_P C_V \frac{T_V^2 S_V^2 \left(T_P \overline{Y}^2 - S_{\overline{y}}\right)}{T_P S_{\overline{y}^2}\left(T_P^2 - T_V S_V^2\right)}}\right)}. \end{aligned}$$

6.2 Construction d'un estimateur du nombre de visiteurs à partir d'un échantillonage 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 j le nombre u_j de passagers de la voiture i qui l'a transportée. Ce nombre u_j est ici égal à v_l.

6.2.1 Définition de $\widehat{\widehat{T}}_P$

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

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

Soit $\overline{v}$ le nombre moyen de passagers dans une voiture défini par

$$\overline{v} = \frac{\sum_{k \in U_V} k t_k}{\sum_{k \in U_V} t_k} = \frac{\sum_{k \in U_P} M_k}{\sum_{k \in U_P} M_k/k}, \tag{6.17}$$

où M_k désigne le nombre de personnes venues dans une voiture à k passagers.

Cette dernière définition permet de donner une dernière écriture de T_P

$$T_P = T_V \overline{v}. \tag{6.18}$$

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

$$\widehat{\widehat{T}}_P = T_V \widehat{\overline{v}}, \tag{6.19}$$

où le nombre total de voitures T_V est parfaitement connu. En observant cette expression, on constate que pour connaître $\widehat{\widehat{T}}_P$, il suffit de déterminer $\widehat{\overline{v}}$. Introduisons alors un estimateur de $\overline{v}$

$$\widehat{\overline{v}} = \frac{\sum_{k \in s_P} m_k}{\sum_{k \in s_P} m_k/k},$$

où m_k est le nombre de personnes de l'échantillon voyageant dans une voiture à k passagers. $\widehat{\overline{v}}$ peut s'écrire également de la façon suivante

$$\widehat{\overline{v}} = \frac{\sum_{j \in s_P} \mathbf{1}}{\sum_{j \in s_P} 1/u_j}$$

ou encore

$$\widehat{\overline{v}} = \frac{m}{\sum_{j \in s_P} 1/u_j}. \tag{6.20}$$

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

$$\frac{1}{\widehat{\overline{v}}} = \frac{1}{m} \sum_{j \in s_P} \frac{1}{u_j}. \tag{6.21}$$

Cette dernière quantité représente la moyenne empirique des $\dfrac{1}{u_j}$. On peut d'ailleurs calculer sa variance qui est égale à

$$\mathrm{Var}\left[\frac{1}{\widehat{\overline{v}}}\right] = \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \tag{6.22}$$

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

Reste à calculer la variance de $\widehat{\overline{v}}$ sachant (6.22). Pour cela, remarquons que l'on peut écrire

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

Par conséquent, on obtient

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

Finalement, on a

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

ou encore, avec (6.22)

$$\mathrm{Var}\left[\widehat{\overline{v}}\right] \simeq \overline{v}^4 \times \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \tag{6.23}$$

Or par définition, $S_{1/u}$ est égale à

$$S^2_{1/u} = \frac{1}{T_P - 1} \sum_{j \in U_P} \left(\frac{1}{u_j} - \frac{1}{\overline{v}}\right)^2. \tag{6.24}$$

Comme T_P est inconnu, cette formule peut être estimée par :

$$\frac{1}{m-1} \sum_{j \in s_P} \left(\frac{1}{u_j} - \frac{1}{\overline{v}}\right)^2. \tag{6.25}$$

Grâce à (6.23) et (6.25) on peut donc connaître facilement la variance de $\widehat{\overline{v}}$ et par conséquent celle de $\widehat{\widehat{T}}_P$ et celle de $\widehat{Y}$.

Remarque 6.3. L'estimateur $\widehat{\widehat{T}}_P$ est biaisé et asymptotiquement sans biais.

Remarque 6.4. Si les variables $\widehat{\widehat{T}}_P$ et $\overline{y}$ ne sont pas indépendantes alors on aurait

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

6.3 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 1 personne, 20% de voitures à 2 personnes, 20% de voitures à 3 personnes, 20% de voitures à 4 personnes, 20% de voitures à 5 personnes. Ainsi, on a 300 visiteurs sur ce site. La variance S_V^2 est égale à 2 en négligeant les corrections de population finie. Le nombre moyen de passagers $\overline{v}$ est de 3. En effet, on a

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

D'où $\overline{v}=3$.
Calculons maintenant une estimation de $S_{1/u}^2$. Après simplifications de (6.24) et en supposant que T_P est suffisamment grand devant 1, on a

$$S_{1/u}^2 = \frac{1}{T_P}\sum_{j\in U_P}\frac{1}{u_j^2} - \left(\frac{1}{\overline{v}}\right)^2.$$

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 $\mathrm{Var}[\widehat{\overline{v}}]$. Ainsi on a

$$\mathrm{Var}[\widehat{\overline{v}}] \simeq 3^4\times\frac{37}{30^2}\times\frac{1}{m}.$$

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 no84-85-86.

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 dénquête vol. 21, p.27-35.

LAVALLÉE, P. (2002) : "Le Sondage Indirect, ou la méthode généralisée du partage des poids", 'Editions de l'Université de bruxelles, Bruxelles.